Supervisor's review of master thesis

Author of the review:	doc. RNDr. Pavel Pecina, Ph.D. Institute of Formal and Applied Linguistics Faculty of Mathematics and Physics, Charles University
Author of the thesis:	Felipe Vianna

Title of the thesis: Expert Classification and Retrieval

The thesis of Felipe Vianna belongs to the large area of Information Retrieval. It tackles two tasks: expert retrieval (also called finding) and expert classification (also called profiling). In expert retrieval, for a given query (topic) the task is to find a (ranked) set of people who are experts on this topic. In expert classification, the task is to assign a set of topics (classes) to each person reflecting their expertise.

The approach adopted in the thesis is to estimate all associations between experts and topics and to use this information for both the tasks: In retrieval, for a given topic a set of highly associated experts are retrieved. In classification, for a given expert highly associated topics are presented as the classes the expert belongs to. This approach is in this work tailored to the specific needs of a private company that provided the data for the experiments. The data is not public, but the source code and results are.

The thesis is structured into six chapters. Chapter one introduces the two tasks and how they can be evaluated. Chapter two reviews related work and state of the art in this field including relevant NLP and ML concepts exploited in the thesis. Chapter three presents the theoretical aspects of the proposed approach. Chapter four than gives practical details on the implementation. Chapter six contains description of the conducted experiments and their results. Chapter seven concludes the work with a summary of achieved results and the most interesting findings. The main text of the thesis spans the total of 82 pages.

The structure of the thesis is not prefect. The description of the methods is (to some extent) unsystematically distributed in chapters three, four and five. A better organization of the text would have helped to present a more coherent and clearer picture of the work done. Also, the data and the way how it was split into subsets is described in multiple places throughout the thesis which makes it difficult to understand. The text is written in English and contains occasional grammatical errors (mostly in preposition and article usage) which makes the text sometimes difficult to follow.

I appreciate the related work chapter which is quite rich and provides a broad context of the work with details on methods and concepts used in the approach taken.

The proposed method is based on multi-class classification by a neural network which assigns an association score to each expert and topic. The entire procedure also involves a num-

ber of data processing steps (such as keyword extraction, keyword augmentation, taxonomy topic matching, topic filtering, etc.). The entire system thus has a lot of (hyper-)parameters requiring optimization. Since automatic optimization was often not possible, the author employed a thorough analysis of intermediate results (supported by a number of plots visualising statistical analysis of the data) to decide the parameter values.

Two novel and interesting ideas were proposed and implemented – the convolutional layers in the NN architecture applied to features ordered based on linearization of the topic hierarchy and the forgetting factor which decreased the effect of older documents. Sadly, none of them showed interesting results.

The overall results are positive and promising for future work. The authors was able to integrate various types of data, link them through a topic taxonomy and train a model which can predict/recommend suitable topics to experts.

The work includes several unclear/questionable points:
1) For a reason which is not explained, the proposed method assumes a closed set of topics (queries). If a user issues a query which does not match with any of the topics the system is trained for, the system will retrieve an empty set of experts (The baseline model presented in Section 4.1 does not suffer from this problem). In addition, the author ignores (removes) topics which are not frequently present in the data. I would understand excluding such topics from the training phase, but not from testing.
2) The data processing and split into sets S_1,...S_4 plus a split into training, validation and test subsets is not very clear. In Section 4.8, the author claims that dimensionality reduction is an important step in date preprocessing, but it is not clear why. The proposed method based on pair-wise correlation is not described sufficiently (it seems that removing one feature does not effect removal of other features, which might result into removing to many features).
3) The topic filtering described in Section 3.3.1 is based on this idea: "If many topics have high cosine similarity to each other, very likely it is a result of bad performance of the encoder.", which does not make much sense.
4) The baseline model is not built using the same data as the other model and it is not fair to compare its results with those of other models. The comparison of the achieved results with the results from the TREC Enterprise shared task seems not fair as well, unless the test data was the same which is probably not the case.

Despite the issues mentioned above, the author demonstrated his creativity and ability to work independently and he delivered a working solution for the given problem. I recommend the thesis to be defended.

Prague 27.1.2020