_____

**We offer two fully funded PhD positions at IRIT (Toulouse, France)**

*Topic 1 - Knowledge-aware few shot learning*
Deep learning models, including transformers, are well-known for their state-of-the art effectiveness in a wide range of tasks including recommendation (Liu, 2021), information retrieval (IR) (Lin and al. 2020) and Natural language processing tasks (NLP) (Chernyavskiy et al., 2021). However, they require to be trained on large amounts of labeled datasets to generalize to new instances and new tasks. Such labeled data is costly in domain-specific scenarios because labeling requires professional expertise (eg., human resource recruitment and management expertise to map relevant profiles to job posts, map relevant skills to job titles, etc.). Besides, data classes and related features evolve over time leading to the prevalence of observations/classes/instances in the testing/deployment phase unseen in the training phase. Training deep learning models with biased available ground truth data distributions lead to skewed predictions and low performance levels (Geng et al. 2019). For instance, the powerful BERT-based neural ranker may often be unstable and overly confident during the learning process (Qiao et al., 2019).
To overcome these issues, Few-shot learning techniques have been proposed in the literature (Wang 2021). Few-shot learning is a type of machine learning problem where training data contains only a limited number of examples with supervised information. Related learning problems include weakly supervised learning, imbalanced learning and transfer learning. For instance, multi-task fine-tuning and unsupervised data augmentation have been shown to be successful when applied to transformers in general domain few-shot settings (Qiao et al., 2019, Finn et al. 2017, Snell et al., 2017, Delpierre et al., 2020).

In this thesis, we would like to tackle in-domain text ranking and text classification tasks in a few-shot learning setting, as follows:

1 . Enrichment of Language model with domain-specific Knowledge Graphs: Two subtasks may be addressed here. First, build a domain-specific KG by extracting information from job offers, career profiles, etc. Second, enriching SoTA PLM models to encode the domain-specific knowledge.

2. Design semantic text ranking models with few labels and weak supervision: 1) learn a constrained pseudo label generation model: for each hard (q,d) label (eg. job post and candidate profile, current job and following job), train a label generator model to learn generating query paraphrases constrained with domain-specific semantic similarities learned in the knowledge graph; 2) learn a combined ranking (of jobs, candidates, skills, etc.) loss based on supervised and unsupervised losses related respectively to classification and clustering techniques with hard labels vs. representative instances of pseudo labels clusters; 3) in addition to ranking, the model could be trained to the auxiliary task of predicting mapping success as reward used to generate new pseudo-labels allowing to re-train knowledge-enriched language models.

*Topic 2 - Information retrieval models for structured and verbose queries*
Although multiple works addressed the verbose queries problem in the past, they are usually limited to queries that nowadays may be considered as short ones. Additionally, domain specific semantics represented on domain knowledge graphs may enriche the

document/query representation. In this thesis, we would like to approach two main problems in IR to correctly identify relevant documents for extremely long and structured queries, but under domain-specific knowledge graphs:
1 .  Enrichment of Language model with domain-specific Knowledge Graphs: Two subtasks may be addressed here. First, build a domain-specific KG by extracting information from job offers, career profiles, etc. Second, enriching SoTA PLM models to encode the domain-specific knowledge.
2 . Query Reduction of Verbose Queries: It implies preprocessing the query to substantially reduce its size in order to be processed by transformers-based models without losing query expressiveness. This step may include query reduction by choosing multiple sub-queries or weighting query concepts by using marking strategies.
3 . Properties of Verbose Queries: The main property to be studied is inherent structure as verbose queries may include indirect references (strong skills in databases of type « big data » ). However, other properties such as length distribution of queries, query types, repetition factor, smoothing mechanisms, one of multiple characteristics (NoSQL skills including “MongoDb, Cassandra, HBase…”), to mention a few.

