REVIEW

of the thesis of Alexander Nikolaev Popov
"*Modeling lexical knowledge for natural language processing*",
presented for awarding the educational and scientific degree "PhD"

by Prof. Galia Mladenova Angelova,
Department of Linguistic Modelling and Knowledge Processing, Institute of Information and Communication Technologies (IICT), Bulgarian Academy of Sciences (BAS)

Pursuant to Order 127 / 12.07.2018 of the Director of IICT-BAS, I am nominated as a member of the Scientific Committee for the defense of the submitted thesis in professional area 4.6 "Informatics and Computer Science". The dissertation is related to some of the hottest problems in Computer Science today: finding new models and methods for efficient processing of large volumes of texts based on vector representation of words (lexical knowledge); creating open-access linguistic resources of word representations as vectors and testing scenarios for comparing and assessing the quality of these vectors; improving the approaches to automatic word sense disambiguation. These issues are part of the dominating "deep learning" research trend. The optimistic expectations for rapid artificial intelligence developments are connected to the successful construction of word representations as vectors as well.

Formally, according the Rules for the Implementation of the Law on the Development of the Academic Staff in the Republic of Bulgaria of July 6, 2018, a PhD candidate in computer science is expected to have minimally at least 30 points in the “D”-group of indicators. The research publications presented with in the Alexander Popov’s thesis have 215 points in total (156 points from papers indexed by Scopus or Web of Science and 59 points from non-indexed papers). Thus the formal requirements for obtaining the educational and scientific degree "doctor" are satisfied and exceeded. The results of the dissertation are published in 14 papers (4 of them authored by the candidate as a single author), among them:

- 1 paper is published in an international scientific journal with SJR rank,
- 4 paper are included in the Proceedings of prestigious international conferences, published by Springer in various series - Computational Intelligence and Lecture Notes on Artificial Intelligence;
- 9 papers appeared in the Proceedings of established international conferences on natural language processing, including RANLP, Global WordNet Conference, TLT and others.

There are 8 citations of the author's papers. In addition to presenting the papers at 6 international conferences, Mr Alexander Popov delivered several talks at the DemoSem seminar, where the results of the dissertation were discussed.

The thesis is written in English and contains 140 pages of main text including 13 pages with references. The bibliography includes 141 titles (127, if we exclude the publications presenting the dissertation results). Seven pages of Annexes contain lists of tables, figures, and abbreviations used in the thesis. The text is organised in 9 chapters: Introduction, 7 Chapters and Conclusion. Chapters 3-8, which present the author's motivation and the research done during his PhD study, end up with summaries of presented results and/or a list of open questions to be investigated in the dissertation (Chapter 3) and as future research (Chapters 4-8).

Chapter 1 (Introduction) discusses the importance of the topic and positions it within the Natural Language Processing (NLP) field. The objectives and tasks of the thesis are presented as well as the place of Bulgarian language processing in the research study. The structure of the dissertation is explained too.

Chapter 2 (Problem Definitions) presents the framework in which the doctoral studies are conducted, with the introduction of basic concepts and formal notations. This Chapter introduces the task of automatic lexical ambiguity resolution (Parts of the speech tagging, POS tagging) as well as the task of word sense disambiguation (WSD) which is central for the research conducted in the thesis. The notions of word similarity and relatedness are considered in detail. Numerous data sets used as linguistic resources to solve these tasks are discussed.

Chapter 3 (Background and Related Work) describes current solutions and language resources for word representations in NLP systems. Since the thesis aims to study and optimise especially word representations, the main subject is the lexicon in principle and WordNet in particular as well as some other linguistic resources. There is an in-depth review of existing WSD approaches and their applications. In addition there is an overview of NLP systems using neural networks (NN) for text processing tasks, focusing on the WSD issues. The review also mentions the results of the author, briefly discussed as part of the state of the art in the field. These hints help for understanding the originality and place of the author's results in the WSD approaches.

