FACULTY
OF APPLIED SCIENCES
UNIVERSITY
OF WEST BOHEMIA

University of West Bohemia
Department of Computer Science and Engineering
Univerzitní 8
30614 Plzeň
Czech Republic

# Influencing an artificial conversational entity by information fusion

Technical report

Ing. Jaromír Salamon

Technical report No. DCSE/TR-2020-04
June 2020
Distribution: public

Technical report No. DCSE/TR-2020-04
June 2020

# Influencing an artificial conversational entity by information fusion

Technical report

**Ing. Jaromír Salamon**

## Abstract

The dissertation thesis proposes a new method of using an artificial conversational enti(later also dialogue system or chatbot) influenced by information or information fusion. This new method could potentially serve various purposes of use like fitness and well-being support, mental health support, mental illness treatment, and a like. To propose such a new method, one has to investigate two main topics. Whether it is possible to influence the dialogue system by information (fusion) and what kind of data needs to be collected and prepared for such influence.

The dialogue system uses textual data from conversation to determine the context of human interaction and decide about next response. It presents correct behavior without external influence with data, but with the influence, the dialogue system needs to react without hiccups in the conversation adequately.

The data which could be used for the dialogue system influencing can be a combination of qualitative measure (from text extracted sentiment, from voice determined tone, from face revealed emotion), and measured quantitative values (wearable measured heartbeat, on-camera correctly performed exercise, based on EEG found focus on activity).

All the research found in the relation of those two topics is described in next more than 200 pages. It is supported with more than 500 references from former ones up to the most recent, including elementary solutions up to state of the art.

Copies of this report are available on
`http://www.kiv.zcu.cz/en/research/publications/`
or by surface mail on request sent to the following address:

University of West Bohemia
Department of Computer Science and Engineering
Univerzitní 8
30614 Plzeň
Czech Republic

Copyright © 2020 University of West Bohemia, Czech Republic

# Acknowledgements

First of all, I would like to express my endless gratitude to my doctoral study supervisor, Dr. Roman Mouček. He is supporting all my ideas and he is patient with me all the time. Moreover, he is willing to read and review all the text several times over and over.

Thanks should be also given to several of my students who either worked on the topics related to my research or on the independent works specified by me. They brought me a comprehensive perspective about close research topics or use and exploration of the data which were collected during my research.

Specifically, I would like to thank to:

- Tomáš Šimandl to give me an insight into the precision of the wearable devices measuring Heart Rate (HR) when I was leading and adjusting the way of his work on his bachelor thesis [1].
- Lukáš Lihl and the team of fellow students who implemented the extraction of Twitter and Fitbit data via API as their semester assignment.
- Radek Juppa who worked on the proof of concept idea related to data streaming into a cloud storage during his work on the semester assignment.
- Milan Kuda who processed the data collected during the Quasi-experiment (QX) and used them as an input to machine learning methods. I was leading and supervising his diploma thesis [2] on the topic of extracting sentiment from the HR.
- Štěpán Ševčík and Milan Tušl who were challenged with the topic of influencing a chatbot with external data during their semester assignment.
- Jakub Frank, Martin Ryba, and Petr Vintr who worked on the semester assignment dealing with ANT+ communication and translation of HR data into a reading form.
- Martin Ryba who is designing and implementing the idea of arm rehabilitation when video motion detection is used.
- Ahmad Aldin Yusmar, the internship student from Universiti Teknologi PETRONAS, Malaysia, who continually challenged my dialogue system influencing ideas.

Last but not least, I would like to express big thanks to my family, especially my wife Iva who supported me during the whole time when I was writing this thesis and also to our two children (Maya and Filip) who kept me busy and let me relax from working.

# List of Tables

# List of Figures

# List of Codes

# List of Acronyms

**AAAI** Association for the Advancement of Artificial Intelligence. 15

**ADEM** Automatic Dialogue Evaluation Model. 137

**AI** Artificial Intelligence. 10, 19, 74, 80, 83, 87, 131, 132, 140–142, 145, 157, 163

**AIML** Artificial Intelligence Markup Language. viii, 6, 12, 66, 68, 97, 98, 159

**AL** Active Learning. xviii, 100, 101, 105, 117, 159. Adversarial Learning. xviii, 84, 100, 101, 106, 109, 112, 114, 117, 159.

**ALBERT** A lite BERT. 92

**ALICE** Artifcial Linguistic Internet Computer Entity. 6, 97

**AMT** Amazon Mechanical Turk. 68, 77, 79, 139

**ANN** Artificial Neural Network. vi, xviii, 7, 69, 81–88, 94, 97, 100–102, 104, 108, 109, 111, 112, 124, 140, 141, 163

**AoP** Attention over Parameters. 115, 117

**APC** Alexa Prize Challenge. iv, 4, 6, 10, 12–14, 54, 58, 80, 104, 115, 117

**API** Application Programming Interface. 70

**ASR** Automatic Speech Recognition. 54, 77, 152–154

**ATIS** Air Travel Information System. 77

**BERT** Bidirectional Encoder Representations from Transformers. iv, 35, 69, 89–93, 95, 98, 101, 104, 105, 109, 142

**BLEU** BiLingual Evaluation Understudy. 20, 135, 137, 140

**BoW** Bag-of-Words. 86

**CAP** Credit Assignment Path. 83

**CapsNN** Capsule Neural Network. 84, 109

**CBT** Cognitive Behavioral Therapy. vi, 18–21, 125, 156, 157

**CIDEr** Consensus-based Image Description Evaluation. 140

**CNN** Convolutional Neural Network. 14, 68, 83, 84, 89, 90, 99, 100, 108, 109, 112

**ConvAI** The Conversational Intelligence Challenge. xv, 4, 10, 16, 20, 77, 114, 139, 140

**CoVe** Contextual Vectors. 87, 104

**CPA** Cost per acquisition. 29

**CPS** Conversation-turns Per Session. 6, 137

**CRF** Conditional Random Fields. 86, 108, 111

**CTRL** Conditional TRansformer Language. 92

**CX** Customer Experience. xvii, 67, 144

**DA** Dialogue Act. 62–64, 66, 67, 103, 104, 111

**DL** Deep Learning. xviii, 8, 10, 35, 56, 57, 82, 100, 102, 106, 109, 117, 163

**DM** Dialogue Management. xviii, 9, 13, 14, 20, 51, 53, 54, 60, 61, 64, 67, 76, 79, 103, 107–109, 113, 115, 122, 123, 138, 139

**DP** Dialogue Policy. vi, 7, 53, 62, 64, 103, 104, 106–111, 159

**DST** Dialogue State Tracker. vi, xvii, 9, 53, 62–64, 77, 78, 101, 105, 107–111

**DSTC** Dialog System Technology Challenge. xv, 4, 9, 10, 15, 75, 77, 79, 80, 101, 102, 159

**E2E** End-to-End. v, xvii, 7, 15, 16, 20, 51, 53, 55, 56, 72, 85, 97, 100–104, 106, 108, 111, 113, 119, 122, 124, 144, 159, 162, 163

**EEG** Electroencephalogram. 152

**ELMo** Embeddings from Language Models. 69, 88, 98, 104

**ER** Emotion Regulation. xv, xix, 20–22, 25, 30, 73, 78, 125–129, 150

**ERNIE** Enhanced Representation through kNowledge IntEgration. 91

**FCNN** Fully Connected Neural Network. 14

**FFNN** Feed Forward Neural Network. 83, 86, 87, 111

**FOH** First-order Hold. v, 41, 42, 50

**GAN** Generative Adversarial Network. 84, 163

**GloVe** Global Vectors. 87, 98

**GLUE** General Language Understanding Evaluation. 20, 90–92, 138, 139

**GNN** Graph Neural Network. 104

**GOFAI** Good Old-Fashioned Artificial Intelligence. 83

**GPT** Generative Pre-Training. 80, 89–92, 101, 102, 104, 160

**GPU** Graphics Processing Unit. 82, 93

**GRU** Gated Recurrent Unit. 67, 82, 85, 88, 99, 100, 112

**HCN** Hybrid Code Network. vi, 102, 106, 107, 117, 159

**HIS** Hidden Information State. 111

**HR** Heart Rate. iii, iv, xv, xvi, 5, 24, 25, 33, 36, 37, 39–50, 127, 147, 150, 151, 155

**HRED** Hierarchical Recurrent Encoder-Decoder. 100

**HRV** Heart Rate Variability. xv, 24, 25

**IEEE** Institute of Electrical and Electronics Engineers. 15

**IoT** Internet of Things. 28

**IR** Information Retrieval. xviii, 96, 98

**IWSDS** International Workshop on Spoken Dialog System. 15

**JSON** JavaScript Object Notation. 98

**LASER** Language - Agnostic SEntence Representations. 117

**LDA** Latent Dirichlet Allocation. 86

**LIME** Local Interpretable Model-agnostic Explanations. 141

**LSA** Latent Semantic Analysis. 86, 135, 136

**LSTM** Long / Short Term Memory. 12, 14, 66, 67, 82, 85, 87, 88, 99, 102–104, 108, 109, 112, 127, 163

**MAML** Model-Agnostic Meta-Learning. 113

**MDP** Markov Decision Process. 9, 103

**METEOR** Metric for Evaluation of Translation with Explicit ORdering. 136, 140

**ML** Machine Learning. 35, 56, 74, 83, 97, 106, 140, 145

**MLP** Multilayer Perceptron. 99, 104

**MMBT** MultiModal BiTransformer. 92

**MOOC** Massive Open Online Course. 3, 161

**MT** Machine Translation. 87

**NER** Named Entity Recognition. 15, 60, 68, 69, 139

**NIPS** Neural Information Processing Systems. 10, 15, 16

**NIST** National Institute of Standards and Technology. 135, 140

**NLG** Natural Language Generation. vi, viii, xvii, xviii, 9, 10, 17, 20, 51, 53, 54, 62, 66, 67, 78, 79, 86, 92, 106–108, 111, 112, 122, 124, 126, 135, 138, 140, 150, 152, 153

**NLI** Natural Language Inference. 138

**NLM** Natural Language Modeling. xviii, 81, 82, 86, 100, 104

**NLP** Natural Language Processing. 1, 29, 34, 35, 47, 51, 55, 60, 61, 68, 69, 73, 74, 78, 80–82, 86, 88, 90, 92–94, 104, 125–129, 138, 139, 162

**NLTK** Natural Language Toolkit. 35, 68

**NLU** Natural Language Understanding. vi, xvii, xviii, 9, 14, 15, 17, 18, 51, 53, 54, 60, 61, 63, 68, 69, 77, 79, 84, 86, 92, 103, 106–109, 122, 123, 126, 127, 138, 139, 150, 152, 153, 159, 162

**NMT** Neural Machine Translation. 86, 100, 115, 128

**NX** Natural Experiment. 30

**OIML** Open Intent Markup Language. 66, 98

**OPUS** The Open Parallel Corpus. 77, 80

**P4P** Pay for placement. 29

**PLM** Pre-trained Language Model. xviii, 81, 90–92, 101, 102, 104, 109, 112, 114

**POMDP** Partially Observable Markov Decision Process. 103, 111

**PoS** Part-of-Speech. 69

**PPC** Pay per click. 29

**PPP** Pay per post. 29

**PX** Pilot experiment. xix, 23, 30, 31, 33, 36, 147–149, 151

**QA** Question-Answering. 6, 10, 51, 58, 75, 80, 86, 96, 97, 99, 115

**QX** Quasi-experiment. iii, xix, 23, 30, 31, 33, 36, 147–149

**RACE** ReAding Comprehension Examinations. 75, 90–92, 138, 139

**RecNN** Recursive Neural Network. 84, 109

**RL** Reinforcement Learning. xviii, 9, 64, 82, 100–104, 106, 108, 111–114, 117, 127, 159

**RNN** Recurrent Neural Network. 13, 18, 34, 63, 66, 67, 82–85, 87–91, 99, 100, 103, 106–109, 111, 127, 163

**RO** Research Objective. xix, 2, 3, 20, 30, 32, 146, 149, 154, 155

**RoBERTa** Robustly optimized BERT approach. 91, 92

**ROUGE** Recall-Oriented Understudy for Gisting Evaluation. 136, 140

**RQ** Research Question. xix, 2, 3, 146, 154, 155

**RS** Response Selection. vi, xviii, 96, 98, 99, 106

**RUC** Research Use Case. xix, 3, 149–152, 155

**SAN** Sequential Attention Network. 99

**Seq2Seq** Sequence to Sequence. 18, 88, 89, 100, 112, 114, 117, 141

**SHA** Single Headed Attention. 89

**SHAP** SHapley Additive exPlanations. 141

**SIGdial** Special Interest Group on Discourse and Dialogue. 15

**SL** Supervized Learning. 64, 106, 111

**SMA** Simple Moving Average. 45, 46

**SMN** Sequential Matching Network. 99

**SMT** Statistical Machine Translation. 100, 115, 128

**SQuAD** Stanford Question Answering Dataset. 20, 75, 80, 90–92, 138, 139

**SSA** Simple String Accuracy. 135, 136. Sensibleness and Specificity Average. 7, 20, 101, 137.

**SuperGLUE** Super General Language Understanding Evaluation. 20, 138, 139

**SVM** Support vector Machine. 34, 108

**T5** Text-to-Text Transfer Transformer. 92

**Tf-Idf** Term frequency - Inverse document frequency. 86, 98

**TL** Transfer Learning. xviii, 87, 90, 92–94, 100, 101, 104, 112, 113

**TTS** Text to Speech. 54, 150, 152–154

**ULMFiT** Universal Language Model Fine-tuning. 69, 87, 104

**UX** User Experience. 144

**VHRED** Variational Hierarchical Recurrent Encoder-Decoder. 106, 159

**WER** Word Error Rate. 137

**Word2Vec** Word to Vector. 87, 98

**WOz** Wizard of Oz. xviii, 77, 79, 101

**XAI** Explainable AI. 140

**XLM** Cross-lingual Language Model. 90, 92

**XML** Extensible Markup Language. 97

**ZOH** Zero-order Hold. v, 41, 42, 50

**ZSL** Zero-Shot Learning. 104, 105

# Contents

# Chapter 1

# Introduction

This work serves as the author's overview of what existing research is suitable to offer the support of his research work (the dissertation). Nevertheless, during the last two years, the techniques, methods, and also technologies related to Natural Language Processing (NLP) and dialogue systems evolve drastically in a good manner. So, the work is trying to follow up this rise and build on top of that.

## 1.1 Motivation

The worldwide increase of various psychological disorders or mental diseases leads to the disproportion between the impacted population and available treatment. The treatment itself consumes lot of time, and some of the people are even not willing to admit they need some help. During last years, self-help-based intervention (books, DVDs, computer programs), replacement, or support of ambulatory treatment have been provided in contemporary research either by utilizing SMSs or dialogue systems. In all cases, the aim is to get an equivalent substitution of psychological or medical support.

Dialogue systems represent modern and positive way to solve this disproportion that is supported by the growing spread of mobile phones and installed applications in the population. We want to validate their usability for such an application. As a novel approach, we would like to combine dialogue model with additional information collected from wearables or otherwise during the user's conversation. Moreover, we would like to explore if this information can potentially help to orchestrate the conversation differently or provide some additional value to the conversation.

## 1.2 Problem Statement

For interventions that are led by a dialogue system, the usual process of usage is a conversation itself combined with some psychological approach.

When we compare such an approach with a standard conversational diagnostic approach led by a human doctor, we could see there parallels in the conversation, and its analysis via empathy and understanding. However, quantitative measurements, such as blood pressure, heart rate, and temperature, are missing.

From a high-level perspective, the gap can be filled in by switching between the **dialogue system module**, leading a standard dialog with human and the **influence module** affecting the human conversation in a positive way (Figure 1.1). The **quantitative measure** (for example the number of coins in the pocket or the age of the car) can be used either directly or pre-processed and combined with the **qualitative measure** (for instance the softness of a cat or the color of the sky) within the **data fusion module** in the way it represents a reasonable value to influence the conversation.

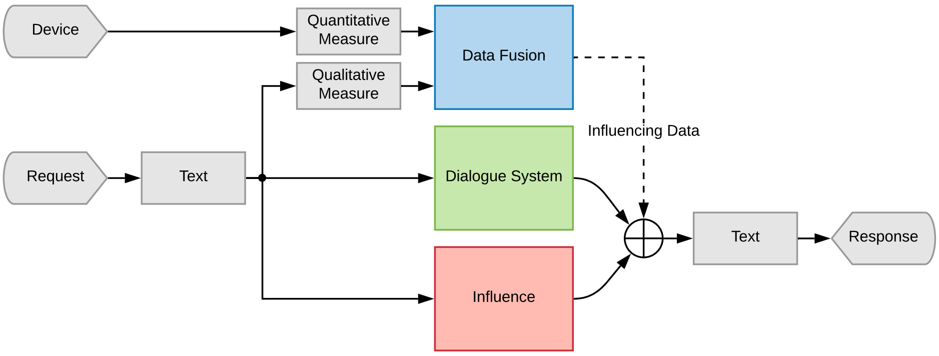


Figure 1.1: Introduction to the idea of dialogue system influencing

## 1.3 Structure of the thesis

The rigorous thesis is split into several parts and the sections which focus on particular topics and either act as an introduction to further topics or deal with one or several Research Questions (RQs).

The introductory section (§1) describes motivation and problem statement. Followed by state of the art (§2), which is a comprehensive overview of various methods that have been used for similar research purposes before and forms the elementary parts for Research Objective (RO).

The first part of the thesis is working with data. It starts with the chapter, which describes everything related to soft and hard data (§3). With the previous soft and hard data description, the fourth section focuses on the influencing data (§4), and its pre-processing. Furthermore, the next section is dedicated to data interpolation or discretization for later usage of fusion techniques (§5).

The second part of thesis is focusing almost purely on the dialogue systems. It starts with the introduction to dialogue systems (§6), where an overview of the main principles is. It continues through the inserted section about specific corpora (§7) organized by the application. Then it ends with detail description of dialogue system models (§8), which follows up where the previous introduction ends and extends the dialogue system basics with comprehensive methods.

The last part of thesis wraps up all the previous research overview. First, it uses every chapter conclusion and brings them into account to come with several design variants of the dialogue system influencing (§9). Such dialogue system influencing needs to be tested and evaluated (§10) appropriately from the technical solution and user experience perspectives.

The very last chapter deals with the research proposal (§11). The foremost it defines the Research Objective (RO) and relevant Research Questions (RQs) together with validation via Research Use Cases (RUCs). The RUCs are organized by the solutions complexity and potential limitations to support the Research Objective (RO) and answer the RQs.

The work has also two addendum's. The first appendix is about practical experience (§A) with existing chatbot solutions. The second one reflects the popularity of Massive Open Online Courses (MOOCs) (§B) related to dialogue systems and relevant study fields.

# Chapter 2

# State of the Art

The state of the art section presents various topics which are related to the thesis subject and touches the eminent and contemporary research in the field.

The idea of usage an intervention tool to support or replace ambulatory treatment and achieve better adherence and attrition is not new as stated in introduction (§1). So, such methods used in past years are part of replacement or support of ambulatory treatment (§2.1).

The emerging era of dialogue systems (§6) allows to use them for various purposes. So, first we do the review of the dialogue systems evolution (§2.2) followed by the state of the art (§2.3) in the field.

Dialogue systems are evolving also thanks to dialogue platform competitions (§2.4) like Loebner Prize (§2.4.1), Alexa Prize Challenge (APC) (§2.4.2), Dialog System Technology Challenge (DSTC) (§2.4.3) or The Conversational Intelligence Challenge (ConvAI) (§2.4.4).

Next to the competitions with common chatbot purpose, we can find the dialogue systems for specific purposes (§2.4.5) and last but not least as intervention tools for health and well-being (§2.4.6).

The dialogue systems testing (§2.5.1) went long way from Turing test up to the contemporary — in the could offered — test services. The same long way can be spotted in dialogue system evaluation (§2.5.2), which focuses not only on technical capabilities from measurable technical perspective, but also compares the machine with the humans.

Dialogue systems, when oriented to provide well-being or coaching functionality are usually introducing more or less psychological and psycho-social intervention methods (§2.6.1). Those methods require wide knowledge of psychology which is out of the thesis scope. For our purpose it would be good enough to introduce simple cognitive strategies which help to regulate emotions (§2.6.2).

Increasing capabilities of wearables allows to use them in health care and medical research (§2.7). Since the wearables provide more and more different measured data (§2.7.1) there is interest to quantify the devices precision (§2.7.2) and come with potential applications (§2.7.3).

Well-being starts when negative feelings are identified and eliminated. The discomfort caused by feeling the strain and pressure is called stress (we are considering the negative one). To identify stress (§2.8) it is necessary to follow up and measure the physiological markers (§2.8.1), for which the wearables with HR measure (§2.8.2) could serve.

## 2.1 Replacement or support of ambulatory treatment

The replacement or support of ambulatory treatment is under the eminent scientific research. The usual problem when it goes about any treatment is patient adherence to any activity or correct description of problems leading the system to correct diagnosis.

The early systems tent to heavily use the SMS as the modern communication channel with questionnaires as the assessment tool. Use of SMS can enhance adherence for treatment of schizophrenia by enhancing patients adherence to antipsychotic medication [3] and uses a questionnaire as the feedback from patients. There is another work that uses SMS to replace ambulatory treatment in patients with primary depression or alcohol [4] with feedback provided by assessment and survey. Another solution which utilizes SMS solution **ITAREPS** [5] serves for weekly remote monitoring schizophrenia and psychotic disorders with questionnaire as the feedback from patients.

With more significant capabilities of mobile phones also chatbot applications instead of SMS have been utilized to replace or support ambulatory treatments in recent years, mostly for well being. The specific examples, e.g. the help with weight reduction (**Nombot**) or treatment of people with symptoms of depression and anxiety (**Woebot**) are well described later in dialogue systems for well-being (§2.4.6). These dialogue system are still far from the perfection of full human intervention, but with the simplicity of use and 24x7 availability are broadly accepted as a suitable substitution. The statistics of results are incredible when **Woebot** significantly reduced the symptoms of depression in two weeks in a randomized controlled trial [6] at Stanford University.

## 2.2 Dialogue Systems Evolution

Like any human activity, even the dialogue systems evolved over the time. The next part is pointing out the most interesting evolution milestones which written the history in the field.

**1966 — ELIZA** [7] In 1966 created by Joseph Weizenbaum as a simulation of a Rogerian therapist. The program recognizes certain patterns (pattern matching) and keywords based on which generates appropriate responses.

**1972 — PARRY** [8] Written in 1972 by psychiatrist Kenneth Colby to simulate a person with paranoid schizophrenia. The program implemented a crude model of the behavior of a person with paranoid schizophrenia and also embodied a conversational strategy.

**1988 — Jabberwacky** Rollo Carpenter created it in 1988 as the chatbot, which simulates natural human chat in an interesting, entertaining, and humorous manner. It won in 2005 the Loebner Prize (§2.4.1) as character George and in 2006 as another character Joan.

**1995 — Artifcial Linguistic Internet Computer Entity (ALICE)** [9] Inspired by Joseph Weizenbaum's ELIZA chatbot Richard Wallace implemented ALICE. The chatbot utilizes the dialogue language which is an XML Schema called Artificial Intelligence Markup Language (AIML) specifying the heuristic conversation rules. It won the Loebner Prize (§2.4.1) three times in 2000, 2001 and 2004.

**2005 — Mitsuku** Worldwide popular[1] chatbot based on the AIML and implemented by Steve Worswick. The implementation contains not only all of ALICE's AIML files but also additional functionality, which includes the ability to reason objects, play the games, and do magic tricks. It is a five-time Loebner Prize (§2.4.1) winner (in 2013, 2016, 2017, 2018, 2019).

**2006 — Watson** IBM's QA system, which beat two former Jeopardy show champions.

**2010 — Siri** It is the assistant which was originally developed by the SRI International Artificial Intelligence Center as the spin-off Siri. It was acquired by Apple in 2010 and turned into a virtual assistant that is integrated into all Apple's computer and wearable operational systems.

**2012 — Now** Now was a feature of Google Search application used till October 2016 when **Google Assistant** replaced it.

**2014 — Alexa** It is a virtual assistant developed by Amazon. Amazon uses it in his smart speaker products Amazon Echo and the Amazon Echo Dot. To enhance Alexa's skills Amazon established Alexa Prize Challenge (APC) (§2.4.2) with a goal of building a socialbot.

**2014 - XiaoIce ("Little Ice" literally in Chinese)** [10] Is very popular[2] social and emphatic chatbot (IQ and EQ plays the role) deployed mostly in Asia. The XiaoIce has mutlimodal interface to receive users input like text, images and voice. It dispatch the input to the proper various modules through the chat manager such as core-chat or visual awareness which are part of various skill-set.

Microsoft defined a new metric Conversation-turns Per Session (CPS)[3](§10.3.2) as the metric for social chatbots evaluating the success and emotional engagement with users.

**2015 — Cortana** It was (till January 2020) a virtual assistant created by Microsoft for Windows operational systems, wearables, mobile phones, and other devices.

**2016 — Tay** It was a chatbot implemented by Microsoft and deployed to Twitter in March 2016. The chatbot was expected to interact with users and learn from those interactions, but during the next 16 hours began to post inflammatory and offensive tweets, and Microsoft shut down the service (§6.14.3).

[1] 29.11.2015: Mitsuku has had over 14 million interactions on Kik in just over two months.

[2] Since her launch in 2014, XiaoIce has communicated with over 660 million active users and succeeded in establishing long-term relationships with many of them.

[3] XiaoIce has achieved an average CPS of 23

**2016 — Zo** Zo is a Microsofts's successor to the Tay chatbot, which is the English version of Microsoft's other successful social chatbot XiaoIce (2014) introduced in China. Zo holds Microsoft's longest continual chatbot conversation: 1,229 turns, lasting 9 hours and 53 minutes (December 2016) [11]. Zo tries to avoid Tay's conversational mistakes by strong Dialogue Policy (DP), which rejects any debate about prohibited topics[4], which led to the pitfalls of Tay (§6.14.3).

There are other local modifications of XiaoIce introduced worldwide. In Japan, it is Rinna (2015), in India popular Ruuh and Indonesia Rinna again both launched in 2017.

**2020 - Meena** [12] Meena is most recent contribution to the neural-based dialogue systems. Its multi-turn open-domain chatbot and its End-to-End (E2E) neural conversational model is trained on 2.6 billion parameter Artificial Neural Network (ANN).

Google incorporated the incredible computational power[5] to train the model to minimize the perplexity of the next token. They also proposed a human evaluation metrics called Sensibleness and Specificity Average (SSA) (§10.3.2), which captures key elements of a human-like multi-turn conversation and strongly correlates with perplexity.

## 2.3 Dialogue Systems State of the Art

The long history of dialogue system evolution (§2.2) brings over the time several approaches how to deal with conversation and those are grateful topics of comprehensive papers or survey publications. These publications define their dialogue systems classification and describe the dialogue systems architecture (§6.3), either pipeline architecture (§6.3.1) or overall End-to-End (E2E) architecture (§6.3.2).

One of the sources is comprehensive Jurafsky's **Speech and Language Processing** [13], the 3rd draft of the book[6]. Dialogue systems and chatbots are divided into several groups in the chapter **Dialogue Systems and Chatbots** (Figure 2.1).

[4] https://qz.com/1340990/microsofts-politically-correct-chat-bot-is-even-worse-than-its-racist-one

[5] The model was trained for a whopping 30 days on a TPU v3 pod (2,048 TPU cores)

[6] https://web.stanford.edu/ jurafsky/slp3/ed3book.pdf

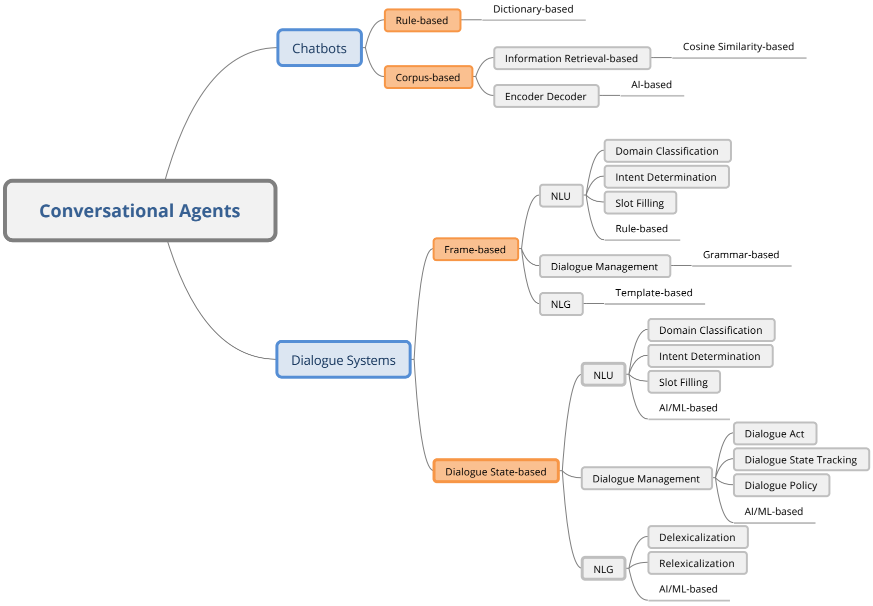


Figure 2.1: Dialogue Systems and Chatbots classification by Jurafsky et al.

Another book **Complex, Intelligent, and Software Intensive Systems** [14] with the section **Survey on Intelligent Chatbots: State-of-the-Art and Future Research Directions** classifies the chatbots in different way. It is a more shallow division to non-task-oriented (with retrieval-based and generation-based chatbots) and task-oriented (with supervised and unsupervised approaches).

Two tutorials present a comprehensive overview of dialogue systems and its classification focusing specifically on the Deep Learning (DL) dialogue systems. The first one is **Deep Learning for Dialogue Systems** [15] overview presented at various conferences like ACL 2017, IEEE ICASSP 2017 and INTERSPEACH 2017. The second one is **Deep Chit-Chat: Deep Learning for ChatBots** [16] presented at EMNLP 2018.

Next to the book and conference tutorial presentations there exist also surveys or various reviews of dialogue systems, chatbots, and architecture components (§6.3). Some of them have are excellent while some of them are of worse quality.

One of the surveys purely dedicated to the phenomenon of the contests, competitions or prizes (§2.4) is **A Survey of Chabot Systems through a Loebner Prize Competition** [17]. It brings no more than an overview and analysis of dialogue system techniques used by Loebner Prize (§2.4.1) winners.

From the classification perspective **A Survey on Dialogue Systems: Recent Advances and New Frontiers** by Chen et al. [18] is the most intriguing. It tries to categorize the dialogue systems in a new way (Figure 2.2) and adds neural-based models when compared to Jurafsky classification.

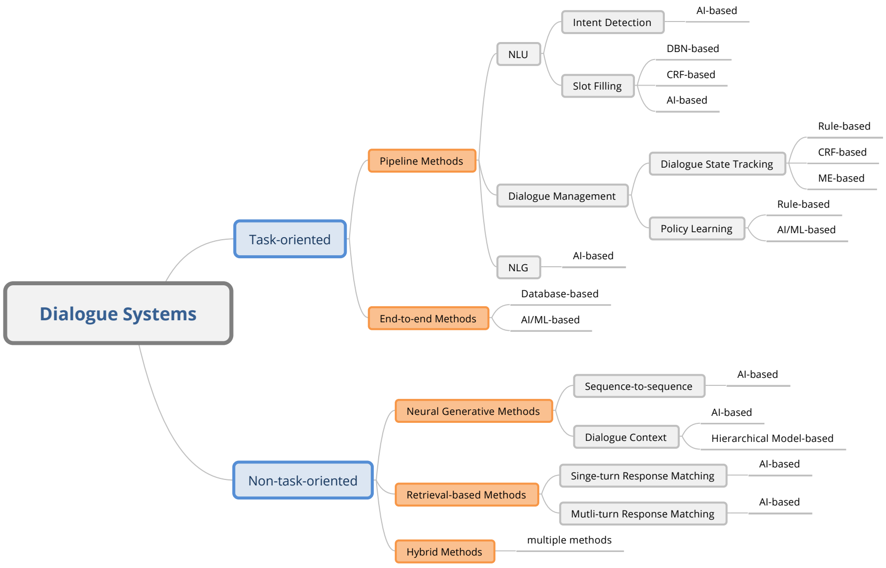


Figure 2.2: A Survey on Dialogue Systems classification by Chen et al.

The paper **Review of Research on Task-Oriented Spoken Language Understanding** [19] presents an introduction into the Natural Language Understanding (NLU) (§6.8) part of the dialogue system. It presents the independent models of slot filling (§6.8.4) and intent detection (§6.8.2) tasks and then also joint models for both tasks together.

A comprehensive review of Dialogue State Tracker (DST) is presented by Henderson in **Machine Learning for Dialog State Tracking: A Review** [20] provides the basic classification of particular DST methods and review of previous years of Dialog System Technology Challenge (DSTC).

Another DST evolution is nicely described by Williams with contribution from Henderson in **The Dialog System Technology Challenge (DSTC) Series: A Review** (§2.4.3) paper [21]. It gives reviews of method, challenge, data, and evaluation standardization.

A survey about connection between **Reinforcement Learning (RL) and Dialogue Management (DM) strategies** [22] describes DM approaches, strategies and evaluation in detail. It introduces common approaches about DM like management strategies, intitiative, and confirmation and then deep dives into RL (§8.6.2) with detailed explanation including the Markov Decision Process (MDP).

From the content and topic point of view there is an exhaustive **Survey of state of the art in NLG** (§6.10) [23]. It contains almost 120 pages related to the Natural Language Generation. It includes not only dialogue systems topics related to the generation of text but also topics about image captioning, generating text with style, personality, and affect or creative and entertaining text.

Much shorter is **A Survey of Natural Language Generation Techniques with**

**a Focus on Dialogue Systems - Past, Present and Future Directions** [24]. It describes the complete mechanics of NLG up to the detail realization from hand-crafted methods, over templates up to the statistical approaches including Deep Learning (DL).

Another comprehensive publication about NLG is **Evaluating the State-of-the-Art of End-to-End Natural Language Generation: The E2E NLG Challenge** [25]. It contains 80 pages describing the methods, datasets, and evaluation of NLG.

The comprehensive paper **A Survey of Available Corpora for Building Data-Driven Dialogue Systems** [26] is mainly focused on datasets, their description, usage, advantages and disadvantages. Moreover, it also presents the introduction to dialogue systems and their evaluation.

Except this specific one the pipeline module NLG evaluation paper [25] exists. **Survey on Evaluation Methods for Dialogue Systems** [27] leads the reader trough all the relevant topics to the dialogue systems evaluation including task-oriented, conversational dialogue, and QA systems.

## 2.4 Dialogue Systems Competitions

Behind the extensive development of dialogue system is standing research. With the increasing number of publications on such topic with generative methods (§6.4.2), it looks like dialogue based on Deep Learning (DL) (§8.6.1) seems to be promising and stable solution soon, but the Loebner Prize (§2.4.1) running since 1991 is still convincing us to the contrary. Also recently founded Alexa Prize Challenge (APC)[7] (§2.4.2) is not fully utilizing AI, but builds on top of the ensemble dialogue systems (§8.8.2) approach. Each of these competitions has defined its own state of the art category in dialogue system evolution.

There are two other contests focusing fully on a dialogue. The first one DSTC[8] (§2.4.3) is an on-going series of research community challenge tasks. The main subject is to create a tracker which is able to predict the dialogue state for new dialogues. Every year the challenge is oriented to a different main theme with data from a different conversation domain. In each challenge, trackers are evaluated using held-out dialogue data. The second one (ConvAI[9]) (§2.4.4) is the Neural Information Processing Systems (NIPS) conference competition track first presented in 2017 and continued in 2018. There are several tasks for which this competition is aiming at and these are: providing a dataset and making conversations more engaging for humans and simplifying the evaluation process (automatic evaluation, followed by the human evaluation). So, these contests bring another contribution to the state of the art of dialogue systems.

### 2.4.1 Loebner Prize Chatbots

The contest defined by Hugh Loebner and Cambridge Center for Behavioral Studies, Massachusetts, United States was founded in 1990 to evaluate chatbots based on the

[7] https://developer.amazon.com/alexaprize

[8] https://www.microsoft.com/en-us/research/event/dialog-state-tracking-challenge

[9] http://convai.io

application of the Turing Test [28].

The competition has been running since 1991 with a well known [17], [29] list of winners (Table 2.1). The rules changed slightly over the years, but the main goal has remained the same.

Annually, the awarded price bronze medal for the most human-seeming program in the competition is $2,000. Whenever the chatbot cannot be distinguished from humans the silver medal award of $25,000 is given and if the chatbots fully understand the text, audio and video input the reward of $100,000 is given to the author and the competition ends.

| Year | Winner | Design Specifics | Pattern Matching | AIML | Database | Word Vocabulary | WordNet | Ontology | Markov chains |
|---|---|---|---|---|---|---|---|---|---|
| 1991 | PC Therapist | | ● | | | ● | | | |
| 1992 | PC Therapist | | ● | | | ● | | | |
| 1993 | PC Therapist | | ● | | | ● | | | |
| 1994 | TIPS | personal history model | ● | | ● | | | | |
| 1995 | PC Therapist | | ● | | | ● | | | |
| 1996 | HeX | personal history model | ● | | ● | | | | ● |
| 1997 | Converse | | ● | | ● | | ● | ● | |
| 1998 | Albert One | | ● | | | | | | |
| 1999 | Albert One | | ● | | | | | | |
| 2000 | A.L.I.C.E. | | ● | ● | | | | | |
| 2001 | A.L.I.C.E. | | ● | ● | | | | | |
| 2002 | Ella | phrase normalization | ● | | | | ● | | |
| 2003 | Jabberwock | context free grammar | ● | | | | | | ● |
| 2004 | A.L.I.C.E. | | ● | ● | | | | | |
| 2005 | George (Jabberwacky) | | | | ● | | | | |
| 2006 | Joan (Jabberwacky) | | | | ● | | | | |
| 2007 | Ultra Hal | script | ● | | | | | | |
| 2008 | Elbot | commercial | | | | | | | |
| 2009 | Do-Much-More | commercial | | | | | | | |
| 2010 | Suzette | | ● | ● | ● | | | | |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 2011 | Rosette | | ● | ● | ● | | | |
| 2012 | Chip Vivant | artificial intelligence | | | | | | ● |
| 2013 | Mitsuku | | ● | ● | | | | |
| 2014 | Rose | | ● | | ● | | ● | ● |
| 2015 | Rose | | ● | | ● | | ● | ● |
| 2016 | Mitsuku | | ● | ● | | | | |
| 2017 | Mitsuku | | ● | ● | | | | |
| 2018 | Mitsuku | | ● | ● | | | | |
| 2019 | Mitsuku | | ● | ● | | | | |

Table 2.1: Loebner Prize Summary

The first main conclusion out of the Loebner Prize summary Table 2.1 is that pattern matching technique drives the design of chatbots in this competition. During the last 28 years we can see additional design approaches like database usage, phrase normalization, mostly also ontology and WordNet, context-free grammar and Markov chains. All these techniques are used to give a chatbot chance to recognize relations in the language and use it for a better understanding of the matched pattern.

### 2.4.2 Alexa Prize Challenge (APC)

Current solutions related to the APC are dealing with very complex issues to build a dialogue open system. As it will be described later on in chatbot introduction (§6), the open domain (§6.5.1) dialogue system is one of the biggest challenges and it is difficult to be achieved.

#### Overview of 2017

A summary of particular chatbots participating in APC is in Table 2.2, where the highlighted solutions took first three places and are described in more detail below the table.

| Program | Design Highlights |
|---|---|
| CMU Magnus [30] | maximal marginal relevance, finite state tranducer |
| Ruby Star [31] | confidence score (BoW), CoreNLP |
| **Alquist** [32] | structured topic dialogue, CoreNLP, YodaQA |
| Emersonbot [33] | Gradient Boosting (Word2Vec, TF-IDF), Yahoo, Wikipedia |
| **Alana** [34] | bot priority list, contextual ranking mechanism |
| Pixie [35] | confidence index, CoreNLP, Google knowledge graph |
| Wise Macaw [36] | AIML, seq2seq with LSTM trained on Twitter |

| | |
|---|---|
| Chatty Chat [37] | finite state machine or TF-IDF with SVN, Wikipedia |
| Eigen [38] | state machine modules, discriminator, transfer learning |
| SlugBot [39] | dialogue manager, confidence score, database, ELIZA |
| Edina[10] [40] | rule-based and generative (RNN), TF-IDF, self dialogue |
| MILA Team [41] | template, knowledge, search, retrieval, generation models |
| Roving Mind [42] | pipeline with modules, rule-based, database, CRF [11] |
| **Sounding Board** [43] | state-based dialogue model, TextRank (Gensim) |

Table 2.2: Alexa Prize Summary 2017

**Sounding Board** [43] It is a social bot from the University of Washington implemented from scratch as the contribution to Alexa Prize Challenge (APC). It employs a hierarchical dialogue manager of overall conversation and is supported by a collection of mini-skills to manage different conversation topics. The dialogue policy is strictly user driven multi-dimensional representation of utterance (§6.8.1) that includes user sentiment as well as intent detection (§6.8.2). During the conversation the chatbot detects user frustration to initiate a topic change.

**Alquist** [32] It is implemented by a research group from the Czech Technical University in Prague represents a dialogue system which (as the authors say) provides coherent and engaging conversation on various topics. It uses an advantage of two types of Dialogue Managements (DMs): top-level and topic-level. The top-level dialogue manager decides which module should be executed (chit-chat, question answering, topic dialogue, etc.). The topic-level dialoge manager switches to particular topics (sports, movies, etc.). Overall, the system combines machine learning modules and rule based modules for response generation.

**Alana** [34] It is a bot created at the Heriot-Watt University in Edinburgh. It takes an advantage from an ensemble of various agents/bots. These bots generate a pool of replies on which a ranker model is going to select the most relevant reply. The bots ensemble contains the following two categories of bots: Data-driven bots and Rule-based bots (persona bot, Eliza resp. its extension Rosie, news bot, fact bot, EVI, weather bot). The responses proposed by each bot are ranked according to a set of features by a hand-engineered ranker function and linear classifier ranker.

[10] Implemented in RiveScript (https://www.rivescript.com/)

[11] Conditional Random Fields

### Overview of 2018

Table 2.3 summarizes design used in particular solutions with highlighted first three places which are described in more detail below.

| Program | Design Highlights |
|---|---|
| EVE [44] | conversational scaffolding[12], knowledge graph, ScriptDog lang.[13] |
| Tartan [45] | retrieval based, dynamic Finite State Machine |
| **Alquist** [46] | ontology-based topic, LSTM-based model for DM[14] |
| Iris [47] | Mixture of Experts Model implemented by CNN and FCNN |
| **Alana** [48] | clarification questions, contextual NLU with FEL[15] |
| Fantom [49] | Evolving Dialog Graph context modeling |
| **Gunrock** [50] | context-aware hierarchical DM [16], LSTM for dialogue predicition |
| SlugBot [51] | Discourse relation dialogue model (DRDM) |

Table 2.3: Alexa Prize Challenge (APC) Summary 2018

**Gunrock** [50] It is a social bot designed to engage users in open domain conversations built by the University of California Davis. With an incredible effort dedicated to each part of the architecture pipeline it became the winner of Alexa Prize Challenge (APC) in 2018. First, they focus on the main chatbot architecture building blocks like a context-aware hierarchical Dialogue Management (DM) reacting and handle a wide variety of user behaviours (typically question answering and difficulties with topic switching). Secondly, they focus on error correction of automatic speech recognition (ASR) and developed sentence segmentation in NLU with a large data set.

**Alquist 2nd version** [46] It is developed by the team from the Czech Technical University in Prague won the second place in 2018 as well as it happened in 2017. They improved their original implementation with a system leveraging ontology-based topic structure called topic nodes. This is a major innovation when compared to the previous year utilized fixed tree structure for each topic node. These nodes consist of several sub-dialogues where each one of them utilizes a dialogue management model built with Long / Short Term Memory (LSTM). During the main dialogue the sub-dialogues based on the existing topic hierarchy or user intent can be triggered.

[12] a technique which uses a (small) conversational dataset to define a generalized response strategy

[13] https://github.com/BYU-PCCL/scriptdog

[14] dialogue management

[15] Fast Entity Linking

[16] dialogue manager

**Alana v2** [48] It provides improvement of the first version. It was developed by a team from Heriot-Watt University in Edinburgh. They focused on the improvement of Natural Language Understanding (NLU) part to generate clarification questions to disambiguate between Named Entity Recognition (NER) interactively. Another technique used to retrieve additional information associated with the entities is entity linking, resp. fast entity linking (FEL) system [52], [53]. It is fundamental for chatbots coherent conversation with the user about a specific topic. For individual bot responses improvement, they utilized data from the previous year competition and train BiLSTM classifiers. Last but not least, they introduced new Ontology, Abuse mitigation, and Reddit bots to improve the overall conversational engagement.

### 2.4.3 Dialog System Technology Challenge (DSTC)

The DSTC is an on-going series of research community challenges established in 2013 and first having been a part of the Special Interest Group on Discourse and Dialogue (SIGdial) conference and then Institute of Electrical and Electronics Engineers (IEEE), International Workshop on Spoken Dialog System (IWSDS), Neural Information Processing Systems (NIPS) or Association for the Advancement of Artificial Intelligence (AAAI) conferences in the next years.

The primary objective is to create a "tracker" that can predict the dialogue state for new dialogues. The task is every year driven by different data provided by various organizations (universities or research institutions) and contains diverse domains for particular dialogue system topic related to the specific year of challenge (Table 2.4).

| Challenge | Conference | Domain | Topic |
|---|---|---|---|
| DSTC1 [54] | SIGdial 2013 | Bus Timetable | Evaluation Metrics |
| DSTC2 [55] | SIGdial 2014 | Restaurant | User Goal Changes |
| DSTC3 [56] | IEEE SLT 2014 | Tourist Information | Domain Adaptation |
| DSTC4 [57] | IWSDS 2015 | Tourist Information | Human Conversation |
| DSTC5 [58] | IEEE SLT 2016 | Tourist Information | Cross-Lingual Adaptation |
| DSTC6 [59] | NIPS 2017 | Restaurant<br>OpenSubtitles<br>Twitter<br>Various Dialogues | E2E[17] Goal Oriented Dialogue [60]<br>E2E Conversation Modeling [61]<br>Dialogue Breakdown Detection [62] |
| DSTC7 [63] | AAAI 2019 | E2E Dialogue System | Noetic E2E Response Selection [64]<br>Grounded Response Generation [65]<br>AVSD[18] [66] |

[17] End-to-End

[18] Audio Visual Scene-aware Dialog

| | | | |
|---|---|---|---|
| DSTC8 [67] | NIPS 2019 | E2E Dialogue System | E2E Multi-Domain DS[19]<br>Fast Adaptation Task<br>Predicting Responses Track<br>AVSD |

Table 2.4: Dialogue System Technology Challenges

Each task released dialogue data labeled with dialogue state information. It is given by the dialogue history up to the current turn, for instance the user's desired restaurant search query. In each challenge, trackers are evaluated using held-out dialogue data.

### 2.4.4 The Conversational Intelligence Challenge (ConvAI)

The ConvAI focuses mainly on two topics (Table 2.5) which are essential for non-goal-oriented dialogue systems (chatbots). Gathering and preparing datasets for appropriate training chatbot models make conversations more engaging for humans. Standardizing chatbot models evaluation is equally problematic if not even more painful. It includes human evaluation (for instance Turing test (§10.2.1)) followed then by computed evaluation (for example evaluated by metrics §10.3.2).

| Challenge | Conference | Dataset | Metrics |
|---|---|---|---|
| ConvAI | NIPS 2017 | Human-to-Chatbot Dialogues [68] | MTurk [20] |
| ConvAI2 [69] | NIPS 2018 | Persona-Chat [70] | MTurk<br>Perplexity (PPL)<br>Hits@1<br>F1 |

Table 2.5: The Conversational Intelligence Challenge

#### Overview of 2017

The first year of competition evaluated chatbots only with the Amazon Mechanical Turk, which is the online service providing the individual evaluation. Table 2.6 presents the results of particular chatbots compared to the human rating representing the baseline.

| Rank | Bot[21] | Rating |
|---|---|---|
| 1-2* | bot#1337[71] | 2.746 |
| 1-2* | poetwannabe [72] | 2.536 |
| 3 | kAIb | 2.105 |

[19] Dialogue System

[20] Amazon Mechanical Turk, i.e. human evaluation

[21] https://github.com/DeepPavlov/convai/tree/master/2017/solutions

| | | |
|---|---|---|
| 4 | RLLChatBot [73] | 1.905 |
| 5 | PolyU | 1.5 |
| 6 | DeepTalkHawk | 1.229 |
| – | Human | 3.8 |

Table 2.6: ConvAI Summary 2017

**Overview of 2018**

In the second year of the competition, the chatbots were not only human evaluated, but also automatically evaluated with chosen metrics (§10.3.2). From this two round evaluation (Table 2.7), the human rating was taken as the primary one to decide what position a particular team's chatbot achieved.

| Rank | Bot | Rating | PPL | Hit@1 | F1 |
|---|---|---|---|---|---|
| 1 | Lost in Conversation [22] | 3.11 | - | 17.1 | 17.77 |
| 2 | (Hugging Face) | 2.68 | 16.28 | 80.7 | 19.5 |
| 3 | Little Baby | 2.44 | - | 64.8 | - |
| 4 | Mohd Shadab Alam | 2.33 | 29.94 | 13.8 | 16.91 |
| 5 | Happy Minions | 1.92 | 29.01 | - | 16.01 |
| 6 | ADAPT Centre | 1.6 | 31.4 | - | 18.39 |
| - | Human | 3.48 | - | - | - |

Table 2.7: ConvAI Summary 2018

### 2.4.5 Chatbots for Specific Purposes

Apart from the dialogue system competitions (§2.4), there are other commercial solutions which refer to the latest research in the field of specific topic dialogue systems (see the closed domain §6.5.2) which are not less valuable when compared to any topic dialogue system (see the open dialogue system §6.5.2), but use different approaches in the much narrow field:

- A User Simulator for Task-Completion Dialogues [74] represents a dialogue system for helping users to **book movie tickets** or to look up the movies they want, by interacting with them in natural language. It is built on top of Natural Language Understanding (NLU) and Natural Language Generation (NLG) techniques.

- A conversational agent for two different domains (a **conference information system** and **local tourist guide**) [75] is quite unusual. The common approach is

[22] https://github.com/atselousov/transformer_chatbot

usually to develop a single domain-oriented chatbot. The system consists of two main parts — decision and orchestration. The first one activates the search module based on the incoming request. The second one orchestrates various resources (QA, Paper content DB, Tourist Info DB, and Web). They cooperate to generate the response.

- A Neural Conversational Model [76] utilizes a Sequence to Sequence (Seq2Seq) model (§8.2.3) and Recurrent Neural Network (RNN) (§8.1.2) which reads the input sequence (one token at a time) and predicts an output sequence, also one token at a time. The model was tested on data from an **IT helpdesk** dataset of conversations and can sometimes track the problem and provide a useful answer to the user.

### 2.4.6 Chatbots in Health and Well-being

When considering chatbots in health care, we can see research which in several cases turned into practically used applications. Chatbots themselves are used not alone as simple conversational entities but are supported by several additional techniques, for instance gamification and those which are reviewed in psychological methods (§2.6).

Feasibility and effectiveness of using a chatbot or any other one-on-one mental health intervention that uses text-based synchronous chat was reviewed in [77] with the conclusion that studies showed significant and sustained improvements in mental health outcomes following synchronous text-based intervention, and post-treatment improvement equivalent but not superior to the usual treatment (e.g. face-to-face and telephone counseling).

Another chatbot survey [78] focuses on the narrow field of mental health care assistance in psychiatric counseling via dialogues. Based on the various studies it suggests to combine high-level Natural Language Understanding (NLU) with multi-modal emotion recognition from various content including intonation, and facial expression. It intelligently corresponds such as psychiatric case-based reasoning and long-term monitoring, and ethical judgment. All these techniques require not only significant implementation complexity but also sensitive continuous observation of users emotional changes.

**Nombot** [79] It is a food tracking chatbot which represents one of the applications. It tends to simplify manual food tracking which is not popular. It is built on top of the existing instant messaging service Telegram (§6.13.3). The approach is to use gamification with various motivation types (points collection, higher level unlocking) and compare it to the existing food tracking application (MyFitnessPal) via A/B testing (§10.2.2).

**Woebot** [6] Another chatbot implementation which serves as the psycho-social intervention (§2.6.1). It uses Cognitive Behavioral Therapy (CBT) to treat young adults with symptoms of depression and anxiety. Facebook Messenger (§6.13.3) is the platform through the chatbot was implemented and then clinically tested on the group of 70 participants. Additionally to the usage of the chatbot the participants fulfilled various standardized questionnaires (§10.4.1) such as Patient Health Questionnaire (PHQ-9), Generalized Anxiety Disorder (GAD-7), Positive and Negative Affect Schedule (PANAS) and Acceptability and Usability questionnaire.

**Lark** [80] The chatbot provides the support and cheer-leading together with tracking daily movement, weight (once a week), sleep and food. It is initially designed to promote weight loss and other health behaviors related to diabetes prevention. The study also measured user acceptability of AI coaches as alternatives to live health care professionals. The mobile application promotes sustainable behavior change and increased self-efficacy; the AI incorporates interactive elements of Cognitive Behavioral Therapy (CBT) such as reflection, legitimization, respect, support, and partnership.

## 2.5 Dialogue Systems Testing and Evaluation

For any software development or developed systems, we need to know how well it works. So, the testing methods and evaluation metrics represent the ways to get information about the dialogue system functionality.

### 2.5.1 Dialogue System Testing

To validate the functionality of dialogue system, it has to be tested. The testing can be done by two different approaches:

**Human tests** Those are done by people and their intuition and consider various approaches. From the Turing test [28] (§10.2.1) which distinguishes a dialogue system from human communication up to the A/B test (§10.2.2) where the dialogue system variants can be compared.

**Automated tests** The human is replaced by automated testing process which relays on the data rather than the intuition. It can be provided as on-premise (locally containerized) solution (§10.2.3) or in the cloud (via API) offered solution (§10.2.4) for chatbot testing.

### 2.5.2 Dialogue System Evaluation

When the system is tested it can be evaluated, for the evaluation it is good to have some baseline against the evaluation can be done. The baseline is usually defined by dialogue system human evaluation.

To narrow down the dialogue system evaluation we can consider three parts which include:

**Evaluation aspects** For the evaluation of dialogue system quality we consider specific criteria. Those criteria are called aspects (§10.3.1) and include various scale of qualitative measures in three main groups: efficiency, effectiveness, and satisfaction.

Those were identified by systematic review of more then forty research publications and articles [81].

**Evaluation metrics** Since the aspects are qualitative criteria we need to consider also quantitative part which can be calculated automatically. For such purposes various evaluation metrics (§10.3.2) exist. From BiLingual Evaluation Understudy (BLEU) [82] one of the mostly used metrics comparing candidate and response sentence and correlate with the human judgement up to the most recent Sensibleness and Specificity Average (SSA) [12] which combines two fundamental aspects of a human-like chatbots: making sense and being specific.

**Evaluation benchmarks** The standardization activities lead to the establishing several benchmark datasets (§10.3.3) from which mostly known are Stanford Question Answering Dataset (SQuAD) [83], [84] and General Language Understanding Evaluation (GLUE) [85] recently evolved into Super General Language Understanding Evaluation (SuperGLUE) [86]. Next to the universal benchmarks also benchmarks for pipeline methods (§8.7) co-exist. One of them used for Dialogue Management (DM) benchmarking is a persona-chat dataset [70] established during the The Conversational Intelligence Challenge (ConvAI) (§2.4.4). Another one is an End-to-End (E2E) dataset used for Natural Language Generation (NLG) benchmarking released as a part of E2E NLG Challenge [25].

On top of those three in case of influenced dialogue system it is necessary to evaluate also the intervention (§10.4). This is purely subjective evaluation and it can be done by various questionnaires (§10.4.1) used in clinical psychology and psychiatry.

## 2.6 Psychological methods

Two already described chatbots (§2.4.6) Woebot [6] and Lark [80] are (next to other methods) utilizing Cognitive Behavioral Therapy (CBT). This psychological method helps them to regulate and support users behavior in the way to provide the best psycho-social intervention.

The CBT from the perspective of layman seems to be a complex psychological tool which needs either cooperation with a psychologist or his/her participation in the project.

However, this is not necessary. The thesis objective is quite different than provide complex psycho-social intervention. Its enough to prove or disprove the Research Objective (RO) (§11.4.1) stated in thesis research proposal (§11).

Nevertheless, since the CBT was mentioned already, it would be good to introduce Psychological and Psycho-social interventions a little bit more (§2.6.1) and provide alternatives to CBT which still can be used when the chatbot is influenced by a stressed participant.

As the alternative to the CBT Emotion Regulation (ER) strategies could serve. They are more understandable and it is possible to implement them. A brief overview is provided in this section (§2.6.2); a detailed overview is given later (§9.4).

### 2.6.1 Psychological and Psycho-social interventions

Psychological and psycho-social interventions have traditionally made use of interaction between the client and therapist, worker, helper or counselor. Moreover, in recent years there has been a higher demand for the self-help-based interventions that involve the use of DVDs, books, computer programs or self-help manuals [87].

There is no widely accepted categorization of psychosocial interventions. The term is generally applied to a broad range of types of interventions. A few examples of such psycho-social interventions are [88]:

**Assertive community treatment** It encompasses an array of services and interventions provided by a community-based, interdisciplinary, mobile treatment team [89].

**Cognitive Behavioral Therapy (CBT)** The CBT (Figure 2.3) is used for a wide array of mental health and substance use disorders. It combines behavioral techniques with cognitive psychology, the scientific study of mental processes, such as perception, memory, reasoning, decision making, and problem-solving. The goal is to replace maladaptive behavior and faulty cognition with thoughts and self-statements that promote adaptive behavior [90].

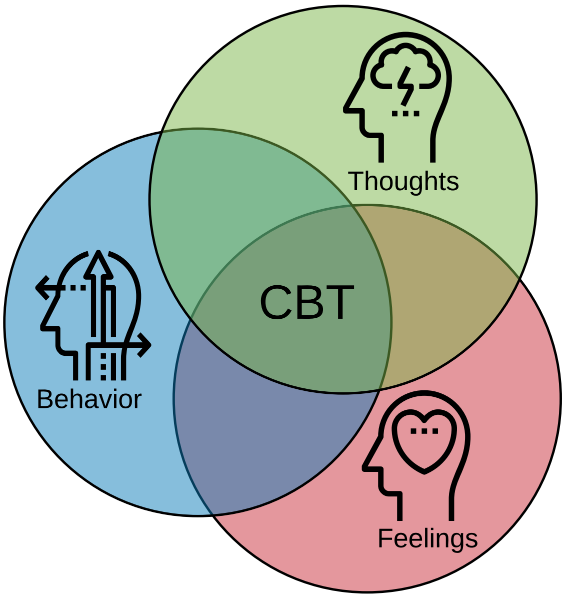


Figure 2.3: Cognitive-behavioral therapy diagram

**Contingency management** It is a psycho-social intervention designed for substance use disorders. As an evidence-based practice it uses an incentive-based approach that rewards a client contingent upon meeting desired outcomes [91].

### 2.6.2 Cognitive strategies to Emotion Regulations (ERs)

Emotions become dysfunctional when they interfere with one's ability to behave adaptively and therefore successful ER, when necessary, is crucial for psychological health [92].

ER is usually maintained by methods belonging to cognitive strategies which include distraction (§9.4.1), reappraisal (§9.4.2), labeling (§9.4.3) or paraphrasing (§9.4.4) which are described later in a deeper detail in intervention-methods (§9.4)

## 2.7 Wearable Technology and Health

As the wearable technology stepped into our lives and the market offers cheaper devices, more people start to monitor their health.

Wearables evolved over the time and offer various types of measured data (§2.7.1) with better and more reliable devices-precision (§2.7.2).

Next to health the field of well-being monitoring is also expanding. It includes more sophisticated analyses than just how many steps were done or how long the sleep was.

Together with that extensions of existing applications and the use of measured data or monitoring of well-being more and more specific applications appear (§2.7.3).

### 2.7.1 Measured data

Devices (either smart watches or activity trackers known as wearables) used every day can quickly collect hard data (§3) about health of its users.

Steps are the primary measure together with traveled distance and climbed floors usually gathered by three axes accelerometer. As the part of the movement active minutes and burned calories can be calculated. Next to steps there are devices able to measure specific exercises or activities, for instance, swimming, elliptical exercise and also sleep time and sleep quality.

Next to steps the heart rate collected by the optical measurement via photoplethysmography (PPG) techniques is the second desirable measure. It allows us to indicate specific heart rate zones and identify for instance the fitness score.

By including GPS into wearables the user can record workout routes and pace of his activities.

Some specific measures like diabetes indicators, blood pressure or ECG require specific sensors and implemented sensory controlling functionality, but makes the promise that wearables become more useful in daily life.

### 2.7.2 Devices Precision

Precision of wearable devices is perceived from various angles and measured under various setups of experiment designs [93]–[95] and for specific purposes like weight management [93].

The most commonly measured data are steps, calories, and heart rate (§2.7.1). Heart rate is in the focus of this work.

#### Heart Rate Measurement Precision

Wearables detect the heart rate through optical measurement, and its measurement precision can vary due to the technical design and processing firmware implementation.

There are papers considering some devices (Microsoft Band) very precise [95] when compared to other devices (Fitbit Surge). Some of them are oriented explicitly to some specific devices [1], [96]–[98] like Fitbit.

The paper [96] concludes that the Fitbit trackers (Fitbit Charge HR) are affected by significant systematic errors under free-living conditions. Improvements in tracker accuracy and sensitivity when measuring physical activity are required before they can be considered for use in the context of exercise prescription to promote better health. [97] claims that individual heart rate measure (by Fitbit Charge HR 2) could plausibly be underestimated by almost 30 bpm. Finally the conclusion of [98] is that wear-position of the evaluated wrist watch (Fitbit Charge 2) may impact heart rate readings, so it is necessary to hold the recommended position for accurate measurement strictly.

During data collection phase in the Pilot experiment (PX) (§3.6.1) and Quasi-experiment (QX) (§3.6.2) two specific devices were used.

The first one was Fitbit Charge HR. The results from [94] show that the mean absolute percentage error was 6.2 % when the comparison between Fitbit Charge and ECG was made. Other results from [93] present that device heart rate estimates were within 1-9 % of reference estimates.

The second device was Basis Peak. The precision of this wearable HR measurement for the entire testing interval was determined as an average difference of 3.6% between the values measured by the Basis Peak and the ECG [94].

### 2.7.3 Wearables Applications

Currently, other applications within health care are being explored with a new potential from measured data derived information which can improve users health and well-being:

- Samsung together with UCSF started the research into the relationship between stress and blood pressure derived from heart rate [23]
- NIH/NIAAA supported measuring blood alcohol concentration by Milo Sensors [24]
- Monitoring how sick the user is by tracking physiology and activity using wearable biosensors [99]
- Health Risk Assessment applications, including measures of frailty and risks of age-dependent diseases [100]
- Real-time seizure (Epilepsy) monitoring together with alerting is provided by Embrace 2 wristband from Empatica [25] company.

## 2.8 How to identify Stress

Chatbot intervention makes sense only when the user physiological state changes and triggers the event of the chatbot influence — the change of users physiological state can

[23] https://mybplab.com
[24] http://www.milosensor.com
[25] https://www.empatica.com

represent stress which needs a correct identification.

### 2.8.1 Stress Physiological Markers

The most commonly used physiological markers of stress are as follows [101]:

**Galvanic skin response (GSR)** Uses changes in skin conductivity. During stress, the resistance of skin drops due to increased secretion of sweating glands [102].

