Deep Learning and Natural Language Processing for Expert Finding from Web Search Results

Nathalie Pernelle[a], **Jorge Garcia Flores**[a], **Pegah Alizadeh**[b], **and Ivan Vladimir Meza**[c]

[a]Laboratoire d'Informatique de Paris Nord (LIPN)–Université Sorbonne Paris Nord; [b]Léonard de Vinci Pôle Universitaire–Centre Recherche; [c]IIMAS–Universidad Nacional Autónoma de México (UNAM)

Our PhD proposal is focused on the development of an end to end Deep Reinforcement Learning model for expert finding, based on state of the art Natural Language Processing (NLP) and Information Extraction (IE) methods. The resulting model will be integrated into an existing expert finding system, that has been used by developing countries in the crucial task of getting in touch with their highly qualified diaspora. Therefore, the PhD student will integrate an ongoing international collaboration (partially funded by an Ecos Nord and a Labex EFL subventions) where she/he will be in touch with foreign partners providing evaluation data and results interpretation.

Research proposal

Most of current expert finding approaches work on highly organized data sources, like scientific publications, semantic web profiles (1) or expert topic relationships (2). These approaches usually require a significant manual data preparation while using scientific data sources such as DBLP, HAL or DOI. Therefore, they are only are able to find experts from the scientific community, but not from other domains (like business or arts) (3).

We propose an original approach based on information extraction from a noisier data source: web search query results. Deep Reinforcement Learning (DRL) has shown promising results on information extraction from this kind of data (4). However, the semantic association of web search results to experts, and the extraction of meaningful semantic information (affiliation institution, city, year, scientific discipline) from the resulting dataset is challenging both from a scientific and a technological perspective. As web search querying and observing all snippets are expensive from an algorithmic point of view, reinforcement learning is intended to optimize the query formulation process and result exploration policy with neural network based algorithms such as Deep-Q networks (5, 6), Monte Carlo tree search (7).

This approach will be evaluated on an existing database (*#AlExtranjero*) of international experts coming from developing countries and having obtained an academic degree abroad (6). An ongoing expert finding system is currently under development (8, 9), and the PhD candidate is expected to participate in this process.

PhD objectives

Diego Martínez was born in Buenos Aires, where he studied up to a Biotechnology Master. In 2010 he got a scholarship to study a Phd in Essex University. Where is he now? Did he stay in the UK, or did he come back to Argentina? That is the kind of use case our expert finding method could be used for.

While the main goal of the PhD is to develop an end to end DRL method for expert finding from web search results, it can be boiled down in the following scientific and technological challenges:

1. *The generation of semantically rich web search queries using machine learning approaches.*
   The queries are generated semantically from a person name, a scientific discipline or a toponym. After taking the expert name as an input, an important challenge is to learn the most informative queries in order to generate snippets related to the given name from the Web (10, 11). We propose to refine the queries by learning effective key words or by reformulating queries using Deep Learning (DL) and NLP methods.

2. *Finding the optimal DRL strategy for selecting among web search results.*
   Since the search space is the Web and it contains too many snippets and documents, observing and analyzing all snippets is not feasible in terms of time and memory complexity. In this work, we want to investigate in using DRL approaches for finding optimal strategies for generating queries from the previous point and observing snippets effectively (4, 7).

3. *Integration of NLP/IE techniques suitable to our method.*
   Study the state-of-the-art NLP/IE techniques for extracting person names, scientific disciplines, institution names, toponyms and dates. Adaptation of the most suitable techniques to our web search oriented approach.

4. *Entity resolution and entity linking.*
   It has been shown recently that both can be more robust when the entity description involves a relational context, and some semantic knowledge such as functional predicates (e.g. birth-date versus research topics) or Local Completeness Assumption (12, 13). In this work we want to investigate for the first time how to enrich DL models with semantic knowledge when data and knowledge are incomplete and evolutive.

5. *Computational evaluation.*
   We evaluate our methods on the *#AlExtranjero* database of 8,000 Latin American researchers who got a scholarship to study a PhD abroad (6).

References

1. B Sateli, F Löffler, B König-Ries, R Witte, Scholarlens: extracting competences from research publications for the automatic generation of semantic user profiles. *PeerJ Comput. Sci.* **3**, e131 (2017).
2. J Liu, et al., A topic rank based document priors model for expert finding. *In Adv. Comput. Methods Life Syst. Model. Simul.*, 334–341 (2017).
3. P Buitelaar, G Bordea, B Coughlan, Hot topics and schisms in NLP: community and trend analysis with saffron on ACL and LREC proceedings in *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC*. pp. 2083–2088 (2014).
4. K Narasimhan, A Yala, R Barzilay, Improving information extraction by acquiring external evidence with reinforcement learning in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. pp. 2355–2365 (2016).
5. V Mnih, et al., Playing atari with deep reinforcement learning. *CoRR* **abs/1312.5602** (2015).
6. P Alizadeh, JG Flores, IV Meza Ruiz, Apprentissage par renforcement pour la recherche d'experts sur le web in *EGC*. (Brussels, Belgium), (2020).
7. G Liu, X Li, J Wang, M Sun, P Li, Extracting knowledge from web text with monte carlo tree search in *Proceedings of The Web Conference 2020*, WWW '20. (Association for Computing Machinery, New York, NY, USA), p. 2585–2591 (2020).
8. J García Flores, P Zweigenbaum, Z Yue, W Turner, Tracking researcher mobility on the web using snippet semantic analysis in *Advances in Natural Language Processing*, eds. H Isahara, K Kanzaki. (Springer Berlin Heidelberg, Berlin, Heidelberg), pp. 180–191 (2012).
9. W Turner, J Garcia Flores, M de Saint Leger, Computer supporting diaspora knowledge networks: a case study in managing distributed collective practices in *Diaspora: towards the new frontier*, ed. JB Meyer. (IRD, Institut de recherche pour le développement, Marseille), pp. 213–243 (2015).
10. R Nogueira, K Cho, Task-oriented query reformulation with reinforcement learning in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. (Association for Computational Linguistics, Copenhagen, Denmark), pp. 574–583 (2017).
11. E Amigó, et al., Weps3 evaluation campaign: Overview of the on-line reputation management task in *CLEF 2010 LABs and Workshops, Notebook Papers*. (2010).
12. V Huynh, P Papotti, A benchmark for fact checking algorithms built on knowledge bases in *ACM SIGMOD International Conference on Management of Data*. (ACM), p. to appear (2020).
13. R Cappuzzo, P Papotti, S Thirumuruganathan, Local embeddings for relational data integration. (2020).