Posudek diplomové práce

Matematicko-fyzikální fakulta Univerzity Karlovy

Autor práce	Lukáš Kyjánek
Název práce	Harmonisation of Language Resources for Word-Formation of Multiple Languages
Rok odevzdání	2020
Studijní program	Informatika **Studijní obor** Matematická lingvistika

Autor posudku	Daniel Zeman **Role** Oponent
Pracoviště	Ústav formální a aplikované lingvistiky

Text posudku:

Předložená diplomová práce se zabývá datovými zdroji pro počítačové zpracování přirozeného jazyka a pro lingvistický výzkum, konkrétně zdroji zachycujícími slovotvorbu. Autor se zaměřuje na inventarizaci různých již existujících zdrojů pro různé jazyky a na jejich harmonizaci ve smyslu nalezení společné datové struktury, aniž by došlo ke ztrátě informace. Toto téma je nanejvýš aktuální. Harmonizované zdroje umožní aplikační využití uvedených jazykových dat pomocí jedné sady softwarových nástrojů, což doposud nebylo myslitelné.

Práce se skládá ze čtyř kapitol. V první jsou položeny teoretické základy slovotvorby s četnými odkazy do odborné literatury. Druhá kapitola přináší přehled a popis hlavních rysů desítek již existujících slovotvorných databází pro různé jazyky. Třetí kapitola navrhuje postup harmonizace, probírá možné alternativy a vyhodnocuje tu část, která byla řešena pomocí strojového učení. Čtvrtá kapitola představuje výslednou kolekci, vydanou pod názvem Universal Derivations. Strukturu textu hodnotím jako dobře zvolenou. Práce má 68 stran, z nichž 24 je věnováno popisu autorovy vlastní práce v kapitole 3. Jednoznačným přínosem je ale i přehled existujících zdrojů v kapitole 2, který v podobném rozsahu nebyl dosud k dispozici. Práce je psána slušnou a dobře srozumitelnou angličtinou, i když v gramatice a stylu jsou ještě určité rezervy.

Ke zvolenému řešení harmonizace nemám zásadní námitky. Cílová datová struktura je zřejmě společným dílem kolektivu autorů z projektu Derinet, autor práce ale implementoval převod zdrojových formátů do cílového formátu, vyvinul anotační nástroj a natrénoval modely strojového učení, které byly v některých částech transformace dat potřeba.

Konkrétní připomínky a otázky:

Na straně 7 se správně říká, že: „lexeme denotes a set of word forms … whereas … lemma refers to one canonical representative form“. Nicméně dále v práci se tento terminologický rozdíl nedodržuje, autor opakovaně (např. bod 3 na straně 39) píše o „the written form of the lexeme“, což nedává smysl (muselo by jít o sadu všech tvarů lexému, tak to ale určitě ani autor nemyslí).

Strana 41: „To avoid damaging of the original data, no new relations are added but, if it is possible, the unification of regular word-formation relations is done, e.g. in the case of capturing negation, and described during the harmonisation procedure.“ Bohužel, tento odstavec neříká, jaké je tedy řešení a co znamená ona „unification“. Co se tedy děje např. se

zachycením negace? Jsou zachovány jak hrany paralelních stromů, tak hrany od každého afirmativního tvaru k odpovídajícímu negativnímu tvaru? A které z těchto dvou sad hran jsou považovány za primární?

Strana 47: „Most of the resources contain loanwords, see example 6 in Figure 3.4. If possible, they were captured as derivation.“ Co tato věta znamená?

Strana 47: „compounds were disconnected from their base lexemes, except for subsequent derivations of compounds“ Takže se ztratila nějaká informace? Nešlo by ukázat nějaký příklad?

Strana 51, popis rysů pro strojové učení: „initial and final character n-grams“ – jak mohou být n-gramy typu Boolean?

Strana 52, definice recallu: jmenovatel ve zlomku nesedí s popisným textem v předcházejícím odstavci, jmenovatel je správně (true positives + false negatives).

Strana 54: „However, some word-formation families cannot be covered by a single tree“ – zde by to chtělo ukázku reálných dat, aby si čtenář mohl představit, o co jde. (Nebo možná alespoň odkaz na řadu B v obrázku 3.6 na následující straně.)

Strana 55, figure 3.6: Po rozdělení rodin slovo z jedné rodiny úplně vypadlo (difference), není to špatně?

Práci doporučuji k obhajobě.

Práci nenavrhuji na zvláštní ocenění.

Pokud práci navrhujete na zvláštní ocenění (cena děkana apod.), prosím uveďte zde stručné zdůvodnění (vzniklé publikace, významnost tématu, inovativnost práce apod.).

Datum 16. června 2020 **Podpis**