

Statistická extrakce idiomů

Jan Bušta
CZPJ FI MU, Brno

PV173
3. 11. 2010

Tento projekt je spolufinancován Evropským sociálním fondem a státním rozpočtem České republiky.



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

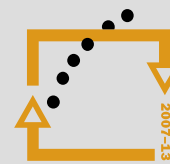
Tento projekt je spolufinancován Evropským sociálním fondem a státním rozpočtem České republiky.



EVROPSKÁ UNIE



MINISTERSTVO ŠKOLSTVÍ,
MLÁDEŽE A TĚLOVÝCHOVY



**OP Vzdělávání
pro konkurenceschopnost**



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

Motivace

Fazly, A. – Stevenson, S.

Automatically Constructing a Lexicon of
Verb Phrase Idiomatic Combinations

(2006)

Cíle

- automaticky najít v textovém korpusu idiomatická spojení ve tvaru

sloveso + (předložka +) jméno

(spadnout z višně, nebud' labuť, zaset nenávist)

- změřit pokrytí a přesnost oproti ruční anotaci a SČFI
- vybrat a přizpůsobit vhodný algoritmus pro extrakci

Jak na to I

- Lexikální pevnost
 - vytvoření množiny „synonymních“ výrazů je jménu ve spojení $\langle v, n \rangle$
 - vytvoření množiny tranzitivních sloves v korpusu
 - výpočet pravděpodobnosti $\langle v, n \rangle$ vzhledem k $\langle *, n \rangle$ a $\langle v, * \rangle$

$$P_{lex}(\langle v, n \rangle) = \left(\frac{|\mathcal{V} \times \mathcal{N}| f(\langle v, n \rangle)}{f(\langle *, n \rangle) f(\langle v, * \rangle)} - \frac{|\mathcal{V} \times \mathcal{N}| f(\langle v, n_j \rangle)}{f(\langle *, n_j \rangle) f(\langle v, * \rangle)} \right) \div s$$

\mathcal{V} – množina tranzitivních sloves

\mathcal{N} – množina synonym k n

$n_j \in \mathcal{N}$ – množina synonym k n

- aneb jak se může měnit jméno ve frázi

Jak na to II

- Syntaktická pevnost
 - pasivizace
 - pluralizace
 - negace
 - změna (přidání) členu

$$P_{syn}(\langle v, n \rangle) = \sum_{pt_k \in \mathcal{PS}} P(pt_k | \langle v, n \rangle) \log \frac{P(pt_k | \langle v, n \rangle)}{P(pt_k)}$$

pt – varianta idiomů

\mathcal{PS} – množina variant idiomů

- aneb v jaké variantě se fráze vyskytuje

Jak na to III

- Kombinace předchozích metod
 - nastavení vah lexikální a syntaktické pevnosti
 - zlepšení výsledků

$$P_{kom}(\langle v, n \rangle) = \alpha P_{syn}(\langle v, n \rangle) + (1 - \alpha) P_{lex}(\langle v, n \rangle)$$

α – číslo v intervalu $\langle 0, 1 \rangle$

- aneb tak dlouho kombinujeme, dokud nám to nevyjde

Závěr

- Funguje to?
 - pro AJ ano, úspěšnost až 74 %
- A pro češtinu?
 - snad, uvidíme v brzké budoucnosti
- A využití?
 - pomoc lexikografům při vytváření slovníků idiomatických frází
 - detekce potenciálních problémů při strojovém překladu

A jak to celé dopadne?

VÍME VŠE: NEVÍME NIC

Cimrmanova teorie poznání

Děkuji za pozornost.

Jan Bušta
xbusta@fi.muni.cz

Tento projekt je spolufinancován Evropským sociálním fondem a státním rozpočtem České republiky.



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