

CJBB105 – 6

Morfologické značkování

Mgr. Dana Hlaváčková, Ph.D.

CJBB105

PRZA009

Morfologické značkování

- mnoho korpusů v mnoha jazycích je obohaceno o morfologické informace o tvarech slov
- tímto značkováním se zvyšuje informační hodnota korpusu a usnadňuje se hledání relevantních informací o užívání jazyka
- jedná se o nejrozšířenější typ vnitrotextového značkování v korpusech

Morfologické značkování

- ke každému tokenu je v označovaném korpusu přiřazen kód, který nese morfologickou informaci o tvaru slova
- např. v ČNK: **koček** *kočka*/NNFP2-----A-----
 - tvar *koček* má lemma *kočka* a jde o substantivum, ženského rodu, v genitivu plurálu, afirmativ (= bez negativní předpony *ne-*)
- co musí uživatel vědět:
 - jak značky vypadají a co kódy znamenají
 - které morfologické kategorie v nich najde
 - které další informace v nich najde (některé značky kombinují informaci o tvaru i o významu)
 - že nejsou v automaticky značkových korpusech přiřazeny 100% správně

Morfologické značkování

- proces automatického značkování
- **token – lemma – tag**
- **tokenizace** – rozdělení textu na pozice/tokeny
- **lemmatizace**
 - přiřazení základního slovního tvaru (jednoslovné) = **lemma**
 - pro subst. – nom.sg., pro adj. – nom. sg. masc., pro verb. – infinitiv
 - problém – význam spojený s tvarem v ustálených slovních spojeních a idiomech – jít na *nervy* (lemma *nerv*), *nechal ho na holičkách* (lemma *holička*), pozdě *bycha* honit (lemma *bycha*) – často neodpovídající nebo uměle vytvořená lemmata
- **tagging**
 - přiřazení morfologické značky (všechny interpretace tvaru **nezávisle na kontextu**, vliv homonymie) = **tag**

Morfologické značkování

- problém – **homonymie**, automatické nástroje v procesu taggingu neberou ohled na kontext
 - např. *ženu* – lemma *žena* (substantivum) i lemma *hnát* (sloveso) – jsou přiřazena všechna lemmata a všechny tagy
- nastupuje další fáze – **desambiguace** – zjednoznačnění lemmat a tagů na základě kontextu
- **tagger** – nástroj, který provádí morfologickou **analýzu** (lemmatizace a přiřazení všech odpovídajících tagů) a **desambiguaci** (výběr správného tagu)
- **„zlaté pravidlo morfologie“** – přiřazení tvaru, lemmatu a tagu je jednoznačné

Morfologické značkování

- na úrovni **slovních druhů**
 - PoS tagging (angličtina)
 - v češtině např. u neohebných slovních druhů (spojky, částice, citoslovce)
 - adverbia – značena navíc negace a stupeň
- **kompletní značkování**
 - všechny morfologické kategorie (slovanské/flektivní jazyky, jazyky s bohatou morfologií)
 - ohebné slovní druhy
 - nutné pro další stupně automatického zpracování jazyka (např. pro syntaktickou analýzu) a navazující aplikace v korpusových manažerech

Morfologické značky

- musí být **transparentní** – tagset (srozumitelné vysvětlení sady tagů)
 - jednoznačná interpretace značky
- zachycují především **morfologické** charakteristiky
 - ale také sémantické vlastnosti (např. druhy zájmen a adverbii)
- musí být **nezávislé** na různých lingvistických teoriích (často se volí úroveň středoškolských znalostí, případné odchylky musí být vysvětleny)
- musí být **orientované na uživatele** a současně **strojově čitelné**
- častá podoba – **kód** sestavený z písmen a čísel
 - **kočka/kočka/NNFS1-----A-----** – ČNK
 - **kočka /kočka/k1gFnSc1** – korpusy na MU v Brně
 - **kot** [kot:subst:sg:nom:m2] – polština
 - **cat /NN/cat** – angličtina
 - **Katze /N.Reg.Nom.Sg.Fem/Katze** – němčina

Homonymie

- v češtině a podobných jazycích znesnadňuje celý proces značkování
- **významová** – obvykle není rozdíl v morfologických kategoriích
 - *koruna* (na hlavě/mince), *sladit* (uvést v soulad/činit sladkým – zde je rozdíl ve vidu)
- **tvárová** – nejfrekventovanější
 - *jarní* (stejný tvar pro různý rod, číslo, pád)
- **slovnědruhová**
 - *jak* (adverbium, spojka, částice) – často těžko rozlišitelné i pro lingvistu
- **kombinovaná**
 - *ženu* (*subst., f, ak., sg./verb., 1. os., sg.*)
- může se kombinovat i v celé větě
- *Sním je místo něho.* – každý tvar v této větě je homonymní
- *Praštil se sluchátkem.* *se* – předložka/zvratné zájmeno – dva různé významy věty

Metody automatického značkování

- morfologické značkování včetně desambiguace
- závisí na velikosti a kvalitě morfologického slovníku
- **Stochastické** (statistické, pravděpodobnostní)
 - založeno na strojovém učení (na referenčních datech)
 - systém se sám učí na základě správně označovaného korpusu
 - aktuálně se začínají využívat i neuronové sítě se slibným výsledkem
- **Pravidlové**
 - využívá pravidla stanovená lingvisty nebo vyvozená z textu
 - pozitivní i negativní pravidla (např. co se může/nemůže vyskytnout ve větě vedle sebe)
- **Hybridní**
 - kombinace obou přístupů, nejúspěšnější

