Posudek diplomové práce

Matematicko-fyzikální fakulta Univerzity Karlovy

Autor práce	Zuzana Šimečková
Název práce	Extrakce vztahů mezi entitami
Rok odevzdání	2020
Studijní program	Informatika **Studijní obor** Umělá inteligence

Autor posudku	Milan Straka **Role** Vedoucí
Pracoviště	Ústav formální a aplikované lingvistiky

Text posudku:

Rozpoznání entit (a speciálně pojmenovaných entit) je jedna z klíčových úloh zpracování textu – v dialogových systémech či kdykoliv v rámci porozumění textu dovoluje odhalit, o jakých osobách, místech či organizacích je pojednáváno. Kromě samotných entit je také důležité dokázat odhalit vzájemné vztahy mezi entitami (a tedy moci reprezentovat vstupní dokument jako graf entit a jejich relací). Tématem diplomové práce byl právě návrh a implementace systému pro extrakci relací mezi entitami v češtině.

Úloha extrakce relací je ve světovém měřítku značně zkoumána, avšak primárně pouze v anglickém jazyce. První a asi nejdůležitější cíl diplomové práce bylo tedy vytvořit český dataset pro extrakci relací. Na jeho základě byl poté navržen a natrénován systém na bázi hlubokých neuronových sítí, který byl kvůli porovnání s existujícími řešeními vyhodnocen kromě vzniklého datasetu i na existujících datasetech anglických.

Práci považuji za velmi zdařilou. Pro tvorbu datasetu, která je popsána v kapitole 2 a 3, studentka zvolila způsob „distant supervision“, tj. využití veřejně dostupné Wikipedie s označenými entitami a znalostní databází Wikidat s označenými relacemi mezi entitami. Touto metodou sice vzniknou zašuměná data, avšak je možné vytvořit velmi velký dataset (což je vhodné vzhledem k následnému použití neuronových sítí) a po drobných úpravách by bylo možné vytvořit data v mnoha dalších jazycích. Tvorba datasetu vyžadovala podstatné množství práce, při které bylo mimo jiné potřeba zpracovat velké kvantity dat, které bylo nutné provádět distribuovaným způsobem. Technické detaily a použité technologie zvoleného řešení jsou podrobně popsané v diplomové práci. Zároveň bylo nutné samostatně provést množství voleb způsobu tvorby nového datasetu. Východiska, možné způsoby řešení a motivaci provedených rozhodnutí studentka také podrobně popisuje v diplomové práci, což velmi oceňuji. Samotný dataset včetně zdrojových kódů je k dispozici nejen jako příloha práce, ale také ve veřejném repozitáři lingvistických dat.

V druhé části diplomové práce jsou popsány metriky používané ve vyhodnocení úlohy extrakcí relací a přehled architektury nejlepších známých řešení. Současné nejlepší modely jsou založené na předtrénovaných modelech kontextualizovaných embeddingů slov, což je jeden z převratných objevů přelomu roku 2018/2019. Řešitelka pomocí této technologie navrhla a implementovala model extrakce relací, který poté vyhodnotila kromě českého na několika anglických datasetech – přestože tento model používá pouze textová trénovací data a je použitelný nejen pro angličtinu, dosahuje na existujících datasetech velmi dobré úspěšnosti v porovnání se specializovanými modely. Natrénovaný model je opět veřejně k dispozici.

Práce je psána srozumitelnou angličtinou. Přestože obsahuje více jazykových chyb, než kdyby byla psána česky, považuji volbu angličtiny za výhodu, díky které může mít práce i mezinárodní dosah.

Celkově hodnotím diplomovou práci jako velmi povedenou a prokazující schopnost samostatné výzkumné činnosti, vzhledem k tomu, že spojuje analýzu a zpracování velkého množství dat pomocí moderních technických prostředků a současně návrh a implementaci pokročilých modelů strojového učení na bázi hlubokých neuronových sítí pomocí GPU akcelerátorů.

Práci doporučuji **k obhajobě.**

Práci nenavrhuji **na zvláštní ocenění.**

Datum 4. září 2020 **Podpis**