Learning-to-Rank Target Types for Entity-Bearing Queries

Darío Garigliotti
University of Stavanger
dario.garigliotti@uis.no

Krisztian Balog
University of Stavanger
krisztian.balog@uis.no

This paper revisits the learning-to-rank approach we proposed for automatically identifying the target entity types of queries [6]. After presenting our contributions and results, we draw on the learned lessons and encountered challenges to identify directions for future enhancements.

A significant portion of information needs in web search target entities [11]. Entities, such as people, organizations, or locations are natural units for organizing information and for providing direct answers. A characteristic property of entities is that they are typed, where types are typically organized in a hierarchical structure, i.e., a *type taxonomy*. Previous work has shown that entity retrieval performance can be significantly improved when a query is complemented with explicit *target type* information, see, e.g., [1, 8, 10]. Recently, Garigliotti and Balog [5] have conducted a systematic evaluation of dimensions of type information. Identifying and exploiting target type information falls within the broad area of *query understanding*, which, according to Croft et al. [4], refers to process of "identifying the underlying intent of the queries, based on a particular representation." In a realistic Web search scenario, automatically detected target types avoid the possible cognitive effort of the user to provide type information for her query. Furthermore, they can be used as facets, for filtering the results. Motivated by the above reasons, our main objective is to generate target type annotations of queries automatically.

Following the *hierarchical target type identification* task proposed in [2], we wish to identify the most specific target types for a query, from a given type taxonomy, such that they are sufficient to cover all relevant results. We introduced a relaxation to the task definition, by allowing for a query to have multiple target types (or none).

One main contribution of this work is a test collection we built for the revised hierarchical target type identification task. We used the DBpedia ontology (version 2015-10) as our type taxonomy and collect relevance labels via crowdsourcing for the 485 queries in the entity ranking DBpedia-Entity collection [3]. We noted that none of the elements of our approach are specific to this taxonomy, and our methods could be applied on top of any type taxonomy.

As our second main contribution, we developed a learning-to-rank (LTR) approach with a rich set of features, including term-based, linguistic, and distributional similarity, as well as taxonomic features. We compared our LTR method versus two competitive baselines from the literature. One approach is an entity-centric model [2, 9, 12], which firstly ranks the entities based on their relevance to the query, then look at what types the top ranked entities have. Alternatively, a type-centric model presented in [2] ranks direct term-based representations (pseudo type description documents), built for each type, by aggregating descriptions of entities of that type. These baseline models also fit, respectively, the late fusion and early fusion design patterns for object retrieval [13]. Our experiments were performed using Nordlys [7], a toolkit for entity-oriented and semantic search. We employed the Random Forest algorithm for regression as the supervised ranking method. We found that our supervised learning approach significantly and substantially outperforms all baseline methods. Also, an analysis of the discriminative power of our features was performed, by sorting them according to their information gain. This showed the effectiveness of textual similarity features, enriched with distributional semantic representations, measured between the query and the type label. Furthermore, we observed the robustness of our proposed method, as it succeeds in automatically detecting target types for a wide variety of queries.

In the forward-looking spirit of the workshop, we identify three main themes for future development. Some of these are closely tied to the target type identification task, while others concern learning-based approaches in general. A first challenge is the generalization of the proposed approach to other type systems. So far, the suitability of our learning-to-rank target type detector was demonstrated only using training data manually labeled with target types from the DBpedia Ontology. It remains an open question whether these particular features are applicable to other type systems. Prior to that, an issue to be addressed is how to ease the acquisition of type labels for training. While manual annotation can be performed via crowdsourcing, with very large type systems it becomes practically unfeasible (selecting a single/handful of target type(s) from several hundred options would create cognitive overload). One possible strategy that we can imagine would be an instance of knowledge transfer from the available test collection of DBpedia target types.

As part of a more general line of discussion, another item to be mentioned has to do with how training data is obtained. It is still to be answered whether more training data helps. And, if that is the case, the acquisition of high-quality labeled data becomes a bottleneck. Automatically obtaining target labels by weak supervision, thereby avoiding human annotation effort, may be a plausible way to alleviate this challenge.

Finally, we draw our attention to more recent developments with an increasing impact in many problem domains within information retrieval. Specifically, using features obtained by representation learning via deep artificial neural networks. Our approach exploits semantic similarities purely based on pre-trained word embeddings. What we do here, using distinguished neural features in a LTR approach, has become a noticeable trend in recent years. The question that arises is whether it is possible to embrace an alternative, fully neural learning approach. We emphasize that this question applies beyond our specific task of target types identification. Indeed, it likely extends to a whole range of tasks where the current state of the art is constituted by a learning-to-rank approach, using a manually engineered set of features.

LEARNER'17, October 1, 2017, Amsterdam, The Netherlands

REFERENCES

[1] Krisztian Balog, Marc Bron, and Maarten De Rijke. 2011. Query Modeling for Entity Search Based on Terms, Categories, and Examples. *ACM Trans. Inf. Syst.* 29, 4 (2011), 22:1–22:31.

[2] Krisztian Balog and Robert Neumayer. 2012. Hierarchical Target Type Identification for Entity-oriented Queries. In *Proc. of CIKM*. 2391–2394.

[3] Krisztian Balog and Robert Neumayer. 2013. A Test Collection for Entity Search in DBpedia. In *Proc. of SIGIR*. 737–740.

[4] W Bruce Croft, Michael Bendersky, Hang Li, and Gu Xu. 2010. Query Representation and Understanding Workshop. In *SIGIR Forum*. 48–53.

[5] Darío Garigliotti and Krisztian Balog. 2017. On Type-Aware Entity Retrieval. In *Proc. of ICTIR*. 27–34.

[6] Darío Garigliotti, Faegheh Hasibi, and Krisztian Balog. 2017. Target Type Identification for Entity-Bearing Queries. In *Proc. of SIGIR*. 845–848.

[7] Faegheh Hasibi, Krisztian Balog, Darío Garigliotti, and Shuo Zhang. 2017. Nordlys: A Toolkit for Entity-Oriented and Semantic Search. In *Proc. of SIGIR*. 1289–1292.

[8] Rianne Kaptein and Jaap Kamps. 2013. Exploiting the Category Structure of Wikipedia for Entity Ranking. *Artificial Intelligence* 194 (2013), 111–129.

[9] Rianne Kaptein, Pavel Serdyukov, Arjen P. De Vries, and Jaap Kamps. 2010. Entity Ranking Using Wikipedia as a Pivot. In *Proc. of CIKM*. 69–78.

[10] Jovan Pehcevski, James A Thom, Anne-Marie Vercoustre, and Vladimir Naumovski. 2010. Entity Ranking in Wikipedia: Utilising Categories, Links and Topic Difficulty Prediction. *Information Retrieval* 13, 5 (2010), 568–600.

[11] Jeffrey Pound, Peter Mika, and Hugo Zaragoza. 2010. Ad-hoc Object Retrieval in the Web of Data. In *Proc. of WWW*. 771–780.

[12] David Vallet and Hugo Zaragoza. 2008. Inferring the Most Important Types of a Query: a Semantic Approach. In *Proc. of SIGIR*. 857–858.

[13] Shuo Zhang and Krisztian Balog. Design Patterns for Fusion-Based Object Retrieval. In *Proc. of ECIR*. 684–690.