

Umělá inteligence a rozpoznávání (KIV/UIR)

13. Zpracování přirozeného jazyka

Ing. Pavel Král, Ph.D.

Katedra informatiky a výpočetní techniky
Západočeská Univerzita

14. května 2014

Zpracování přirozeného jazyka (*Natural Language Processing (NLP)*)

- = schopnost počítačového programu rozumět mluvené i psané lidské řeči
- patří do oblasti umělé inteligence (UI)

- Rozpoznávání řeči
- Syntéza řeči
- Porozumění přirozenému jazyku (*Natural language understanding*)
- Komunikace člověk-počítač (řízení dialogu)
- Generování přirozeného jazyka
- Strojový překlad
- Odpovídání na dotazy
- Získávání informací (*Information retrieval*)
- Extrakce znalostí (*Information extraction*)
- ...

Aplikace

Umělá
inteligence a
rozpoznávání
(KIV/UIR)

Ing. Pavel
Král, Ph.D.

Úvod

Rozpoznávání
mluvené řeči

■ ???

Rozpoznávání přirozeného jazyka

Umělá
inteligence a
rozpznávání
(KIV/UIR)

Ing. Pavel
Kráal, Ph.D.

Úvod

Rozpoznávání
mluvené řeči

- $W = \{w_1, \dots, w_n\}$... sekvence n slov
- $O = \{o_1, \dots, o_m\}$... sekvence m (pozorování) z řečového signálu

Cíl:

- nalezení nejpravděpodobnější sekvence slov W k dané sekvenci O z řeč. signálu:

$$\hat{W} = \arg \max_W P(W|O) \quad (1)$$

Rozpoznávání přirozeného jazyka

Umělá
inteligence a
rozpznávání
(KIV/UIR)

Ing. Pavel
Král, Ph.D.

Úvod

Rozpoznávání
mluvené řeči

- řešení pomocí Bayesovy věty

$$\hat{W} = \arg \max_W \frac{P(O|W).P(W)}{P(O)} \quad (2)$$

- $P(O)$ nehraje roli při maximalizaci přes $P(W) \rightarrow$

$$\hat{W} = \arg \max_W P(O|W).P(W) \quad (3)$$

Základní úlohy rozpoznávání přirozeného jazyka

Umělá
inteligence a
rozpoznávání
(KIV/UIR)

Ing. Pavel
Král, Ph.D.

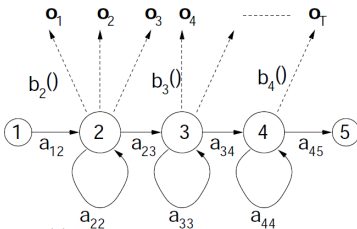
Úvod

Rozpoznávání
mluvené řeči

- 1 zpracování akustického signálu
 - nahrávání
 - digitalizace
 - předzpracování
 - parametrizace
- 2 vytvoření akustického modelu $P(O|W)$
- 3 vytvoření jazykového modelu $P(W)$
- 4 nalezení nejpravděpodobnější posloupnosti slov

Akustický model

- nejčastěji skryté Markovovy modely



Obrázek: Ukázka skrytých Markovových modelů

- stochastický konečný automat
- generativní model
- normální rozdělení (směs Gaussových funkcí)
- akustické jednotky: fóny, slabiky, 3-fóny,...

Cíl:

- určení pravděpodobnosti posloupnosti n slov
- předpověď dalšího slova na základě předchozích

Další aplikace:

- Korektury pravopisu a gramatiky
- Dopňování diakritiky
- Automatické určení jazyka
- ...

N-gramový jazykový model

- pravděpodobnostní přístup
- dnes nejrozšířenější

$$P(W) = P(w_1, \dots, w_n) \quad (4)$$

- w_i = slovo na pozici i
- n = počet slov ve větě
- $P(W)$ = pravděpodobnost sekvence slov W

$$P(W) = P(w_1) \cdot P(w_2|w_1) \dots P(w_n|w_1, \dots, w_{n-1}) \quad (5)$$

- závislost pravděpodobnosti slova w_i na všech předchozích
- nemožno odhadnout $P(w_i|w_1, \dots, w_{i-1})$, když $i \rightarrow \infty$

N-gramový jazykový model

Umělá
inteligence a
rozpoznávání
(KIV/UIR)

Ing. Pavel
Král, Ph.D.

Úvod

Rozpoznávání
mluvené řeči

- v praxi se historie “zkrátí” na k předchozích slov
- Markovův předpoklad nezávislosti

→

- unigram: $P(w_i)$
- bigram: $P(w_i|w_{i-1})$
- trigram: $P(w_i|w_{i-1}, w_{i-2})$

N-gramový jazykový model

- výpočet z relativní frekvence sekvencí slov na trénovacím korpusu

→

- unigram: $P(w_i) = \frac{\#w_i}{N}$
- bigram: $P(w_i|w_{i-1}) = \frac{\#(w_{i-1}, w_i)}{\#w_{i-1}}$
- trigram: $P(w_i|w_{i-1}, w_{i-2}) = \frac{\#(w_{i-2}, w_{i-1}, w_i)}{\#(w_{i-2}, w_{i-1})}$

Př:

- Šel pes domů

N-gramový jazykový model

Př. - unigram (příp. bigram)

Umělá
inteligence a
rozpoznávání
(KIV/UIR)

Ing. Pavel
Kráľ, Ph.D.

Úvod

Rozpoznávání
mluvené řeči

Trénovací korpus

- Veronika je v kině
- Jana je v zahradě

Testovací sekvence W

- Veronika je v zahradě
- $P(W) = ?$

Unigramový jazykový model

Př. - určení jazyka

Umělá
inteligence a
rozpoznávání
(KIV/UIR)

Ing. Pavel
Král, Ph.D.

Úvod

Rozpoznávání
mluvené řeči

- Pozn.: použít písmenné unigramy