FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG
TECHNISCHE FAKULTÄT

Universität Erlangen-Nürnberg, LS Informatik 5, Martensstr. 3, 91058 Erlangen, Germany

Západočeská univerzita v Plzni
Děkanát FAV
Ing. Jaroslav Toninger
Univerzitní 8
306 14 Plzeň
Czech Republic

Department Informatik

Lehrstuhl für Informatik 5
(Mustererkennung)

PD Dr.-Ing. Tino Haderlein
Martensstr. 3, 91058 Erlangen
Telefon +49 9131 85-27872
Fax +49 9131 85-27270
Tino.Haderlein@fau.de
www5.cs.fau.de

Erlangen, den 24.10.2019

Opponent's Review for the Ph. D. Thesis of Ing. Lukáš Svoboda, "Distributional Semantics Using Neural Networks"

a) Meaning of the Thesis for the Field

Automatic analysis of the similarity of word and phrase meanings is important for applications of automatic speech processing in order to improve the accuracy of translations or the assessment of linguistic information in the given data. Since linguistic approaches require an expert in the respective field, unsupervised methods for the creation of such methods and models are strongly desired. Especially for highly inflectional languages, such as Czech or Croatian, this is an essential issue. Mr. Svoboda's thesis contributes successfully to the solution of this problem by applying word embedding methods and creating new corpora for testing them, by evaluating sentiment analysis methods on sentence level, and by introducing a new approach of learning word embeddings enriched with global information. The latter improves the respective results for smaller corpora of languages with extensive inflection.

b) Method of Problem Solving, Used Methods, and Fulfillment of Targets

Several established word embedding methods (CBOW, Skip-gram, GloVe, FastText) have been used in the thesis, at first mostly for data in English. The focus of the work was then turning towards inflectional languages, which meant a substantial increase in problem complexity. For classification purposes, several established approaches, such as Support Vector Machines, Feed-Forward Neural Networks, Convolutional Neural Networks, and Recursive Neural Networks, have been applied and thoroughly evaluated with different parameter settings.

Remarkable contributions of this thesis are a Czech and a Croatian word analogy corpus and several Croatian word similarity corpora, which are the first of their kind. The presented corpora and data are available online for other research groups. The most meaningful novelty of the thesis is an approach for training word embeddings, using global information extracted from large text corpora (here: Wikipedia), that has been applied to data in Czech language.

The thesis goals were the analysis of the influence of rich morphology on the quality of meaning representation, the improvement of the meaning representation for inflectional languages with new approaches based on neural networks, and the improvement of natural language processing tasks (NLP) by using distributional semantic models. All thesis goals have been fulfilled very well.

The quality of the developed methods has been proven by one second and one first place among more than 100 approaches in the SemEval competition of the International Workshop on Semantic Evaluation, organized by the Association for Computational Linguistics (ACL).

c) Results of the Thesis

In the analysis of cross-lingual word analogies (Chapter 4.4), six different languages have been transformed into a shared semantic space using dictionaries of word translations. On the monolingual task, 51.1% of accuracy were reached. For the multilingual task, where all semantic spaces were transformed to the English one, the respective accuracy was 38.2%.

In the problem of semantic textual similarity (Chapter 5), the newly created set for Czech achieved a Pearson's correlation of 0.80 to the reference, which is not much below the value that had been determined for English data (0.88). However, the results showed substantial differences in individual experiments, depending on the topic addressed in the data.

In the aspect-based sentiment analysis (Chapter 6), word clusters and stemming improved the F-measures which reached about 80% for both English and Czech data.

For the word embeddings trained with global information (Chapter 7), word similarity and word analogy results showed accuracies of above 80% for some experiments with the English data. The results for the Czech data were similar for some subtasks. Nevertheless, the results of the entire thesis are remarkable in the view of the complexity of Czech morphology and the large amount of out-of-vocabulary words in some constellations.

d) Systematics, Clarity, Formal Elaboration, and Language Level

The main part of the thesis is composed of several publications. This structure makes it sometimes a bit difficult to see the relation between the single chapters and their embedding into the entire research task. Some methods and approaches are expected to be known by the reader; more details would have been desirable. Table captions do often not say whether correlations or accuracies are presented. The single parts of the thesis, on the other hand, show a clear structure and writing, the mathematical notation is also mostly clear. The use of English is adequate with a few minor mistakes. In the Bibliography section, there are a few details missing.

e) Publications of the Author

The publication list contains four journal articles; for two of them, Mr. Svoboda is the first author, the only other author is the consulting specialist of the thesis, Dr. Brychcín. Additionally, five conference papers have been published between 2016 and 2018. Mr. Svoboda is the first author of three of them; most papers have only one co-author. Two papers have been published by Springer, another one is from the well-known LREC conference, and the remaining two were published at the SemEval 2016 conference in San Diego, California.

f) Recommendation for the Acceptance of the Thesis

The thesis and the publication list show that Mr. Svoboda is able to perform research independently in a structured and thorough way. For this reason, I recommend the acceptance of his thesis for the granting of the academic title Ph. D.

FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG
TECHNISCHE FAKULTÄT

Questions for the Defense of the Thesis

- Chapter 4.2 (p. 36):
  "*The dataset contains only frequent-enough words from the Czech Wikipedia.*"
  What exactly does "frequent enough" mean?
- Chapter 4.2.1 (p. 38):
  The paragraph "Models settings" says:
  "*We also explore results with different vector dimension (set to 100, 300, and 500).*"
  The next paragraph ("Results") says:
  "*[...] we present results for different vector dimension ranging between 50 and 500 [...]*"
  How does this match?
- Chapter 5.2.4 (p. 56/57):
  "*Firstly, we translated Spanish sentences to English via Google translator.*"
  How was ensured that this translation was correct?
  What influence could a wrong translation have on the task?
  (analogous: Chapter 5.3.3 with Czech-to-English translation)
- Chapter 5.2.5 (p. 57):
  "*As an evaluation measure we use Pearson correlation between system output and human annotations.*"
  How do you obtain a numerical value to get a correlation from?
- Chapter 7.3 (p. 81):
  Eq. 7.1: Where do i and $\mathbf{a}_j$, that are mentioned right before, appear in the equation?
  Eq. 7.2: "*The CBOW model optimizes [the] following function [...].*"
  This is not a function, just a term.
  What does "optimize" mean? Is there a minimum/maximum computed?
- Chapter 7.6.1 (p. 88):
  About the training sample, it says:
  "*The set of negative samples N is always sampled from unigram word distribution raised to 0.75 [...].*"
  What does that mean?

Tino Haderlein

prof. Ing. Dušan Krokavec, CSc. Katedra kybernetiky a umelej inteligencie,
Fakulta elektrotechniky a informatiky TU v Košiciach, Letná 9, 042 00 Košice

OPONENTSKÝ POSUDOK

dizertačnej práce

Autor:	Ing. Lukáš Svoboda
Téma práce:	**DISTRIBUTIONAL SEMANTICS USING NEURAL NETWORKS** **(Distribučná sémantika s využitím neurónových sietí)**
Školiteľ:	prof. Ing. Václav Matoušek, CSc.
Konzultant:	Ing. Tomáš Brychcín, Ph.D.
	Katedra informatiky a výpočetní techniky Fakulta aplikovaných věd Západočeská univerzita v Plzni
Obor štúdia:	Informatika a výpočetní technika
Rozsah práce:	116 str., 8 kapitol, 11 obrázkov, 26 tabuliek

1. Aktuálnosť zvolenej témy a jej význam

Pretože súčasný trend v syntaxi a sémantike sekvencie slov, alebo celých viet, možno charakterizovať výhradne zovšeobecnenými prístupmi odrážajúcimi ich číselnú reprezentáciu, štruktúru a podmienenosť, cieľovo orientované programové prostriedky, tak ako v tomto prípade použitie postupov využívajúcich na akceleráciu výpočtov a klasifikácie princípy umelých neurónových sietí, možno z realizačného pohľadu vždy považovať za úlohu aplikačne aktuálnu. Pri známych teoretických východiskách, a ich priemetu do vytvorených korpusov a testovacích experimentov, bola takto orientovaná téma práce reálne podmienená a metodicky zvládnutá. Práca spadá do oboru, prináša do neho nové interpretácie teoretických východísk a klasifikačných postupov.

2. Cieľ dizertácie a jeho splnenie

Doktorand má ciele dizertačnej práce pevne fixované na problémy sémantickej reprezentácie dát. V kontexte základných princípov strojového učenia sa postupne snaží o zdôvodnenie platformy zlepšenia porozumenia sémantiky a syntaxe pomocou metód predspracovania a klasifikácie nasadením neurónových sietí, resp. štandardných postupov regresnej analýzy (LR, GP, SVM). Som toho názoru, že hlavne kvôli už na úvod potenciálne špecifikovanej platforme (neurónové siete) a zvolenej metodike riešenia (distribuovaná sémantika), bol cieľ práce chápaný predovšetkým ako vytvorenie reprezentácie v zmysle tých postupov sémantickej analýzy, ktoré takúto štruktúru vyžadujú, s potenciálnym rozšírením do reprezentatívnych modifikácii s reflexiou syntaxe. Takto formulovaný cieľ (ako realizačnú ideu) možno chápať ako vecne primeraný a v zmysle výsledkov riešenia za splnený.

3. Metódy spracovania

Riešenie, ako to možno vyvodzovať z daného cieľa, bolo ohraničené na výber učiacich (trénovacich) súborov dát s distribuovanou reprezentáciou slov a ich okolia, prípravu štruktúr dát s kódovaním viacerých lingvistických vlastností a optimalizáciu sekvenčnej klasifikácie pre prípady, kde je možné použiť metódu hĺbkového učenia neurónovej siete, resp. na architektúry s potenciálnu kompozíciou vektorov slov až po úroveň vektorovej reprezentácie vety. Tým, že tok dát je spravidla limitovaný, použitý prístup bolo možné redukovať na heuristickú optimalizáciu v zmysle účelových funkcií odpovedajúcich architektúram CBOW a Skip-gram. Metodiky sú v súlade s teoretickými východiskami a pri vlastnostiach týchto architektúr je takýto postup implementovateľný. Z praktického hľadiska reprezentácie sémantiky slov českého jazyka ide zaujímavý návrh riešenia, aj keď zatiaľ len v ohraničenej konfigurácii.