Metody automatického značkování

- v textu mohou být neznámé tvary, které systém nemá ve slovníku
 - nástroj **guesser** – automaticky odhadne možné lemma a tag
 - často se netrefí – např. v brněnských korpusech *mývalí* kočka (lemma *mývalit*)
- úspěšnost taggerů až 98 %, měří se:
 - **pokrytí** (recall) – kolik tokenů dostalo značku (může být až 100%, nějakou značku dostává i interpunkce, čísla, znaky, neznámá slova)
 - **přesnost** (precision) – kolik značek je správných (nikdy není 100 %)

Morfologická analýza v ČR

- v ČR existují **dva systémy značkování** (Praha, Brno)
- za základní a rozšířenější se považuje pražský systém
 - také má vyšší úspěšnost, udržovaný slovník a nyní se do analýzy zapojují neuronové sítě
- brněnský systém je často označen za srozumitelnější
 - jsou jím označovány velké miliardové korpusy na MU
- pro českého lingvistu je výhodou znát oba systémy a moci pracovat se všemi dostupnými českými korpusy

Morfologická analýza v ČR

- ÚČNK Praha – ČNK, manažer KonText
 - Ústav formální a aplikované lingvistiky MFF UK
 - Ústav teoretické a počítačové lingvistiky FF UK
- **poziční systém**
 - značka se skládá z 16 pozic, každá vyjadřuje jednu morfologickou charakteristiku
 - 2 rezervní (13. a 14.), 1 stylová (15.), 1 smíšená (2.)
 - SYN2020 – 15 pozic, vid na 13. pozici
- analyzátor hybridní – stochastický i pravidlový **MorphoDiTa** (se slovníkem Morfflex)

, ty kvalitní , na nichž se dá sedět i osm hodin denně , stojí **kolem** /ko1em/RR--2----- 5000 až 6000 korun . Za plně vybaven
běrové řízení na komplexní informační systém , jehož prvním **kolem** /ko1o/NNNS7-----A----- prošli čtyři výrobci . Průběh implemen

Morfologická analýza v ČR

- MU Brno, manažer Sketch Engine
 - Centrum zpracování přirozeného jazyka FI MU
 - Ústav českého jazyka FF MU (formální popis české morfologie – doc. K. Osolsobě)
 - Lexical Computing
- **atributivní systém**
 - **atribut** – morfologická kategorie obecně (např. c = pád)
 - **hodnota** – morfologická kategorie konkrétně (c1–c7)

...vý výsledek nebyl ovlivněn . Druhým **kolem** /kolo/k1gNn5c7 prezidentských voleb se Rusko ve středu
luma požaduje šetření korupce kliky **kolem** /kolem/k7c2 Gračova Jako člověka po uši zapleteného

Syntaktická analýza

- v korpusech SYN2015 a SYN2020
- zobrazení **závislostních vztahů** mezi slovy ve větě
 - **závislostní strom**
- vychází z **PDT** (Prague Dependency Treebank)
z ÚFAL MFF UK
 - manuálně označovaná data
 - východiskem je syntax VI. Šmilauera
 - syntaktický analyzátor (parser) – úspěšnost cca 80 %
- zobrazení v KonTextu, možnosti vyhledávání podle syntaktických atributů (*parent*, *afun*)

Práce s morfologickými značkami

- uživatel – lingvista
- důležitá je znalost tagsetu a principu analýzy
- je pak možné vyhledávání v korpusu podle morfologických charakteristik
- kontrola správnosti značkování
- jazykové a frekvenční studie
- důležitá je schopnost interpretace značky a nalezených informací
- v současnosti probíhá projekt NovaMorf – nové přepracování značkování pro ČNK, nový přístup a nový tagset (výsledky budou v nejbližší době)

Prohlížení lemmat a tagů

- v obou manažerech musíte mít zaškrtnutou možnost zobrazit **lemma** a **tag**
- **KonText** – Zobrazení – Korpusová nastavení
- **Sketch Engine** – Možnosti zobrazení (ikonka oka)
- je také možné nastavit možnost zobrazit lemma a tag **jen pro KWIC** nebo **pro všechny tokeny**

Hledání podle lemmat a tagů

- využívá se dotazovací jazyk **CQL** (Corpus Query Language)
- formální podoba dotazu např.
 - [lemma=„kočka“]
 - [tag=„N.*F.*“] – najdi všechna substantiva v ženském rodě (poziční systém), kombinace .* je **regulární výraz** = jakákoli kombinace znaků, můžeme jím nahradit části značky, které v dotazu nejsou důležité

Hledání podle tagů

- oba manažery pomáhají s konstrukcí dotazu
- **KonText**
 - Typ dotazu – CQL
 - Dotaz – **Vložit tag** (uvidíte celý tagset, který vám pomůže vložit tag), přeskočené části značky se samy nahradí regulárním výrazem .*
 - **popis morfologických značek** (odkaz na tagset)
- **Sketch Engine**
 - Typ dotazu – CQL
 - Vložit (formální znaky)
 - Značky (celý tagset)
 - **CQL Builder** – konstrukce značky

Odkazy

- popis pozičního systému v ČNK
 - <https://wiki.korpus.cz/doku.php/seznamy:tagy>
- Kurz práce s ČNK – 5. lekce
 - [https://wiki.korpus.cz/doku.php/kurz:pokrocile d
otazy](https://wiki.korpus.cz/doku.php/kurz:pokrocile_d
otazy)
- Vyzkoušejte si hledání podle morfologické značky v obou manažerech, korpus SYN2020 a Czech Web 2017