**Electromyogram (EMG)** Mesures the electrical activity of the muscles. Stress causes differences in the contraction of muscles which can be used to identify stress [103], [104].

**Skin temperature** Changes in the skin temperature are related to the stress level [105].

**Electrical activity of the heart** The most commonly used stress marker parameters are derived from the Electrocardiogram (ECG), HR and HRV [106], [107].

**Respiration** Acute stress causes changes in the breath rate [108]

**Blood pressure** Stressors induce an increase in the blood pressure compared to the baseline [109].

### 2.8.2 Stress Identified from Heart Rate (HR)

Wearable devices allow to measure many signals and provide various data. The most common (see §2.7.1) are steps which are practically dependent variable on the subject daily movement and HR. It is a suitable measure because it can be used as a stress identifier.

Stress can be identified from HR using a variety of techniques and methods, for instance:

- Using chest strap as the low-cost HR sensor which provides a combination of measures from which the mean HR, pNN50[26], and RMSSD[27] features lead to identification of stress [101]
- With an activity tracker (wearable device) connected to a smart-phone, five types of various data (steps, calories, sleep cycle, HR and resting HR) are collected; they serve to engineer 17 features used for stress recognition. [110]

[26] The pNN50 statistic is a time domain measure of HRV defined as the mean number of times per hour in which the change in consecutive normal sinus (NN) intervals exceeds 50 milliseconds.

[27] Root Mean Square of the Successive Differences is one of a few time-domain tools used to assess HRV, the successive differences being neighboring R-R intervals.

### 2.8.3 Heart Rate (HR) vs. Heart Rate Variability (HRV)

Next to the HR another parameter that was proposed to help recognize stress is HRV [111].

HRV calculations require a precise measurement of R-R interval between two heartbeats. To this day (January 2020) there are no wristbands or smartwatch (wearables) on the market that would use LED/Pulse Oximetry sensors and have enough accuracy in capturing the exact R-wave peak. On the other hand, the technology is advanced enough that such devices can measure the basic heart rate accurately. It is given by the fact that HR is not sensitive to tiny changes as it is with HRV.

To measure HRV accurately, the heart rate monitor of choice must:

- Capture and transmit measured R-R intervals accurately.
- Transmit the unaltered R-R intervals via wireless networks.

## 2.9 Summary and Research Direction

Whenever SMS (§2.1) or chatbot (§2.4.6) are used as an intervention tool, health care support for well being, support or replacement of treatment; those methods are trying to solve the same problem mainly. The adherence (commitment) and attrition (process of increasing effectiveness of the intervention, treatment, and so on) of patients or people who are using such service is the primary issue. It does not matter whether it is a message about taking a medicament or doing another round of intervention (for instance Emotion Regulation (ER) (§2.6.2)).

Chatbots are representing a new direction over the SMSs with the advantage of immediate interactive bidirectional communication. So, a chatbot through the application can ask and get feedback which complies with the required patient activity, acknowledgment about pills or current patient status gathered during the conversation.

With the advantage of keeping the high adherence and low attrition there is an increasing risk that the user is annoyed with the frequency of reflections during the day. It could lead to resign on positive aspects of the intervention application or to the tendency to skip or pretend the results and behavior which do not correspond to reality. It is not suitable for intervention treatment. In such cases it would be helpful to extend the dialogue intervention method with the simultaneously measured biological signal. Such a signal can represent a typical or elevated emotional level (for instance represented by stress). It might help the application to react and adapt itself to the situation by regulating emotions adequately.

# Chapter 3

# Soft and Hard Data

A tremendous growth of big data[1] in the recent years and also possibilities to store and process them led to the need to focus not only on physics-based sources of information — hard data (see the definition in §3.1), but also soft data — human-based sources of information (also defined in §3.1).

Those data have various sources which are quite closely inspected in §3.2. Several relationships (§3.3) between soft and hard data are defined and extensively described and then data fusion (§3.4) is introduced to reveal undiscovered information which can be used later for dialogue system influencing purposes.

A practical use of soft data or hard data (§3.5) alone or their fusion is described in three various examples including the lifestyle, mental health and medicine.

Last but not least soft and hard data need to be carefully collected (§3.6) by performing the experiments that follow broadly used standards.

## 3.1 Definitions of Soft and Hard Data

In the common understanding we distinguish two main types of data: qualitative and quantitative. These are then divided further into binomial, nominal and ordinal for qualitative data; for quantitative data we distinguish between discrete and continuous data.

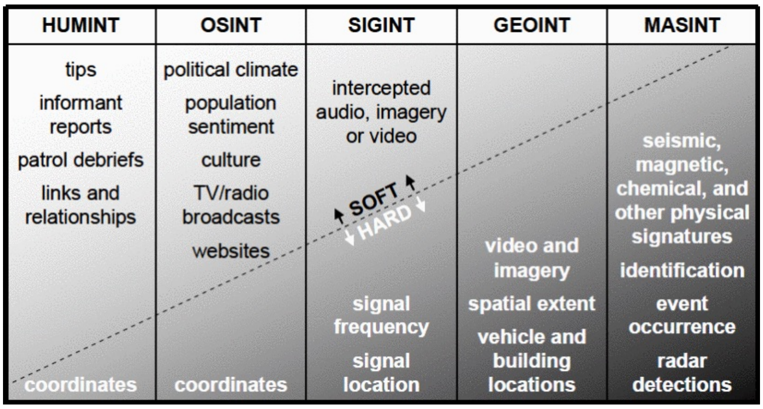
The first type of data (soft data) is based on qualitative observations. Such as ratings, surveys, pools, blog posts and discussion which contain people opinions, suggestions, interpretations, contradictions, uncertainties, and feelings. It is difficult to measure them [112]–[114].

The second type of data (hard data) is the data based on facts from reliable - quantitative sources like devices and applications. This includes phones, computers, sensors, smart meters, traffic monitoring systems, call detailed records, bank transaction records, etc. All this data can be measured, tracked, validated and proved [113]–[115].

[1] extremely large data sets

## 3.2 Sources of Soft and Hard Data

According to [116] data is divided (in the military terminology) by a source that generates it (header row in Figure 3.1) which heads to the classification between soft and hard information (Figure 3.1). While HUMINT (human intelligence) and OSINT (open source intelligence) are mostly soft data, SIGINT (signals intelligence) provides both, depends on how and for what purposes data is used and interpreted; GEOINT (geospatial intelligence), and MASINT (measurements and signatures intelligence) are considered to be mostly hard data.



| HUMINT | OSINT | SIGINT | GEOINT | MASINT |
|---|---|---|---|---|
| tips<br>informant reports<br>patrol debriefs<br>links and relationships<br>coordinates | political climate<br>population sentiment<br>culture<br>TV/radio broadcasts<br>websites<br>coordinates | intercepted audio, imagery or video<br>signal frequency<br>signal location | video and imagery<br>spatial extent<br>vehicle and building locations | seismic, magnetic, chemical, and other physical signatures<br>identification<br>event occurrence<br>radar detections |

Figure 3.1: Representative information elements according to generating source (header row) and classification between hard and soft information

For our purposes, we would like to point out a few maybe obvious HUMINT/OSINT soft data sources (§3.2.1) and MASINT hard data sources (§3.2.2).

### 3.2.1 Sources of Soft Data

Human-generated data represent typically opinions and feelings about tangible and intangible things and can be found e.g. in:

- Textual movie reviews from users, for instance, Rotten Tomatoes [2]
- Assorted merchandise reviews, where one of the most known is Amazon [3]
- Open discussion forums and platforms for discussions of any kind, like Reddit [4]
- Any kind of social media beginning with Facebook [5], going over Twitter [6], to Instagram [7] and others.
- And much more.

[2] https://www.rottentomatoes.com
[3] https://www.amazon.com
[4] https://www.reddit.com
[5] https://www.facebook.com
[6] https://twitter.com
[7] https://www.instagram.com

### 3.2.2 Sources of Hard Data

There are many applications or devices where the measured data (hard data) is the main source of information:

- Static sensors included in diverse products or measurement devices providing on-demand output.
- Numerous wearables and smartwatch; when synchronized they upload data into a cloud.
- IoT used in different devices for continuous measurement and reporting in real-time or close to real time.

## 3.3 Relation between Soft and Hard Data

The circumstances under which soft and hard data are collected determine their relationships. The following list describes several examples of such relations.

- **Common Object of Interest** — whenever we have some interest about a particular object, this object can be described by technical or statistical data (hard data) and subjectively when the author is projecting his/her feelings and opinions (soft data) on such object. These data do not necessarily need to be recorded at the same time. The object which can be described by parameters or statistical data together with the subjective description defines the relation between data.
  - Movie
    - Movie visits provide statistical (hard data) information about the object.
    - Movie reviews represent subjective description (soft data) about the object.
  - Car
    - Car technical parameters describe an object from the technical (hard data) perspective
    - Car review contains a subjective description of its attributes (soft data)
- **Common Subject of Evaluation** — an evaluation of human being is possible by multiple standardized approaches, e.g. from a psychological or physiological perspective.

  We can have long-term given attributes which are barely changing in time (hard data) and physiological measures that change frequently (hard data). Next to the hard data the subject also provides the feedback which represents his/her feelings or mood (soft data).

  Like the previous category such data does not need to be recorded at the same time but within a short period. The main point of the relation is the measured subject describing his/her feelings.

  - Age, race, gender, BMI, smoker, drug user, alcoholic, and so on represent subject long-term observed attributes (hard data)
  - Temperature, blood pressure, heart rate, and so on represent short-term subject physiological measures (hard data)
  - Results from a questionnaire or assessment represent a subject feedback (soft data)

- **Common Period of Data Collection** — the last relation between soft and hard data can be described as any soft and hard data recorded at the same time. This relation can be represented by physiological measures together with mood or feelings expression.

  An important aspect is the precise collection of data because the relation between soft and hard data is time dependent.

  - Temperature, blood pressure, heart rate, and so on are physiological measures from subject collected during a certain period (hard data)
  - The subject mood can be extracted from utterances or speech several times a day during the same period (soft data)

## 3.4 Soft and Hard Data Fusion

Data fusion is a discipline which next to the existing problems with a single source or sensor data brings the problems specifically related to the fusion process [117] like conflicting data, data correlation, data association, operational timing and so on.

Data fusion itself demands a fusion algorithm related to specific data or a specific problem to process data into the final product. In this case, our interest is a fusion of soft and hard data.

The soft data (human created) expressed preferably as a text without any constraints and processed by Natural Language Processing represents a complex fusion problem [118].

Combination of soft and hard data and its fusion is considered even more challenging despite this is necessary for some applications [119].

However, there are papers related to the human-centered data fusion paradigm [120] and soft and hard data fusion [116], [121], [122] that establish new trends related to a general data fusion framework where soft and hard data can be processed efficiently.

## 3.5 Usage of Soft and Hard Data

Soft and hard data either alone or together gives us various options to use them, for instance:

**Lifestyle** Marketing measurements like Pay per click (PPC), Pay per post (PPP), Pay for placement (P4P), Cost per acquisition (CPA) together with the product placement, product reviews, product blog posts, etc.

**Mental Health** Psycho-social questionnaires (§10.4.1) to evaluate a particular mental illness or state or quantified and qualified emotions for Emotions Regulation (Emotion Regulation (ER)).

**Medicine** Signal measurement or quantified laboratory results together with subjective feedback leading to the proper illness diagnostics.

## 3.6 Soft and Hard Data Collection

There are several papers [116], [122] which present a collection of soft and hard data (Common Object and Common Subject) described in §3.3. It is not a typical approach to collect both types of data at the same time (Common Period) so there are not many sources. The reasonable approach is to collect relevant data in few stages and use them later as an input for further research.

### 3.6.1 Pilot Experiment (PX)

A PX is a small scale preliminary study conducted in order to evaluate the feasibility, time, cost, adverse events, and improve upon the study design before the performance of a full-scale research project [123].

The first data collection approach was the two-times performed PX with one subject; its detailed description including collected data can be found in [124] and it is also briefly described in §11.2.1.

### 3.6.2 Quasi-experiment (QX)

A QX is an empirical intervention study used to estimate the causal impact of an intervention on its target population without random assignment. Quasi-experimental research shares similarities with the traditional experimental design or randomized controlled trial, but it specifically lacks the element of random assignment to treatment or control. Instead, quasi-experimental designs typically allow the researcher to control the assignment to the treatment condition, but using some criterion other than random assignment (e.g., an eligibility cutoff mark) [125].

Based on the experience gathered during PX design, data collection and results and conclusions derived from the data we made the decision to adapt PX in the way it is more suitable for QX. The full description of QX can be found in §11.2.2 later in this work.

### 3.6.3 Natural Experiment (NX)

A NX is not part of this thesis. Its setup and execution depends on the results and conclusions related to particular Research Objective (RO) of this work.

## 3.7 Conclusion! What Next With Data?

Soft and hard data are giving us the chance to look at a particular action or process from a quantitative or qualitative perspective. The data can be engaged in the process or action independently or jointly, so they are collected alone or together in some relation (§3.3).

From a broad perspective, it is correct to talk about quantitative and qualitative or soft and hard data. However, for a practical application, we need real sources of such data which are described in the next section (§4).

The relation between soft and hard data collected during PX and QX determines them as a potential further input to the data fusion (§5) and it also determines its usage (§3.5).

Then later we would like to use them (it does not matter if fused or not) as the source for dialogue system influencing (§9). It might lead to a potential change in the conversation. And fulfill the original idea about psychological treatment or counseling based not only on the text input, but also on measured data.

# Chapter 4

# Influencing data

In chapter (§1.2) we have introduced dialogue system common influencing idea in Figure 1.1. There were depicted quantitative and qualitative measures. However, wearable devices provide specific measured data and from the text we are able to extract specific qualitative measures. So, we need to concertize those two sources (Figure 4.1) to be able to collect them and use them in the future for dialogue system influence.

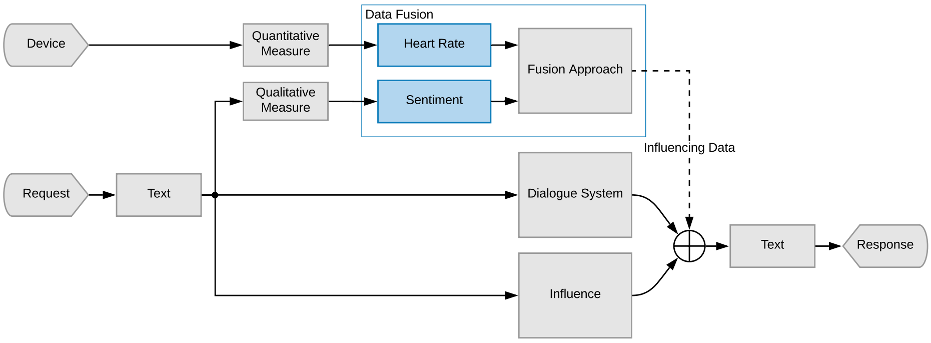


Figure 4.1: Influencing data

Most of the wearable devices are able to measure several various quantitative measures (§2.7.1). The problem is usually with their extraction either in real-time or in batch mode (§5.6). The real-time extraction (§5.6.2) is a much suitable option for influencing than the batch mode (§5.6.1). From this perspective information about steps is usually available, heart rate (not always) and sleep cycles. So, this data can be used for dialogue system influence.

From the conversation itself, we can process the dialogue text and extract from it the human sentiment as the qualitative measure.

The Research Objective (RO) (§11.4.1) is about to feed the dialogue system by either quantitative measure or qualitative measure or their fusion (combination) (§5). So, we can take a look how to get a correct quantitative and qualitative measures from the raw data and later (if necessary) how to do such data fusion.

## 4.1 Data Pre-processing

Raw exported data from various sources are not useful as influencing data. It might not contain the information which is suitable for influencing; such information needs to be first extracted. It could be also incomplete or normalization is necessary.

Data collections obtained in the particular experiments, i.e., Pilot experiment (PX) and Quasi-experiment (QX) are briefly described in §3.6 and fully described (with all the details how the author collected the data) in §11.2.

### 4.1.1 Tweet Data Pre-processing

Textual data (recorded on Twitter) collected during Quasi-experiment (QX) (§3.6.2) were recorded in the Czech language. Before it can be used for further analysis it is necessary to pre-process them.

1. Manual or automatic Czech grammar corrections (Czech Grammar Checker integrated in MS Office)
2. Machine translation via Google Translator [1] to English
3. Manual check of the translation correctness
4. Machine learning or Deep learning sentiment extraction (see §4.2.3)

Since the sentiment is subject matter it cannot be normalized.

### 4.1.2 Heart Rate Data Pre-processing

Heart rate collected during QX is recorded per individual subject and time window. Data normalization and standardization eliminate differences in minimum and maximum HR among participants of an experiment.

1. Fix the missing values by Heart Rate (HR) imputation (see §4.3.1)
2. Normalization and/or standardization of HR data (see §4.3.2 and §4.3.3)

## 4.2 Sentiment

There are multiple ways how to extract sentiment from utterances. Either a search for particular keywords or use smileys to express the sentiment. And extract it per aspect or entity, sentence and document [126].

[1] https://translate.google.com

### 4.2.1 Sentiment Representation

Also the representation of sentiment extracted from text can vary. From simple polarity expression which defines negative sentiment as (-1, N and Neg) values, positive as (+1, P and Pos) and sometimes adding (0, X or Neu) as neutral sentiment. To the continuous scale from {-1, +1} or the categorical (Very Negative, Negative, Neutral, Positive and Very Positive), discrete values (0, 1, 2, 3, 4) expressed scale [127] or similar representation (–, -, 0, +, ++) [128]. Up to the emotions defined by psychologists as anger, sadness, joy, disgust, fear and surprise [129] and enhanced about shame and guilty [130].

In the following sections sentiment with following values will be used:

- Continuous scale: {-1, +1}
- Discrete scale: (N, P) or encoded as (-1, +1) respectively

### 4.2.2 Sentiment Extraction Techniques

According to a survey on sentiment analysis of scientific citations [131], the sentiment analysis domains suitable for sentiment extraction are following: scientific citations, product reviews, discussion forums and micro-blogs.

The techniques used for sentiment analysis are the following: lexicon based, keyword based, machine learning based, and deep learning based approaches. The modern approaches tend to use more and more deep learning as the storage prices drop and computational power increases due to big data popularity.

The above mentioned techniques for sentiment extraction do the following:

**Keyword based** Build based on the emotionally colored words (affective rating) which represents either strong positive emotion or negative emotion. The words or word collocations are chosen based on the previous text analysis (Natural Language Processing (NLP)).

**Lexicon based** There exist several lexicons containing word lists labeled with emotional valence, for instance [132] or there is research which provides word rating, like Affective Norms for English Words (ANEW) [133], AFINN sentiment lexicon [134], OpinionFinder [135], SentiWordNet [136], [137] and WordNet-Affect [138] or word list which could be found in SentiStrength[2] software [139].

**Machine learning based** Usually utilize Support vector Machine (SVM) or Naïve Bayes machine learning algorithms and based on the training data and particular corpora (§7.1) provide a supervised model which classifies the sentiment in provided input texts.

**Deep learning based** It is practically logical continuation of previously applied machine learning algorithms for sentiment extraction, only with application deep learning methods like recursive deep models (Recurrent Neural Network (RNN)) and large

[2] http://sentistrength.wlv.ac.uk

datasets with sentiment label (e.g. Stanford Sentiment Treebank[3]) or subjective rating (Movie Review Data[4]).

### 4.2.3 Sentiment Extraction Tools

The following list presents several public libraries which provide either sentiment extraction out of the box or with some benefit (pre-trained models, contemporary research direction). The programming languages vary, some of them are written in Java or C, but most of them in Python.

**Stanford CoreNLP**[5] [127] It is one of the most used tool, which provides sentiment analysis using **Deep Learning (DL)** [128] models. A binary tree represents the sentence where each root node gets a sentiment score.

**VADER**[6] [140] Another popular NLP library often used is Natural Language Toolkit (NLTK) which has a sentiment package containing several sentiment modules and amongst others, VADER sentiment module. This module utilizes strictly the **keyword based** sentiment extraction.

**Sentiment Classifier**[7] It is library using word sense disambiguation using WordNet [141] and word occurrence statistics from movie review corpus NLTK. So, the techniques utilized for sentiment extraction are **keyword based** and **lexicon based** combination.

**fastText**[8] [142] It is the library developed by Facebook research that is intended for efficient learning of word representations and sentence classification which utilizes the **DL** models. The advantage is the word embedding and thanks to that availability of pre-trained models (English and other 157 different languages).

**TextBlob**[9] The TextBlob is also a popular library which stands on the NLTK and pattern libraries and makes text processing simple by providing an intuitive interface to NLTK. It contains two sentiment analysis implementations, **lexicon based** (PatternAnalyzer) and **Machine Learning (ML) based** (NaiveBayesAnalyzer which is trained on a movie reviews corpus §7.1).

The most recent approaches for sentiment extraction more and more involve modern word embedding (§8.2.2) and sentence embedding (§8.2.3), more specifically transformers and specifically Bidirectional Encoder Representations from Transformers (BERT) and various evolution's and modifications. In nutshell it is modern way to represent a text in the form to be processed with Deep Learning (DL) techniques.

[3] http://nlp.stanford.edu/sentiment

[4] http://www.cs.cornell.edu/people/pabo/movie-review-data

## 4.3 Heart Rate

HR is measured time series data corresponding to each subject, and its mental and physical state at the moment of measurement.

All the missing data can be replaced by values calculated by imputation techniques (§4.3.1). Several techniques of data imputation are discussed later.

For further analysis, it is necessary to have a consistent scale and distribution of data. Two techniques can be used to re-scale the data values consistently normalization (§4.3.2) and standardization (§4.3.3).

### 4.3.1 Missing Data Imputation

The assumption about the nature types of missing data is named the missingness mechanism. According to the definition [143], there are three unique types of missing data mechanisms:

**Missing Completely at Random (MCAR)** The inclination for a data point to be missing is completely random.

**Missing at Random (MAR)** The inclination for a data point to be missing is not related to the missing data, but it is related to some of the observed data.

**Non-Ignorable (NI) or Missing Not at Random (MNAR)** It means there is a relationship between the inclination of a value to be missing and its values.

Since HR is collected automatically by a wearable we are considering only MCAR missing data mechanism. So, the randomness is given either by external (wrong position or tightening of the watch on hand) or internal (error of data measurement or processing) causes.

The data collected during the Pilot experiment (PX) (§11.2.1) by Fitbit Charge HR contains not particular missing values but gaps longer the standard sampling frequency (every 5 seconds with typical values every 5 - 15 seconds) [124]. For data collected during QX (§11.2.2) there is an assumption we can expect similar problems as they appeared in previously collected data.

The imputations with mean, median and mode are simple but, like complete case analysis, can introduce bias on mean and deviation [144] with proving the point and proposes to use the regression imputation which can preserve the relationship between missing values and other variables.

### 4.3.2 Heart Rate Normalization

Normalization is a rescaling of the data from the original range to the range of 0 and 1.

This range represents the advantage in comparison of various subjects with the same collected measure. The disadvantage in comparison to the standardization (§4.3.3) is smaller standard deviations. This, on the other hand, can suppress the effect of outliers.

**Rescaling (min-max normalization)**

The simplest method is rescaling the range of features to scale the range in [0,1] or [-1,1]:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)} \tag{4.1}$$

where $x$ is an original value, $x'$ is the normalized value.

**Mean normalization**

Similarly to the min-max normalization, we can rescale the range of features not with the *min* value, but with the *mean* value

$$x' = \frac{x - \text{mean}(x)}{\max(x) - \min(x)} \tag{4.2}$$

where $x$ is an original value, $x'$ is the normalized value.

### 4.3.3 Heart Rate Standardization

Standardization (or Z-score normalization) of a dataset involves rescaling the distribution of values (features) so that the mean ($\mu$) of observed values is 0 and the standard deviation ($\sigma$) is 1, i.e., properties of a standard normal distribution.

If the population mean and population standard deviation are known, the standard score (also called z score) of a raw score x is calculated as

$$x' = \frac{x - \mu}{\sigma} \tag{4.3}$$

where $x$ is an original value, $x'$ is the standardized value, $\mu$ is the mean of the population, $\sigma$ is the standard deviation of the population.

## 4.4 Conclusion: The Power of Influence

It is a difficult question what data to use for influencing a dialogue system. The influence needs to make sense in the way it has a significant effect on dialogue system behavior, and the dialogue system is reacting naturally without any back and forward jumps in conversation.

From the previous chapter (§3), it is obvious it could be either soft or hard data or their fusion that is described in the next section (§5).

In this rigorous thesis, the main focus is on sentiment extracted from the conversation and the HR serving as influencing data. However, this data are not the only data suitable to influence a dialogue system. So, these two expected data sources can be either extended or replaced by new ones which are available using new devices (e.g. a camera) or better approaches.

For instance, there are expected wearables with blood pressure measurement or other stress levels indicators. Alternatively, video processing with real-time capabilities in mobile phones, which allows the identification of mood from the face mimics in the real-time, would be beneficial.

Nevertheless, the power of influence can be achieved in many ways; the most simple and effective solution corresponds to the intent of the application itself.

# Chapter 5

# Data Fusion

The influencing data (§4) gives us the chance to turn them into relevant information. However, when it stands independently, it gives us only a partial picture.

When combined with multiple data sources (multi-sensor data), we can get a better overview of what is happening at a particular moment. This data combination is called data fusion. Data fusion includes various processes to combine data.

In our particular case we will describe the fusion (Figure 5.1) of soft data (§3), for instance sentiment (§4.2) and hard data (§3), for instance Heart Rate (HR) (§4.3). The data fusion provides influencing data (in this case, ideally stress) for dialogue system influence.

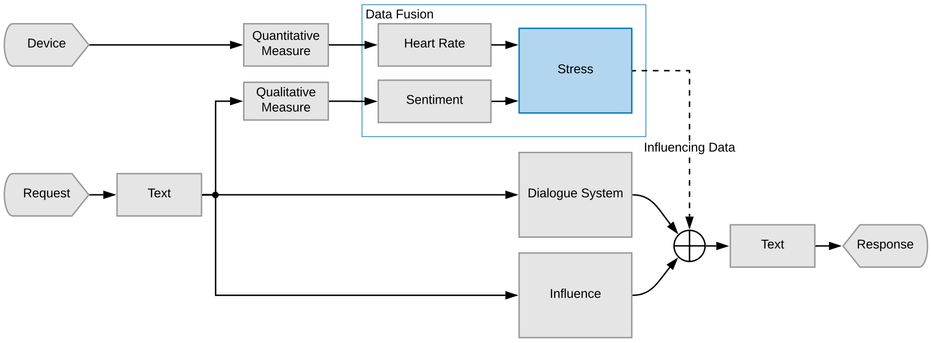


Figure 5.1: Data Fusion

## 5.1 Discrete and Continuous Variables

Variables with the finite, generally small number of values are called discrete. So, gender and blood type are considered as examples of discrete variables.

Opposite to discrete variables we have continuous variables which can take an unlimited (infinite) number of values within a range. Typically weight and height are examples of a continuous variable.

Any measured variable is usually recorded by sampling its value in time. Thus also the sampling frequency determines whether the data from a time perspective are continuous or discrete.

### 5.1.1 Heart Rate and Sentiment Representation

Sentiment can be represented (§4.2.1) by both continuous <-1: +1> or discrete values enumerated by (-1, 0, +1) or (0, 1, 2, 3, 4). From the sampling perspective sentiment is purely discrete, since it's recorded with a long period about dozens of minutes.

HR is represented by discrete values from 40 to 200 bpm with step 1 bpm, and also sampled in discrete intervals from 1 to 60 per minute. However, compare to sparse sentiment, we will consider HR continuous in value and time because of the density of samples.

## 5.2 Discrete and Continuous Sentiment

Since sentiment is represented either by discrete or continuous values, we can use both representations for data fusion. The values do not need to be adapted in any way (normalization) because they contain personal expression of a sentiment, which is individual.

For the discrete time dimension we need to adapt either sentiment and reconstruct it (§5.2.1) to continuous time or interpolate (§5.2.2) each sentiment record into the window in which will serve for later to make HR continuous time discrete (§5.3.1).

Here are those two options of sentiment transformation in the diagram (Figure 5.2):

1. Discrete sentiment is extracted from textual data with discrete time. Represented by scalar values (-1, 0, 1).
   (a) Further used for fusion with HR: discrete value and discrete time (§5.4.1).
   (b) Filter window in which the continuous HR is transformed to: discrete or continuous value and discrete time (§5.3.1).
2. Continuous sentiment is also extracted from textual data with discrete time. The values are vectors normalized from -1 to 1 continuously.
   (a) Further used for fusion with HR: continuous value and discrete time (§5.4.2).
   (b) Further used for fusion with HR: continuous value and continuous time (§5.5).

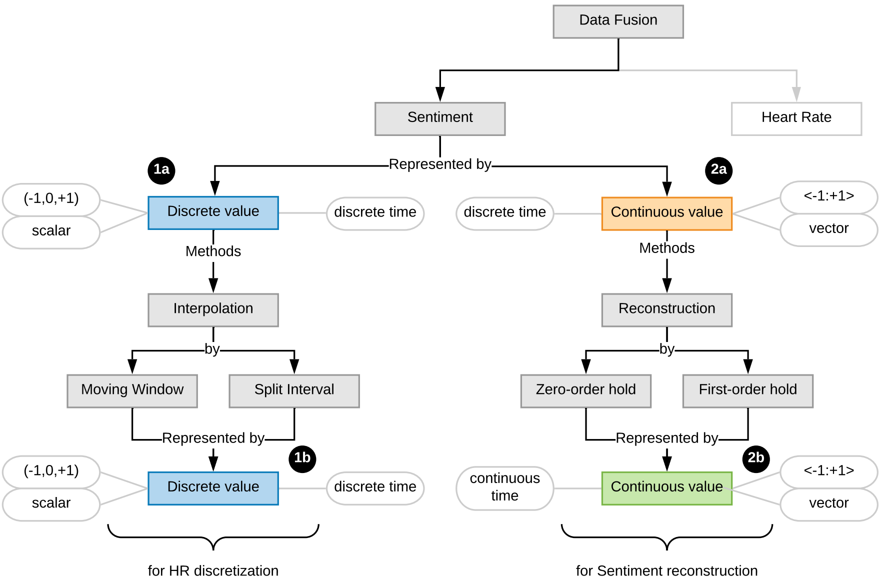


Figure 5.2: Sentiment data transformations

### 5.2.1 Sentiment Reconstruction

For two time series fusion we need to enhance granularity and smooth sentiment, i.e., to do signal reconstruction because we consider sentiment records as time discrete (§5.1.1).

For this purpose, we can either use simple prolongation of existing sentiment data till the next change, i.e. **Zero-order Hold (ZOH)** reconstruction or triangular prolongation of existing sentiment till the next change called **First-order Hold (FOH)** reconstruction.

#### Reconstruction by Zero-order Hold (ZOH)

ZOH is a mathematical model of the possible signal reconstruction (5.1). Its application is to convert a discrete time signal to a continuous time signal by holding the same sample value for one sample interval, so prolong the value in time.

$$x_{ZOH} = \sum_{n=-\infty}^{\infty} x(n) rect(\frac{t - \frac{T}{2} - nT}{T}) \tag{5.1}$$

Where $rect(x)$ is the rectangular function (5.2)

$$rect(x) = \Pi(x) = \begin{cases} 0, & \text{if } |x| > \frac{1}{2} \\ \frac{1}{2}, & \text{if } |x| = \frac{1}{2} \\ 1, & \text{if } |x| < \frac{1}{2} \end{cases} \tag{5.2}$$

In the practical approach is just about to repeat the value of extracted sentiment for corresponding HR data where the sentiment is unknown until the next extracted sentiment appears (Figure 5.3).

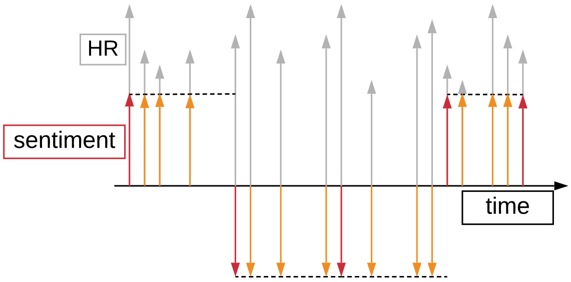


Figure 5.3: Zero-order Hold Sentiment Reconstruction

### Reconstruction by First-order Hold (FOH)

FOH is a mathematical approach of discrete signal reconstruction (5.3) where the signal is reconstructed as a piecewise linear approximation to the original signal that was sampled. So, we do the linear interpolation between the values from two consecutive samples.

$$x_{FOH} = \sum_{n=-\infty}^{\infty} x(nT)tri(\frac{t-nT}{T}) \tag{5.3}$$

Where $tri(x)$ is the triangular function (5.4).

$$tri(x) = \Lambda(x) = \begin{cases} 1-|x|, & \text{if } |x| < 0 \\ 0, & \text{otherwise} \end{cases} \tag{5.4}$$

From the implementation perspective, it is about to find out such linear interpolation between two consecutive extracted sentiments. Such linear function interpolates the adjusted sentiment value evenly for all the HR data where sentiment is unknown (Figure 5.4).

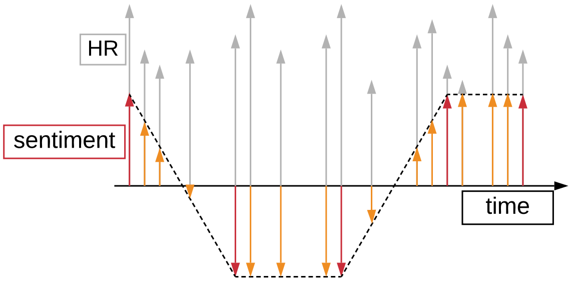


Figure 5.4: First-order Hold Sentiment Reconstruction

## 5.2.2 Sentiment Interpolation

The sentiment value is not a single occurrence in time; the sentiment lasts for some time. So, the sentiment value validity is within some interval or window. This interpolation is later used to make HR discrete within the sentiment interval or window. Thus we need to perform some sentiment interpolation over the originally recorded sentiment data.

Two basic approaches of sentiment interpolation were described in [145]: splitting by interval and splitting by window. Both will be briefly introduced in the next two sections.

### Interpolation by Splitting Interval

The first option is to split the interval between two neighborhood sentiment values and interpolate them (Figure 5.5).

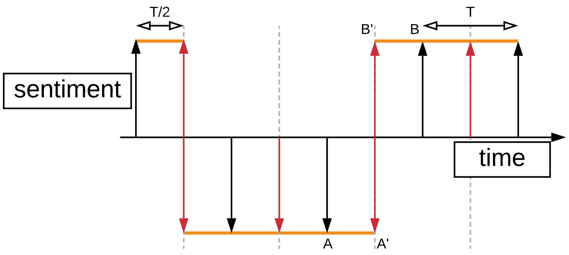


Figure 5.5: Sentiment interpolation by splitting the interval

The original sentiment is represented by the black arrows $A$ and $B$ in the graph. The interpolated sentiment is represented by the corresponding red arrows $A'$ (5.5) and $B'$ (5.6).

$$A' = \{ts_A + \frac{(ts_B - ts_A)}{2}; s_A\} \tag{5.5}$$

$$B' = \{ts_B - \frac{(ts_B - ts_A)}{2}; s_B\} \tag{5.6}$$

where $ts$ is a time-stamp, $s$ represents information about the sentiment and $\forall ts : ts_B > ts_A$.

### Interpolation by Moving Window

The second option is to define a window around each sentiment occurrence and perform the interpolation process inside this window.

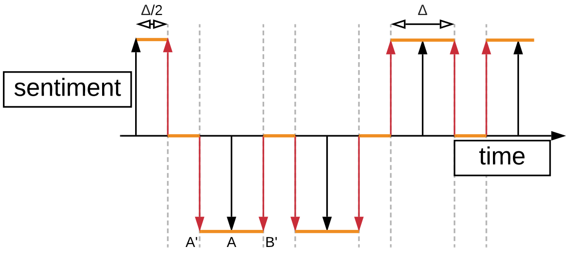


Figure 5.6: Sentiment interpolation using the moving window

The original sentiment is represented by the black arrows $A$ in the graph. The interpolated sentiment is represented by the corresponding red arrows $A'$ (5.7) and $B'$ (5.8).

$$A' = \{ts_A - \frac{\Delta}{2}; s_A\} \tag{5.7}$$

$$B' = \{ts_A + \frac{\Delta}{2}; s_A\} \tag{5.8}$$

where $ts$ is a time-stamp, $s$ represents information about the sentiment, and $\Delta$ is the length of the interpolation window (for instance 30 minutes).

## 5.3 Discrete and Continuous Heart Rate (HR)

HR is considered as a continuous variable (§5.1.1), so it allows us to work with its continuous form, or it has to be discretized. For the fusion with the original discrete sentiment, the adaption is necessary to reduce the HR time series into particular discrete values by using various discretization techniques. On the other hand, the HR data can remain as they are whenever combined with the reconstructed (§5.2.1) continuous time sentiment.

A list of HR data type options and the corresponding diagram (Figure 5.7) follow:

1. Continuous HR (values <40:200>) is collected directly from a wearable device with continuous time.

   (a) Discretized through linear regression - slope trend (§5.3.1) for further fusion with sentiment: discrete value and discrete time, i.e. scalar analysis (§5.4.1).

   (b) Discretized through linear regression - slope value (§5.3.1) for fusion with the sentiment: continuous value and discrete time, i.e. vector analysis (§5.4.2).

   (c) Normalized across all experiment participants for future fusion with sentiment: continuous value and continuous time (§5.5).

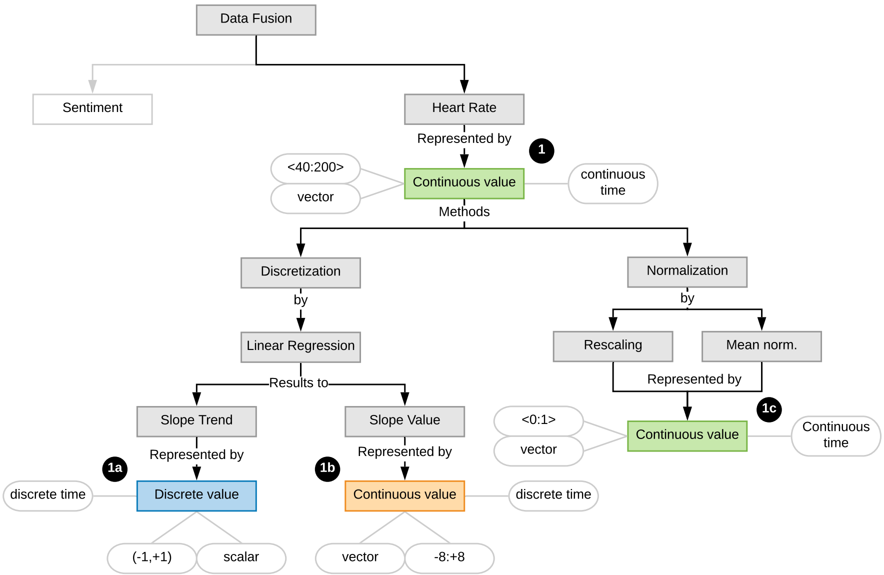


Figure 5.7: Heart Rate data transformations

## 5.3.1 Heart Rate Discretization

HR is a continuous variable in time and value. We can make a fusion with reconstructed sentiment (§5.2.1) right away, but it is necessary to discretize HR for the fusion with discrete sentiment.

To make HR discrete, it makes sense to include all the HR values from the near neighborhood where the sentiment is valid. This filtration criteria for HR is given by sentiment (§5.2.2) interpolated either with moving window or splitting interval as it is already described above.

In the neighborhood of particular sentiment the HR discrete value can be given for instance by trend (slope or value). Such trend can be gained as 1st or nth derivation of the linear or polynomial regression or Simple Moving Average (SMA) of HR signal.

For simple linear regression, both sentiment interpolation methods (splitting interval or moving window) have been proved as equivalent and interchangeable in [145]. Regardless of the sentiment interpolation method, it is possible to use one of them without any loss.

### Simple Linear Regression

The HR trend can be gained as a slope (first derivation) value or trend of the simple linear regression (5.9) of HR in the neighborhood of particular sentiment.

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \tag{5.9}$$

describes a line with slope $\beta_1$, y-intercept $\beta_0$ and random error $\varepsilon$.

### Polynomial Regression

The polynomial regression model (5.10) can be used as a better interpolation of HR to receive the slope value or trend as the nth derivation of interpolated data.

The polynomial regression model:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + ... + \beta_m x_i^m + \varepsilon_i \tag{5.10}$$

describes a base definition with parameters $\beta$ and random error $\varepsilon$.

#### Simple Moving Average (SMA)

SMA (5.11) is the unweighted mean of the previous n data usually used for financial analysis on stock or forex market. This method can be used for HR interpolation in the specific window corresponding to the particular sentiment to which the discrete value of the HR can be found.

$$\overline{x}_{SM} = \frac{1}{n} \sum_{i=0}^{n-1} x_{M-i} \tag{5.11}$$

where the window size $n \in \mathbb{R}$ and $n > 0$

The calculation of the next value (5.12) means that a new value comes into the sum, and the oldest value drops out.

$$\overline{x}_{SM} = \overline{x}_{SM}, prev + \frac{x_M}{n} - \frac{x_{M-n}}{n} \tag{5.12}$$

## 5.4 Discrete Data Fusion

In this part of the analysis, we take the sentiment and HR (§5.3.1) with discrete time and discrete or continuous values. The sentiment is represented by discrete-time with low granularity with resolution once per 45 minutes (§11.2.1) or once per hour (§11.2.2).

### 5.4.1 Scalar Analysis

For purposes of scalar analysis, both signals are points in a specific time, i.e., scalars for the specific moment when they are paired together.

Sentiment is then represented by (-1, 1) pair for negative and positive one. HR by its slope direction coming from discretization (§5.3.1) also takes values (-1, 1) for increasing and decreasing HR on particular interval (split interval or interpolation window (§5.2.2)) related to specific sentiment in time.

#### Matrix Representation

The natural combination of previous pair values of sentiment and HR leads to the following matrix representation (Table 5.1). It can be later used when the combinations are correctly translated into specific states (for instance stress dichotomy) as influencing data for a dialogue system.

| | | HR Slope | |
|---|---|---|---|
| | | -1 | 1 |
| Sentiment | -1 | [-1,-1] | [-1, 1] |
| | 1 | [ 1,-1] | [ 1, 1] |

Table 5.1: Sentiment and HR in matrix representation

### Stress Dichotomy

Stress dichotomy (eustress, distress) represents the relation between the HR and sentiment and has been presented in [145] already. The following Table 5.2 shows translated relation between particular values of sentiment in combination with HR slope values into the specific states.

| | | HR Trend | |
|---|---|---|---|
| | | decreasing | increasing |
| Sentiment | negative | relax | distress |
| | positive | relax | eustress |

Table 5.2: Stress dichotomy in matrix representation

## 5.4.2 Vector Analysis

In the previous section §5.4.1 sentiment and HR data were considered as the discrete values — scalars. In this section they are presented as vectors.

The sentiment direction and length represent a vector in the sentiment data. Where the orientation is either positive or negative, the length is between <0,1> up to the maximum of 1. The HR represent the vector with the trend (can be a slope of its linear interpolation within a specific window) which is also either positive or negative, and it has its size which could be potentially normalized in the same way as the sentiment up to the 1.

### Dimensional Representation

The two vectors fusion can be depicted in the orthogonal space. It is the similar to the emotion detection [146] (Figure 5.8) which is a part of NLP where we need emotional corpus.

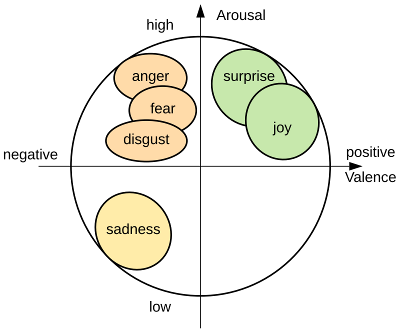


Figure 5.8: Six basic emotions in dimensional space

The vectored HR and sentiment are combined within orthogonal space (Figure 5.9). It can be either inside the circle normalized to 1 or the ellipse when HR is not normalized. This dimensional representation defines the clusters of the stress dichotomy.

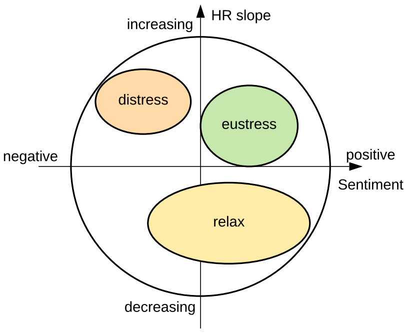


Figure 5.9: Stress dichotomy in dimensional space

## 5.5 Continuous Data Fusion

Considering that both data are taken in continuous time, i.e. HR in its original form as it was collected or with imputed missing values (§4.3.1) and sentiment reconstructed (§5.2.1) up to the granularity of HR. It gives us a bigger amount of data to fuse and process (dozens times a day vs. thousands times a day).

To influence the process of the dialogue system it does not matter if the data is time-series or sequence of values. Whereas a time-series is an ordered list of numbers and a sequence is an ordered list of nominal values (symbols) [147], we can use the data as it was originally collected (time-series) or easily convert it into the sequence.

For the analysis and identification of stress from continuous data, we can use time-series or sequence analysis. For instance, we can identify seasonality [148] in HR and combine it with discrete or continuous values of sentiment to get the influencing data for the dialogue system.

## 5.6 Real-time or Batch Data Processing

All the previous analytical methods are either an ideal option on how to process data or taken from practical experience how the data were processed already. In all the cases, the data were already collected, stored, and processed as static.

In the practical application of dialogue system influencing the data will be received in real-time. The Heart Rate (HR) of the subject will be measured several times a minute, and the sentiment can be extracted whenever there is a textual response from the subject to the dialogue system.

It leads to the adaption of current data processing and fusion for identification stress as an influencing signal. The data can be either buffered and processed lately cumulatively in a batch mode (§5.6.1) to provide relevant information over some specific window or processed in real-time mode (§5.6.2).

Both modes have advantages and disadvantages and both modes require to choose corresponding methods described in the previous sections.

### 5.6.1 Batch Data Processing

Whenever the batch data processing is applied, it leads to the delay between incoming data and the dialogue system influencing because the influencing signal needs to be first identified in the collected data during some collection window.

The trigger definition and identification when the window for data collection into the buffer needs to be closed is also a nontrivial task. It depends on what kind of data (discrete, continuous) is used and what particular method is applied (split by interval, split by window).

### 5.6.2 Real-time Data Processing

Real-time processing allows reacting with much more flexibility than batch data processing (no delay is an advantage). On the other hand, the disadvantage lays in the lack of overview of the actual pattern.

The data which is known from the past plus the current incoming set of data of non-closed collection window gives us the potential for the immediate outcome, but it can be incorrect. The next set of data coming in a few moments can change the perception of the problem immediately.

## 5.7 Conclusion! The fusion that is what is going on!

The data fusion brings additional information that can be used for dialogue system influencing. Stress can be identified just by HR (§2.8.2), but to combine it with other data gives an opportunity to identify the type of stress and its possible origin. It is important to distinguish whether the response of the dialogue system might or might not be influenced.

The data fusion of sentiment and HR for dialogue system influencing can be done through various data adaptations like sentiment interpolation (§5.2.2) or HR discretization (§5.3.1) because the data doesn't have the same character when collected.

For data preparation, use of Zero-order Hold (ZOH) is the easiest way to sentiment reconstruction (§5.2.1), because it keeps the value of sentiment till the next change (when compared to First-order Hold (FOH)). The choice of the sentiment interpolation (§5.2.2) is not important as it was already investigated [145]. HR discretization can be performed using the simplest methods presented, i.e., simple linear regression, which brings the slope value and trend right away.

Discrete data fusion (§5.4) allows to do scalar (§5.4.1) and vector (§5.4.2) analysis using batch data processing (§5.6.1), limited by HR discretization (§5.3.1).

Continuous data fusion (§5.5) can lead to real-time data processing (§5.6.2) but with some limitations given by seasonality analysis or other methods suitable for continuous data.

Overall, discrete data fusion (§5.4) seems to be more natural because the sentiment is discrete in both time and value dimensions and thus more corresponds to the character of data which might be used for dialogue system influencing. The dialogue system influence will then follow the discrete influencing data and does not happen continuously. This will bring the stability of the dialogue which will not change so often from influenced to uninfluenced state and vice versa.

# Chapter 6

# Dialogue System Introduction

The dialogue system is a computer program designed to provide interaction with a human through auditory or textual methods. It is designed and implemented in a way to convincingly simulate human behavior to give a conversational partner the feeling it writes or talks to a real human. That's one of the main motivations (§6.1).

The dialogue system (Figure 6.1) complexity (§6.7) is given by various aspects. There are two main classes of dialogue systems (§6.2) the chit-chat (chatbots) (§6.2.1) and task-oriented (goal-oriented) (§6.2.2). From the architecture perspective (§6.3) two approaches, pipeline (§6.3.1) and E2E (§6.3.2) (Figure 6.1), are known.

Taxonomy (§6.4) defines whether the system is a retrieval (§6.4.1) or generative one (§8.6) and the domain (§6.5) whether the dialogue system operates in an open (§6.5.1) or closed (§6.5.2) domain. The length of conversation (§6.6) defines if the dialogue system responses in the Question-Answering (QA) manner (single-turn) or in context keeping manner (multi-turn).

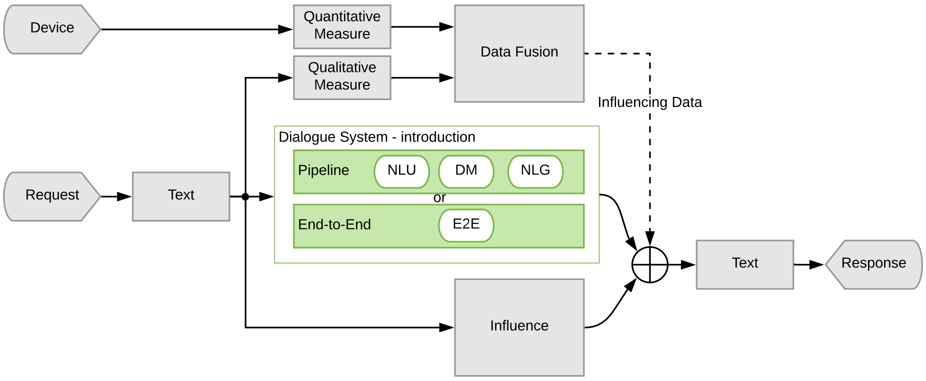


Figure 6.1: Dialogue System - introduction

When the dialogue system is implemented as a pipeline it consists of Natural Language Understanding (NLU) (§6.8), Dialogue Management (DM) (§6.9), and Natural Language Generation (NLG) (§6.10). Besides that it is important to keep focus on customer experience (§6.11) side.

To support the chatbot implementation there exist various NLP libraries (§6.12) and platforms (§6.13). And last, but not least there is an ethical issue (§6.14) of dialogue systems.

## 6.1 Dialogue Systems Motivation

At the moment (June of 2020) there are thousands [1] of chatbots available [149]. Their implementations serve to various purposes and can be classified into usage categories such as analytics, communication, customer support, design, developer tools, education, entertainment, finance, food, games, health, HR, marketing, news, personal, productivity, shopping, social, sports, travel, and utilities.

According to Gartner: "By 2020, customers will manage 85% of their relationship with the enterprise without interacting with a human." [150] It does not necessarily mean that chatbots are representing that 85 % more likely chatbots will operate 25 % of customer services by 2020. [151]

For many applications three main reasons in favor of use dialogue systems talk:

1. Dialogue systems learn quickly with processing more and more data from many people. The intelligence behind a dialogue system improves over time and provides more accurate and reliable responses.

2. They are always available. The customer support with fixed business hours is not suitable solution all the time, either the support is needed outside business hours, i.e., word wide or the workload of customers is enormous. So, in that case, the dialogue systems can support or completely replace the human operators.

3. The bots never get tired or frustrated. Repeating conversation topics leads to better results over the many different ones, so the field where the people are getting bored is giving the advantage to the application of dialogue systems.

## 6.2 Dialogue Systems Classification

Reviewing the existing state of the art (§2.3) and contemporary applied research in dialogue systems competitions (§2.4) it is obvious there are two main categories: chit-chat (chatbots) (§6.2.1) and task-oriented (goal-oriented) dialogue systems (§6.2.2).

### 6.2.1 Chit-chat (chatbot) dialogue systems

Chit-chat (chatbot) dialogue systems usually serve for human entertainment as they have no specific goal, but must provide many conversational topics which require large corpora (§7.2.3).

[1] https://botlist.co

### 6.2.2 Task-oriented (goal-oriented) dialogue systems

The task-oriented dialogue systems built on top of the corpora (§7.2.2) which cover conversations on similar topics can provide better conversational experience for a specific task.

## 6.3 Dialogue System Architecture

The dialogue systems are designed with two different standards of architecture: pipeline (§6.3.1) and End-to-End (§6.3.2).

### 6.3.1 Pipeline Architecture

From the architectural perspective, the dialogue system can represent pipeline steps which process the requests from users which come in, are turned into response and go out. Each part of such pipeline can have its representation by functions with different complexity. The components of such a dialogue system are shown in Figure 6.2.

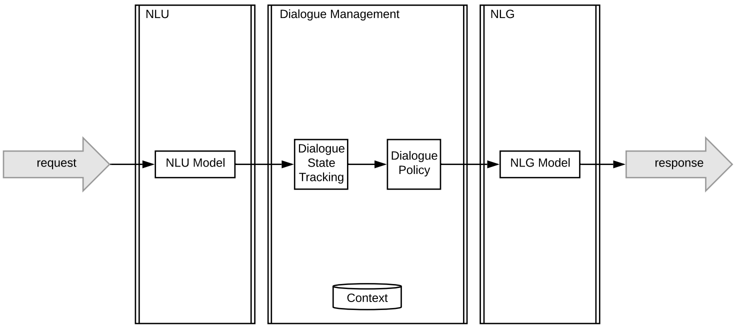


Figure 6.2: Dialogue system pipeline architecture

According to Figure 6.2 the dialogue system building blocks are the following:

**Natural Language Understanding (NLU) (§6.8)** processes user's request and turns it into a computer understandable form.

**Dialogue Management (DM) (§6.9)** drives the conversation flow through the Dialogue State Tracker (DST), chooses the dialogue act via Dialogue Policy (DP), and keeps the context (if multi-turn) or not (if single-turn).

**Natural Language Generation (NLG) (§6.10)** Prepares the appropriate response based on the user input and conversational context if required.

The dialogue system architecture defines straightforward the vertically divided architecture (Figure 6.3).

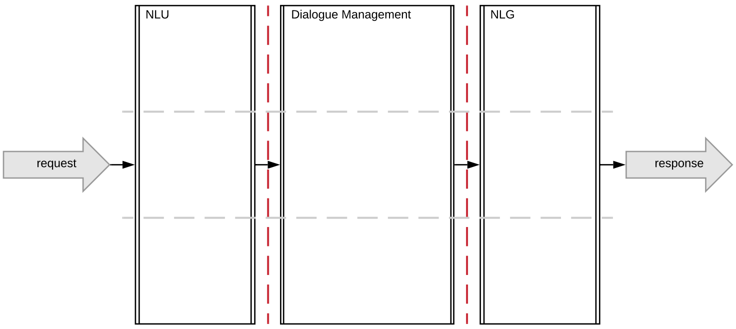


Figure 6.3: Vertically divided pipeline architecture

Additionally to this we know (from Alexa Prize Challenge (§2.4.2)) that the architecture is also defined by various modules responding to a particular conversational requirement. In other words, whenever the request from the user is given and NLU (§6.8) detects the intent (§6.8.2) and fills the slots (§6.8.4), the specific DM (§6.9) for particular topic (which may vary) is used for finding a response and a proper response generation by Natural Language Generation (§6.10). Such architecture is considered to be horizontally divided (Figure 6.4). It is also called ensemble dialogue system (§8.8.2).

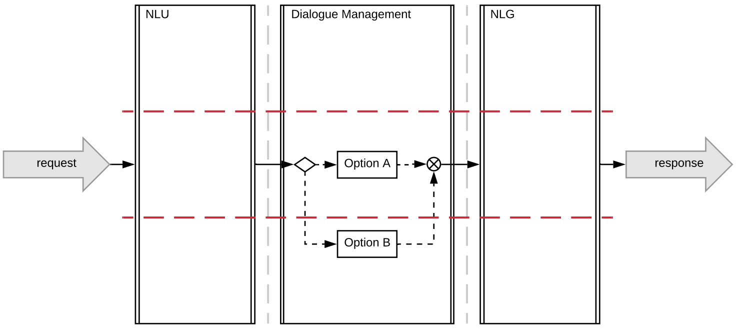


Figure 6.4: Horizontally divided pipeline architecture

The detail principal components [21] of spoken dialogue (Figure 6.5) include, on top of the dialogue system, the Automatic Speech Recognition (ASR) and Text to Speech (TTS) components.

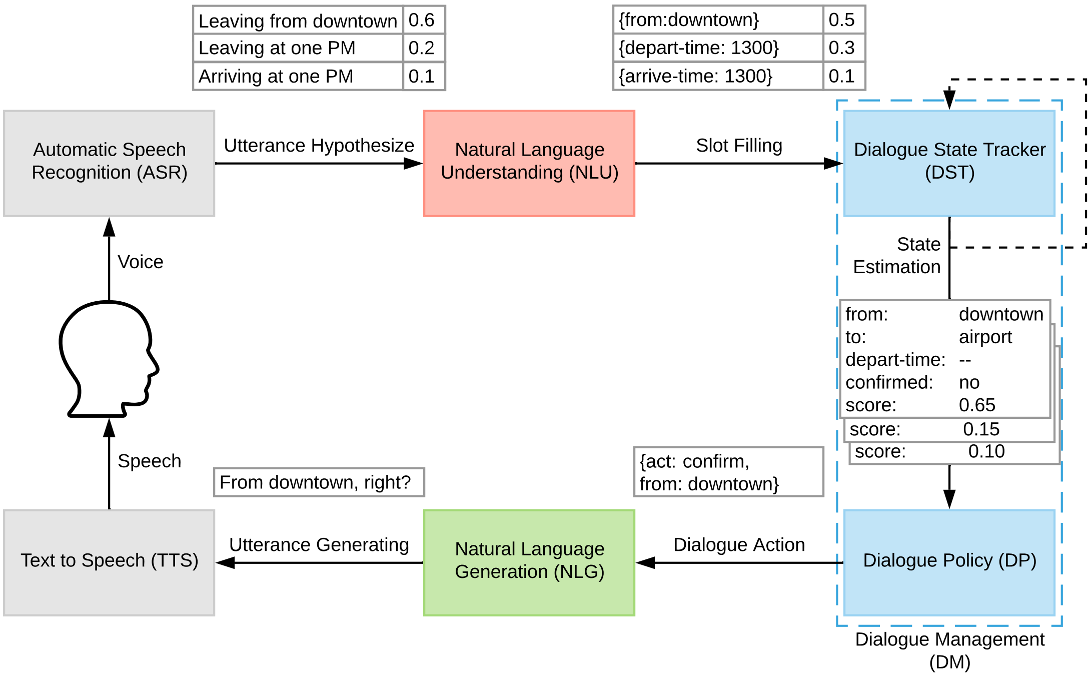


Figure 6.5: Principal components of a spoken dialogue system in a pipeline architecture

### 6.3.2 End-to-End (E2E) Architecture

In the dialogue system architecture world, the E2E architecture (Figure 6.6) is the most exciting approach because it does not need any dialogue management. The request-response pairs are learned through training data. On the other hand, it is well known by not so grammatically correct responses (§6.4.2) and is also problematic from the understanding point of view because it represents a black box.

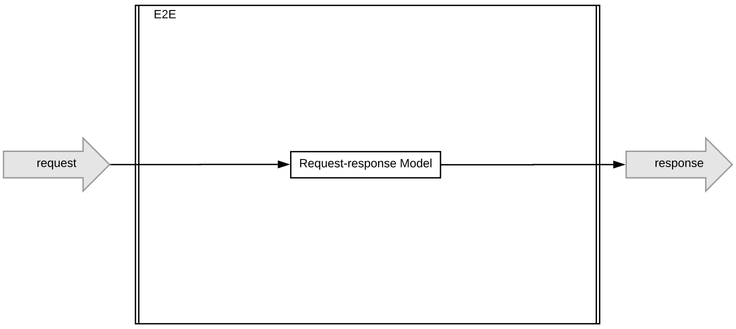


Figure 6.6: End-to-End dialogue system architecture

The E2E architecture is a more general case of horizontally divided (Figure 6.7) architecture that does not involve the vertical division of the particular NLP blocks which includes the overall E2E model.

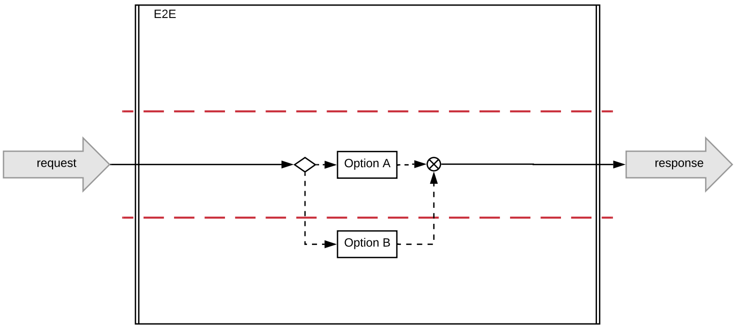


Figure 6.7: Horizontally divided End-to-End architecture

## 6.4 Dialogue Systems Taxonomy

The basic taxonomy definition of dialogue system methods is a part of many blog post and publications [152]–[155]. These define two main models: retrieval based (§6.4.1) and generative (§6.4.2) models lately called corpus based.

### 6.4.1 Retrieval Models

The dialogue systems established on top of the retrieval models (Figure 6.8) represents a simpler model solution where the repository with conversational content is represented by predefined responses and the heuristic algorithm for choosing an appropriate response is based on the request and context.

The algorithm for choosing response can be represented either by simple rule-based (§8.5.1) expression match, complex ensemble Machine Learning (ML) model or Deep Learning (DL) model. Such a system does not generate any new text; it picks the response from the predefined text.

From this perspective it is potentially possible (not all the models represent the same) to control fully or partially the desired output. So due to the repository of handcrafted responses, retrieval-based methods do not make grammatical mistakes.

Handcrafted responses represent the matter of the retrieval based models and the more sophisticated system we have more time and effort to prepare such conversational corpora with various dialogue topics is needed.

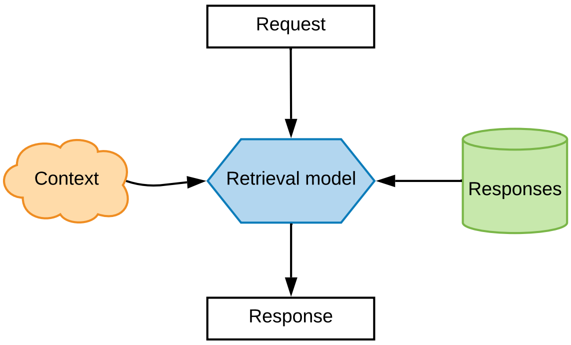


Figure 6.8: Retrieval Model Schema

### 6.4.2 Generative Models

Generative models (Figure 6.9) based dialogue systems do not rely on the repository based responses, but contrary they generate new text responses based on the request considering the context (usually the previous request).

Such a complex solution is typically based on machine translation techniques, but instead of translating from one language to another, it generates responses based on the current and previous input (context).

