

Morfologická analýza Desambiguace

PLIN059

Mgr. Dana Hlaváčková, Ph.D.

Morfologie

- slovní druhy
 - *substantiva, adjektiva, pronomina, numeralie, verba, adverbia, prepozice, konjunkce, partikule, interjekce*
- pád
 - *nominativ, genitiv, dativ, akuzativ, vokativ, lokál, instrumentál*
- číslo
 - *singulár, plurál, duál*
- rod
 - *maskulinum, femininum, neutrum*

Proč to potřebujeme?

- **morfologické značkování korpusů**
 - zvýšená informační hodnota korpusu
- možnost hledání v korpusu podle morfologických kategorií
- možnost samostatného použití analyzátoru jako morfologické databáze (slovník)
- předpoklad pro **další stupně analýzy** jazyka
 - syntaktická, sémantická analýza
- předpoklad pro navazující aplikace
 - např. *Word Sketch*, *Morfio*
- zapojení do dalších nástrojů pro práci s jazykem
 - kontrola pravopisu, slovníky, webové prohlížeče
- možnost adaptace pro jiné slovanské jazyky

Základní pojmy

- **tag** (*morfologická značka, index*)
 - kód přiřazený k jednotlivým tvarům slov nesoucí informaci o jejich morfologických charakteristikách
- **tagset**
 - soubor používaných morfologických značek
- **značkování** (*tagování, tagging, anotace, indexování*)
 - automatické přiřazení lemmatu a tagu
 - přiřazení všech interpretací daného tvaru (homonymie)
 - *ženu – žena – k1gFnSc4*
 - *ženu – hnát – k5eAaImIp1nS*
 - *Sním je místo něho.*

Základní pojmy

- **morfológický analyzátor** (*tagger*), obsahuje slovník
- **desambiguace** (*disambiguace, disambiguation*)
 - zjednoznačnění, výběr správné morfológické značky v závislosti na kontextu slova
 - **pravidlová** – tvoří lingvisté, nebo se vyvozují automaticky (Ajka + desamb, Majka)
 - **statistická** (Morče)
 - **hybridní**
 - pravidla + statistika (MorphoDiTa)
 - **pravidla + neuronové síť** (MorphoDiTa)
- **guesser** – nástroj, který analyzuje neznámé tvary

Vyhodnocování úspěšnosti

- **pokrytí/recall** = (v %) = poměr získaných výsledků ke všem možným výsledkům
 - **přesnost/precision** (v %) = poměr výsledků získaných správně ke všem nalezeným výsledkům
1. *true positives*, **TP** – relevantní výsledky
 2. *false positives*, **FP** – nesprávné výsledky
 3. *false negatives*, **FN** – nesprávná vynechání
 4. *true negatives*, **TN** – správná vynechání

Systemy morfologických značek pro češtinu

- **poziční systém** (ČNK)
 - Jan Hajič, Jaroslava Hlaváčová, ÚFAL MFF UK
 - tagger **MORČE** (MORfologie ČEštiny)
 - včetně desambiguace (pravděpodobnostní model), Jan Raab
 - tagger **MorphoDiTa**, morfologický slovník **MorfFlex**
 - Milan Straka, Jana Straková, ÚFAL MFF UK
 - značky mají **16 pozic**
 - ženu **NNFS4-----A-----**

Systemy morfologických značek pro češtinu

- **atributivní systém** (korpusy ve Sketch Engine)
 - Klára Osolobě (FF MU) – algoritmický popis morfologie
 - tagger **AJKA** (Analyzátor JazyKA), Radek Sedláček (FI MU) + **Desamb**
 - tagger **MAJKA** (Morfologický Analyzátor JazyKa) – Pavel Šmerk, odlišný přístup (FI MU)
 - systém **atribut – hodnota**
 - např. atribut **c** s hodnotami **1–7**
 - ženu **k1gFnSc4**

Problémy

- pojetí slovních druhů, slovnědruhové přechody, forma – význam
- co je lemma
 - jednoslovný základní tvar
 - MWE – Multiword Expressions
- homonymie (*nominativ – akuzativ*)
- nedostatečný slovník
- neznámá slova a guesser
- mluvené korpusy, korpusy korespondence

Desambiguace

- Některé tvary nelze desambiguovat – není možné jednoznačně vybrat správnou značku ani na základě kontextu

Německá firma Tebis v Hannoveru představila kompaktní zařízení pro firemní modelárny.

Technické řešení těsnění nádrží a podlah...

Myrha je přírodní pryskyřice, aloe je vonné dřevo.

V osmi letech měl za sebou účinkování v mnoha televizních show...

Dolní listy jsou obvejčité, čepel se zužuje v ouškatý řapík.

Jak lze z názvu vytušit, jde o nástroje pro zprostředkování databázových transakcí a tvorbu dotazů prostřednictvím standardu SQL.

Jak nám řekl ředitel tohoto závodu, nebyla to jejich chyba...

jak – k1, k6, k8, k9

Odkazy

- <http://nlp.fi.muni.cz/projekty/wwwajka>
- CQL [tag=„“]
 - *Corpus Query Language*