Posudek diplomové práce

Matematicko-fyzikální fakulta Univerzity Karlovy

Autor práce	Jonáš Vidra		
Název práce	Morphological segmentation of Czech Words		
Rok odevzdání	2018		
Studijní program	Informatika	**Studijní obor**	Matematická lingvistika

Autor posudku	David Mareček	**Role**	oponent
Pracoviště	Ústav formální a aplikované lingvistiky		

Text posudku:

Cílem diplomové práce Jonáše Vidry je vytvoření nástroje, který bude využitím derivační sítě DeriNet automaticky segmentovat česká slova na morfémy.

Práce je členěna do sedmi kapitol. Po úvodu do problematiky následuje kapitola definující základní lingvistické pojmy jako, derivace, inflexe, morfém, apod. Třetí kapitola se zabývá souvisejícími pracemi. Nejdříve popisuje existující anotovaná data zachycující derviační vztahy mezi slovy nebo jejich morfologickou segmentaci, dále pak existující nástroje pro neřízenou segmentaci slov. Řízená morfologická segmentace, která by se učila z anotovaných dat se prakticky nedělá a to z důvodů nedostatečných trénovacích dat, které jsou navíc závislé na použitých lingvistických teoriích které jsou ohledně segmentace slov často velmi vágní. Hlavním zdrojem, o který se práce opírá, je DeriNet, ve kterém je více jak milion českých lemmat pospojováno derivačními vztahy do orientovaných stromů.

Ve čtvrté kapitole je navržena metoda segmentace využívající EM algoritmu. Je zde popsán pravděpodobnostní model, který využívá toho, že odvozené slovo zpravidla se svým předkem sdílí kmen (až na drobné hláskové změny) a liší se pouze v předponách a příponách. Popisovaná metoda je však schopna segmentovat pouze slova obsažená v DeriNetu. V šesté kapitole proto navrhuje metodu využívající rekurentních neuronových sítí, která predikuje morfologickou segmentaci a současně hledá předka v derivačním stromě pro jakýkoliv vstup. Neuronová síť se učí řízeně, proto využívá data nasegmentovaná modelem s předchozí kapitoly. V šesté kapitole jsou všechny navržené metody shrnuty a jejich výsledky vyhodnoceny na manuálně anotovaných 14 tisících českých slovesech (Slavíčková, 1975). Poslední kapitola pak práci uzavírá a navrhuje další možný vývoj.

Práce má celkem 54 stránek čistého textu a 9 stránek referencí. Je psána dobře srozumitelnou angličtinou s pouze drobnými chybami nebo překlepy. Čitelnost však zhoršuje hutnost především čtvrté kapitoly. Chybí mi tam podrobný popis (například pseudokód) EM algoritmu, který je jádrem práce. Neznalý čtenář by pouze podle práce asi měl problémy algoritmus implementovat. Některá notace ve vzorcích je trochu zavádějící. Například ve vzorcích 4.25 a 4.26 jsou proměnné u a v jednou znaky a jednou celé affixy. Dále výraz $c(s \to t)$ neudává počet (count), jak bych s označení předpokládal, ale součet pravděpodobností. Navíc proměnné s a t nejsou nikde v práci definovány, předpokládám že má jít o zkratky 'source' a 'target', o kterých se ale v textu píše jako o 'parent' a 'child'. Až do rovnice 4.5 jsou navíc písmenem s označovány také suffixy. Pro lepší čtení by také nebylo na škodu některé důležité věci občas zopakovat, aby se nemusel čtenář pořád vracet zpátky a hledat vysvětlení. I přes tyto nedostatky jsem však, myslím, všemu v práci porozuměl.

Z obsahového hlediska má práce velmi dobrou úroveň. Diplomant prokázal, že umí vytvářet a používat jak pravděpodobnostní modely použité v neřízeném učení, tak plně řízené a dnes velmi populární neuronové sítě. Provedl velké množství experimentů, všechny správně vyhodnotil na testovacích datech a diskutoval výsledky. Autor přiznává že pravděpodobnostní model má některé nedostatky. Ideální by byl čistě generativní model, který však možná není jednoduché v tomto případě najít. Model během práce postupně vylepšuje, aby mu dával dobré výsledky. Například zjišťuje, že je třeba v modelování alternací kmene nahradit složenou pravděpodobnost za podmíněnou, přidává vyhlazování, nebo ručně prořezává vzniklou tabulku afixů. V přehledné tabulce ukazuje, že má lepší výsledky než často používaný nástroj Morfessor, který pro své učení využívá velkých neanotovaných korpusů a dokazuje tak, že s pomocí derivačních sítí lze dosáhnout kvalitnější morfologocké segmentace.

Dotazy a poznámky:

1. V analýze chyb zmiňujete, že se slovo dvojice *oko* $\to$ *očko* špatně segmentuje na *o.ko* $\to$ *o.čko*. Nehodilo by se, podobně jako prořezáváte nesprávné affixy, prořezávat i nesprávné substituce? Šlo by to i automaticky, například vyhodit všechny substituce mezi samohláskami a souhláskami, které by se, myslím, vyskytovat nikde neměly.

2. Zkoušel jste použít podmíněnou pravděpodobnost i pro affixy? Přijde mi v tomto případě přirozenější. Chápu ale, že vyšší pravděpodobnosti afixů mohou způsobovat zkracování kmenů.

3. Pro modelování alternací bych navrhoval sloučit substituci s mazáním (přidat do substituce znak λ). Vkládání bych pak modeloval pomocí distribuce obsahující i prázdný znak (žádné vložení), který by byl samozřejmě nejpravděpodobnější, a aplikoval bych ji na všechny možné pozice v kmeni.

Práci doporučuji k obhajobě.

Práci nenavrhuji na zvláštní ocenění.

V Praze dne 4. 9. 2018

Podpis: