

Posudek oponenta bakalářské práce

Autor práce: Jiří Trefil

Název práce: **Vícejazyčné rozpoznávání dialogových aktů**

Tématem práce Jiřího Trefila je **rozpoznávání dialogových aktů ve více jazycích**. Autor pro tuto úlohu použil modely hlubokých neuronových sítí a otestoval je na datových sadách obsahující 3 jazyky (angličtinu, němčinu, španělštinu). Práce obsahuje 8 kapitol a je dobře strukturovaná. V úvodu autor práce dobře vysvětluje motivaci včetně důležitosti vícejazyčného zpracování.

Teoretická část

V teoretické části autor vysvětluje problematiku dialogů a dialogových aktů, modely neuronových sítí a použité datové sady. Vyzdvihl bych kapitolu o datových sadách. Je výborně popsána a nelze ji téměř nic vytknout. Popisy modelů neuronových sítí obsahují různé architektury, včetně těch nejmodernějších (tj. síť typu *Transformer*). Překvapil mě nicméně model *ESEC* (resp. jeho zkratka), se kterou jsem se nesetkal.

Experimenty

Experimenty jsou poměrně dobře navrženy, autor nejprve zkouší modely na jednom jazyku, poté až vícejazyčný scénář. Zarazil mě nicméně **nízký počet trénovacích epoch**, které autor uvádí (2, 4 resp. 9 u modelu *ESEC*). Autor tato čísla zdůvodňuje regularizací modelu pomocí mechanismu „Early Stopping“ na validační sadě, kterou tvoří 10 % trénovacích dat. Doporučil bych udělat několik scénářů s vypnutým mechanismem „Early Stopping“, modely nechat trénovat např. 50 až 100 epoch a výsledky porovnat. V případě použití náhodných slovních vektorů si úplně nedovedu představit, že by konvoluční či rekurentní síť (*LSTM*) dosahovaly deklarovaných úspěšností po 2 epochách trénování.

Autor dále uvádí, že jsou datové sady nevyvážené, což je pro dialogy a promluvy pochopitelné. Bylo by proto vhodné uvést i výsledky modelu „*majority classifier*“, který vždy predikuje nejčastější třídu. Čtenář by pak získal lepší představu o tom, jak jednotlivé modely fungují a kolik procent překonají tento „baseline model“.

Dodané zdrojové kódy v jazyku Python jsou funkční, podařilo se mi bez problémů skripty spustit. Autor využívá knihovny Keras a Tensorflow pro implementaci modelů neuronových sítí. Výsledky v tabulkách jsem nicméně nebyl schopen ověřit, protože z licenčních důvodů nemám k dispozici úplné datové sady, ale pouze jejich fragmenty na vyzkoušení.

Formální úroveň

Práce je psaná českým jazykem a obsahuje minimum překlepů. Vytknul bych kvalitu některých rastrových obrázků a snahu za každou cenu uvádět český název pro některé anglické termíny (např. „sdrůžovací vrstva“ nebo „vnoření slov“). V seznamu literatury jsou autoři některých referencí s pomlčkou, u publikace č. 12 je dokonce název velkými písmeny. Seznam literatury a jednotlivé reference pak působí nejednotně.

Myslím, že na bakalářskou práci autor odvedl velmi dobrý výkon a celý problém uchopil dobře. Ve výsledných tabulkách nicméně postrádám **Macro F-míru**, která se zpravidla používá obzvlášť, když je datová sada nevyvážená. Postrádám rovněž hlubší diskusi k výsledkům, např. pro model CNN_1 jsou

vidět výrazně nižší hodnoty metrik než u ostatních konfigurací (viz *Tabulka 7.3 a 7.4*, druhý řádek). Očekával bych k tomuto nějaké vysvětlení v textu, jestli je to způsobeno nedostatečným natrénováním modelu nebo jednoduše tato konfigurace nefunguje.

Zadání bylo z mého pohledu **splněno** v plném rozsahu. Autor práce provedl velkou sérii experimentů, seznámil se s modely neuronových sítí, které se používají v oblasti zpracování přirozeného jazyka a aplikoval je na vícejazyčnou úlohu.

Vzhledem k výše uvedeným připomínkám navrhuji hodnocení známkou **velmi dobře** a práci doporučuji k obhajobě. K práci mám následující dotazy:

- 1) V sekci 3.3.2 *Konvoluční neuronová síť* vysvětlujete, že konvoluční síť je široce využívána zejména v oblasti počítačového vidění. Obrázek 3.3 ukazuje úlohu rozpoznávání číslic z obrázků. Proč jste použil i tuto architekturu, když řešíte textovou úlohu z oblasti zpracování přirozeného jazyku?
- 2) V tabulkách ve výsledcích experimentů máte kromě úspěšnosti (Accuracy) také F-míru, která je zřejmě průměrovaná přes jednotlivé vzorky (Micro F1). O kolik by se zhoršily výsledky, kdyby se použilo průměrování přes jednotlivé třídy (Macro F1)?
- 3) Se kterými třídami (dialogovými akty) mají modely problémy a nedaří se je spolehlivě předpovědět?

V Plzni 27.5.2022

Ing. Jiří Martínek, Ph.D.