Overview of The First Workshop on Knowledge Graphs and Semantics for Text Retrieval and Analysis (KG4IR)

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

Chenyan Xiong
Carnegie Mellon University
Pittsburgh, PA, USA
cx@cs.cmu.edu

Edgar Meij
Bloomberg L.P.
London, United Kingdom
edgar.meij@acm.org

Abstract

Knowledge graphs have been used throughout the history of information retrieval for a variety of tasks. Advances in knowledge acquisition and alignment technology in the last few years have given rise to a body of new approaches for utilizing knowledge graphs in text retrieval tasks. This report presents the motivation, output, and outlook of the first workshop on Knowledge Graphs and Semantics for Text Retrieval and Analysis which was co-located with SIGIR 2017 in Tokyo, Japan. We aim to assess where we stand today, what future directions are, and which preconditions could lead to further performance increases.

1 Introduction

The past decade has witnessed the emergence of publicly available knowledge graphs (KGs) such as DBpedia, Freebase, and WikiData and also proprietary KGs such as Google's Knowledge Graph and Microsoft's Satori. The availability of large knowledge graphs and semantic annotation techniques gave rise to successful approaches for many information retrieval (IR) tasks. It has been shown that heterogeneous information in knowledge graphs and entity annotations can help to significantly improve the performance of information retrieval tasks. In particular, the semantics encoded in knowledge graphs have been effectively integrated in various aspects of IR systems, including query representation (Meij et al., 2012; Xiong and Callan, 2015b; Hasibi et al., 2015), retrieval models (Liu and Fang, 2015; Dalton et al., 2014; Raviv et al., 2016), learning-to-rank (Xiong and Callan, 2015a), and document/result (re)presentations (Voskarides et al., 2017; Raviv et al., 2016).

At the KG4IR workshop,[1] researchers from different fields came together to discuss how to improve the end-to-end utilization of knowledge graphs and semantics in text retrieval and IR-related downstream applications. The scope included the **acquisition**, **alignment**, and **utilization** of knowledge graphs and semantic resources for the purpose of optimizing end-to-end system performance, with a focus on information retrieval and text analysis applications.

Acquisition includes knowledge graph population and semantic resource construction with a special focus on enabling IR-related techniques and applications. Examples include domain/task-specific knowledge graph construction, knowledge representation, and query-time knowledge extraction.

Alignment includes the semantic annotation process such as entity linking of short keyword queries or relation extraction for satisfying information needs. It also includes information integration, ontology matching, entity search, and knowledge graph selection based on an information need.

Utilization includes the use of knowledge graphs and semantics in text-centric tasks. Examples are utilizing the knowledge graph to improve document retrieval, question answering, factoid search, dialogue systems, event tracking, and retrieval of complex answers.

The workshop featured discussions and presentations on innovative ideas for new methods, suggestions for shared tasks and benchmarks, position papers, as well as reports on practical experiences with knowledge graph technology in academia and industry.

2 Keynotes

We invited 6 keynote speakers from both academia and industry and interspersed their presentations with those selected through peer review. Every keynote speaker provided their unique angle on the theme of the workshop, allowing for a rich and diverse program.

Bogdan Arsintescu, LinkedIn: 14 True Facts About Knowledge Graphs

Bogdan spoke about how LinkedIn manages their large and inherently graph-structured data, including their LinkedIn Economic Graph. In all their data, entities are first-class citizens. The system's heavy reliance on edges renders relational databases not ideal, and their nodes identities makes it difficult to apply standard IR indexing technology. Instead, tasks are addressed with graph databases, which are fundamentally different—yet sometimes graph databases would benefit from a better integration with IR techniques.

[1] See http://kg4ir.github.io and https://groups.google.com/forum/#!forum/kg4ir.

Jun Xu, Chinese Academy of Sciences: Deep Approaches to Semantic Text Matching

In semantic matching the task is to predict relevance or similarity given two text documents. This task arises in many text applications such as paraphrase identification, information retrieval, and question answering. Focusing on neural network methods for semantic matching, Jun takes a new perspective on word-level matching and sentence-level matching. In both cases he applies word representations, proximity and multi-word patterns to bridging the semantic gap.

Jeff Dalton, University of Glasgow: From Facts to Acts: Knowledge Graphs for Personal Assistant

A lot of work on question answering finds their motivation in personal assistants. Jeff discusses similarities and differences between factoid question answering and personal assistants even though both approaches utilize knowledge graphs and react to user input. However, while question answering focuses on explicit factual knowledge, personal assistants are volatile and need to incorporate knowledge about the user, such as their state, what they know, and their level of expertise. Given the maturity of the question answering field, Jeff discusses necessary steps to translate the research findings from question answering to personal assistants.