4. Výsledky práce a ich charakteristika

Výsledky práce doktoranda sú prezentované v priereze kapitol Kap. 4-7 dizertačnej práce a ich stručné záverečné zhrnutie je uvedené v Kap. 8. V prvých troch kapitolách sú uvádzané súvislosti, z ktorých vychádzajú nosné myšlienky reprezentácie syntaktických a sémantických štruktúr a základné charakteristiky platforiem pre reprezentáciu, učenie a testovanie. Závery o modifikácii a optimalizácii riešenia, uvedené v Kap. 6-7 v súvislosti distribuovanými sémantickými modelmi a globálnym kontextom, sú z pohľadu cieľov práce očakávateľné (s. 92). Na všetky koncepčné východiská nadväzuje opis experimentov a spôsob testovania vytvorených štruktúr. Použité modifikácie metód pre český jazyk sa javia ako akceptovateľné.

Hoci sú vlastné výsledky charakterizované predovšetkým strohou formuláciou riešenia a tabuľkovou prezentáciou najdôležitejších testovacích výsledkov, implementačná platforma v akceptovateľnej miere zviditeľňuje použiteľnosť navrhovaných metodík a postupov. Na úrovni záverečného spracovania sa prejavila značná konzervatívnosť autora s interpretáciou získaných výsledkov, čo bránilo zvýrazniť hlavne implementačnú pôvodnosť výsledkov práce.

5. Prínos pre ďalší rozvoj vedy a techniky

Autor uvádza v dizertačnej práci niekoľko novovytvorených korpusov, použitých predovšetkým na vkladanie slov, na určovanie sémantickej textovej podobnosti a analýzu sentimentu a aplikovaných v súvislosti s optimalizáciou funkčnosti v štruktúrach architektúr CBOW a Skip-gram, resp. sekvenčného behu nad množinou testovacích súborov. Ich návrhom a využitím sa dopĺňajú poznatky z oblasti počítačovej reprezentácie slov českého a chorvátskeho jazyka, rozširuje sa štandard syntézy systémov analýzy sentimentu založenej na aspektoch a vytvára sa formalizovaný základ pre ďalšie potenciálne zovšeobecnenie prezentovaných metodík. Aj keď ide o relatívne ohraničenú teoretickú oblasť, zo získaných výsledkov vyplýva výrazný autorov praktický prínos v oblasti formalizácie riešenia zadanej úlohy. Možno ešte vyzdvihnúť nový prístup k určeniu číselnej reprezentácie slov, vylepšený o globálne textové informácie, pre jazyky charakteristické bohatou morfológiou slov.

6. Publikačná aktivita autora

Doktorand uvádza spoluautorstvo v piatich prácach publikovaných v zborníkoch z vedeckých konferencií a v štyroch prácach publikovaných v časopisoch. Všetky práce sú v databáze SCOPUS, v databáze Web of Science sú všetky časopisecké príspevky, ktorých je spoluautorom. V zmysle uvedeného možno jeho publikačnú aktivitu hodnotiť ako veľmi dobrú.

7. Stanoviská a pripomienky k práci

Práca je napísaná v anglickom jazyku, odborne prijateľným štýlom, s dobrou vysvetľujúcou charakteristikou riešenia, prakticky bez chýb (1 preklep na strane 78). Zásadné pripomienky k práci nemám, otázky a námety do diskusie možno zhrnúť nasledovne:

- lokálny kontext sa spravidla aplikuje, ak náklady na implementáciu globálneho kontextu sú vyššie ako prahová hodnota nákladov na lokálny kontext; existuje pri týchto typoch úloh prahová hodnota nákladov globálneho kontextu a vzhľadom k čomu môže byť definovaná?
- aká je opakovateľnosť výsledkov vašich riešení pri tej istej topológii neurónovej siete, ale s inou inicializáciou ako je rovnomerná? Čo predovšetkým môže ovplyvniť „streaming"?
- uveďte hlavné výhody a vašu prioritnú motiváciu využívať metódy nekontrolovaného učenia!

8. Záver

V dizertačnej práci „Distribučná sémantika s využitím neurónových sietí" Ing. Lukáš Svoboda prezentuje prínosné spôsoby číselnej reprezentácie sémantickej informácie a syntaxe, predovšetkým českého jazyka. Práca prináša nové poznatky v odbore, formálne spĺňa podmienky pre doktorandské dizertačné práce a keďže nemám pripomienky ktoré by znižovali metodický prínos práce,

odporúčam

aby práca bola predložená k obhajobe v obore Informatika a výpočetní technika.

Nadväzne, po úspešnej obhajobe,

súhlasím

s udelením akademického titulu Ph.D.

Košice, 21.11.2019 prof. Ing. Dušan Krokavec, CSc.