REVIEW

by Ivan Koychev, PhD, Professor at FMI of Sofia University "St. Kl. Ohridski "
for the PhD thesis entitled:
"Modeling Lexical Knowledge for Natural Language Processing"
Author: Alexander Nikolaev Popov
Supervisor: Assoc. Prof. Kiril Simov
for acquiring the educational and scientific degree "doctor" in:
Area of higher education 4. Natural sciences, mathematics and informatics;
Professional field 4.6. Informatics and Computer Science;

This review was prepared on the basis of order 127/12.07.2018 of the Director of IICT-BAS, which included me in the scientific jury for the above-mentioned thesis, and the decision of the first jury meeting, which assigned this task to me.

Natural language modeling is an important field for the disciplines studying natural languages, as well as for a number of modern computer applications. An important task for Natural Language Processing (NLP) is to find an appropriate formal presentation of lexical knowledge. Very often, the dictionary (lexicons) is seen as a means for presenting multiple types of information, such as: syntactic structure, semantic relations, co-occurrences, etc. The present thesis offers modern approaches for the automated creation and improvement of such dictionaries, which are widely used in many modern software systems. Consequently, we can conclude that the subject of the presented dissertation is in a very modern research field.

The dissertation has a total volume of 147 pages in which about 120 pages constitute the dissertation itself. It is written in English, which I think is very good. It is well structured, consisting of an introduction, seven chapters and a conclusion.

Chapter 1 (introduction) gives a brief motivation of the topic, clearly defines the aims of the dissertation and the subsequent specific tasks. The structure of the dissertation is briefly described.

Chapter 2 gives formal definitions of the assigned tasks by introducing the basic concepts and mathematical inscriptions.

Chapter 3 gives an overview of the main results and current research in the field.

Chapter 4 presents the developed approach based on recurrent neural networks for the purpose of determining the parts of the speech of words. This chapter also presents the conducted experiments for verifying and comparing the proposed approach.

Chapter 5 presents an approach to enriching the semantic graph of WordNet through additional links that better take into account the context. It is suggested that the Hypernym-Hyponym link in the WordNet can be explicitly represented, in the case of the Bulgarian language, as well as that syntactic relations from other sources can be added. At the end of this chapter are also presented the results of an experimental examination of the developed methods. Experiments were carried out using data from the part-of-speech and word-sense-annotated corpus BulTreeBank for Bulgarian and data from the SemCor corpus for English.

Chapter 7 presents two recurrent neural networks approaches which learn models of lemmas, synonyms and contexts. It presents also experiments that are conducted to evaluate the proposed approaches.

Chapter 8 presents a multitask machine learning approach that uses a hybrid architecture combining the previously presented approaches. This chapter also presents results of experiments aiming to evaluate the proposed approaches.

Each of the chapters 2-8 ends with a sub-chapter discussing the proposed approaches and experimental results and proposes ideas for further development.

Chapter 9 (Conclusion) summarizes the thesis results, presents the concrete contributions of the PhD student and proposes ideas for future research.

The list of literature used ("Bibliography") has a volume of 13 pages and contains over 140 titles: scientific articles, books, online publications and documentation. I have not noticed any that are not referenced in the dissertation text.

There are two appendixes with a total length of 6 pages, the first is a list of tables, and the second is a list of figures.

The PhD student demonstrates in-depth knowledge in the scientific fields of the dissertation. The reviews of related works are in-depth and cover a wide range of methods, technologies and software tools. These overviews adequately represent the current state of the art in the relevant areas. The approaches and models developed are well explained and illustrative examples are given. The applied algorithms have been tested in a number of experiments that are well planned and conducted, and the results are well presented and discussed. The easy understanding of the presented methods and experimental results is facilitated by a proper amount of figures and tables in the dissertation (8 figures, 23 tables).