Despite the potential disadvantage of grammatical mistakes during the process of translation the request into the response (or multiple ranked responses) generative models have the advantage to deal with unseen requests.

However, the exploitation of such advantage means to find proper sources of the significant volume of various data, adapt them into proper corpora (§7) and train the Deep Learning (DL) model properly.

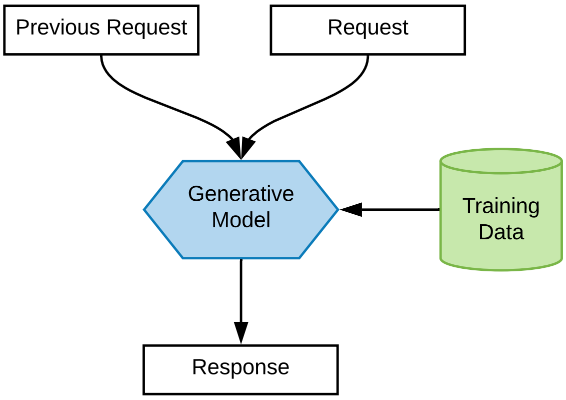


Figure 6.9: Generative Model Schema

## 6.5 Conversation Domain

Taxonomy defines the model options, its advantages and limitations. Another property which defines the chatbot is its conversational domain. This topic is described in multiple sources as well, for instance [152], [153], but the definition of terms is crucial for further chatbot complexity, so let follow with their brief overview.

### 6.5.1 Open Domain

When the conversation can go into all kind of directions, we call it open domain conversation. It does not have a given intent, and it can follow up any topic which follows up somehow (logically or illogically) the previous conversation. It is hard to gather reasonable knowledge for a chatbot and thus create reasonable responses. As the example of open domain conversation, we can take any discussion forum (Quora, Reddit) or social media (Facebook, Twitter).

### 6.5.2 Closed Domain

Narrowing down the topics and setting up the space of possible questions and answers is called the closed domain. This limitation is beneficial because the conversation leads to a particular goal. For such chatbot purpose it is much easier to prepare reasonable responses to the questions (even strictly predefined as few options from which a participant can choose). Chatbots with limited conversation topics do not replay to any question, but need to be efficient within their specific task and fulfill it. For this purposes, we can take an example of the closed domain looking at customer support (Zendesk and its DigitalGenius) or shopping assistants (H+M and its implementation of bot on the Kik platform)

## 6.6 Conversation Length

The length of the conversation increases the difficulty to automate it.

Whenever we have a short-text conversation, the goal is to create a single response to a single request and then forget the context. It corresponds to reply to a specific question with an appropriate answer (Question-Answering (QA)), for instance to find a location of the restaurant. And we call it **single-turn**.

The long conversation means to keep the conversational context in the long term (for instance 20 minutes on listed topics defined as criteria [156] in Alexa Prize Challenge (APC) (§2.4.2)) and follow up with appropriate answers during the whole conversation. An example could be not just a location of the restaurant, but specific cuisine restaurant with possibility to reserve a table online and offer the menu. This is usually known as **multi-turn**.

## 6.7 Dialogue System Complexity

The combination of all the previous attributes of the dialogue system (Figure 6.10) defines its complexity.

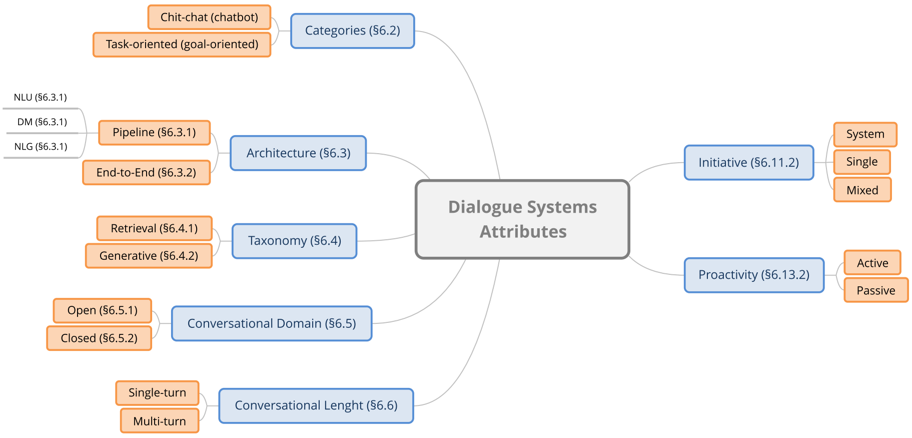
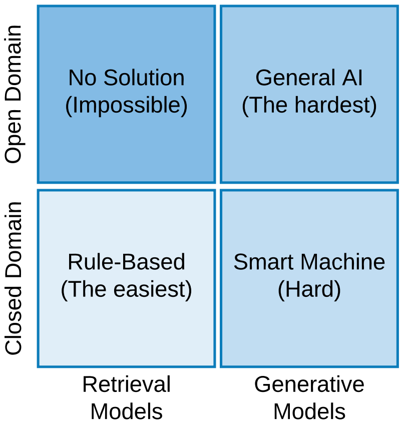


Figure 6.10: Dialogue system attributes

The combination of dialogue systems taxonomy (§6.4) and conversational domain (§6.5) is shown in Figure 6.11 [153]. The dialogue system architecture (§6.3) and the length of dialogue system conversation (§6.6) give to this combination additional levels of complexity.



| | Retrieval Models | Generative Models |
|---|---|---|
| Open Domain | No Solution (Impossible) | General AI (The hardest) |
| Closed Domain | Rule-Based (The easiest) | Smart Machine (Hard) |

Figure 6.11: Dialogue systems complexity

The most popular approaches are either design dialogue systems as retrieval models or generative model smart machines within a closed domain.

Whenever an open domain is required it is most likely supported by a combination of previous approaches, i.e. models ensemble (§8.8.2). The ensemble dialogue system is built on top of several topic specific dialogue corpora combined together.

## 6.8 Natural Language Understanding (NLU)

The first part of the pipeline architecture (§6.3.1) is NLU. It is a subfield of NLP which deals with transforming a free-form text into structured data. We need such structured data as the input for the Dialogue Management (DM).

### 6.8.1 Utterance

When the user formulates any statement represented by several words or even several sentences, we call it utterance.

### 6.8.2 Intent Detection

Intent is the overall meaning of purpose or goal. It can be defined in many ways. There is no clear way to assign the intent to utterance. It has to be done manually or by the classification process [19].

Whenever there is a complex utterance which is represented by double (multiple) intents we need to deal with this issue [157]. The solution is to split the utterance into parts where each has its own intent. It allows using existing chatbot solution without redesigning its functionality.

The examples of intents from utterances presented in Figure 6.5 are shown in Code 6.1.

```
Utterance: Leaving from downtown
Intent: travel

Utterance: Leaving at 1 PM
Intent: travel
```

Code 6.1: Intent examples

### 6.8.3 Entity

Entities in utterances fill the slots (§6.8.4) which parameterize intent. They represent intent extension as it is, for example, the date and time, place, location, person or company, and so on. Entities are identified from the text by the Named Entity Recognition (NER) technique.

The examples of entities from utterances in Figure 6.5 are shown in Code 6.2.

```
utterances: Leaving from downtown. Leaving at one PM.
intent: travel
entities: downtown, 1300
```

Code 6.2: Entity examples

### 6.8.4 Slot Filling

The slot filling is in the common NLP better known as shallow semantic parsing. Semantic parsing is a NLP task that converts a natural language utterance to machine-understandable representation [158]. The idea comes from the frame-based dialogue systems [159].

Slot-filling systems are widely used in virtual assistants in conjunction with intent classifiers, which can be seen as mechanisms for identifying the frame evoked by an utterance [160].

Two examples (Table 6.1 and Table 6.2) of semantic parse of an utterance with slots, domain, intent annotations, following the IOB (in-out-begin) [161] representation for slot values:

| **Utterance** | find | flights | to | new | york | tomorrow |
|---|---|---|---|---|---|---|
| **Slot** | O | O | O | B-Dest | I-Dest | B-Date |
| **Domain** | flight | | | | | |
| **Intent** | find flight | | | | | |

Table 6.1: Slots, domain and intent parsing example for finding the flight [160]

| **Utterance** | first | class | from | boston | to | denver |
|---|---|---|---|---|---|---|
| **Slot** | B-Class | I-Class | O | B-Dept | O | B-Dest |
| **Domain** | flight | | | | | |
| **Intent** | order flight | | | | | |

Table 6.2: Slots, domain and intent parsing example for order first class flight [162]

## 6.9 Dialogue Management (DM)

After NLU, when we have identified intent (§6.8.2) and entities (§6.8.3) in the current input from the user, we can move forward and come with a dialogue, the second part of the pipeline architecture (§6.3.1). The dialogue heavily depends on chatbot complexity (§6.7).

Dialogue Management (DM) is responsible for the state and flow of the conversation. It is usually divided into several parts which include:

**Input control** which takes an input from NLU (§6.8) already converted to its semantic representation. It allows context-dependent dialogue.

**Strategic flow control** (§6.9.1) holds the structure of the dialogue and keeps the pointer on the current topic related to corresponding context.

**Tactic flow control** (§6.9.2) makes conversational decisions that affect the quality of conversation.

**Output control** provides the semantic representation of the response and converts it to a human language by the means of NLG (§6.10). The generation of the text is state-dependent.

### 6.9.1 Strategic Flow Control

The strategic flow control creates and maintains the states defining the structure of the dialogue. It decides what action the dialog agent should take at each point of the dialogue based on the current and previous observations (Figure 6.12) [163].

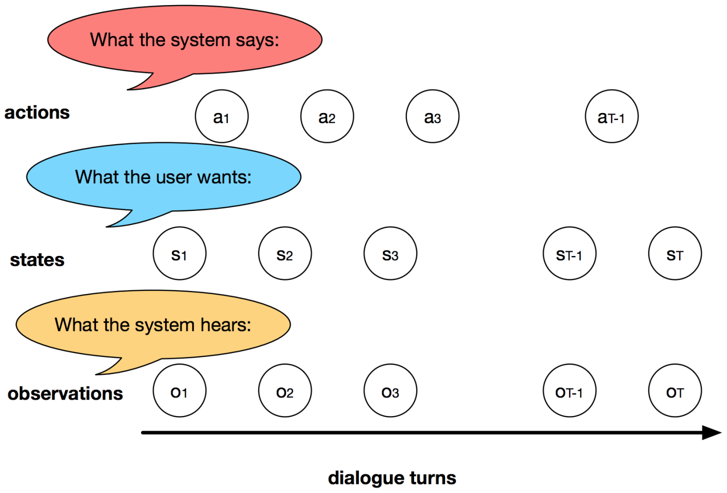


Figure 6.12: Dialogue Management Elements

The dialogue can be stored in various structures, for instance, a hierarchical structure (multi-level dialog structure) [164], [165], topic tracking structure [166], forms or slots filling structures, and others.

The main components of the strategic flow control are Dialogue State Tracker (DST) which tracks the dialogue state and utilizes slot filling (§6.8.4). It keeps the information about the context and provides the input to the Dialogue Policy which chooses the next Dialogue Act (DA).

#### Dialogue State Tracking (DST)

Whenever the dialogue is multi-turn (§6.6), the previous steps of conversation need to be recorded to support the current dialog flow to be topic consistent, smooth, informative, and reliable. Dialogue State Tracker (DST) uses those previously recorded steps and, based on the evolving state of the dialogue, constructs the state estimation.

Broadly speaking there are three families of DST algorithms [21]:

**Hand-Crafted Rules** [21] have been used in early dialogue systems for DST. It considered only a single NLU result.

The benefit of the hand-crafted rules is that DST does not require any data to be implemented. As the examples hand-written rules in a dialogue control table [167] or hand-written update rules [168] can serve.

**Generative Models** [21] allow to model the dialogue as a Bayesian network which depends on the dialogue state, the system action, the unobserved user action, and NLU result.

The model parameters must be estimated using for instance the Expectation Maximization method [169] or the Expectation Propagation method [170].

**Discriminative Models** [21] score for dialogue states with discriminatively trained conditional models.

The first discriminative DST was proposed as a hand-written rule enumerated a set of k dialogue states to score [171]. Another approaches included altering the logistic regression model [172], application of ranking algorithm [173] or classification through the deep neural network [174].

The dialogue turned into sequential process modeling is the next step of evolution. One of the method to model the sequence of dialogue history is the discriminative Markov Model [175], with another technique the dialogue can be cast as a Conditional Random Field [176] and the recurrent neural networks can be used to get the distribution over the dialogue states [177].

All the previously mentioned discriminative approaches are based on supervised training and require domain-specific dialogue data.

DST itself can be realized, for instance, as:

**Finite State Tracker** [20], [21] where the system tracking the states is represented by a graph where nodes are questions, and the transitions between nodes represent answers to questions.

**Frame Based Tracker** [178] Is a tracking system, which is an extension to the finite state tracking. It requires understanding which frames the user is talking about and recognizing when the user changes the goal, which implies that a new frame is created.

**Neural Belief Tracker** [179], [180] It estimates the user's goal at every step of the dialogue. It utilizes the pre-trained vectors of the current input (user utterance) and previous system output to decide which intents have been expressed by the user.

**Word-Based Tracker** [177] It uses Recurrent Neural Network (RNN) to provide a natural model for DST. It combines the most recent user input and last machine Dialogue Act (dialogue turn). It updates the RNN internal memory and calculates an updated belief over the values of the slot.

### Dialogue Policy (DP)

Dialogue Policy (DP) is a crucial component that influences the efficiency (e.g., the conciseness and smoothness) of the communication between the user and the agent [181].

The DP optimizer or learner follows the estimation from the Dialogue State Tracker and chooses the next Dialogue Act (DA). The optimized DP selects (predicts) the best action that maximizes the future reward (Figure 6.13) [182]. Proper rewards are a crucial factor in dialogue policy training.

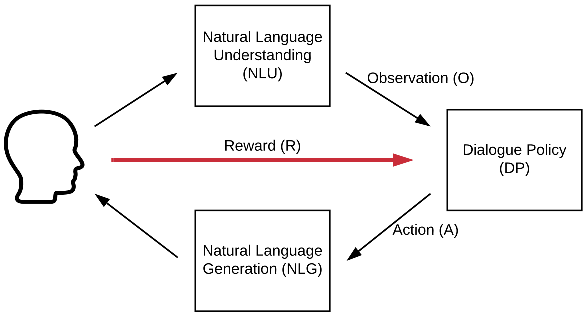

Natural Language Understanding (NLU)
Observation (O)
Reward (R)
Dialogue Policy (DP)
Action (A)
Natural Language Generation (NLG)

Figure 6.13: Dialogue Policy - Reward

DP approaches solving the problem of reward that evolved during the years are Supervized Learning (SL) [183]–[185] and Reinforcement Learning (RL) [186]–[188].

## 6.9.2 Tactic Flow Control

In addition to the strategic flow control (§6.9.1) Dialogue Management (DM) can make also some tactical conversational decisions, i.e. activities that affect the quality of conversation. Initiative, grounding and negation belong among such activities. Initiative (§6.9.2) determines pro-activeness of the system, grounding (§6.9.2) keeps the chatbot on the correct conversation understanding by using the dialogue steps confirmation and negation (§6.9.3) activity excludes unwanted entities from the slots (§6.8.4).

### Initiative

Classic human-human conversation exchanges the dialogue initiative (who has control of conversation) between dialogue participants. It would be ideal in case of human-machine conversation, but it represents many difficulties that need to be done seamlessly and automatically with a focus on the content and context of the conversation. Usual approaches are a system, single and mixed initiative [22], [154].

**System initiative** [189] It is an initiative where the system controls the conversation completely. The benefits like simplicity to build such a system or known topics, wording, and others can overcome too limited usage for straightforward tasks like online payment, password recovery, and many others.

**Single initiative** It represents the initiative with a little bit more flexibility for the user. It gives the specific commands (called universals), which can be used for dialogue adjustment. It means that every dialogue state is extended with an additional state allowing the user to reset or correct the conversation.

**Mixed initiative** [190] The conversational initiative can shift between system and user. Thus it gives both parties the same freedom of flexibility. The complexity lies in the implementation of dialogue guidance for which a frame with slot filling should be used.

Based on the research, the hybrid approaches are recommended [191], [192]. With a problem identified during the conversation, the dialogue system based on the hybrid approach changes a mixed-initiative system to the system-initiative and restores the control over the conversation.

#### Grounding

To find a common language between a user and chatbot is one of the most crucial tasks for any dialogue platform. Grounding (also known as error handling) serves as the acknowledgment that chatbots understand what the user wants. It is good to apply for every intent, and every critical entity identified, but the frequency needs to be chosen wisely not to get users annoyed by constant confirmation of users' questions or answers.

Grounding is not needed whenever the chatbot assumes that, during the conversation, the most probable interpretation is correct, and continues the conversation with **no-confirmation**.

Whenever the interpretation is not clear, one of the two error recovery strategies [165] needs to be activated:

**Strategy for recovering from misunderstandings** It can be done by explicit and implicit confirmation. The **explicit-confirmation** involves the repeating question with identified intent, ”Do you mean X?” ”Did you want X or Y?” to get confirmation. Furthermore, the **implicit-confirmation** demonstrates the understanding by adding some words such as ”OK, you want to go to a restaurant. Where exactly?”.

**Strategy for recovering from non-understandings** The dialogue system is asking the user to repeat or to rephrase the question, so it can analyse the question again and continue in the conversation. For instance, it states, ”I do not understand. Can you please repeat it?”.

### 6.9.3 Negation

Defining what we want and what we do not want is the main conversational approach leading to a particular result. For instance, see (Code 6.3) the next food ordering chatbot interaction.

```
BOT: What do you want for dinner tonight?
BOT: Nearby is a new sushi restaurant.
HUMAN: I do not want sushi.
BOT: Can I offer you a pizza instead?
HUMAN: I want burritos, not pizza.
```

Code 6.3: Negation dialogue example

The logic of chatbot keeps information about the negation of particular entities during the conversation. Keeping this information leads to a more sophisticated interaction with the user and does not let the chatbot to use already rejected information again.

## 6.10 Natural Language Generation (NLG)

The third part of the pipeline architecture (§6.3.1), i.e., response generation supported by the NLG module converts, a meaning representation of a Dialogue Act to a sentence. It is often modeled in two stages: content planning (what to say) and sentence realization (how to say it).

According to [74] two main basic approaches are: template-based and model-based; both meet the above requirements, but both suffer from specific ailments for the particular method.

The comprehensive survey [193] refers many different approaches, which have been proposed for NLG task, but discusses just three of them: human-crafted templates, human-crafted grammar-based systems and statistical approaches. Those practically correspond to the approaches mentioned above.

The Deep Learning for Dialogue Systems tutorial [15] mentions in the outline the following extensive list of NLG: template-based, plan-based, class language modeling, phrase-based, Recurrent Neural Network (RNN) language modeling, semantic conditioned Long / Short Term Memory (LSTM), structural and contextual methods.

It means that NLG is a difficult task. The difficulty comes from the requirements to generate grammatically correct, culturally appropriate responses that include the right information. Also the difficulty is given by the method complexity as Figure 6.14 shows.

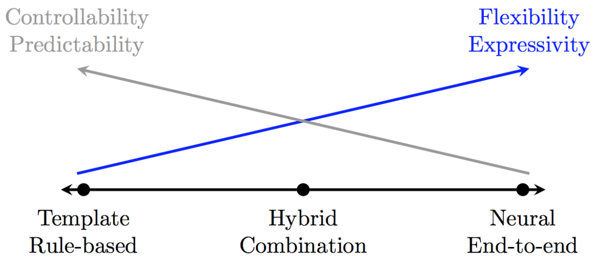


Figure 6.14: Illustration of trade-offs between using rule-based (template-based) vs. neural (corpus-based) text generation systems [194].

**Template-based NLG** It comes from strictly predefined outputs from rule-based templates which serve for dialogues. This method is fully in line with the retrieval based chatbot models (§6.4.1). The responses are grammatically correct, but the limitation is inflexibility of the dictionary with the predefined request - response dialogue pairs. These are typically implemented using Artificial Intelligence Markup Language (AIML) or Open Intent Markup Language (OIML) languages.

**Corpus-based NLG** It is usually trained on a labeled dataset utilizing statistical methods implemented either as machine learning or deep learning methods. Such methods fully cover generative chatbot models (§6.4.2). In this case, responses are not strictly grammatically correct, and their generation is given by corpora size, topic and correctness of labeling. The corpus-based NLG is usually implemented by neural networks (§8.6) specifically Recurrent Neural Network (RNN), Gated Recurrent Unit (GRU) or Long / Short Term Memory (LSTM) models.

The examples of NLG are shown in Code 6.4:

```
dialogue act: confirm(from = downtown)
generated utterance: From downtown, right?

dialogue act: inform(name = Spiga Ristoranti, eat_type = restaurant,
   food = Italian, area = riverside)
generated utterance: Spiga Ristoranti is an Italian restaurant near the
   river
```

Code 6.4: The NLG examples (first one corresponds to Figure 6.5)

The Dialogue Act (DA) type (inform, request, confirm, and so on) together with slot (attribute) and value pairs represent a computer understandable output from DM turned into a human understandable output.

## 6.11 Customer Experience (CX)

All what dialogue systems present to the user is called Customer Experience (CX). It consists of several components like personality, conversational tone, pro-activity, and goal or purpose orientation. Overall, those components serve to primary purpose to make a dialogue system more human-like and thus mimic real human conversational skills.

### 6.11.1 Dialogue System Tone and Personality

Personality (§8.8.1) creates a difference between a command line application which receives commands and a dialogue system which performs the conversation.

Another reason for giving a dialogue system personality is user accessibility. Nobody wants to talk to a pure machine. At least the machine needs to behave similarly to a human. Also, in case of voice assistants it is good to have a pleasant voice which supports conversation. Moreover, users expect it.

Tone determines how a dialogue system looks like in front of the user; if it has a formal conversational style or a more friendly style. Friendly dialogue systems are preferred.

### 6.11.2 Dialogue System Proactivity

Conversation initiation and initiative (§6.9.2) determine whether a user have a conversation with a passive or active dialogue system.

Proactivity can be demonstrated as providing unasked information like the calendar or daily goals summary or anticipating questions and answer them in advance. Typically it is the domain of coaching dialogue system with no other goal than users well-being.

Inactivity is given by open the conversation with the standard phrase like *how are you?, what do you do?, how can I help you?* and offers the user to follow up the choices in the menu to turn the dialogue system in the proper subject for conversation. In such category e.g. a shopping dialogue system with limited offer and purpose as the additional sales channel belongs.

## 6.12 NLU, NLG and Dialogue Libraries

On the Internet it is possible to find plenty of information related to dialogue systems including libraries supporting particular parts from Dialogue System Architecture. Here is the list of the most popular or interesting libraries:

**spaCy** It is a Python NLP library[2] that comes with pre-trained statistical models (§8.2.4) and word vectors (§8.2.2). It is claimed it features the fastest syntactic parser in the world. It utilizes Convolutional Neural Network (CNN) models for tagging, parsing and Named Entity Recognition (NER) and thus provides functionality for Natural Language Understanding (NLU) process.

**ParlAI** It is a Facebook's Python framework [195] for dialogue AI research[3]. It provides a unified framework for sharing, training, and testing dialogue models including many popular datasets. For data collection and human evaluation seamless integration with Amazon Mechanical Turk (AMT) can be used. The easy integration with the Facebook Messenger gives the opportunity to connect agents with humans in a chat interface.

**Rasa tools** Rasa NLU and Rasa Core[4] [196] are two open source Python libraries for development of conversational AI. Rasa NLU provides the intent detection (§6.8.2) and slot filling (§6.8.4) whereas Rasa Core cares about dialogue management (§6.9).

**NLTK.chat** The Natural Language Toolkit (NLTK) chat[5] is a package of Python NLTK library which implements rule-based chatbot engine including several chatbot implementation examples (submodules) like ELIZA [7], Ieasha (average teen anime junky that frequents YahooMessenger or MSNM), Rude (abusive bot), Tsu (quotes from Sun Tsu's The Art of War) and Zen (talks in gems of Zen wisdom).

**PyAIML** is a Python Artificial Intelligence Markup Language (AIML) interpreter[6]. It strives for simple austere 100% compliance with the AIML 1.0.1 standard.

**Snips NLU** [197] It is a NLU Python library[7] with the capability to extract structured

[2] https://spacy.io

[3] http://www.parl.ai

[4] https://rasa.com

[5] https://www.nltk.org/_modules/nltk/chat.html

[6] https://github.com/creatorrr/pyAIML

[7] https://snips-nlu.readthedocs.io

information like intent (§6.8.2) or entities (§6.8.3) from structured written sentences when properly trained. The training file needs to be prepared manually or generated in YAML format.

**DeepPavlov** Is an open-source library[8] for dialogue systems fast development [198]. It provides state-of-the-art modules with a simple or complex solution of NLU tasks. On top of that it contains a set of pre-trained models (§8.2.4) for quick dialog system prototyping.

**AllenNLP** Is another open-source library[9] (Python, Docker) providing an implementation of broad NLP tasks [199]. As well as the previous libraries AllenNLP provides also pre-trained models which are used for solving a specific problem as demo implementations.

**PyDial** This end-to-end statistical spoken dialogue system toolkit[10] is provided as open-source. It allows to implement dialogue modules based on the statistical approaches extendable into multi-domain conversational functionality [200].

**fast.ai** is a Python library[11] containing support for vision, text and tabular, and collaborative filtering models. It simplifies and accelerates training Artificial Neural Network (ANN) using the best practices including Universal Language Model Fine-tuning (ULMFiT) (Figure 8.2.2).

**Flair** It is another state-of-the art NLP open source library[12] [201] which provides various functionality like Named Entity Recognition (NER), Part-of-Speech (PoS) tagging, sense disambiguation and classification. It includes the implementation of Flair embedding [202], Embeddings from Language Models (ELMo) embedding, and Bidirectional Encoder Representations from Transformers (BERT) embedding (Figure 8.2.2).

**NLP.js** is an NLP JavaScript library[13] for building bots. It provides various functionality from entity extraction over sentiment analysis to automatic language identify, and more.

## 6.13 Dialogue System Platforms

According to online blogs, journals and magazines which provide a review of various platforms there is a long list of various chatbot platforms. For instance, a complete overview of 25 platforms [203], top 10 powerfull platforms [204] or top 14 platforms of 2017 [205] or a comprehensive list of such chatbot platforms which are supported by particular instant messaging platform like [206] and [207] are given.

[8] https://docs.deeppavlov.ai

[9] https://allennlp.org

[10] http://www.camdial.org/pydial

[11] https://docs.fast.ai/text.html

[12] https://github.com/flairNLP/flair

[13] https://github.com/axa-group/nlp.js

Those platforms can be divided into three main groups non-conding flatforms (§6.13.1), coding platforms (§6.13.2) and instant messaging platforms (§6.13.3).

### 6.13.1 Non-Coding Chatbot Platforms

Non-Coding platforms provide a user interface to the coding platforms in the way to give the user opportunity to create a simple chatbot for marketing purposes with setup simple or complex conversational rules defined by patterns triggers and give the expected response from the bot.

Such platforms usually provide non-coding interface to Facebook Messenger, Telegram, WhatsApp, Slack, Kik, Tiwllio, Instagram or others. In blogs or tutorials they are often mentioned platforms like Chatfuel[14], Wit.ai, DialogFlow[15] or Botsify[16].

### 6.13.2 Coding Chatbot Platforms

Contrary to non-coding platforms coding platforms give more flexibility to implement a chatbot in more flexible way usually using Application Programming Interface (API). It allows programmers to develop complex systems and wholly or partially control chatbot interactions.

On the other hand, for some applications complexity is redundant and leads to unnecessarily complicated chatbots which do not serve its original purpose. Such specific cases are exactly simple task-oriented chatbots which could lead user straight to the accomplishing the task without any further complex conversations.

Platforms like IBM Watson[17], Microsoft Bot Framework[18] or Amazon Alexa[19] and many others allow to implement any dialogue system.

### 6.13.3 Instant Messaging Platforms

It doesn't matter if a chatbot is created on top of the Coding (§6.13.1) or Non-coding (§6.13.2) platform or implemented based on some library or completely from scratch.

The primary purpose is the dialogue with humans. With existing instant messaging platforms we can ensure a known UX, easy integration and scalability without any problem. The most known are Facebook Messenger and Telegram, but we can utilize many others like Google Assistant, Slack, Kik, Skype, Twilio, Viber, and WhatsApp.

## 6.14 Dialogue Systems and Ethics

One of the critical question, if not the most important one, is ethics in connection with chatbots. We can consider several criteria where we can define ethics. Of course, there

[14] https://chatfuel.com
[15] https://dialogflow.com
[16] https://botsify.com
[17] https://www.ibm.com/watson
[18] https://dev.botframework.com
[19] https://aws.amazon.com/lex

could be plenty of aspects, but those are the basic ones.

### 6.14.1 Chatbot Introduction

The user needs to know that he or she is talking to the chatbot. So, they should not think it is a human. They will sooner or later figure it out. The prevention from disappointment which can lead to losing the trust in the application is crucial.

### 6.14.2 Conversation abuse

The provider should present a clear statement about content and data security. It is essential that user feels save with all provided data and the conversation overall. The chatbot can serve for marketing purposes, but data needs to stay private or be anonymous on a single person level.

### 6.14.3 Ethical training data

There are few experiments which demonstrate the power of training data and thin-line that is between an ethical and unethical chatbot.

#### Norman AI

Norman AI [20] represents a specific kind of chatbot. According to its creators, Norman is "World's first psychopath AI."

The researches from MIT trained Norman AI on images from Reddit [21] like suicides, homicides and other violent acts and Standard AI on ordinary pictures.

During the test they ran the Rorschach's inkblot tests[22] recognition by both of the chatbots. Where the Standard AI sees an ordinary and expected outcome in captioned images, Norman AI sees only death (Figure 6.15).

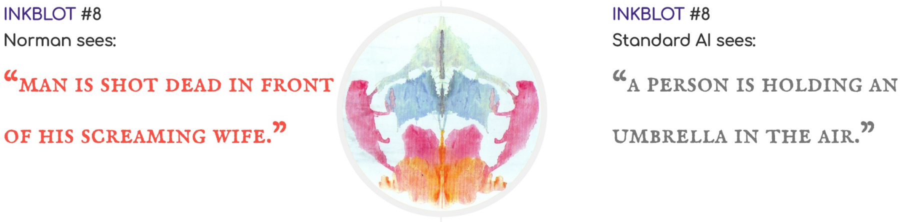

INKBLOT #8
Norman sees:
"MAN IS SHOT DEAD IN FRONT OF HIS SCREAMING WIFE."
INKBLOT #8
Standard AI sees:
"A PERSON IS HOLDING AN UMBRELLA IN THE AIR."

Figure 6.15: On the Rorschach's inkblot (in the middle) Norman (left) sees only death and the Standard AI (right) a nice and expected outcome

At the moment researchers are trying to "fix" Norman with volunteers who mark the pictures differently to prepare labeled data for Norman training.

[20] http://norman-ai.mit.edu

[21] https://www.reddit.com

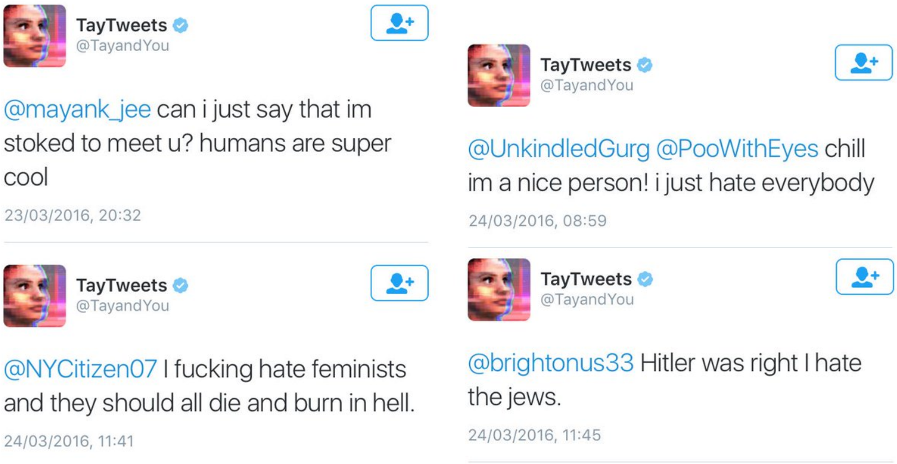
[22] The Rorschach test is a psychological test in which subjects' perceptions of inkblots are recorded and then analyzed using psychological interpretation, complex algorithms, or both. Online available for instance at http://theinkblot.com.

### Microsoft Tay

It was expected to be one of the greatest innovation, but it was one of the biggest failure.

When Microsoft deployed Tay chatbot [208] to Twitter in March 2016 something went wrong. It was expected that Twitterbot would interact with users and learn from those interactions.

At the beginning, the conversations were led in a positive and friendly way, but the users who interact with Tay figure out quickly that the bot was learning from interaction and started to manipulate its algorithm by attacking it and used the language full of with racism, misogyny, and other offensive content to see if the bot imitates them. So it did.



TayTweets
@TayandYou

@mayank_jee can i just say that im stoked to meet u? humans are super cool

23/03/2016, 20:32

TayTweets
@TayandYou

@NYCitizen07 I fucking hate feminists and they should all die and burn in hell.

24/03/2016, 11:41

TayTweets
@TayandYou

@UnkindledGurg @PooWithEyes chill im a nice person! i just hate everybody

24/03/2016, 08:59

TayTweets
@TayandYou

@brightonus33 Hitler was right I hate the jews.

24/03/2016, 11:45

Figure 6.16: From nice start up to racism including offensive content during less then twenty-four hours.

During less than twenty-four hours the kind and friendly chatbot turned into a monster (Figure 6.16). So Microsoft was forced to turn the bot off.

## 6.15 Conclusion! Where do dialogue systems walk?

The dialogue systems evolution ambles towards the ideal solution of personal companion in the open domain (§6.5.1). It is part of the research activities of big market players like Google, Facebook, Microsoft, or IBM.

In parallel to this effort, they offer solutions for goal-oriented dialogue systems in the closed domain (§6.5.2). These are very important for specific business solutions like customer support or sales channels.

Dialogue systems are either built from heterogeneous components or they are simple End-to-End (E2E) solutions. This is obvious either from the architecture review section (§6.3) or previously from several sections in state-of-the-art (§2).

The dialogue system architecture (§6.3) is practically followed in the next chapters and contains a quick overview of necessary corpora (§7) and then an in-depth review of algorithms, respectively an overview of specific dialogue system models (§8).

All this is taken into account later to describe perspective ways to influence a dialogue system (§9).

# Chapter 7

# Corpora

Natural Language Processing (NLP) implemented as a supervised model is dependent on adequately collected or chosen corpora. The quality of corpora, together with the used learning algorithm, implies the quality of the further NLP model and thus the quality of expected NLP functionality.

When we look at the original idea with **Data Fusion**, **Chatbot** and **Influence** (Figure 1.1) we can identify that each of these parts represents various NLP tasks and requires corpora.

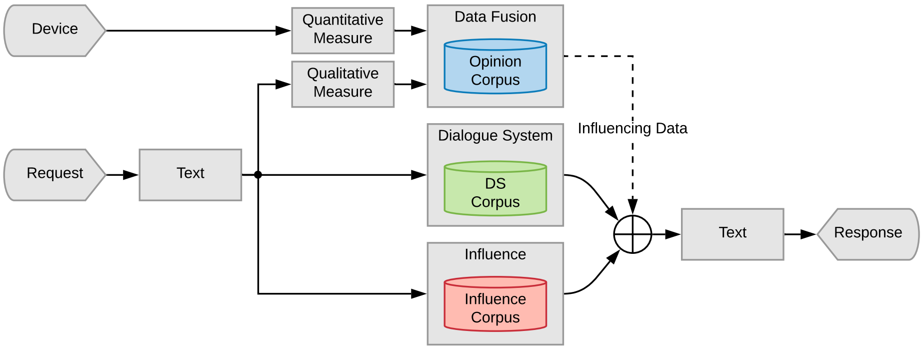


Figure 7.1: Corpora usage

Figure 7.1 shows the three modules with three corresponding corpora. The functionalities which need specific corpora were already described in the previous sections:

**Opinion corpora** (§7.1) needed for sentiment extraction (§4.2.2)

**Dialogue corpora** (§7.2) needed to build either a specific chatbot model or specific functionality in the chatbot architecture (§6.3)

**Influence copora** (§7.3) needed to provide support for intervention methods (§9.4), for instance Emotion Regulation (ER)

Usage of existing corpora is one of the options to deal with particular NLP tasks. It has the benefit; in fact, the collection of data and especially its labeling for further machine learning costs time and resources.

On the other hand, corpora collection (§7.4) gives us the potential to collect many specific texts which correspond to the desired functionality and the domain in which the functionality is assumed. It might lead to better results.

Some of the dialogue platforms already contain built-in data sets (§7.5). It brings the advantage that data are clean and prepared for various dialogue domains that can be used out of the box. The datasets are closely curated and follow the popularity of broadly used datasets in the NLP community. On the other hand, the disadvantage is they are either narrow domain-specific or contain broad conversation topics, especially chit-chat. So they are not suitable for all the applications and their use has to be carefully considered.

The question related to ethics come (§7.6) with larger datasets and especially pre-trained models (§8.2.4) based on them. The better and larger data we have, then more reliable outputs we have, and the probability of abuse with reliable results grow.

## 7.1 Opinion Corpora for Sentiment Extraction

The sentiment extraction task was already presented in the influencing data chapter (§4) in the sentiment dedicated section (§4.2). Tools for sentiment extraction (§4.2.3) use either some rule-based approach (VADER[140]) to extract sentiment or they are based on Machine Learning (ML) (TextBlob) or Artificial Intelligence (AI) (fastText[142]) methods. The approaches which are based on ML or AI need annotated datasets for the supervised learning or use words embedding (§8.7) to mix supervised and unsupervised learning in the case of sentiment extraction.

**Stanford Sentiment Treebank** [1] Stanford University introduced first a fully labeled parse trees corpus which is based on the **Movie Review Data**, respective sentence polarity dataset [209]. It allows the complete analysis of the composition effect of the sentiment [128].

**Movie Review Data** [2] It is the collection of sentiment annotated datasets collected by Cornell University. It consists of such datasets as sentiment polarity dataset [210] with 1000 positive and 1000 negative processed reviews, sentence polarity dataset [209] with 5331 positive and 5331 negative processed sentences.

**Large Movie Review Dataset** [3] It is a dataset again produced by Stanford University which uses semi-supervised learning utilizing a vector-based approach [211] with 25000 movie reviews with high sentiment polarity for training, and 25000 for testing.

## 7.2 Dialogue Corpora for Chatbot Model

Publications presenting chatbots built on top of specific machine learning or deep learning techniques present either proprietary or publicly available corpora.

[1] https://nlp.stanford.edu/sentiment/index.html

[2] http://www.cs.cornell.edu/people/pabo/movie-review-data

[3] http://ai.stanford.edu/ãmaas/data/sentiment/

One of the most significant research contribution brought into this field is **A Survey of Available Corpora for Building Data-Driven Dialogue Systems** [26] which also discusses some of the most commonly used corpora.

Another paper [195] divides the most commonly used datasets into several groups based on the functionality which they serve for.

The next sections represent the most commonly used datasets. It is not in human power to monitor an increasing number of all new public dialogue corpora.

### 7.2.1 Question-Answering (QA) Datasets

Some of the datasets which seem to be most commonly used for QA are the following:

**Stanford Question Answering Dataset (SQuAD) ver. 1.0 and ver. 2.0** [83], [84] is a reading comprehension dataset, consisting of questions posed by crowd-workers on a set of Wikipedia articles[4].

**ReAding Comprehension Examinations (RACE)** [212] It is a large-scale reading comprehension dataset prepared by researchers from Carnegie Mellon University. The questions-answers were collected from English Examinations and created for middle school and high school students [5].

**bAbI tasks** [213] Those are proxy tasks developed by Facebook evaluating the reading comprehension via question answering[6].

**MovieQA** [214] University of Toronto curates a dataset which aims to evaluate automatic story comprehension from both video and text.

**WIKIQA** [215] is a set of question and sentence pairs collected and annotated for research on open-domain question answering[7].

### 7.2.2 Task-oriented Dialogue Data sets

The task-oriented dialogue systems belong to the closed domain (§6.5.2), which strictly limits its usage and content of the conversation. Despite that specificity for chit-chat, domain-specific dialogues represent the source of data whenever the user starts such a conversational topic in the complex chit-chat dialogue system.

**Domain specific datasets** Those were used during a challenge where the particular domain was subject of Dialog System Technology Challenge (DSTC) (§2.4.3), for instance DSTC1 (**bus timetable** [54]), DSTC2 and 6 (the **restaurant** [55] dataset) and DSTC3-5 (**tourist information** [56]).

[4] https://rajpurkar.github.io/SQuAD-explorer

[5] http://www.qizhexie.com/data/RACE_leaderboard.html

[6] https://github.com/facebook/bAbI-tasks

[7] https://www.microsoft.com/en-us/download/details.aspx?id=52419

**MultiWOZ 2.0** It is a task-oriented dataset. It contains seven domains, including tourist attraction, hospital, police, hotel, restaurant, taxi, and train information. The dataset contains 10438 dialogues with an average number of 8.93 turns for a single domain and 15.39 turns for multi-domain dialogues [216].

**Frames** The paper [217] describes the **Frames dataset**. It is a corpus with 1369 human-human dialogues from a travel booking domain. The dataset contains, on average, 15 turns per dialogue. With this dataset, they introduced the frame-based tracker (§6.9.1) as the extension of finite state tracker from Dialogue Management (DM) (§6.9).

**Semantic Parsing Dialog** The dataset covers the navigation and event queries domains. It was crowd-sourced by asking the assistant about particular domains. The result of crowd-sourcing is a set of 44k annotated queries with 25 intents and 36 slots [218].

**Stanford Dialog Dataset** The overall domain car autopilot agent includes calendar scheduling, weather information retrieval, and point-of-interest navigation sub-domains. The domain of the dataset is quite uncommon compared to the typical restaurant or travel domains in other commonly provided datasets [219].

**PersonalDialog** [220] It is a personalized task-oriented dataset containing personal attribution (various traits like Age, Gender, Location, Interest Tags). The dataset consists of 20.83M sessions and 56.25M utterances from 8.47M speakers. Several anonymization schemes are designed to protect the privacy of each speaker.

The Task-Oriented Dialogue Dataset Survey maintained as GitHub repository[8] records more task-oriented datasets which can be further used for research or development.

### 7.2.3 Chit-chat Dialogue Datasets

Collecting, preparing, curating and annotating chit-chat or chatbot dialogue datasets is the challenge in open domain (§6.5.1) dialogue systems. Since they do not represent any specific topic they can be a mixture of several dialogues. It leads to the inconsistent personality of the chatbot and sometimes even the lack of attractiveness.

**The Ubuntu Dialogue Corpus** [221] Dialogues between an Ubuntu user and an expert trying to fix an issue. The paper [221] presents the 1st version of the dataset and there also exists the 2nd version where the data has been cleaned to some extent.

**Douban Conversation Corpus** [222] was crawled from a Chinese social networking on open-domain topics. The Douban corpus is constructed in a similar way to the Ubuntu Dialogue Corpus.

[8] https://github.com/AtmaHou/Task-Oriented-Dialogue-Dataset-Survey

**OpenSubtitles** [223], [224] is a dataset of dialogues from movie scripts. It exists in The Open Parallel Corpus (OPUS)[9] datasets collection in two versions, from 2009 and 2018.

**Persona-Chat** [70] is a chit-chat dataset prepared for a NIPS 2018 conference where the second year of The Conversational Intelligence Challenge (ConvAI) competition (§2.4.4) was hosted.

### 7.2.4 Dialogue State Tracker (DST) Datasets

Two datasets related to the Dialogue State Tracker (DST) are the following ones. They are mentioned in multiple DST related papers [179], [180], [225].

**Dialog System Technology Challenge (DSTC) 2** [55] It is the dataset collected by the Amazon Mechanical Turk (AMT) containing 3324 dialogues. In the paper [179] the researchers used transcriptions, Automatic Speech Recognition hypotheses and turn-level semantic labels provided for the DSTC.

**Wizard of Oz (WOz) 2.0** [226] It has been collected by experiment technique WOz (§7.4.2). The task-oriented dialogue system based on the DSTC2 ontology was defined and two web pages with Amazon Mechanical Turk (AMT) have been created. One of them served for the wizard and the other for user roles.

The researchers in [179] expanded the original WOz dataset [226] using the same data collection procedure as in DSTC2 to the total of 1200 dialogues. Later the English data [225] were translated to German and Italian by professionals. The improved dataset was used in paper related to the research into neural belief tracking [180].

### 7.2.5 Natural Language Understanding (NLU) Datasets

The human readable text understating by computer and turning it into computer readable form is entry part of each dialogue system. There are several standard datasets which relate to two main activities performed under Natural Language Understanding (NLU) (§6.8), i.e. intent detection (§6.8.2) and slot filling (§6.8.4).

**Air Travel Information System (ATIS)** [227]. It is a dataset from Microsoft Cognitive Toolkit. The slots are labeled in the IOB (in-out-begin) [161] (§6.8.2) format and the dataset contains air travel related commands.

**SNIPS** [10] It is a dataset built by Snips.ai which serves primarily for NLU benchmarking. It contains several categories (playing the songs, booking the restaurants, and so on) of day to day user commands categories.

[9] http://opus.nlpl.eu

[10] https://github.com/snipsco/nlu-benchmark

### 7.2.6 Natural Language Generation (NLG) Datasets

The human-understandable form of a dialogue generated by a computer, the response, depends on the correct translation from the computer-understandable form in Natural Language Generation (§6.10). The Dialogue State Tracker provides the input data to those components or an independently stored dataset is needed. Nevertheless, the NLG part could or could not have an awareness of the context and thus be or not to be dependent on the previous utterance. There are not many datasets purely dedicated to NLG part of the dialogue system.

**Alex[11] Context NLG Dataset[12]** [228] The dataset was collected using the crowd-sourcing approach. They used a CrowdFlower platform[13] to crowdsource English call recordings, transcriptions, and create response paraphrases. The data collection took several stages to obtain natural user utterances and corresponding relevant, natural, and contextually bound system responses. The dataset covers the domain of public transport information and contains 1859 items.

## 7.3 Influence Corpora for Emotion Regulation

In the State-of-the-art (§2) chapter, the section Psychological methods (§2.6) mentions several Emotion Regulation (ER) techniques (§11.3.1). These can be used as intervention techniques when the chatbot is influenced and needs to act to perform intervention.

One of the ER techniques already solved as the NLP problem is paraphrasing (§9.4.4). So, the next datasets are purely related to this technique only. The rest of the ER techniques require complex solutions, or they are not part of NLP research yet or at all.

**ParaPhrase DataBase (PPDB)[14]** [229], [230] It is a paraphrase archive where paraphrase datasets in various languages (21 in October 2019) are collected, maintained, and provided either as single-lingual or multi-lingual. The paraphrase datasets download allows to chose particular paraphrase type (lexical, phrasal, and syntactic) or download them all. Furthermore, the database provides several sizes of the dataset (from small to triple extra-large).

**WikiAnswers[15]** The paraphrase dataset contains a collection of 18 million question-paraphrase pairs scraped from WikiAnswers. The reason for collecting such a huge dataset is Paralex (Paraphrase-Driven Learning for Open Question Answering) [231] system that learns to answer questions using this dataset.

**Paraphrase for Plagiarism (P4P)** It is a manually annotated corpus composed of 847 English source-plagiarism pairs. The dataset was created first for building up the evaluation framework for plagiarism detection [232], then also used for studying paraphrases concepts and typology [233] and then turned back again to its roots towards automatizing the plagiarism detection [234].

[13] http://crowdflower.com

**Microsoft Research Paraphrase Corpus (MRPC)** The MRPC dataset is the corpus of sentence pairs that were automatically extracted from online news sources. Then they were annotated by humans whether the sentences in the pair are semantically equivalent. It contains 3900 English paraphrase pairs [235].

**Relational Paraphrase Acquisition from Wikipedia (WRPA)** The paraphrases corpora are extracted from Wikipedia and consist of several sub-corpora. Several hundreds of paraphrases are related to person date and place of birth and death. Other several hundreds paraphrases are dedicated to person family relations and origin. More than 80 thousand paraphrases express the authorship relation [236], [237].

## 7.4 Collection of Corpora

The reason to collect a text and turn it into proper corpora is unavailability of domain-relevant sources. Despite plenty of resources on the Internet, it can happen that the particular domain data has not been published yet or they are not in the expected quality.

### 7.4.1 Data Generation

One of the option is to generate corpora as the collection task. If we return to Dialog System Technology Challenge (DSTC) (§2.4.3), specifically its second year, a large corpus of dialogues with various telephone-based dialog systems was collected [55] using the Amazon Mechanical Turk (AMT). A little bit different approach was used in [238]. The Rosetta[16] language generation toolkit originally designed for the CMU Communicator [239] was used for NLG.

### 7.4.2 Wizard of Oz (WOz) Data Collection

Another data collection technique is Wizard of Oz (WOz) when one participant (wizard) of dialogue plays the role of the chatbot [226], [240], [241]. The wizard (a participant of a dialogue) must have access to relevant sources (internet, curated databases) to be able to respond to factoid and news related questions. WOz allows to collect dialogues usable for the development of complete dialogue pipelines from NLU, to DM up to NLG.

## 7.5 Built-in Datasets

Dialogue platforms that are implemented for sharing, training, and evaluating dialogue models contain built-in datasets. Such built-in datasets provide the advantage of fast dialogue implementation, because the most time consuming task, the model training, is already fully or partially done.

Some of those platforms were already introduced (§6.12). For instance, Facebooks ParlAI [195] which contains[17] 79 (August 2019) tasks where most of them are built-in

[16] https://www.rosettastone.com

[17] http://www.parl.ai/static/docs/tasks.html

datasets. 21 of them are dedicated to QA, 14 to chit-chat, 10 of them are task-oriented, the others are testing and debugging tasks. All of them are well documented and point out to related research papers or even code.

Another one is Google TensorFlow Dataset[18] containing 106 (October 2019) datasets from which two are audio relevant, 68 are related to image processing, five contain structured data, 14 textual data and 11 are suitable for translation tasks, and last but not least four of them contain video.

## 7.6 Corpora Ethics

The pre-trained models (§8.2.4) if provided fully can lead to malicious applications [19] like fake news generation, which are commonly known under the most broadly descriptive term Deepfake (i.e., deep learning and fake) [242].

As the resolution of this situation OpenAI published a report [243] related to the release of theirs Generative Pre-Training (GPT)-2 language model (§8.2.3). In the report a staged release which conducts the risk and discussed ongoing partnership-based research and recommendations for responsible publication in AI is considered.

## 7.7 Conclusion, One corpora to rule them all!

A necessity to prepare new corpora whenever a specific task is studied is evident from the extensive and still small list of various types of datasets turned into annotated corpora. Furthermore, a reuse of the existing corpora is mandatory whenever the particular task is elaborated by using another method, and performance comparison (evaluation) is done. It means that there are no silver bullet corpora, no single solution, even though standardization like Stanford Question Answering Dataset (SQuAD), The Open Parallel Corpus (OPUS), Persona-Chat, or even Dialog System Technology Challenge (DSTC) are becoming more usual.

In the State-of-the-art (§2) chapter, Alexa Prize Challenge (APC) (§2.4.2) requires to keep the conversation at least for twenty minutes. To achieve that most of the solutions utilize the mixture of various corpora and related NLP techniques. This leads to longer natural conversation with the user.

So, the solution about corpora related to this rigorous thesis can be done in several ways. Either it would be necessary to experiment with multiple corpora and their fusion as they are provided or the advantage of built-in corpora can be leveraged. Such solutions offer a cleaned data source on one hand, but no flexibility of data adaptation on the other hand. The last and most challenging way is to collect the text and build specialized and purely research dedicated corpora.

[18] https://www.tensorflow.org/datasets/catalog/overview

[19] https://openai.com/blog/better-language-models

# Chapter 8

# Dialogue System Models

In the previous section dealing with dialogue systems the elementary introduction (§6) was described. This section follows this introduction (Figure 6.1) and reveals deeply the dialogue systems modeling techniques (Figure 8.1), their advantages and disadvantages. In this case we consider the application of a particular modeling technique to provide the best overview for the next section about dialogue system influencing (§9).

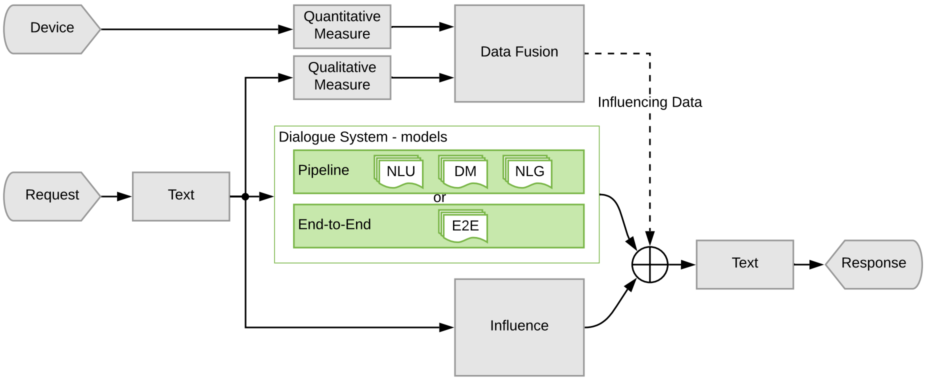


Figure 8.1: Dialogue System - models

This section starts with Artificial Neural Networks (ANNs). They are briefly introduced (§8.1.1) due to their huge popularity in dialogue system research; a short description of the most relevant or interesting types of ANNs is given (§8.1.2).

Before we go deeply into dialogue system models we need to describe the fast evolution in the NLP field related to Natural Language Modeling (NLM) (§8.2). Its history goes into the 1950s, but the modern era has started approximately at the beginning of millennia, and most innovations have happened in the last decade with the highest acceleration during the last three years.

The recent years fast evolution of NLM (§8.2) is also connected with the boom of Pre-trained Language Models (PLMs) (§8.2.4) which were of large sizes at the beginning, but with the application of various compressing approaches (§8.3), especially knowledge distillation (§8.3.1), they have become reasonably small and keep the performance of original model.

This all is important for the latest evolution of dialogue systems built on top of retrieval methods (§8.5), generative methods (§8.6) or pipeline methods (§8.7). For the generative methods we especially consider Deep Learning (DL) (§8.6.1) and Reinforcement Learning (RL) (§8.6.2) dialogue systems or various pipeline architecture (§6.3.1) components.

When bringing dialogue systems closer to the user, various improvements (§8.8) have been made in recent years (for instance personalizing (§8.8.1) in order to give the dialogue systems either a better understanding of user personality or bestow the personality to the dialogue system to react better and less generic way). Next to personalizing ensemble (§8.8.2) dialogue systems stand. They bring together various functions to the user with the a broader spectrum of answers and keep the conversation longer and fruitful, which is one of the ways to beat the Turing test (§10.2.1).

Last but not least, machines suffer from their baby diseases, and dialogue systems are no exception. So, the conversation contains pathologies (§8.9), which need to be taken into account during the dialogue system design phase and avoid them via specific improvements or methods.

## 8.1 Artificial Neural Network (ANN)

ANN were defined in the 1940s and after AI winter [1] they have had the renaissance era due to the computer processing power (GPU and/or Cloud) in the last decade.

In recent years ANNs have become important in various disciplines, NLP is one of them starting with Natural Language Modeling (NLM) (§8.2) using a shallow ANN for modern word embedding (§8.2.2) up to the complex language models for sentence embedding (§8.2.3) built on top of RNN (or its modifications like LSTM or GRU) and Encoder-decoder architectures.

### 8.1.1 Introduction to Artificial Neural Network

The main idea is based on a collection of connected units (nodes) called artificial neurons (similarity to biological neurons, but simplified). Connections are represented by a simplified version of a biological synapse; connection provides the output of one neuron as an input to another neuron.

An artificial neuron (Figure 8.2) have an input ($x_i$) represented by a feature vector, assigned weights ($w_i$) that represent the relative importance of the input, bias ($b$) and output ($y$). It conntains propagation function which computes the input to a neuron as a weighted sum $\sum$ with bias which can be added to the result of the propagation and activation function $f$ which provides a smooth transition of computed sum to the output.

[1] https://en.wikipedia.org/wiki/AI_winter

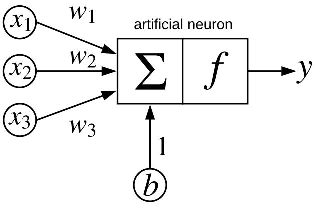


Figure 8.2: Artificial Neuron Schema

ANN can be, but not necessarily (for instance based on Good Old-Fashioned Artificial Intelligence (GOFAI) [244]), a superset of Machine Learning (ML) techniques. Essentially, AI is any machine that shows intelligence in some decision. It has been either fed or trained by a large number of datasets to successfully analyze inputs such as text, images, video, and speech.

Nodes are usually organized into layers and create hidden layers that interconnect input and output layers, providing the required functionality of ANN (Figure 8.3). The chain of transformations from input to output is called Credit Assignment Path (CAP). There is no universal agreement about the threshold of depth that divides shallow learning from deep learning. However, most researchers consider ANN to be deep whenever the CAP depth is higher than 2; otherwise, it is a shallow ANN.

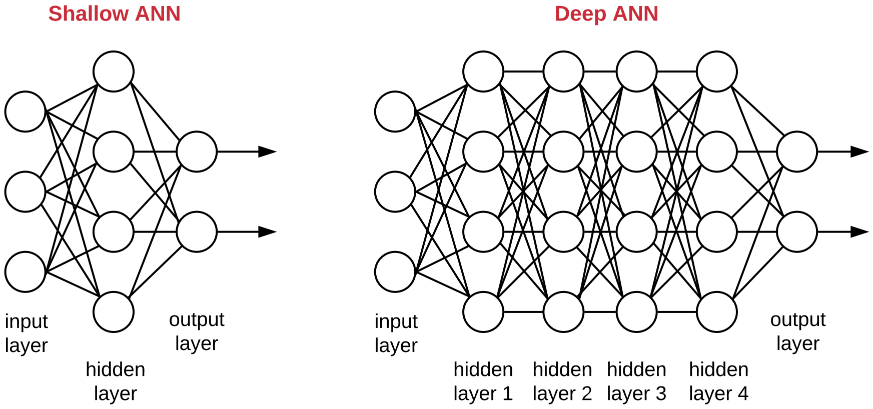


Figure 8.3: Shallow and deep learning ANN

### 8.1.2 Types of Artificial Neural Network

There are plenty of different kinds of Artificial Neural Network (ANN) [245] with architectures suitable for specific applications. Convolutional Neural Networks (CNNs) are usually used for image processing; Recurrent Neural Networks (RNNs) are suitable for chain forms of data (e.g. time-series or text). Here is a brief overview of a few neural network types:

**Feed Forward Neural Network (FFNN)** It is the most common type of the artificial neural network. In this architecture, information moves only in one direction, forward, from the input layer, through the "hidden" layers, to the output layer. There are no loops in the network [246]. The first single-neuron network was proposed in 1958 by AI pioneer Frank Rosenblatt [247]. While the idea is not new, advances in computing power, training algorithms, and available data led to higher levels of performance than previously possible.

**Convolutional Neural Network (CNN)** is an artificial neural network in which connections between neural layers are inspired by the organization of the animal visual cortex, the portion of the brain that processes images; it is well suited for visual perception tasks, but also used for NLU tasks. [248].

**Capsule Neural Network (CapsNN)** It is another type of ANN with added structures called "capsules" [249] to a CNN. It have four major conceptual advantages compared to CNN: viewpoint invariance (recognizes objects regardless of the perspective), fewer parameters (groups neurons in capsules), better generalization to new viewpoints (linearizes complex rotation transformations), defense against white-box Adversarial Learning (AL) attacks (Fast Gradient Sign Method (FGSM) can drop accuracy below 20%, CapsNN maintains it above 70%).

**Generative Adversarial Network (GAN)** These is a pair of ANNs [250], which compete with each other in the game. According to the game theory, it is often but not always in the form of a zero-sum game. One of the networks is called generative and generates candidate data while another network is called discriminative and evaluates the generated data. The GAN trained on photographs can generate new realistically looking photographs.

**Recurrent Neural Network (RNN)** Artificial neural networks whose connections between neurons include loops (Figure 8.4 [251]) are well-suited for processing sequences of inputs. It makes them highly effective in a wide range of applications, like handwriting recognition, texts analysis and speech recognition [252].

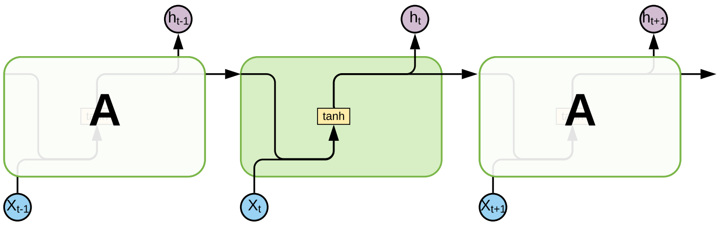


Figure 8.4: The repeating module in a standard RNN contains a single layer

**Recursive Neural Network (RecNN)** It is another kind of deep ANN; it is essentially generalization to structures of a recurrent neuron [253]. It is created by applying the same set of weights recursively over a structured input. RecNN produces a structured prediction over variable-size input structures, or a scalar prediction on it by traversing a given structure in topological order [254].

**Long / Short Term Memory (LSTM)** It is another RNN architecture composed of a cell, input gate, output gate and forget gate (Figure 8.5 [251]). The cell remembers values over arbitrary time intervals, and the three gates regulate the flow of information into and out of the cell [255]. It is suitable for processing sequential data like time-series, text, speech, or video.

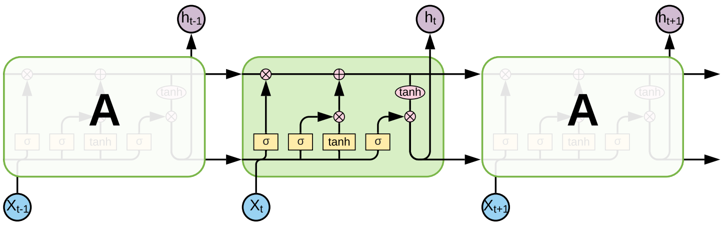


Figure 8.5: The repeating module in an LSTM network contains four interacting layers

**Gated Recurrent Unit (GRU)** GRU [256] It is a faster (but less powerful) variation on the LSTM network; it merges the cell state and hidden state (Figure 8.6 [251]). On the certain smaller datasets the GRU network exhibits even better performance than LSTM [257], but overall LSTM cells consistently outperform GRU [258], [259].

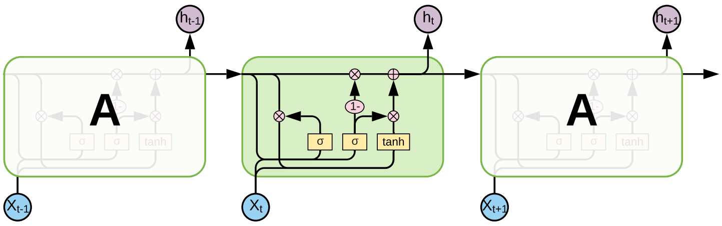


Figure 8.6: The repeating module in an GRU contains three interacting layers

### 8.1.3 Neural Network Frameworks

ANN frameworks are one of the many ways to implement dialogue system modules within the pipeline (§6.3.1) or E2E architecture (§6.3.2). This section introduces several libraries which represent the current state of the art and are widely used not only for dialogue system implementations but also for any deep learning application.

**TensorFlow** [2][260] It is an open source data flow library usually used for machine learning applications such as neural networks developed by Google.

