Deep learning for academic expert search

HOUATI Chakib Mouloud, MEZIANI Serine

University Of Science And Technology Houari Boumediene
Chakibhouati@gmail.com mzserine@gmail.com

Abstract Scientific expert search is a recurrent task in the academic world. Indeed, it is often necessary to look for supervisors, jury members for end-of-study or thesis projects, evaluators for research project proposals, referees for articles submitted to scientific journals or members of conference program committees. This research is made very difficult by the increase in the number of researchers and scientific work as well as the great diversification and specialization of the scientific research fields.
The objective of this work will be to design an expert finding system that enhanced through the expansion of queries expressed in natural language. This system will be based on the data of the articles published by the researchers (title, abstract, keywords). The proposed system will integrate the most recent techniques in natural language processing (NLP) based on deep-learning, in particular the transformer-based machine learning technique *BERT* (*SciBERT* and *RoBERTa*), developed by *Google*, and published in 2018. We will present three main contributions: a new indexation, expansion of queries by using the definition, and a score distribution method. In addition to this, the construction of a new test corpus based on ACM.
A national platform for expert research was created at the end of this project. This portal integrates the methods that we proposed, and validates them on the case of the Algerian scientific production.

Keywords: Scientific Expert Finding ; Deep Learning ; Word Embeddings ; Bibliographic Databases; Voting Models; Query Expansion; BERT; SciBERT; RoBERTa; FAISS

1 Introduction

In 2019, Algeria had more than 36,000 active researchers spread across 1470 laboratories and research centres at national level. Nevertheless, the Minister of Higher Education and Scientific Research at the time declared that "an enormous gap exists between economic enterprise and scientific research", and the same source deplored a potential that remains "under-exploited by the economic sector and public enterprises in particular" [1]. Human potential is a valuable resource, which, if managed properly, will allow for more efficient management of different projects, thus increasing the productivity of the company or other public organisation. However, in order to exploit this resource, the first step is to find the right people to trust, which is difficult and not very clear.
Indeed, until recently, finding the right people for a task required manually contacting individuals, asking them for suggestions, in the hope that the judgement made on others would be relevant and justified. However, in the academic field, a researcher's activity is sometimes not well known by his or her peers, or there are individuals willing to collaborate but they are not well known. In order to automate this process, and make it more accurate, the last few years have seen the emergence of so-called expert search systems. In the literature, the term expert refers to a person who has knowledge, skills and experience in a given field. In Algeria, this type of system is not yet used, but their adoption will allow a better

connection between the scientific community and the various public and private bodies. Thus, not only will a private company, for example, be able to more easily find the best available skills in a newly launched technology, but it will also be possible to better distribute academic tasks such as the training of judges for thesis, the selection of speakers for conferences, or even to efficiently find the necessary personnel for a multidisciplinary scientific collaboration, by means of a simple query containing the necessary key words. The aim of our second year Master project is to set up a national expert search portal, which brings together all Algerian researchers, and which uses the latest advances in automatic language processing. Expert search is part of information retrieval. This discipline is one of the most prolific in computer science. Recently, we have seen the emergence of several models using deep learning for automatic natural language processing tasks such as *Word2Vec* or more recently *GPT-3*. Since their appearance, these models have been very successful, especially in machine translation. *BERT* [2] is one of these models which appeared in 2018, and was made freely available by *Google*. It is a language model, which uses a structure of *transform*. It is used in order to capture the semantics of a word, in a vector, according to its context of appearance. This technique is known as *embedding*. In any scientific expert search system, the first step is to represent the expert's profile (his expertise), depending on the type of system, this requires obtaining data about him. In our case, we have represented the profile of each researcher by the documents he has published. We were inspired by the work of Berger et al. [3], who were one of the first to use embedding with *BERT* in expert search. We will use their best result as a base model for comparisons. In order to produce a more efficient system we will propose three main contributions, each concerning a step of the search process. Firstly, a new approach to indexing, which is based on individual sentences instead of the whole document, during this phase we will compute plots with two variants of the *BERT* model which are *SciBERT* and *RoBERTa* of *Facebook*. This will allow us to compare the efficiency of these two models. Secondly, we will modify the query initially introduced by the user, using its definition, and see if this improves the final result. This technique is called query expansion. Thirdly, we will propose a new way of distributing a score over the authors of the same document, in order to limit the influence of very prolific authors. The tests with the basic model will be done in a first step on the database published in open access by the authors: Arxiv+MAG. Some limitations have been noticed in this database. For this reason, we decided to create a new dataset using the technique of *web scraping* from the ACM library. This new database will be more representative of the authors' expertise. We will make it public for future research. During the computation of the plunges of *SciBERT*[4] and *RoBERTa*[5], we found that using *CUDA* technology from *Nvidia* for parallel computation, brings a remarkable acceleration compared to the CPU.

