Posudek diplomové práce

Matematicko-fyzikální fakulta Univerzity Karlovy

Autor práce	Ján Faryad
Název práce	Extraction of multilingual valency frames from dependency treebanks
Rok odevzdání	2024
Studijní program	Informatika **Studijní obor** Matematická lingvistika

Autor posudku	RNDr. Daniel Zeman, Ph.D. **Role** Vedoucí
Pracoviště	Ústav formální a aplikované lingvistiky

Text posudku:

Předložená práce zkoumá možnosti extrakce informací o valenci sloves ze závislostních korpusů pro různé jazyky. Konkrétně autor pracuje s kolekcí Universal Dependencies (UD). Své postupy testuje na češtině, slovenštině a angličtině s tím, že výsledný softwarový nástroj lze použít na libovolný další jazyk v UD. První část práce se věnuje extrakci informací z jednojazyčných dat: jak identifikovat slovesa, jak mezi jejich závislými uzly identifikovat argumenty a jakými vlastnostmi tyto argumenty zachytit v rámci slovesa. Druhá část zkoumá možnosti propojení takto získaných rámců mezi dvěma jazyky tak, aby bylo možné paralelně zobrazit sloveso a rámec s odpovídajícím významem, včetně propojení sobě odpovídajících argumentů. Potenciálním využitím takového propojení je rozšíření jazykových zdrojů na nové jazyky, např. projekce funktorů z českého valenčního slovníku Vallex do slovenštiny.

I když myšlenka extrakce valenčních informací z treebanku není nová, její zacílení na mnohojazyčnou kolekci treebanků nové je. Přínosem práce je už samotný rozbor jazykových jevů a anotačních rozdílů, které mohou vést k odlišným výsledkům při aplikaci algoritmu na různé jazyky. Druhým přínosem je softwarový nástroj, který identifikuje slovesné rámce, mapuje je na zdroje typu Vallex, pokud existují, a mapuje na sebe odpovídající rámce napříč jazyky. Takový nástroj je užitečný pro vizualizaci dat při srovnávacím lingvistickém výzkumu (vizualizace pomocí HTML je také jedním z výstupů nástroje). Data na výstupu se samozřejmě nemohou měřit kvalitou a hloubkou anotace s ručně vytvořenými valenčními slovníky, to je dáno i povahou vstupních dat; na druhou stranu se otevírá cesta pro přibližnou projekci ručních slovníků na nové jazyky.

Velkým plusem je striktně modulární architektura a široká konfigurovatelnost nástroje, který lze podle potřeby přizpůsobovat vstupním datům a novým jazykům.

Práce má 123 stran, z toho 75 stran připadá na stěžejní kapitoly 2 a 3, které popisují autorovy experimenty. Organizaci textu do kapitol a oddílů hodnotím jako převážně dobře zvolenou, s jednou výjimkou: Vzhledem k množství výsledkových tabulek objevujících se už v průběhu kapitoly 2 (což samo o sobě hodnotím kladně) by bylo vhodné popsat metody vyhodnocení experimentů už na začátku této kapitoly, ne až na jejím konci v části 2.7. Důraz na pečlivé vyhodnocení všech experimentů a alternativních postupů je jinak velkou devizou celého textu, stejně jako řada ilustračních příkladů nejen ze tří hlavních zkoumaných jazyků, ale i z mnohých dalších. Práce s barvami v tabulkách a grafická podoba doprovodných schémat naznačuje autorovu schopnost srozumitelně vysvětlit, co přesně se dělá a proč; tento dojem je ale bohužel zkalen zjevným spěchem, ve kterém finální podoba textu vznikala. Jde jednak o jazykovou podobu – práce je psána angličtinou, která je sice poměrně dobře srozumitelná, ale

zanesená nepříjemným množstvím překlepů a jiných chyb – a jednak o příležitostné chyby v sazbě (např. tabulka 2.7 na straně 58 nebo opakované vadné odkazy, snad na původně plánovanou čtvrtou kapitolu).

Autor odvedl velké množství práce, a i když se do odevzdaného textu nakonec část z ní nevešla, to, co se vešlo, je stále rozsahem spíše nadprůměrné. Autor prokázal jak orientaci v lingvistických teoriích, tak schopnost návrhu robustního modulárního softwarového řešení. Podle mého názoru předkládaná práce splňuje standardy oboru a doporučuji ji k obhajobě.

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 2. února 2024 **Podpis**