---

Venue:
Toulouse is at the heart of sultry Southwestern France not far from the border with Spain. The balmy climate and friendly locals give Toulouse an inviting ambience. Toulouse is the center of the European aerospace industry, with the headquarters of Airbus (formerly EADS), the SPOT satellite system, ATR and the Aerospace Valley. It hosts the CNES's Toulouse Space Centre (CST) which is the largest space center in Europe, but also, on the military side, the newly created NATO space center of excellence and the French Space Command and Space Academy. Thales Alenia Space, ATR, SAFRAN, Liebherr-Aerospace and Airbus Defence and Space also have a significant presence in Toulouse. Although it is an industrial city, Toulouse is one of the most pleasant cities in France and offers multiple tourist attractions including medieval pilgrimages at the UNESCO-listed Basilique Saint-Sernin. Next, visitors can explore a 13th-century convent exemplifying Southern Gothic style, or spend time walking around the Place du Capitole, lined with red-brick architectural landmarks. Toulouse is renowned for its archaeology and fine arts museums, as well as its local culture. The douceur de vivre (good life) is felt at the sunny terraces of outdoor cafés and savored in the regional cuisine.

The University of Toulouse is one of the oldest in Europe (founded in 1229). Toulouse is also the home of prestigious higher education schools. Together with the university, they have turned Toulouse into the fourth-largest student city in France, with a university population of nearly 140,000 students.

Location:
Institut de Recherche en Informatique de Toulouse (IRIT)
Université Toulouse 3 Paul Sabatier (UT3)
118 Route de Narbonne, F-31062 TOULOUSE CEDEX 9, France
*Comment: Nantes (France) is also an alternative location if requested by the applicant as Synergie has also offices in Nantes.*

Starting date:
Fall 2022

Research team:
IRIS at IRIT : Lynda Tamine / José G Moreno / Taoufiq Dkaki
Synergie: Christophe Thovex

Profile:

- Master's level or engineering school in Computer Science, with skills in Information Extraction/Research and Text Mining
- Good English skills (written and oral)
- Good skills in advanced programming (Python, Pytorch, sklearn, ...)
- Good knowledge in Machine Learning, deep learning is a plus

Funding:
Total gross salary for 3 years : 105 401,50 € / Note that final monthly salary depends of personal situation in France

Application instructions:
All applications must include the following to be considered: detailed CV, cover letter, transcripts (with rankings), contacts for recommendation. Please use “PhD application - Synergie” as subject of the email and select a preferred topic (1 or 2) if any.
Applications to be sent by mail to Lynda Tamine, José G Moreno, and Taoufiq Dkaki (lynda.tamine@irit.fr, jose.moreno@irit.fr, taoufiq.dkaki@irit.fr).
All applications will be processed as they arise until the positions will be filled.

---

References

Samarth Bhargav, Georgios Sidiropoulos, and Evangelos Kanoulas. 2022. 'It's on the tip of my tongue': A new Dataset for Known-Item Retrieval. In Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining (WSDM '22).

Minghan Li and Eric Gaussier. 2021. KeyBLD: Selecting Key Blocks with Local Pre-ranking for Long Document Information Retrieval. Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval.

Manish Gupta and Michael Bendersky. 2015. Information Retrieval with Verbose Queries. In Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '15).

Jaime Arguello, Adam Ferguson, Emery Fine, Bhaskar Mitra, Hamed Zamani, and Fernando Diaz. 2021. Tip of the Tongue Known-Item Retrieval: A Case Study in Movie Identification. In Proceedings of the 2021 Conference on Human Information Interaction and Retrieval (Canberra ACT, Australia) (CHIIR ’21).

A. Arampatzis and J. Kamps. A Study of Query Length. In Proc. of the 31st Annual Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR), pages 811–812, 2008.

N. Balasubramanian, G. Kumaran, and V. R. Carvalho. Exploring Reductions for Long Web Queries. In Proc. of the 33rd Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR), pages 571–578, 2010.

M. Bendersky and W. B. Croft. Discovering Key Concepts in Verbose Queries. In Proc. of the 31st Annual Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR), pages 491–498, 2008.