2 Chapter one

In the first chapter, we first discussed the basic concepts of information retrieval (IR). Expert search is part of it, and several of these concepts are present in the literature. A section dedicated to the methods used in IR allowed us to introduce the concepts of vectors and similarity, in addition to the concept of folding, which we detailed further when we talked about neural networks in IR, we also presented in this section a classification of existing models. After defining some of the evaluation metrics used in our work, we discussed the concept of query expansion, its utility, and its state of the art. As mentioned before, a section dedicated to the use of neural networks in IR is presented, where we have focused on the different ways to learn good semantic representations of terms and documents, and we have dis-

cussed the different language models used in our project. Finally, the last section of this chapter presents expert research. We have mentioned the areas where it is most present and the factors that differentiate it from other IR tasks. We have also discussed the models used, as well as the available databases.

3 Chapter two

In this chapter we have presented the three main parts of our contribution, each of which touches on a level of research. Starting with the sentence indexing stage. Then, the modification step to the voting formula in order to limit the impact of very prolific authors, and finally at the level of the introduced query, to which we add its definition in the search process, in two different ways.

4 Chapter three

In this chapter, we first presented the different tools used in our project. Then, we presented the results of the evaluations of each of the two databases used (Arxive+MAG, ACM) with our expert search system, which uses sentence indexing, and the authors' system [3] which uses document indexing, by introducing for each system other contributions, namely the proposed query expansion techniques. In other words, we will compare the two types of indexing, and for each type, we experiment with the contribution of the different types of query expansion: Average, Hybrid. Finally, we will talk about the robustness of our system, to the change of the databases, as well as the future work to be done.

5 Chapter four

In this chapter we come to the implementation of our portal for the search of Algerian academic experts. This portal will allow us to use the proposed approaches with Algerian specialists and experts. In the first part of this chapter, we presented the tools and the development language used for the realization of the website. We then presented the different web pages of our site. At the end, we demonstrated all the functionalities of our website, illustrating it with screenshots and explanations of an example.

6 Conclusion

The work we have described in this thesis was aimed at searching for scientific experts using deep learning, and in particular pre-trained language models based on the *BERT* model proposed by *Google*.We have built on the work of Berger et al. [3]. We explored three ways to improve on this work: 1) a new indexing method for documents, which is phrase indexing, 2) use of two methods to augment the query, by extracting the definition from *Wikipedia*, 3) modification of the voting formula by incorporating the author's domain information.We combined all these methods with two variants of the *BERT* model, namely *SciBERT* and *RoBERTa*, to compare with our proposed improvements. The conclusion is that our best performing model has surpassed the state of the art on the database, Arxiv+MAG, used by the authors of this work. The best combination obtained is the one using phrase indexing, with *SciBERT* as the folding model, and with query augmentation.

We have redone the evaluation on a new database, which we created ourselves using *web scraping* on the ACM library, and the results confirm overall that the methods we have proposed are better than the basic model. Another interesting result is that using the definition to augment the query was very beneficial. The modification we made to the voting formula is in a preliminary stage, and we found only minimal improvements, and in some cases none at all. Nevertheless, this modification is a promising avenue for further investigation. Finally, and to our surprise, *RoBERTa* was in most cases better for indexing than *SciBERT*.

The methods proposed in our project have been integrated into a platform that we have developed for the search of Algerian academic experts. The task of searching for experts being a recurrent task, the automation of which would give many advantages, we hope through our work to contribute to a better valorisation of the Algerian academic environment.

Finally, the results of our research are promising and can be improved by considering the following perspectives: 1) Use of other words close to the query (by metaheuristics for example) in addition to its definition, we could also consider the use of embedding by *SciBERT* when determining the words close to the query, 2) More precise use of the author's domain by considering the case of *clustering*, for example, 3) Amplification of the distance between the author's domain and the document to make the changes more significant, and thus obtain a greater difference in scores.

References

[1] APS. Interministerial Council on the promotion of research and development in enterprises. https://www.aps.dz/economie/96376 , 27-10-2019.

[2] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. Natural questions : a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7 :453-466, 2019.

[3] Mark Berger, Jakub Zavrel, and Paul Groth. Effective distributed representations for academic expert search. *arXiv preprint arXiv : 2010.08269*, 2020

[4] Iz Beltagy, Kyle Lo, and Arman Cohan. Scibert : A pretrained language model for scientific text.*arXiv preprint arXiv :1903.10676*, 2019.

[5] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta : A robustly optimized bert pretraining approach. preprint arXiv :1907.11692 , 2019