Skoltech

Skolkovo Institute of Science and Technology

Jury Member Report – Doctor of Philosophy thesis.

Name of Candidate: Irina Nikishina

PhD Program: Computational and Data Science and Engineering

Title of Thesis: Learning linguistic tree structures with text and graph methods

Supervisor: Assistant Professor Alexander Panchenko

Name of the Reviewer:

I confirm the absence of any conflict of interest (Alternatively, Reviewer can formulate a possible conflict)	**Date: 19-09-2002**

The purpose of this report is to obtain an independent review from the members of PhD defense Jury before the thesis defense. The members of PhD defense Jury are asked to submit signed copy of the report at least 30 days prior the thesis defense. The Reviewers are asked to bring a copy of the completed report to the thesis defense and to discuss the contents of each report with each other before the thesis defense.

If the reviewers have any queries about the thesis which they wish to raise in advance, please contact the Chair of the Jury.

Reviewer's Report

Reviewers report should contain the following items:

- Brief evaluation of the thesis quality and overall structure of the dissertation.
- The relevance of the topic of dissertation work to its actual content
- The relevance of the methods used in the dissertation
- The scientific significance of the results obtained and their compliance with the international level and current state of the art
- The relevance of the obtained results to applications (if applicable)
- The quality of publications

The summary of issues to be addressed before/during the thesis defense

The PhD thesis of Mrs. Irina Nikishina approaches the problem of Taxonomy Enrichment (TE). The TE the methods for automated extension of existing taxonomies by adding new concepts. The TE problem has received attention in both the research and industrial community. Thus, there is no doubt that the developed datasets and methods along with experimental results make a valuable contribution to the current state of Natural Language Processing (NLP).

The thesis is well-structured, so that every Chapter presents with a standalone application. Chapter 1 makes an introduction to the problem and outlines main research questions, tackled in the PhD project. Chapter 2 provides an extensive related works review, comparing different TE setups and two main research directions, namely, use of either word or graph embeddings. Chapter 3 introduces two novel datasets, designed for the Russian and English languages, to serve as a test-bed for TE methods further. These datasets rely on different versions of both Russian and English taxonomies, used to simulate the process of adding new concepts. Chapter 4 introduces the formal TE task setup and lifts the limitations used in previous works, which heavily rely on having access to word definition. At the same time, this Chapter introduces the TE baseline DWRank, which runs Logistic Regression with features, derived from distributional similarity. Chapter 5 extends the DWRank method with features, derived from hierarchical representations, such as hierarchical embeddings and graph neural networks. Chapter 6 conducts an extensive comparison of DWRank and its extension to most recent TE methods and shows that DWRank establishes new state-of-the-art performance. This Chapter continues with a detailed error analysis with respect to the candidate's number of senses and part of speech, discovering 5 main reasons for errors. Section 6 lifts further restrictions and demolishes the need for candidates for new words. To this end, pre-trained language models combined with graph neural networks generate [mask] replacements for carefully selected patterns. As a backbone technology, hidden state projections methods are used. Last, but not least, the Chapter 8 show-cases the use of developed methods in the TaxFree toolkit.

The PhD thesis is well-written and the writing style is solid. The dissertation meets expectations for PhD levels in Data science. The topic of the thesis is relevant to its content. The choice of methods is substantiated. The results are published in top-tier venues. The Chapter 8 of the dissertation show-cases a real-life application which can be used by NLP practitioners.

Nevertheless there are several minor drawbacks in the dissertation. In particular, the dissertation could be written in a more formal way. There are at least two examples where the writing could be improved.

1. The tasks and evaluation metrics could be formulated in a more formal way. Chapter 6 uses MAP, while Chapter 7 uses MRR. It would be beneficial to use a consistent set of metrics throughout experimental evaluation.
2. Section 4.2 states that the baseline method utilizes Linear Regression, while further Section 4.4 builds upon Logistic Regression. Anyway, it should be stated in a clearer way a) what method is used, b) what are the target values and what learning objective is minimized to train the model.

Overall, I believe that Mrs. Irina Nikishina has composed a high-quality dissertation. The dissertation explains how a taxonomy can be extended with novel concepts without having access to concepts’ definition or even without accessing potential candidates. The scientific contributions of this dissertation are aimed at utilizing unstructured information stored in form of word or graph embeddings with the aim of enriching a given taxonomy. Moreover, this dissertation has a number of practical applications that NLP practitioners may benefit from. The dissertation makes a methodological contribution to inducing linguistic structures from texts.

Provisional Recommendation

X *I recommend that the candidate should defend the thesis by means of a formal thesis defense*
☐ *I recommend that the candidate should defend the thesis by means of a formal thesis defense only after appropriate changes would be introduced in candidate’s thesis according to the recommendations of the present report*
☐ *The thesis is not acceptable and I recommend that the candidate be exempt from the formal thesis defense*

Dr. Ekaterina Artemova

Faculty of Computer Science

National Research University Higher School of Economics

email: elartemova@hse.ru

tel.: 8 916 972 86 46