Neuro-Symbolic Representations for Information Retrieval

Laura Dietz
University of New Hampshire, USA
dietz@cs.unh.edu

Hannah Bast
University of Freiburg, Germany
bast@cs.uni-freiburg.de

Shubham Chatterjee
University of Glasgow, Scotland
shubham.chatterjee@glasgow.ac.uk

Jeff Dalton
University of Glasgow
Jeff.Dalton@glasgow.ac.uk

Jian-Yun Nie
University of Montreal
nie@iro.umontreal.ca

Rodrigo Nogueira
State University of Campinas
rfn@unicamp.br

ABSTRACT

This tutorial will provide an overview of recent advances on neuro-symbolic approaches for information retrieval. A decade ago, knowledge graphs and semantic annotations technology led to active research on how to best leverage symbolic knowledge. At the same time, neural methods have demonstrated to be versatile and highly effective.

From a neural network perspective, the same representation approach can service document ranking or knowledge graph reasoning. End-to-end training allows to optimize complex methods for downstream tasks.

We are at the point where both the symbolic and the neural research advances are coalescing into neuro-symbolic approaches. The underlying research questions are how to best combine symbolic and neural approaches, what kind of symbolic/neural approaches are most suitable for which use case, and how to best integrate both ideas to advance the state of the art in information retrieval.

Materials are available online: **https://github.com/laura-dietz/neurosymbolic-representations-for-IR**

CCS CONCEPTS

• **Information systems → Information retrieval**.

KEYWORDS

neuro-symbolic representation, neural networks, document representation, knowledge graph, entities

ACM Reference Format:
Laura Dietz, Hannah Bast, Shubham Chatterjee, Jeff Dalton, Jian-Yun Nie, and Rodrigo Nogueira. 2023. Neuro-Symbolic Representations for Information Retrieval. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23), July 23–27, 2023, Taipei, Taiwan.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3539618.3594246

SIGIR '23, July 23–27, 2023, Taipei, Taiwan

ACM ISBN 978-1-4503-9408-6/23/07…$15.00
https://doi.org/10.1145/3539618.3594246

1 MOTIVATION

Being able to reason on what is relevant for an information need is important for all kinds of information retrieval tasks: web search, question answering, dialogues, image search, task assistance, or e-commerce. As traditional keyword-matching approaches are successively being replaced with neural-representation approaches [1, 2], the question is whether symbolic approaches still have merit and how it can be combined with neural approaches.

A decade ago, advances in knowledge graphs and semantic annotations, such as via entity linking, led to significant improvements in text ranking tasks [3, 4]. These, in turn, set new standards for entity-oriented downstream tasks like question answering [5, 6]. Now, neural representations for semantic annotations or other kinds of symbols have taken hold in the field of knowledge management. The information retrieval community is split into research that solely relies on neural representations (abandoning symbols altogether) and research that integrates neural and symbolic approaches.

Symbolic approaches have been studied in information retrieval over the years. The IR community has had a continued interest in entity retrieval tasks [7–9]. Sometimes, information needs are best answered using knowledge from external databases [10]—other times text can contextualize knowledge [11]. Furthermore, effective query expansion via pseudo-relevance feedback relies upon approaches that analyze retrieved documents—and reason about why these are relevant.

2 OBJECTIVES

The goal of this tutorial is to inspire more IR researchers to include symbolic approaches into their neural ranking algorithms. By consolidating findings and initiate a synergistic transfer across different IR-relevant use cases with respect to neuro-symbolic approaches.

Given the abundance of entity-bearing search queries asked on web search engines or entity-centric questions asked to chatbots, there is a clear potential for improvement. However, incorporating symbolic knowledge requires a few extra steps in the pipeline. We will educate the audience on those steps, how to best implement them, along with empirical results on what works and what doesn't.

3 FORMAT AND DETAILED SCHEDULE

In this full-day tutorial, we will provide different perspectives on neural-symbolic methods, provide different perspectives on the topic, and discuss customizations for different use cases.

Our goal is to provide useful information to a *wide variety* of audiences. The first part of the tutorial will be *introductory*, designed to bring audience members up to speed who have only basic knowledge in neural representations and/or symbolic approaches, such as knowledge graphs and entity linking.

The second part will of interest to both a *beginners and intermediate audience*, where different speakers provide their own perspective on the topic and look at different use cases where we need to reason on what is relevant.

To conclude the tutorial, we will invite all speakers and some additional guests to a panel discussion. Our goal is to spur a discussion of what works where, when, and why.

