Z Wikipedie, otevřené encyklopedie

Lemmatizace je určení lemmatu (základního slovního tvaru) k ohýbanému slovnímu tvaru. **Lemmatizátor** je nástroj (např. počítačový program), který vytvoří (vyhledá v databázi) k určitému tvaru slova základní tvar, tzv. lemma. [1]

Doplňkovou funkcí lemmatizátoru jsou informace o mluvnických kategoriích (např. jmenných a slovesných) k danému tvaru.[pozn. 1] Např. pro tvar „barvě“ lemmatizátor vrátí tvar „barva“, případně doplňkovou informaci *podstatné jméno, ženský rod, jednotné číslo, 3./6. pád*.

Lemmatizace se např. využívá se pro vyhledávání ve fulltextových databázích. Pro fulltextové vyhledávání se ovšem využívají i podobně strukturovaná data sloužící k automatické kontrole pravopisu (např. slovníky pro hunspell).[3]

Využití lemmatizace

1. Fulltextové vyhledávání – např. pro zadanou frázi „sběrný dvůr“ se vyhledají i dokumenty obsahující tato slova v jiných pádech a číslech (*sběrné dvory*, *umístění sběrných dvorů*).[4]
2. Korpusová lingvistika – [illegible] lingvistického analyzátoru se využívá při značkování korpusů.[5] Lemmatizaci využívá software QUITA (Quantitative Index Text Analyzer), který dokáže posuzovat a analyzovat rozsáhlé texty, např. bohatost slovní zásoby a další lingvistické ukazatele.[6]
3. Dalším nástrojem využívající lemmatizaci je latentní sémantická analýza (LSA). „Latentní sémantická analýza je technika, která zobrazuje dokumenty a dotazy do prostoru latentních sémantických dimenzí, přičemž slova, která jsou sémanticky podobná (měřeno mírou souvýskytů v dokumentech) jsou zobrazována do stejných dimenzí a slova sémanticky odlišná do různých dimenzí.“[7] LSA pro každé slovo vytváří další dimenzi, dokumenty se tak mohou nacházet až v několika [illegible] dimenzí. Lemmati

na základní tvar. Tím se nevytváří různé dimenze pro stejná slova v jiném slovním tvaru. „Díky tomu mohou mít velkou sémantickou podobnost i dokumenty (případně dotaz a dokument), které spolu nesdílejí žádná slova.“[7]

Úskalí lemmatizátoru

Některá slova jsou mnohoznačná (v češtině např. ženu, stát, tancích) a pokud lemmatizátor neposoudí nebo nemůže posoudit kontext, není schopen zvolit zamýšlený význam. Např. „Jeden z nejhodnotnějších zdrojů o maďarských tancích“ zpracuje takto: „Jeden/jíst z hodnotný zdroj o maďarský tank/tanec“.

Obtížným specifikem jsou taktéž *víceslovná spojení*, tj. vytváření lemmat i tam, kde to není možné, např. zdvořilá prosba *Dovolíte?* se nenachází v žádném z registrovaných významů slova dovolit, dále se může jednat o frazémy, např. nechat na holičkách, popř. se jedná o idiomy např. z někoho si vystřelit.[1]

Dostupné lemmatizátory pro češtinu

Neúplný výčet podle bakalářské práce *Lemmatizace češtiny*:[6]

České lemmatizátory

- Ajka[8]
- Majka[9]
- Morče[10]
- MorphoDiTa[11]
- Czech HMM tagger[12]
- Czech "Free" Morphology[13]
- Morfo[14]

Zahraniční lemmatizátory

- Cistern[15] (Lemming[16] + Marmot[17])
- LemmaGen[18]

Ostatní nástroje

- QUITA[19] (Quantitative Indicator Text Analyzer)
- RDRPOSTagger[20] (Ripple Down Rules Part-Of-Speech Tagger) – Tagger založený na Ripple Down Rules

Poznámky

1. ↑ Tento proces (přiřazení morfologických kategorií) se – v technickém smyslu – nazývá morfologická analýza.[2]

Reference

1. ↑ **a** **b** CVRČEK, Václav; RICHTEROVÁ, Olga. *Slovníček pojmů* [online]. Český národní korpus [cit. 2016-06-21]. Kapitola Lemma. Dostupné online.
2. ↑ RUSÍNOVÁ, Zdenka; PETKEVIČ, Vladimír. *Nový encyklopedický slovník češtiny*. Příprava vydání Petr Karlík, Marek Nekula, Jana Pleskalová. Praha: NLN, 2017. ISBN 978-80-7422-480-5. Heslo Morfologická analýza.

13. ↑ Czech "Free" Morphology
14. ↑ Morfo
15. ↑ Cistern
16. ↑ Lemming
17. ↑ Marmot
18. ↑ LemmaGen. *lemmatise.ijs.si* [online]. [cit. 2017-05-11]. Dostupné v archivu pořízeném z originálu dne 2017-06-06.
19. ↑ QUITA
20. ↑ RDRPOSTagger