Posudek oponenta diplomové práce

Autor práce: Bc. Petr Kopal

Název práce: Propojení témat zpravodajských článků mezi jazyky

Obsah práce
Práce se zabývá metodami určenými pro propojení tematicky podobných shluků zpravodajských článků mezi různými jazyky. Práce je spíše výzkumného charakteru a autor musel prostudovat oblast zpracování přirozeného jazyka, což je poměrně náročné.

První polovina práce obsahuje některé základní pojmy z oblasti zpracování přirozeného jazyka a strojového učení. Dále autor popisuje možnosti pro sémantickou reprezentaci slov a dokumentů, metody pro podobnost dokumentů napříč jazyky. Následuje opravdu velmi stručný návrh systému zahrnující pouze seznam vybraných metod, popis použitých dat pro trénování a jejich předzpracování. Za velký nedostatek považuji chybějící diskuzi (porovnání, odůvodnění, zhodnocení vhodnosti) vybraných metod. Implementace systému obsahuje pouze popis jednotlivých Java tříd, který je však vhodně doplněn UML diagramy.

Autor definuje model skip-gram s negativním vzorkováním (z angl. skip-gram with negative sampling - SGNS) algoritmu word2vec jako vzorec 4.5 (str. 14). Tento vzorec ale nelze použít pro zápis uvedeného modelu, protože označuje cenovou funkci modelu (nikoliv formální zápis modelu). Cílem trénování modelu je pak minimalizace této funkce. Vzorec 4.5 (resp. cenová funkce) není vzorcem pro cenovou funkci SGNS, jak autor popisuje, ale vzorcem pouze pro cenovou funkci základní verze skip-gram modelu (bez negativního vzorkování). Podobně vzorec 4.7 nedefinuje model CBOW (z angl. Continuous Bag of Words), ale jeho cenovou funkci. Obrázek 4.3 popisuje model CBOW (nikoliv SGNS) a obrázek 4.4 popisuje model SGNS.

Na základě výše popsaných chyb, nepřesností a celkového dojmu z teoretické části se domnívám, že autor plně neporozuměl všem metodám, které v práci popisuje.

Struktura práce a řazení kapitol je logické. Text diplomové práce je srozumitelný, ale autor často používá velmi vágní definice a výrazy (např. str. 22 “Obecně tato metoda nefunguje vůbec špatně”, str. 22 “Potřebovali bychom nějak…”, str. 60 “Pokud podobnost shluků byla dostatečně velká” apod.), které jsou pro text diplomové práce nevhodné. Autor také v textu používá zkratky, které předtím nikde nedefinuje a často nevhodně mísí české, anglické nebo počeštěné výrazy a slova (např. str. 32 “shluků - clusterů”, str. 50 “hashovací mapou”, str. 54 “clusterer”, str. 31 “ground truth” apod.). Drobným nedostatkem je nekonzistence použitých výrazů napříč celým textem.

Také bych vytknul použití rastrových obrázků místo obrázků vektorových, které mohly být použity ve většině případů. Obrázky 8.1, 8.2, 8.6 mohly být vysázeny jako text, např. pomocí balíčku Listings v LaTeXu.

Další poznámky k textu
- Kapitola 5 obsahuje jen dva odstavce textu a jeden obrázek. Vhodnější by bylo zařadit ji jako podkapitolu.
- Na str. 60 není vůbec zřejmé, jak byl použit word2vec pro nalezení nejpodobnějších článků.
- Statistiky uvedené na str. 40 až 42 by bylo vhodnější umístit do tabulky.

Kvalita řešení a dosažených výsledků
Realizaci a postup při experimentech považuji za vhodně zvolené a v souladu s cílem práce. Autor experimentuje s vybranými metodami a různými kombinacemi příznaků a postupně vylepšuje úspěšnost modelu.

Výrazný nedostatek vidím v popisu experimentů, který je velmi strohý a neobsahuje důležité detaily implementace (předzpracování textu, konkrétní konfigurace jednotlivých metod, na základě čeho autor vybral počty dimenzí a velikosti použitých matic apod.). Popis postrádá smysluplnou diskuzi, zhodnocení a porovnání výsledků jednotlivých metod. Z uvedeného popisu bych nebyl schopen

experimenty zopakovat. Dále mi chybí porovnání výsledků s aktuálními state-of-the-art metodami nebo podobným systémem.

Experimenty jsou implementovány v jazyce Java. Zdrojové soubory jsem bez problémů přeložil nástrojem maven. Zdrojový kód je dobře strukturovaný a čitelný (ocenil bych více komentářů především v důležitých částech implementace). Implementované metody jsem bez potíží spustil, ale nebyl jsem schopen zreplikovat experimenty popsané v textu. Potřebné konfigurační soubory pro tyto experimenty nebyly přiloženy a podle textu diplomové práce je nebylo možné vytvořit.

Formální úroveň
Formální úroveň práce je dostatečná. Dokument diplomové práce je vysázen v TeXu a dokument neobsahuje typografické chyby. V textu jsem objevil minimum překlepů nebo gramatických chyb.

Práce s literaturou
Citovaná literatura je vzhledem k tématu diplomové práce relevantní a citace v textu jsou v pořádku. Autor ovšem necituje původní články pro *Explicitní sémantickou analýzu* (str. 10), *Word2vec* (str. 13) a *GloVe* (str. 15) i přes to, že u prvních dvou jmenovaných metod uvádí autory i rok publikování článků. Dále chybí citace nebo alespoň odkaz pro tyto použité zdroje: *Eurovoc, Snowball stemmer* a Java knihovny *jaxb* a *dom4j*.

Splnění zadání
Zadání bylo splněno i přes výše uvedené nedostatky a poznámky.

Dotazy k práci
1. Vysvětlete, co myslíte tvrzením, že korpus nepoužíváte z důvodu vysoké paměťové náročnosti? (poznámka na str. 41)
2. V diskuzi tvrdíte, že implementačně nenáročné metody fungují lépe než uvedené "embeddingové" metody pro trénovací data o menší velikosti. Na základě čeho vyvozujete tento závěr? (str. 62 kap. 9.4 odst. 3)
3. Proč a v čem je strojový překlad "těžší" úloha než řešený problém? (str. 20 kap. 6.1 odst. 2)
4. Jaké konkrétní předzpracování textu dokumentů jste použil pro experimenty uvedené v práci?

Navrhuji hodnocení známkou **dobře** a práci doporučuji k obhajobě.

V Plzni dne 25. srpna 2019

Ing. Pavel Přibáň