Posudek diplomové práce

Matematicko-fyzikální fakulta Univerzity Karlovy

Autor práce	Kateřina Macková
Název práce	Mezijazykový přenos znalostí v úloze odpovídání na otázky
Rok odevzdání	2020
Studijní program	Informatika **Studijní obor** Umělá inteligence

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

Text posudku:

Cílem diplomové práce bylo natrénovat systém pro úlohu „čtení s porozuměním“ v českém jazyce. V úloze „čtení s porozuměním“ obdrží systém text a otázku k tomuto textu, a má na tuto otázku vyznačit v daném textu odpověď. Jedná se tedy o zjednodušenou variantu systému pro odpovídání na otázky. I když se jedná o úlohu často řešenou na mezinárodní úrovni, existující datasety jsou primárně v anglickém jazyce a pro češtinu žádný rozsáhlejší dataset neexistuje. Cílem práce tedy bylo navrhnout a vyhodnotit metody, které dokáží systém natrénovat za pomoci anglických dat.

První část práce shrnuje současný stav poznání v této oblasti. Kapitola 3 prezentuje přehled existujících anglických dat a detailní popis datasetu SQuAD, který je v publikacích nejčastěji používaný. Kapitoly 4 a 5 popisují dva modely hlubokých neuronových sítí řešící cílovou úlohu. První z nich je BiDirectional Attention Flow, což je standardní model využívající rekurentních neuronových sítí, který dosahoval v roce 2016 nejlepších známých výsledků. Druhý popsaný model je řešení založené na modelu BERT, které bylo představeno na konci roku 2018 a dosáhlo tehdy nejlepších známých výsledků. Toto řešení je navíc možné použít s předtrénovaným modelem mBERT, který vznikl zpracováním prostého textu Wikipedie ve ~100 nejčastějších jazycích.

Samotný příspěvek této práce shrnují kapitoly 6 a 7. V kapitole 6 je popsán způsob konstrukce české varianty datasetu SQuAD pomocí nejlepšího známého překladového systému mezi angličtinou a češtinou. Text, otázka i odpověď jsou přeloženy a přeložená odpověď je poté vyhledána v přeloženém textu, protože v rámci SQuAD musí být odpověď vždy souvislý úsek vstupního textu. Přeložená odpověď se samozřejmě v přeloženém původním textu nemusí objevit doslova, ať už kvůli skloňování, použití synonym či chybám v překladu. Ideální řešení by byla ruční kontrola přeložených otázek, která je ale vzhledem k množství (300 tisíc slov) mimo rozsah diplomové práce. Místo toho je využité přímočaré nalezení nejpodobnějšího úseku daného textu přeložené odpovědi s co největší znakovou shodou. Pro vyhodnocení systémů je pak použita podmnožina dat se 100% shodou, pro trénování se uvažuje jak podmnožina dat se 100% shodou, tak odpovědi se shodou alespoň 80%. Tento způsob nalezení odpovědí je nezávislý na jazyce a data se 100% shodou jsou velmi pravděpodobně korektní (což je vhodné pro evaluaci), jen je dat s perfektní shodou menší množství, než by bylo možné dosáhnout pokročilejšími metodami – nabízí se použití lemmatizace, která je navíc již využita při vyhodnocení. K popsané perfektní shodě nicméně dojde přibližně v polovině případů, což považuji za dostatečné.

Těžiště práce je v kapitole 7, která popisuje a vyhodnocuje několik navržených systémů. Nejprve jsou popsána dvě přímočará řešení využívající překladový systém – první z nich trénuje systém na českých přeložených datech, druhé z nich využívá systému natrénovaném na anglickém jazyce tak, že český vstup přeloží do angličtiny, použije natrénovaný systém, a přeloží výsledky nazpět. Oba tyto postupy jsou vyhodnoceny jak na modelu BiDAF tak BERT.

Z mého pohledu nejzajímavější systém využívá model mBERT, který je předtrénován učením bez učitele na velkém množství čistého textu přibližně stovky nejrozsáhlejších jazykových mutací Wikipedie. V průběhu roku 2019 se začalo ukazovat, že ačkoliv je tento model trénovaný bez paralelních dat (tj. bez explicitního překladového slovníku či označených vět stejného významu ve více jazycích), dokáže reprezentovat podobné věty v různých jazycích podobným způsobem. V diplomové práci byl na tomto základě dotrénován model mBERT na anglických trénovacích datech SQuAD a poté vyhodnocen na českých evaluačních datech – tento systém tedy nevyžaduje žádná anotovaná data v češtině, pouze v angličtině. Přitom jsou výsledky tohoto přístupu srovnatelné se systémem trénovaným na přeložených českých datech. Výsledky diplomové práce potvrzuje i paralelně vzniklý článek Lewis et al. (2019) z konce října 2019, která provedla podobnou evaluaci na 6 jazycích (ani jeden z nich není čeština) s použitím modelu mBERT a XLM – větší počet jazyků a modelů je daný velikostí autorského kolektivu. Tato publikace nicméně potvrzuje, že téma zpracované v diplomové práci je vysoce aktuální.

Za největší slabinu diplomové práce považuji samotný text. Práce je psaná anglicky, a i když je jazyková úroveň nižší, považuji volbu angličtiny za výhodu, protože dosah práce je tak podstatně větší. Ovšem kapitola 2 obsahuje obecný úvod do zpracování přirozeného jazyka a question answeringu, který není nijak relevantní ke zbytku práce. Dále popis obou evaluačních metrik v sekci 3.2.3 je velmi nejasný (a pravděpodobně také chybný). V kapitole 4 představující architekturu BiDAF chybí podrobnější informace, ze současné podoby textu by nebylo možné ji reimplementovat. Kapitola 5 popisující stěžejní model BERT je také extrémně stručná a navíc sekce 5.2 chybně popisuje způsob predikce odpovědí – model BERT předpovídá pozici odpovědi v daném textu podobně jako BiDAF, a neprovádí rescoring dané odpovědi, jak popis sekce 5.2 naznačuje. Ke kapitolám 6, 7 a 8 nemám větší výhrady, jen některé hodnoty v obrázku 8.1 neodpovídají hodnotám v tabulce 8.1, konkrétně modely M-BERT{c,u}t{CZ100,EN}dCZ. V neposlední řadě by bylo vhodné vylepšit formátování odkazů na literaturu.

Před tyto nedostatky považuji práci za dostatečně kvalitní a doporučuji k obhajobě.

Bibliografie

Patrick Lewis, Barlas Oğuz, Ruty Rinott, Sebastian Riedel, Holger Schwenk: MLQA: Evaluating Cross-lingual Extractive Question Answering https://arxiv.org/abs/1910.07475

Práci doporučuji **k obhajobě.**

Práci nenavrhuji **na zvláštní ocenění.**

Datum 27. ledna 2020 **Podpis**