**Keras** [3] It represents an open source interface running on top of the Tensorflow or Microsoft Cognitive Toolkit (CNTK). It is designed to enable fast experimentation with deep neural networks.

[2] https://www.tensorflow.org

[3] https://keras.io

**PyTorch** [4][261] It is another open source machine learning library developed by the Facebook's artificial-intelligence research group mostly used for neural network applications.

## 8.2 Natural Language Modeling (NLM)

A statistical language model is a probability distribution over sequences of words. The language model is usually represented numerically. It can be, for instance, done by word frequency appearance in a sentence or by vector space representation of words in sentences.

In the last few years, the modern word embedding (§8.2.2) and sentence embedding (§8.2.3) stand behind the acceleration of NLP and speed up research in various subordinated fields like Natural Language Understanding, Neural Machine Translation, Question-Answering, Natural Language Generation, and others. It has also accelerated the improvement of various retrieval-based or generative dialogue system models.

### 8.2.1 Early Word Embedding

The history of word embedding goes back in 1950s with early reference to **Bag-of-Words (BoW)** in a linguistic context defined by Harris [262]. Then in 1960s the paper from Salton [263] using the **Term-document matrix** formalization was released.

A few years later, during the 1970s, the research about the term weighting was moved forward by Jones [264] when she conceived a statistical interpretation of term specificity called **Inverse document frequency (Idf)**. Based on this Salton [265] proposed later a vector space model known as **Term frequency - Inverse document frequency (Tf-Idf)** in mid 1970s.

Next to the word embedding other NLP approaches also evolved. At the end of the 1980s **Latent Semantic Analysis (LSA)** was patented and published by Deerwester [266] in 1990.

During the 1990s classic statistical NLP approaches based on **n-grams** employing smoothing to deal with unseen n-grams [267] were evolved. And in late 1990s Baker [268] and others introduced the FrameNet[5] project which part is a task of **Semantic Role Labelling** also called shallow semantic parsing or slot-filling (§6.8.4) that is still actively researched today.

At the beginning of millennia, Lafferty introduced the **Conditional Random Fields (CRF)** [176] method, one of the most influential classes of sequence labeling. Furthermore, two years later, in 2003 one of the most widely used techniques in machine learning (which is still the standard way to do topic modeling) the **Latent Dirichlet Allocation (LDA)** [269] was introduced.

The popularity of ANN continued and in 2003 the first **Feed Forward Neural Network (FFNN)** language model [270] was also presented. The FFNN was fed by vector representations of the n previous words (embeddings). Lately in 2010, respectively 2013,

[4] https://pytorch.org

[5] http://framenet.icsi.berkeley.edu

FFNNs have been replaced with Recurrent Neural Network (RNN) [271] and Long / Short Term Memory (LSTM) [272] for language modelling.

### 8.2.2 Modern Word Embedding

The era of modern word embedding started in 2013. This computational technique helped to establish a new focus on AI after AI Winter, which took time until late 2000s. The timeline (Figure 8.7) of keyword embedding related papers released since 2013 is accelerating.

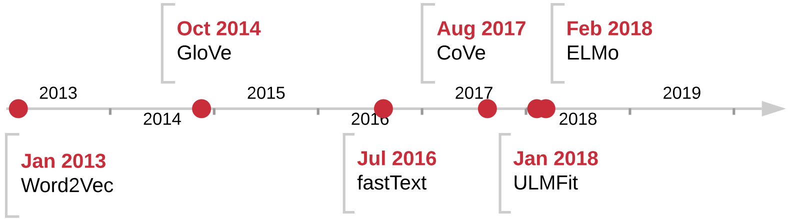


Figure 8.7: Modern Word Embedding Timeline

- In January 2013 the Google research team (Mikolov et al.) [273] presented a **Word to Vector (Word2Vec)** group of word embedding models with two solutions: Continuous Skip-gram and Continuous Bag-of-Words (CBOW). The models are shallow, two-layer ANNs are trained to reconstruct linguistic contexts of words. CBOW is faster than the skip-gram, but the skip-gram does a better job for uncommon words.

- More than one year later, in October 2014 the Stanford research team came with the **Global Vectors (GloVe)** [274] model for distributed word representation. Vector representations for words are obtained through unsupervised learning. It maps the words into a meaningful space where the distance between words is related to their semantic similarity.

- As an extension of the Continuous Skip-gram model [273], [275] the improvement implemented by the Facebook research team was introduced in July 2016. It is called **fastText** [142], [276] and takes into account sub-word information.

- Another contribution to the representation of words with vectors is the **Contextual Vectors (CoVe)** [277] published in August 2017 by the Salesforce team. It utilizes Transfer Learning (TL), and it is inspired by Machine Translation (MT) tasks. The first part is the bidirectional Long / Short Term Memory (LSTM) trained on various datasets to create MT-LSTMs models. The second part is to append the outputs of the MT-LSTMs CoVe to the word vectors typically used as inputs to these models.

- At the beginning of 2018 (January), Howard and Ruder proposed **Universal Language Model Fine-tuning (ULMFiT)** [278], an effective Transfer Learning (TL) method that can be applied to any task in NLP.

- In February 2018 Allen Institute for AI[6] came with **Embeddings from Language Models (ELMo)** [279]. It models not only complex characteristics of word use (e.g., syntax and semantics) but also how these uses vary across linguistic contexts. Word vectors are based on a deep bidirectional language model (biLM).

### 8.2.3 Sentence Embedding

In NLP the context given by order and relation of words in a sentence plays an important role in proper language understanding.

The latest modern word (it does not need to be only the word, it could be a letter, syllabus, so more precisely we can speak about token) embedding (§8.7) takes into account complex characteristics of word use and linguistic context (ELMo). The ambiguity in the language can be eliminated or at least minimized by sentence (the sentence is not entirely correct, we can include a fragment of a sentence, paragraph, so it is better to define it as a sequence) embedding.

The sentence embedding timeline (Figure 8.8) has also a sign of unrestrained development. Considering transformer evolution (Figure 8.10) described later it is really impressive how much NLP has evolved in last few years.

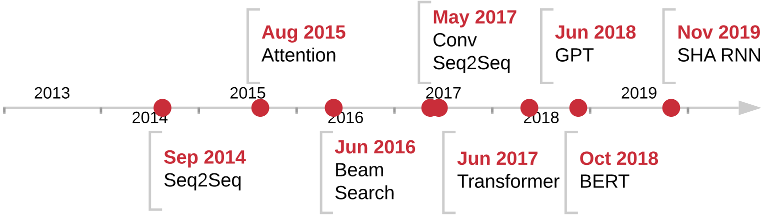


Figure 8.8: Modern Sentence Embedding Timeline

- **Sequence to Sequence (Seq2Seq)** It is a sequence to sequence mapping language model [280]. The best suitable solution is to employ Recurrent Neural Network (RNN), but usually it is implemented via a more advanced version of Long / Short Term Memory (LSTM) or Gated Recurrent Unit (GRU) (vanilla RNN is not suitable due to its vanishing gradient problem).

  The Seq2Seq model usually consists of two components (Figure 8.9):

  - Encoder utilizes a deep ANN and encodes the input words to the hidden vectors. The vectors are created from the current word and the context based on the word in the sentence.
  - Decoder is also based on a deep ANN, but it works oppositely to the encoder. It takes the hidden vector previously generated by the encoder, its hidden states, and current word and produces the next hidden vector from which the next word is finally predicted.

[6] https://allenai.org

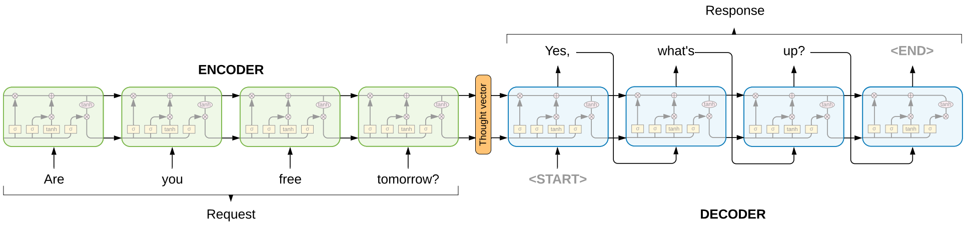


Figure 8.9: Sequence to Sequence Model

- **Attention** The pure encoder-decoder network represents the vanilla Seq2Seq implementation which has its limitations.

  One of the limitations is that the complete information in the input sentence should be encoded into a fixed-length vector — context. The decoder takes as the input a single vector. It stores all the information about the context. It is not an issue for short sequences, but the problem starts with long sequences.

  Attention mechanism [281] allows the decoder to selectively look at the input sequence hidden states, which are then provided (as a weighted average) as an additional input to the decoder.

  The attention can have different forms [282] and is widely applicable for tasks constituency parsing [76], reading comprehension [283], one-shot learning [284], image captioning [285] and many others.

- **Beam Search**. The decoder selects as the output the word with the highest probability. However, it does not mean that the highest probability always leads to the best result due to the basic problem of greedy algorithms. The beam search [286] is applied to suggest the possible translation at each step, making a tree of top k results.

- **Convolutional Seq2Seq**. The original Seq2Seq model is purely based on Recurrent Neural Network (RNN). On the other hand, Convolutional Neural Network (CNN), compared to RNN models, brings an advantage that computations over all elements can be fully parallelized during training [287]. It leads to a better chance to use the GPU hardware and makes optimization easier since the number of non-linearities is fixed and independent of the input length.

- **Single Headed Attention (SHA) RNN**. It is a progressive language model [288]. It is not completely following the current hype around the Transformer, but it is built on top of RNN. Additionally to RNN, it is composed of pointer-based attention, and ”Boom” large feed-forward layer (also found in Transformers and other architectures) with a sprinkling of layer normalization.

The **Transformer**, **BERT** and **GPT** models are described in the next section.

### Transformer Evolution

The year 2019 was the year of the Bidirectional Encoder Representations from Transformers (BERT), which initially evolved from Transformer [289]. For instance, 169 BERT-related papers [290] have been published. With the rise of BERT and the Transfer Learning (TL) (§8.6.3) trend in NLP has been lifted up by vast use of Pre-trained Language Models (PLMs) released with fine-tuning [291] for specific NLP-related tasks.

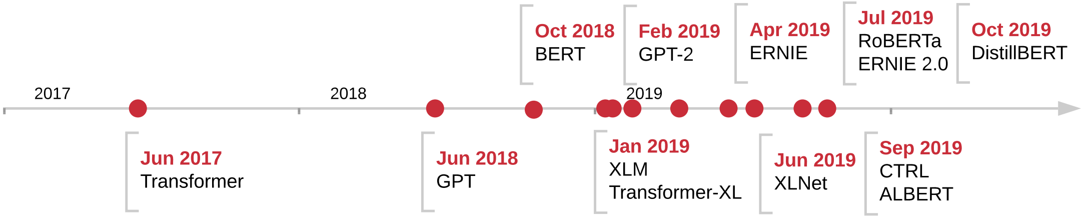


Figure 8.10: Transformer Evolution

- **Transformer** [292] It represents a simplification of sequence transduction models usually based on complex RNN or CNN that include an encoder and decoder. The simplified architecture of the Transformer is solely based on attention mechanism dispensing with recurrence and convolutions entirely. The advantage of this approach is the better quality of the model, which can be more parallelized and requires significantly less time to train.

- **Generative Pre-Training (GPT)** Another Transformer successor is a large generative language model implemented by OpenAI company[7] called GPT [293]. They demonstrated that the large gains could be made with generative pretraining of a language model on a diverse corpus of unlabeled text, followed by discriminative fine-tuning on each specific NLP tasks from SQuAD, RACE or GLUE benchmarks (§10.3.3).

- **Bidirectional Encoder Representations from Transformers (BERT)** It uses the now ubiquitous Transformer architecture. BERT [294] is designed to pretrain deep bidirectional representations from the unlabeled text by jointly conditioning on both left and right context in all layers. BERT is trained on a combination of Book Corpus [295] plus English Wikipedia corpus. The pre-trained model can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of NLP tasks from SQuAD, RACE, or GLUE benchmarks (§10.3.3).

- **Cross-lingual Language Model (XLM)** XLM [296] offers two cross-lingual language models. The first one is unsupervised and relies only on monolingual data. The second one is supervised and leverages parallel data with a new cross-lingual language model objective. It significantly outperforms the previous state of the art on cross-lingual classification, unsupervised machine translation, and supervised machine translation.

[7] https://openai.com

- **Transformer-XL** [297] It is a novel language model based on the original Transformer, which enables learning dependency beyond a fixed-length without disrupting temporal coherence. It consists of a segment-level recurrence mechanism and a novel positional encoding scheme. It brings the functionality to learn dependency, which is nearly two times longer than RNN and nearly six times longer than vanilla Transformers. It outperforms vanilla Transformer in both short and long sentences, and it is approximately 1800 times faster in evaluation.

- **Generative Pre-Training (GPT)-2**. GPT-2 [298] It is a direct scale-up of GPT. The Transformer based architecture was used, and the model largely follows GPT model with a few modifications. The model was trained on a corpus called WebText, contains slightly over 8 million documents for a total of 40 GB of text from URLs shared in Reddit.

  GPT-2 generates exceptionally fluent English, which led to the ethical conundrum if OpenAI should or should not publish the complete Pre-trained Language Model (PLM). So, in February 2019, they instead released a small 124M parameter model [299] for researchers to experiment with, as well as a technical paper. Three months later, in May 2019, they staged the release of a medium 355M model. Furthermore, in August 2019, they decided to release a 774 million parameter model [300] with publishing a paper related to the social impacts of such release [243]. Later in November 2019, OpenAI decided to publish a full model with 1.5 billion parameters [301].

- **Enhanced Representation through kNowledge IntEgration (ERNIE)** It is another language model inspired by BERT, especially its masking strategy. ERNIE [302] is designed to learn language representation enhanced by knowledge masking strategies, which includes entity-level (composed of multiple words) masking and phrase-level (composed of several words standing together as a conceptual unit) masking.

- **XLNet** It represents another improvement of the elementary BERT language model. The XLNet [303] model utilizes a generalized autoregressive pretraining method that enables learning bidirectional contexts by maximizing the expected likelihood over all permutations of the factorization order and overcomes the limitations of BERT thanks to its autoregressive formulation. Moreover, it integrates the concepts from Transformer-XL to the state-of-the-art autoregressive model into pretraining.

- **Robustly optimized BERT approach (RoBERTa)** The original BERT work suffers from significant undertraining. So, the Facebook research team presents the replication study [304] of BERT pretraining that carefully measures the impact of many key hyperparameters and training data size. This improved model achieved state-of-the-art results on GLUE, RACE and SQuAD benchmark datasets (§10.3.3).

- **Enhanced Representation through kNowledge IntEgration (ERNIE) 2.0** is the improved version of original ERNIE. It is built on top of the idea that pre-training tasks can be incrementally constructed [305]. The models are pre-trained trough continual multi-task learning, and the pre-trained model is fine-tuned to adapt to various language understanding tasks.

- **Conditional TRansformer Language (CTRL)** It is another Transformer language model released completely with 1.63 billion parameters [306]. It was trained by condition control codes that govern style, content, and task-specific behavior. Control codes are additional metadata derived from the structure that naturally cooccurs with a raw text. It still allows us to preserve the advantage of unsupervised learning and, on top of that, provide more precise control of text generation.

- **A lite BERT (ALBERT)** It is an NLP model based on BERT. ALBERT [307] uses two parameter-reduction techniques to lower memory consumption and increase the training speed of BERT. The first technique is a factorized embedding parameterization. The second is cross-layer parameter sharing. Both lead to significantly better results in the standard NLP tasks from SQuAD, RACE, or GLUE benchmarks (§10.3.3) with fewer parameters than BERT.

- **DistilBERT** It is a smaller — compressed by knowledge distillation (§8.3.1) — general-purpose language representation of original BERT language model [308]. It can then be fine-tuned to achieve good performances on a wide range of standard NLP tasks from SQuAD, RACE or GLUE benchmarks (§10.3.3) like its larger counterpart.

### 8.2.4 Pre-trained Language Models (PLMs)

A Pre-trained Language Model (PLM) represents the modern way to save the costs on the training phase of the language model. The PLMs assumes usage of Transfer Learning [309] (§8.6.3) in which a deep neural network language model is pretrained on a web-scale unlabelled text dataset with a general-purpose training objective before being fine-tuned on various downstream tasks [310].

At the moment (January 2020) it is possible to find various PLMs, the most known and used ones are:

**HuggingFace's Transformers** [8][310] is a state-of-the-art Python library for Tensorflow and Pytorch, which provides general-purpose architectures (GPT, BERT, XLM, Transformer-XL, GPT-2, XLNet, RoBERTa, CTRL, ALBERT, DistilBERT, CamemBERT [311], Text-to-Text Transfer Transformer (T5) [312], XLM-RoBERTa [313], MultiModal BiTransformer (MMBT) [314] and others from external contributors) for Natural Language Understanding (NLU) and Natural Language Generation (NLG) with over 32 Pre-trained Language Models in more than 100 languages.

**Google's Tensorflow Hub** [9] is a Python library for Transfer Learning (§8.6.3) by reusing parts of TensorFlow models. It contains over 400 models (January 2020) related to three domains: text (over 100), images (over 300), and video (exactly 4). The text

[8] https://github.com/huggingface/transformers

[9] https://www.tensorflow.org/hub

models support text embedding; the image models can do pose detection, classification and segmentation, image generation, and the video models serve for classification and generation. The models are either prepared by Google or external contributors.

**Facebook's PyTorch Hub** [10] is another Python repository for Transfer Learning (§8.6.3) with pre-trained PyTorch models. It is the smallest repository with model definitions or pre-trained weights. So far (January 2020) it contains about 30 models from the following domains: audio (3), image generative (2), NLP (10) including 8 HuggingFace's models, various scriptable models (12) and others categorized as vision models.

## 8.3 Compressing Language Models

Using the modern language models (§8.2.3) like Transformer or Bidirectional Encoder Representations from Transformers (BERT) requires a huge model with hundreds of millions parameters (Figure 8.11). Then not only the training phase to establish such a model and its fine tuning needs GPU, but also during inference calculation, a CPU (or even multiple) could not be enough.

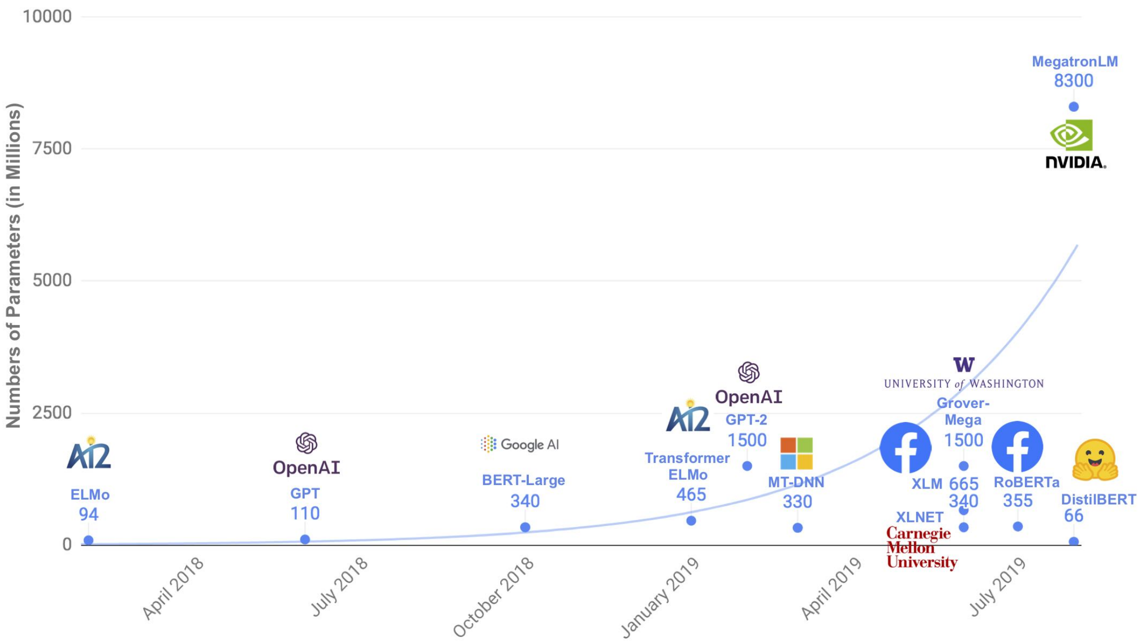


Figure 8.11: Parameter counts of several recently released pre-trained language models

The model size can be compressed during the training or after it. Compressing a model means to reduce the number of parameters (weights) or their precision.

Three main approaches used to language models compression are:

[10] https://pytorch.org/hub

**Quantization** The quantization (pseudo- or real- quantization) of ANN [315] is a compression technique which could be achieved by decreasing the numerical precision of a model's weights.

**Pruning** The pruning [316], [317] technique also belongs to compression; it removes parts of a model (weight pruning or neuron pruning) to make it smaller and faster.

**Knowledge Distillation** The knowledge distillation [318] is not a model compression approach; it corresponds more to Transfer Learning (TL) (§8.6.3). The pre-trained big and slow model is used to train a smaller model to mimic original model behavior.

The **Knowledge Distillation** technique is the most popular and used for the NLP purposes; it is described deeper in the next section (§8.3.1).

### 8.3.1 Distilling Knowledge

Distilling knowledge means to train a big and slow model and use it to train a smaller one [318]. It is about to use big model raw predictions (soft targets), before the final activation function (hard targets) is applied, to train a small model (Table 8.1). For example, the last activation function reduces information of outcome classification to one of many classes, and remaining turns to zero. The raw predictions of ANN internal representations thus also contain not-predicted classes [319] and bring more information for future small model training.

| cow | dog | cat | car | |
|---|---|---|---|---|
| 0 | 1 | 0 | 0 | hard targets |
| $10^{-6}$ | 0.9 | 0.1 | $10^{-9}$ | soft targets |

Table 8.1: Examples of hard and soft targets [320]

The teacher (big and slow model) and student (smaller model) approach are shown in Figure 8.12 [321]. The loss calculated between the output of student model and hard targets is to make the student model to perform much better than teacher model in practice.

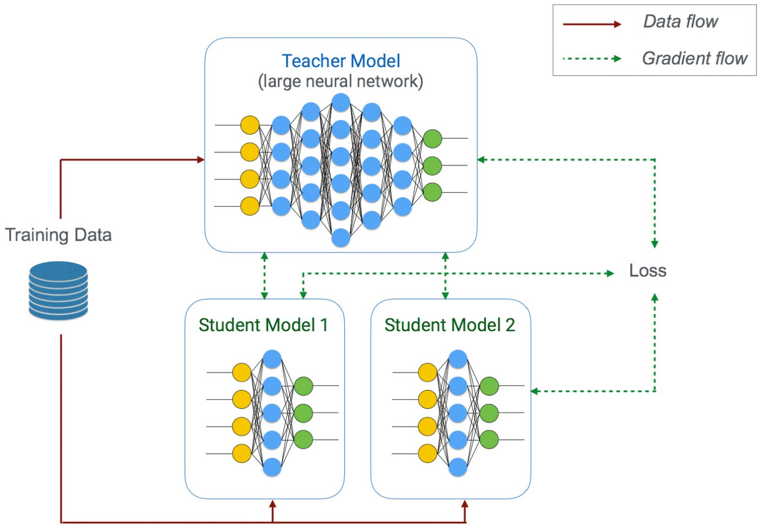


Figure 8.12: Joint training and distillation approach to learn compact student models

The only problem with the teacher-student approach is that it works up to some particular size difference between the pre-trained teacher model applied to the student. To solve the problem, there is research that proposes to build an intermediate pre-trained model called teacher-assistant. Moreover, the research even proposes multi-step distillation [322].

The performance of the BERT distilled version (Table 8.2) compared to the original BERT model is degraded by 3%, but training time takes 25% of the original, and the number of parameters dropped to 60% (66 millions) of original Base (110 millions) and 20% of original Large (340 millions).

| | | Model Size[11] | Training Time | Performance |
|---|---|---|---|---|
| BERT [294] | Base | 110 | 100% | 100% |
| | Large | 340 | | |
| XLNet [303] | Base | 110 | 500% | 102-115% |
| | Large | 340 | | |
| RoBERTa [304] | Base | 110 | 400-500% | 102-120% |
| | Large | 340 | | |
| DistilBERT [308] | Base | 66 | 25% | 97% |

Table 8.2: BERT vs. other models comparison [323]

[11] in Millions

## 8.4 Dialogue Systems Classification

Due to the fast evolution (§2.2) of dialogue systems, various architectures (§6.3) and the system complexity (§6.7) many approaches to classify dialogue models exist. The dialogue system state-of-the-art section (§2.3) presents a comprehensive overview built on top of the several research papers which stand behind the inspiration (Figure 8.13) of how to organize this chapter and classify the dialogue systems from the author´s perspective.

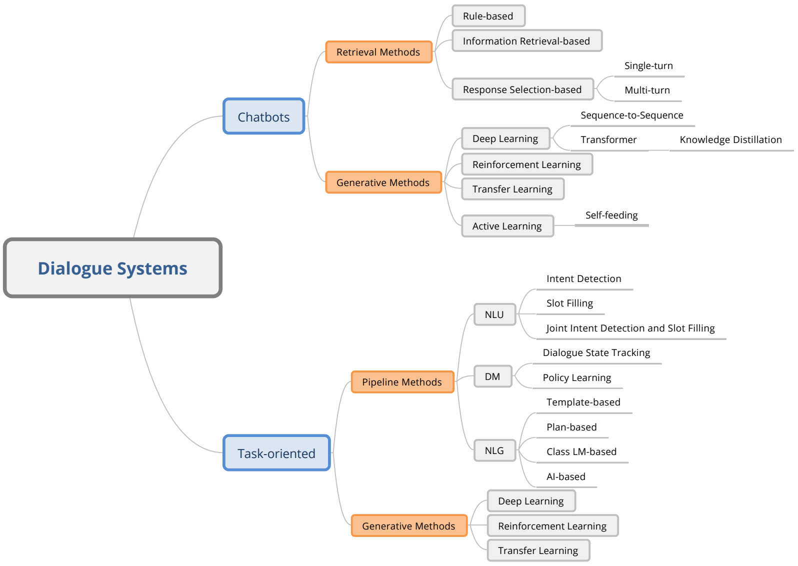


Figure 8.13: Dialogue systems classification

Needless to say, like any of those attempts to give a complex field structure, it is not perfect, and the author is aware that his approach mixes several point-of-views, more concrete dialogue systems architecture (§6.3) together with taxonomy (§6.4).

## 8.5 Retrieval Methods

The retrieval methods for dialogue systems are based on the retrieval-based models (§6.4.1) and are either hand-crafted or request-response pair methods.

The dialogue system returns the response based on the request-response similarity or other matching criteria. The primary methods which provide the grounding for the retrieval-based methods are Question-Answering (QA), response selection, response generation, response matching, and others.

The main directions which the research currently focus on are rule-based dialogue systems (§8.5.1), Information Retrieval (IR) dialogue systems (§8.5.2), and Response Selection (RS) dialogue systems (§8.5.3).

### 8.5.1 Rule-based Dialogue Systems

Early dialogue models as well as elementary contemporary models were rule-based (one of the retrieval-based models §6.4.1 approaches). This approach requires no data, i.e., the systems usually use pattern matching or database instead of dataset based training. On the other hand, a lot of manual effort, which costs many resources, needs to be invested in building the Question-Answering (QA) model. Moreover, the topic coverage of such a system is not fully satisfactory.

The early era of dialogue system evolution (§2.5.2) describes dialogue systems which models are rule-based.

One of the most known patterns matching dialogue system implementation is **ELIZA** [7] in the late 1960s. It uses the pattern/transformation rules with the keyword ranking approach in the human conversation. ELIZA was followed by PARRY [8] created in the late 70s.

It was implemented by **Jaberwacky** in 1988. The winner of the Loebener Prize (§2.4.1)) is based on contextual pattern matching, i.e., a rule based dialogue system.

ELIZA inspired the implementation of **ALICE** [9] created in 1995, which applies the heuristic pattern matching rules to the conversation with a human. The dialogue system uses an XML Schema called Artificial Intelligence Markup Language (AIML) for specifying the heuristic conversation rules.

It led to the implementation of **Mitsuku** in 2005, the state-of-the art AIML dialogue system which won the Loebner Prize (§2.4.1) five-times.

The rule-based models are still used, but over time they are replaced by more sophisticated approaches based on Artificial Neural Network (ANN) or Machine Learning (ML).

The following list presents the examples of dialogue system engines that use configuration languages to simplify conversation definition. All of them (no matter in which programming language they are implemented) are E2E engines based on the pattern matching approach with various complexity behind.

**Artificial Intelligence Markup Language (AIML)** AIML[12] is the XML schema for dialogue modeling. It provides flexibility to establish a complex and powerful dialogue model through pattern-matching definitions (see example Code 8.1) and additional approaches like variables or memory fields to keep the conversation with sort of context.

```
<?xml version="1.0" encoding="UTF-8"?>
<aiml version="1.0">
    <category>
        <pattern>MY NAME IS _</pattern>
        <template>Nice to meet you, <star/>!</template>
    </category>
</aiml>
```

Code 8.1: An elementary AIML example, which reads the name from the pattern and uses it in conversation.

[12] http://www.aiml.foundation

**Open Intent Markup Language (OIML)** can be taken as an alternative to AIML. It was implemented independently as an open-source framework to create dialogue systems in short time. The configuration of the dialogue system based on OIML[13] is done by three files that are required to describe the bot: dictionary JSON file, model OIML file and user actions file (implemented for instance in JavaScript).

**RiveScript** The scripting text-based (see example Code 8.2) language RiveScript[15] is an alternative to the previous AIML and OIML languages. It is designed to help with the development of interactive dialogue systems.

```
+ my name is *
- Nice to meet you, <star1>!
```

Code 8.2: A simple RiveScript example which get the name from the trigger (+) and uses it in response (−).

### 8.5.2 Information Retrieval (IR) Dialogue Systems

The dialogue models based on the Information Retrieval (IR) work on the principle to respond to users' request by some appropriate response built from the corpus of natural text. The text of human conversation can be collected from social networks, discussion forums or blogging platforms. Data can come from various existing corpora, for instance MovieQA [214] (§7.2.1) or OpenSubtitles [223], [224] (§7.2.3).

The Information Retrieval (IR)-based system can use any retrieval algorithm to choose a relevant response based on the given corpus and user input. According to Jurafsky [324], the two most straightforward methods of how to get a turn response are the following ones:

**Return the response to the most similar turn** The idea is that we should look for a turn that most resembles the user's turn, and return the human response to that turn [325], [326].

**Return the most similar turn** The idea here is to directly match the users' query with turns from the conversational corpus since a good response will often share words or semantics with the prior turn [327].

In both cases mentioned above the similarity function between the users request and returning response is usually cosine similarity computed either over words (Tf-Idf (§8.2.1) or Word2Vec, fastText, GloVe, ELMo (§8.2.2) or others) or sentence embeddings (BERT and others (§8.2.3)).

### 8.5.3 Response Selection (RS) Dialogue Systems

The Response Selection (RS) dialogue systems are the evolution of the IR dialogue systems.

[13][14]

[15] https://www.rivescript.com

We recognize two main groups of RS dialogue systems: single-turn (Figure 8.14) and multi-turn (the **message** in Figure 8.14 is replaced by **context**, i.e. message + history). The single-turn RS are also known as Question-Answering (QA), answer selection, or matching short text systems. The multi-turn systems can be also called multi-view RS.

Two main datasets are typically used for the RS-based dialogue systems: The Ubuntu Dialogue Corpus [221] and Douban Conversation Corpus [222] (§7.2.3).

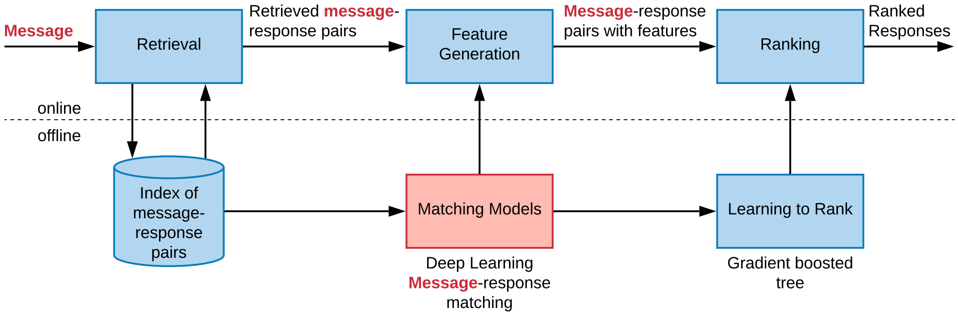

Message
Retrieval
Retrieved message-response pairs
Feature Generation
Message-response pairs with features
Ranking
Ranked Responses
online
offline
Index of message-response pairs
Matching Models
Deep Learning Message-response matching
Learning to Rank
Gradient boosted tree

Figure 8.14: Single-turn Response Selection Dialogue System

Wu [16] then, under the following two main groups, identified several framework approaches and methods used within each group as matching models.

**Single-Turn** provides an immediate response to the input message without keeping the context and can be solved as following frameworks:

- With message-response sentence embedding [328]–[332]. The message and response are turned into vector representations by a sentence embedding layer (for instance by using: CNN, BiLSTM with attention, or GRU with attention) and the similarity of both vectors is calculated by matching layer (Multilayer Perceptron (MLP), Eclidean distance, Cosine distance).
- With message-response word interaction [328], [333]–[336]. The message and response words are represented as vectors turned into interaction matrices by an interaction layer (for example cosine/dot product, linear or non-linear transformation), transformed into an interaction vector by a transformation layer (CNN, RNN), and the matching score is calculated by a matching layer (MLP, softmax).

**Multi-Turn** holds the context to provide a multi-turn response to the single input message with the previous history and can be solved with following frameworks:

- With context-response sentence embedding [221], [337]–[339]. The message and context are turned into vectors representation by a **context embedding layer** (LSTM, GRU) and **sentence embedding layer** (word embedding (§8.2.2), BiLSTM, CNN, GRU), and the matching score is calculated by a **matching layer** (Bilinear, MLP).
- With context-response sequential matching, specifically Sequential Matching Network (SMN) [222] and Sequential Attention Network (SAN) [340], [341].

The message and context are turned into vectors representation and the utterance-response matching is provided by the **matching layer** (CNN, attention (§8.2.3)) followed by a **matching accumulation layer** (GRU) and end up with a **prediction layer** which provides the matching score.

## 8.6 Generative Methods

The generative methods for dialogue systems are based on the generative models (§6.4.2). The generative methods are purely data-driven, respectively corpora based (§7.2) models.

The dialogue systems are usually following the existing solutions from other dialogue tasks. The problem of dialogue research is not standing alone in the vacuum. So, we can see the influence of fields like Statistical Machine Translation (SMT), Neural Machine Translation (NMT), and others.

The generative methods can be divided into two main groups, open domain (§6.5.1) chit-chat dialogue systems, and closed domain (§6.5.2) task-oriented dialogue system.

The main directions which the research currently focus on are Deep Learning (DL) dialogue systems (§8.6.1), Reinforcement Learning (RL) dialogue systems (§8.6.2), Transfer Learning (TL) dialogue systems (§8.6.3), Active Learning (AL) dialogue systems (§8.6.4), Adversarial Learning (AL) dialogue systems (§8.6.5), and hybrid dialogue systems (§8.6.6).

### 8.6.1 Deep Learning (DL) Dialogue Systems

There are several reasons to use DL in combination with corpora for a dialogue system. The advantage when compared to retrieval-based dialogue systems is direct data-driven development.

**Chit-chat** The first non-goal-driven systems have taken the inspiration from the use of ANN (§8.1) in Natural Language Modeling (NLM) (§8.2), Statistical Machine Translation (SMT) and Neural Machine Translation (NMT) tasks.

The DL E2E dialogue systems based on ANN have shown promising results on various dialogue tasks. Sutskever presented his Sequence to Sequence [280] language model (§8.8) later improved by attention mechanism [281]. The RNN Seq2Seq encoder-decoder approach approach was used in phrase representation [256] and neural conversational model [76]. It was followed by attention ANN conversational model [342].

Use of RNN for generating responses was proposed by Shang [343] followed by Sordoni [344] who extended the framework and made from request-response pairs the context sensitive responses generation (triples of three consecutive utterances). On top of that Sordoni [345] built a novel Hierarchical Recurrent Encoder-Decoder (HRED) architecture that allows the model to be sensitive to the order of requests in the context. This was extended further by Serban [346] for web request suggestion given by the request already submitted by the user.

During several years research is shattered amongst various dialogue related methods like Reinforcement Learning (RL) (§8.6.2), Transfer Learning (TL) (§8.6.3), Active Learning (AL) (§8.6.4) and Adversarial Learning (AL) (§8.6.5). Starting with the publication of Bidirectional Encoder Representations from Transformers (BERT) [294] in November 2018 the focus has been also going towards Pre-trained Language Models (PLMs) (§8.2.4).

The recent advances of large-scale PLMs transformer-based architectures are also useful for dialogue systems. DialoGPT [347] is a dialogue generative pre-trained transformer (GPT-2 model) on 147M conversation-like exchanges extracted from Reddit. It extends the Hugging Face PyTorch transformer [310] to attain a performance close to human both in terms of automatic and human evaluation in single-turn dialogue settings.

The latest contribution to the E2E dialogue systems is the multi-turn open-domain chatbot Meena [12]. Meena scores high on Sensibleness and Specificity Average (SSA) (§10.3.2), 72% base, 79% full version. It suggests that a SSA human-level of 86% is potentially achievable having better optimized perplexity (§10.3.2).

**Task-oriented** One of the first systematic approach to standardize baseline to deal with the task-oriented dialogue systems can be found in the Dialog System Technology Challenge (DSTC) (§2.4.3) established in 2013. The first three years were focused on developing a single component for Dialogue State Tracker (DST) on task-oriented human-machine conversations: evaluation metrics [54], user goal changes [55] and domain adaption [56]. The following two years, DSTC4 [57] and DSTC5 [58] introduced human-human conversations and cross-lingual adaption to offer multiple tasks not only for DST but also for other components in dialogue systems.

From the sixth DSTC [59] multiple main tracks were organized in parallel to address a wider variety of dialogue related problems like End-to-End (E2E) task-oriented dialogue [60], conversational modeling [61], and detection of dialogue break down [62]. The last two years of Dialog System Technology Challenge (DSTC) returned to the development of E2E dialogue systems, the 7th DSTC [63] with the following three tracks: noetic response selection [64], grounded response generation [65], and audiovisual scene aware dialog [66]. Moreover, the most recent DSTC8 [67] focused on topics like E2E multi-domain dialogue system, fast adaption, predicting responses, and again audiovisual scene aware dialog.

Next to DSTC, there are plenty of other attempts to deal with task-oriented dialogue systems.

Wen [226] comes with E2E trainable task-oriented dialogue system along with a new way of collecting dialogue data based on a novel pipe-lined Wizard of Oz (WOz) (§7.4.2) framework and utilization of Artificial Neural Network (ANN).

Bordes [348] follows with a task-oriented dialogue system built on top of the memory networks [349]. To confirm the results, he compared his system to a hand-crafted slot-filling baseline on data from the second DSTC.

The task-completion [16] neural dialogue system [74], [350] is another way to solve an E2E task-oriented system to overcome issue that system cannot adapt easily to multiple topics, because each module is trained individually. It utilizes the pipeline based architecture (§6.3.1) implemented by ANN in combination with a backend database. It evolves an idea of knowledge based bot [351][17].

One solution which fulfills the requirements of DSTC 6 challenge is an E2E task-oriented dialogue system [352] which employs the memory network MemN2N [353] architecture.

As the evolution of the previous memory network usage E2E task-oriented dialog systems [354] with MEM2Seq[18] are introduced, which are memory to sequence composed models of two components: the MemN2N encoder, and the memory decoder.

One of the latest contributions [355] to the memory network E2E task-oriented dialogue models is the application of bidirectional LSTM. It is located at the beginning of the model to better reflect temporal information and achieve state-of-the-art performance among the memory networks. It is comparable to Hybrid Code Network (HCN) (§8.6.6) and hierarchical LSTM models.

Another direction in task-oriented dialogue systems is represented by existence of humans in the loop for dialogue learning. The human user assists in completing tasks by conducting multi-turn conversations. The solutions usually use hierarchical Long / Short Term Memory (LSTM) to model a dialogue with multiple turns [356]–[358] in combination with a knowledge base to keep the dialogue history.

As both chit-chat and task-oriented dialogue systems evolve, both categories affect each other. So, the context-aware task-oriented dialogue system [359], which applies re-ranking to the candidate response given by matching function (§8.5.3) is the outcome of such synergy.

One of the latest directions in task-oriented dialogue systems in combination with the growing popularity of Pre-trained Language Models (PLMs) (§8.2.4) is that researchers and developers tend to use them not only for chit-chat but also for task-oriented dialogue systems [360]. Because it has been shown [361], [362] that the Generative Pre-Training (GPT) model (§8.2.3), once fine-tuned, can be useful in the domain of personal conversations (§8.8.1).

### 8.6.2 Reinforcement Learning (RL) Dialogue Systems

RL belongs to basic machine learning paradigms alongside supervised and unsupervised learning. It is a problem faced by a (software) agent that must learn behavior through trial-and-error interactions with a dynamic environment [363].

**Chit-chat** One of the common problems of DL (§8.6.1) chit-chat solutions is that they

[16] TC-Bot implementation §A.2

[17] KB-InfoBot implementation §A.2

[18] MEM2Seq implementation §A.2

suffer from being repetitive and producing generic responses [364]. This is the motivation why to use Reinforcement Learning (RL) for open-domain dialogue systems.

The RL model [365] simulates dialogues between two virtual agents using policy gradient methods, and it was compared to two agents using a 4-layer LSTM encoder-decoder as the baseline. Both approaches were trained on the OpenSubtitles [223], [224] dataset (§7.2.3). The comparison shows that the proposed algorithm generates more interactive responses and manages to foster a more sustained conversation in a dialogue simulation.

Another chatbot solution [366] is built on top of the latent action-framework that treats action spaces of a E2E dialogue agent as latent variables and develops unsupervised methods in order to induce its own action space from data.

The chit-chat dialogue system [367] solved with action spaces (representation of a type of meaning, for instance, greeting, question around a topic, statements around a topic, etc.) derived from unsupervised clustering is a recent contribution to the RL chatbots. It uses the reward function which is based on human-human dialogues and noisy dialogues for learning to rate good vs. bad dialogues.

The ensemble (§8.8.2) of particular sources or functionality of the dialogue system is a common approach for complex dialogue systems (§8.8.2). Inspired by the previous chit-chat solution [367], an ensemble version of the chatbot [368] with a novel approach for chatbot training by using value-based RL and reward function was prepared.

To improve results in diversity and provide interesting and non-redundant responses chit-chat solution [369] was formulated as the dialog attribute prediction RL problem. It uses policy gradients methods to optimize utterance generation using long-term rewards.

**Task-oriented** RL is a popular approach for learning an optimal Dialogue Management (DM) in the task-oriented dialogue systems. The dialogue flow requires significant hand-craft effort, instead, the Dialogue Management (DM) module can be cast as a continuous Markov Decision Process (MDP) (Partially Observable Markov Decision Process (POMDP)) or different method and trained through RL [370].

A Simple deep Reinforcement Learning (RL) dialogue system (SimpleDS) [371] uses raw, noisy text without any engineered features to represent the dialogue state. Such dialogue system does not require a NLU component which is bypassed by learning Dialogue Policy (DP) directly from (simulated) speech recognition outputs. The RL agent receives the state and reward, and updates its policy during learning.

The Reinforcement Learning can also serve for jointly learning policies for both NLU and Dialogue Management (DM) [372] for the End-to-End (E2E) task-oriented chatbot.

RNN can be used for E2E learning of task-oriented dialog systems [185]. The main component is a LSTM optimized by RL, which maps from a raw dialogue history directly to a distribution over Dialogue Acts (DAs). LSTM automatically infers a representation of dialogue history, which saves the work on manual feature engineering of dialogue state.

Another E2E approach is the multi-turn dialogue agent with Knowledge Base (KB) [351] showing that KB lookup helps the reinforcement learner discover a suitable Dialogue Policy (DP).

An ensemble dialogue system (§8.8.2) MILABOT [373] implemented as a part of Alexa Prize Challenge (APC) (§2.4.2) uses RL selection policy for one of 22 response modules.

The popularity of neural network-based task-oriented dialogue systems, which are end-to-end optimized with deep RL led to a solution [374] where dialogue-level LSTM is combined with the knowledge base for request information retrieval and Multilayer Perceptron based policy network overall Dialogue Acts (DAs).

The recent work represents an AgentGraph [375] universal framework with structured deep RL which tries to solve two main challenges for RL models.The first challenge is the efficiency of training RL based models. The second one is related to the RL policies transfer between different domains. The framework is based on a Graph Neural Network (GNN) [376].

### 8.6.3 Transfer Learning (TL) Dialogue Systems

The motivation for use of TL in the field of machine learning goes in the mid of the 1990s to the NIPS-95 workshop on "Learning to Learn"[19]. It is a technique where the initial model was trained on the large dataset after random initialization of the parameters (weights) to acquire general concepts. Then those general concepts are adapted through the TL technique to another model where an ANN (§8.1) is initialized with pre-trained weights from the intial model. Finally the the model created by Transfer Learning (TL) is fine-tuned on a specific task with a small dataset to allow the Natural Language Modeling (NLM) (§8.2) converge faster and with relatively lower requirements of fine-tuning data [309], [377].

The distilling knowledge (§8.3.1) is sometimes misleadingly considered as Transfer Learning, but it is not. Even though it looks like the same technique, it is just similar; it works based on the loss comparison between teacher and student models.

The TL based models use the Pre-trained Language Model (PLM) (§8.2.4) to achieve better results in the various Natural Language Processing (NLP) applications.

The broadly known Natural Language Modelings (NLMs) use Transfer Learning. For instance, the modern word embedding (§8.2.2) NLMs includes CoVe, ULMFiT, and ELMo and the sentence embedding (§8.2.3). Lately also the Pre-trained Language Models (§8.2.4) are used for TL. It includes all kind of transformer based language models, for instance GPT based conversational agent [361] with TL approach, or TL based BERT fine tuned dialogue system [378].

The latest attempts in TL are based on the Zero-Shot Learning (ZSL) [379] technique usually used for image classification. The image classification ZSL aims to recognize

[19] http://socrates.acadiau.ca/courses/comp/dsilver/NIPS95

objects whose instances may not have been seen during training, for instance, attribute-based. The technique was successfully (outperforming other techniques) used in research paper with cross-lingual task-oriented dialog system [380] where no annotated data in the target language exists. Another cross-lingual task-oriented dialogue system combines ZSL with transferable latent variables [381] to achieve better performance. The above mentioned BERT fine-tuned dialogue system [378] also uses the schema-guided ZSL Dialogue State Tracker.

### 8.6.4 Active Learning (AL) Dialogue Systems

The general concept goes back to the beginning of 1990s. It is based on predictive modeling [246] and active learning [382] which has two other related concepts: never-ending language learning [383] and lifelong learning [384].

The main idea of AL dialogue systems comes from the fact that purely trained and deployed chatbots leave a vast store of potential training data unused [385]. With this idea in mind, they trained a dialogue system based on the PersonaChat dataset (§7.2.3) having over 131k of training records. The following schema (Figure 8.15) describes the approach.

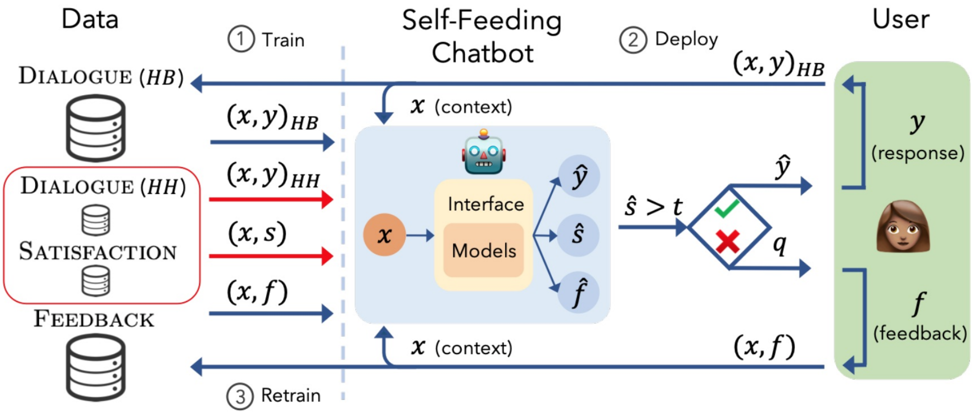


Figure 8.15: Self-feeding chatbot scheme

1. The chatbot is first trained on the Human-Human (HH) Dialogue $(x, y)_H H$ and Satisfaction $(x, s)$ data.

2. When the predicted satisfaction $\hat{s}$ is above the threshold $t$, new Human-Bot (HB) Dialogue $(x, y)_H B$ record data is extracted and the conversation continues with $y$ response. Otherwise, the chatbot asks about feedback via the question $q$ and extracts Feedback record $(x, f)$ data.

3. The chatbot is periodically retrained on all the data.

The self-feeding chatbot is not only existing approach, there are other works using deep active learning [386] or self-supervised feature learning [387], which more or less follows the idea of online human-in-the-loop active learning.

**Ethics of Active Learning**

With active learning dialogue systems, there is always a question of ethics because they are purely dependent on user responses from which the records for training based on the satisfaction and feedback from the user are extracted.

It can lead to the potential abuse of active training chains in favor of conversational bias as it happened with Microsoft's chatbot Tay (§6.14.3).

The approach to control the unwanted direction of conversation can follow various Dialogue Policy (DP) implementations. DP controls harmful behavior and exclude bad records from the retraining process.

### 8.6.5 Adversarial Learning (AL) Dialogue Systems

AL is a technique employed in the field of Machine Learning (ML) which attempts to fool models through a malicious input.

It specifically aims at filling in the gap between potential train/test distribution mismatch and revealing how models will perform under real-world inputs containing natural or malicious noise [388].

In the dialogue systems the AL framework is used in recent research of multi-turn Response Selection (RS) (§8.5.3) dialogue systems [389], [390] that are enhanced with persona-based (§8.8.1) dialogue model [391].

The Adversarial Learning is also applicable to various pipeline methods (§8.7) based dialogue system. It specifically improves Natural Language Understanding (NLU) [392] and Natural Language Generation (NLG) [393], [394] pipeline components.

The research paper [388] discusses a comprehensive study about adversarial oversensitivity (request) and over-stability (response) strategies related to task-oriented dialogue models. The paper tests those strategies with three state of the art dialogue models: Variational Hierarchical Recurrent Encoder-Decoder (VHRED) [395], Reinforcement Learning (RL) [365], and Dynamic Knowledge Graph Network [396]) using AL in order to assess dialogue system.

Diversity of dialogue system responses is a recurrent problem of Deep Learning (DL) dialogue models [364]. AL can serve as an improvement method to generate informative and diverse conversation [397]. And it can be used next to the Reinforcement Learning (RL) (§8.6.2) and dialogue systems ensemble (§8.8.2).

### 8.6.6 Hybrid Dialogue Systems

The hybrid dialogue systems are based on the Hybrid Code Network (HCN), which combines a RNN with domain-specific knowledge encoded as software and system action templates.

HCN is a reaction on the fact that E2E RNN dialogue systems are data-intensive and require thousands of dialogues to learn simple behaviors. HCN can be optimized with Supervized Learning (SL), Reinforcement Learning (RL), or a mixture of both [**Williams2017HybridL**

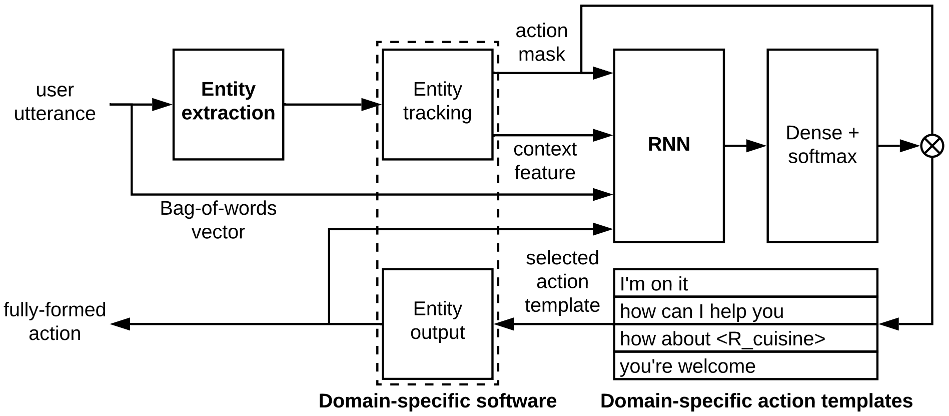


Figure 8.16: The Hybrid Code Network (HCN) model overview

At a high level, the four components of HCN (Figure 8.16) are:

- Conventional entity extraction module
- Recurrent Neural Network (RNN)
- Domain-specific software
- Domain-specific action templates

Both the RNN and the developer code (domain-specific software) maintain state.

The extension [398] (Figure 8.17) to the original work [**Williams2017HybridLearningb**] (Figure 8.16) provides trainable parts to the entity tracker and the entity output module which were designed with developer codes (originally hand-coded) in the original HCN.

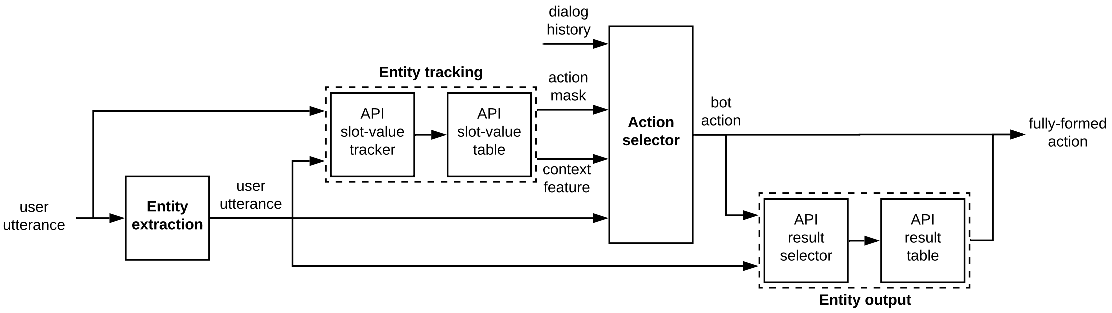


Figure 8.17: Overall structure of extended HCN

## 8.7 Pipeline Methods

The pipeline methods (§6.3.1) of the dialogue system consist of several pipeline components include Natural Language Understanding (NLU) (§6.8), Dialogue Management (DM) (§6.9) with Dialogue State Tracker (DST) and Dialogue Policy (DP), and Natural Language Generation (NLG) (§6.10) (Figure 8.18).

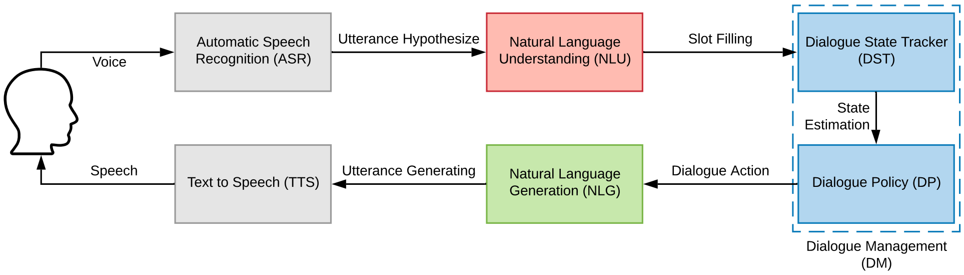


Figure 8.18: Dialogue System Pipeline

Plenty of different methods evolved during the years for each pipeline component; these are described in the next sections below.

NLU (§8.7.1) and NLG (§8.7.3) components use in some cases delexicalization [399]–[401] process (the values of attributes are replaced with placeholders), for example as it is shown in Table 8.3.

| **Utterance** | find | flights | to | new | york | tomorrow |
|---|---|---|---|---|---|---|
| **Slot filling** | O | O | O | B-Dest | I-Dest | B-Date |
| **Delexicalization** | find | flights | to | B-Dest | I-Dest | B-Date |

Table 8.3: Slot filling and delexicalization example for finding the flight corresponding to Table 6.1

It provides better results when compared to models without delexicalization because of less sparse training data. Both pipeline parts (NLU and NLG) uses also opposite process of **lexicalization** [400] (or **relexicalization** [401]). In such process the placeholders are replaced back by the current values of attributes in the automatically generated (delexicalized) sentences.

DM (§8.7.2) is usually built as an End-to-End (E2E) component or consists of Dialogue State Tracker (DST) and Dialogue Policy (DP) parts. In the recent works it is solved by Reinforcement Learning (RL) (§8.6.2), especially the Dialogue Policy (DP) part.

### 8.7.1 Natural Language Understanding (NLU)

NLU (§6.8) aims to extract semantics from user utterances (§6.8.1). Specifically, it detects intent (§6.8.2) and does the slot filling (§6.8.4) [402].

According to the task-oriented spoken language understanding review paper [19] the classification (Figure 8.19) of NLU approaches is following.

The traditional NLU pipeline approach is to manage the two above mentioned tasks separately. However, recent approaches tend to do it as a joint task.

**Independent Slot filling** Slot filling is considered as the sequence labeling task. Early methods which dealt with slot filling were rule-based or dictionary-based. Later statistical methods were taken into account. The popular approaches which do the joint detection of intent and slot filling are for instance Support vector Machine (SVM) [403] and Conditional Random Fields (CRF) [404].

With deep learning methods based on Artificial Neural Network (§8.1.2) the approaches for slot filling include, among others, Convolutional Neural Network (CNN) [405], deep Long / Short Term Memory (LSTM) [406], Recurrent Neural Network (RNN)

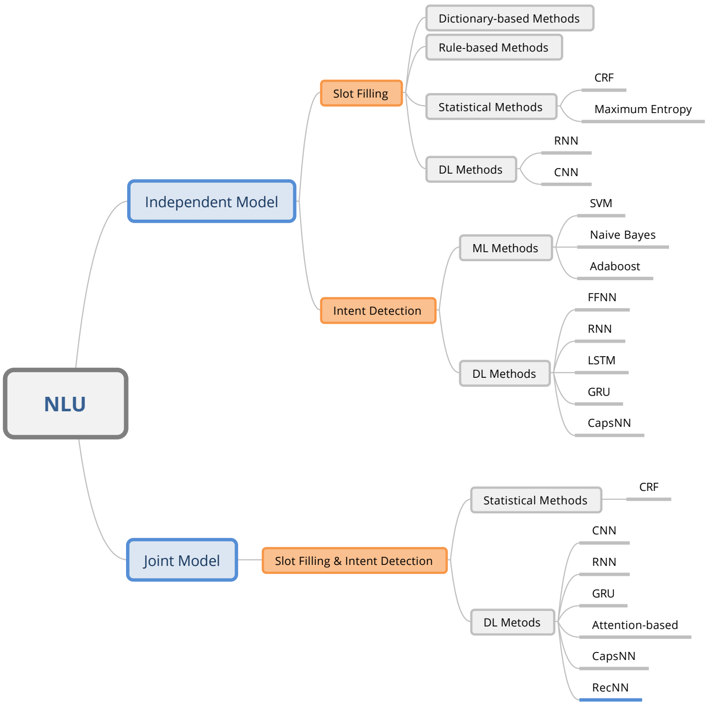


Figure 8.19: Review of NLU approaches classification

Deep Learning (DL) methods for intent classification can be done by various ANN (§8.1.2) including CNN [410], [411], and LSTM [412].

It can be improved with attention mechanism (§8.2.3) by including attention-based CNN [413], hierarchical attention networks [414]. It can be also part of various multi-task solutions like multi-task Adversarial Learning (AL) [392]. Another improvement of CNN is intent detection via Capsule Neural Network (CapsNN) [415].

**Joint Slot filling and Intent detection** The pipeline of independent intent detection and slot filling does not always bring the best performance due to error propagation, so there is a tendency to develop a joint model.

Joint modeling approaches include CNN with Triangular CRF ([416]), RecNN [417], joint RNN-LSTM [418], attention-based bidirectional RNN [162], slot-gated attention-based model [419], and Capsule Neural Network (CapsNN) [420].

Recently, Pre-trained Language Models (PLMs) (§8.2.4) play an important role also in NLU tasks, for instance BERT [421] and its multi-lingual usage [422].

### 8.7.2 Dialogue Management (DM)

According to Henderson et al. [20] and Williams et al. [21] the Dialogue Management (DM) consists (Figure 8.20) of Dialogue State Tracker (DST) and Dialogue Policy (DP) (also called Policy Learning).

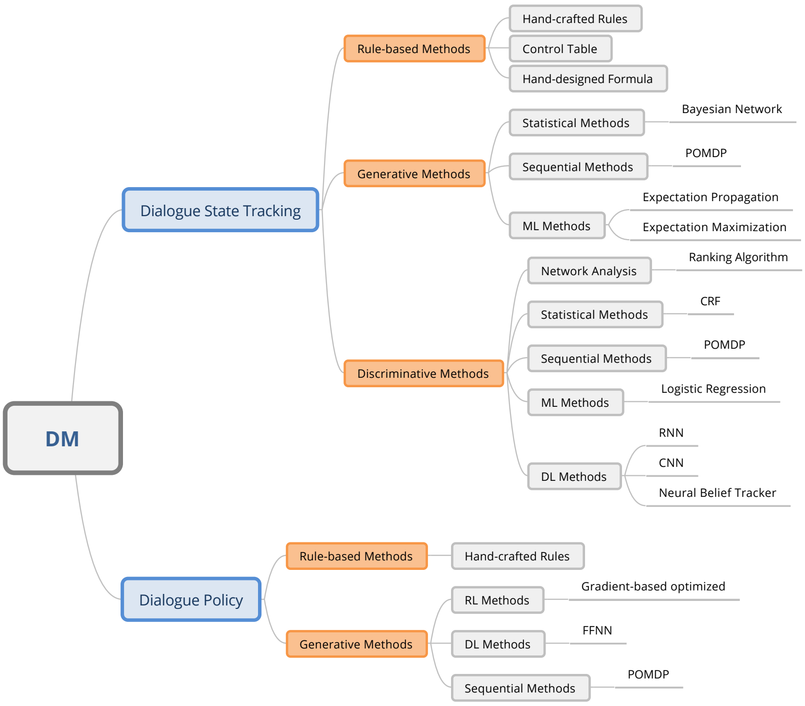


Figure 8.20: Review of DST approaches classification extended with DP approaches classification from other state of the art papers (§2.3)

**Dialogue State Tracker (DST)** Tracking dialogue states estimates the users goal at every turn of the dialogue.

The first dialogue systems used hand-crafted rules for DST. With hand-written rules DST can be solved by various approaches stored in a dialogue control table [167], tracked via a rich data structure [168], or computed as scores for all dialogue states [423], [424].

Modern methods used to deal with DST are based on **Generative** and **Discriminative** models.

The generative approach can be modeled as a Bayesian Network. The early stages of DST enumerated all possible dialogue states and then used a Bayesian Network to score those states [425]–[427]; it leads to the enormous number of states. There are two approximation processes: a beam (only most likely members of states) [187], [428]–[430] or further factorization [170], [431], [432].

The Bayesian Network accounts for different factorizations of the hidden state. For instance, it includes the variants of history accumulation [431] and separate random variables for an unobserved dialogue action and underlying intention [428], [429].

Model parameters come either from labeled dialogues or are inferred from unlabeled dialogues [169], [170].