M. Bendersky, D. Metzler, and W. B. Croft. Parameterized Concept Weighting in Verbose Queries. In Proc. of the 34th Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR), pages 605–614, 2011.

S. Huston and W. B. Croft. Evaluating Verbose Query Processing Techniques. In Proc. of the 33rd Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR), pages 291–298, 2010.

G. Kumaran and J. Allan. Effective and Efficient User Interaction for Long Queries. In Proc. of the 31st Annual Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR), pages 11–18, 2008.

G. Kumaran and V. R. Carvalho. Reducing Long Queries using Query Quality Predictors. In Proc. of the 32nd Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR), pages 564–571, 2009.

M. Lease. An Improved Markov Random Field Model for Supporting Verbose Queries. In Proc. of the 32nd Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR), pages 476–483, 2009.

J. H. Paik and D. W. Oard. A Fixed-Point Method for Weighting Terms in Verbose Informational Queries. In Proc. of the 23rd ACM Conf. on Information and Knowledge Management (CIKM), pages 131–140, 2014.

J. H. Park and W. B. Croft. Query Term Ranking based on Dependency Parsing of Verbose Queries. In Proc. of the 33rd Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR), pages 829–830, 2010.

N. Phan, P. Bailey, and R. Wilkinson. Understanding the Relationship of Information Need Specificity to Search Query Length. In Proc. of the 30th Annual Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR), pages 709–710, 2007.

Chernyavskiy, A., Ilvovsky, D., Nakov, P. (2021). Transformers: “The End of History” for Natural Language Processing?. In: Oliver, N., Pérez-Cruz, F., Kramer, S., Read, J., Lozano,

J.A. (eds) Machine Learning and Knowledge Discovery in Databases. Research Track. ECML PKDD 2021

Decorte, J., Hautte, J.V., Demeester, T., & Develder, C. (2021). JobBERT: Understanding Job Titles through Skills. ArXiv, abs/2109.09605.

Thomas Dopierre, Christophe Gravier, Julien Subercaze, and Wilfried Logerais. 2020. Few-shot Pseudo-Labeling for Intent Detection. In COLING pages 4993–5003, 2020

Chelsea Finn, Pieter Abdeel, Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks CORR, abs/1703.03400

Ruiying Geng, Binhua Li, Yongbin Li, Xiaodan Zhu, Ping Jian, Jian Sun. Induction Networks for Few-Shot Text Classification. Archiv 2019

Tianyu Gao, Adam Fisch, Danqi Chen. Making Pre-trained Language Models Better Few-shot Learners. ACL 2021

Jimmy Lin, Rodrigo Nogueira, and Andrew Yates. 2020 a. Pretrained Transformers for Text Ranking: BERT and Beyond. arXiv:2010.06467 (2020)

Davide Liu, George Philippe Farajalla, and Alexandre Boulenger. 2021. Transformer-based Banking Products Recommender System. SIGIR'21

Yifan Qiao, Chenyan Xiong, Zheng-Hao Liu and Zhiyuan Liu. Understanding the behaviors of BERT in ranking, arXiv 2019

Jake Snell, Kevin Swersky, Richard S. Zemel. Prototypical networks for few-shot learning. CORR, abs/1703.05175

Yaqing Wang, Quanming Yao, James T. Kwok, and Lionel M. Ni. 2020. Generalizing from a Few Examples: A Survey on Few-shot Learning. ACM Comput. Surv. 53, 3, Article 63 (May 2021)

Michiharu Yamashita, Yunqi Li, Than Tran, Yongefeng Zhang, Dongwon Lee. Looking Further into the Future: Career Pathway Prediction. Workshop, Job in the market place, WSDM'22.

Denghui Zhang, Junming Liu, Hengshu Zhu, Yanchi Liu, Lichen Wang, Pengyang Wang, Hui Xiong. Job2Vec Job Title Benchmarking with Collective Multi View Representation Learning. CIKM'19