CPA, PDEV, SA, LCP, LFP

Vít Baisa

seminář NLP

listopad 2012

CPA a PDEV

- aktuálně projekt DVC – disambiguation of verbs by collocations
- `http://clg.wlv.ac.uk/projects/DVC/`
- Mitkov, Hanks, Orasan (Wolverhampton)
- role FI: technická podpora, hosting, vývoj nového rozhraní
- role Wolverhamptonu: lexikografická anotace

CPA – anotace slovesa *zlomit*

- anotace v czes2
- dělá se většinou sample (250–500)
- anotace špatně označkovaných dat `x`
- anotace patternů
- sémantické typy se do ontologie přidávají pouze tehdy, kdy pomáhají rozlišit dva různé významy (patterny)
- podkategorie `s`, `a`, `f`, `w`

Jazykový a překladový model pomocí SA, LCP a LFP

- Potřebuji pro libovolný řetězec v textu s četností $> n$ seznam všech po něm následujících řetězců s četností $> n$ a seznam po těchto řetězcích následujících řetězců s četností $> n$.
- Pokud se vyskytuje řetězec více jak $n\times$, je to jedna paměťová jednotka *chunk*.
- m (3) po sobě následujících chunků – *pracovní paměť*.
- SA – suffix array
- LCP – longest common prefix
- LFP^n – longest frequent prefix
- revSA – reverse suffix array

Překladový model

- SA, LCP a LFP
- explicitní hranice mezi jazyky v datech
- krást | to steal || zmatený | confused || děláš | you are doing ||

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