Vyjádření vedoucího k disertační práci

Jana Straková: Neural Network Based Named Entity Recognition

Jan Straková se ve své práci zaměřila na rozpoznávání pojmenovaných entit, což je jedna ze základních úloh ve zpracování přirozeného jazyka, nezbytná pro následné úlohy porozumění přirozenému jazyku zejména v oblasti vazby na reálný svět. Použití umělých neuronových sítí pro tuto úlohu je motivováno výbornými výsledky, kterých tato metoda dosáhla i v jiných úlohách zpracování přirozeného jazyka v poslední době, a je tedy zcela na místě, že autorka ji pro své experimenty využívala.

Práce má 110 stran a je členěna na obvyklý úvod, šest obsahových kapitol, závěr, poděkování, literaturu, seznamy obrázků, tabulek a zkratek a dvě krátké přílohy. V úvodu autorka popisuje problém rozpoznávání pojmenovaných entit, shrnuje vlastní příspěvek k problému, popisuje organizaci disertační práce a poskytuje i krátký „návod“ pro čtenáře podle jeho zájmu. Ve druhé kapitole popisuje problém rozpoznávání pojmenovaných entit, zejména v češtině, možné metody a obvyklé postupy. Ve třetí kapitole se věnuje českému korpusu s manuální anotací pojmenovaných entit (CNEC), který používala pro experimenty v češtině. Ve čtvrté kapitole pak popisuje nezbytné základy umělých neuronových sítí, ale také vlastní experimenty s úpravami architektury těchto sítí (tzv. drop-in, kterým nahrazuje loglineární model), včetně softmax výstupní vrstvy trénovaný metodou klesání podle gradientu. V páté kapitole popisuje jednu sadu experimentů rozpoznávání pojmenovaných entit z roku 2013, a v šesté kapitole pak vytvořený Open Source software NameTag, který je na této metodě založen. V sedmé kapitole prezentuje nejnovější výsledky založené právě na použití umělých neuronových sítí včetně slovních i znakových vektorů a bez použití rysů (features). Osmá kapitola shrnuje výsledky práce, dotýká se kvality a složení použitého korpusu a ukazuje možné pokračování výzkumu směrem k navazující úloze tzv. Named Entity Linking (navázání pojmenovaných entit na strukturované databáze znalostí), a na úplný závěr znovu shrnuje vlastní přínos disertační práce k dané problematice. Poté následuje seznam grantů, které daný výzkum podporovaly, použitá literatura, seznamy obrázků, tabulek a zkratek, a dvě přílohy obsahující nastavení pro loglineární klasifikátor popsaný v kap. 4 a seznamy rysů (features) pro všechny systémy popsané v práci. Je třeba rovněž uvést, že kap. 2, 3, 5, 6 a 7 jsou založeny na publikovaných pracích, disertace je jinak pojata jako samostatná práce (tj. nejde o kolekci publikací).

Hodnocení:

Práce je psána dobrou angličtinou. Úvodní přehledová kapitola je v zásadě převzata a adaptována z autorčina velmi dobře zpracovaného encyklopedického hesla z velké české jazykovědné encyklopedie P. Karlíka, o které byla před časem požádána. Kapitoly 3 až 7 popisují vlastní práci autorky při experimentech i přípravě a zpracování dat. Všechny tyto kapitoly jsou psány velmi srozumitelně a obsahují dostatečnou informaci o implementaci zvolených metod. Vzhledem k tomu, že v dnešní době je téměř jakákoli práce v oblasti zpracování přirozeného jazyka týmovou prací, nebo přebírá osvědčené moduly nebo systémy realizované jako Open Source knihovny a nástroje, autorka také spolupracovala na experimentech s dalšími členy týmu na ÚFAL MFF UK a používala i Open Source software (např. Lua, Torch). Svůj přesný podíl a podíl dalších členů týmu na výzkumu, experimentech i implementaci autorka přesně definuje na konci kapitoly 1.1. Je třeba rovněž konstatovat, že autorka v disertační práci popisuje pouze tu část svého výzkumu v rámci doktorandského studia, která se bezprostředně týkala rozpoznávání pojmenovaných entit, ale v době tohoto studia se podílela na vynikajících výsledcích i v dalších oblastech oboru, jako je Semantic Role Labeling a zejména pak výzkum neuroligvistický, ve kterém i publikovala v ceněném časopise. Rovněž v práci nejsou uvedeny starší experimenty v oblasti rozpoznávání pojmenovaných entit, neboť dle vlastních slov autorky metodologicky

zastaraly (byť byly v době vzniku rovněž publikovány). Z formálního hlediska práce rovněž vyhovuje všem požadavkům a standardům ohledně disertační práce: formát práce je obvyklý, grafická úprava dobrá a pomáhá srozumitelnosti, seznamy obrázků, tabulek a zkratek úplné, seznam literatury je relevantní a více než dostatečný.

Dotazy:

1. V závěru je zmínka o některých potenciálně negativních vlastnostech korpusu CNEC. Navrhla byste dnes postupovat při výběru a anotaci takového korpusu jinak, ovšem s ohledem na efektivitu manuální anotace?
2. Práce směřuje (zejména v kap. 7) k dnes často studovaným „end-to-end“ systémům založeným na umělých neuronových sítích. Jaký je Váš názor na propojení rozpoznávání pojmenovaných entit a napojení na databáze znalostí (Named Entity Linking) z tohoto hlediska – budou to dva navazující systémy s rozdílnými metodami, nebo to rovněž bude směřovat k end-to-end systému, kde krok rozpoznávání bude spíše „skryt“ v jedné velké síti? Jaké důsledky budou mít různá řešení pro multilingvální systémy nebo dokonce jazykově nezávislé systémy?

Závěr:

Disertační práce Jany Strakové, zejména vzhledem ke komplexnosti a objemu provedených experimentů, navržených úprav v oblasti umělých neuronových sítí a výborných výsledků ve srovnání se „state of the art“ v oblasti rozpoznávání pojmenovaných entit, a v neposlední řadě proto, že její výsledky již v nadstandardním počtu článků publikovala, splňuje všechna kritéria kladená na disertační práci v oboru Matematická lingvistika na MFF UK v Praze a doporučuji ji proto k obhájení.

.....

Praha, 20. 5. 2017, Jan Hajič, ÚFAL MFF UK