David Carmel, Yahoo! Research: What People are Asking About You: Mining Entity Search Intents in CQA Sites

Users repeatedly ask the same questions on community question answering sites. Past work developed approaches that extract typical questions that are asked about a given entity. These are called entity search intents. David presents his work on utilizing entity search intents to derive a measure of entity relatedness. The work is based on the assumption that people ask similar questions about strongly related entities.

Ian Soboroff, NIST: Overview of and Lessons from the FEIII Challenge

FEIII (pronounced: "Eff eeh triple I") is a community challenge on extraction and linking of entities and relations in the financial domain.[2] The challenge focuses on semantic annotations in publicly available company filings. The challenge entails several tasks such as entity tagging, entity alignment across structured databases, and relation extraction. The challenge is led by Louiqa Raschid from University of Maryland. It was motivated by the US Department of the Treasury with the ultimate goal of preventing financial disasters such as in previous sub-prime mortgage crisis, through the development of automated early warning systems.

Hannah Bast, Universität Freiburg: Semantic Search on Text and Knowledge Bases

Hannah talks about how to craft systems that support search with meaning and gives an overview over her recent survey (Bast et al., 2016). While knowledge graphs are the preferred way of storing structured knowledge, she believes that most of the world's knowledge will continue to be in text form. To support seamless search over both data sources, she presents two approaches: a system that supports a user in incrementally building a complex semi-structured query (such as the search in DBLP) and a system that allows free-form natural queries that are interpreted with respect to text and structured knowledge.

[2]See `https://ir.nist.gov/feiii`.

3 Contributed Papers

The wealth of contributed papers were actively discussed at the workshop. All accepted contributed papers are available in the proceedings (Dietz et al., 2017).[3]

Acquisition of knowledge graphs

- Blanco, Joho, Jatowt, and Yu introduce a test collection for constructing knowledge graphs on actions that can be applied to certain entities.
- Rastogi and Durme explain how to complete knowledge bases with transitive, yet asymmetric relations.

Alignment with knowledge graphs

- Rastogi, Poliak, and Van Durme train a relation embedding from a knowledge graph with consistency constraints.
- Ali, Caputo, and Lawless propose a method for identifying the most salient attributes for entities, using training data from Wikipedia infoboxes.
- Li, Xiong, and Callan suggest an approach for annotating questions with matching relations and annotating relations with keywords for sentence retrieval.

Utilizing knowledge graphs

- Saleiro, Milic-Frayling, Mendes Rodrigues, and Soares propose a method for retrieving related entity sets that match a question.
- Gupta, Radhakrishnan, Gupta, Varma, and Gupta explain how to categorize scientific papers by learning which entities are representative for each class.
- He and Bron predict which domains users are knowledgeable in by analyzing their search logs with knowledge graphs.

4 Panel Discussions

We merged with the workshop on Open Knowledge Base and Question Answering (OKBQA) for our panel discussion at the end of the workshop which included the following panelists.

- Hannah Bast, Universität Freiburg, Germany
- Noriko Kando, National Institute of Informatics, Japan
- Jaap Kamps, University of Amsterdam, The Netherlands
- Edgar Meij, Bloomberg L.P., U.K.
- Bogdan Arsintescu, LinkedIn, U.S.A
- David Carmel, Yahoo!, Israel

[3] `http://ceur-ws.org/Vol-1883/`

The panel speakers discussed opportunities for future improvements on knowledge graphs and alignment techniques for the purposes of IR and QA tasks. Topics ranged from definitions of an “entity” to how provenance and conflicts can be effectively handled and how “alternate facts” can be dealt with. On a final note, the panel speakers agreed that despite the advances in deep learning, knowledge graph technology are unlikely to become obsolete anytime soon.

5 Conclusion

The workshop generated a lot of interest, with a room full of researchers with expertises ranging from question answering to graph databases discussing emerging trends and opportunities for knowledge graphs and their underlying technology. The overarching question was how knowledge graphs can be most effectively utilized. This discussion included a wide range of basic alignment techniques which are capable of semantically annotating questions, extracting knowledge triples, and learning embeddings from knowledge graphs. Techniques for identifying relevant elements in the knowledge graph included the retrieval of relevant entity pairs, identification of salient attributes for entities, and entity relatedness measures. Many uses of knowledge graphs benefit from combining unstructured text and structured knowledge, for instance through entity linking, semantic matching, as well as leveraging queries that explicitly include both modalities. The presented applications were plentiful, ranging from categorization over conversational search agents to inference of user expertises.