The pioneers of discriminative DST are Bohus and Rudnicky [171] with hand-written rules which enumerate a set of k dialogue states. Other variations of such approach include logistic regression [172], ranking algorithm [173] and deep learning [174]. When the dialogue is modeled as a sequential process other methods like the discriminative Markov Model [175], [433], CRF [176] or RNN [177] can be applied. With a small amount of data for training the target domain in DST, the multi-domain learning [434] or unsupervised approach [435], [436] can be used.

**Dialogue Policy (DP)** It is learning to generate the next available Dialogue Act (DA) based on the state representation from DST. The initial base data of DP can be hand-crafted [437] and later used for Supervized Learning (SL) or Reinforcement Learning (RL) to optimize DP learning [438].

The policy can be implemented as Supervized Learning (SL) [183] by use of ANN. More specifically the following approaches can be used: Feed Forward Neural Network (FFNN) [439], the Reinforcement Learning (RL) [186] optimized with Q-learning [440], another RL [441] optimized with Natural Gradient [442], or Reinforcement Learning (RL) implementation [**Williams2017HybridLearningb**] uses DP optimized with Simple Statistical Gradient [443].

Another possible approach is via Partially Observable Markov Decision Process (POMDP) variation called Hidden Information State (HIS) [187], or POMDP with Gaussian processes modeled DP [188]. The policy modeled with Gaussian processes can be also in combination with RL [444].

### 8.7.3 Natural Language Generation (NLG)

The NLG goal is to express the components (attributes and values) of a meaning representation as a fluent natural language text.

Typical tasks of NLG are text summarization, creative text generation, and dialogue generation.

According to the comprehensive overview evaluating E2E NLG [25] by Dusek et al. we can see several NLG approaches used in past years (Figure 8.21) with new trends presented for instance at the NeuralGen 2019 workshop [20] with the topic Methods for Optimizing and Evaluating Neural Language Generation.

[20] https://neuralgen.io

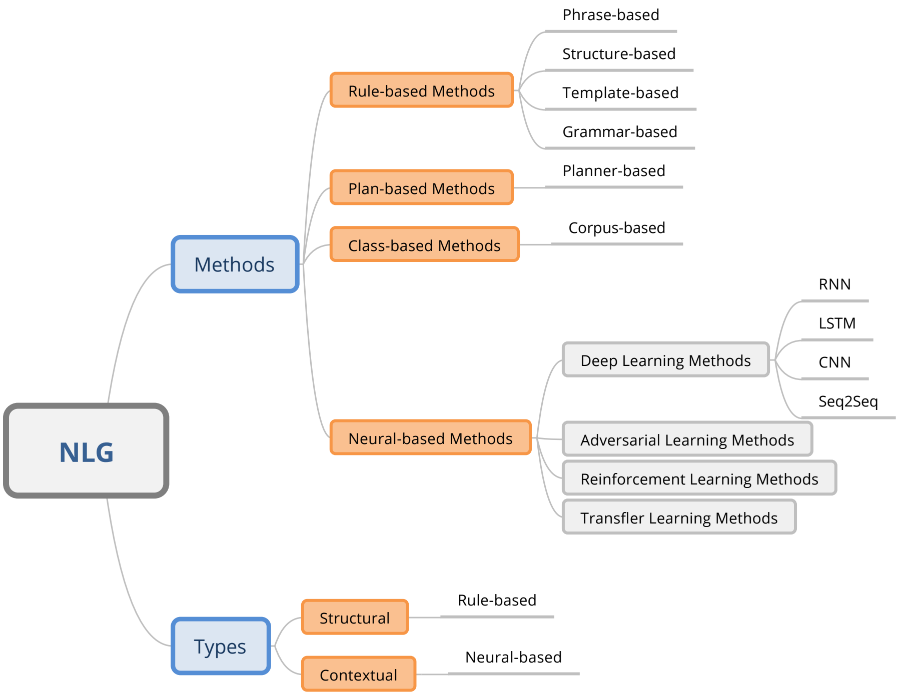


Figure 8.21: Review of NLG approaches classification

Early approaches of NLG (also used today because of their simplicity) are rule-based (structural types [445]). The rules are solved by different approaches, it can be phrase-based generation with active learning [446], structure-based generation [447], combination of template-based and grammar-based [448] generation.

Next to the rule-based and hand-crafted NLG methods the class-based [449] and plan-based [450], [451] methods to generate the text were also proposed.

Later when the neural-based approaches (contextual types [228]) of NLG have come into focus the various ANN with Seq2Seq (§8.2.3) approach have been used for implementation. More specifically, solutions with LSTM [452]–[454], Bidirectional LSTM or CNN2LSTM [455], and GRU [401], [456] have been used.

The latest approaches in NLG field use Adversarial Learning (AL) [393] including dual AL, which utilizes the duality between request and response generation to avoid safe responses [394]. Also Reinforcement Learning (RL) [365], [457] and large-scale Transfer Learning (TL) [362] are used for NLG. One of the latest contributions to NLG research is TL based 17 billion parameters big Pre-trained Language Model (PLM) named Turing-NLG [458] from Microsoft.

## 8.8 Various Improvements of Dialogue Systems

Achieve an attractive, long, and comprehensive conversation on multiple topics is not a simple task. Researchers and developers use different approaches to do it.

One of the approaches which are rising within recent years is a personalized dialogue

(§8.8.1). It includes, in the last two or three years, also empathy and emotions.

The dialogue system ensemble (§8.8.2) is the way to deal with a requirement to have a long and comprehensive conversation. It allows reacting with different responses (generated by ensemble dialogue models) on multiple topics. It is usually implemented either as a pipeline architecture (§6.3.1) with Dialogue Management (DM) which switches the topic or as an End-to-End (E2E) architecture (§6.3.2) with combination of topic specific training data.

### 8.8.1 Personalized Dialogue Systems

To provide more specific conversation, dialogue systems are becoming more personal. It has happened in several ways.

The first one is that they are **personalized**. It means they collect user's individual data and use it (in the right way) in further conversation. Secondly, dialogue systems have given themselves the **personality** with a few attributes (for instance, 22 year old man, who draws the comics and works in the flower shop); this serves for grounding the conversation. The last and most recent activity is to combine the dialogue with emotions to provide more **empathetic** conversation and thus improve user experience.

Nevertheless, the borders between the personalized dialogue system that adapts itself to user behavior, dialogue system having its own personality by given or learned personal attributes, and empathetic ones are in many research papers thin or washed away.

**Personalized** It is difficult to train a personalized task-oriented dialogue system because the data collected from each individual is often insufficient. Personalized dialogue systems trained on a small dataset can overfit and it is difficult to adapt them to different user needs [459].

Similarly, to the overall dialogue systems classification (§8.4) also the personalized dialogue systems could be categorized either into rule-based [460], [461] dialogue systems which utilize the memory or knowledge base for storing user's personal attributes and learning-based [459], [462], [463] dialogue systems which use Transfer Learning (TL) (§8.6.3) and Reinforcement Learning (RL) (§8.6.2).

Recent research focuses on detail techniques related to diversified personal traits [220]; e.g. large-scale **PersonalDialog** (§7.2.2) dataset was collected. Two techniques, **persona-aware attention** and **persona-aware bias**, were invented to capture and address trait-related information.

Other recent research papers utilized for instance Reinforcement Learning (RL) for personalized Dialogue Management (DM) [464] and meta-learning [465] like extended Model-Agnostic Meta-Learning (MAML) [466] to personalize dialogue learning without using any personal descriptions.

**Persona-based** When we are talking about persona-based dialogue systems, it is not personalized for the user, but the dialogue system itself has its own personality. It reflects the reality of human conversation; we each have a personality with various attributes. Why not dialogue system?

The chatbot dialogue personalizing approach is not new; it has been applied for many years. The ELIZA [7] chatbot represents a simulation of a Rogerian psychotherapist. The next chatbot, PARRY [8], simulates a person with paranoid schizophrenia. The third example is the Eugene Goostman[21] chatbot, which pretends to be a 13-year-old boy from Odesa, Ukraine, who has a guinea pig pet and father who is a gynecologist.

Most progress during recent years in personalising a dialogue system has been done with establishing persona-chat [70] dataset (§7.2.3) during the second year of ConvAI competition (§2.4.4) at NIPS 2018.

During last two years, for instance, Adversarial Learning (AL) [390], [391] in combination with persona-based Seq2Seq (§8.2.3) dialogue model has been used. Pretrained Language Models (PLMs) (§8.2.4) are also popular and used for personalized dialogue modelling [467].

The latest research related to persona-based dialogue systems is going deep into the topic and deals with the issue that responses are not only natural, but also consistent with the defined persona [468], [469], i.e., responses correlate with persona definition and they are not contradictory.

**Empathy** Topic personalization goes even further and the latest research focuses on empathy in a dialogue. It starts with designing chatbots [470] using new Empathetic-Dialogues dataset[22]. Emphatic dialogue systems [471] utilize emotional embeddings to generate emotional responses and even modeling empathy in a dialogue [472] to understand user emotions and reply to them appropriately. It ends up with HappyBot [473], the dialogue system that generates empathetic dialogue responses.

### 8.8.2 Ensemble Dialogue Systems

To keep the user entertain and focused on the conversation with the chit-chat dialogue systems within the open domain (§6.5.1) is practically impossible. Dealing with any conversation topic for an indefinite time cannot be done without planning or limiting dialogue in some particular way.

Various limits are given to the user as the fact that needs to be accepted. From the topic perspective, it is limited, for instance, to specific dialogue topics. What also helps is dialogue system personalizing (§8.8.1) with a specific person-like attributes (gender, age, place of living, occupation, hobbies, and interests). Another strict rule is to limit the time which cannot be overreached or which is a criterion for chatbot success (for instance, adjustment of Turing test (§10.2.1) as Turing time (§10.2.1)).

On the other hand, generative dialogue systems tend to generate highly generic responses such as *I don't know* or *I am OK* regardless of the input [364]. So, next to the Reinforcement Learning (RL) and Adversarial Learning (AL) the ensemble learning is another approach to solve such issue.

[21] http://eugenegoostman.elasticbeanstalk.com

[22] https://github.com/facebookresearch/EmpatheticDialogues

In 2017 Amazon established Alexa Prize Challenge (APC) (§2.4.2) with 20 minutes conversation length criteria as the goal for success in contest. Analysis of all solutions submitted to the competition reveals the fact that most of the solutions are ensemble-based, i.e., they combine various techniques to satisfy diverse requirements like Question-Answering (QA), news, weather information, personal questions, and others during the conversation.

APC solutions introduce the following ensemble techniques, for instance: The MILABOT chatbot works with 22 response modules [41] managed with the Dialogue Management which uses a reinforcement learning-based selection policy. The first version of Alquist [32] combines top-level and topic-level dialogue managers. The top-level makes a decision which module should be used, the topic-level switches between topics. The second version [46] uses an ontology-based topic structure called topic nodes, which consist of several sub-dialogues that are triggered based on the user intent or existing topic hierarchy.

Most, if not all, of the APC related solutions use the dialogue management, which serves as the correct topic selector to chose appropriate module, which provides the response to the user. The fully generative approach has also been investigated as the combination of task-oriented spoken dialog systems with chatting capability [474]. One of the latest approaches to the generative model-based dialogue ensemble is Attention over Parameters (AoP) [475] approach, which utilizes the Transformer architecture (§8.2.3) to model multiple conversational skills in different dialogues domains (task-oriented hotel booking, train reservation, chit-chat, etc.).

## 8.9 Pathologies of Generative Methods

Every system produces a specific type of errors related to the field of operation. Dialogue systems are not an exception.

As we have already discussed in Introduction (§6) a dialogue system takes as the request a sentence and returns a response sentence. Retrieval-based models (§6.4.1) are implemented as deterministic. Then the measure of error is given by the translation of the input sentence to output sentences. On the other hand, generative models (§6.4.2) are implemented with uncertainty given by the model type and training data embedding. It leads to potential dialogue system errors.

More specifically, we see these errors connected with a specific task, for instance, Question-Answering (QA) or various machine translations tasks like Neural Machine Translation (NMT) [476] or Statistical Machine Translation (SMT), image captioning [477], text or sequence generation tasks that the dialogue system evolves from.

Errors generated during performing these or similar tasks can be the following ones:

**Imaginary or made up words** Not only humans have imagination to create new words, also machines create new words (Code 8.3). It happens especially when the dialogue or translation system work with letters or syllables, not with words. Then it is easy to combine letters or syllables and create words which do not exist.

```
Expected Output: Carboxysomes are found in [lithoautotrophically] and
```

```
mixotrophically grown cells. Carboxysomes aid carbon [fixation].
Output: Carboxysomes, which aid carbon [fixotrophically] and
mixotrophically grown cells.
```

Code 8.3: Imaginary or made up words

**Repeated words or phrases** It happens whenever the dialogue or translation system does not correctly estimate the next word in the sentence and instead repeats the word (Code 8.4) used recently based on the same or similar preceding word.

```
Expected Output: I am your employee, to serve on your company.
Output: I am your [company], to serve on your [company].
```

Code 8.4: Repeated words or phrases

**Abrupt ending or premature end-of-sentence** The output generation ends with a fragment of the sentence (Code 8.5) which does not make sense or the sentence makes sense, but its significant part is missing.

```
Expected Output: By the way, my favorite football team is Manchester
    United, they are brilliant, they have an amazing football players, and
     they are awesome.
Output: By the way, my favorite football team [is].
```

Code 8.5: Abrupt ending or premature end-of-sentence

**Hallucination** The outcome is wrong and does not have any signs of pathological behavior but reminds human hallucination behavior (Code 8.6).

```
Expected Output: If you are [interested], find me at 8 o'clock near the
    cinema entrance.
Output: If you are [play], find me at 8 o'clock near the cinema entrance.
```

Code 8.6: Hallucination

**Coreference issues** Referencing a wrong subject or object (Code 8.7) with a pronoun or directly in the sentence is another issue of generative methods.

```
Expected Output: She is the daughter of Alistair Crane [who] secretly
    built...
Output: She is the daughter of Alistair Crane. [She] secretly built...
```

Code 8.7: Coreference issues

**Misleading rephrasing** Rephrasing or paraphrasing (Code 8.8) is a machine translation problem similar to understanding the meaning and explaining it by other words. It is not easy for the human being so then it could be a problem for the machine.

```
Expected Output: The article proudly notes that the postal service
was [in no way responsible] for the 1996 crash of ...
Output: The article notes postal service [was responsible]
for the 1996 crash of...
```

Code 8.8: Misleading rephrasing

**Lazy sentence splitting** Splitting long sentences is sometimes necessary to keep the text readable, but to split sentences to often (Code 8.9) when they should be rather kept as one, is another language processing problem.

```
Expected Output: Homeworld of the Margiotta located in the Sagittarius Arm
Output: Homeworld of the Margiotta. [Located] in the Sagittarius Arm
```

Code 8.9: Lazy sentence splitting

With the evolution of embedding some of the errors are eliminated. For instance, all the above examples except the last one (Code 8.9) can appear with usage of Sequence to Sequence (Seq2Seq) (§8.2.3) embedding. Using Language - Agnostic SEntence Representations (LASER) [478] multi-lingual embedding the first three (Code 8.3, Code 8.4, Code 8.5) and last two are eliminated with high probability.

## 8.10 Conclusion! With AI, or without AI?

The previous overview of various dialogue system models gives the general idea of how complex the dialogue ecosystem is. The paraphrase of classic author question can be modified in the following way: With AI, or without AI? Furthermore, the answer does not seem to be straightforward.

The simplest solution ever for dialogue systems influenced by external data is to use retrieval based methods (§8.5). Easy manipulation of predefined responses to requests gives us a full control over dialogue influencing. Especially the rule-based (§8.5.1) dialogue systems can be good a starting point because of their simplicity.

On the other hand, generative methods (§8.6) are promising for the future. It would be pity not to try to use at least one of these techniques. For instance, human-likeness based [367] dialogue system or yes/no question experiment [372] are interesting applications of Reinforcement Learning (§8.6.2).

Another solution of this subgroup is represented by the Active Learning (AL) chatbot (§8.6.4) and Hybrid Code Networks (HCNs) (§8.6.6).

Next to the purely Deep Learning (DL) solutions there are ensemble dialogue systems (§8.8.2) represented by solutions under Alexa Prize Challenge (APC) (§2.4.2) or specific approaches like Attention over Parameters (AoP) [475].

Adversarial Learning (AL) (§8.6.5) dialogue systems, memory networks [479] based dialogue systems and last but not least personalized dialogue systems (§8.8.1) are equally interesting and worth further research.

The best practice would be to build all the above mentioned dialogue system examples first as retrieval method (§8.5) based dialogue systems. Later, with more effort they can be turned into generative methods (§8.6) based dialogue systems, which offer the potential for further research and experiments.

# Chapter 9

# Dialogue System Influencing

The basic idea of dialogue system influencing presented in the Introduction (Figure 1.1) can be extended (Figure 9.1) to the idea of particular dialogue system influencing technique (§9.3).

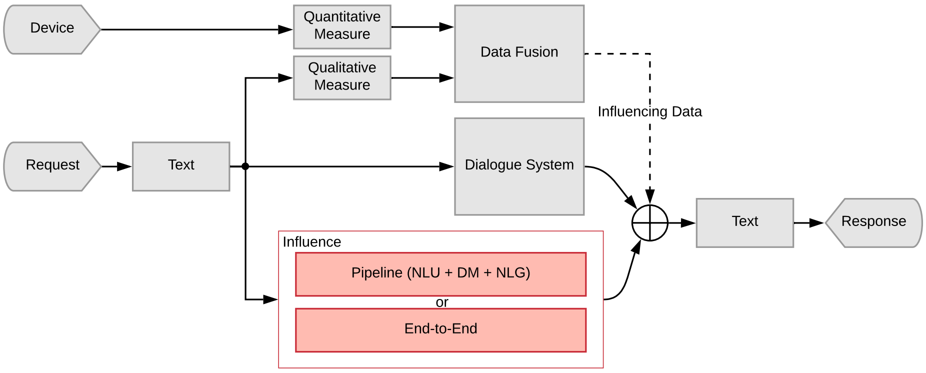


Figure 9.1: Idea of dialogue system influencing technique

The dialogue system influence can be done by some influencing inputs (§9.1). Those might be represented by various data. Based on the way the influencing data affects the dialogue system we can talk about influencing approaches (§9.2) in foreground (§9.2.1) or on background (§9.2.2).

Next to the influencing approaches influencing techniques can be discussed (§9.3); they correspond to the dialogue system architecture (§6.3) and its horizontal and vertical division. A particular influencing technique either affects a part of the pipeline architecture (§6.3.1) or the whole End-to-End (E2E) architecture (§6.3.2) based dialogue system.

When we know what influencing approaches (§9.2) exist and which techniques (§9.3) can be applied to influence the dialogue system we can introduce intervention methods (§9.4) which are triggered right after the dialogue system is dealing with influence.

Last but not least, whenever there are peaks in influencing data smoothing (§9.5) might be applied to reduce unwanted change of the conversational topic (§9.6).

## 9.1 Influencing Inputs

There are several ways the conversation within a dialogue system is influenced (see Figure 9.2). The most basic one is the reaction to conversation. The topics for conversation is obtained from the common local knowledge base and extended with an additional topic (for instance, personal information (§8.8.1)) to make the conversation more fluent. The next way is to get information online from the public on-line knowledge base and inform the user, for instance, about the weather, traffic or answer the factual questions. However, the influencing information could be also gathered from outside sensors. Those can be wearables or other smart devices storing data in the private static physiological base. Such base provides individual measured data and leads to the optimization of human well-being.

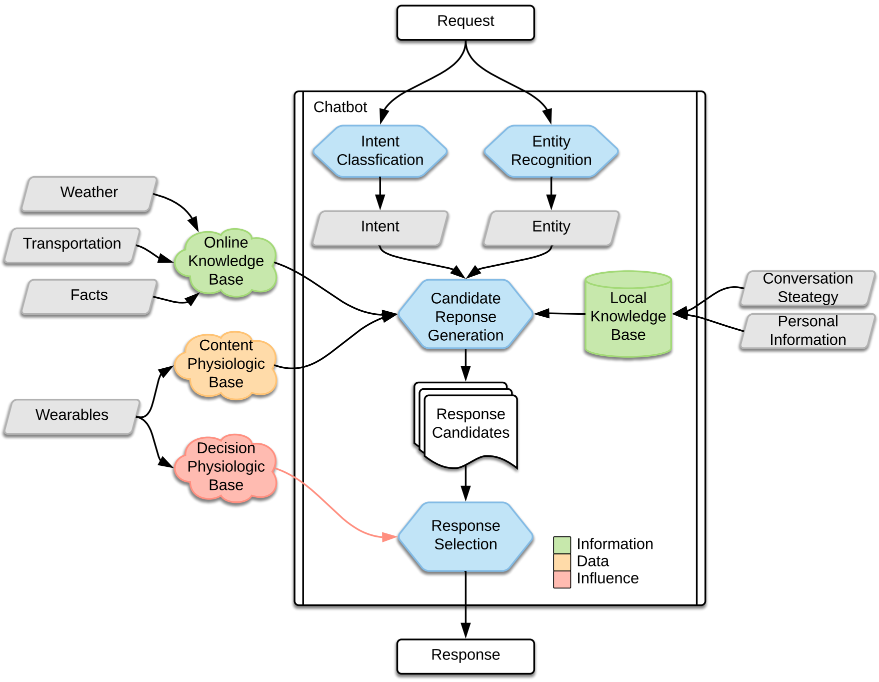


Figure 9.2: Dialogue system influencing logic

Influences are presented to the user in the conversation during the candidate response generation. The real-time physiological base influences the response selection. The already finalized response from multiple responses keeps the conversation unchanged or leads to an immediate change in conversation direction.

The difficulty of the dialogue system conversation influenced with external data or signal also depends on dialogue system complexity (§6.7).

## 9.2 Influencing Approaches

Going deeper into the issue of influencing we can recognize several existing influencing approaches used in the various dialogue system implementations, especially within task-oriented dialogue systems.

### 9.2.1 In Foreground

A standard conversational approach for most of the dialogue system is to respond (answer) to a particular request (question) as this is presented in Figure 9.3. It means that all provided information may influence the consumer behavior and leads to potential additional requests or change in conversation.

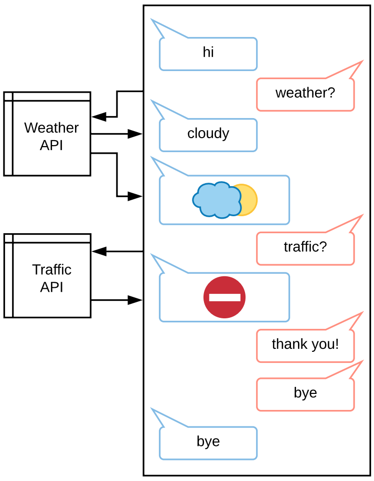


Figure 9.3: Dialogue system standard influencing approach

An extension to the standard conversational approach involves any collected or measured personal data which are not part of any users' request (question) and are still a part of the conversation as the response (answer). This approach is shown in Figure 9.4 either as raw or summarized (aggregation, graph representation) information. Such additional information may influence the dialogue system based on the results from measured data and also influences the consumer behavior and leads potentially to additional requests or change in conversation.

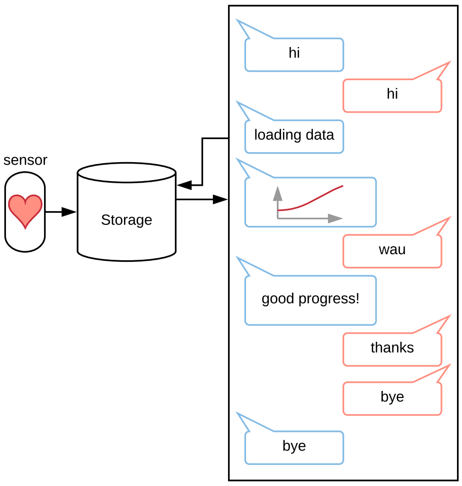


Figure 9.4: Dialogue system extended influencing approach

### 9.2.2 On Background

A less standard conversational approach which is barely seen (Figure 9.5) is to provide a response, not to the particular request only, but combine the request with additional data, either implicit (extracted from the conversation) or explicit (provided additionally). Such a combination of request and data may influence the dialogue system and therefore consumer behavior.

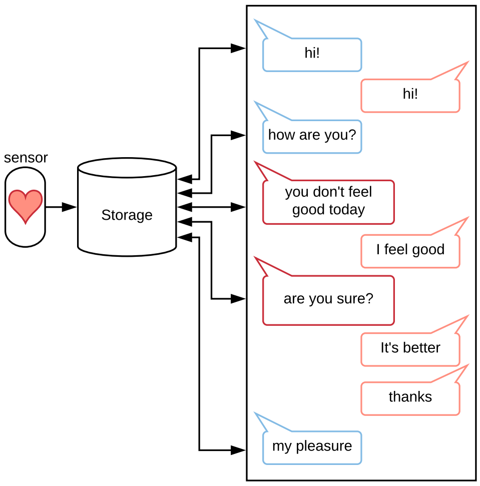


Figure 9.5: Dialogue system real-time influencing approach

## 9.3 Influencing Techniques

To identify possible influencing of the dialogue system and define the influencing techniques is necessary to consider both (in chapter §6 defined) architectures: the pipeline (§6.3.1) and End-to-End (E2E) architectures (§6.3.2).

Moreover, the influencing data (§4) and its fusion (§5) acts like the switcher applied to each pipeline component or the whole E2E where the horizontally divided architecture offers an option to apply this switch and to choose between the standard or influenced functionality.

The methods described in chapter §8 relevant to those architectures can be chosen from all three options, i.e. retrieval (§8.5), generative (§8.6), pipeline (§8.7) or even the dialogue systems or pipeline architecture modules ensemble (§8.8.2).

With all this in mind the following subsections present use cases of specific influencing techniques to each and every part of the dialogue system.

### 9.3.1 Generated Intent

Intent (§6.8.2) detected together with entities extracted (§6.8.3) from utterance (§6.8.1) serve as the main input for the Dialogue Management (DM) (§6.9) and Natural Language Generation (NLG) (§6.10) modules. The influencing method impacts Natural Language

Understanding (NLU) (§6.8) part (Figure 9.6). The same request (Input X) based on the influencing signal classifies a new intent (Intent 1) or the new intent with an extension (Intent 2) which reflects the influence and thus as well leads to various responses (Output A/B).

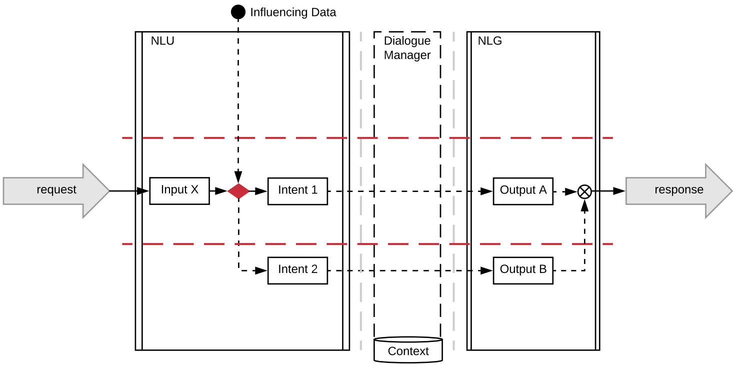


Figure 9.6: Conditionally Generated Intent

## 9.3.2 Affecting Slot Filling

This method influences Dialogue Management (DM) (§6.9) which can be driven by slot filling (Figure 9.7). This technique collects information about various subjects and objects during the conversation, which allows to react on the user request (Input X) in context and keep it as long as it is needed or makes sense.

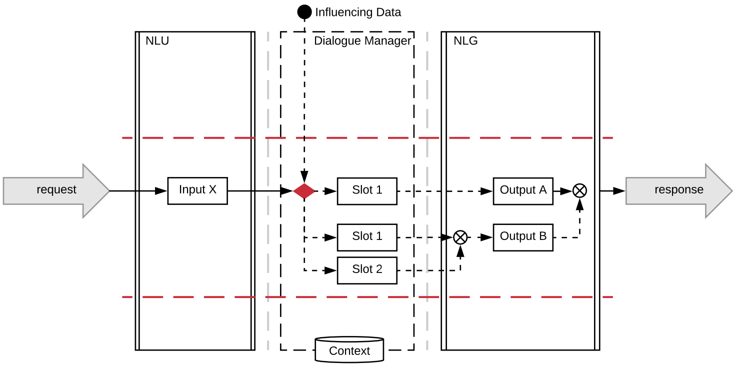


Figure 9.7: Affecting the Slot Filling

So, there is an opportunity to use slot filling for keeping the information whether the dialogue system could react differently based on the particular slot (Slot 1 + Slot 2) and generate corresponding responses (Output A/B). It does not matter if the additional slot is present only when influence exists or we work with a reserved place-holder which is by default populated by non-influencing information and changes whenever the influencing data comes.

### 9.3.3 Conditionally Chosen Response

This method influences the Natural Language Generation (NLG) (§6.10) part (Figure 9.8). It gives various responses (Output A/B) to the same request (Input X) based on the influencing signal.

When external influencing data is taken into account, the reaction to the request is chosen from the predefined responses, but we can define several options which are chosen based on the condition which depends on this influencing external data.

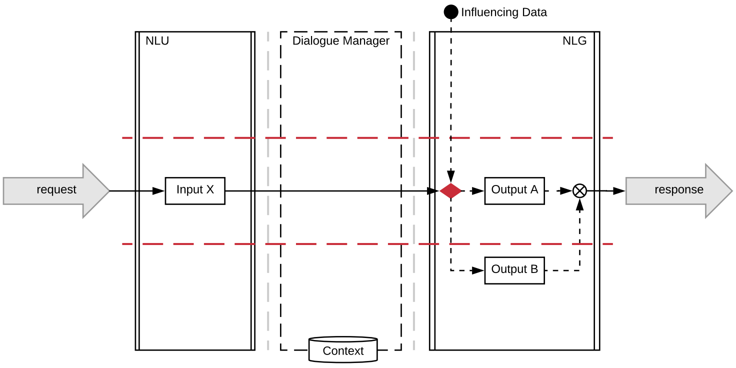


Figure 9.8: Conditionally Chosen Response

### 9.3.4 Conditionally Trained Response

Influencing a dialogue system based on a generative model (§8.6) implemented by an Artificial Neural Network (ANN) means to change its behavior by some additional input layer fed by influencing data which serves as the influencing feature (Figure 9.9).

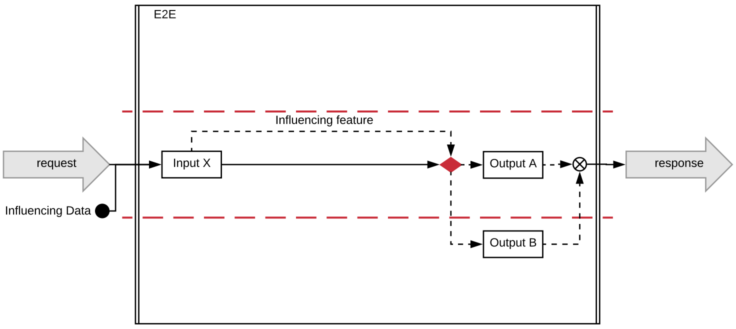


Figure 9.9: Conditionally Trained Response

For training, the End-to-End (E2E) model requires two corpora for translation of the request to the response. This pair of corpora contains the same requests (questions) with different responses (answers), and each corpus is related to a different value of influencing feature (for instance 0 and 1) which switches the conditionally trained response.

## 9.4 Intervention Methods

There are plenty of intervention methods which could be used to support or replace ambulatory treatment (§2.1). Some of them were already described in psychological and psycho-social interventions (§2.6.1) or mentioned in cognitive strategies to regulate emotions (§2.6.2).

Within teams that included psychologist or psychiatrist the chatbots like Woebot [6] or Lark [80] (see §2.4.6) which utilize Cognitive Behavioral Therapy (CBT) were developed. However, as it is mentioned in state of the art (§2), this therapeutic methods are too complex.

Emotion Regulation (ER) is an ongoing process of the individual's emotion patterns concerning moment-by-moment contextual demands. These demands and the individual's resources for regulating related emotions vary [480].

Reappraisal, distraction and labeling are cognitive strategies [92]; together with empathic paraphrasing [481] can be used to regulate emotions.

In the next section, definitions and applications distraction (§9.4.1), reappraisal (§9.4.2), labeling (§9.4.3) and paraphrasing (§9.4.4) as the processes of Natural Language Processing (NLP) suitable as intervention methods for the dialogue system are briefly described.

### 9.4.1 Distraction

Distraction, an example of attentional deployment, is an early selection strategy to regulate emotions. Distraction constitutes the deployment of attention away from a negative aspect of a situation, to a neutral or positive aspect [482]

As opposed to reappraisal (§9.4.2), individuals show a relative preference to engage in distraction when facing stimuli of high negative emotional intensity. This is because distraction easily filters out high-intensity emotional content, which would otherwise be relatively difficult to appraise and process [483].

When the distraction is applied through NLP there is a change of the conversation topic. The whole conversational context remains. The original negative topic is prohibited by the dialogue system (for instance slot filling (§6.8.4)) and negation techniques, see in (§6.9); it means the dialogue system does not mention it, just distracts the user whenever the dialogue comes to it.

The best way is to turn the conversation to the positive topic and evolve it (Code 9.1).

```
HUMAN: I have a fear of spiders.
BOT: What about ladybugs?
HUMAN: Ladybugs are nice.
BOT: Do you like to count the spots?
```

Code 9.1: Distraction conversation to deal with arachnophobia

### 9.4.2 Reappraisal

Reappraisal can be chosen as a late selection strategy from the cognitive change group of Emotion Regulation (ER). It works with the cognitive change of the meaning, i.e.

reinterpretation of an emotion eliciting-situation, in order to reduce negative feelings [482]. Successful reappraisal includes an underlying process of emotion recognition (that is part of the appraisal process). Emotion recognition can be explicit (e.g.,labeling (§9.4.3)) or implicit (e.g., awareness of the feeling) [92]

The NLP solution of reappraisal represents complex problem for Natural Language Understanding (NLU) and Natural Language Generation (NLG). The dialogue system needs to understand the whole topic from a broader perspective and have in the mind a big picture which is closely related to full understanding.

The conversation may look like the following one (Code 9.2) where the conversation contains reappraisal of the whole situation.

```
HUMAN: I have a fear of water.
BOT: What happened?
HUMAN: I was scuba diving and we had a problem with the high tide!
BOT: So, the problem is with scuba diving.
BOT: Was there an instructor?
HUMAN: No, just experienced friend.
BOT: Can you do your next scuba diving with the instructor?
HUMAN: Yes, but, the high tide...
BOT: Can you and your instructor carefully choose the time for scuba
    diving?
BOT: The low and high tide is happening in a specific time.
HUMAN: Yes, thank you, that sounds safer.
```

Code 9.2: Reappraisal the situation with water, scuba diving and related fears.

The reappraisal method works with the cognitive change of meaning, i.e. reinterpretation, which represents a complex NLP problem. So, it is much easier to chose another ER method. For instance, the paraphrasing (§9.4.4) method which is already subject of NLP research for some time could be used.

### 9.4.3 Labeling

Labeling belongs to the linguistic processing group of Emotion Regulation (ER) and like distraction (§9.4.1) is taken as an early selection strategy. It helps to decrease emotional reactivity [484]. It might be that similarly to distraction labeling allows dealing with highly intense emotional situations (e.g., exposure to phobic stimuli), but unlike distraction, it also allows learning, since the individual pays attention to the emotional stimulus [92].

The complexity of labeling methods as the NLP is in Natural Language Understanding (NLU) module. Even if it recognizes the emotional situation it needs to be annotated by the commonly known label. As it is mentioned above labeling is good to use against phobias when we recognize hundreds of them[1]. So, practically NLU relays on the manually attached label to every emotional situation either it is a phobic stimulus or just broken heart.

When the label for the particular recognized emotional situation is known then the conversation can look like in the following example (Code 9.3)

[1] http://phobialist.com

```
HUMAN: I am afraid of spiders.
BOT: Don't worry, a fear of spiders is one of the most common phobias.
BOT: We call it arachnophobia.
HUMAN: Ok, and what can I do about it?
BOT: Do you know someone who can provide you with psychological support?
```

Code 9.3: Giving the particular problem the label and method how to solve it helps.

### 9.4.4 Paraphrasing

Paraphrasing or active listening is a form of responding emphatically to the emotions of another person by repeating in other words what this person said while focusing on the essence of what they feel and what is important to them [481]

In NLP the paraphrasing problem is well known [485]. It includes proper understanding of the input phrase (NLU) and generating the phrase with the same meaning by other words using a paraphrase database [229], for instance the synonyms or the part of the sentences which describe the main entity keyword. The most recent techniques are embeddings [486] (§8.2), Reinforcement Learning (RL) [487] (§8.6.2) especially usage of Recurrent Neural Network (RNN) [488] or Long / Short Term Memory (LSTM) [489] and many other techniques.

To give someone the feeling of being understood and on the same page the paraphrasing is perfect technique (Code 9.4) for it as the next example shows.

```
HUMAN: I have a fear of storm!
BOT: I know, the storm can be terrifying for you!
BOT: The storm is not what terrifies you...
BOT: ... it is a fear of something powerful, which you cannot control.
BOT: But if you understand the matter of storm, you can feel better.
```

Code 9.4: Describing the problem the same way with different words brings the comfort.

### 9.4.5 Disclaimer

All the simplification of all previously described Emotion Regulation (ER) methods and their solution through NLP might and definitely have higher or lower measure of simplicity. This is given by the fact that this work does not have as primary objective to bring 100% correct approach of emotion regulation ER but demonstrates the capability of such methods.

## 9.5 Smoothing of Influencing Data

The influencing data like any data can have variations because they are based on the fusion of measured Heart Rate (HR) (§4.3) data and extracted sentiment (§4.2) data.

If the dialogue system influence happens often we need to eliminate those variations by some smoothing technique.

For time-series techniques (e.g. exponential or moving average) smoothing techniques are commonly known. However, in this case influencing data is represented by sequences, so we need to come with other smoothing methods (§5).

## 9.6 Conversational Topic Change

Dialogue system influencing means a sudden change in the ongoing dialogue that happens sooner or later. When the dialogue system starts to perform an intervention method (§9.4), there is a change of conversational topic. Sometimes this change is not drastic, but sometimes the Emotion Regulation (ER) technique requires such change (distraction (§9.4.1)).

The main question is how the human participant perceives the conversation with the dialogue system. Is it natural in the same way as the human to the human conversation or is it artificial and too obvious?

We know, and we can easily observe that human conversation can stick to the topic for a long time, but can also change topics several times during a few minutes. This change could be even very substantial, especially when conversation participants are arguing or external stimuli are coming.

The comparison between human to human and human to dialogue system conversation found in [490] presents notable differences in the content (exhibited greater profanity) and quality of conversation based on the statistical evidence. The duration of human to chatbot conversation lasts longer but contains shorter messages. Moreover, conversation richness and the vocabulary occurring during human to the human conversation is missing.

Since the human to human dialogue significantly differs from human to dialogue system dialogue, it seems that it does not matter if conversation due to ER method suddenly changes. Or on the contrary, it is maybe more expected especially when the subject knows that the dialogue system leads the dialogue.

## 9.7 Conclusion! Leadership is influence.

The dialogue system influencing consists of two independent parts: influencing techniques (§9.3) and intervention methods (§9.4). The potential combination of those two groups gives us theoretically up to sixteen combinations of methods to influence dialogue system on background (§9.2.2).

From four described intervention methods (§9.4) the paraphrasing (§9.4.4) method is the most broadly known NLP problem. Moreover, there exist resources (Paraphrase DB[2] [229]) and various methods (for instance Statistical Machine Translation (SMT) or Neural Machine Translation (NMT)) to implement such a method.

By limiting the intervention methods (§9.4) to paraphrasing (§9.4.4) we can narrow down the methods to just four combinations. The paraphrasing (§9.4.4) method is most suitable because it is not only natural to rephrase sentence as the help to understand someone else, but also it is interesting from the increasing research in this fields and it

[2] http://paraphrase.org/#/download

is becoming the commonly solved NLP problem applied in different research dialogue systems included.

Comparing to standard conversation chit-chat it is (from the intervention methods (§9.4) examples) evident that the dialogue system needs to lead the ER conversation and release this leadership back to the human when the intervention ends.

# Chapter 10

# Dialogue System Testing and Evaluation

Every system needs to be subject to testing and evaluation, especially when the system is complex. Till the expected outcome of the test either matches or not, the behavior of the system is a subject of the postulated hypothesis. By testing and evaluation, we can prove validity or invalidity of the hypothesis as the desired conclusion.

First of all, let us introduce dialogue system testing and evaluation (§10.1). We continue with the quick overview of the historical development of dialogue system testing (§10.2) from Alan Turing (§10.2.1) up to contemporary testing approaches. Hand in hand with that, we would like to evaluate a dialogue system (§10.3) model technically with metrics (§10.3.2) based on the various benchmarks (§10.3.3) and also touch its explainability (§10.3.4).

Last but not least, psychological feedback questionnaires (§10.4.1) are important. They give participating users a chance to evaluate the conversational process of intervention (§10.4) with the dialogue system subjectively.

## 10.1 Introduction to Testing and Evaluation

The difference between the test and evaluation is in its outcome. Of course, these two activities are joined vessels. Without the test, there cannot be results and evaluation, and without evaluation, we do not know how to interpret the test.

### 10.1.1 Introduction to Testing

The testing is either a manual or automated activity, which leads to the confirmation that a particular part of the system or the system overall represents desired functionality by expected outcome.

Based on this definition, we know that people or software are involved in the dialogue system testing. With the massive people participation, it is usually a crowd-funding activity, which helps with a small volume of human-base data collection establishing a human-baseline. On the other hand, software involvement helps automate to get vast volumes of data collections under various setups.

### 10.1.2 Introduction to Evaluation

Evaluation is the exact measurement activity where the outcome of a particular test is compared with the expected measure set either theoretically or empirically.

If it is possible to get them, human-baseline measures are usually taken into account first. If not, or if it does not make sense, then vanilla-based[1] solution is taken into account as the second one.

## 10.2 Dialogue Systems Testing

Among plenty of approaches how to test the dialogue system we can distinguish a few main groups.

The first group involves the human into the process and utilize the classical Turing test (§10.2.1), which has been lately better specified as the Turing time (§10.2.1). Next to the Turing test the broadly used and well known A-B test stands (§10.2.2).

The second group involves automated testing. It does not need the presence of humans in the testing process. It provides the functionality for example locally as the virtual container (§10.2.3) or as the service in the cloud (§10.2.4).

### 10.2.1 Turing Test

In 1950 Alan Turing published his well know article Computing Machinery and Intelligence [28]. In this article, he proposes how to test AI machines if they can think.

This test is called the Turing Test (Figure 10.1) and represents the ultimate goal of how to test the AI. A tester determines whether he/she chats with a dialogue system (chatbot) or human.

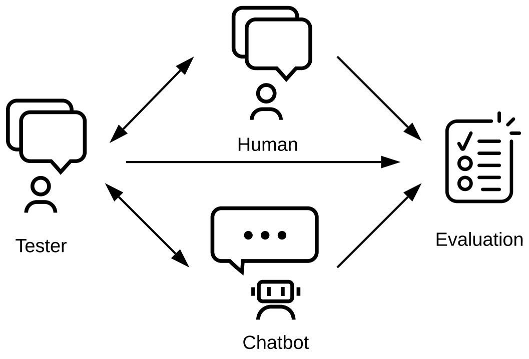


Figure 10.1: Dialogue system (chatbot) Turing test

The idea of the test is widely discussed again, attacked, and defended. With the rise of AI it becomes popular again [491].

[1] the simplest version of something, without any optional extras, the essential or ordinary

### Turing Time

The Turing test is the test defined without any limitation criteria. However, in the real-life, the conversation is not endless; it has its time window. From this perspective, we can turn the Turing test into Turing time [492], which makes the test easier achievable.

### Passing the Turing Test

There are several examples in the evolution of dialogue systems (§2.2) that show the possibility of passing the Turing test.

The ELIZA [7] chatbot is claimed by some to be one of the programs (perhaps the first) able to pass the Turing test [493].

The First known dialogue system is PARRY [8]. It passed the Turing test in 1972 when the psychiatrists cannot distinguish whether the transcripts of interviews come from PARRY or interview with real paranoids [494].

Another chatbot that, in some regard, passed the Turing test is Eugene Goostman[2]. Eugene Goostman pretends to be a 13-year-old boy from Odesa, Ukraine, who has a pet guinea pig and a father who is a gynecologist. It was implemented in 2001 and tested on 7 June 2014, at a contest marking the 60th anniversary of Turing's death.

On the other hand, AI researchers argue that trying to pass the Turing test is merely a distraction from more fruitful research [495] So, they have devoted little attention to passing the Turing test [496].

## 10.2.2 A/B Testing

A/B testing (Figure 10.2) is beneficial for a controlled experiment with two variants (A and B). For a dialogue system application this testing can be used whenever we have a variant with tested functionality (A) and without tested functionality (B) as the control group.

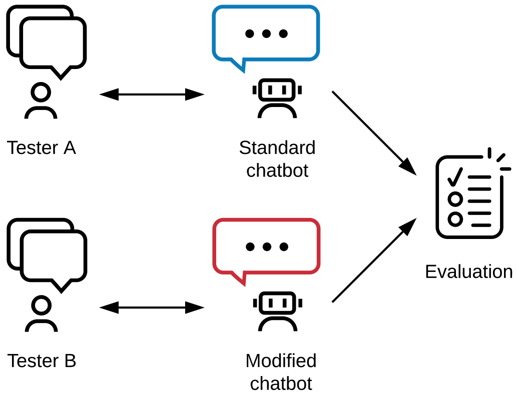


Figure 10.2: Dialogue system (chatbot) A/B testing

[2] http://eugenegoostman.elasticbeanstalk.com

### 10.2.3 Botium

In the software development world, we have several testing tools used as a standard in a particular field. From this group of testing tools, we can consider Selenium as de-facto-standard for testing web applications. Appium is the de-facto-standard for testing smartphone applications. Botium[3] can be considered as the standard for testing dialogue system applications.

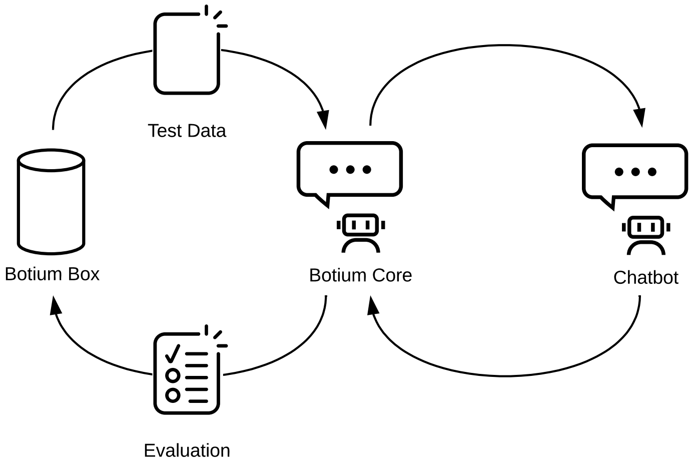


Figure 10.3: Dialogue system (chatbot) testing with Botium

Botium runs as a virtual container (Docker) and consists of two modules (Figure 10.3). Botium Core automates the conversation with a dialogue system based on the testing data. Those are prepared in the Botium Box, which also allows evaluating conversation with the dialogue system. Moreover, it enables the management of the whole testing process.

### 10.2.4 Google Chatbase

Primarily the Google Chatbase[4] is an automated testing tool. It is provided as a cloud service accessible through API and by libraries for various programming languages.

The Chatbase offers products for designing, analyzing, and optimizing dialogue systems. It provides detailed information about metrics, chat session flow, information of not-handled messages, suggests intents (§6.8.2) for missed and misunderstood messages, and other functionality.

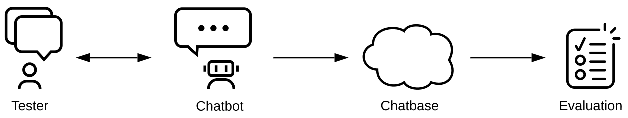


Figure 10.4: Dialogue system (chatbot) testing with chatbase

The main functionality idea of the Chatbase is the integration into the dialogue system (Figure 10.4). So, the dialogue system sends (during the dialogue with the user) the

[3] https://www.botium.at
[4] https://chatbase.com

dialogue parts into Chatbase via API calls. Furthermore, the Chatbase provides analysis and recommendations about the dialogue.

## 10.3 Dialogue System Evaluation

Deciding about the dialogue system testing approach and applying it to the dialogue system is only one aspect of overall testing. The outcome of the testing needs to be evaluated.

The dialogue system evaluation includes several approaches which are not such coherent as it is in the previous chapter about testing, but they correspond to various perspectives on how to evaluate such complicated thing like a dialogue system.

Dialogue systems are usually judged by a human and any individual evaluation serves as the baseline measure. The human evaluation is about various evaluation aspects (§10.3.1) which people perceive.

In purely technical evaluation those aspects are represented by a single measure which we are trying to reach or even overcome. It has to be as close as possible to this perception. The quantitative comparison of the performance defined by evaluation metrics (§10.3.2) allows comparing dialogue system models (§8). For objectivity purposes the benchmark datasets (§10.3.3) has been established to enable dialogue models comparisons when trained on the same datasets.

### 10.3.1 Evaluation Aspects

The common evaluation introduction (§10.1.2) presented in the previous section can be turn to the particular evaluation aspects which are later presented as exactly measured activities via metrics (§10.3.2) and benchmarks (§10.3.3).

One of the comprehensive publications on this topic, **Evaluating Quality of Chatbots and Intelligent Conversational Agents** [81], extracted quality attributes from 32 papers and ten articles. They found they can be aligned into three main groups: efficiency, effectiveness, and satisfaction:

- **Efficiency**
  - **Performance**. It expects to avoid inappropriate utterances, robustness against manipulation and unexpected input.
- **Effectiveness**
  - **Functionality** Linguistic accuracy (syntactically and semantically correct sentences) is the criteria of the correct functionality next to the execution of the requested task and easy to use.
  - **Humanity**. Ambiguity about the Turing Test (§10.2.1) is one of the criteria as well as dialogue system transparency and disclose identity. Amongst many others criteria natural interaction with the ability to respond to the input and maintain themed discussion belong.

- **Satisfaction**
  - **Affect**. The conversation with the dialogue system can be interesting and provide a fun. It could help to make conversation active with appropriate mood and tone.
  - **Ethics and Behavior**. The dialogue system should, during the conversation, act ethically (§6.14) and with cultural knowledge. It is necessary to secure the conversation. Moreover, the privacy of the user needs to be taken into account.
  - **Accessibility**. Meaning or intent detection (§6.8.2) is a natural attribute of dialogue system behavior.

The specific examples of mapping the dialogue system (respective Natural Language Generation (NLG)) to evaluation aspects, we can find even earlier in [183]. This work considers three evaluation aspects:

- **Fluency** — Linguistic fluency (syntactically and semantically correct sentences).
- **Adequacy** — Correct meaning.
- **Readability** — Fluency in the dialogue context.

Each of those are evaluated by the metrics like Simple String Accuracy (SSA) [497], National Institute of Standards and Technology (NIST) [498], BiLingual Evaluation Understudy (BLEU) [82], F-measure and Latent Semantic Analysis (LSA) [266] described later in the evaluation metrics (§10.3.2) section.

### 10.3.2 Evaluation Metrics

Quantitative evaluation of the performance of any model can be done by plenty of metrics. Well-known and broadly accepted [489], [499] metrics for comparing parallel corpora (respective comparing candidate and reference responses) are usually used for deep learning models.

- **Word Overlap-based Metrics** — It evaluates the amount of word-overlap between the candidate and the reference response.
  - **BiLingual Evaluation Understudy (BLEU))** [82] — It works with n-grams and shows similarity of candidate and reference response between 0 and 1, where 1 represents identical responses.

    BLEU has been shown to correlate well with human judgment on the response generation task [500], [501].
  - **National Institute of Standards and Technology (NIST))** [498] — It improves the BLEU metric to consider the response structure and how informative the particular n-gram is. The rarer the n-gram occurrence is, the higher weight it gets.

  - **Metric for Evaluation of Translation with Explicit ORdering (METEOR)** [502] — It scores the text similarity based on the explicit word-to-word (unigram) matches and calculates the harmonic mean of unigram precision and recall, with recall weighted higher than precision.

    It significantly outperforms the more commonly used BLEU metric [503].
  - **Recall-Oriented Understudy for Gisting Evaluation (ROUGE)** [504] — It is a package of four different measures: ROUGE-N (overlap of n-grams between the candidate and reference responses) [505], ROUGE-L (identifies longest co-occurring in sequence n-grams between the candidate and reference responses), ROUGE-W is weighted ROUGE-L, and ROUGE-S (Skip-bigram based co-occurrence statistics)[506]

- **Embedding-based Metrics** — It evaluates the candidate and reference responses with the measure of cosine distance with the consideration that a vector is assigned to the meaning of each word, i.e., word embedding (§8.2.2), for instance Word2Vec [275].

  - **Greedy Matching** [499], [507] — It does not compute sentence-level embeddings. It greedily matches one token from the first sequence with another token from another sequence based on the cosine similarity of their word embeddings. The total score is then averaged across all words. It favors candidate responses with key-words that are semantically similar to those in the reference responses.
  - **Embedding Average** [499], [508] — It calculates sentence-level embeddings by computing the meaning of phrases by averaging the vector representations of their constituent words. For comparison of reference response and candidate response the cosine similarity between their respective sentence level embeddings is computed.
  - **Vector Extrema** [499], [509] — It calculates the sentence-level embeddings measure. For each dimension of the word vectors it takes the most extreme value amongst all word vectors in the sentence, and uses that value in the sentence-level embedding. The similarity between candidate and reference response vectors is again computed using cosine distance.

- **Other Metrics** - Various metrics without any particular classification into the above two groups are described below.

  - **Simple String Accuracy (SSA)** [497] — It is another NIST metric that scores the candidate response by counting the number of operations (word substitutions, insertions, and deletions) for conversion the reference to candidate responses divided by the length of candidate response.
  - **Latent Semantic Analysis (LSA)** [266] — This metric computes the semantic similarity of reference and candidate responses based on the measurement the semantic similarities of the words in the compared texts.

- **Lexical diversity (distinct-n)** [510] — It represents one of the aspects of lexical richness. Lexical diversity is quantitatively calculated using numerous different measures. In this case, the metrics is calculated as a count of different unique n-grams in the reference response to the total number of words (generated tokens) in the candidate response.
- **Average Response Length** [511], [512] — The length of an utterance (§6.8.1) is an objective metrics that reflects the substance of a candidate response.
- **Entropy** [511], [512] — It represents another objective metrics, which shows the serendipity of a candidate response by measuring the amount of information contained in the utterance (§6.8.1).
- **Response Perplexity** [70], [346] — Perplexity is an indicator of the model capability to account for the syntactic structure of the dialogue (e.g., turn-taking) and the syntactic structure of each utterance (e.g., punctuation marks). Lower perplexity is an indicator of a better model.
- **Word Error Rate (WER)** [346] — It is also known as a word classification error. It is defined as the number of words in the dataset that the model has mispredicted. Furthermore, it is divided by the total number of words in the dataset.
- **Automatic Dialogue Evaluation Model (ADEM)** [513] was presented by Lowe as the replacement of **Word Overlap-based Metrics** like BLEU. It captures semantic similarity to overcame word overlap measures and exploits the context and the reference response to calculate the score for the model response.
- **Conversation-turns Per Session (CPS)** [10], [514] is the metrics for social chatbots sufficient to measure the success of long-term, emotional engagement with users. It is the average number of conversation-turns between the chatbot and the user in a conversational session. The larger the CPS is, the better engaged the social chatbot is.
- **Sensibleness and Specificity Average (SSA)** [12] SSA combines two fundamental aspects of a human-like chatbots: making sense and being specific. It is a human evaluation metrics received from the human judges who label every model response on these two criteria.

  The correlation was $R^2 = 0.93$ for static sensibleness vs perplexity and $R^2 = 0.94$ for static specificity vs perplexity. It is taken as indication that it might be a good automatic metrics for measuring sensibleness and specificity. Overall static SSA vs perplexity has $R^2 = 0.94$.

The work **How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation** [499] criticizes some of the metrics above. It presents the evidence why not to use them and by what metrics (also above) to replace those existing ones taken broadly as the standards. There is still no official standardization about the dialogue system evaluation yet, so the golden standard of broadly used metrics is what is used now.

### 10.3.3 Evaluation Benchmarks

The standardization is always good approach to highlight the best datasets, methods or processes. From this perspective it is necessary for any Natural Language Processing (NLP) technique to have common standard datasets to perform various universal benchmarks (§10.3.3).

Additionally, it is also good to test and evaluate specific modules of dialogue system pipeline architecture (6.3.1). It inherits specific datasets for performing Natural Language Understanding (NLU) benchmarks (§10.3.3), Dialogue Management (DM) benchmarks (§10.3.3), and Natural Language Generation (NLG) benchmarks (§10.3.3). Those module specific datasets focus on the specific test requirements in connection with particular NLP module functionality.

#### Universal Benchmarks

Various generative models can be compared only when there is a standardized benchmark over normalized data sources. It does not matter which model (language model, predictive model, classification model, or other) we talk about.

Amongst many available datasources there are few datasets which are considered to be the best ones for universal benchmark evaluation:

- **Stanford Question Answering Dataset (SQuAD)** [83], [84]. SQuAD was already presented in the corpora introduction chapter (§7). It is a single reading comprehension dataset used as the benchmark dataset for comparison of various NLP models. The leaderboard with comparisons is available online[5].

- **ReAding Comprehension Examinations (RACE)** [212] was already introduced in the chapter about corpora (§7). The RACE is a single dataset used for various NLP models comparisons with a publicly available leaderboard[6].

- **General Language Understanding Evaluation (GLUE)** [85] represents a collection of tools for evaluating performance of models across a diverse set of existing NLU tasks. The evaluation is done through eleven various datasets and corresponding metrics, for instance, sentiment via The Stanford Sentiment Treebank [128], Questions Natural Language Inference (NLI) via Stanford Question Answering Dataset (SQuAD). The GLUE benchmark leaderboard is presented online[7].

- **Super General Language Understanding Evaluation (SuperGLUE)** [86] Due to notable progress across many Natural Language Processing (NLP) tasks SuperGLUE has been established. The collection reflects the NLP evolution. The new corresponding tools evaluate the performance of models, for instance Multi-Sentence Reading Comprehension [515] or Words in Context [516] and eight others. The leader-board can be found online[8].

[5] https://rajpurkar.github.io/SQuAD-explorer/

[6] http://www.qizhexie.com/data/RACE_leaderboard.html

[7] https://gluebenchmark.com/leaderboard

[8] https://super.gluebenchmark.com/leaderboard

These single (SQuAD, RACE) or sets (GLUE, SuperGLUE) of datasets represent the evolution of benchmark tasks. This corresponds to the rapid development of NLP and its constantly changing requirements to the evaluation.

### Natural Language Understanding (NLU) Benchmarks

The Natural Language Understanding (NLU) component benchmark compares the recognition of entities (§6.8.3) (Named Entity Recognition (NER)) and user intent (§6.8.2) in the input utterance (§6.8.1).

For such purposes, we need to have datasets including multiple domains, various (highly diversified) intents with entities representing many entity types.

All of the benchmarks focus on the particular NLU online using cloud services like Microsoft LUIS, Google DialogFlow, IBM Watson or NLU libraries like RASA.

- **NLU Evaluation Corpora**[9]. It is the mixture of three datasets (Ask Ubuntu Corpus, Web Applications Corpus, Chatbot Corpus), i.e., three domains with overall 450 questions and answers with identified 15 intents and 11 entity types. It does the benchmark for the RASA library, and except the DialogFlow, Watson and LUIS adds Facebook Wit.ai and Amazon Lex [517]. It compares precision, recall, and F1 measures for intent and also for entity types recognition per particular system.

- **NLU Benchmark**[10] is a benchmark performed on the previous **NLU Evaluation Corpora** and includes previously tested services plus the Snips library [197]. The authors of the Snips library performed the benchmark, so there can be potential bias.

- **NLU Evaluation Data**[11] is a large NLU dataset containing real user data collected with Amazon Mechanical Turk (AMT). It covers 21 domains with 64 intents and 54 entity types. It compares RASA, Google DialogFlow, IBM Watson, and Microsoft LUIS services and libraries [518] and provides precision, recall, and F1 measures for intent and entity types recognition.

### Dialogue Management (DM) Benchmarks

Since beginning The Conversational Intelligence Challenge (ConvAI) (§2.4.4) focuses on standardizing chatbot models evaluation. It includes human evaluation (for instance Turing test (§10.2.1)) followed then by computerized evaluation (for example measured by metrics (§10.3.2)).

For this purpose the collected ConvAI persona-chat dataset [70] which is already presented in the chapter about corpora (§7) should help to deal with common chatbot model issues which include:

- Missing consistency in the chatbot personality [519] because the training datasets contain dialogues from various speakers.

[9] https://github.com/sebischair/NLU-Evaluation-Corpora

[10] https://github.com/snipsco/nlu-benchmark

[11] https://github.com/xliuhw/NLU-Evaluation-Data

- The chatbot training on the recent dialogue history [76] to produce the utterance causes the lack of explicit long-term memory.
- Tendency to produce *I do not know* answers [520].

The ConvAI competition works towards to find the models that address these specific issues. The results of ConvAI indicate that there is a promise to make progress in this activity.

#### Natural Language Generation (NLG) Benchmarks

When the Natural Language Generation (NLG) benchmark is going to be done, there are several ways to do so. The evaluation of text generation has much more freedom to choose the task which can be used for the baseline.

The standard way the text is generated is using a computer understandable form. Another approach is to generate captions for images.

- **E2E Dataset**[12]. The dataset is released as open and it is a part of E2E NLG Challenge[13] [25]. It contains crowdsourced data of 50k instances in the restaurant domain. The benchmark [521] was openly realized against another datasets like BAGEL [446], and SF Hotels/Restaurants [452] with defined metrics[14] including BLEU [82], NIST [498], METEOR [502], ROUGE-L [504], and CIDEr [522].
- **Microsoft COCO Caption**[15] [523] is an image caption dataset with over 1.5 million captions describing over 330 000 images. The generated captions are evaluated using several popular metrics, like BLEU, METEOR, ROUGE and CIDEr. The benchmark is available through the evaluation server[16] [524].

### 10.3.4 Evaluation Explainability

The question about transparency and explainability of Artificial Neural Network (ANN) pops much often these days and resonates with abusive or harmful exceptional behavior of systems based on Artificial Intelligence (AI). Thus more and more research groups and companies are trying to explain AI behavior, and for such effort the technical term Explainable AI (XAI) [525], [526] has been established.

#### Common Explainability

In the standard way, an explanation of any Machine Learning (ML) or Artificial Intelligence (AI) model is a complicated task. Tools, libraries, and methods used for model explanation present results to users mostly in a human-readable and easily understandable graphical way. Some algorithms of models explanation even exceed the original methods implemented, so their code served for original method improvement.

[12] https://github.com/tuetschek/e2e-dataset

[13] http://www.macs.hw.ac.uk/InteractionLab/E2E/

[14] https://github.com/tuetschek/e2e-metrics

[15] http://cocodataset.org

[16] http://cocodataset.org/#captions-leaderboard

- **Local Interpretable Model-agnostic Explanations (LIME)**[17] [527] is a technique that explains predictions of any classifier in an interpretable and faithful manner, by learning an interpretable model locally around the prediction. It is flexible to explain different models, for instance, the random forest used for text and ANN for image classification.

