Oponentský posudek disertační práce

Název:	*Discovering the structure of natural language sentences by semi-supervised methods*
Autor:	Rudolf Rosa
Obor:	matematická lingvistika, MFF UK
Rozsah:	133 stran textu

Práce se zabývá výrazně aktuálními tématy z oblasti syntaktické analýzy jazyků s nedostatečnými jazykovými zdroji (*under-resourced languages*) využívající zdroje z jiných jazyků (*cross-lingual parsing*). Autor v práci nejprve předkládá velice podrobný a čtivý přehled historického vývoje tvorby datových zdrojů a technik v dané oblasti, poté prezentuje vlastní experimenty založené na důsledném zpracování posledních mezinárodních výsledků. V první části (kap. 3 a 4) je představena míra KL_{cpos^3} jako jeden z hlavních výsledků celé práce a jsou prezentovány výsledky analýz bez určení typů hran (*unlabelled*). V navazujících částech jsou pak replikovány experimenty s lexikalizací přenesené analýzy a s přenosem morfologického značkování (také výrazně využívající navrženou vlastní podobnostní míru).

Text práce je psán vynikající angličtinou, je v něm vidět velký důraz na kompletnost a preciznost. V textu se téměř nevyskytují překlepy ani gramatické chyby. Většina otázek, na které čtenář při čtení narazí, je následně alespoň stručně zodpovězena.

Přednosti a přínos práce:

- přehledný a podrobný souhrn historického vývoje v oblasti *cross-lingual parsing*,
- důsledné a podrobné vyhodnocení jednotlivých experimentů,
- mezinárodně kompetitivní výsledky s využitím navržené míry podobnosti stromových korpusů (*treebanks*) KL_{cpos^3} a experimenty přenosu analýzy blízkých jazyků.

Nedostatky a nejasnosti v práci:

a) Potenciálně problematický se v práci jeví fakt, že prezentované experimenty zřejmě spadají do víceletého období a výsledky v klíčových kapitolách 3 a 4 jsou založené na poměrně omezeném a dnes překonaném nástroji *MSTperl*. Některé z experimentů jsou v dalších kapitolách opakovány i se současnými nástroji, ale např. základní vyhodnocení míry KL_{cpos^3} je převzato. Z textu je občas zřetelně vidět použití starších textů ("now-emerging Enhanced UD" z roku 2016).

b) Všechny uvedené experimenty se opírají exkluzivně o výsledky porovnání shody s dostupnými stromovými korpusy. Z praktického hlediska je tento přístup pochopitelný a dostupný, z širšího pohledu

ale v práci chybí alespoň zmínka o vnějším vyhodnocení syntaktické analýzy (*extrinsic evaluation*). Zejména v situacích, kdy nastavení použitých měr nebo parametrů není zcela průkazné napříč jazyky, může takové vyhodnocení poskytnout jiný pohled na situaci a ukázat, kdy je rozdíl ve výsledcích podstatný a kdy ne. Při striktním spoléhání pouze na uvedené zdroje mohou být výsledky ovlivněny zcela neočekávanými příčinami, jak autor sám na několika místech ukazuje (závislostní anotace neslovních uzlů Prague vs. Stanford, odlišné lemmatizace negací v českém a slovenském korpusu).

c) V rámci velmi dobrého stylu autora působí rušivě nadužívání termínu *noise/noisy* (cca třicetkrát). Autor sice na str. 42 upřesňuje popis zamýšleného užití tohoto termínu, ale na čtenáře působí spíše jako označení situací, které by vyžadovaly podrobnější zkoumání jiných přístupů (např. str. 86, kde je tento termín použitý sedmkrát).

d) Rovnice (4.3) na str. 60 (hlavní výsledek práce) neuvádí, jaká je hodnota míry pro $\hat{P}_{src}(cpos^3) = 0$. Dále v textu se zmiňuje použití "add 1" vyhlazování, správný vzorec by tedy měl obsahovat toto přičtení.

e) Na str. 70 autor říká, že vážená kombinace stromů je v provedeném experimentu horší pouze v případě japonštiny, z tabulky 4.5 ale vychází lépe varianta s jedním zdrojem také u perštiny.

Otázky k obhajobě:

1. U lexikalizovaných technik autor porovnává různé způsoby využití strojového překladu v rámci mezijazykového přenosu. Nikde v textu se ovšem nezmiňuje možnost využití dnes naprosto převažujícího strojového překladu s využitím neuronových sítí (NMT). V architekturách typu *encoder-decoder with attention* přitom nepřímo dochází k využití interní reprezentace věty v podobě jazykově nezávislé reprezentace, tedy to, co autor zmiňuje (nedůvěřivě) na str. 87 jako *pivot language*. Část sítě označovaná jako *attention* poskytuje informace o zarovnání slov ve zdrojovém a cílovém jazyce. Navíc u NMT je prokázáno právě kombinované využití a přenos jazykových informací mezi různými jazyky. Zvažoval autor tento nebo podobný přístup?

2. Použitá míra KL_{cpos^3} je poměrně jednoduchá (což je pozitivní), v uvedených experimentech ale většinou nepokrývá optimálně všechny jazyky. Proč v ní (nebo ve výpočtu vah) nejsou zahrnuté další důležité faktické parametry jako třeba dostupné velikosti korpusů?

3. Na str. 77 autor uvádí, že v případě přechodového analyzátoru (*transition-based parser*) není žádný jednoduchý způsob, jak získat skóre jednotlivých hran. Nebylo by možné toto skóre odvodit z hodnot interních skóre analyzátoru pro konkrétní operace "left-arc" a "right-arc"?

4. Na str. 86 je v rámci (správného) odůvodnění zarovnání přeložených slov 1:1 zmíněno, že je triviální zahrnout další texty v cílovém jazyce do procesu překladu zdrojových stromových korpusů. Jak přesně?

Závěrečné hodnocení:

Práce se zabývá výrazně aktuálními tématy z oblasti syntaktické analýzy jazyků s nedostatečnými jazykovými zdroji s využitím zdrojů z jiných jazyků. Výsledky práce spočívají v precizní analýze, nastavení a vyhodnocení přístupů k tomuto tématu na velkém množství dat dostupných v rámci kolekce *Universal Dependencies 1.4*. Mezi vlastní kompetitivní výzkumné výsledky autora patří zejména návrh míry KL_{cpos^3} pro vyhodnocení podobnosti zdrojových syntaktických korpusů a podrobný návrh a vyhodnocení technik pro analýzu blízkých jazyků.

Celkově konstatuji, že předložená dizertační práce **prokazuje** předpoklady autora k samostatné tvořivé práci a představuje vhodný podklad pro udělení titulu Ph.D.

V Brně dne 28. května 2018

doc. RNDr. Aleš Horák, Ph.D.
Fakulta informatiky
Masarykova univerzita, Brno