Chapter 4 (Recurrent Neural Networks for Part-of-Speech Tagging) presents the use of recurrent neural networks (RNNs) for resolving the POS-ambiguity. The training of RNNs for building vector representations of the words was accomplished on a Bulgarian corpus with about 220 million words, and the training of the vectors presenting the suffixes - on a subset with about 10 million words. As a result, vector representations were obtained: the words are 200-dimensional vectors, and their suffixes – 50-dimensional vectors. The RNN hidden layers utilize the LSTM mechanism in order to dynamically handle long sequences. A multilayer architecture with bidirectional RNNs is proposed, which allows concatenation of word and suffix vectors. The BulTreeBank Corpus was used for the testing. In training with 10,000 iterations, the best accuracy of 78.16% for POS tagging was obtained. Adding the suffix vectors with a hidden layer of 125 neurons, an accuracy of 94.47% was achieved. This accuracy is comparable to the best solutions for English texts and shows that the basic language technologies for Bulgarian are already of world-class quality. Another interesting issue here is the empirical evidence that morphological information about words can be represented as embeddings as successfully as syntactic-semantic information. This is important knowledge about languages with rich morphology similar to the Bulgarian one. I believe that the presented model for solving the POS-tagging task will be useful for scientists developing NLP systems for other inflexional languages because the model can be easily transferred to other languages given the availability of the necessary linguistic resources.

Chapter 5 (Graph-based modeling of lexical semantics) presents one approach to WSD by using a lexicon built as a graph. The idea is that integrating knowledge (in the form of specific semantic networks) into the lexicon will contribute to a more successful automatic identification of the word meanings. The idea is to infer automatically relations from texts and it is implemented for both Bulgarian and English. The comparative analysis of the behavior of the proposed algorithms for these two languages makes a very good impression because it demonstrates the deep and comprehensive study of graph-based linguistic resources. The detailed experiments show which relationships are the most appropriate improvement of the lexicon to support the WSD approaches in a knowledge-based scenario. The accuracy of over 68% in automatic WSD means that adding knowledge - information about relations derived from graphical representations - significantly improves the potential for solving this task. This chapter demonstrates the broad expertise of Mr Popov in the NLP techniques: for example in parsing, as well as his extensive knowledge about resources created within the BulTreeBank projects and other large graph-based semantic models that are popular lately especially for English. As far as I know, the presented dissertation for the first time examines graph-based representations of Bulgarian text in such depth and volume. Furthermore, due to the size of the experiments, after reading Chapters 4 and 5, the reader assesses the candidate's skills as a programmer and his admirable expertise as a professional computer scientist who is well informed about a variety of techniques and applies them with ease. This impression is deepened with the following chapters and a positive opinion on Alexander Popov's extensive knowledge in linguistics and the existing language resources for different languages is added.

Chapter 6 (Distributed Representation of Words, Lemmas and Senses Based on Lexical Resources) explores ways of presenting words in context, the main object being the vectors embeddings. Here the models use NNs, trained on artificial lexical unit sequences generated by walking on graph-based text representations (pseudo-corpora). The knowledge graphs are generated with different sets of relations by random walks with different lengths. Word2vec provides the vector representations of the lexicon. The resulting model is evaluated on the task of measuring word similarity and relatedness. It is shown how the knowledge graph density can be increased by including relations between predicates and arguments given certain filter that helps extracting meaningful relationships. In the present work, the relations "subject", "direct object" and "indirect object" have been used. The assessment of the received representations' quality is done again using the task of measuring word similarity and relatedness; it shows an improvement of over 1-3% depending on the inputs. This is an indication about the productivity of the approach and raises the question of how to include other meaningful semantic relationships in the construction. In principle such relations can be derived from various texts, including the definitions in WordNet or Simple English Wikipedia (the latter texts are generally simplified and their NLP analysis should be easier).

