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: Anh-Huy Phan

I confirm the absence of any conflict of interest (Alternatively, Reviewer can formulate a possible conflict)	**19.09.2022** **Date: DD-MM-YYYY**

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

The Phd thesis of Irina Nikishina contains 9 chapters and an appendix for additional experimental results. The thesis covers the methods that Nikishina developed (and published) for the problem of Taxonomy Enrichment, i.e., adding new words to existing taxonomy. The main achievement is the fusion method which incorporates various best deep representations of graph structures, like node2vec, Poincare embeddings, GCN, text-based contextualized representations from transformer networks, to achieve the state of the art results.

Nikishina has published 9 publications with 4 as the first author. Nikishina has improved structure and presentation of the thesis. I appreciate the author presented a comprehensive overview of the related methods, provided very illustrative examples for the studied problem, as well as detailed analysis of the different common errors in the experiments.

In this version of the Phd thesis, the reference section is still not good as the other chapters because of duplicate references.

For example

Georgeta Bordea, Paul Buitelaar, Stefano Faralli, and Roberto Navigli. SemEval-2015 task 17: Taxonomy extraction evaluation (TExEval). ... June 2015.

Georgeta Bordea, Paul Buitelaar, Steven Faralli, and Roberto Navigli. Semeval-2015 task 17: Taxonomy extraction evaluation (texeval). ... pages 1081–1091, 2016a

Georgeta Bordea, Els Lefever, and Paul Buitelaar. SemEval-2016 task 13: Taxonomy extraction evaluation (TExEval-2). 2016b.

Georgeta Bordea, Els Lefever, and Paul Buitelaar. Semeval-2016 task 13: Taxonomy extraction evaluation (texeval-2). Evaluation (SemEval-2016), pages 1081–1091, 2016c.

Jose Camacho-Collados, Claudio Delli Bovi, Luis Espinosa-Anke, ... SemEval-2018 task 9: Hypernym discovery. 2018a.

Jose Camacho-Collados, Claudio Delli Bovi, Luis Espinosa-Anke, ... Semeval-2018 task 9: Hypernym discovery. 2018b.

David Dale. A simple solution for the Taxonomy enrichment task: Discovering hypernyms using nearest neighbor search. In Computational Linguistics and Intellectual Technologies: papers from the Annual conference “Dialogue”, 2020a.

David Dale. A simple solution for the taxonomy enrichment task: Discovering hypernyms using nearest neighbor search. In Computational Linguistics and Intellectual Technologies: papers from the Annual conference “Dialogue”, 2020b.

David Jurgens and Mohammad Taher Pilehvar. SemEval-2016 task 14: Semantic taxonomy enrichment. June 2016a.

David Jurgens and Mohammad Taher Pilehvar. Semeval-2016 task 14: semantic taxonomy enrichment. Semantic Evaluation (SemEval-2016), pages 1092–1102, 2016b.

Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. 2016a.

Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In International Conference on Learning Representations (ICLR),2017.

Maria Kunilovskaya, Andrey Kutuzov, and Alister Plum. Taxonomy Enrichment: Linear Hyponym-Hypernym Projection vs Synset ID Classification. 2020a.

Maria Kunilovskaya, Andrey Kutuzov, and Alister Plum. Taxonomy enrichment: Linear hyponym-hypernym projection vs synset id classification. 2020b.

.

Tom´as Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. ICLR 2013, 2013a.

Tom´as Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. 2013b.

George A Miller. Nouns in wordnet. WordNet: An electronic lexical database, pages 23–46, 1998a.

George A Miller. WordNet: An electronic lexical database. MIT press, 1998b

Irina Nikishina, Varvara Logacheva, et al . Russe'2020: Findings of the first taxonomy enrichment task for the russian language.

Irina Nikishina, Varvara Logacheva, et al. RUSSE'2020: Findings of the First Taxonomy Enrichment Task for the Russian Language.

Ashish Vaswani, Noam Shazeer, Niki Parmar, et al. Attention is all you need. 2017a.

Ashish Vaswani, Noam Shazeer, et al.. Attention is all you need. , Inc., 2017b.

- Chapter 2 introduces problems of Taxonomy Enrichment (TE), Taxonomy Induction, background of DNNs applied to TE, presents comprehensive review on existing representation methods for taxonomies.

The thesis may formulate the mathematical model for Taxonomy enrichment problem. This will help readers to understand the general framework for TE.

- Captions of figures should be placed at bottom of the figures.

- Section 1.2.

Define the entity set and the relation set

- Page 30, Line 2

Remove the redundant close curly bracket

- Page 30. duplicate text

"The authors did not manage to outperform the first-word baseline, however, they computed the upper bound which is 0.7, which means that the developed methods still demonstrate decent results."

- Page 36

"Compared to the above mentioned competitions, RUSSE'2020 is closely related to the SemEval-2016 Taxonomy Enrichment Task"

RUSSE'2020 is compared with other tasks, but it has not been introduced yet on page 36.

- Page 39, introduction to Transformer

Write full name Recurrent Neural Networks before the first use.

"Transformers have classic encoder-decoder architecture"

"classic" -> "classical" or "ordinary"

- Page 41

why are there vertical bars in (2.4), (2.5)?

- Page 42

Review section for BERT: should mention applications of BERT to TE?

- Pages 45, 46 and 47

Different notation for the Inner products in (2.7), (2.10), (2.11) (2.12), (2.15), (2.16). Should unify the notation.

- Page 51

"in this paragraph" -> in this section

- (2.12) and (2.13)

v_c, v_w and z_g are vectors, they should be denoted by similar notation.

- Page 52 (2.17)

the matrix M is not defined yet in (2.17)

- Page 54 : What is the role of the vertical bar in (2.20)?

- Page 56, inconsistent notation for vectors z in (2.16), u and v in (2.19), (2.20) and h_i on page 56.

- (2.23) and (2.21) : Unify and correct notation for matrices

- Chapter 3: What make the new dataset unique, different from the other datasets?

- Page 74

"RUSSE'2020 Taxonomy Enrichment task for the Russian Language. The method achieved the best result on the nouns track. Therefore, we consider it the state-of-the-art method for Russian."

Please correct if I misunderstood

- The student created a new dataset (presented in Chapter 3), organised the RUSSE-2020 competition on Taxonomy Enrichment for Russian language, proposed an approach [Nikishina et al., 2020b]) for their own dataset in the same competition, and claimed it the state-of-the-art method for Russian language?

Has the proposed method been tested for other datasets, e.g. for other languages? Did it achieve SOTA result?

- Page 78: Chapters 2 and 4 review existing methods for TE. Why don't write them in one chapter?

- Page 95: Figure 6-5 should be on top of the next page.

- Error analysis

Type 1

" Extracted nearest neighbours can be semantically related words but are not necessary co-hyponyms"

Does it mean the mathematical model for this problem is incorrect?

The example for "delist" (WordNet) is not clear, since predicted senses are quite close to the expected senses.

"delist (WordNet); expected senses: get rid of; predicted senses: remove, delete"

Type 2. "Distributional models are unable to predict multiple senses for one word"

Is it because of small sample size? Have the authors checked prediction accuracy for various sample sizes?

- Page 134

"SKOS Knowledge Organization Systems" should be Simple Knowledge Organization Systems

Provisional Recommendation

☐ *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*