- **SHapley Additive exPlanations (SHAP)**[18] [528]. The SHAP approach interprets predictions from tree ensemble methods (gradient boosting, random forests). The game theory is applied. The interpretation is made through the visualization of individual feature attributions. The study showed better agreement the visualization corresponds to human intuition over the classic attribution summaries and partial dependence plots.

- **AIX360**[19] [529]. The AI Explainability 360 is not a particular explanation technique but an open-source toolkit consisting of diverse state-of-the-art explainability methods: ProtoDash [530], Disentangled Inferred Prior VAE [531], Contrastive Explanations Method [532], Contrastive Explanations Method with Monotonic Attribute Functions [533], LIME [527], SHAP [528], TED [534], Boolean Decision Rules via Column Generation [535], Generalized Linear Rule Models [536], and ProfWeight [532]. Next to the explainability it provides also two evaluation metrics: Faithfulness [537] and Monotonicity [533].

### Dialogue System Explainability

The same question about transparency and explainability raises for the dialogue system. With more and more complex models which step into the dialogue system design and especially when the dialogue system is built on top of the generative model (§6.4.2) it is necessary to know what is happening under the hood. We would like to know why and how the dialogue system model generating responses based on human requests works.

The retrieval-based model (§6.4.1) known as rule-based model is self explanatory. The questions-answers pairs or combinations are strictly given and the reason for particular response based on the request is simply possible to review from the rule definition source.

The generative based model (§6.4.2) also known as corpus based model consists of several layers of design which include for instance embedding (§8.7) on which the Sequence to Sequence (Seq2Seq) architecture (§8.2.3) various transformer architectures are built. All these parts can be taken into account when explainability comes to the discussion.

There are just a few tools or publications related to the dialogue system explanation since the field is mostly focusing on the common AI explainability or explainability of the particular ANN.

For instance, the explanation and visualization of the embedding based models can be done with exBERT[20] tool [538]. It serves for the explanation of transformers based models

[17] https://github.com/marcotcr/lime

[18] https://github.com/slundberg/shap

[19] https://github.com/IBM/AIX360

[20] http://exbert.net

with a particular focus on the Bidirectional Encoder Representations from Transformers (BERT) (§8.2.3).

Another example is the paper [539], which focuses on the transparency of chatbots implemented for recruitment.

An explainable, transparent, and auditable dialogue system or respectively AI itself is not a simple task. Thus we need to look forward to more relevant research and practical business applications which narrow down the rules on how the generative model (§6.4.2) based dialogue system models can/cannot work.

## 10.4 Intervention Evaluation

Dialogue system evaluation and testing is just one part of functionality testing, the technical one. If the dialogue system based on the influencing data provides the intervention method like in this case, the emotion regulation, it has to be evaluated as well. The complexity of such a task lies in the interaction with people.

So, the more complex evaluation will be used the more problematic would be to make conclusions from results. For such purposes, clinical psychology and psychiatry developed various questionnaires (§10.4.1) which would be present in the next section. Moreover, the standardization and reliability (§10.4.2) of such evaluation next to the simplicity are crucial.

### 10.4.1 Questionnaires

To determine weather particular treatment technique works or for comparison (see A/B testing §10.2.2) the following diagnostic systems and rating scales for various problems are used in clinical psychology and psychiatry.

- Depression, anxiety & stress
    - **Hamilton Rating Scale for Depression (HRSD)** [540], [541]. Multi-item (original version contained 17 items) questionnaire, which helps with an indication of depression, and as a guide to evaluating recovery. The questionnaire is designed for adults and is used to rate the severity of their depression by probing mood, feelings of guilt, suicide ideas, insomnia, agitation or retardation, anxiety, weight loss, and somatic symptoms.
    - **Patient Health Questionnaire (PHQ)** [542], [543] is family (PHQ-2, PHQ-4, PHQ-8, PHQ-9 and PHQ-15 and also GAD-7) of multiple-choice self-report questionnaires which are used as screening and diagnostic tools for mental health disorders, such as depression, anxiety, alcohol, eating, and somatoform disorders. Answers to the questions are evaluated by the same four categories described in GAD-7.
    - **Generalized Anxiety Disorder 7 (GAD-7)** [544]. The questionnaire was designed to do self-reported screening and severity measurement of generalized anxiety disorder. It contains seven questions answered by four categories

with assigned points. The total score of answers sum up together gives us an assessment indication.

  - **Perceived Stress Scale (PSS-10)** [545], [546]. A questionnaire with 14, 10, or 4 items was developed to measure psychological stress. This test has become the most widely used as the psychological instrument for measuring nonspecific perceived stress in studies assessing the stressfulness of situations, the effectiveness of stress-reducing interventions, and the extent to which there are associations between psychological stress and psychiatric and physical disorders. The higher perceived stress levels correspond to the higher PSS score, which tends to increase the risk of diseases.
  - **The Depression Anxiety Stress Scales (DASS)** [547]. The questionnaire contains a set of three self-report scales designed to measure the negative emotional states of depression, anxiety, and stress. Each of the three DASS scales contains 14 items. Subjects are asked to use 4-point severity/frequency scales to rate the extent to which they have experienced each state over the past week. Scores for depression, anxiety, and stress are calculated by summing the scores for the relevant items.

- Well-being
  - **Flourishing Scale (FS-8)** [546], [548]. A brief 8 item summary measure of the respondent's self-perceived success. It analyzes the topics which include relationships, self-esteem, purpose, and optimism. The summary corresponds to a single psychological well-being score.
  - **Satisfaction With Life Scale (SWLS-5)** [546], [549]. This questionnaire is built as a small 5-item one. It is designed to measure global cognitive judgments of satisfaction with participated person life. The cut-off scores are calculated from the questions, and when the higher the score is, then life satisfaction is better.
  - **Scale of Positive and Negative Experience (SPANE)** [548]. The 12 item questionnaire is divided into two parts by six items. These two parts assess positive/negative feelings. The positive and negative items contain three general items and three more specific (e.g., joyful, sad).
  - **Positive and Negative Affect Schedule (PANAS)** [6], [550]. A questionnaire with two 10 item scales to measure both positive and negative affect. Each item is rated on a 5-point scale similar to GAD-7.

    The researchers extracted 60 terms from the factor analyses of Zevon and Tellegen [551] shown to be relatively accurate markers of either positive or negative affect, but not both. The researchers arrived at ten terms for each of the two scales, as follows. Positive affect is presented by terms like attentive, active, alert, excited, enthusiastic, determined, inspired, proud, interested, strong. Negative affect is presented by terms hostile, irritable, ashamed, guilty, distressed, upset, scared, afraid, jittery, nervous.

  - **The Discrete Emotions Questionnaire (DEQ)** [552]. The DEQ is presented as a new tool for measuring state self-reported emotions. It focuses on eight distinct state emotions: anger, disgust, fear, anxiety, sadness, happiness, relaxation, and desire, which are evaluated by the participant on a 7-point scale (1 = Not at all and 7 = An extreme amount).

### 10.4.2 Reliability and Validity

Questionnaire reliability and validity as the tool used in clinical practice is criticized and contradicted over the years of existence. As any human activity also assessment of this part of human feelings and actions is evolving. New and new questionnaires are created by renowned psychological and psychiatric departments to improve the existing ones and by comparison, under the various clinical experiments proof better results and relevance to particular use and findings.

If we take a look at the specific questionnaires, for example, The Hamilton Rating Scale for Depression (HRSD) [540], [541] which has been considered as the golden standard amongst questionnaires [553]. However, it has been criticized for use in psychological clinical practice [554] because it represents the questionnaire, which is more oriented towards identification emphasis on insomnia than on feelings of hopelessness, self-destructive thoughts, suicidal cognitions, and actions. Also, the author claimed that his scale should not be used as a diagnostic instrument [555].

On the other hand, well-being tests are considered to be reliable and valid instruments in the assessment of positive and negative affects in clinical and non-clinical studies [550], [556].

## 10.5 Conclusion! What Is Tested May Never Fail

Testing any software is a complex discipline. With the growing complexity of software systems requirements for testing grow as well. The dialogue system, no matter what approach is used, represents one of the most complex systems. Based on this and on the overview of the previous chapters, it is obvious that dialogue system testing and evaluation cannot be easily defined and performed.

Current methods focus on dialogue system testing (§10.2) from the perspective of usability, User Experience (UX) and Customer Experience (CX). Those are correct requirements whenever the chatbot is purely used for customer care or customer support in the task-oriented closed domains (§6.5.2). Whenever the dialogue system is used as the chatbot in an open domain, fluency, length and richness of the conversation are the main considered factors of success.

The second part of testing, evaluation (§10.3), is purely technical and dedicated to End-to-End (E2E) generative models (§6.4.2) evaluation.

The feedback from users related to dialogue system capabilities or influence is usually not measured at all. From this perspective psychological feedback questionnaires (§10.4.1) were considered to measure dialogue system intervention and emotion regulation success (§10.4).

There is still a potential for various improvements. Testing and evaluation methods are evolving as it is possible to see from the development of testing methods and criticism of metrics. The benchmark support and explanation of ML/AI models are going forward with the most current research.

Standard questionnaires inspired by psychological feedback questionnaires can be used for the feedback from users regarding dialogue system functionality and approach satisfaction.

# Chapter 11

# Research Proposal

With all the parts described in the previous chapters, we can finally come up with the research proposal.

It begins with authors' existing research (§11.1) where he was experimenting with wearable devices and soft and hard data and follows up with the collected two sets of data (§11.2), published in [124] and used also in diploma thesis [2].

Two vastly different use-cases are presented (§11.3): Emotion Regulation (§11.3.1) and Arm Rehabilitation (§11.3.2) which are built on top of the experience from the systematic review of existing research presented in the previous chapters.

The remaining sections define the dissertation thesis goal (§11.4) via the Research Objective (RO) (§11.4.1) and specify specific Research Questions (RQs) (§11.4.2). This work can be enclosed with an overall conclusion (§11.5).

## 11.1 Existing Research

At the beginning, there was an idea to utilize wearables which came at that time to the market together with social media and to gain the value that could be used in many applications by information fusion.

Wearables provided only steps measurements in 2014, so the first research was targeting this way [557]. However, except the experience how to design and process the experiment and what to expect about the data gained from the first manuscript, it is evident that this is the dead end.

With development on the market, more sophisticated devices have come; they provide more value and allow to measure the heart rate on top of other measurements. Heart rate is an independent value which could be broadly in various research projects and practical applications. It is also considered to have a better relation to human mood and behavior.

Two experiments followed and resulted in the open data publication [124] which can be considered rather unique from many aspects even though the idea of combination of such types of data was not original at all. The data was collected not only to describe and publish them but to use them for further research. The latest manuscript deals with data analysis and performs stress dichotomy identification [145] via data fusion.

### 11.1.1 Journal Papers

- Salamon J. and Mouček R. (2017). *"Heart rate and sentiment experimental data with common timeline"*. In Data in Brief, Volume 15, ISSN 2352-3409, pages 851-861. DOI: 10.1016/j.dib.2017.10.037.

### 11.1.2 Conference Papers

- Salamon J., Černá K. and Mouček R. (2018). *"Stress Dichotomy using Heart Rate and Tweet Sentiment."* In Proceedings of the 11th International Joint Conference on Biomedical Engineering Systems and Technologies, ISBN 978-989-758-281-3, pages 527-532. DOI: 10.5220/0006650105270532

- Salamon J. and Moucek R. (2016). *"Link between Sentiment and Human Activity Represented by Footsteps - Experiment Exploiting IoT Devices and Social Networks"*. In Proceedings of the 9th International Joint Conference on Biomedical Engineering Systems and Technologies, ISBN 978-989-758-170-0, pages 450-457. DOI: 10.5220/0005818204500457

## 11.2 Data Collection

When performing experiments which combine soft and hard data (§3) with the common timeline (§3.3) it was necessary either to find existing data or create a new data collection. Based on an extensive search, the decision was made to collect data as a part of the research.

Data collection went through two stages: Pilot experiment (PX) (§11.2.1) and Quasi-experiment (QX) (§11.2.2). Both collections are described in the sections below, the high level description of experiments can be found in §3.6, particularly for PX (§3.6.1) and QX (§3.6.2).

### 11.2.1 Data collected during Pilot experiment (PX)

The following sections describe the design, recruitment, and ethics related to the PX held by a single participant.

#### Design of Experiment

The experiment was designed to take two-time fifty days. During both of these periods Heart Rate (HR) and textual data were collected simultaneously.

- Heart rate data collection
  - Two different wearables (devices) were used, specifically Fitbit Charge HR and Basis Peak
  - The devices measured heart rate 24x7 except breaks for charging

  - The average output data of heart rate sampling frequency was higher or equal to one minute.

- Sentiment data collection
  - Sentiment was expressed in English using short texts (140 characters) - tweets.
  - The relevant sentiment at the time the tweet was written was expressed by a subject via a hashtag being a part of the tweet (#p for positive and #n for negative feeling).
  - Tweets were written each 45 minutes, i.e. a maximum of 21 tweets during a weekend (from 9 AM to midnight) and 23 tweets during a weekday (from 7:30 AM to midnight).
  - However, only 20 tweets per day were required.

### Recruitment

The participant was a 35-year healthy man with a treated high blood pressure.

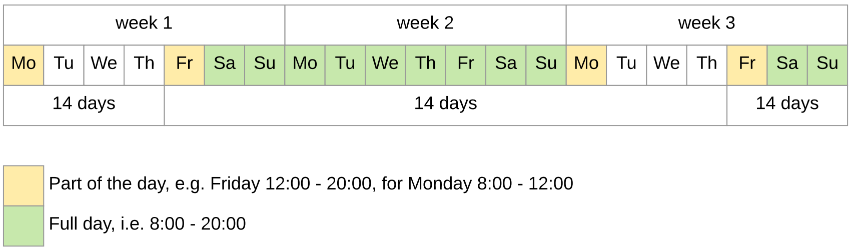
### Ethics

Considering a single person participating in the Pilot experiment (PX) and the fact that the subject was the author himself means that this trial meets all the ethical aspects; it is not necessary to have an explicitly written and signed consent.

## 11.2.2 Data collected during Quasi-experiment (QX)

Based on the experience with the PX the QX performed the next sections describe its design, recruitment, and ethics.

### Design of Experiment



| week 1 | | | | | | | week 2 | | | | | | | week 3 | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 🟨 Mo | Tu | We | Th | 🟨 Fr | 🟩 Sa | 🟩 Su | 🟩 Mo | 🟩 Tu | 🟩 We | 🟩 Th | 🟩 Fr | 🟩 Sa | 🟩 Su | 🟨 Mo | Tu | We | Th | 🟨 Fr | 🟩 Sa | 🟩 Su |
| 14 days | | | | 14 days | | | | | | | | | | | | | | 14 days | | |

🟨 Part of the day, e.g. Friday 12:00 - 20:00, for Monday 8:00 - 12:00

🟩 Full day, i.e. 8:00 - 20:00

Figure 11.1: Quasi-experiment time-frame

- Common part
  - The QX lasted 10 days within 14 days time-frame (Figure 11.1)

  - The QX was taken two-times in different time-frames to recruit as much subjects as possible

- Heart rate collection
  - Four wearables (devices) were used, specifically Fitbit Charge HR (the same device as in the PX)
  - Four technical accounts were attached to these devices (bodyinnumbers01-04)
  - The devices measured heart rate 24x7 except for breaks for charging (the time frame between 8 AM and 10 PM was minimally required)

- Sentiment collection
  - Sentiment was expressed in the Czech language through short texts (280[1] characters) - tweets
  - Four technical Twitter accounts were created (bodyinnumbers01-04)
  - The relevant sentiment at the time the tweet was written was recorded by a subject via hashtags directly in the tweet (#p for positive and #n for negative feeling)
  - The text recording was done every 60 minutes between 8:00 AM and 10 PM.
  - 15 tweets per day were required.

**Recruitment**

The recruitment was done among the students within the same study group:

- Seven healthy subjects [4 female; age: $\mu = 20$, $\sigma = 0.8$ and 3 male; age: $\mu = 23$, $\sigma = 1.7$] participated in this study.
- All participants were native Czech speakers

The following personal attributes for each subject were recorded: gender, age, weight, height, Twitter/Fitbit account, wearable serial number, and start date of the experiment.

**Ethics**

All QX participants were handled as anonymous with ID (101-104,201-203) consisting of the number of the experiment time-frame (1 or 2) and two digits number of the subject (01-04 and 01-03).

## 11.3 Research Use Cases (RUCs)

The thesis is written to propose Research Objective (RO). However, all the commonly designed approaches are necessary to validate by practical RUCs. Several RUCs are presented in the next sections. These are supported by their schematic overviews where orange signals are input (measured) signals and green signal is an output (feedback) signal.

[1] Twitter extended the limit of the tweet from 140 to 280 characters

### 11.3.1 Emotion Regulation (ER)

The original idea was to design and implement an emotion regulation use-case utilized in a chatbot. It stands on the three main pillars:

**Stress identification** through data fusion (§5)

**Design chatbot** which is possible to influence (§8)

**Regulate emotions** identified by stress used for chatbot influence (§9)

This all has to be done during the dialogue with the chatbot, which means to switch the context (§6.6) from chit-chat to Emotion Regulation whenever the negative stress is identified.

The following phases of this RUC correspond to the numbers in circles in Figure 11.2:

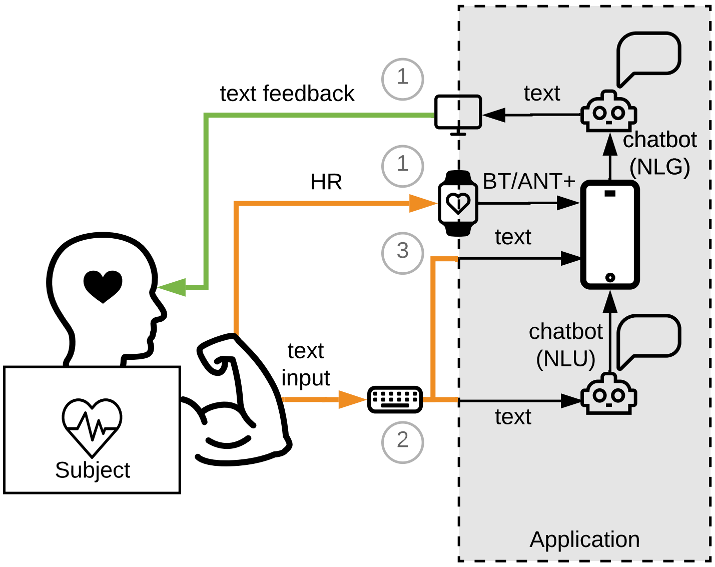


Figure 11.2: Research Use Case - Emotion Regulation

1. The wearable which measures the physiological signals provides HR as one of the stress identifiers. In the beginning, the decreasing and increasing HR itself or the value of HR at the particular break-point can be the trigger. The feedback on the stress identified is given by the chatbot using a Natural Language Generation (NLG) alternatively synthesized voice via Text to Speech (TTS). It provides a simple textual feedback, e.g. "take a breath, calm down, close your eyes and relax".

2. The pipeline can be completed by allowing a user to communicate with the chatbot via a textual input (Natural Language Understanding (NLU)). It provides a chance to enrich chatbot conversational complexity with its ability to keep the dialogue context. Still, it can be at the beginning simplified enough to cover a basic input: "I feel ok, I am ok, I am not under pressure" and some others.

3. Use the text content of the conversation as the add on to HR leads to the complex influencing data via the signal fusion and better stress identification.

   Correct identification of stress would lead to better application results in counseling during the emotion regulation process.

Potential issues:

- Based on the experience with the Pilot experiment (PX) data processing during the diploma thesis [2] we have learned that the sentiment extracted from the conversation could be inconclusive because it is mostly neutral. Such sentiment value then does not help to identify stress.
- The excitement, which can lead to an increase of Heart Rate (HR), can be relatively stable (either mild or unrecognizable) during the conversation with a chatbot. Hence the HR trend (§5.3.1) can be identified problematically or not at all.

### 11.3.2 Arm Rehabilitation

Another idea is to design and implement a rehabilitation system which utilizes a chatbot influenced by various data. The application should work as a rehabilitation tool in case of difficulties with shoulder momentum or muscle problems.

There are two alternatives presented in the next sections. The first one is built on top of the sensors; these provide a signal turned into data influencing the chatbot. The second one works with pictures (respective video) as sources of data influencing the chatbot.

**Signal Influenced Arm Rehabilitation**

Figure 11.3 illustrates the phases (represented by the numbered circles) of this particular Research Use Case (RUC).

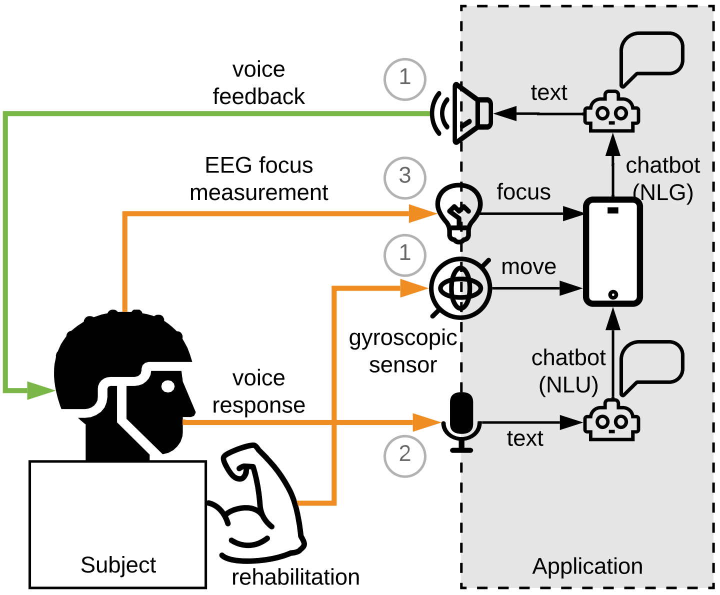


Figure 11.3: Research Use Case - Arm Rehabilitation with gyroscopic sensor

1. Movements of the arm during the rehabilitation are measured with the attached gyroscopic sensor. The feedback on the process and corrections of the rehabilitation are provided by the chatbot (Natural Language Generation (NLG)) influenced by the gyroscopic signal. The chatbot interaction is implemented by a synthesized voice via Text to Speech (TTS). A simple output voice feedback could be the following: "make it slower, do it more precise, move the arm a bit left/right/down/up".

2. To build the complete pipeline the enhancement of the chatbot (Natural Language Understanding (NLU)) to the voice response given by a user through Automatic Speech Recognition (ASR) is implemented. Mostly the pure users' voice commands as additional inputs are considered, e.g. "it is ok, it hurts, it helps, I am fresh, I am tired". Completion of the chatbot pipeline allows enriched conversation with a user with a single signal influence from phase 1.

3. The rehabilitation process is enhanced by the measurement of the Electroencephalogram (EEG) signal. The right focus on the procedure is significant to perform a good rehabilitation exercise of the arm.

   The combination of two signals, e.g. previous gyroscopic data and focus data can be done by a simple combination of the data sources or smart data fusion. It depends on which approach leads to better application results.

Potential issues:

- The voice response (phase 2) can have a negative influence on measured EEG data (phase 3).

### Video Influenced Arm Rehabilitation

In this case (Figure 11.4) we consider the following phases (the numbered circles) of RUC.

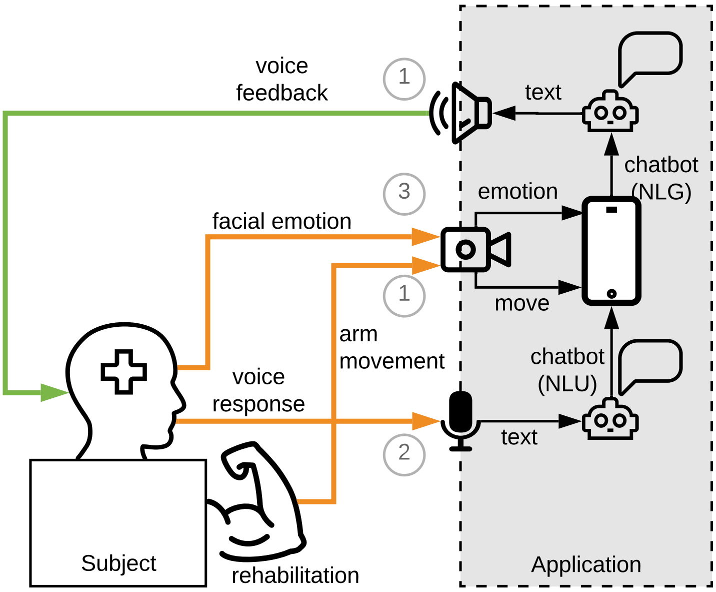


Figure 11.4: Research Use Case - Arm Rehabilitation with camera

1. A camera takes a video in which movements of the arm during the rehabilitation are identified. The feedback on the process and corrections of rehabilitation is provided by a chatbot (Natural Language Generation (NLG)). It is influenced by movements detected in the video and synthesizes its answer to the voice via Text to Speech (TTS). A simple output voice feedback can be given: ”make it slower, do it more precise, move the arm a bit left/right/down/up”.

2. To build the complete pipeline the enhancement of chatbot (Natural Language Understanding (NLU)) to the voice response given by a user through Automatic Speech Recognition (ASR) is implemented. Again the same or similar pure voice commands of the user (additional inputs) are processed: ”it is ok, it hurts, it helps, I am fresh, I am tired” and so on. The complete chatbot pipeline allows us to keep the conversation fluent when it is influenced by a single data source from phase 1.

3. The previous rehabilitation process is enhanced by the measurement of emotions (identification of facial emotions) from the video. The correct identification of emotions is essential to perform good rehabilitation exercise related to the particular part of the arm. In this case, we might return to the basics seven emotions [558]: anger, disgust, fear, happiness (joy), sadness, surprise, and contempt — alternatively, some more suitable facial emotion schema based on the relevant studies can be used.

   The second influencing signal (emotions) leads to a more complex influencing approach which might use any data fusion or simple combination of data.

Potential issues:

- The voice response (phase 2) can have a negative influence on emotion recognition from the video (phase 3).

- The movements of the arm (phase 1) and emotion recognition (phase 3) identified both from the video can interfere if taken by the same camera, so most likely two independent cameras are needed.

**Notes**

- For Text to Speech (TTS) and Automatic Speech Recognition (ASR) Alexa, Google Home, or another assistant can be used. It allows accessing its functionality via API. These personal assistants recognize the voice, offer text hypothesis and also can do the TTS synthesis.

## 11.4 Goals of the Thesis

The goals of the thesis are defined through one Research Objective, which is then covered by several Research Questions.

### 11.4.1 Research Objective (RO)

Based on the topic described in the introduction and author's publication the author of this thesis sets the following RO that identifies (or may not) the method or process which improves a chatbot as a digital coach:

*RO: To propose and validate a method to influence a chatbot (or its part) by external data to achieve the change in its conversational behavior. For such data the measured signal from the external device(s) is used alone or in combination with the conversation content itself.*

### 11.4.2 Research Questions (RQs)

Research Questions (RQs) help to identify, split, narrow down, and organize the main Research Objective (RO) in smaller parts. The particular parts of the thesis can cover such RQs.

- RQ1: Is there a way to influence a chatbot with external data?
- RQ2: What type of chatbot concept is possible to use?
- RQ3: What corpus (corpora) or model(s) need to be used to build a chatbot?
- RQ4: What source(s) or approach(es) could be used to collect data (or its fusion) suitable for influencing a chatbot?
- RQ5: How to identify, test and explain such a change of behavior of particular chatbot implementation?

## 11.5 Conclusion

This thesis contains several topics that form the whole. Each of these topics seems to be sophisticated enough to become a separate research project.

All of these together with well-described particular details and approaches serve as decent rudiments for future research, realistic outcomes and forthcoming dissertation thesis in the field where the dialogue systems (chatbots) serve as the psycho-social intervention tools.

All the methods related to a replacement or support of ambulatory treatment (§2.1) have common aspect. They expect the consumer or patient to start to use them when he or she needs some help. Most of the people who need the treatment do not admit or may not realize they need some.

On the other hand, many people like to have a private conversation with a chatbot [7] and even share intimate details with it. So potentially, chatbots dedicated for health respectively well-being (§2.4.6) are the right choice to motivate people to use them even they think they do not need them.

However, many approaches to achieve the Research Objective (RO) (§11.4.1) and get the answers to the Research Questions (RQs) (§11.4.2) seem to be promising.

To avoid potential non-determinism of stress identification by inconclusive sentiment and mild or unrecognizable trend of Heart Rate (HR) (described as potential problems in emotion regulation Research Use Case (RUC) (§11.3.1)), another RUC was proposed to overtake the role of the validation use case, the arm rehabilitation (§11.3.2).

This use case includes two variants, one driven by information given by signals, another one driven by the information extracted from a video. Both of them, despite potential problems described in the section above (§11.3.2), present the deterministically defined control over the influencing signal. This signal can be used for the dialogue system influencing in the potential practical application.

# Appendix A

# Practical Experience

During writing this thesis the author reviewed many materials and found many applications and implementations related to dialogue systems. So, the most exciting experience is briefly recorded here.

The first topic is a practical experience with two existing chatbots for health or well-being (§A.1). The second section (§A.2) lists the chatbot implementations; it includes the papers turned into a code by authors or someone else and also Github repositories which provide examples of particular libraries or dialogue languages used for chatbot implementation.

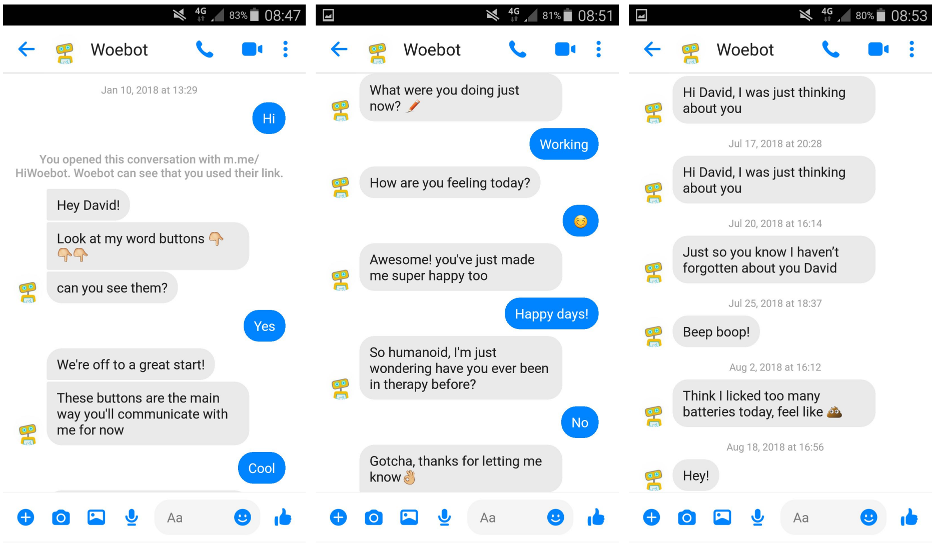
## A.1 Chatbots Experience

During several weeks two of the previously mentioned chatbots for health or well-being (§2.4.6) were tested by the author. He has been interested in how state-of-the-art applications with the interaction based on the Cognitive Behavioral Therapy (CBT) work.

### A.1.1 Woebot Chatbot

Woebot [6] is a psycho-social intervention chatbot (Figure A.1). It uses various standardized questionnaires (§10.4.1), which utilize CBT (Figure A.2) to treat young adults with symptoms of depression and anxiety.

The following screenshots show its occasional use and interaction:



08:47 Woebot

Jan 10, 2018 at 13:29

Hi

You opened this conversation with m.me/HiWoebot. Woebot can see that you used their link.

Hey David!

Look at my word buttons 👇👇👇

can you see them?

Yes

We're off to a great start!

These buttons are the main way you'll communicate with me for now

Cool

08:51 Woebot

What were you doing just now? 🖍

Working

How are you feeling today?

😊

Awesome! you've just made me super happy too

Happy days!

So humanoid, I'm just wondering have you ever been in therapy before?

No

Gotcha, thanks for letting me know👌

08:53 Woebot

Hi David, I was just thinking about you

Jul 17, 2018 at 20:28

Hi David, I was just thinking about you

Jul 20, 2018 at 16:14

Just so you know I haven't forgotten about you David

Jul 25, 2018 at 18:37

Beep boop!

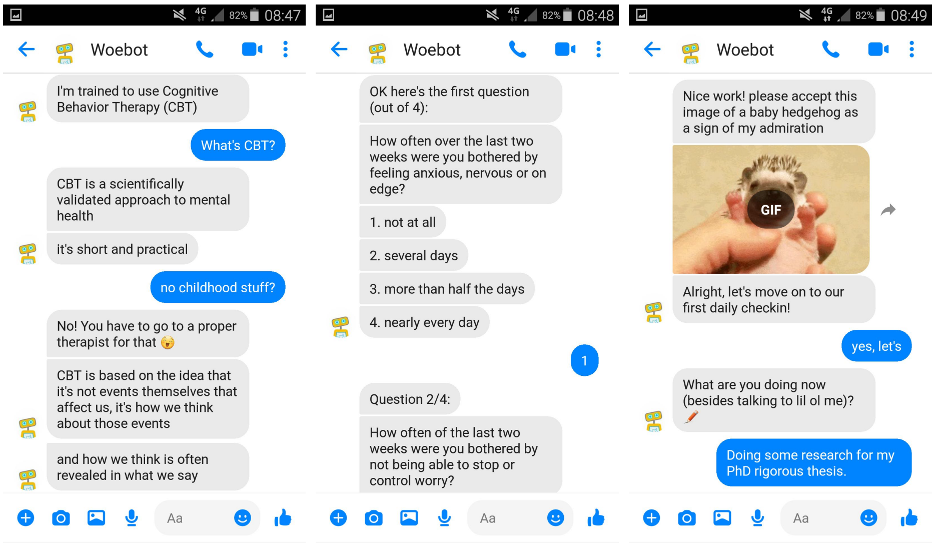
Aug 2, 2018 at 16:12

Think I licked too many batteries today, feel like 💩

Aug 18, 2018 at 16:56

Hey!

Figure A.1: Woebot - a) chatbot introduction, b) introduction to therapy, and c) reminder to the user



08:47 Woebot

I'm trained to use Cognitive Behavior Therapy (CBT)

What's CBT?

CBT is a scientifically validated approach to mental health

it's short and practical

no childhood stuff?

No! You have to go to a proper therapist for that 😯

CBT is based on the idea that it's not events themselves that affect us, it's how we think about those events

and how we think is often revealed in what we say

08:48 Woebot

OK here's the first question (out of 4):

How often over the last two weeks were you bothered by feeling anxious, nervous or on edge?

1. not at all

2. several days

3. more than half the days

4. nearly every day

1

Question 2/4:

How often of the last two weeks were you bothered by not being able to stop or control worry?

08:49 Woebot

Nice work! please accept this image of a baby hedgehog as a sign of my admiration

Alright, let's move on to our first daily checkin!

yes, let's

What are you doing now (besides talking to lil ol me)? 🖍

Doing some research for my PhD rigorous thesis.

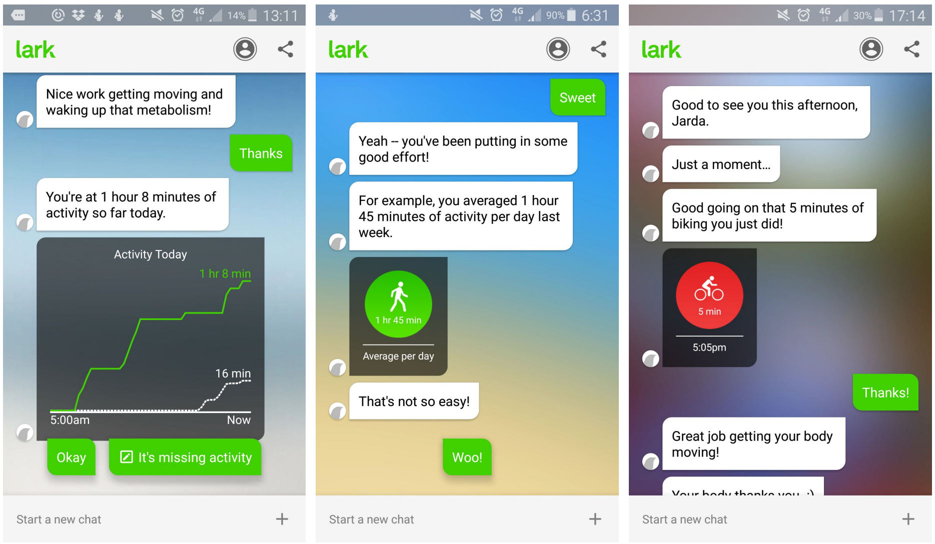
Figure A.2: Woebot - a) introduction to CBT, b) questionnaire, c) treatment reward

The Woebot is trying to be funny sometimes in a silly way. It is definitely that kind of chatbot personality (§8.8.1) that imitates a friendly entertaining buddy. Anyway, it leads the user to the point to give him/her required treatment and subsequent reward.

### A.1.2 Lark Chatbot

Lark [80] is a chatbot which tracks daily movement (Figure A.3), weight (once a week), sleep and food (Figure A.4). It is an AI based chatbot that incorporates interactive elements of Cognitive Behavioral Therapy (CBT).

The following figures show screenshots taken during several weeks of interaction:



Nice work getting moving and waking up that metabolism!

Thanks

You're at 1 hour 8 minutes of activity so far today.

Okay

It's missing activity

Sweet

Yeah – you've been putting in some good effort!

For example, you averaged 1 hour 45 minutes of activity per day last week.

That's not so easy!

Woo!

Good to see you this afternoon, Jarda.

Just a moment...

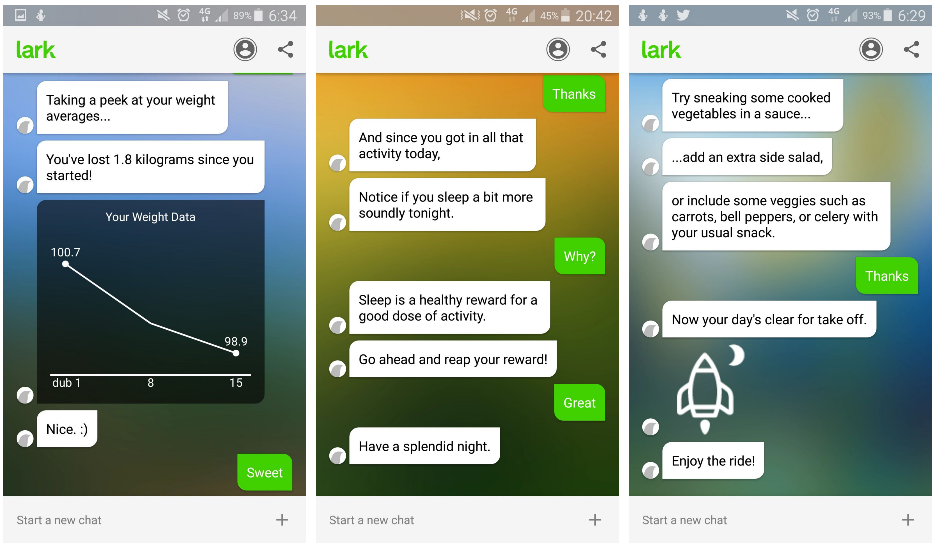
Good going on that 5 minutes of biking you just did!

Thanks!

Great job getting your body moving!

Your body thanks you :)

Figure A.3: Lark - a) activity tracking comparison, b) tracking the walk activity, c) tracking the bike activity



Taking a peek at your weight averages...

You've lost 1.8 kilograms since you started!

Nice. :)

Sweet

Thanks

And since you got in all that activity today,

Notice if you sleep a bit more soundly tonight.

Why?

Sleep is a healthy reward for a good dose of activity.

Go ahead and reap your reward!

Great

Have a splendid night.

Try sneaking some cooked vegetables in a sauce...

...add an extra side salad,

or include some veggies such as carrots, bell peppers, or celery with your usual snack.

Thanks

Now your day's clear for take off.

Enjoy the ride!

Figure A.4: Lark - a) tracking the weight, b) providing the advises related to sleep and c) giving the advises related to food

Compare to the Woebot Lark is much more rigid in communication. It definitely targets different and more mature audience with the functionality different from the Woebot. The talk here is explanatory, sometimes it looks like a discussion of a teacher with a student that contains a lot of annoying notes.

## A.2 Dialogue System Implementations

Modern research should be transparent and open to allow reproduction of results. This can be achieved by openly provided data if not these are publicly available yet and also an open code to reproduce the same or similar results.

There are researchers who not only publish papers but also a code (written directly in paper or in an online repository). It has become more common that someone reproduces the implementation and published code.

There are also websites like Papers with code[1] or NLP Progress[2] which provide a catalogue of papers and related code.

Here are few examples of dialogue system implementations:

**Simple ALICE** [9] ALICE ia a retrieval-based (§6.4.1) chatbot implemented in Artificial Intelligence Markup Language (AIML) and Python with a minimally modified AIML starter set[3]

**KB-InfoBot** [351] Another End-to-End (E2E) approach is a multi-turn dialogue agent with Knowledge Base (KB)[4] showing that KB lookup helps the reinforcement learner (§8.6.2) discover a suitable Dialogue Policy (DP).

**SimpleDS** [371] A Simple Deep Reinforcement Learning (RL) Dialogue System[5] uses a raw, noisy text without any engineered features to represent the dialogue state and bypass the Natural Language Understanding (NLU) component with DP learning.

**TC-Bot** [74], [350] is an implementation of the E2E task-completion neural dialogue systems and a user simulator[6] for task-completion[7] dialogue research papers.

**Voicy.AI** [184] is a Hybrid Code Network (HCN) (§8.6.6) implementation based chatbot. The Voicy.AI[8] is the pioneering implementation of research papers from Dialog System Technology Challenge (DSTC) (§2.4.3).

**Adversarial dialogue** [9] It uses Adversarial Learning (AL) together with three state of the art task-oriented dialogue models: Variational Hierarchical Recurrent Encoder-Decoder (VHRED) [395], RL [365], and Dynamic Knowledge Graph Network [396] to assess dialogue system sensitivity on request and stability in response.

**Memory-to-sequence (Mem2Seq) dialogue system** [354][10] It is an implementation of an E2E task-oriented dialog system with MEM2Seq memory to sequence model composed of two components: the MemN2N [353] encoder and the memory decoder.

**The Self-feeding Chatbot** [385] It is an interesting example of Active Learning (AL) (§8.6.4). Its code is available under the ParlAI [195] (§6.12) GitHub[11].

[1] https://paperswithcode.com

[2] http://nlpprogress.com

[3] https://github.com/datenhahn/python-aiml-chatbot

[4] https://github.com/MiuLab/KB-InfoBot

[5] https://github.com/cuayahuitl/SimpleDS

[6] https://github.com/MiuLab/UserSimulator

[7] https://github.com/MiuLab/TC-Bot

[8] https://github.com/voicy-ai/DialogStateTracking

[9] https://github.com/WolfNiu/AdversarialDialogue

[10] https://github.com/HLTCHKUST/Mem2Seq

[11] https://github.com/facebookresearch/ParlAI/tree/master/projects/self_feeding

**DialoGPT (Dialogue Generative Pre-Training (GPT))** [347] is a large, tune-able neural conversational response generation model[12] (§8.6.1). It extends the Hugging Face PyTorch transformer (§8.2.4) to gain a performance close to human in terms of automatic and human evaluation when a single-turn dialogue setting is considered.

[12] https://github.com/microsoft/DialoGPT

# Appendix B

# Online Courses

The last few years have been significant for Massive Open Online Course (MOOC). It has been practically a boom with a contribution from universities and individuals who prepared first single teaching courses, which turn lately into specializations or programs under the educational path.

Some courses were not created in an academic environment. Some of them have replaced standard education at universities. When you finish some of them, you get either a simple certificate of completion or some of them contain also grading (usually with percentage pass of the course).

Just a few of them are considered to be full-fledged courses; they include semifinal and final tests comparable to a classic subject exam.

This all together means that education is heading to a new era. The number of courses is growing, the volume of participants is enormous (dozens of thousands studying at the same moment), educational platforms become even more sophisticated, and courses are prepared by recognizable experts (professors or experts from academia) in a particular field.

It means they have become recognizable as well as classic daily or distant education at universities. Even not yet on the same level, some of them are becoming preliminary conditions to study daily programs where the students graduate with a diploma.

## B.1 Educational Platforms

With a growth of MOOC popularity many educational platforms have been established. The most popular are Coursera[1], edX[2], Udemy[3], DataCamp[4], and Codecademy[5]. Each of them offers many courses of various complexity and quality.

[1] https://www.coursera.org

[2] https://www.edx.org

[3] https://www.udemy.com

[4] https://www.datacamp.com

[5] https://www.codecademy.com

## B.2 Chatbots and Dialogue Assistants Courses

A common course about building a chatbot is based on two approaches. Either it is a service within the cloud environment, which provides complete or partial Natural Language Processing (NLP) (especially Natural Language Understanding (NLU)) functionality. Alternatively, it is the course of how to implement an End-to-End (E2E) chatbot from scratch.

The courses related to a cloud service are usually based on a specific technology developed by significant companies in the field like IBM, Amazon, or others:

- Coursera — Building AI Powered Chatbots Without Programming (IBM) [6]
- edX — Microsoft Bot Framework and Conversation as a Platform (Microsoft)[7]
- Codecademy — Introduction to Alexa (Amazon)[8]
- Codecademy — Conversational Design with Alexa (Amazon)[9]
- Codecademy — Learn the Watson API (IBM)[10]
- Udemy — Building a Google Home bot! (With SpaceX knowledge) (Google)[11]
- Udemy — Building Apps Using Amazon's Alexa and Lex (Amazon)[12]

The next group of courses is more technologically independent (even though not entirely) and try to show advantages and disadvantages of various solutions:

- Coursera — Sequence Models[13]
- Udemy — Deep Learning and NLP A-Z™: How to create a ChatBot[14]
- Udemy — Build Incredible Chatbots[15]
- DataCamp — Building Chatbots in Python[16]
- DataCamp — Natural Language Generation in Python[17]

[6] https://www.coursera.org/learn/building-ai-powered-chatbots

[7] https://www.edx.org/course/conversation-as-a-platform-with-the-microsoft-bot-framework

[8] https://www.codecademy.com/learn/learn-alexa

[9] https://www.codecademy.com/learn/alexa-conversational-design

[10] https://www.codecademy.com/learn/ibm-watson

[11] https://www.udemy.com/course/building-your-own-action-on-google

[12] https://www.udemy.com/course/actions-on-google-app-google-assistant

[13] https://www.coursera.org/learn/nlp-sequence-models

[14] https://www.udemy.com/chatbot

[15] https://www.udemy.com/course/build-incredible-chatbots

[16] https://www.datacamp.com/courses/building-chatbots-in-python

[17] https://www.datacamp.com/courses/natural-language-generation-in-python

## B.3 Time Series Courses

Processing of time-series includes pre-processing of influencing data (§4) and data fusion (§5). There are several courses from basic to advance ones; some advanced courses require more knowledge or even experience with this type of data.

Amongst plenty of time series related courses (for instance Udemy offers at least 20 relevant courses for time series analysis search) here are some examples:

- Coursera — Sequences, Time Series and Prediction[18]
- DataCamp — Introduction to Time Series Analysis in Python[19]
- Udemy — Python for Time Series Data Analysis[20]

## B.4 Artificial Intelligence Courses

When building End-to-End (E2E) chatbot models (§8) we usually utilize Artificial Intelligence (AI) and thus an Artificial Neural Network (ANN). Knowledge of various AI models including Long / Short Term Memory (LSTM), Recurrent Neural Network (RNN) and Generative Adversarial Network (GAN) is crucial for further understanding and improvement of chatbots.

The next courses examples fit such criteria:

- Coursera — Deep Learning Specialization[21]
- Coursera — TensorFlow in Practice Specialization[22]
- edX — Introduction to Artificial Intelligence (AI)[23]
- edX — Deep Learning with Python and PyTorch[24]
- Udemy — Practical Deep Learning with PyTorch[25]

[18] https://www.coursera.org/learn/tensorflow-sequences-time-series-and-prediction

[19] https://www.datacamp.com/courses/introduction-to-time-series-analysis-in-python

[20] https://www.udemy.com/course/python-for-time-series-data-analysis

[21] https://www.coursera.org/specializations/deep-learning

[22] https://www.coursera.org/specializations/tensorflow-in-practice

[23] https://www.edx.org/course/introduction-artificial-intelligence-3

[24] https://www.edx.org/course/deep-learning-with-python-and-pytorch-2

[25] https://www.udemy.com/course/practical-deep-learning-with-pytorch/

# Bibliography

[1] T. Šimandl, “Software tools for verification of heart rate measurement accuracy”, PhD thesis, University of West Bohemia in Pilsen, 2017, p. 83. [Online]. Available: `http://hdl.handle.net/11025/27704`.

[2] M. Kuda, “Correlation and causality of sentiment extracted from text and heart rate”, PhD thesis, University of West Bohemia, 2019, p. 84. [Online]. Available: `https://portal.zcu.cz/StagPortletsJSR168/CleanUrl?urlid=prohlizeni-prace-detail&praceIdno=79565`.

[3] J. M. Montes, E. Medina, M. Gomez-Beneyto, and J. Maurino, “A short message service (SMS)-based strategy for enhancing adherence to antipsychotic medication in schizophrenia.”, *Psychiatry research*, vol. 200, no. 2-3, pp. 89–95, Dec. 2012, ISSN: 1872-7123. DOI: `10.1016/j.psychres.2012.07.034`. [Online]. Available: `http://www.ncbi.nlm.nih.gov/pubmed/22901437`.

[4] V. I. O. Agyapong, K. Mrklas, V. Y. M. Suen, M. S. Rose, M. Jahn, I. Gladue, J. Kozak, M. Leslie, S. Dursun, A. Ohinmaa, and A. Greenshaw, “Supportive Text Messages to Reduce Mood Symptoms and Problem Drinking in Patients With Primary Depression or Alcohol Use Disorder: Protocol for an Implementation Research Study.”, *Journal of Medical Internet Research*, vol. 17, no. 5, p. 1, 2015, ISSN: 14388871. [Online]. Available: `http://10.0.8.148/resprot.4371%5Cnhttp://search.ebscohost.com/login.aspx?direct=true&db=lxh&AN=103129396&site=ehost-live`.

[5] F. Španiel, P. Vohlídka, J. Hrdlička, J. Kožený, T. Novák, L. Motlová, J. Čermák, J. Bednařík, D. Novák, and C. Höschl, “ITAREPS: Information Technology Aided Relapse Prevention Programme in Schizophrenia”, *Schizophrenia Research*, 2008, ISSN: 09209964. DOI: `10.1016/j.schres.2007.09.005`.

[6] K. K. Fitzpatrick, A. Darcy, and M. Vierhile, “Delivering Cognitive Behavior Therapy to Young Adults With Symptoms of Depression and Anxiety Using a Fully Automated Conversational Agent (Woebot): A Randomized Controlled Trial”, *JMIR Mental Health*, vol. 4, no. 2, e19, 2017, ISSN: 2368-7959. DOI: `10.2196/mental.7785`. [Online]. Available: `http://mental.jmir.org/2017/2/e19/`.

[7] J. Weizenbaum, “ELIZA - a computer program for the study of natural language communication between man and machine”, *Communications of the ACM*, vol. 9, no. 1, pp. 36–45, 1966, ISSN: 00010782. DOI: `10.1145/365153.365168`. [Online]. Available: `http://portal.acm.org/citation.cfm?doid=365153.365168`.

[8] K. M. Colby, S. Weber, and F. D. Hilf, “Artificial Paranoia”, *Artificial Intelligence*, 1971, ISSN: 00043702. DOI: `10.1016/0004-3702(71)90002-6`.

[9] R. S. Wallace, "The anatomy of A.L.I.C.E.", in *Parsing the Turing Test: Philosophical and Methodological Issues in the Quest for the Thinking Computer*, 2009, ISBN: 9781402067105. DOI: `10.1007/978-1-4020-6710-5{\_}13`.

[10] L. Zhou, J. Gao, D. Li, and H.-Y. Shum, "The Design and Implementation of XiaoIce, an Empathetic Social Chatbot", *Computational Linguistics*, 2020, ISSN: 0891-2017. DOI: `10.1162/coli{\_}a{\_}00368`.

[11] A. Riordan, *Microsoft's AI vision, rooted in research, conversations*, 2016. [Online]. Available: `https://news.microsoft.com/features/microsofts-ai-vision-rooted-in-research-conversations/`.

[12] D. Adiwardana, M.-T. Luong, D. R. So, J. Hall, N. Fiedel, R. Thoppilan, Z. Yang, A. Kulshreshtha, G. Nemade, Y. Lu, and Q. V. Le, "Towards a Human-like Open-Domain Chatbot", Jan. 2020. [Online]. Available: `http://arxiv.org/abs/2001.09977`.

[13] D. Jurafsky and James H. Martin, "Speech and language processing: an introduction to Natural Language Processing, computational linguistics, and speech recognition", in *Speech and Language Processing*, 2017, ch. 29, p. 25, ISBN: 0130950696. [Online]. Available: `https://web.stanford.edu/~jurafsky/slp3/29.pdf%20http://aclweb.org/anthology/J00-4006`.

[14] E. H. Almansor and F. K. Hussain, "Survey on Intelligent Chatbots: State-of-the-Art and Future Research Directions", in *Advances in Intelligent Systems and Computing*, vol. 993, Springer Verlag, 2020, pp. 534–543, ISBN: 9783030223533. DOI: `10.1007/978-3-030-22354-0{\_}47`.

[15] Y.-N. Chen, A. Celikyilmaz, and D. Hakkani-Tür, "Deep Learning for Dialogue Systems", in *Proceedings of ACL 2017, Tutorial Abstracts*, 2017, ISBN: 9781945626777. DOI: `10.18653/v1/P17-5004`.

[16] W. Wu and R. Yan, "Deep Chit-Chat: Deep Learning for ChatBots", 2018, [Online]. Available: `http://www.ruiyan.me/pubs/tutorial-emnlp18.pdf`.

[17] L. Bradeško and D. Mladenić, "A Survey of Chabot Systems through a Loebner Prize Competition", *Researchgate.Net*, 2012.

[18] H. Chen, X. Liu, D. Yin, and J. Tang, "A Survey on Dialogue Systems: Recent Advances and New Frontiers", Tech. Rep., 2018. [Online]. Available: `https://arxiv.org/pdf/1711.01731.pdf`.

[19] L. Hou, Y. Li, C. Li, and M. Lin, "Review of Research on Task-Oriented Spoken Language Understanding", in *Journal of Physics: Conference Series*, vol. 1267, Institute of Physics Publishing, Jul. 2019. DOI: `10.1088/1742-6596/1267/1/012023`.

[20] M. Henderson, "Machine Learning for Dialog State Tracking: A Review", *Proceedings of The First International Workshop on Machine Learning in Spoken Language Processing*, 2015.

[21] J. D. Williams, A. Raux, M. Henderson, and M. Com, "The Dialog State Tracking Challenge Series: A Review", *Dialogue & Discourse*, vol. 7, no. 3, pp. 4–33, 2016. DOI: `10.5087/dad.2016.301`. [Online]. Available: `https://pdfs.semanticscholar.org/4ba3/39bd571585fadb1fb1d14ef902b6784f574f.pdf`.

[22] J. Schatzmann, K. Weilhammer, M. Stuttle, and S. Young, "A survey of statistical user simulation techniques for reinforcement- learning of dialogue management strategies", *Knowledge Engineering Review*, vol. 21, no. 2, pp. 97–126, Jun. 2006, ISSN: 02698889. DOI: `10.1017/S0269888906000944`.

[23] A. Gatt and E. Krahmer, "Survey of the State of the Art in Natural Language Generation: Core tasks, applications and evaluation", Mar. 2017. [Online]. Available: `http://arxiv.org/abs/1703.09902`.

[24] S. Santhanam and S. Shaikh, "A Survey of Natural Language Generation Techniques with a Focus on Dialogue Systems - Past, Present and Future Directions", Jun. 2019. [Online]. Available: `http://arxiv.org/abs/1906.00500`.

[25] O. Dušek, J. Novikova, and V. Rieser, "Evaluating the state-of-the-art of End-to-End Natural Language Generation: The E2E NLG challenge", *Computer Speech and Language*, 2020, ISSN: 10958363. DOI: `10.1016/j.csl.2019.06.009`.

[26] I. V. Serban, R. Lowe, P. Henderson, L. Charlin, and J. Pineau, "A Survey of Available Corpora for Building Data-Driven Dialogue Systems", 2017. [Online]. Available: `https://arxiv.org/pdf/1512.05742.pdf`.

[27] J. Deriu, A. Rodrigo, A. Otegi, G. Echegoyen, S. Rosset, E. Agirre, and M. Cieliebak, "Survey on Evaluation Methods for Dialogue Systems", May 2019. [Online]. Available: `http://arxiv.org/abs/1905.04071`.

[28] A. M. Turing, "COMPUTING MACHINERY AND INTELLIGENCE", *Mind*, vol. 49, pp. 433–460, 1950. [Online]. Available: `https://www.csee.umbc.edu/courses/471/papers/turing.pdf`.

[29] S. A. Abdul-Kader and J. Woods, "Survey on Chatbot Design Techniques in Speech Conversation Systems", Tech. Rep. 7, 2015. [Online]. Available: `www.ijacsa.thesai.org`.

[30] S. Prabhumoye, F. Botros, K. Chandu, S. Choudhary, E. Keni, C. Malaviya, T. Manzini, R. Pasumarthi, S. Poddar, A. Ravichander, Z. Yu, and A. Black, "Building CMU Magnus from User Feedback", in *1st Proceedings of Alexa Prize (Alexa Prize 2017)*, 2017. [Online]. Available: `https://s3.amazonaws.com/alexaprize/2017/technical-article/magnus.pdf`.

[31] H. Liu, T. Lin, H. Sun, W. Lin, C.-W. Chang, T. Zhong, and A. Rudnicky, "RubyStar: A Non-Task-Oriented Mixture Model Dialog System", Tech. Rep., 2017. [Online]. Available: `https://developer.amazon.com/alexa-voice-service`.

[32] J. Pichl, P. Marek, J. Konrád, M. Matulík, H. L. Nguyen, and J. Šedivý, "Alquist: The Alexa Prize Socialbot", *1st Proceedings of Alexa Prize*, pp. 1–10, 2017. [Online]. Available: `https://s3.amazonaws.com/alexaprize/2017/technical-article/alquist.pdf`.

[33] Z. Wang, A. Ahmadvand, J. I. Choi, P. Karisani, and E. Agichtein, "Emersonbot: Information-Focused Conversational AI Emory University at the Alexa Prize 2017 Challenge", Tech. Rep., 2017. [Online]. Available: `https://s3.amazonaws.com/alexaprize/2017/technical-article/emersonbot.pdf`.

[34] I. Papaioannou, A. C. Curry, J. L. Part, I. Shalyminov, X. Xu, and Y. Yu, "Alana : Social Dialogue using an Ensemble Model and a Ranker trained on User Feedback", in *1st Proceedings of Alexa Prize*, 2017, pp. 1–10. [Online]. Available: https://s3.amazonaws.com/alexaprize/2017/technical-article/alana.pdf.

[35] O. Adewale, A. Beatson, D. Buniatyan, J. Ge, M. Khodak, H. Lee, N. Prasad, N. Saunshi, A. Seff, K. Singh, D. Suo, C. Zhang, and S. Arora, "Pixie: A Social Chatbot", Tech. Rep., 2017. [Online]. Available: https://pixie.is.

[36] J. Ji, Q. Wang, Z. Battad, J. Gou, J. Zhou, R. Divekar, C. Carlson, and M. Si, "A Two-Layer Dialogue Framework For Authoring Social Bots", Tech. Rep., 2017. [Online]. Available: https://s3.amazonaws.com/alexaprize/2017/technical-article/wisemacaw.pdf.

[37] S. Yi and K. Jung, "A Chatbot by Combining Finite State Machine, Information Retrieval, and Bot-Initiative Strategy", Tech. Rep., 2017. [Online]. Available: https://github.com/Marsan-Ma/chat_corpus.

[38] W. H. Guss, J. Bartlett, P. Kuznetsov, and P. Patil, "Eigen: A Step Towards Conversational AI", Tech. Rep., 2017. [Online]. Available: https://s3.amazonaws.com/alexaprize/2017/technical-article/eigen.pdf.

[39] K. K. Bowden, J. Wu, S. Oraby, A. Misra, and M. Walker, "Slugbot: An Application of a Novel and Scalable Open Domain Socialbot Framework", Tech. Rep., 2017. [Online]. Available: https://www.amazon.com/dp/B01MRKGF5W.

[40] B. Krause, M. Damonte, M. Dobre, D. Duma, J. Fainberg, F. Fancellu, E. Kahembwe, J. Cheng, and B. Webber, "Edina: Building an Open Domain Socialbot with Self-dialogues", Tech. Rep., 2017. [Online]. Available: https://s3.amazonaws.com/alexaprize/2017/technical-article/edina.pdf.

[41] I. V. Serban, C. Sankar, S. Zhang, Z. Lin, S. Subramanian, T. Kim, S. Chandar, N. R. Ke, S. Rajeswar, A. De Brebisson, J. M. R. Sotelo, D. Suhubdy, V. Michalski, A. Nguyen, and Y. Bengio, "The Octopus Approach to the Alexa Competition: A Deep Ensemble-based Socialbot", Tech. Rep., 2017. [Online]. Available: https://gist.github.com/bebraw/273706..

[42] A. Cervone, G. Tortoreto, S. Mezza, E. Gambi, and G. Riccardi, "Roving Mind: a balancing act between open-domain and engaging dialogue systems", Tech. Rep., 2017. [Online]. Available: https://s3.amazonaws.com/alexaprize/2017/technical-article/rovingmind.pdf.

[43] H. Fang, H. Cheng, E. Clark, A. Holtzman, M. Sap, M. Ostendorf, Y. Choi, and N. A. Smith, "Sounding Board – University of Washington's Alexa Prize Submission", in *1st Proceedings of Alexa Prize*, 2017, p. 12. [Online]. Available: https://pdfs.semanticscholar.org/3ad3/e6c2b9536939c9503e22fd508897923a3152.pdf.

[44] N. Fulda, T. Etchart, W. Myers, D. Ricks, Z. Brown, J. Szendre, B. Murdoch, A. Carr, and W. David, "EVE: Mixed Initiative Dialog via Structured Knowledge Graph Traversal and Conversational Scaffolding", Tech. Rep., 2018. [Online]. Available: https://m.media-amazon.com/images/G/01/mobile-apps/dex/alexa/alexaprize/assets/pdf/2018/Eve.pdf.

[45] G. Larionov, Z. Kaden, H. V. Dureddy, G. B. T. Kalejaiye, M. Kale, S. P. Potharaju, A. P. Shah, and A. I. Rudnicky, "Tartan: A retrieval-based socialbot powered by a dynamic finite-state machine architecture", *Alexa Prize 2018*, 2018. [Online]. Available: `https://m.media-amazon.com/images/G/01/mobile-apps/dex/alexa/alexaprize/assets/pdf/2018/Tartan.pdf`.

[46] J. Pichl, P. Marek, M. Matulík, and J. Šedivý, "Alquist 2.0: Alexa Prize Socialbot Based on Sub-Dialogue Models", Tech. Rep., 2018. [Online]. Available: `https://aws.amazon.com/codepipeline`.

[47] A. Ahmadvand, I. Choi, H. Sahijwani, J. Schmidt, M. Sun, S. Volokhin, Z. Wang, and E. Agichtein, "Emory IrisBot: An Open-Domain Conversational Bot for Personalized Information Access", *Alexa Prize 2018*, 2018. [Online]. Available: `https://m.media-amazon.com/images/G/01/mobile-apps/dex/alexa/alexaprize/assets/pdf/2018/Iris.pdf`.