Chapter 7 (Recurrent Neural Networks for Word Sense Disambiguation) presents two neural architectures A and B that approach the WSD task in a different way and corresponding experiments with evaluation. Architecture B differs from Architecture A in the final phase of context representation. Architecture B has tested a context-enhancing strategy with more semantic features, as English Wikipedia has been lemmatised and concatenated to the training corpus, over which 300-dimensional vector representations have been created. Three models of vector representations have been trained with architectures A and B, and the assessment of their quality shows that they are close to the best known models in the field. The concatenation of the lemmatised Wikipedia to the pseudo-corpus significantly improves the vector quality and the model surpasses the features of two well-known models presented in 2015 and 2016.

Chapter 8 (Multi-task Learning with Recurrent Neural Networks) presents results related to studies of neural networks trained in parallel to share parameters between the two. Here the author explores the possibilities of doing WSD in parallel with other NLP tasks. It has been shown that combining a WSD-classifier with a learner of context embeddings provides a more accurate model for both tasks than NNs trained for each task individually. Experiments were also conducted to parallel NNs trained to recognise POS and to solve WSD. The model trained on both tasks permits sharing of principles and parameters and performs better than individually trained models. As the author writes, “a unified solution that is able to model language in many different ways, while sharing most of its parameters amongst the kinds of analyses it produces, would be a serious step towards building multi-purpose and complexly structured linguistic and conceptual representations that resemble human thought”.

The conclusion (Chapter 9, Summary and Outlook) summarises the results obtained and discusses their importance. The thesis contributions are related to the improved vision of the lexicon as a set of vector models in a mixed space of words, basic forms, grammatical and semantic information. Applied achievements are related to the proposed NN text processing architectures and the experimental evidence of their useful features and application to key NLP tasks. It is obvious (as evidenced by the immediately occurring citations in publications by foreign authors) that some of the suggestions for the representation of the lexicon through vector models are innovative and contribute to the international state of the art: these are the augmentations of morphological information (suffixes) and grammatical roles to the model as well as the training of distributive representations of basic forms and synsets by generating pseudo-corpora. It was a very good idea to include the table on pages 111-115 with author’s publications on the subject and a brief summary of the content - it allows to follow the development of the research work over time. The list of ideas for future work shows an ambition to carry out world-class research and compare the obtained results to the best-known achievements. The tasks concern both the technological side of the IT tools used and the quality of the linguistic resources.

In general, the dissertation text is tight and well organised, with a clear chapter and subject division, and a consistency of expression. As a technical comment, I would point out that the footnote numbering starts from '1' in the individual chapters (which is not typical of a holistic, integrated text). The Autoreferat in Bulgarian reflects correctly the content of the dissertation. Translation equivalents of the English terms have been thoroughly sought, and I think the current version of the Autoreferat is much better than the previous one in this regard. My comments on the version of the dissertation that was submitted for the preliminary defense are also reflected.

According to my personal observations from meetings and seminars, Alexander Popov is the author of the presented innovative results, which is confirmed by the declaration of originality. I have witnessed the great interest in his presentations and posters at prestigious international conferences. During his four years of doctoral studies, Alexander Popov has accumulated the capacity of a mature specialist in computational linguistics and NN-based text processing applications. He is an excellent example of an interdisciplinary expert between linguistics and computer science, and the thesis text is a witness of his competence in both areas.

Conclusion. I believe that the results obtained and the published papers prove the author's expertise and capacity for conducting independent scientific and applied work, which are required by the Law on the Development of the Academic Staff in the Republic of Bulgaria for awarding the educational and scientific degree "doctor" (PhD). The dissertation is impressive with its ambition for positioning on a world level in comparison to the best achievements in the field. On these grounds, I will vote positively to award the degree and I propose with conviction to the honourable **Scientific Committee to award to Alexander Popov the educational and scientific degree "doctor" (PhD) in computer science**.

October 2018

Member [redacted]