6 Acknowledgements

This work was supported in part by National Science Foundation (NSF) grant IIS-1422676. Any opinions, findings, and conclusions in this paper are the author’s and do not necessarily reflect those of the sponsors. The peer review process was supported by EasyChair and our online proceedings are provided by CEUR-WS. We thank our PC members for their excellent reviews and last but not least, we wholeheartedly thank all workshop attendees and speakers for the many interesting discussions!

References

Esraa Ali, Annalina Caputo, and Séamus Lawless. 2017. Entity Attribute Ranking Using Learning to Rank. In *KG4IR@SIGIR (CEUR Workshop Proceedings)*, Vol. 1883. CEUR-WS.org, 19–24.

Hannah Bast, Björn Buchhold, Elmar Haussmann, and others. 2016. Semantic Search on Text and Knowledge Bases. *Foundations and Trends® in Information Retrieval* 10, 2-3 (2016), 119–271.

Roi Blanco, Hideo Joho, Adam Jatowt, and Haitao Yu. 2017. Test Collection for Evaluating Actionable Knowledge Graphs. In *KG4IR@SIGIR (CEUR Workshop Proceedings)*, Vol. 1883. CEUR-WS.org, 32–37.

Jeffrey Dalton, Laura Dietz, and James Allan. 2014. Entity Query Feature Expansion using Knowledge Base Links. In *Proceedings SIGIR 2014*. ACM, 365–374.

Laura Dietz, Chenyan Xiong, and Edgar Meij (Eds.). 2017. *Proceedings of the First Workshop on Knowledge Graphs and Semantics for Text Retrieval and Analysis (KG4IR 2017) co-located with the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2017), Shinjuku, Tokyo, Japan, August 11, 2017*. CEUR Workshop Proceedings, Vol. 1883. CEUR-WS.org. `http://ceur-ws.org/Vol-1883`

Shashank Gupta, Priya Radhakrishnan, Manish Gupta, Vasudeva Varma, and Umang Gupta. 2017. Enhancing Categorization of Computer Science Research Papers using Knowledge Bases. In *KG4IR@SIGIR (CEUR Workshop Proceedings)*, Vol. 1883. CEUR-WS.org, 38–42.

Faegheh Hasibi, Krisztian Balog, and Svein Erik Bratsberg. 2015. Entity Linking in Queries: Tasks and Evaluation. In *Proceedings of ICTIR 2015*. ACM, 171–180.

Jiyin He and Marc Bron. 2017. Measuring Demonstrated Potential Domain Knowledge with Knowledge Graphs. In *KG4IR@SIGIR (CEUR Workshop Proceedings)*, Vol. 1883. CEUR-WS.org, 13–18.

Hongyu Li, Chenyan Xiong, and Jamie Callan. 2017. Natural Language Supported Relation Matching for Question Answering with Knowledge Graphs. In *KG4IR@SIGIR (CEUR Workshop Proceedings)*, Vol. 1883. CEUR-WS.org, 43–48.

Xitong Liu and Hui Fang. 2015. Latent entity space: A novel retrieval approach for entity-bearing queries. *Information Retrieval Journal* 18, 6 (2015), 473–503.

Edgar Meij, Wouter Weerkamp, and Maarten de Rijke. 2012. Adding Semantics to Microblog Posts. In *WSDM*.

Pushpendre Rastogi and Benjamin Van Durme. 2017. Predicting Asymmetric Transitive Relations in Knowledge Bases. In *KG4IR@SIGIR (CEUR Workshop Proceedings)*, Vol. 1883. CEUR-WS.org, 7–12.

Pushpendre Rastogi, Adam Poliak, and Benjamin Van Durme. 2017. Training Relation Embeddings under Logical Constraints. In *KG4IR@SIGIR (CEUR Workshop Proceedings)*, Vol. 1883. CEUR-WS.org, 25–31.

Hadas Raviv, Oren Kurland, and David Carmel. 2016. Document retrieval using entity-based language models. In *Proceedings of SIGIR 2016*. ACM, 65–74.

Pedro Saleiro, Natasa Milic-Frayling, Eduarda Mendes Rodrigues, and Carlos Soares. 2017. Early Fusion Strategy for Entity-Relationship Retrieval. In *KG4IR@SIGIR (CEUR Workshop Proceedings)*, Vol. 1883. CEUR-WS.org, 49–54.

Nikos Voskarides, Edgar Meij, and Maarten de Rijke. 2017. Generating descriptions of entity relationships. In *ECIR 2017: 39th European Conference on Information Retrieval (LNCS)*. Springer.

Chenyan Xiong and Jamie Callan. 2015a. EsdRank: Connecting query and documents through external semi-structured data. In *Proceedings of CIKM 2015*. ACM, 951–960.

Chenyan Xiong and Jamie Callan. 2015b. Query expansion with Freebase. In *Proceedings of ICTIR 2015*. ACM, 111–120.