

Posudek oponenta diplomové práce

Student: Pastorek Peter, Bc.
Téma: Strojové učení pro odpovídání na otázky v češtině (id 22913)
Oponent: Fajčík Martin, Ing., UPGM FIT VUT

- 1. Náročnost zadání** **obtížnější zadání**
Študent v práci musel zostaviť systém, ktorý bude schopný získavať dokumenty zo širokého portfólia otázok, a zároveň systém, ktorý bude schopný extrahovať odpovede z dokumentov, alebo odpovedať na otázky áno/nie na základe získaných dokumentov. Zadanie je obtiažnejšie.
- 2. Splnení požadavků zadání** **zadání téměř splněno**
Bod č.1 bol splnený iba čiastočne. Práca obsahuje popis niektorých základných častí neurónových sietí (LSTM, dynamic memory network), avšak neobsahuje popis celých systémov, ktoré by sa zameriavali na odpovedanie na otázky z otvorenou doménou (teda s vyhľadávaním nad veľkou datovou sadou, napr. wikipedia).
- 3. Rozsah technické zprávy** **je v obvyklém rozmezí**
- 4. Prezentací úroveň předložené práce** **49 b. (F)**
 - Práca je písaná veľmi nejasne, obsahuje subjektívne konštrukcie, vágne termíny a na mnohých miestach obsahuje nepravdy, resp. polopravdy.
 - Práca neobsahuje ani jeden obrázok, schému, a dokonca ani označenú formulu (z výnimkou špecifických výpočtov na strane 33).
 - Kapitola 5 nie je štruktúrovaná a neobsahuje žiadne podkapitoly.
 - Práca vôbec nepopisuje implementačné detaily a zo samotného textu je veľmi obtiažne výsledky replikovať.
 - Logická návaznosť je nelineárna; práca v návrhu obsahuje tvrdenia a rozhodnutia, ktoré sú však podložené až ďalej v práci, a niekedy ani podložené nie sú.
 - Popis modelov je vytvorený slovne (viď strany 22 a 23) a pre čitateľa, ktorý doménu nepozná a s autorom sa nerozprával, je iba veľmi ťažko pochopiteľný.
 - Abstrakt práce pojednáva o implementačných detailoch, avšak vôbec nereflektuje výsledky a prínosy práce, ani nešpecifikuje konkrétnu myšlienku využitia pri návrhu práce (napr. systém je rozdelený na 5 podčastí, kde každá rieši samostatný problém).
 - Tabuľky z výsledkami obsahujú veľké množstvo dát a nasledujú až za diskusiou na samostatných stranách, anotácia stĺpcov v mnohých prípadoch nie je explicitne vysvetlená.
 - Práca obsahuje silné a nepodložené tvrdenia, napr. str. 54 "Odpovedanie na otázky je stále skúmaná oblasť, ktorá nemá optimálne riešenie."
- 5. Formální úprava technické zprávy** **50 b. (E)**
 - Jazyková stránka práce je silne podpriemerná, práca obsahuje mnohé nespisovné slová a slovné spojenia (napr. str. 12 "v celku").
 - Lexikálne a gramatické chyby (napr. "stratila bi sa", str. 53 záver, "veľkí rozdiel", str. 15 "jedno slovných" str. 15, "vedomostné databáze" - správne databázy, vzor žena, str. 15).
 - Typografická stránka práce obsahuje závažné pochybenia.
 - Citovanie je vždy realizované iba za paragrafom, aj keď sa citácia vzťahuje iba ku krátkej časti paragrafu (napr. str. 5). Citácie za paragrafmi sú často identické a prekopírované paragraf po paragrafe (napr. kapitola 2.3, str. 6).
 - Pri použití úvodzoviek nenasleduje za uzatvárajúcimi úvodzovkami medzera **v žiadnom pozorovanom prípade** v práci (napr. "áno"*alebo* "nie"*a systém...*, str. 13).
- 6. Práce s literaturou** **60 b. (D)**
 - V práci nie sú citácie na niektoré základné stavebné bloky neurónových sietí, ktoré študent popisuje a využíva (napr. neurónové siete LSTM, či GRU, Word2Vec).
 - Citácie sú vždy udávané až za paragrafom, niekedy až 7 citácií za sebou, je preto veľmi obtiažne pre čitateľa dohľadať, ktorá citácia sa viaže ku akej časti paragrafu.
- 7. Realizační výstup** **63 b. (D)**
 - Celý zdrojový kód je napísaný ako veľký python notebook (teda 1 súbor) dostupný v službe Google Colab (spustiteľný online). Zadanie nie je spustiteľné, pretože je závislé od Gdrive účtu študenta. S pomocou

študenta sa ho ale podarilo spustiť. Neexistuje demonštračný program, v ktorom by užívateľ mohol interagovať s finálne natrénovaným systémom, k dispozícii sú mu iba výsledky na validačnej sade.

- Systém je schopný odpovedať iba na otázky s odpoveďou nachádzajúcou sa v úzkej vybranej podmnožine Wikipédie.
- Riešenie využíva poznatky adekvátne k stavu poznania v rokoch 2013-2016 a je založené na mojom jednoduchom riešení pripravenom pre demonštračné účely pre letnú školu BISSIT19. Techniky aktuálneho stavu poznania v práci nie sú ani spomenuté.
- Výhodou takéhoto riešenia je však jeho nízka výpočtová náročnosť a možná nasaditeľnosť, napr. na webovej platforme.

8. Využitelnost výsledků

Práce může posloužit jako baseline systém pro systémy tvořená na české datové sadě SQAD.

9. Otázky k obhajobě

1. V práci využíváte dva typy rekurentních sítí - LSTM a GRU, bez ich hlbšieho popisu. Popíšte rozdiely medzi týmito sieťami. Ktorá sieť má koľko parametrov?

10. Souhrnné hodnocení

54 b. dostatečně (E)

Textová část práce je velmi těžko pochopitelná, obsahuje vágně aj nepravdivé termíny. Neobsahuje jedinou schému, obrázok ani formulu, aj keď si to text vyžaduje. Vytvorený plagát je na tom rovnako. Niektoré citácie chýbajú a formálna úprava práce je vysoko podpriemerná. Výsledky práce sa javia funkčne, ale kvôli chýbajúcemu demu a kódu s nedostupnými závislosťami s nimi nie je možné interagovať.

Výsledkom práce je rýchle a jednoduché riešenie pre odpovedanie na otázky v prirodzenom jazyku na vybranej podmnožine Wikipédie. Na základe týchto faktov prácu hodnotím ako podpriemernú.

Prohlášení: Uděluji VUT v Brně souhlas ke zveřejnění tohoto posudku v listinné i elektronické formě.

V Brně dne: 28. června 2020

Fajčík Martin, Ing.
oponent