Lexikální disambiguace anglických sloves

Team

Silvie Cinková, Martin Holub, Vincent Kríž, Lenka Smejkalová

Pavel Rychlý, Adam Rambousek, Vít Baisa
infrastruktura pro anotace korpusu MU Brno

Ema Krejčová, Anna Vernerová, Jonáš Thál
anotační práce

Pavel Pecina
vedoucí výzkumné skupiny pro grantový projekt
Center for Multimodal Interpretation of Large Scale Data
GAČR, 2012-2018

Proč lexikální disambiguace?

- Východisko: text jako posloupnost slov. Co ta slova znamenají?

- Člověk významu slov "rozumí", počítač ne.

- Řada počítačových aplikací vyžaduje, aby počítač pracoval/reagoval podobně jako člověk, který významy slov rozlišuje
  - vyhledávání/extrakce informací, dialogové systémy, analýza a reagování na textové/řečové zprávy, strojový překlad, . . .

- Modelování lexikálního významu je snaha napodobit člověka v rozlišování významu slov.

Lexikální disambiguace – příklad

Slovo v textu: *ženou*

- žena (s)
- hnát (v)
  - *hnát se* (= spěchat)
  - *hnát se za* něčím (ziskem, slávou, penězi, apod.)
  - *hnát se za* někým (chtít jej dostihnout)
  - *hnát se do* něčeho (horlivě se pouštět)
  - *hnát* někoho někam (pobízet/nutit k pohybu nebo k jiné činnosti)
  - *hnát* zvířata (na pastvu, apod.)
  - *hnát* něco (pohánět; předávat mechanickou práci)
  - *hnát* někoho odněkud *sviňským krokem*
  - *hnát* (o rostlinách: prudce vyrážet, e.g. obilí žene do klasů)
  - *hnát vodu na něčí mlýn*
  - *hnát* věci *na ostří nože*
  - . . .

Dvě fáze lexikální disambiguace

- Slovo (libovolný slovní tvar)
  → Lemma (morfologicky základní slovní tvar)
  → Sémantická kategorie (popisuje význam)

- První krok řeší nejednoznačnost *morfologickou*, druhý nejednoznačnost *sémantickou*.

- *Morfologická disambiguace* vs. *sémantická disambiguace*
  - sémantická disambiguace je obecně *mnohem* obtížnější

Problémy lexikálně-sémantické disambiguace

- **forma definice** sémantických kategorií
  - velmi nejasné, jaký způsob definice je pro NLP adekvátní
  - historicky mnoho pokusů a rozdílných přístupů: e.g. tradiční slovníková hesla, tezaury, *word senses*, WordNet, valenční slovníky, FrameNet, sémantické konkordance, distribuční sémantické modely
  - naše cesta = *sémantické vzory typického užívání*
    - Corpus Pattern Analysis (Hanks)
- **obsah definice** sémantických kategorií
  - různí lexikografové typicky vytvoří různá slovníková hesla
  - naše cesta = definice důsledně podloženy statistcky významným nálezem v korpusu
- **intersubjektivní shoda** při určování sémantických kategorií
  - mnohem problematičtější než na rovině morfologie nebo syntaxe
  - množství různých příčin neshody
  - testujeme/analyzujeme pomocí manuálních anotací
- **granularita** sémantických kategorií
  - příliš nízká implikuje malou míru získané informace o významu
  - příliš vysoká implikuje nemožnost jednoznačného určení kategorie
  - optimum zjevně závislé na aplikaci
- vysoká **variabilita** – slovo od slova se hodně liší
  - je obtížné nalézt jednotné/univerzální zásady pro zpracování různých slov