Korpusová lingvistika – 5

Mgr. Dana Hlaváčková, Ph.D.

CJBB105

Morfologické značkování

- token – lemma – tag
- tokenizace
  - rozdělení na pozice
- lemmatizace
  - přiřazení základního slovního tvaru (jednoslovné)
  - subst. – nom.sg., adj. – nom. sg. masc., verb. – infinitiv
  - význam spojený s tvarem – *nerv/nervy, na holičkách*
- tagging
  - přiřazení morfologické značky (všechny interpretace tvaru)
- desambiguace – zjednoznačnění na základě kontextu

Morfologické značkování

- na úrovni **slovních druhů**
  - PoS tagging (angličtina)
  - neohebné slovní druhy (spojky, částice, citoslovce)
    - adverbia – značena negace a stupeň
- **kompletní**
  - všechny morfologické kategorie (slovanské jazyky)
    - ohebné slovní druhy
  - nutné pro další stupně automatického zpracování jazyka a navazující aplikace

Morfologické značky

- transparentní – tagset
  - jednoznačná interpretace značky
- zachycují morfologické charakteristiky
  - křížení se sémantickými vlastnostmi (např. druhy zájmen a adverbií)
- nezávislé na lingvistických teoriích
  - orientované na uživatele
- podoba – kód sestavený z písmen a čísel

Homonymie

- významová – rozdíl v morfologických kategoriích
  - *sladit (vid)*
- tvarová – nejfrekventovanější
  - *jarní (rod, číslo, pád)*
- slovnědruhová
  - *jak (adverbium, spojka, částice)*
- kombinovaná
  - *ženu (subst., f, ak., sg./verb., 1. os., sg.)*
- *Sním je místo něho. Praštil se sluchátkem.*

Metody automatického značkování

- závisí na velikosti a kvalitě morfologického slovníku
- Stochastické (statistické, pravděpodobnostní)
  - strojové učení (referenční data)
- Pravidlové
  - pravidla stanovená lingvisty nebo vyvozená z textu
  - pozitivní i negativní
- Hybridní
  - kombinace obou přístupů, nejúspěšnější
- neznámé tvary – Guesser – odhadne možné lemma a tag
- úspěšnost taggerů – přesnost (precision) a pokrytí (recall)

Morfologická analýza v ČR

- Praha – Český národní korpus, manažer KonText
  - Ústav formální a aplikované lingvistiky MFF UK
  - Ústav teoretické a komputační lingvistiky FF UK
- poziční systém
  - značka se skládá z 16 pozic, každá vyjadřuje jednu morfologickou charakteristiku
  - 2 rezervní, 1 stylová, 1 smíšená
- analyzátor stochastický, pravidlový, hybridní

Morfologická analýza v ČR

- Brno – MU, manažer Sketch Engine
  - Centrum zpracování přirozeného jazyka
- atributivní systém
  - atribut – morfologická kategorie obecně (c = pád)
  - hodnota – morfologická kategorie konkrétně (1–7)
- analyzátor pravidlový, hybridní

Přístup

- KonText – ČNK – http://www.korpus.cz
  - registrace https://www.korpus.cz/toolbar/signup.php
- Sketch Engine – http://ske.fi.muni.cz
  - volně dostupné na MU
  - vlastní přístup viz Register