1. Knowledge Graphs and Entities. (2h)

Neural Text Representations and Semantic Annotations. BERT and other large neural language models (LLMs) of text have led to tremendous increases in performance improvements. LLM-based document re-ranking models are either based on Siamese-models like the Duet Model [12], or transformers, such as mono-BERT or duo-BERT [2]. We also cover neural query expansion [13] and query rewriting [14]. At their core, all these methods rely on metric learning abilities of neural networks, placing relevant documents close to queries, just like auto-regressive LLMs place the next word close to previous words.
The task of Entity Linking [15] (aka Wikification) is to annotate unstructured text with detected and disambiguated entity identifiers. Such entity links in the text can serve as a set of logical symbols to reason with. The entity links also provide a means to align the text with nodes in a knowledge graph to perform inferences in. Neural alignment methods allow the utilization of information in text and symbols for better ranking quality, such as EM-BERT [7].

Knowledge Graphs and Question Answering. Knowledge graphs provide background information that can be directly accessed and reasoned with. For example, a question can be given in arbitrary natural language, which should then be translated to a corresponding structured query (for example, in SPARQL) and executed on a given knowledge graph. The currently best approaches for solving this problem [6] are all inherently neuro-symbolic: the knowledge is given in strongly structured (symbolic) form, yet the learning is neural. Correspondingly, the challenges are twofold. The *symbolic* challenge is to understand the nature of the structured queries, which are often surprisingly complex and non-trivial—even for seemingly simple questions. The *neural* challenge is to learn a high-quality translation model that can handle also complex questions and requires little supervision. There has been impressive progress on this problem over the last decade, but the problem is still very far from being solved satisfactorily.

Ranking Wikipedia Entities/Aspects. Given a query and a knowledge graph, the entity ranking task is to retrieve and rank entities from the knowledge graph according to their relevance to the query. Entity ranking has also been shown to be useful for tasks that require an explicit semantic understanding of text [16]. Two broad directions for entity ranking are (1) Non-neural approaches that leverage symbols and semantic annotations in text, and (2) Neural approaches that leverage dense representations of entities learnt using neural networks. Finally, we discuss future directions for learning better entity representations for IR.

2. Neuro-Symbolic Foundations. (75 min)

Infusion of Symbolic Knowledge into Text Representation. Knowledge can be used to enhance text representations learned from raw texts. Related concepts in a knowledge graph can be incorporated into text representations so that one can match texts with similar concepts. Various approaches have been proposed to infuse knowledge into text representations. One can use graph neural networks (GNN) ([17, 18]) to train node representations and combine them with text representations with BERT [19], or to train a text representation jointly with knowledge encoding [20–22]. Another approach injects symbolic relations into texts before training text representations (e.g. K-BERT [23]). These knowledge infusing approaches have improved text representations for text matching and classification.

Reasoning about Relevance. Retrieval models aim to reason about what is relevant. Hence, we summarize related ideas from other areas, such as logic-based reasoning in knowledge graphs as well as Natural Language Inference (NLI). Some systems include retrieval into their neural inference models, such as REALM [24]. Following on research on probabilistic reasoning with neural approaches for logic-based reasoning in knowledge bases, such as FuzzQE [25]. Chain-of-Thought reasoners [26] are leveraging neural few-shot learners to generate well-reasoned arguments.

2. Practical Advantages for Ad Hoc Retrieval. (75 min)

Denoising Dense Representations with Symbols. Currently, there are three common ways to store pieces of information in modern search engines: As continuous vectors generated by dense retrievers, as symbols generated by sparse retrievers, or as knowledge graphs. All types of representations show great in-domain effectiveness. However, we will see that the performance of dense representations degrades quickly when applied on a different domain than originally trained for. This behavior is less than ideal for commercial search engines. We will discuss how symbols from knowledge graphs and sparse representations such as those from SPLADE [27], can help to reduce the noise of dense retrievers leading to more resilient approaches.

Explainability for Pseudo-relevance Feedback. Traditionally, pseudo-relevance feedback (PRF/RM3) is a technique to identify relevant terms for query expansion. This idea has been generalized to identify relevant entities for expanding queries with expansion entities [4], or augmenting neural representations [13].

In "Explainability" the focus shifts from making correct predictions to explaining why a prediction was made. One explainability approach is to analyze model gradients to approximate input importance [28]. In a PRF setting, such information can be used to glean information on why a document was deemed relevant, with the goal to augment and refine the search query.