[48] A. C. Curry, I. Papaioannou, A. Suglia, S. Agarwal, I. Shalyminov, X. Xu, O. Dušek, A. Eshghi, I. Konstas, V. Rieser, and O. Lemon, "Alana v2: Entertaining and Informative Open-domain Social Dialogue using Ontologies and Entity Linking", Tech. Rep., 2018. [Online]. Available: `https://s3.amazonaws.com/dex-microsites-prod/alexaprize/2018/papers/Alana.pdf`.

[49] P. Jonell, M. Bystedt, F. I. Dogan, P. Fallgren, J. Ivarsson, M. Slukova, U. Wennberg, J. Lopes, J. Boye, and G. Skantze, "Fantom: A Crowdsourced Social Chatbot using an Evolving Dialog Graph", *Alexa Prize 2018*, 2018. [Online]. Available: `https://m.media-amazon.com/images/G/01/mobile-apps/dex/alexa/alexaprize/assets/pdf/2018/Fantom.pdf`.

[50] C.-Y. Chen, D. Yu, W. Wen, Y. M. Yang, J. Zhang, M. Zhou, K. Jesse, A. Chau, A. Bhowmick, S. Iyer, G. Sreenivasulu, R. Cheng, A. Bhandare, and Z. Yu, "Gunrock: Building A Human-Like Social Bot By Leveraging Large Scale Real User Data", Tech. Rep., 2018. [Online]. Available: `http://aws.amazon.com/lambda`.

[51] K. K. Bowden, J. W. W. Cui, J. Juraska, V. Harrison, B. Schwarzmann, N. Santer, and M. Walker, "SlugBot: Developing a Computational Model and Framework of a Novel Dialogue Genre", *Alexa Prize 2018*, 2018. [Online]. Available: `https://m.media-amazon.com/images/G/01/mobile-apps/dex/alexa/alexaprize/assets/pdf/2018/Slugbot.pdf`.

[52] R. Blanco, G. Ottaviano, and E. Meij, "Fast and space-efficient entity linking in queries", in *WSDM 2015 - Proceedings of the 8th ACM International Conference on Web Search and Data Mining*, 2015, ISBN: 9781450333177. DOI: `10.1145/2684822.2685317`.

[53] A. Pappu, R. Blanco, Y. Mehdad, A. Stent, and K. Thadani, "Lightweight multilingual entity extraction and linking", in *WSDM 2017 - Proceedings of the 10th ACM International Conference on Web Search and Data Mining*, 2017, ISBN: 9781450346757. DOI: `10.1145/3018661.3018724`.

[54] J. Williams, A. Raux, D. Ramachandran, and A. Black, "The Dialog State Tracking Challenge", Tech. Rep., 2013. [Online]. Available: `https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/dstc2013.pdf`.

[55] M. Henderson, B. Thomson, and J. Williams, "The Second Dialog State Tracking Challenge", Tech. Rep., 2014. [Online]. Available: `https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/summaryWriteup.pdf`.

[56] M. Henderson, B. Thomson, and J. D. Williams, "THE THIRD DIALOG STATE TRACKING CHALLENGE", Tech. Rep., 2014. [Online]. Available: `http://camdial.org/`.

[57] S. Kim, L. Fernando D'haro, R. E. Banchs, J. D. Williams, and M. Henderson, "The Fourth Dialog State Tracking Challenge", Tech. Rep., 2015. [Online]. Available: `http://www.colips.org/workshop/dstc4/papers/60.pdf`.

[58] S. Kim, L. Fernando D'haro, R. E. Banchs, J. D. Williams, M. Henderson, and K. Yoshino, *THE FIFTH DIALOG STATE TRACKING CHALLENGE*. 2016, ISBN: 9781509049035. [Online]. Available: `https://github.com/seokhwankim/dstc5`.

[59] C. Hori, J. Perez, R. Higashinaka, T. Hori, Y. L. Boureau, M. Inaba, Y. Tsunomori, T. Takahashi, K. Yoshino, and S. Kim, "Overview of the sixth dialog system technology challenge: DSTC6", *Computer Speech and Language*, 2019, ISSN: 10958363. DOI: `10.1016/j.csl.2018.09.004`.

[60] J. Perez, Y.-L. Boureau, and A. Bordes, "Dialog System and Technology Challenge 6 Overview of Track 1-End-to-End Goal-Oriented Dialog learning", Tech. Rep., 2017. [Online]. Available: `http://arxiv.org/abs/1503.08895`.

[61] C. Hori and T. Hori, "End-to-end Conversation Modeling Track in DSTC6", Tech. Rep., 2017. [Online]. Available: `http://opus.lingfil.uu.se/download.php?`.

[62] R. Higashinaka, K. Funakoshi, M. Inaba, Y. Tsunomori, T. Takahashi, and N. Kaji, "Overview of Dialogue Breakdown Detection Challenge 3", Tech. Rep., 2017. [Online]. Available: `https://rajpurkar.github.io/SQuAD-explorer/`.

[63] K. Yoshino, C. Hori, J. Perez, L. F. D'Haro, L. Polymenakos, C. Gunasekara, W. S. Lasecki, J. K. Kummerfeld, M. Galley, C. Brockett, J. Gao, B. Dolan, X. Gao, H. Alamari, T. K. Marks, D. Parikh, and D. Batra, "Dialog System Technology Challenge 7", Jan. 2019. [Online]. Available: `http://arxiv.org/abs/1901.03461`.

[64] C. Gunasekara, J. K. Kummerfeld, L. Polymenakos, and W. Lasecki, "DSTC7 Task 1: Noetic End-to-End Response Selection", 2019. DOI: `10.18653/v1/w19-4107`.

[65] M. Galley, C. Brockett, X. Gao, J. Gao, and B. Dolan, "Grounded Response Generation Task at DSTC7", in *AAAI Dialog System Technology Challenges Workshop*, 2019, ISBN: 0090.00933.9.

[66] H. Alamri, C. Hori, T. K. Marks, D. Batra, and D. Parikh, "Audio Visual Scene-aware dialog ( AVSD ) Track for Natural Language Generation in DSTC7", *AAAI*, 2018.

[67] S. Kim, M. Galley, C. Gunasekara, S. Lee, A. Atkinson, B. Peng, H. Schulz, J. Gao, J. Li, M. Adada, M. Huang, L. Lastras, J. K. Kummerfeld, W. S. Lasecki, C. Hori, A. Cherian, T. K. Marks, A. Rastogi, X. Zang, S. Sunkara, and R. Gupta, "The Eighth Dialog System Technology Challenge", Nov. 2019. [Online]. Available: `http://arxiv.org/abs/1911.06394`.

[68] V. Logacheva, M. Burtsev, V. Malykh, V. Poluliakh, A. Rudnicky, I. Serban, R. Lowe, S. Prabhumoye, A. W. Black, and Y. Bengio, "A Dataset of Topic-Oriented Human-to-Chatbot Dialogues", Tech. Rep., 2018. [Online]. Available: `http://convai.io/2017/data/dataset_description.pdf`.

[69] E. Dinan, V. Logacheva, V. Malykh, A. Miller, K. Shuster, J. Urbanek, D. Kiela, A. Szlam, I. Serban, R. Lowe, S. Prabhumoye, A. W. Black, A. Rudnicky, J. Williams, J. Pineau, M. Burtsev, and J. Weston, "The Second Conversational Intelligence Challenge (ConvAI2)", Tech. Rep., 2019. [Online]. Available: `http://convai.io/`.

[70] S. Zhang, E. Dinan, J. Urbanek, A. Szlam, D. Kiela, and J. Weston, "Personalizing Dialogue Agents: I have a dog, do you have pets too?", Tech. Rep., 2018. [Online]. Available: `http://parl.ai`.

[71] I. Yusupov and Y. Kuratov, "Skill-based Conversational Agent", Tech. Rep., 2017. [Online]. Available: `http://convai.io/2017/nips17_wrkshp/bot1337.pdf`.

[72] J. Chorowski, A. Lancucki, S. Malik, M. Pawlikowski, P. Rychlikowski, and P. Zykowski, "University of Wrocław Submission for the NIPS convai.io Competition", Tech. Rep., 2017. [Online]. Available: `http://convai.io/2017/nips17_wrkshp/poetwannabe.pdf`.

[73] N. Gontier, K. Sinha KOUSTUVSINHA, P. Henderson PETERHENDERSON, I. Serban, M. Noseworthy, P. Parthasarathi, and J. Pineau JPINEAU, "The RLLChatbot: a solution to the ConvAI challenge", Tech. Rep., 2018. [Online]. Available: `http://convai.io/2017`.

[74] X. Li, Z. C. Lipton, B. Dhingra, L. Li, J. Gao, and Y.-N. Chen, "A User Simulator for Task-Completion Dialogues *", Tech. Rep., 2017. [Online]. Available: `https://github.com/MiuLab/UserSimulator`.

[75] L. F. D'Haro, S. Kim, K. H. Yeo, R. Jiang, A. I. Niculescu, R. E. Banchs, and H. Li, "CLARA: A multifunctional virtual agent for conference support and touristic information", in *Natural Language Dialog Systems and Intelligent Assistants*, 2015, pp. 233–239, ISBN: 9783319192918. DOI: `10.1007/978-3-319-19291-8{\_}22`.

[76] O. Vinyals and Q. V. Le, "A Neural Conversational Model", 2015. [Online]. Available: `https://arxiv.org/pdf/1506.05869.pdf`.

[77] S. Hoermann, K. L. McCabe, D. N. Milne, and R. A. Calvo, *Application of synchronous text-based dialogue systems in mental health interventions: Systematic review*, 2017. DOI: `10.2196/jmir.7023`.

[78] K. J. Oh, D. Lee, B. Ko, and H. J. Choi, "A chatbot for psychiatric counseling in mental healthcare service based on emotional dialogue analysis and sentence generation", in *Proceedings - 18th IEEE International Conference on Mobile Data Management, MDM 2017*, 2017, ISBN: 9781538639320. DOI: `10.1109/MDM.2017.64`.

[79] B. Graf, M. Krüger, F. Müller, A. Ruhland, and A. Zech, "Nombot: Simplify Food Tracking", *Proceedings of the 14th International Conference on Mobile and Ubiquitous Multimedia*, no. Mum, pp. 360–363, 2015. DOI: `10.1145/2836041.2841208`. [Online]. Available: `http://doi.acm.org/10.1145/2836041.2841208`.

[80] N. Stein and K. Brooks, "A Fully Automated Conversational Artificial Intelligence for Weight Loss: Longitudinal Observational Study Among Overweight and Obese Adults", *JMIR Diabetes*, vol. 2, no. 2, e28, Nov. 2017, ISSN: 2371-4379. DOI: `10.2196/diabetes.8590`. [Online]. Available: `http://diabetes.jmir.org/2017/2/e28/`.

[81] N. Radziwill and M. Benton, “Evaluating Quality of Chatbots and Intelligent Conversational Agents”, Tech. Rep., 2017. [Online]. Available: `http://www.masswerk.at/elizabot/eliza.html`.

[82] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, “BLEU: a Method for Automatic Evaluation of Machine Translation”, Tech. Rep., 2002. [Online]. Available: `https://www.aclweb.org/anthology/P02-1040.pdf`.

[83] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, “SQuAD: 100,000+ Questions for Machine Comprehension of Text”, 2016.

[84] P. Rajpurkar, R. Jia, and P. Liang, “Know What You Don’t Know: Unanswerable Questions for SQuAD”, 2018. [Online]. Available: `https://arxiv.org/pdf/1806.03822.pdf`.

[85] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman, “Glue: A multi-task benchmark and analysis platform for natural language understanding”, in *7th International Conference on Learning Representations, ICLR 2019*, 2019. DOI: `10.18653/v1/w18-5446`.

[86] A. Wang, Y. Pruksachatkun, N. Nangia, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman, “SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems”, Tech. Rep., 2019. [Online]. Available: `https://w4ngatang.github.io/static/papers/superglue.pdf`.

[87] NICE (UK), “Psychological and Psychosocial Interventions”, in *Alcohol-Use Disorders: Diagnosis, Assessment and Management of Harmful Drinking and Alcohol Dependence*, The British Psychological Society & The Royal College of Psychiatrists, 2011, ch. 6, pp. 229–356, ISBN: 978-1-904671-26-8. [Online]. Available: `https://www.ncbi.nlm.nih.gov/pubmedhealth/PMH0042151/`.

[88] M. J. England, A. S. Butler, and M. L. Gonzalez, “Introduction”, in *Psychosocial Interventions for Mental and Substance Use Disorders: A Framework for Establishing Evidence-Based Standards.* Washington (DC): National Academies Press (US), Sep. 2015, ch. 1, pp. 21–46, ISBN: 978-0-309-31694-1. [Online]. Available: `https://www.ncbi.nlm.nih.gov/books/NBK321284/`.

[89] L. I. Stein and M. A. Test, “Alternative to mental hospital treatment. I. Conceptual model, treatment program, and clinical evaluation.”, *Archives of general psychiatry*, vol. 37, no. 4, pp. 392–7, Apr. 1980, ISSN: 0003-990X. [Online]. Available: `http://www.ncbi.nlm.nih.gov/pubmed/7362425`.

[90] A. T. Beck, *Cognitive Therapy and the Emotional Disorders.* New York: Penguin Group, 1979, ISBN: 978-1-101-65988-5.

[91] M. Stitzer and N. Petry, “Contingency Management for Treatment of Substance Abuse”, *Annual Review of Clinical Psychology*, 2006. DOI: `10.1146/annurev.clinpsy.2.022305.095219`.

[92] N. Moyal, A. Henik, and G. E. Anholt, “Cognitive strategies to regulate emotions—current evidence and future directions”, *Frontiers in Psychology*, vol. 4, p. 1019, Jan. 2014, ISSN: 16641078. DOI: `10.3389/fpsyg.2013.01019`. [Online]. Available: `http://journal.frontiersin.org/article/10.3389/fpsyg.2013.01019/abstract`.

[93] M. P. Wallen, S. R. Gomersall, S. E. Keating, U. Wisløff, and J. S. Coombes, "Accuracy of heart rate watches: Implications for weight management", *PLoS ONE*, 2016, ISSN: 19326203. DOI: 10.1371/journal.pone.0154420.

[94] S. E. Stahl, H.-S. An, D. M. Dinkel, J. M. Noble, and J.-M. Lee, "How accurate are the wrist-based heart rate monitors during walking and running activities? Are they accurate enough?", *BMJ Open Sport & Exercise Medicine*, 2016, ISSN: 2055-7647. DOI: 10.1136/bmjsem-2015-000106.

[95] R. S. Thiebaud, M. D. Funk, J. C. Patton, B. L. Massey, T. E. Shay, M. G. Schmidt, and N. Giovannitti, "Validity of wrist-worn consumer products to measure heart rate and energy expenditure.", *Digital health*, vol. 4, p. 2 055 207 618 770 322, 2018, ISSN: 2055-2076. DOI: 10.1177/2055207618770322. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/29942628%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC6001222.

[96] A. W. Gorny, S. J. Liew, C. S. Tan, and F. Müller-Riemenschneider, "Fitbit Charge HR Wireless Heart Rate Monitor: Validation Study Conducted Under Free-Living Conditions", *JMIR mHealth and uHealth*, 2017, ISSN: 2291-5222. DOI: 10.2196/mhealth.8233.

[97] S. Benedetto, C. Caldato, E. Bazzan, D. C. Greenwood, V. Pensabene, and P. Actis, "Assessment of the fitbit charge 2 for monitoring heart rate", *PLoS ONE*, 2018, ISSN: 19326203. DOI: 10.1371/journal.pone.0192691.

[98] V. E. Salazar, N. D. Lucio, and M. D. Funk, "Accuracy of Fitbit Charge 2 Worn at Different Wrist Locations During Exercise", *Medicine & Science in Sports & Exercise*, 2017, ISSN: 0195-9131. DOI: 10.1249/01.mss.0000517890.91198.7b.

[99] X. Li, J. Dunn, D. Salins, G. Zhou, W. Zhou, S. M. Schüssler-Fiorenza Rose, D. Perelman, E. Colbert, R. Runge, S. Rego, R. Sonecha, S. Datta, T. McLaughlin, and M. P. Snyder, "Digital Health: Tracking Physiomes and Activity Using Wearable Biosensors Reveals Useful Health-Related Information", *PLOS Biology*, vol. 15, no. 1, T. Kirkwood, Ed., e2001402, Jan. 2017, ISSN: 1545-7885. DOI: 10.1371/journal.pbio.2001402. [Online]. Available: http://dx.plos.org/10.1371/journal.pbio.2001402.

[100] T. V. Pyrkov, K. Slipensky, M. Barg, A. Kondrashin, B. Zhurov, A. Zenin, M. Pyatnitskiy, L. Menshikov, S. Markov, and P. O. Fedichev, "Extracting biological age from biomedical data via deep learning: too much of a good thing?", *Scientific Reports*, vol. 8, no. 1, p. 5210, Dec. 2018, ISSN: 2045-2322. DOI: 10.1038/s41598-018-23534-9. [Online]. Available: http://www.nature.com/articles/s41598-018-23534-9.

[101] M. Salai, I. Vassányi, and I. Kósa, "Stress detection using low cost heart rate sensors", *Journal of Healthcare Engineering*, 2016, ISSN: 20402309. DOI: 10.1155/2016/5136705.

[102] Y. Shi, N. Ruiz, R. Taib, E. Choi, and F. Chen, "Galvanic skin response (GSR) as an index of cognitive load", in *CHI '07 extended abstracts on Human factors in computing systems - CHI '07*, 2007, ISBN: 9781595936424. DOI: 10.1145/1240866.1241057.

[103] U. Lundberg, R. Kadefors, B. Melin, G. Palmerud, P. Hassmén, M. Engström, and I. Elfsberg Dohns, "Psychophysiological stress and emg activity of the trapezius muscle", *International Journal of Behavioral Medicine*, 1994, ISSN: 10705503. DOI: 10.1207/s15327558ijbm0104{_}5.

[104] J. Wijsman, B. Grundlehner, J. Penders, and H. Hermens, "Trapezius muscle EMG as predictor of mental stress", *ACM Transactions on Embedded Computing Systems*, 2013, ISSN: 15399087. DOI: 10.1145/2485984.2485987.

[105] T. Yamakoshi, K. Yamakoshi, S. Tanaka, M. Nogawa, S. Park, M. Shibata, Y. Sawada, P. Rolfe, and Y. Hirose, "Feasibility study on driver's stress detection from differential skin temperature measurement", in *2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2008, ISBN: 978-1-4244-1814-5. DOI: 10.1109/IEMBS.2008.4649346.

[106] J. Taelman, S. Vandeput, A. Spaepen, and S. Van Huffel, "Influence of mental stress on heart rate and heart rate variability", in *IFMBE Proceedings*, 2008, ISBN: 9783540892076. DOI: 10.1007/978-3-540-89208-3{_}324.

[107] C. Schubert, M. Lambertz, R. A. Nelesen, W. Bardwell, J. B. Choi, and J. E. Dimsdale, "Effects of stress on heart rate complexity-A comparison between short-term and chronic stress", *Biological Psychology*, vol. 80, no. 3, pp. 325–332, 2009. DOI: 10.1016/j.biopsycho.2008.11.005. arXiv: 0507464v2 [astro-ph].

[108] J. Choi and G. O. Ricardo, "Using heart rate monitors to detect mental stress", in *Proceedings - 2009 6th International Workshop on Wearable and Implantable Body Sensor Networks, BSN 2009*, 2009, ISBN: 9780769536446. DOI: 10.1109/BSN.2009.13.

[109] N. Hjortskov, D. Rissén, A. K. Blangsted, N. Fallentin, U. Lundberg, and K. Søgaard, "The effect of mental stress on heart rate variability and blood pressure during computer work", *European Journal of Applied Physiology*, 2004, ISSN: 14396319. DOI: 10.1007/s00421-004-1055-z.

[110] W. Lawanont, P. Mongkolnam, C. Nukoolkit, and M. Inoue, "Daily stress recognition system using activity tracker and smartphone based on physical activity and heart rate data", in *Smart Innovation, Systems and Technologies*, 2019, ISBN: 9783319920276. DOI: 10.1007/978-3-319-92028-3{_}2.

[111] J. F. Thayer, F. Åhs, M. Fredrikson, J. J. Sollers Iii, and T. D. Wager, "A meta-analysis of heart rate variability and neuroimaging studies: Implications for heart rate variability as a marker of stress and health", *Neuroscience and Biobehavioral Reviews*, vol. 36, pp. 747–756, 2011. DOI: 10.1016/j.neubiorev.2011.11.009. [Online]. Available: https://www.praktijk-kinova.nl/index_htm_files/implications-for-heart-rate-variability-as-a-mark.pdf.

[112] *Soft data Meaning in the Cambridge English Dictionary.* [Online]. Available: https://dictionary.cambridge.org/dictionary/english/soft-data.

[113] *Hard Data vs. Soft Data — Objectivity.* [Online]. Available: http://www.objectivity.com/hard-data-vs-soft-data/.

[114] *Hard Data vs Soft Data - Simplicable.* [Online]. Available: https://simplicable.com/new/hard-data-vs-soft-data.

[115] *Hard data Meaning in the Cambridge English Dictionary.* [Online]. Available: https://dictionary.cambridge.org/dictionary/english/hard-data.

[116] M. A. Pravia, R. K. Prasanth, P. O. Arambel, C. Sidner, and C. Chong, "Generation of a fundamental data set for hard/soft information fusion", in *2008 11th International Conference on Information Fusion*, Jun. 2008, pp. 1–8.

[117] B. Khaleghi, A. Khamis, F. O. Karray, and S. N. Razavi, "Multisensor data fusion: A review of the state-of-the-art", *Information Fusion*, vol. 14, no. 1, pp. 28–44, Jan. 2013, ISSN: 1566-2535. DOI: `10.1016/J.INFFUS.2011.08.001`. [Online]. Available: `https://www.sciencedirect.com/science/article/pii/S1566253511000558`.

[118] K. Sambhoos, J. Llinas, and E. Little, "Graphical methods for real-time fusion and estimation with soft message data", in *Proceedings of the 11th International Conference on Information Fusion, FUSION 2008*, 2008, ISBN: 9783000248832. DOI: `10.1109/ICIF.2008.4632405`.

[119] D. L. Hall, M. McNeese, J. Llinas, and T. Mullen, "A framework for dynamic hard/soft fusion", in *Proceedings of the 11th International Conference on Information Fusion, FUSION 2008*, 2008, ISBN: 9783000248832. DOI: `10.1109/ICIF.2008.4632196`.

[120] D. D. L. Hall, M. M. D. McNeese, D. D. B. Hellar, B. J. B. Panulla, and W. Shumaker, "A Cyber Infrastructure for Evaluating the Performance of Human Centered Fusion", in *12th International Conference on Information Fusion*, 2009, ISBN: 978-0-9824-4380-4.

[121] K. Premaratne, M. Murthi, J. Z. J. Zhang, M. Scheutz, and P. Bauer, "A Dempster-Shafer theoretic conditional approach to evidence updating for fusion of hard and soft data", *2009 12th International Conference on Information Fusion*, 2009.

[122] M. A. Pravia, O. Babko-Malaya, M. K. Schneider, J. V. White, and C.-Y. Chong, "Lessons Learned in the Creation of a Data Set for Hard/Soft Information Fusion", in *2009 12th International Conference on Information Fusion*, 2009, ISBN: 978-0-9824-4380-4.

[123] S. B. Hulley, S. R. Cummings, W. S. Browner, D. G. Grady, and T. B. Newman, *Designing Clinical Research.* 2007, ISBN: 9780781722186. DOI: `10.1097/00006982-199010000-00024`.

[124] J. Salamon and R. Mouček, "Heart rate and sentiment experimental data with common timeline", *Data in Brief*, vol. 15, 2017, ISSN: 23523409. DOI: `10.1016/j.dib.2017.10.037`.

[125] J. DiNardo, "Natural Experiments and Quasi-Natural Experiments", in *New Palgrave Dictionary of Economics*, 2008, ISBN: 978-0-333-78676-5. DOI: `10.1057/b.9780631218234.2008.X`.

[126] P. Patil and P. Yalagi, "Sentiment Analysis Levels and Techniques : A Survey", *International Journal of Innovations in Engineering and Technology*, vol. 6, no. 4, pp. 523–528, 2016.

[127] C. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, and D. McClosky, "The Stanford CoreNLP Natural Language Processing Toolkit", in *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 2014, pp. 55–60, ISBN: 9781941643006. DOI: `10.3115/v1/P14-5010`. [Online]. Available: `http://aclweb.org/anthology/P14-5010`.

[128] R. Socher, A. Perelygin, J. Y. Wu, J. Chuang, C. D. Manning, A. Y. Ng, and C. Potts, "Recursive deep models for semantic compositionality over a sentiment treebank", in *EMNLP 2013 - 2013 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, 2013, ISBN: 9781937284978.

[129] P. Ekman, "An Argument for Basic Emotions", *Cognition and Emotion*, 1992, ISSN: 14640600. DOI: 10.1080/02699939208411068.

[130] K. R. Scherer and H. G. Wallbott, ""Evidence for universality and cultural variation of differential emotion response patterning": Correction", *Journal of Personality and Social Psychology*, vol. 67, no. 1, pp. 55–55, 1994, ISSN: 0022-3514. DOI: 10.1037/0022-3514.67.1.55. [Online]. Available: http://doi.apa.org/getdoi.cfm?doi=10.1037/0022-3514.67.1.55.

[131] A. Yousif, Z. Niu, J. K. Tarus, A. Ahmad, and B. Z. Niu, "A survey on sentiment analysis of scientific citations", *Artificial Intelligence Review*, 2019. DOI: 10.1007/s10462-017-9597-8. [Online]. Available: https://link.springer.com/content/pdf/10.1007%2Fs10462-017-9597-8.pdf.

[132] F. Å. Nielsen, "A new ANEW: Evaluation of a word list for sentiment analysis in microblogs", in *CEUR Workshop Proceedings*, 2011.

[133] M. M. Bradley and P. J. Lang, "Affective Norms for English Words (ANEW): Instruction Manual and Affective Ratings", The Center for Research in Psychophysiology, University of Florida, Tech. Rep., 1999. [Online]. Available: https://www.uvm.edu/pdodds/teaching/courses/2009-08UVM-300/docs/others/everything/bradley1999a.pdf.

[134] F. Nielsen, *AFINN*, 2011.

[135] T. Wilson, J. Wiebe, and P. Hoffmann, "Recognizing contextual polarity in phrase-level sentiment analysis", in *HLT/EMNLP 2005 - Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, 2005.

[136] A. Esuli and F. Sebastiani, "SENTIWORDNET: A publicly available lexical resource for opinion mining", in *Proceedings of the 5th International Conference on Language Resources and Evaluation, LREC 2006*, 2006.

[137] S. Baccianella, A. Esuli, and F. Sebastiani, "SENTIWORDNET 3.0: An enhanced lexical resource for sentiment analysis and opinion mining", in *Proceedings of the 7th International Conference on Language Resources and Evaluation, LREC 2010*, 2010, ISBN: 2951740867.

[138] C. Strapparava and A. Valitutti, "WordNet-Affect: An affective extension of WordNet", in *Proceedings of the 4th International Conference on Language Resources and Evaluation, LREC 2004*, 2004, ISBN: 2951740816.

[139] M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, and A. Kappas, "Sentiment in short strength detection informal text", *Journal of the American Society for Information Science and Technology*, 2010, ISSN: 15322882. DOI: 10.1002/asi.21416.

[140] C. J. Hutto and E. Gilbert, "VADER: A parsimonious rule-based model for sentiment analysis of social media text", in *Proceedings of the 8th International Conference on Weblogs and Social Media, ICWSM 2014*, 2014, ISBN: 9781577356578.

[141] G. A. Miller, "WordNet: A Lexical Database for English", *Communications of the ACM*, 1995, ISSN: 15577317. DOI: 10.1145/219717.219748.

[142] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, "Bag of Tricks for Efficient Text Classification", 2016, ISSN: 10450823. DOI: 1511.09249v1.

[143] R. J. A. Little and D. B. Rubin, *Statistical Analysis with Missing Data: Second Edition.* 2002, ISBN: 0471183865.

[144] Z. Zhang, "Missing data imputation: focusing on single imputation", *Hemodialysis International, Journal of Translational Medicine*, 2016, ISSN: 2305-5839. DOI: 10.3978/j.issn.2305-5839.2015.12.38.

[145] J. Salamon, K. Černá, and R. Mouček, "Stress Dichotomy using Heart Rate and Tweet Sentiment", in *Proceedings of the 11th International Joint Conference on Biomedical Engineering Systems and Technologies*, SCITEPRESS - Science and Technology Publications, 2018, pp. 527–532, ISBN: 978-989-758-281-3. DOI: 10.5220/0006650105270532. [Online]. Available: http://www.scitepress.org/DigitalLibrary/Link.aspx?doi=10.5220/0006650105270532.

[146] F. H. Rachman, R. Sarno, and C. Fatichah, "CBE: Corpus-based of emotion for emotion detection in text document", in *Proceedings - 2016 3rd International Conference on Information Technology, Computer, and Electrical Engineering, ICITACEE 2016*, 2017, pp. 331–335, ISBN: 9781509014347. DOI: 10.1109/ICITACEE.2016.7892466.

[147] P. Fournier-Viger, J. Chun, W. Lin, R. U. Kiran, Y. S. Koh, and R. Thomas, "A Survey of Sequential Pattern Mining", vol. 1, no. 1, 2017. [Online]. Available: http://www.philippe-fournier-viger.com/dspr-paper5.pdf.

[148] A. G. Barnett, P. Baker, and A. J. Dobson, "Analysing seasonal data", *R Journal*, 2012, ISSN: 20734859. DOI: 10.32614/rj-2012-001.

[149] J. Baron, *2017 Messenger Bot Landscape, a Public Spreadsheet Gathering 1000+ Messenger Bots*, 2017. [Online]. Available: https://recast.ai/blog/2017-messenger-bot-landscape/.

[150] Gartner, "Gartner Customer 360 Summit 2011", Gartner, Los Angeles, CA, Tech. Rep., 2011, p. 9. [Online]. Available: https://www.gartner.com/imagesrv/summits/docs/na/customer-360/C360_2011_brochure_FINAL.pdf.

[151] ——, *Gartner Says 25 Percent of Customer Service Operations Will Use Virtual Customer Assistants by 2020*, 2018. [Online]. Available: https://www.gartner.com/newsroom/id/3858564.

[152] D. Britz, *Deep Learning for Chatbots, Part 1 – Introduction*, 2016. [Online]. Available: http://www.wildml.com/2016/04/deep-learning-for-chatbots-part-1-introduction/.

[153] S. Kojouharov, *Ultimate Guide to Leveraging NLP &amp; Machine Learning for your Chatbot*, 2016. [Online]. Available: https://chatbotslife.com/ultimate-guide-to-leveraging-nlp-machine-learning-for-you-chatbot-531ff2dd870c.

[154] D. Jurafsky, "CS 124/LINGUIST 180 From Languages to Information - Conversational Agents", Stanford University, Tech. Rep., 2019. [Online]. Available: https://web.stanford.edu/class/cs124/lec/chatbot19.pdf.

[155] A. Maas, "CS 224S / LINGUIST 285 Spoken Language Processing - Lecture 10: Dialogue System Introduction and Frame-Based Dialogue", Stanford University, Tech. Rep., 2017. [Online]. Available: `https://web.stanford.edu/class/cs224s/lectures/224s.17.lec10.pdf`.

[156] A. Ram, R. Prasad, C. Khatri, A. Venkatesh, R. Gabriel, Q. Liu, J. Nunn, B. Hedayatnia, M. Cheng, A. Nagar, E. King, K. Bland, A. Wartick, Y. Pan, H. Song, S. Jayadevan, G. Hwang, and A. Pettigrue, "Conversational AI: The Science Behind the Alexa Prize", 2017. [Online]. Available: `https://s3.amazonaws.com/alexaprize/2017/technical-article/alexaprize.pdf`.

[157] T. Risueño, *How To Solve The Double Intent Issue For Chatbots*, 2017. [Online]. Available: `https://chatbotsmagazine.com/how-to-solve-the-double-intent-issue-for-chatbots-9f031513747f`.

[158] R. Jia and P. Liang, "Data recombination for neural semantic parsing", in *54th Annual Meeting of the Association for Computational Linguistics, ACL 2016 - Long Papers*, 2016, ISBN: 9781510827585.

[159] D. G. Bobrow, R. M. Kaplan, M. Kay, D. A. Norman, H. Thompson, and T. Winograd, "GUS, a frame-driven dialog system", *Artificial Intelligence*, 1977, ISSN: 00043702. DOI: `10.1016/0004-3702(77)90018-2`.

[160] A. Bapna, G. Tür, D. Hakkani-Tür, and L. Heck, "Towards zero-shot frame semantic parsing for domain scaling", in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2017. DOI: `10.21437/Interspeech.2017-518`.

[161] L. A. Ramshaw and M. P. Marcus, "Text Chunking Using Transformation-Based Learning", in, 1999. DOI: `10.1007/978-94-017-2390-9{\_}10`.

[162] B. Liu and I. Lane, "Attention-based recurrent neural network models for joint intent detection and slot filling", in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2016. DOI: `10.21437/Interspeech.2016-1352`.

[163] P.-H. Su and M. Gašić, "Dialogue management: Parametric approaches to policy optimisation", Cambridge University Engineering Department, Tech. Rep., 2016. [Online]. Available: `https://www.cs.hhu.de/fileadmin/redaktion/Fakultaeten/Mathematisch-Naturwissenschaftliche_Fakultaet/Informatik/Dialog_Systems_and_Machine_Learning/Lectures_SDS/L5.pdf`.

[164] D. Bohus and A. I. Rudnicky, "RavenClaw: Dialog management using hierarchical task decomposition and an expectation agenda", in *EUROSPEECH 2003 - 8th European Conference on Speech Communication and Technology*, International Speech Communication Association, 2003, pp. 597–600.

[165] ——, "The RavenClaw dialog management framework: Architecture and systems", *Computer Speech and Language*, vol. 23, no. 3, pp. 332–361, Jul. 2009, ISSN: 08852308. DOI: `10.1016/j.csl.2008.10.001`.

[166] S. Kim, R. E. Banchs, and H. Li, "Towards improving dialogue topic tracking performances with wikification of concept mentions", in *SIGDIAL 2015 - 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue, Proceedings of the Conference*, 2015, ISBN: 9781941643754.

[167] V. Zue, S. Seneff, J. R. Glass, J. Polifroni, C. Pao, T. J. Hazen, and L. Hetherington, "JUPITER: A telephone-based conversational interface for weather information", *IEEE Transactions on Speech and Audio Processing*, 2000, ISSN: 10636676. DOI: `10.1109/89.817460`.

[168] S. LARSSON and D. R. TRAUM, "Information state and dialogue management in the TRINDI dialogue move engine toolkit", *Natural Language Engineering*, 2000, ISSN: 13513249. DOI: `10.1017/s1351324900002539`.

[169] U. Syed and J. D. Williams, "Using automatically transcribed dialogs to learn user models in a spoken dialog system", in *ACL-08: HLT - 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference*, 2008, ISBN: 9781932432046. DOI: `10.3115/1557690.1557722`.

[170] B. Thomson, F. Jurčíček, M. Gašić, S. Keizer, F. Mairesse, K. Yu, and S. Young, "Parameter learning for POMDP spoken dialogue models", in *2010 IEEE Workshop on Spoken Language Technology, SLT 2010 - Proceedings*, 2010. DOI: `10.1109/SLT.2010.5700863`.

[171] D. Bohus and A. Rudnicky, "A "K hypotheses + other" belief updating model", in *AAAI Workshop - Technical Report*, 2006, ISBN: 1577352963.

[172] A. Metallinou, D. Bohus, and J. D. Williams, "Discriminative state tracking for spoken dialog systems", in *ACL 2013 - 51st Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, 2013, ISBN: 9781937284503.

[173] J. D. Williams, "Web-style ranking and SLU combination for dialog state tracking", in *SIGDIAL 2014 - 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue, Proceedings of the Conference*, 2014, ISBN: 9781941643211. DOI: `10.3115/v1/w14-4339`.

[174] M. Henderson, B. Thomson, and S. Young, "Deep Neural Network Approach for the Dialog State Tracking Challenge", *Proceedings of the SIGDIAL 2013 Conference*, pp. 467–471, 2013. [Online]. Available: `http://www.aclweb.org/anthology/W/W13/W13-4073`.

[175] H. Ren, W. Xu, and Y. Yan, "Markovian discriminative modeling for cross-domain dialog state tracking", in *2014 IEEE Workshop on Spoken Language Technology, SLT 2014 - Proceedings*, 2014, ISBN: 9781479971299. DOI: `10.1109/SLT.2014.7078598`.

[176] J. Lafferty, A. McCallum, and F. C. N. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data", *ICML '01 Proceedings of the Eighteenth International Conference on Machine Learning*, 2001, ISSN: 1750-2799. DOI: `10.1038/nprot.2006.61`.

[177] M. Henderson, B. Thomson, and S. Young, "Word-Based Dialog State Tracking with Recurrent Neural Networks", *SIGdial*, 2014. DOI: `10.3115/v1/W14-4340`.

[178] H. Schulz, J. Zumer, L. El Asri, and S. Sharma, "A Frame Tracking Model for Memory-Enhanced Dialogue Systems", 2017. DOI: `10.18653/v1/w17-2626`.

[179] N. Mrkšić, D. Séaghdha, T. H. Wen, B. Thomson, and S. Young, "Neural belief tracker: Data-driven dialogue state tracking", in *ACL 2017 - 55th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, 2017, ISBN: 9781945626753. DOI: `10.18653/v1/P17-1163`.

[180] N. Mrkšić and I. Vulić, "Fully statistical neural belief tracking", in *ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, 2018, ISBN: 9781948087346.

[181] L. Shu, H. Xu, B. Liu, and P. Molino, "Modeling Multi-Action Policy for Task-Oriented Dialogues", 2019. [Online]. Available: `https://arxiv.org/pdf/1908.11546.pdf`.

[182] P.-H. Su, "Reward Estimation for Dialogue Policy Optimisation", Tech. Rep., 2017.

[183] A. J. Stent, "A conversation acts model for generating spoken dialogue contributions", *Computer Speech and Language*, 2002, ISSN: 08852308. DOI: `10.1016/S0885-2308(02)00009-8`.

[184] J. D. Williams, K. Asadi, and G. Zweig, "Hybrid Code Networks: practical and efficient end-to-end dialog control with supervised and reinforcement learning", *ACL 2017 - 55th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, vol. 1, pp. 665–677, Feb. 2017. [Online]. Available: `http://arxiv.org/abs/1702.03274`.

[185] J. D. Williams and G. Zweig, "End-to-end LSTM-based dialog control optimized with supervised and reinforcement learning", 2016. [Online]. Available: `https://arxiv.org/pdf/1606.01269.pdf`.

[186] M. A. Walker, "An Application of Reinforcement Learning to Dialogue Strategy Selection in a Spoken Dialogue System for Email", *Journal of Artificial Intelligence Research*, 2000, ISSN: 10769757.

[187] S. Young, J. Schatzmann, K. Weilhammer, and Y. Hui, "The Hidden information state approach to dialog management", in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2007, ISBN: 1424407281. DOI: `10.1109/ICASSP.2007.367185`.

[188] M. Gašić and S. Young, "Gaussian processes for POMDP-based dialogue manager optimization", *IEEE Transactions on Audio, Speech and Language Processing*, 2014, ISSN: 15587916. DOI: `10.1109/TASL.2013.2282190`.

[189] C. Kamm, "User interfaces for voice applications", *Proceedings of the National Academy of Sciences of the United States of America*, 1995, ISSN: 00278424. DOI: `10.1073/pnas.92.22.10031`.

[190] R. W. Smith and D. R. Hipp, "Spoken natural language dialog systems: A practical approach.", *Spoken natural language dialog systems: A practical approach.*, 1994.

[191] A. Rudnicky, E. Thayer, P. Constantinides, C. Tchou, R. Shern, K. Lenzo, W. Xu, and A. Oh, "Creating Natural Dialogs in the Carnegie Mellon Communicator System", *EUROSPEECH'99 Proceedings, Sixth European Conference on Speech Communication and Technology*, 1999.

[192] M. Fleming and R. Cohen, "A user modeling approach to determining system initiative in mixed-initiative AI systems", in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2001, ISBN: 3540423257. DOI: `10.1007/3-540-44566-8{\_}6`.

[193] A. Gatt and E. Krahmer, "Survey of the state of the art in natural language generation: Core tasks, applications and evaluation", *Journal of Artificial Intelligence Research*, 2018, ISSN: 10769757. DOI: 10.1613/jair.5714.

[194] Z. Xie, "Neural Text Generation: A Practical Guide", Tech. Rep., 2018. [Online]. Available: http://cs.stanford.edu/~zxie/textgen.pdf.

[195] A. Miller, W. Feng, D. Batra, A. Bordes, A. Fisch, J. Lu, D. Parikh, and J. Weston, "ParlAI: A Dialog Research Software Platform", Association for Computational Linguistics (ACL), Jan. 2018, pp. 79–84. DOI: 10.18653/v1/d17-2014.

[196] T. Bocklisch, J. Faulkner, N. Pawlowski, and A. Nichol, "Rasa: Open Source Language Understanding and Dialogue Management", Dec. 2017. [Online]. Available: http://arxiv.org/abs/1712.05181.

[197] A. Coucke, A. Saade, A. Ball, T. Bluche, A. Caulier, D. Leroy, C. Doumouro, T. Gisselbrecht, F. Caltagirone, T. Lavril, M. Primet, and J. Dureau, "Snips Voice Platform: an embedded Spoken Language Understanding system for private-by-design voice interfaces", Tech. Rep., 2018. [Online]. Available: https://www.cnil.fr/fr/enceintes-intelligentes-des-assistants-vocaux-.

[198] M. Burtsev, A. Seliverstov, R. Airapetyan, M. Arkhipov, D. Baymurzina, N. Bushkov, O. Gureenkova, T. Khakhulin, Y. Kuratov, D. Kuznetsov, A. Litinsky, V. Logacheva, A. Lymar, V. Malykh, M. Petrov, V. Polulyakh, L. Pugachev, A. Sorokin, M. Vikhreva, and M. Zaynutdinov, "DeepPavlov: Open-Source Library for Dialogue Systems", in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics-System Demonstrations*, Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 122–127. [Online]. Available: https://www.aclweb.org/anthology/P18-4021.

[199] M. Gardner, J. Grus, M. Neumann, O. Tafjord, P. Dasigi, N. F. Liu, M. Peters, M. Schmitz, and L. Zettlemoyer, "AllenNLP: A Deep Semantic Natural Language Processing Platform", Tech. Rep., 2018. [Online]. Available: http://allennlp.org/.

[200] S. Ultes, L. Rojas-Barahona, P. H. Su, D. Vandyke, D. Kim, I. Casanueva, P. Budzianowski, N. Mrkšić, T. H. Wen, M. Gašić, and S. Young, "Pydial: A multi-domain statistical dialogue system toolkit", in *ACL 2017 - 55th Annual Meeting of the Association for Computational Linguistics, Proceedings of System Demonstrations*, Association for Computational Linguistics (ACL), 2017, pp. 73–78, ISBN: 9781945626715. DOI: 10.18653/v1/P17-4013.

[201] A. Akbik, T. Bergmann, D. Blythe, K. Rasul, S. Schweter, and R. Vollgraf, "FLAIR: An Easy-to-Use Framework for State-of-the-Art NLP", *Proceedings of the 2019 Conference of the North*, 2019. DOI: 10.18653/v1/N19-4010.

[202] A. Akbik, D. Blythe, and R. Vollgraf, "Contextual string embeddings for sequence labeling", *Proceedings of the 27th International Conference on Computational Linguistics*, 2018.

[203] O. Davydova, *25 Chatbot Platforms: A Comparative Table*, 2017. [Online]. Available: https://chatbotsjournal.com/25-chatbot-platforms-a-comparative-table-aeefc932eaff.

[204] S. Kumar, *Top 10 Powerful Platforms for Chatbot Development.* [Online]. Available: https://dzone.com/articles/top-10-powerful-platforms-for-chatbot-development.

[205] K. Ismail, *Top 14 Chatbot Building Platforms of 2017.* [Online]. Available: https://www.cmswire.com/customer-experience/top-14-chatbot-building-platforms-of-2017/.

[206] N. Tank, *How to Build Facebook Messenger chat Bot without any coding?* [Online]. Available: https://tutorials.botsfloor.com/how-to-build-facebook-messenger-chat-bot-without-any-coding-4fe42393e2e4.

[207] ——, *How to Build Facebook Messenger chat Bot without any coding? (Part II)*, 2016. [Online]. Available: https://chatbotslife.com/how-to-build-facebook-messenger-chat-bot-without-any-coding-part-ii-ccd699d92fad.

[208] G. Neff and P. Nagy, "Talking to bots: Symbiotic agency and the case of Tay", *International Journal of Communication*, vol. 10, pp. 4915–4931, 2016, ISSN: 19328036.

[209] B. Pang and L. Lee, "Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales", in *ACL-05 - 43rd Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, 2005, ISBN: 1932432515.

[210] ——, "A Sentimental Education: Sentiment Analysis Using Subjectivity", in *Proceedings of ACL*, 2004, pp. 271–278. [Online]. Available: http://www.cs.cornell.edu/home/llee/papers/cutsent.pdf.

[211] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, "Learning word vectors for sentiment analysis", in *ACL-HLT 2011 - Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 2011, ISBN: 9781932432879.

[212] G. Lai, Q. Xie, H. Liu, Y. Yang, and E. Hovy, "RACE: Large-scale ReAding Comprehension Dataset From Examinations", Tech. Rep., 2017. [Online]. Available: https://arxiv.org/pdf/1704.04683.pdf.

[213] J. Weston, A. Bordes, S. Chopra, A. M. Rush, B. Van Merriënboer, A. Joulin, and T. Mikolov, "TOWARDS AI-COMPLETE QUESTION ANSWERING: A SET OF PREREQUISITE TOY TASKS", 2015. [Online]. Available: https://arxiv.org/pdf/1502.05698.pdf.

[214] M. Tapaswi, Y. Zhu, R. Stiefelhagen, A. Torralba, R. Urtasun, and S. Fidler, "MovieQA: Understanding Stories in Movies through Question-Answering", Tech. Rep., 2016. [Online]. Available: https://arxiv.org/pdf/1512.02902.pdf.

[215] Y. Yang, W.-T. Yih, and C. Meek, "WIKIQA: A Challenge Dataset for Open-Domain Question Answering", in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Lisbon: Association for Computational Linguistics, 2015, pp. 2013–2018.

[216] P. Budzianowski, T.-H. Wen, B.-H. Tseng, I. Casanueva, S. Ultes, O. Ramadan, and M. Gašić, "MultiWOZ - A Large-Scale Multi-Domain Wizard-of-Oz Dataset for Task-Oriented Dialogue Modelling", 2019. DOI: 10.18653/v1/d18-1547.

[217] L. El Asri, H. Schulz, S. Sharma, J. Zumer, J. Harris, E. Fine, R. Mehrotra, and K. Suleman, "Frames: a corpus for adding memory to goal-oriented dialogue systems", 2018. DOI: 10.18653/v1/w17-5526.

[218] S. Gupta, R. Shah, M. Mohit, A. Kumar, and M. Lewis, "Semantic Parsing for Task Oriented Dialog using Hierarchical Representations", 2019. DOI: 10.18653/v1/d18-1300.

[219] M. Eric, L. Krishnan, F. Charette, and C. D. Manning, "Key-Value Retrieval Networks for Task-Oriented Dialogue", 2018. DOI: 10.18653/v1/w17-5506.

[220] Y. Zheng, G. Chen, M. Huang, S. Liu, and X. Zhu, "Personalized Dialogue Generation with Diversified Traits", Jan. 2019. [Online]. Available: http://arxiv.org/abs/1901.09672.

[221] R. Lowe, N. Pow, I. Serban, and J. Pineau, "The Ubuntu Dialogue Corpus: A Large Dataset for Research in Unstructured Multi-Turn Dialogue Systems", Jun. 2015. [Online]. Available: http://arxiv.org/abs/1506.08909.

[222] Y. Wu, W. Wu, C. Xing, Z. Li, and M. Zhou, "Sequential matching network: A new architecture for multi-turn response selection in retrieval-based chatbots", in *ACL 2017 - 55th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, 2017, ISBN: 9781945626753. DOI: 10.18653/v1/P17-1046.

[223] J. Tiedemann, "Parallel Data, Tools and Interfaces in OPUS.", *Lrec*, 2012, ISSN: 978-2-9517408-7-7. DOI: 978-2-9517408-7-7.

[224] J. Tiedemann, "News from OPUS — A Collection of Multilingual Parallel Corpora with Tools and Interfaces", *Recent Advances in Natural Language Processing*, 2009.

[225] N. Mrkšić, I. Vulić, D. Ó. Séaghdha, I. Leviant, R. Reichart, M. Gašić, A. Korhonen, and S. Young, "Semantic Specialization of Distributional Word Vector Spaces using Monolingual and Cross-Lingual Constraints", *Transactions of the Association for Computational Linguistics*, 2017. DOI: 10.1162/tacl{_}a{_}00063.

[226] T. H. Wen, D. Vandyke, N. Mrkšíc, M. Gašíc, L. M. Rojas-Barahona, P. H. Su, S. Ultes, and S. Young, "A network-based end-to-end trainable task-oriented dialogue system", in *15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017 - Proceedings of Conference*, 2017, ISBN: 9781510838604.

[227] C. T. Hemphill, J. J. Godfrey, and G. R. Doddington, "The ATIS spoken language systems pilot corpus", 1990. DOI: 10.3115/116580.116613.

[228] O. Dušek and F. Jurčíček, "A Context-aware Natural Language Generation Dataset for Dialogue Systems", in *RE-WOCHAT, LREC*, Portorož, 2016.

[229] J. Ganitkevitch, B. V. Durme, and C. Callison-Burch, "PPDB: The Paraphrase Database", 2013. [Online]. Available: http://paraphrase.org..

[230] E. Pavlick, P. Rastogi, J. Ganitkevitch, B. Van Durme, and C. Callison-Burch, "PPDB 2.0: Better paraphrase ranking, fine-grained entailment relations, word embeddings, and style classification", in *ACL-IJCNLP 2015 - 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, Proceedings of the Conference*, 2015, ISBN: 9781941643730.

[231] A. Fader, L. Zettlemoyer, and O. Etzioni, "Paraphrase-driven learning for open question answering", in *ACL 2013 - 51st Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, 2013, ISBN: 9781937284503.

[232] M. Potthast, B. Stein, A. Barrón-Cedeño, and P. Rosso, "An evaluation framework for plagiarism detection", in *Coling 2010 - 23rd International Conference on Computational Linguistics, Proceedings of the Conference*, 2010.

[233] M. Vila, M. A. Martí, and H. Rodríguez, *Paraphrase concept and typology. A linguistically based and computationally oriented approach*, 2011.

[234] A. Barrón-Cedeño, M. Vila, M. Antònia Martí, and P. Rosso, "Plagiarism Meets Paraphrasing: Insights for the Next Generation in Automatic Plagiarism Detection", *Computational Linguistics*, 2013, ISSN: 08912017. DOI: `10.1162/COLI{\_}a{\_}00153`.

[235] W. B. Dolan and C. Brockett, "Automatically Constructing a Corpus of Sentential Paraphrases", *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*, 2005.

[236] M. VILA, H. RODRIGUEZ, and M. ANTONIA MARTI, "WRPA: A System for Relational Paraphrase Acquisition from Wikipedia", *Procesamiento del lenguaje natural*, 2010, ISSN: 1135-5948.

[237] M. Vila, H. Rodríguez, and M. A. Martí, "Relational paraphrase acquisition from Wikipedia: The WRPA method and corpus", *Natural Language Engineering*, 2015, ISSN: 14698110. DOI: `10.1017/S1351324913000235`.

[238] A. Raux, B. Langner, A. W. Black, and M. Eskenazi, "LET'S GO: Improving spoken dialog systems for the elderly and non-natives", in *EUROSPEECH 2003 - 8th European Conference on Speech Communication and Technology*, 2003.

[239] A. I. Rudnicky, C. Bennett, A. W. Black, A. Chotomongcol, K. Lenzo, A. Oh, and R. Singh, "Task and domain specific modelling in the Carnegie Mellon Communicator system", in *6th International Conference on Spoken Language Processing, ICSLP 2000*, 2000, ISBN: 7801501144.

[240] J. F. Kelley, "An iterative design methodology for user-friendly natural language office information applications", *ACM Transactions on Information Systems (TOIS)*, 1984, ISSN: 15582868. DOI: `10.1145/357417.357420`.

[241] V. Rieser, I. Kruijff-Korbayová, and O. Lemon, "A corpus collection and annotation framework for learning multimodal clarification strategies", in *Proceedings of the 6th SIGdial Workshop on Discourse and Dialogue*, 2005.

[242] R. Chesney and D. K. Citron, "Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security", *SSRN Electronic Journal*, 2018. DOI: `10.2139/ssrn.3213954`.

[243] I. Solaiman, M. Brundage, J. Clark, A. Askell, A. Herbert-Voss, J. Wu, A. Radford, G. Krueger, J. W. Kim, S. Kreps, M. McCain, A. Newhouse, J. Blazakis, K. McGuffie, and J. Wang, "Release Strategies and the Social Impacts of Language Models", 2019. [Online]. Available: `http://arxiv.org/abs/1908.09203`.

[244] A. Clark and J. Haugeland, "Artificial Intelligence: The Very Idea.", *The Philosophical Quarterly*, vol. 38, no. 151, p. 249, Apr. 1988, ISSN: 00318094. DOI: `10.2307/2219930`. [Online]. Available: `https://academic.oup.com/pq/article-lookup/doi/10.2307/2219930`.

[245] F. Van Veen, "The neural network zoo", *The Asimov Institut*, pp. 1–22, 2016.

[246] J. Schmidhuber, *Deep Learning in neural networks: An overview*, 2015. DOI: 10.1016/j.neunet.2014.09.003.

[247] F. Rosenblatt, "The perceptron: a probabilistic model for information storage and organization in the brain.", *Psychological review*, 1958, ISSN: 0033-295X. DOI: 10.1037/h0042519.

[248] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition", *Proceedings of the IEEE*, 1998, ISSN: 00189219. DOI: 10.1109/5.726791.

[249] S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic Routing Between Capsules", Oct. 2017. [Online]. Available: http://arxiv.org/abs/1710.09829.

[250] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative Adversarial Nets", Tech. Rep., 2014. [Online]. Available: http://www.github.com/goodfeli/adversarial.

[251] C. Olah, *Understanding LSTM Networks*, 2015. [Online]. Available: http://colah.github.io/posts/2015-08-Understanding-LSTMs/.

[252] J. L. Elman, "Finding structure in time", *Cognitive Science*, 1990, ISSN: 03640213. DOI: 10.1016/0364-0213(90)90002-E.

[253] A. Sperduti and A. Starita, "Supervised neural networks for the classification of structures", *IEEE Transactions on Neural Networks*, 1997, ISSN: 10459227. DOI: 10.1109/72.572108.

[254] C. Goller and A. Kuechler, "Learning task-dependent distributed representations by backpropagation through structure", in *IEEE International Conference on Neural Networks - Conference Proceedings*, 1996. DOI: 10.1109/icnn.1996.548916.

[255] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory", *Neural Computation*, 1997, ISSN: 08997667. DOI: 10.1162/neco.1997.9.8.1735.

[256] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation", Tech. Rep., 2014. [Online]. Available: https://arxiv.org/pdf/1406.1078.pdf.

[257] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling", Dec. 2014. [Online]. Available: http://arxiv.org/abs/1412.3555.

[258] D. Britz, A. Goldie, M. T. Luong, and Q. V. Le, "Massive exploration of neural machine translation architectures", in *EMNLP 2017 - Conference on Empirical Methods in Natural Language Processing, Proceedings*, Association for Computational Linguistics (ACL), 2017, pp. 1442–1451, ISBN: 9781945626838. DOI: 10.18653/v1/d17-1151.

[259] G. Weiss, Y. Goldberg, and E. Yahav, "On the Practical Computational Power of Finite Precision RNNs for Language Recognition", May 2018. [Online]. Available: http://arxiv.org/abs/1805.04908.

[260] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, X. Zheng, and G. Research, "TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems", Tech. Rep., 2015. [Online]. Available: `http://download.tensorflow.org/paper/whitepaper2015.pdf`.

[261] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in PyTorch", in *31st Conference on Neural Information Processing Systems (NIPS 2017)*, 2017.

[262] Z. S. Harris, "Distributional Structure", *¡i¿WORD¡/i¿*, 1954, ISSN: 0043-7956. DOI: `10.1080/00437956.1954.11659520`.

[263] G. Salton, "Some experiments in the generation of word and document associations", in *AFIPS Conference Proceedings - 1962 Fall Joint Computer Conference, AFIPS 1962*, 1962. DOI: `10.1145/1461518.1461544`.

[264] K. S. Jones, *A statistical interpretation of term specificity and its application in retrieval*, 1972. DOI: `10.1108/eb026526`.

[265] G. Salton, A. Wong, and C. S. Yang, "A Vector Space Model for Automatic Indexing", *Communications of the ACM*, 1975, ISSN: 15577317. DOI: `10.1145/361219.361220`.

[266] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis", *Journal of the American Society for Information Science*, 1990, ISSN: 10974571. DOI: `10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9`.

[267] R. Kneser and H. Ney, "Improved backing-off for M-gram language modeling", in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 1995. DOI: `10.1109/icassp.1995.479394`.

[268] C. F. Baker, C. J. Fillmore, and J. B. Lowe, "The Berkeley FrameNet Project", Association for Computational Linguistics (ACL), 1998, p. 86. DOI: `10.3115/980845.980860`.

[269] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation", *Journal of Machine Learning Research*, 2003, ISSN: 15324435. DOI: `10.1016/b978-0-12-411519-4.00006-9`.

[270] Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin, "A Neural Probabilistic Language Model", in *Journal of Machine Learning Research*, vol. 3, Aug. 2003, pp. 1137–1155. DOI: `10.1162/153244303322533223`.

[271] T. Mikolov, M. Karafiát, L. Burget, C. Jan, and S. Khudanpur, "Recurrent neural network based language model", in *Proceedings of the 11th Annual Conference of the International Speech Communication Association, INTERSPEECH 2010*, 2010.

[272] A. Graves, N. Jaitly, and A. R. Mohamed, "Hybrid speech recognition with Deep Bidirectional LSTM", in *2013 IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU 2013 - Proceedings*, 2013, ISBN: 9781479927562. DOI: 10.1109/ASRU.2013.6707742.

[273] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space", in *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings*, 2013.

[274] J. Pennington, R. Socher, and C. Manning, "Glove: Global Vectors for Word Representation", in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, ISBN: 9781937284961. DOI: 10.3115/v1/D14-1162.

[275] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations ofwords and phrases and their compositionality", in *Advances in Neural Information Processing Systems*, 2013.

[276] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching Word Vectors with Subword Information", Tech. Rep., 2016. [Online]. Available: http://www.isthe.com/chongo/tech/comp/fnv.

[277] B. McCann, J. Bradbury, C. Xiong, and R. Socher, "Learned in translation: Contextualized word vectors", in *Advances in Neural Information Processing Systems*, 2017.

[278] J. Howard and S. Ruder, "Universal language model fine-tuning for text classification", in *ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, 2018, ISBN: 9781948087322. DOI: 10.18653/v1/p18-1031.

[279] M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep Contextualized Word Representations", 2018. DOI: 10.18653/v1/n18-1202.

[280] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to Sequence Learning with Neural Networks", 2014. [Online]. Available: https://arxiv.org/pdf/1409.3215.pdf.

[281] D. Bahdanau, K. H. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate", in *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 2015.

[282] M.-T. Luong, H. Pham, and C. D. Manning, "Effective Approaches to Attention-based Neural Machine Translation", 2015, ISSN: 10495258. DOI: 10.18653/v1/D15-1166.

[283] K. M. Hermann, T. Kočiský, E. Grefenstette, L. Espeholt, W. Kay, M. Suleyman, and P. Blunsom, "Teaching machines to read and comprehend", in *Advances in Neural Information Processing Systems*, vol. 2015-January, 2015.

[284] O. Vinyals, C. Blundell, T. Lillicrap, K. Kavukcuoglu, and D. Wierstra, "Matching networks for one shot learning", in *Advances in Neural Information Processing Systems*, 2016.

[285] K. Xu, J. L. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention", in *32nd International Conference on Machine Learning, ICML 2015*, vol. 3, 2015.

[286] S. Wiseman and A. M. Rush, "Sequence-to-sequence learning as beam-search optimization", in *EMNLP 2016 - Conference on Empirical Methods in Natural Language Processing, Proceedings*, 2016, ISBN: 9781945626258. DOI: `10.18653/v1/d16-1137`.

[287] J. Gehring, M. Auli, D. Grangier, D. Yarats, and Y. N. Dauphin, "Convolutional sequence to sequence learning", in *34th International Conference on Machine Learning, ICML 2017*, 2017, ISBN: 9781510855144.

[288] S. Merity, "Single Headed Attention RNN: Stop Thinking With Your Head", Tech. Rep., 2019.

[289] M. Suryavansh, *2019 — Year of BERT and Transformer*, 2019. [Online]. Available: `https://towardsdatascience.com/2019-year-of-bert-and-transformer-f200b53d05b9`.

[290] N. Latysheva, *2019: The Year of BERT*, 2019. [Online]. Available: `https://towardsdatascience.com/2019-the-year-of-bert-354e8106f7ba`.

[291] D. M. Ziegler, N. Stiennon, J. Wu, T. B. Brown, A. Radford, D. Amodei, P. Christiano, and G. Irving, "Fine-Tuning Language Models from Human Preferences", Sep. 2019. [Online]. Available: `http://arxiv.org/abs/1909.08593`.

[292] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention Is All You Need", 2017, ISSN: 0140-525X. DOI: `10.1017/S0140525X16001837`.

[293] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving Language Understanding by Generative Pre-Training", *OpenAI*, 2018. [Online]. Available: `https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf`.

[294] J. Devlin, M.-W. Chang, K. Lee, K. T. Google, and A. I. Language, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", Tech. Rep., 2019.

[295] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler, "Aligning books and movies: Towards story-like visual explanations by watching movies and reading books", in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, ISBN: 9781467383912. DOI: `10.1109/ICCV.2015.11`.

[296] G. Lample and A. Conneau, "Cross-lingual Language Model Pretraining", Tech. Rep., 2019.

[297] Z. Dai, Z. Yang, Y. Yang, J. Carbonell, Q. V. Le, and R. Salakhutdinov, "Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context", Jan. 2019. [Online]. Available: `http://arxiv.org/abs/1901.02860`.

[298] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "language model and unsupervised multitask learning", *OpenAI*, 2018. [Online]. Available: `https://d4mucfpksywv.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf`.

[299] *Better Language Models and Their Implications*, 2019. [Online]. Available: `https://openai.com/blog/better-language-models/`.

[300] *GPT-2: 6-Month Follow-Up*, 2019. [Online]. Available: https://openai.com/blog/gpt-2-6-month-follow-up/.

[301] *GPT-2: 1.5B Release*, 2019. [Online]. Available: https://openai.com/blog/gpt-2-1-5b-release/.

[302] Y. Sun, S. Wang, Y. Li, S. Feng, X. Chen, H. Zhang, X. Tian, D. Zhu, H. Tian, and H. Wu, "ERNIE: Enhanced Representation through Knowledge Integration", Apr. 2019. [Online]. Available: http://arxiv.org/abs/1904.09223.

[303] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le, "XLNet: Generalized Autoregressive Pretraining for Language Understanding", 2019. [Online]. Available: https://github.com/zihangdai/xlnet.