The chosen research methodology is suitable for the purposes of the dissertation. It includes: an analytical overview of the field, a motivated choice of appropriate methods and their creative development, and their application for solving particular tasks. The developed methods are implemented in software tools and verified in experiments. It is clear that the PhD student has gained experience throughout the research: from the creation of abstract models to their application for solving specific tasks in real projects, such as the QTLeap project.

The dissertation is very well formatted and aesthetically shaped. It is also good from the point of view of spelling, grammar and stylistics. The important units in the text are indicated and marked with appropriate formatting. The figures and tables are duly numbered, have explanatory captions and are commented on in the text. Enclosed are an index of figures, an index of tables and a list of abbreviations.

The scientific and applied-scientific contributions of the dissertation can be summarized as follows:

1. Enriching WordNet with multiple semantic and lexical links. It has been found that these new connections have greatly improved the results in some lexicon resolving tasks using knowledge bases.
2. Several neural network architectures of the sequence-to-sequence type are proposed and tested within three lexical tasks: identifying parts of speech, solving lexical ambiguity and representing the context.
3. Various vector-space models (using words, lemmas, synonyms, and suffixes) were studied for presenting input for supervised machine learning algorithms. Some of these models are based on the WordNet semantic network structure and on its enriched variants. It has been found that these models show very good results when solving the tasks of calculating relatedness and similarity between words. A new distributed representation of grammatical roles has been created and used in a new way to enrich the links in the WordNet.
4. Two approaches based on multitask machine learning for lexical modeling were compared. The results show that models using lexical dependencies are performing better. Namely, these models show improvements in "context-sensitive lexical disambiguation" and "part-of-speech-sensitive lexical disambiguation ".

There is a list of 14 articles in which parts of the dissertation results are published. In two of them the PhD student is the single author and in 3 of them he is the lead author. All publications are in international peer reviewed issues. Eight of these are indexed in SCOPOS, including both ones authored by the PhD student alone, and five are indexed in the Web of Science, also including both ones authored by the PhD student alone. The PhD student has delivered 4 presentations on the topic of the dissertation at internationally recognized conferences, 2 at student workshops and one at an internal seminar.

Evidence for the significance of the results obtained are the 8 citations of articles in which contributions from the dissertation have been published, 3 of the citations being of the papers where the doctoral candidate is a lead author. The publications presented, in terms of number and quality, not only fulfill the formal requirements of the law and the rules for dissertations for the obtaining of the educational and scientific degree of doctor, but also significantly exceed them.

Some of the results presented in the thesis have been used in several research projects, which is a testament to their applicability. For example, approaches for processing language sequences for Bulgarian and English are used in the European projects QTLeap and EUCases.

The extended abstract in Bulgarian is well-done and accurately reflects the content of the dissertation. The terms are thoroughly translated into Bulgarian, but in my opinion there is room for improvement. The abbreviations are correctly introduced, with an associated glossary, but I think there are too much of them and that makes the text a bit difficult to the reader. In general, the text of the extended abstract can be improved with some editing, by clarifying the terminology and description.

I have no significant remarks about the thesis.

I do not know the PhD student personally, but the professional biography and other documents attest that he is a highly qualified specialist with a valuable experience and research spirit.

Conclusion: My assessment of Alexander Popov's dissertation thesis, the extended abstract, scientific publications and scientific contributions **is positive**.

The presented dissertation works in full compliance with all the requirements, conditions and criteria of the Law on the Development of the Academic Staff in the Republic of Bulgaria (SRARPD), the Regulations for the implementation of the SRARPD, the Regulations for the Conditions and Procedure for Acquisition of Academic Degrees at the Bulgarian Academy of Sciences, and the Regulations for the Specific Conditions for Acquisition of Academic Degrees at IICT-BAS. On this basis I propose to the honorable jury to award the academic and scientific degree "doctor" to Alexander Popov in the field of higher education 4. Natural sciences, mathematics and informatics; professional field 4.6. Informatics and Computer Science.

09.10.2018

Sofia