4. Use Cases. (30 mins)

Use Case: Task-based Assistance. Information agents support complex real-world tasks and must not only retrieve relevant information, but also perform complex tasks using external symbolic tools and computation. This requires grounded reasoning about information and world state that is multimodal across text, images, video, and structured knowledge. Further, they must support the user in explainable and controllable fashion that involves eliciting structured information and storing it in personal knowledge graphs to incorporate structured symbolic constraints ("make it vegan") as well as being adaptable to mood, situation, and skill level.

Use Case: Generating Relevant Articles. Some usage scenarios ask for long, multi-faceted answers without the need for a user to interact. The goal is to foresee obvious next questions, and be forthcoming with such information without being explicitly asked. To satisfy this use case requires to solve a range of inter-dependent tasks: (1) high-recall retrieval with broad coverage, (2) query-specific clustering for subtopic-detection, (3) organization of content into a sequential structure, and (4) summarization and natural language generation.

4. Discussion Panel. (1h): The goal is to identify synergistic opportunities across different use cases. We are discussing approaches that (according to the literature) are supposed to work, but do not yet yield satisfactory results, leaving ample room for improvement. We also debate some controversial questions, such as "Since we have neural text representations, do we really need symbolic approaches?" The tutorial presenters and panel speakers are selected because they represent a broad spectrum of expert opinions on the topic.

4 PRESENTERS

Laura Dietz, Associate Professor, University of New Hampshire (main contact). Dr. Dietz focuses on integrating relevance-oriented tasks, using full-text search, Wikipedia knowledge, semantic annotations, along with organizing and generating content based on subtopics. She organized the KG4IR Tutorial and Workshop Series (ICTIR 2016, WSDM 2017, SIGIR 2017, SIGIR 2018) and the TREC Complex Answer Retrieval track (2017–2019).

Hannah Bast, Full Professor, University of Freiburg. Dr. Bast works on all aspects of information retrieval, with a focus on efficiency, ease of use, and fully functional systems. Her search systems power DBLP, Google Maps, and maybe soon Wikidata. Her work combines indexing and search in full text and structured knowledge for downstream applications such as question answering.

Shubham Chatterjee, Research Associate, University of Glasgow. Dr. Chatterjee focuses on neural entity-oriented information retrieval and extraction, particularly text understanding using entities and entity ranking. His goal is to build an intelligent search system that can answer open-ended information needs.

Jeff Dalton, Senior Lecturer, University of Glasgow. Dr. Dalton focuses on methods for effectively leveraging knowledge for complex information-seeking tasks. His work on entity-based query feature expansion published at SIGIR in 2014 is one of the first to demonstrate the effectiveness of using general-purpose knowledge graphs for search. He is a Turing AI Acceleration Fellow at the Turing Institute with a prestigious UKRI fellowship, and the lead organizer of the TREC Conversational Assistance track.

Jian-Yun Nie, Full Professor, University of Montreal. Dr. Nie is interested in all aspects of semantic matching based on reasoning and knowledge. He has worked on logical IR models and neural models based on knowledge graphs and applied them to various tasks such as medical IR. He has served for many conferences, including as SIGIR 2019 PC chair.

Dr. Nogueira, Consultant and Adjunct Professor, UNICAP University in Brazil. Dr. Nogeira will be speaking about his pioneering work on using Transformers for search, with many central publications such as monoT5, doc2query, and InPars, and is a co-author of the book "Pretrained Transformers for Text Ranking." After several years in academia, he is now consulting search engine companies.

ACKNOWLEDGMENTS

This work is supported in part by the Engineering and Physical Sciences Research Council grant EP/V025708/1 and in part by the National Science Foundation under Grant No. 1846017.

REFERENCES

[1] Bhaskar Mitra, Nick Craswell, et al. An introduction to neural information retrieval. *Foundations and Trends® in Information Retrieval*, 13(1):1–126, 2018.

[2] Jimmy Lin, Rodrigo Nogueira, and Andrew Yates. Pretrained transformers for text ranking: BERT and beyond. *CoRR*, abs/2010.06467, 2020.

[3] Ridho Reinanda, Edgar Meij, Maarten de Rijke, et al. Knowledge graphs: An information retrieval perspective. *Foundations and Trends® in Information Retrieval*, 14(4):289–444, 2020.

[4] Jeffrey Dalton, Laura Dietz, and James Allan. Entity query feature expansion using knowledge base links. In *Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '14, page 365–374, New York, NY, USA, 2014. Association for Computing Machinery.