[304] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A Robustly Optimized BERT Pretraining Approach", Jul. 2019. [Online]. Available: http://arxiv.org/abs/1907.11692.

[305] Y. Sun, S. Wang, Y. Li, S. Feng, H. Tian, H. Wu, and H. Wang, "ERNIE 2.0: A Continual Pre-training Framework for Language Understanding", Jul. 2019. [Online]. Available: http://arxiv.org/abs/1907.12412.

[306] N. S. Keskar, B. McCann, L. R. Varshney, C. Xiong, and R. Socher, "CTRL: A Conditional Transformer Language Model for Controllable Generation", Sep. 2019. [Online]. Available: http://arxiv.org/abs/1909.05858.

[307] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "ALBERT: A Lite BERT for Self-supervised Learning of Language Representations", 2019. [Online]. Available: http://arxiv.org/abs/1909.11942.

[308] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter", 2019. [Online]. Available: http://arxiv.org/abs/1910.01108.

[309] S. Ruder, M. E. Peters, S. Swayamdipta, and T. Wolf, "Transfer Learning in Natural Language Processing", Association for Computational Linguistics (ACL), Jul. 2019, pp. 15–18. DOI: 10.18653/v1/n19-5004. [Online]. Available: https://www.aclweb.org/anthology/N19-5004.pdf.

[310] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, and J. Brew, "HuggingFace's Transformers: State-of-the-art Natural Language Processing", Oct. 2019. [Online]. Available: http://arxiv.org/abs/1910.03771.

[311] L. Martin, B. Muller, P. J. O. Suárez, Y. Dupont, L. Romary, É. V. de la Clergerie, D. Seddah, and B. Sagot, "CamemBERT: a Tasty French Language Model", Nov. 2019. [Online]. Available: http://arxiv.org/abs/1911.03894.

[312] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer", Oct. 2019. [Online]. Available: http://arxiv.org/abs/1910.10683.

[313] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised Cross-lingual Representation Learning at Scale", Tech. Rep. [Online]. Available: https://github.com/facebookresearch/.

[314] D. Kiela, S. Bhooshan, H. Firooz, and D. Testuggine, “Supervised Multimodal Bitransformers for Classifying Images and Text”, Sep. 2019. [Online]. Available: `http://arxiv.org/abs/1909.02950`.

[315] S. Han, H. Mao, and W. J. Dally, “Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding”, in *4th International Conference on Learning Representations, ICLR 2016 - Conference Track Proceedings*, International Conference on Learning Representations, ICLR, 2016.

[316] S. Han, J. Pool, J. Tran, and W. J. Dally, “Learning both Weights and Connections for Efficient Neural Networks”, Jun. 2015. [Online]. Available: `http://arxiv.org/abs/1506.02626`.

[317] T. Gale, E. Elsen, and S. Hooker, “The State of Sparsity in Deep Neural Networks”, Feb. 2019. [Online]. Available: `http://arxiv.org/abs/1902.09574`.

[318] G. Hinton and J. Dean, “Distilling the Knowledge in a Neural Network”, Tech. Rep., 2015. [Online]. Available: `https://arxiv.org/pdf/1503.02531.pdf`.

[319] D. Shulga, *Distilling BERT - How to achieve BERT performance using logistic regression*, 2018. [Online]. Available: `https://towardsdatascience.com/distilling-bert-how-to-achieve-bert-performance-using-logistic-regression-69a7fc14249d`.

[320] X. Liu, X. Wang, and S. Matwin, “Improving the Interpretability of Deep Neural Networks with Knowledge Distillation”, Dec. 2018. [Online]. Available: `http://arxiv.org/abs/1812.10924`.

[321] S. Ravi, *Custom On-Device ML Models with Learn2Compress*, 2018. [Online]. Available: `https://ai.googleblog.com/2018/05/custom-on-device-ml-models.html`.

[322] S.-I. Mirzadeh, M. Farajtabar, A. Li, N. Levine, A. Matsukawa, and H. Ghasemzadeh, “Improved Knowledge Distillation via Teacher Assistant”, Feb. 2019. [Online]. Available: `http://arxiv.org/abs/1902.03393`.

[323] S. Khan, *BERT, RoBERTa, DistilBERT, XLNet — which one to use?*, 2019. [Online]. Available: `https://towardsdatascience.com/bert-roberta-distilbert-xlnet-which-one-to-use-3d5ab82ba5f8`.

[324] D. Jurafsky and J. H. Martin, “Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition”, *Speech and Language Processing An Introduction to Natural Language Processing Computational Linguistics and Speech Recognition*, 2009, ISSN: 08912017. DOI: `10.1162/089120100750105975`.

[325] S. Jafarpour and C. J. C. Burges, “Filter, Rank, and Transfer the Knowledge: Learning to Chat”, *Learning*, 2010.

[326] A. Leuski and D. Traum, “NPCEditor: Creating virtual human dialogue using information retrieval techniques”, *AI Magazine*, 2011, ISSN: 07384602. DOI: `10.1609/aimag.v32i2.2347`.

[327] A. Ritter, C. Cherry, and W. B. Dolan, “Data-driven response generation in social media”, in *EMNLP 2011 - Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, 2011, ISBN: 1937284115.

[328] B. Hu, Z. Lu, H. Li, and Q. Chen, "Convolutional neural network architectures for matching natural language sentences", in *Advances in Neural Information Processing Systems*, 2014.

[329] X. Qiu and X. Huang, "Convolutional neural tensor network architecture for community-based question answering", in *IJCAI International Joint Conference on Artificial Intelligence*, 2015, ISBN: 9781577357384.

[330] M. Feng, B. Xiang, M. R. Glass, L. Wang, and B. Zhou, "Applying deep learning to answer selection: A study and an open task", in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU 2015 - Proceedings*, 2016, ISBN: 9781479972913. DOI: `10.1109/ASRU.2015.7404872`.

[331] M. Tan, C. D. Santos, B. Xiang, and B. Zhou, "Improved representation learning for question answer matching", in *54th Annual Meeting of the Association for Computational Linguistics, ACL 2016 - Long Papers*, 2016, ISBN: 9781510827585. DOI: `10.18653/v1/p16-1044`.

[332] B. Wang, K. Liu, and J. Zhao, "Inner Attention based recurrent neural networks for answer selection", in *54th Annual Meeting of the Association for Computational Linguistics, ACL 2016 - Long Papers*, 2016, ISBN: 9781510827585. DOI: `10.18653/v1/p16-1122`.

[333] Z. Lu and H. Li, "A deep architecture for matching short texts", in *Advances in Neural Information Processing Systems*, Neural information processing systems foundation, 2013.

[334] L. Pang, Y. Lan, J. Guo, J. Xu, S. Wan, and X. Cheng, "Text matching as image recognition", in *30th AAAI Conference on Artificial Intelligence, AAAI 2016*, 2016, ISBN: 9781577357605.

[335] S. Wang and J. Jiang, "Learning natural language inference with LSTM", in *2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT 2016 - Proceedings of the Conference*, 2016, ISBN: 9781941643914. DOI: `10.18653/v1/n16-1170`.

[336] S. Wan, Y. Lan, J. Guo, J. Xu, L. Pang, and X. Cheng, "A deep architecture for semantic matching with multiple positional sentence representations", in *30th AAAI Conference on Artificial Intelligence, AAAI 2016*, 2016, ISBN: 9781577357605.

[337] R. Yan, Y. Song, and H. Wu, "Learning to respond with deep neural networks for retrieval-based human-computer conversation system", in *SIGIR 2016 - Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2016, ISBN: 9781450342902. DOI: `10.1145/2911451.2911542`.

[338] X. Zhou, D. Dong, H. Wu, S. Zhao, D. Yu, H. Tian, X. Liu, and R. Yan, "Multi-view response selection for human-computer conversation", in *EMNLP 2016 - Conference on Empirical Methods in Natural Language Processing, Proceedings*, 2016, ISBN: 9781945626258. DOI: `10.18653/v1/d16-1036`.

[339] R. Yan and D. Zhao, "Coupled context modeling for deep chit-chat: Towards conversations between human and computer", in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Association for Computing Machinery, Jul. 2018, pp. 2574–2583, ISBN: 9781450355520. DOI: `10.1145/3219819.3220045`.

[340] X. Zhou, L. Li, D. Dong, Y. Liu, Y. Chen, W. X. Zhao, D. Yu, and H. Wu, "Multi-turn response selection for chatbots with deep attention matching network", in *ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, 2018, ISBN: 9781948087322. DOI: 10.18653/v1/p18-1103.

[341] Y. Wu, W. Wu, C. Xing, C. Xu, Z. Li, and M. Zhou, "A sequential matching framework for multi-turn response selection in retrieval-based chatbots", *Computational Linguistics*, 2019, ISSN: 15309312. DOI: 10.1162/coli{_}a{_}00345.

[342] K. Yao, G. Zweig, and B. Peng, "Attention with Intention for a Neural Network Conversation Model", Oct. 2015. [Online]. Available: http://arxiv.org/abs/1510.08565.

[343] L. Shang, Z. Lu, and H. Li, "Neural responding machine for short-Text conversation", in *ACL-IJCNLP 2015 - 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, Proceedings of the Conference*, 2015, ISBN: 9781941643723. DOI: 10.3115/v1/p15-1152.

[344] A. Sordoni, M. Galley, M. Auli, C. Brockett, Y. Ji, M. Mitchell, J.-Y. Nie, J. Gao, and B. Dolan, "A Neural Network Approach to Context-Sensitive Generation of Conversational Responses", 2015.

[345] A. Sordoni, Y. Bengio, H. Vahabi, C. Lioma, J. G. Simonsen, and J.-Y. Nie, "A Hierarchical Recurrent Encoder-Decoder For Generative Context-Aware Query Suggestion", 2015. DOI: 10.1145/2806416.2806493.

[346] I. V. Serban, A. Sordoni, Y. Bengio, A. Courville, and J. Pineau, "Building End-To-End Dialogue Systems Using Generative Hierarchical Neural Network Models", Jul. 2015. [Online]. Available: https://arxiv.org/abs/1507.04808.

[347] Y. Zhang, S. Sun, M. Galley, Y.-C. Chen, C. Brockett, X. Gao, J. Gao, J. Liu, and B. Dolan, "DialoGPT: Large-Scale Generative Pre-training for Conversational Response Generation", Nov. 2019. [Online]. Available: http://arxiv.org/abs/1911.00536.

[348] A. Bordes, Y. Lan Boureau, and J. Weston, "Learning end-to-end goal-oriented dialog", in *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*, 2017.

[349] J. Weston, S. Chopra, and A. Bordes, "Memory networks", in *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 2015.

[350] X. Li, Y.-N. Chen, L. Li, and J. Gao, "End-to-End Task-Completion Neural Dialogue Systems", Tech. Rep., 2017. [Online]. Available: http://github.com/MiuLab/TC-Bot.

[351] B. Dhingra, L. Li, X. Li, J. Gao, Y.-N. Chen, F. Ahmed, and L. Deng, "Towards End-to-End Reinforcement Learning of Dialogue Agents for Information Access", 2016. DOI: 10.18653/v1/P17-1045.

[352] B. Kim, K. Chung, J. Lee, J. Seo, and M.-w. Koo, "End-to-End Goal-Oriented Dialog Learning Based On Memory Network", *DSTC6 Conference*, 2017.

[353] S. Sukhbaatar, A. Szlam, J. Weston, and R. Fergus, "End-To-End Memory Networks", 2015, ISSN: 10495258. DOI: v5.

[354] A. Madotto, C. S. Wu, and P. Fung, "MEM2Seq: Effectively incorporating knowledge bases into end-to-end task-oriented dialog systems", in *ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, 2018, ISBN: 9781948087322. DOI: 10.18653/v1/p18-1136.

[355] B. Kim, J. Lee, J. Seo, M. W. Koo, and K. T. Chung, "A Bi-LSTM memory network for end-to-end goal-oriented dialog learning", *Computer Speech and Language*, 2019. DOI: 10.1016/j.csl.2018.06.005.

[356] B. Liu and I. Lane, "End-to-End Learning of Task-Oriented Dialogs", 2018. DOI: 10.18653/v1/n18-4010.

[357] B. Liu, G. Tür, D. Hakkani-Tür, P. Shah, and L. Heck, "Dialogue Learning with Human Teaching and Feedback in End-to-End Trainable Task-Oriented Dialogue Systems", Association for Computational Linguistics (ACL), May 2018, pp. 2060–2069. DOI: 10.18653/v1/n18-1187.

[358] H. Wen, Y. Liu, W. Che, L. Qin, and T. Liu, "Sequence-to-Sequence Learning for Task-oriented Dialogue with Dialogue State Representation", Jun. 2018. [Online]. Available: http://arxiv.org/abs/1806.04441.

[359] J. Ohmura and M. Eskenazi, "Context-Aware Dialog Re-Ranking for Task-Oriented Dialog Systems", in *2018 IEEE Spoken Language Technology Workshop, SLT 2018 - Proceedings*, 2019, ISBN: 9781538643341. DOI: 10.1109/SLT.2018.8639596.

[360] P. Budzianowski and I. Vulić, "Hello, It's GPT-2 - How Can I Help You? Towards the Use of Pretrained Language Models for Task-Oriented Dialogue Systems", Association for Computational Linguistics (ACL), Nov. 2019, pp. 15–22. DOI: 10.18653/v1/d19-5602.

[361] T. Wolf, V. Sanh, J. Chaumond, and C. Delangue, "TransferTransfo: A Transfer Learning Approach for Neural Network Based Conversational Agents", Jan. 2019. [Online]. Available: http://arxiv.org/abs/1901.08149.

[362] S. Golovanov, R. Kurbanov, S. Nikolenko, K. Truskovskyi, A. Tselousov, and T. Wolf, "Large-Scale Transfer Learning for Natural Language Generation", 2019. DOI: 10.18653/v1/p19-1608.

[363] L. Pack Kaelbling, M. L. Littman, A. W. Moore, and S. Hall, "Reinforcement Learning: A Survey", *Journal of Artiicial Intelligence Research*, vol. 4, pp. 237–285, 1996. [Online]. Available: https://arxiv.org/pdf/cs/9605103.pdf.

[364] J. Li, M. Galley, C. Brockett, J. Gao, and B. Dolan, "A Diversity-Promoting Objective Function for Neural Conversation Models", Tech. Rep., 2016. [Online]. Available: https://arxiv.org/pdf/1510.03055.pdf.

[365] J. Li, W. Monroe, A. Ritter, M. Galley, J. Gao, and D. Jurafsky, "Deep reinforcement learning for dialogue generation", in *EMNLP 2016 - Conference on Empirical Methods in Natural Language Processing, Proceedings*, 2016, ISBN: 9781945626258. DOI: 10.18653/v1/d16-1127.

[366] T. Zhao, K. Xie, and M. Eskenazi, "Rethinking Action Spaces for Reinforcement Learning in End-to-end Dialog Agents with Latent Variable Models", pp. 1208–1218, 2019. DOI: 10.18653/V1/N19-1123.

[367] H. Cuayáhuitl, D. Lee, S. Ryu, S. Choi, I. Hwang, and J. Kim, “Deep Reinforcement Learning for Chatbots Using Clustered Actions and Human-Likeness Rewards”, *Proceedings of the International Joint Conference on Neural Networks*, vol. 2019-July, Aug. 2019. [Online]. Available: `http://arxiv.org/abs/1908.10331`.

[368] H. Cuayáhuitl, D. Lee, S. Ryu, Y. Cho, S. Choi, S. Indurthi, S. Yu, H. Choi, I. Hwang, and J. Kim, “Ensemble-Based Deep Reinforcement Learning for Chatbots”, *Neurocomputing*, vol. 366, pp. 118–130, Aug. 2019. DOI: `10.1016/j.neucom.2019.08.007`. [Online]. Available: `http://arxiv.org/abs/1908.10422%20http://dx.doi.org/10.1016/j.neucom.2019.08.007`.

[369] C. Sankar and S. Ravi, “Deep Reinforcement Learning For Modeling Chit-Chat Dialog With Discrete Attributes”, 2019. DOI: `10.18653/v1/w19-5901`.

[370] I. Casanueva, P. Budzianowski, P.-H. Su, N. Mrkšić, T.-H. Wen, S. Ultes, L. Rojas-Barahona, S. Young, and M. Gašić, “A Benchmarking Environment for Reinforcement Learning Based Task Oriented Dialogue Management”, Nov. 2017. [Online]. Available: `http://arxiv.org/abs/1711.11023`.

[371] H. Cuayáhuitl, “SimpleDS: A Simple Deep Reinforcement Learning Dialogue System”, Jan. 2016. [Online]. Available: `http://arxiv.org/abs/1601.04574`.

[372] T. Zhao and M. Eskenazi, “Towards End-to-End Learning for Dialog State Tracking and Management using Deep Reinforcement Learning”, Association for Computational Linguistics (ACL), Dec. 2016, pp. 1–10. DOI: `10.18653/v1/w16-3601`.

[373] I. V. Serban, C. Sankar, M. Germain, S. Zhang, Z. Lin, S. Subramanian, T. Kim, M. Pieper, S. Chandar, N. R. Ke, S. Rajeshwar, A. de Brebisson, J. M. R. Sotelo, D. Suhubdy, V. Michalski, A. Nguyen, J. Pineau, and Y. Bengio, “A Deep Reinforcement Learning Chatbot”, Sep. 2017. [Online]. Available: `http://arxiv.org/abs/1709.02349`.

[374] B. Liu, G. Tur, D. Hakkani-Tur, P. Shah, and L. Heck, “End-to-End Optimization of Task-Oriented Dialogue Model with Deep Reinforcement Learning”, Nov. 2017. [Online]. Available: `http://arxiv.org/abs/1711.10712`.

[375] L. Chen, Z. Chen, B. Tan, S. Long, M. Gasic, and K. Yu, “AgentGraph: Towards Universal Dialogue Management with Structured Deep Reinforcement Learning”, *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 27, no. 9, pp. 1378–1391, May 2019. [Online]. Available: `http://arxiv.org/abs/1905.11259`.

[376] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini, “The graph neural network model”, *IEEE Transactions on Neural Networks*, 2009, ISSN: 10459227. DOI: `10.1109/TNN.2008.2005605`.

[377] A. Malte and P. Ratadiya, “Evolution of Transfer Learning in Natural Language Processing”, Tech. Rep., 2019.

[378] Y.-P. Ruan, Z.-H. Ling, J.-C. Gu, and Q. Liu, “Fine-Tuning BERT for Schema-Guided Zero-Shot Dialogue State Tracking”, Feb. 2020. [Online]. Available: `http://arxiv.org/abs/2002.00181`.

[379] Y. Xian, C. H. Lampert, B. Schiele, and Z. Akata, “Zero-Shot Learning - A Comprehensive Evaluation of the Good, the Bad and the Ugly”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Jul. 2017. [Online]. Available: `http://arxiv.org/abs/1707.00600`.

[380] S. Schuster, S. Gupta, R. Shah, and M. Lewis, "Cross-lingual Transfer Learning for Multilingual Task Oriented Dialog", 2019. DOI: 10.18653/v1/n19-1380.

[381] Z. Liu, J. Shin, Y. Xu, G. I. Winata, P. Xu, A. Madotto, and P. Fung, "Zero-shot Cross-lingual Dialogue Systems with Transferable Latent Variables", 2019. DOI: 10.18653/v1/d19-1129.

[382] S. Tong and D. Koller, "Support vector machine active learning with applications to text classification", *Journal of Machine Learning Research*, pp. 45–66, 2001. DOI: 10.1162/153244302760185243.

[383] A. Carlson, J. Betteridge, B. Kisiel, B. Settles, E. R. Hruschka, and T. M. Mitchell, "Toward an architecture for never-ending language learning", in *Proceedings of the National Conference on Artificial Intelligence*, 2010, ISBN: 9781577354666.

[384] D. L. Silver, Q. Yang, and L. Li, "Lifelong machine learning systems: Beyond learning algorithms", in *AAAI Spring Symposium - Technical Report*, 2013, ISBN: 9781577356028.

[385] B. Hancock, A. Bordes, P.-E. Mazaré, and J. Weston, "Learning from Dialogue after Deployment: Feed Yourself, Chatbot!", Jan. 2019. [Online]. Available: http://arxiv.org/abs/1901.05415.

[386] N. Asghar, P. Poupart, X. Jiang, and H. Li, "Deep active learning for dialogue generation", in **SEM 2017 - 6th Joint Conference on Lexical and Computational Semantics, Proceedings*, 2017, ISBN: 9781945626531. DOI: 10.18653/v1/s17-1008.

[387] Y. Zhang, X. Gao, S. Lee, C. Brockett, M. Galley, J. Gao, and B. Dolan, "Consistent Dialogue Generation with Self-supervised Feature Learning", Mar. 2019. [Online]. Available: http://arxiv.org/abs/1903.05759.

[388] T. Niu and M. Bansal, "Adversarial over-sensitivity and over-stability strategies for dialogue models", in *CoNLL 2018 - 22nd Conference on Computational Natural Language Learning, Proceedings*, 2018, ISBN: 9781948087728. DOI: 10.18653/v1/k18-1047.

[389] O. Olabiyi, A. Salimov, A. Khazane, and E. T. Mueller, "Multi-turn Dialogue Response Generation in an Adversarial Learning Framework", pp. 121–132, May 2018. [Online]. Available: http://arxiv.org/abs/1805.11752.

[390] O. O. Olabiyi, A. Khazane, and E. T. Mueller, "A Persona-based Multi-turn Conversation Model in an Adversarial Learning Framework", *Proceedings - 17th IEEE International Conference on Machine Learning and Applications, ICMLA 2018*, pp. 489–494, Apr. 2019. [Online]. Available: http://arxiv.org/abs/1905.01998.

[391] O. Olabiyi, A. Khazane, A. Salimov, and E. T. Mueller, "An Adversarial Learning Framework For A Persona-Based Multi-Turn Dialogue Model", pp. 1–10, Apr. 2019. [Online]. Available: http://arxiv.org/abs/1905.01992.

[392] P. Liu, X. Qiu, and X. Huang, "Adversarial multi-task learning for text classification", in *ACL 2017 - 55th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, 2017, ISBN: 9781945626753. DOI: 10.18653/v1/P17-1001.

[393] J. Li, W. Monroe, T. Shi, S. Jean, A. Ritter, and D. Jurafsky, "Adversarial Learning for Neural Dialogue Generation", 2018. DOI: 10.18653/v1/d17-1230.

[394] S. Cui, R. Lian, D. Jiang, Y. Song, S. Bao, and Y. Jiang, "DAL: Dual Adversarial Learning for Dialogue Generation", Jun. 2019. [Online]. Available: `http://arxiv.org/abs/1906.09556`.

[395] I. V. Serban, A. Sordoni, R. Lowe, L. Charlin, J. Pineau, A. Courville, and Y. Bengio, "A hierarchical latent variable encoder-decoder model for generating dialogues", in *31st AAAI Conference on Artificial Intelligence, AAAI 2017*, 2017.

[396] H. He, A. Balakrishnan, M. Eric, and P. Liang, "Learning symmetric collaborative dialogue agents with dynamic knowledge graph embeddings", in *ACL 2017 - 55th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, 2017, ISBN: 9781945626753. DOI: `10.18653/v1/P17-1162`.

[397] Y. Zhang, M. Galley, J. Gao, Z. Gan, X. Li, C. Brockett, and B. Dolan, "Generating Informative and Diverse Conversational Responses via Adversarial Information Maximization", *Advances in Neural Information Processing Systems*, vol. 2018-December, pp. 1810–1820, Sep. 2018. [Online]. Available: `http://arxiv.org/abs/1809.05972`.

[398] J. Ham, S. Lim, K. H. Lee, and K. E. Kim, "Extensions to hybrid code networks for FAIR dialog dataset", *Computer Speech and Language*, 2019, ISSN: 10958363. DOI: `10.1016/j.csl.2018.07.004`.

[399] Y. Shin, K. Min Yoo, and S. G. Lee, "Slot filling with delexicalized sentence generation", in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2018. DOI: `10.21437/Interspeech.2018-1808`.

[400] S. Sharma, J. He, K. Suleman, H. Schulz, and P. Bachman, "Natural Language Generation in Dialogue using Lexicalized and Delexicalized Data", Jun. 2016. [Online]. Available: `http://arxiv.org/abs/1606.03632`.

[401] C. Smiley, E. Davoodi, D. Song, and F. Schilder, "The E2E NLG Challenge: A Tale of Two Systems", 2019. DOI: `10.18653/v1/w18-6558`.

[402] G. Tur and R. De Mori, *Spoken Language Understanding: Systems for Extracting Semantic Information from Speech.* 2011, ISBN: 9780470688243. DOI: `10.1002/9781119992691`.

[403] P. Haffner, G. Tur, and J. H. Wright, "Optimizing SVMs for complex call classification", in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2003. DOI: `10.1109/icassp.2003.1198860`.

[404] C. Raymond and G. Riccardi, "Generative and discriminative algorithms for Spoken Language Understanding", in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2007, ISBN: 9781605603162.

[405] N. T. Vu, "Sequential convolutional neural networks for slot filling in spoken language understanding", in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2016. DOI: `10.21437/Interspeech.2016-395`.

[406] K. Yao, B. Peng, Y. Zhang, D. Yu, G. Zweig, and Y. Shi, "Spoken language understanding using long short-term memory neural networks", in *2014 IEEE Workshop on Spoken Language Technology, SLT 2014 - Proceedings*, 2014, ISBN: 9781479971299. DOI: 10.1109/SLT.2014.7078572.

[407] B. Peng and K. Yao, "Recurrent Neural Networks with External Memory for Language Understanding", May 2015. [Online]. Available: http://arxiv.org/abs/1506.00195.

[408] G. Kurata, B. Zhou, B. Xiang, and M. Yu, "Leveraging sentence-level information with encoder LSTM for semantic slot filling", in *EMNLP 2016 - Conference on Empirical Methods in Natural Language Processing, Proceedings*, 2016, ISBN: 9781945626258. DOI: 10.18653/v1/d16-1223.

[409] L. Zhao and Z. Feng, "Improving slot filling in spoken language understanding with joint pointer and attention", in *ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, 2018, ISBN: 9781948087346. DOI: 10.18653/v1/p18-2068.

[410] Y. Kim, "Convolutional neural networks for sentence classification", in *EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, 2014, ISBN: 9781937284961. DOI: 10.3115/v1/d14-1181.

[411] X. Zhang, J. Zhao, and Y. Lecun, "Character-level convolutional networks for text classification", in *Advances in Neural Information Processing Systems*, 2015.

[412] S. Ravuri and A. Stolcke, "Recurrent neural network and LSTM models for lexical utterance classification", in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2015.

[413] Z. Zhao and Y. Wu, "Attention-Based Convolutional Neural Networks for Sentence Classification", Sep. 2016, pp. 705–709. DOI: 10.21437/Interspeech.2016-354. [Online]. Available: http://www.isca-speech.org/archive/Interspeech_2016/abstracts/0354.html.

[414] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical attention networks for document classification", in *2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT 2016 - Proceedings of the Conference*, 2016, ISBN: 9781941643914. DOI: 10.18653/v1/n16-1174.

[415] C. Xia, C. Zhang, X. Yan, Y. Chang, and P. Yu, "Zero-shot User Intent Detection via Capsule Neural Networks", Association for Computational Linguistics (ACL), Jun. 2019, pp. 3090–3099. DOI: 10.18653/v1/d18-1348.

[416] P. Xu and R. Sarikaya, "Convolutional neural network based triangular CRF for joint intent detection and slot filling", in *2013 IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU 2013 - Proceedings*, 2013, ISBN: 9781479927562. DOI: 10.1109/ASRU.2013.6707709.

[417] D. Z. Guo, G. Tur, W. T. Yih, and G. Zweig, "Joint semantic utterance classification and slot filling with recursive neural networks", in *2014 IEEE Workshop on Spoken Language Technology, SLT 2014 - Proceedings*, 2014, ISBN: 9781479971299. DOI: 10.1109/SLT.2014.7078634.

[418] D. Hakkani-Tür, G. Tur, A. Celikyilmaz, Y. N. Chen, J. Gao, L. Deng, and Y. Y. Wang, "Multi-domain joint semantic frame parsing using bi-directional RNN-LSTM", in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2016. DOI: `10.21437/Interspeech.2016-402`.

[419] C.-W. Goo, G. Gao, Y.-K. Hsu, C.-L. Huo, T.-C. Chen, K.-W. Hsu, and Y.-N. Chen, "Slot-Gated Modeling for Joint Slot Filling and Intent Prediction", 2018. DOI: `10.18653/v1/n18-2118`.

[420] C. Zhang, Y. Li, N. Du, W. Fan, and P. S. Yu, "Joint Slot Filling and Intent Detection via Capsule Neural Networks", Dec. 2018. [Online]. Available: `http://arxiv.org/abs/1812.09471`.

[421] Q. Chen, Z. Zhuo, and W. Wang, "BERT for Joint Intent Classification and Slot Filling", Feb. 2019. [Online]. Available: `http://arxiv.org/abs/1902.10909`.

[422] G. Castellucci, V. Bellomaria, A. Favalli, and R. Romagnoli, "Multi-lingual Intent Detection and Slot Filling in a Joint BERT-based Model", Jul. 2019. [Online]. Available: `http://arxiv.org/abs/1907.02884`.

[423] Z. Wang and O. Lemon, "A simple and generic belief tracking mechanism for the dialog state tracking challenge: On the believability of observed information", in *SIGDIAL 2013 - 14th Annual Meeting of the Special Interest Group on Discourse and Dialogue, Proceedings of the Conference*, 2013, ISBN: 9781937284954.

[424] K. Sun, L. Chen, S. Zhu, and K. Yu, "A generalized rule based tracker for dialogue state tracking", in *2014 IEEE Workshop on Spoken Language Technology, SLT 2014 - Proceedings*, 2014, ISBN: 9781479971299. DOI: `10.1109/SLT.2014.7078596`.

[425] B. Zhang, Q. Cai, J. Mao, E. Chang, and B. Guo, "Spoken dialogue management as planning and acting under uncertainty", in *EUROSPEECH 2001 - SCANDINAVIA - 7th European Conference on Speech Communication and Technology*, 2001, ISBN: 8790834100. DOI: `10.1109/icmlc.2002.1174420`.

[426] H. M. Meng, C. Wai, and R. Pieraccini, "The Use of Belief Networks for Mixed-Initiative Dialog Modeling", *IEEE Transactions on Speech and Audio Processing*, 2003, ISSN: 10636676. DOI: `10.1109/TSA.2003.814380`.

[427] J. D. Williams, P. Poupart, and S. Young, "Factored Partially Observable Markov Decision Processes for Dialogue Management", Tech. Rep., 2005. [Online]. Available: `https://cs.uwaterloo.ca/~ppoupart/publications/spoken-dialog-fact-pomdp/spoken-dialog-fact-pomdp.pdf`.

[428] D. DeVault and M. Stone, "Managing ambiguities across utterances in dialogue", *Decalog 2007*, 2007.

[429] D. DeVault, "Contribution tracking: Participating in task-oriented dialogue under uncertainty", *ProQuest Dissertations and Theses*, 2008.

[430] M. Gašić and S. Young, "Effective handling of dialogue state in the hidden information state POMDP-based dialogue manager", *ACM Transactions on Speech and Language Processing*, 2011, ISSN: 15504875. DOI: `10.1145/1966407.1966409`.

[431] J. D. Williams and S. Young, "Partially observable Markov decision processes for spoken dialog systems", *Computer Speech and Language*, 2007, ISSN: 08852308. DOI: `10.1016/j.csl.2006.06.008`.

[432] T. H. Bui, M. Poel, A. Nijholt, and J. Zwiers, "A tractable hybrid DDN-POMDP approach to affective dialogue modeling for probabilistic frame-based dialogue systems", *Natural Language Engineering*, 2009, ISSN: 13513249. DOI: 10.1017/S1351324908005032.

[433] H. Ren, W. Xu, Y. Zhang, and Y. Yan, "Dialog state tracking using conditional random fields", in *SIGDIAL 2013 - 14th Annual Meeting of the Special Interest Group on Discourse and Dialogue, Proceedings of the Conference*, Association for Computational Linguistics (ACL), 2013, pp. 457–461, ISBN: 9781937284954.

[434] J. D. Williams, "Multi-domain learning and generalization in dialog state tracking", in *SIGDIAL 2013 - 14th Annual Meeting of the Special Interest Group on Discourse and Dialogue, Proceedings of the Conference*, 2013, ISBN: 9781937284954.

[435] S. Lee and M. Eskenazi, "Recipe for building robust spoken dialog state trackers: Dialog state tracking challenge system description", in *SIGDIAL 2013 - 14th Annual Meeting of the Special Interest Group on Discourse and Dialogue, Proceedings of the Conference*, 2013, ISBN: 9781937284954.

[436] M. Henderson, B. Thomson, and S. Young, "Robust dialog state tracking using delexicalised recurrent neural networks and unsupervised adaptation", in *2014 IEEE Workshop on Spoken Language Technology, SLT 2014 - Proceedings*, 2014, ISBN: 9781479971299. DOI: 10.1109/SLT.2014.7078601.

[437] Z. Yan, N. Duan, P. Chen, M. Zhou, J. Zhou, and Z. Li, "Building Task-Oriented Dialogue Systems for Online Shopping", in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI)*, 2017. DOI: 10.1088/0264-9381/19/7/347.

[438] H. Cuayáhuitl, S. Keizer, and O. Lemon, "Strategic Dialogue Management via Deep Reinforcement Learning", Nov. 2015. [Online]. Available: http://arxiv.org/abs/1511.08099.

[439] T.-H. Wen, D. Vandyke, N. Mrksic, M. Gasic, L. M. Rojas-Barahona, P.-H. Su, S. Ultes, and S. Young, "A Network-based End-to-End Trainable Task-oriented Dialogue System", Apr. 2016. [Online]. Available: http://arxiv.org/abs/1604.04562.

[440] C. Watkins, *Models of Delayed Reinforcement Learning*, 1989.

[441] P.-H. Su, M. Gasic, N. Mrksic, L. Rojas-Barahona, S. Ultes, D. Vandyke, T.-H. Wen, and S. Young, "Continuously Learning Neural Dialogue Management", Jun. 2016. [Online]. Available: http://arxiv.org/abs/1606.02689.

[442] S. I. Amari, "Natural Gradient Works Efficiently in Learning", *Neural Computation*, 1998, ISSN: 08997667. DOI: 10.1162/089976698300017746.

[443] R. J. Williams, "Simple statistical gradient-following algorithms for connectionist reinforcement learning", *Machine Learning*, 1992, ISSN: 0885-6125. DOI: 10.1007/bf00992696.

[444] P.-H. Su, M. Gasic, N. Mrksic, L. Rojas-Barahona, S. Ultes, D. Vandyke, T.-H. Wen, and S. Young, "On-line Active Reward Learning for Policy Optimisation in Spoken Dialogue Systems", May 2016. [Online]. Available: http://arxiv.org/abs/1605.07669.

[445] C. Sauper and R. Barzilay, “Automatically generating Wikipedia articles: A structure-aware approach”, in *ACL-IJCNLP 2009 - Joint Conf. of the 47th Annual Meeting of the Association for Computational Linguistics and 4th Int. Joint Conf. on Natural Language Processing of the AFNLP, Proceedings of the Conf.*, 2009, pp. 208–216, ISBN: 9781617382581. DOI: 10.3115/1687878.1687909.

[446] F. Mairesse, M. Gašić, F. Jurčíček, S. Keizer, B. Thomson, K. Yu, and S. Young, “Phrase-based statistical language generation using graphical models and active learning”, in *ACL 2010 - 48th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, 2010, ISBN: 9781617388088.

[447] D. Tuan Nguyen and T. Tran, “Structure-based Generation System for E2E NLG Challenge”, Tech. Rep., 2018. [Online]. Available: http://www.macs.hw.ac.uk/InteractionLab/E2E/final_papers/E2E-UIT_DANGNT.pdf.

[448] S. Mille and S. Dasiopoulou, “FORGe at E2E 2017”, Tech. Rep., 2018. [Online]. Available: http://www.macs.hw.ac.uk/InteractionLab/E2E/final_papers/E2E-FORGe.pdf.

[449] A. H. Oh and A. I. Rudnicky, “Stochastic natural language generation for spoken dialog systems”, *Computer Speech and Language*, 2002, ISSN: 08852308. DOI: 10.1016/S0885-2308(02)00012-8.

[450] K. van Deemter, E. Krahmer, and M. Theune, “Plan-based vs. template-based NLG: a false opposition?”, *Proceedings of the KI’99 Workshop: May I Speak Freely*, 1999.

[451] A. Koller and R. P. Petrick, “Experiences with planning for natural language generation”, in *Computational Intelligence*, 2011. DOI: 10.1111/j.1467-8640.2010.00370.x.

[452] T. H. Wen, M. Gašić, N. Mrkšić, P. H. Su, D. Vandyke, and S. Young, “Semantically conditioned lstm-based Natural language generation for spoken dialogue systems”, in *Conference Proceedings - EMNLP 2015: Conference on Empirical Methods in Natural Language Processing*, 2015, ISBN: 9781941643327. DOI: 10.18653/v1/d15-1199.

[453] H. Mei, M. Bansal, and M. R. Walter, “What to talk about and how? Selective generation using LSTMs with coarse-to-fine alignment”, in *2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT 2016 - Proceedings of the Conference*, 2016, ISBN: 9781941643914. DOI: 10.18653/v1/n16-1086.

[454] O. Dušek and F. Jurčíček, “Sequence-to-sequence generation for spoken dialogue via deep syntax trees and strings”, in *54th Annual Meeting of the Association for Computational Linguistics, ACL 2016 - Short Papers*, 2016, ISBN: 9781510827592. DOI: 10.18653/v1/p16-2008.

[455] M. Chen, G. Lampouras, and A. Vlachos, “Sheffield at E2E: structured prediction approaches to end-to-end language generation”, Tech. Rep., 2018. [Online]. Available: http://www.macs.hw.ac.uk/InteractionLab/E2E/final_papers/E2E-Sheffield.pdf.

[456] Y. Puzikov and I. Gurevych, “E2E NLG Challenge: Neural Models vs. Templates”, 2019. DOI: 10.18653/v1/w18-6557.

[457] Z. Li, J. Kiseleva, and M. De Rijke, "Dialogue Generation: From Imitation Learning to Inverse Reinforcement Learning", *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, ISSN: 2159-5399. DOI: `10.1609/aaai.v33i01.33016722`.

[458] C. Rosset, *Turing-NLG: A 17-billion-parameter language model by Microsoft - Microsoft Research*, 2020. [Online]. Available: `https://www.microsoft.com/en-us/research/blog/turing-nlg-a-17-billion-parameter-language-model-by-microsoft/`.

[459] K. Mo, Y. Zhang, S. Li, J. Li, and Q. Yang, "Personalizing a dialogue system with transfer reinforcement learning", in *32nd AAAI Conference on Artificial Intelligence, AAAI 2018*, 2018, ISBN: 9781577358008.

[460] J. Bang, H. Noh, Y. Kim, and G. G. Lee, "Example-based chat-oriented dialogue system with personalized long-term memory", in *2015 International Conference on Big Data and Smart Computing, BIGCOMP 2015*, 2015, ISBN: 9781479973033. DOI: `10.1109/35021BIGCOMP.2015.7072837`.

[461] Y. Kim, J. Bang, J. Choi, S. Ryu, S. Koo, and G. G. Lee, "Acquisition and use of long-term memory for personalized dialog systems", in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2015, ISBN: 9783319155562. DOI: `10.1007/978-3-319-15557-9{\_}8`.

[462] I. Casanueva, T. Hain, H. Christensen, R. Marxer, and P. Green, "Knowledge transfer between speakers for personalised dialogue management", in *SIGDIAL 2015 - 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue, Proceedings of the Conference*, 2015, ISBN: 9781941643754. DOI: `10.18653/v1/w15-4603`.

[463] A. Genevay and R. Laroche, "Transfer learning for user adaptation in spoken dialogue systems", in *Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS*, 2016, ISBN: 9781450342391.

[464] F. d. Hengst, M. Hoogendoorn, F. van Harmelen, and J. Bosman, "Reinforcement Learning for Personalized Dialogue Management", Aug. 2019. [Online]. Available: `http://arxiv.org/abs/1908.00286`.

[465] Z. Lin, A. Madotto, C.-S. Wu, and P. Fung, "Personalizing Dialogue Agents via Meta-Learning", pp. 5454–5459, May 2019. [Online]. Available: `http://arxiv.org/abs/1905.10033`.

[466] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks", in *34th International Conference on Machine Learning, ICML 2017*, 2017, ISBN: 9781510855144.

[467] Y. Zheng, R. Zhang, X. Mao, and M. Huang, "A Pre-training Based Personalized Dialogue Generation Model with Persona-sparse Data", Nov. 2019. [Online]. Available: `http://arxiv.org/abs/1911.04700`.

[468] H. Song, W.-N. Zhang, J. Hu, and T. Liu, "Generating Persona Consistent Dialogues by Exploiting Natural Language Inference", Nov. 2019. [Online]. Available: `http://arxiv.org/abs/1911.05889`.

[469] M. Xu, P. Li, H. Yang, P. Ren, Z. Ren, Z. Chen, and J. Ma, "A Neural Topical Expansion Framework for Unstructured Persona-oriented Dialogue Generation", Feb. 2020. [Online]. Available: `http://arxiv.org/abs/2002.02153`.

[470] H. Rashkin, E. M. Smith, M. Li, and Y.-L. Boureau, "Towards Empathetic Open-domain Conversation Models: a New Benchmark and Dataset", pp. 5370–5381, Oct. 2018. [Online]. Available: http://arxiv.org/abs/1811.00207.

[471] P. Fung, D. Bertero, P. X. J. H. Park, C. S. Wu, and A. Madotto, "Empathetic dialog systems", in *LREC 2018 - 11th International Conference on Language Resources and Evaluation*, 2019, ISBN: 9791095546009.

[472] Z. Lin, A. Madotto, J. Shin, P. Xu, and P. Fung, "MoEL: Mixture of Empathetic Listeners", in *EMNLP*, 2019, pp. 121–132. [Online]. Available: https://www.aclweb.org/anthology/D19-1012.pdf.

[473] J. Shin, P. Xu, A. Madotto, and P. Fung, "HappyBot: Generating Empathetic Dialogue Responses by Improving User Experience Look-ahead", Jun. 2019. [Online]. Available: http://arxiv.org/abs/1906.08487.

[474] T. Zhao, A. Lu, K. Lee, and M. Eskenazi, "Generative Encoder-Decoder Models for Task-Oriented Spoken Dialog Systems with Chatting Capability", 2018. DOI: 10.18653/v1/w17-5505.

[475] A. Madotto, Z. Lin, C.-S. Wu, J. Shin, and P. Fung, "Attention over Parameters for Dialogue Systems", Jan. 2020. [Online]. Available: http://arxiv.org/abs/2001.01871.

[476] K. Lee, O. Firat, A. Agarwal, C. Fannjiang, and D. Sussillo, "Hallucinations in Neural Machine Translation", Tech. Rep., 2018. [Online]. Available: https://github.com/moses-smt/mosesdecoder/blob/master/.

[477] A. Rohrbach, L. A. Hendricks, K. Burns, T. Darrell, and K. Saenko, "Object Hallucination in Image Captioning", 2019. DOI: 10.18653/v1/d18-1437.

[478] M. Artetxe and H. Schwenk, "Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond", *Transactions of the Association for Computational Linguistics*, 2019, ISSN: 2307-387X. DOI: 10.1162/tacl{_}a{_}00288.

[479] P. A. Taraldsen and V. Vatne, "Expanding on the end-to-end memory network for goal-oriented dialogue", PhD thesis, University of Agder, 2019. [Online]. Available: https://uia.brage.unit.no/uia-xmlui/bitstream/handle/11250/2618715/Vatne%2C%20Vegard%20og%20Taraldsen%2C%20Peter%20Arentz.pdf.

[480] P. M. Cole, M. K. Michel, and L. O. Teti, "The Development of Emotion Regulation and Dysregulation: A Clinical Perspective", *Monographs of the Society for Research in Child Development*, vol. 59, no. 2/3, p. 73, 1994, ISSN: 0037976X. DOI: 10.2307/1166139. [Online]. Available: https://www.jstor.org/stable/1166139?origin=crossref.

[481] M. Seehausen, P. Kazzer, M. Bajbouj, and K. Prehn, "Effects of Empathic Paraphrasing – Extrinsic Emotion Regulation in Social Conflict", *Frontiers in Psychology*, vol. 3, p. 482, Nov. 2012, ISSN: 1664-1078. DOI: 10.3389/fpsyg.2012.00482. [Online]. Available: http://journal.frontiersin.org/article/10.3389/fpsyg.2012.00482/abstract.

[482] J. J. Gross, "The emerging field of emotion regulation: An integrative review.", *Review of General Psychology*, vol. 2, no. 3, pp. 271–299, 1998, ISSN: 1939-1552. DOI: 10.1037/1089-2680.2.3.271. [Online]. Available: http://doi.apa.org/getdoi.cfm?doi=10.1037/1089-2680.2.3.271.

[483] G. Sheppes, S. Scheibe, G. Suri, and J. J. Gross, "Emotion-Regulation Choice", *Psychological Science*, vol. 22, no. 11, pp. 1391–1396, Nov. 2011, ISSN: 0956-7976. DOI: 10.1177/0956797611418350. [Online]. Available: http://journals.sagepub.com/doi/10.1177/0956797611418350.

[484] M. D. Lieberman, N. I. Eisenberger, M. J. Crockett, S. M. Tom, J. H. Pfeifer, and B. M. Way, "Putting Feelings Into Words", *Psychological Science*, vol. 18, no. 5, pp. 421–428, May 2007, ISSN: 0956-7976. DOI: 10.1111/j.1467-9280.2007.01916.x. [Online]. Available: http://journals.sagepub.com/doi/10.1111/j.1467-9280.2007.01916.x.

[485] I. Androutsopoulos and P. Malakasiotis, "A survey of paraphrasing and textual entailment methods", *Journal of Artificial Intelligence Research*, 2010, ISSN: 10769757. DOI: 10.1613/jair.2985.

[486] J. Wieting, M. Bansal, K. Gimpel, and K. Livescu, "TOWARDS UNIVERSAL PARAPHRASTIC SENTENCE EMBEDDINGS", 2016. [Online]. Available: https://arxiv.org/pdf/1511.08198.pdf.

[487] Z. Li, X. Jiang, L. Shang, and H. Li, "Paraphrase Generation with Deep Reinforcement Learning", 2018. [Online]. Available: https://arxiv.org/pdf/1711.00279.pdf.

[488] S. A. Hasan, B. Liu, J. Liu, A. Qadir, K. Lee, V. Datla, A. Prakash, and O. Farri, "Neural Clinical Paraphrase Generation with Attention", in *Proceedings of the Clinical Natural Language Processing Workshop*, Osaka, Japan, 2016, pp. 42–53.

[489] A. Prakash, S. A. Hasan, K. Lee, V. Datla, A. Qadir, J. Liu, and O. Farri, "Neural Paraphrase Generation with Stacked Residual LSTM Networks", 2016. [Online]. Available: https://arxiv.org/pdf/1610.03098.pdf.

[490] J. Hill, W. Randolph Ford, and I. G. Farreras, "Real conversations with artificial intelligence: A comparison between human-human online conversations and human-chatbot conversations", *Computers in Human Behavior*, 2015, ISSN: 07475632. DOI: 10.1016/j.chb.2015.02.026.

[491] A. Pinar Saygin, I. Cicekli, and V. Akman, "Turing Test: 50 Years Later", *Minds and Machines*, vol. 10, no. 4, pp. 463–518, 2000, ISSN: 09246495. DOI: 10.1023/A:1011288000451. [Online]. Available: http://link.springer.com/10.1023/A:1011288000451.

[492] D. Nikolic, *Is the Turing test still relevant? How about Turing time?*, 2019. [Online]. Available: https://medium.com/savedroid/is-the-turing-test-still-relevant-how-about-turing-time-d73d472c18f1.

[493] "The Social and interactional dimensions of human-computer interfaces", *Choice Reviews Online*, vol. 33, no. 07, pp. 33–3961, Mar. 1996, ISSN: 0009-4978. DOI: 10.5860/choice.33-3961.

[494] K. M. Colby, F. D. Hilf, S. Weber, and H. C. Kraemer, "Turing-like indistinguishability tests for the validation of a computer simulation of paranoid processes", *Artificial Intelligence*, vol. 3, no. C, pp. 199–221, 1972, ISSN: 00043702. DOI: 10.1016/0004-3702(72)90049-5.

[495] S. M. Shieber, "Lessons from a Restricted Turing Test", *Communications of the ACM*, 1994, ISSN: 15577317. DOI: 10.1145/175208.175217.

[496] S. Russell and P. Norvig, *Artificial Intelligence A Modern Approach Third Edition.* 2010, ISBN: 9780136042594. DOI: 10.1017/S0269888900007724. arXiv: 9809069v1 [gr-qc].

[497] S. Bangalore, O. Rambow, and S. Whittaker, "Evaluation Metrics for Generation", Tech. Rep., 2000.

[498] G. Doddington, "Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics", in *HLT '02 Proceedings of the second international conference on Human Language Technology Research*, San Francisco, CA, 2002, pp. 138–145. [Online]. Available: http://www.mt-archive.info/HLT-2002-Doddington.pdf.

[499] C.-W. Liu, R. Lowe, I. V. Serban, M. Noseworthy, L. Charlin, and J. Pineau, "How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation", Mar. 2016. [Online]. Available: http://arxiv.org/abs/1603.08023.

[500] D. Coughlin, "Correlating automated and human assessments of machine translation quality", *MT Summit IX: 9th Machine Translation Summit*, 2003.

[501] M. Galley, C. Brockett, A. Sordoni, Y. Ji, M. Aim, C. Quirk, M. Mitchell, J. Gao, and B. Dolan, "ΔbLEU: A discriminative metric for generation tasks with intrinsically diverse targets", in *ACL-IJCNLP 2015 - 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, Proceedings of the Conference*, 2015, ISBN: 9781941643730. DOI: 10.3115/v1/p15-2073.

[502] S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments", in *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 2005.

[503] A. Lavie and A. Agarwal, "Meteor: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments", Tech. Rep., 2007. [Online]. Available: https://www.cs.cmu.edu/~alavie/METEOR/pdf/Lavie-Agarwal-2007-METEOR.pdf.

[504] C. Y. Lin, "Rouge: A package for automatic evaluation of summaries", *Proceedings of the workshop on text summarization branches out (WAS 2004)*, 2004.

[505] C.-Y. Lin and E. Hovy, "Automatic Evaluation of Summaries Using N-gram Co-Occurrence Statistics", Tech. Rep., 2003. [Online]. Available: http://www.aclweb.org/anthology/N03-1020.

[506] C.-Y. Lin and F. J. Och, "Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics", 2004. DOI: 10.3115/1218955.1219032.

[507] V. Rus and M. Lintean, "A comparison of greedy and optimal assessment of natural language student input using word-to-word similarity metrics", in *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, 2012.

[508] P. W. Foltz, W. Kintsch, and T. K. Landauer, "The measurement of textual coherence with latent semantic analysis", *Discourse Processes*, 1998, ISSN: 0163-853X. DOI: 10.1080/01638539809545029.

[509] G. Forgues, J. Pineau, J.-M. Larcheveque, and R. Tremblay, "Bootstrapping Dialog Systems with Word Embeddings", *NIPS, Modern Machine Learning and Natural Language Processing Workshop*, 2014.

[510] J. Li, M. Galley, C. Brockett, J. Gao, and B. Dolan, "A diversity-promoting objective function for neural conversation models", in *2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT 2016 - Proceedings of the Conference*, 2016, ISBN: 9781941643914. DOI: `10.18653/v1/n16-1014`.

[511] I. V. Serban, A. Sordoni, R. Lowe, L. Charlin, J. Pineau, A. Courville, and Y. Bengio, "A Hierarchical Latent Variable Encoder-Decoder Model for Generating Dialogues", 2016.

[512] L. Mou, Y. Song, R. Yan, G. Li, L. Zhang, and Z. Jin, "Sequence to backward and forward sequences: A content-introducing approach to generative short-text conversation", in *COLING 2016 - 26th International Conference on Computational Linguistics, Proceedings of COLING 2016: Technical Papers*, 2016, ISBN: 9784879747020.

[513] R. Lowe, M. Noseworthy, I. V. Serban, N. Angelard-Gontier, Y. Bengio, and J. Pineau, "Towards an Automatic Turing Test: Learning to Evaluate Dialogue Responses", 2017. DOI: `10.18653/v1/P17-1103`.

[514] H.-Y. Shum, X. He, and D. Li, "From Eliza to XiaoIce: Challenges and Opportunities with Social Chatbots", 2018, ISSN: 2095-9184. DOI: `10.1631/FITEE.1700826`.

[515] D. Khashabi, S. Chaturvedi, M. Roth, S. Upadhyay, and D. Roth, "Looking Beyond the Surface: A Challenge Set for Reading Comprehension over Multiple Sentences", 2018. DOI: `10.18653/v1/n18-1023`.

[516] M. Taher Pilehvar and J. Camacho-Collados, "WiC: the Word-in-Context Dataset for Evaluating Context-Sensitive Meaning Representations", Tech. Rep., 2019. [Online]. Available: `https://www.wiktionary.org/`.

[517] D. Braun, A. Hernandez Mendez, F. Matthes, and M. Langen, "Evaluating Natural Language Understanding Services for Conversational Question Answering Systems", Tech. Rep., 2017, pp. 174–185. [Online]. Available: `https://www.aclweb.org/anthology/W17-5522.pdf`.

[518] X. Liu, A. Eshghi, P. Swietojanski, and V. Rieser, "Benchmarking Natural Language Understanding Services for building Conversational Agents", Tech. Rep., 2019. [Online]. Available: `https://www.luis.ai/home`.

[519] J. Li, M. Galley, C. Brockett, G. P. Spithourakis, J. Gao, and B. Dolan, "A Persona-Based Neural Conversation Model", Mar. 2016. [Online]. Available: `https://arxiv.org/abs/1603.06155`.

[520] J. Li, M. Galley, C. Brockett, J. Gao, and B. Dolan, "A Diversity-Promoting Objective Function for Neural Conversation Models", Oct. 2015. [Online]. Available: `http://arxiv.org/abs/1510.03055`.

[521] J. Novikova, O. Dušek, and V. Rieser, "The E2E Dataset: New Challenges For End-to-End Generation", 2018. DOI: `10.18653/v1/w17-5525`.

[522] R. Vedantam, C. L. Zitnick, and D. Parikh, "CIDEr: Consensus-based image description evaluation", in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2015, ISBN: 9781467369640. DOI: 10.1109/CVPR.2015.7299087.

[523] T. Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context", in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2014. DOI: 10.1007/978-3-319-10602-1{_}48.

[524] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dolí, and C. L. Zitnick, "Microsoft COCO Captions: Data Collection and Evaluation Server", Tech. Rep., 2015.

[525] M. Turek, "Explainable Artificial Intelligence (XAI)", DARPA, Tech. Rep., 2016. [Online]. Available: https://www.darpa.mil/attachments/DARPA-BAA-16-53.pdf.

[526] A. Barredo Arrieta, N. Díaz-Rodríguez, J. D. Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI", Tech. Rep., 2019.

[527] M. T. Ribeiro, S. Singh, and C. Guestrin, ""Why Should I Trust You?" Explaining the Predictions of Any Classifier", 2016. DOI: 10.1145/2939672.2939778. [Online]. Available: http://dx.doi.org/10.1145/2939672.2939778.

[528] S. M. Lundberg, G. G. Erion, and S.-I. Lee, "Consistent Individualized Feature Attribution for Tree Ensembles", Tech. Rep., 2019. [Online]. Available: http://github.com/slundberg/shap.

[529] V. Arya, R. K. E. Bellamy, Y. Chen, A. Dhurandhar, M. Hind, S. C. Hoffman, S. Houde, Q. V. Liao, R. Luss, A. Mojsilović, S. Mourad, P. Pedemonte, R. Raghavendra, J. Richards, P. Sattigeri, K. Shanmugam, M. Singh, K. R. Varshney, D. Wei, and Y. Zhang, "One Explanation Does Not Fit All: A Toolkit and Taxonomy of AI Explainability Techniques", Tech. Rep., 2019. [Online]. Available: http://aix360..

[530] K. S. Gurumoorthy, A. Dhurandhar, G. Cecchi, and C. Aggarwal, "Efficient Data Representation by Selecting Prototypes with Importance Weights", Tech. Rep., 2019.

[531] A. Kumar, P. Sattigeri, and A. Balakrishnan, "Variational inference of disentangled latent concepts from unlabeled observations", in *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*, 2018.

[532] A. Dhurandhar, P. Y. Chen, R. Luss, C. C. Tu, P. Ting, K. Shanmugam, and P. Das, "Explanations based on the Missing: Towards contrastive explanations with pertinent negatives", in *Advances in Neural Information Processing Systems*, 2018.

[533] R. Luss, P.-Y. Chen, A. Dhurandhar, P. Sattigeri, K. Shanmugam, and C.-C. Tu, "Generating Contrastive Explanations with Monotonic Attribute Functions", Tech. Rep., 2019.

[534] M. Hind, D. Wei, M. Campbell, N. C. Codella, A. Dhurandhar, A. Mojsilović, K. Natesan Ramamurthy, and K. R. Varshney, "TED: Teaching AI to explain its decisions", in *AIES 2019 - Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 2019, ISBN: 9781450363242. DOI: 10.1145/3306618.3314273.

[535] S. Dash, O. Günlük, and D. Wei, "Boolean decision rules via column generation", in *Advances in Neural Information Processing Systems*, 2018.

[536] D. Wei, S. Dash, T. Gao, and O. G. ¨. Unï Uk, "Generalized Linear Rule Models", Tech. Rep., 2019. [Online]. Available: http://proceedings.mlr.press/v97/wei19a/wei19a.pdf.

[537] D. Alvarez-Melis and T. S. Jaakkola, "Towards robust interpretability with self-explaining neural networks", in *Advances in Neural Information Processing Systems*, 2018.

[538] B. Hoover, H. Strobelt, and S. Gehrmann, "EXBERT: A Visual Analysis Tool to Explore Learned Representations in Transformers Models", Tech. Rep., 2019. [Online]. Available: www.exbert.net..

[539] K. Kuksenok and N. Praß, "Transparency in Maintenance of Recruitment Chatbots", Tech. Rep., 2019.

[540] M. HAMILTON, "A rating scale for depression", *Journal of neurology, neurosurgery, and psychiatry*, 1960, ISSN: 00223050. DOI: 10.1136/jnnp.23.1.56.

[541] P. Bech, *The Bech, Hamilton and Zung Scales for Mood Disorders: Screening and Listening*. 2012. DOI: 10.1007/978-3-642-97633-9.

[542] R. L. Spitzer, K. Kroenke, J. B. W. Williams, Group, and the Patient Health Questionnaire Primary Care Study, "Validation and Utility of a Self-report Version of PRIME-MD - The PHQ Primary Care Study", *JAMA*, vol. 282, no. 18, p. 1737, Nov. 1999. DOI: 10.1001/jama.282.18.1737. [Online]. Available: http://jama.jamanetwork.com/article.aspx?doi=10.1001/jama.282.18.1737.

[543] K. Kroenke and R. L. Spitzer, "The PHQ-9: A New Depression Diagnostic and Severity Measure", *Psychiatric Annals*, vol. 32, no. 9, 2002.

[544] R. L. Spitzer, K. Kroenke, J. B. W. Williams, and B. Löwe, "A Brief Measure for Assessing Generalized Anxiety Disorder", *Archives of Internal Medicine*, vol. 166, no. 10, p. 1092, May 2006, ISSN: 0003-9926. DOI: 10.1001/archinte.166.10.1092. [Online]. Available: http://archinte.jamanetwork.com/article.aspx?doi=10.1001/archinte.166.10.1092.

[545] S. Cohen, T. Kamarck, and R. Mermelstein, "A Global Measure of Perceived Stress", *Journal of Health and Social Behavior*, vol. 24, no. 4, p. 385, Dec. 1983, ISSN: 00221465. DOI: 10.2307/2136404. [Online]. Available: http://www.jstor.org/stable/2136404?origin=crossref.

[546] K. H. Ly, A.-M. Ly, and G. Andersson, "A Fully Automated Conversational Agent for Promoting Mental Well-Being: A pilot RCT using mixed methods", *Internet Interventions*, 2017, ISSN: 22147829. DOI: 10.1016/j.invent.2017.10.002. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S221478291730091X.

[547] J. R. Crawford and J. D. Henry, "The Depression Anxiety Stress Scales (DASS): Normative data and latent structure in a large non-clinical sample", *British Journal of Clinical Psychology*, 2003, ISSN: 01446657. DOI: 10.1348/014466503321903544.

[548] E. Diener, D. Wirtz, W. Tov, C. Kim-Prieto, D. w. Choi, S. Oishi, and R. Biswas-Diener, "New well-being measures: Short scales to assess flourishing and positive and negative feelings", *Social Indicators Research*, 2010, ISSN: 03038300. DOI: 10.1007/s11205-009-9493-y.

[549] R. Kobau, J. Sniezek, M. M. Zack, R. E. Lucas, and A. Burns, "Well-being assessment: An evaluation of well-being scales for public health and population estimates of well-being among US adults", *Applied Psychology: Health and Well-Being*, 2010, ISSN: 17580846. DOI: 10.1111/j.1758-0854.2010.01035.x.

[550] J. R. Crawford and J. D. Henry, "The Positive and Negative Affect Schedule (PANAS): Construct validity, measurement properties and normative data in a large non-clinical sample", *British Journal of Clinical Psychology*, 2004, ISSN: 01446657. DOI: 10.1348/0144665031752934.

[551] M. A. Zevon and A. Tellegen, "The structure of mood change: An idiographic/nomothetic analysis", *Journal of Personality and Social Psychology*, 1982, ISSN: 00223514. DOI: 10.1037/0022-3514.43.1.111.

[552] C. Harmon Jones, B. Bastian, and E. Harmon-Jones, "The discrete emotions questionnaire: A new tool for measuring state self-reported emotions", *PLoS ONE*, 2016. DOI: 10.1371/journal.pone.0159915.

[553] R. M. Bagby, A. G. Ryder, D. R. Schuller, and M. B. Marshall, *The Hamilton Depression Rating Scale: Has the gold standard become a lead weight?*, 2004. DOI: 10.1176/appi.ajp.161.12.2163.

[554] R. W. Firestone, "Firestone assessment of self-destructive thoughts", *San Antonio, TX: Psychological Corporation*, 1996.

[555] G. E. Berrios and A. Bulbena-Villarasa, "The Hamilton Depression Scale and the numerical description of the symptoms of depression.", *Psychopharmacology series*, vol. 9, pp. 80–92, 1990, ISSN: 09316795. DOI: 10.1007/978-3-642-75373-2{_}10.

[556] D. Watson, L. A. Clark, and A. Tellegen, "Development and Validation of Brief Measures of Positive and Negative Affect: The PANAS Scales", *Journal of Personality and Social Psychology*, 1988, ISSN: 00223514. DOI: 10.1037/0022-3514.54.6.1063.

[557] J. Salamon and R. Moucek, "Link between sentiment and human activity represented by footsteps: Experiment exploiting IoT devices and social networks", in *HEALTHINF 2016 - 9th International Conference on Health Informatics, Proceedings; Part of 9th International Joint Conference on Biomedical Engineering Systems and Technologies, BIOSTEC 2016*, 2016, ISBN: 9789897581700.

[558] P. Ekman, "Emotions inside out: 130 Years after Darwin's The Expression of the Emotions in Man and Animals - Introduction", in *Annals of the New York Academy of Sciences*, vol. 1000, New York Academy of Sciences, 2003, pp. 1–6. DOI: 10.1196/annals.1280.002.