[5] Emma J Gerritse, Faegheh Hasibi, and Arjen P de Vries. Graph-embedding empowered entity retrieval. In *Advances in Information Retrieval, Proceedings of the 42nd European Conference on Information Retrieval (ECIR 2020)*, Lecture Notes in Computer Science, pages 97–110, Cham, 2020. Springer.

[6] Hannah Bast and Elmar Haussmann. More accurate question answering on freebase. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 1431–1440, 2015.

[7] Emma J. Gerritse, Faegheh Hasibi, and Arjen P. de Vries. Entity-aware transformers for entity search. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 1455–1465, New York, NY, USA, 2022. Association for Computing Machinery.

[8] Shubham Chatterjee and Laura Dietz. BERT-ER: Query-Specific BERT Entity Representations for Entity Ranking. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 1466–1477, New York, NY, USA, 2022. Association for Computing Machinery.

[9] Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. Autoregressive entity retrieval. *CoRR*, abs/2010.00904, 2020.

[10] Chenyan Xiong, Zhengzhong Liu, Jamie Callan, and Eduard Hovy. Jointsem: Combining query entity linking and entity based document ranking. In *Proceedings of the 2017 ACM SIGIR Conference on Information and Knowledge Management*, CIKM '17, page 2391–2394, New York, NY, USA, 2017. Association for Computing Machinery.

[11] Marco Ponza, Diego Ceccarelli, Paolo Ferragina, Edgar Meij, and Sambhav Kothari. Contextualizing trending entities in news stories. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, pages 346–354, 2021.

[12] Bhaskar Mitra and Nick Craswell. An updated duet model for passage re-ranking. *arXiv preprint arXiv:1903.07666*, 2019.

[13] Ronak Pradeep, Rodrigo Nogueira, and Jimmy Lin. The expando-mono-duo design pattern for text ranking with pretrained sequence-to-sequence models. *arXiv e-prints*, pages arXiv–2101, 2021.

[14] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, 2020.

[15] Johannes M van Hulst, Faegheh Hasibi, Koen Dercksen, Krisztian Balog, and Arjen P de Vries. Rel: An entity linker standing on the shoulders of giants. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2197–2200, 2020.

[16] Shubham Chatterjee and Laura Dietz. Predicting Guiding Entities for Entity Aspect Linking. In *Proceedings of the 31st ACM International Conference on Information and Knowledge Management*, CIKM '22, New York, NY, USA, 2022. Association for Computing Machinery.

[17] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks, 2018.

[18] William L. Hamilton, Rex Ying, and Jure Leskovec. Inductive representation learning on large graphs, 2018.

[19] Chanwoo Jeong, Sion Jang, Hyuna Shin, Eunjeong Park, and Sungchul Choi. A context-aware citation recommendation model with bert and graph convolutional networks, 2019.

[20] Xiaozhi Wang, Tianyu Gao, Zhaocheng Zhu, Zhengyan Zhang, Zhiyuan Liu, Juanzi Li, and Jian Tang. KEPLER: A Unified Model for Knowledge Embedding and Pre-trained Language Representation. *Transactions of the Association for Computational Linguistics*, 9:176–194, 03 2021.

[21] Donghan Yu, Chenguang Zhu, Yiming Yang, and Michael Zeng. Jaket: Joint pre-training of knowledge graph and language understanding. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):11630–11638, Jun. 2022.

[22] Zhibin Lu, Pan Du, and Jian-Yun Nie. Vgcn-bert: Augmenting bert with graph embedding for text classification. In Joemon M. Jose, Emine Yilmaz, João Magalhães, Pablo Castells, Nicola Ferro, Mário J. Silva, and Flávio Martins, editors, *Advances in Information Retrieval*, pages 369–382, Cham, 2020. Springer International Publishing.

[23] Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, and Ping Wang. K-bert: Enabling language representation with knowledge graph. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(03):2901–2908, Apr. 2020.

[24] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. Realm: retrieval-augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning*, pages 3929–3938, 2020.

[25] Xuelu Chen, Ziniu Hu, and Yizhou Sun. Fuzzy logic based logical query answering on knowledge graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 3939–3948, 2022.

[26] Abulhair Saparov and He He. Language models are greedy reasoners: A systematic formal analysis of chain-of-thought. *arXiv preprint arXiv:2210.01240*, 2022.

[27] Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. Splade: Sparse lexical and expansion model for first stage ranking. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2288–2292, 2021.

[28] Thorben Funke, Megha Khosla, Mandeep Rathee, and Avishek Anand. Zorro: Valid, sparse, and stable explanations in graph neural networks. *IEEE Transactions on Knowledge and Data Engineering*, 2022.