Learning Semantic Query Suggestions*

Edgar Meij
University of Amsterdam
Science Park 107
Amsterdam, The Netherlands
edgar.meij@uva.nl

Marc Bron
University of Amsterdam
Science Park 107
Amsterdam, The Netherlands
m.m.bron@uva.nl

Laura Hollink
VU University Amsterdam
de Boelelaan 1081a
Amsterdam, The Netherlands
hollink@cs.vu.nl

Bouke Huurnink
University of Amsterdam
Science Park 107
Amsterdam, The Netherlands
bhuurnink@uva.nl

Maarten de Rijke
University of Amsterdam
Science Park 107
Amsterdam, The Netherlands
derijke@uva.nl

* Human-defined concepts are fundamental building blocks of the semantic web. When used as annotations for documents or text fragments they can provide explicit anchoring in background knowledge and enable intelligent search and browsing facilities. As such, an important application of ontological knowledge is augmenting unstructured text with links to relevant, human-defined concepts. For the author or reader of the text, this augmentation may supply useful pointers, for example to the concepts themselves or to other concepts related to the ones found. For ontology learning applications, such links may be used to learn new concepts or relations between them [11]. Recently, data-driven methods have been proposed to generate links between phrases appearing in full-text documents and a set of ontological concepts known a priori. Mihalcea and Csomai [8] propose the use of several linguistic features in a machine learning framework to link phrases in full-text documents to Wikipedia articles and this approach is further improved upon by Milne and Witten [9]. Because of the connection between Wikipedia and DBpedia, such data-driven linking methods help us establish links between textual documents and Linked Open Data [1, 2, 4, 10].

Another, more challenging instantiation of linking text to human-defined concepts in a knowledge source is *semantic query suggestion*. Query suggestion is a strategy to derive terms that are able to return more relevant results than the initial query. Commonly used approaches to query suggestion (sometimes referred to as a form of query expansion) are highly data-driven and based mostly on term frequencies [5, Chapter 9]. *Semantic* query suggestion, in contrast, tries to understand (or learn) which concepts the user used in her query or, phrased alternatively, the concepts she is interested in and wants to find.[1] Moreover, the properties of each concept, and any other resources associated with it, could serve as additional, useful information for the user. In our current work, we use DBpedia as our target ontology. As an example of our task, consider the query "obama white house". A semantic query suggestion algorithm should return suggestions in the form of the (DBpedia) instances labeled "Barack Obama" and "White House". Identifying such semantic suggestions serves multiple purposes: it can (i) help the user acquire contextual information, (ii) suggest related concepts or associated terms that may be used for search, and (iii) provide valuable navigational suggestions.

In this paper we address the semantic query suggestion task and automatically link queries to DBpedia concepts. The specific task we address in this paper is the following. Given a query that is submitted to a search engine, identify the relevant concepts that the user entered in her query where the concepts are taken from an existing knowledge base or ontology. We address our task in the setting of a digital archive, specifically of the Netherlands Institute for Sound and Vision ("Sound and Vision"). Sound and Vision maintains a large digital audiovisual collection, currently containing over a million objects and daily updated with new broadcasts.

Our approach to suggesting DBpedia concepts for user queries consists of two stages. In the first stage, a ranked list of possible concepts for the query is generated using a language modeling framework for each full query and for each n-gram (i.e., contiguous sequence of n words) in the query. We use various textual representations of each DBpedia concept, including the Wikipedia article text, its label, and the text used in the hyperlinks pointing to it.

Once we have obtained a ranked list of possible concepts for each n-gram in the query, we turn to concept selection. In this stage we need to decide which of the candidate concepts are most viable. We use a supervised machine learning approach, which takes as input a set of labeled examples (query to concept mappings) and several features of these examples. We choose to compare a Naive Bayes (NB) classifier, with a Support Vector Machine (SVM) classifier and a decision tree classifier (J48)—a set representative of the state-of-the-art in classification. We experiment with multiple classifiers in order to confirm that our results are generally valid, i.e., not dependent on any machine learning algorithm. In order to

*This an extended abstract of Meij et al. [7].

[1] We use "ontology" to refer to the full spectrum of conceptualizations, ranging from glossaries to formal ontologies [6]. We refer to an instance in DBpedia as "concept" [10].

train the machine learning algorithms, we examined close to 1000 queries from a search engine for a digital multimedia archive and manually linked over 600 of these to relevant concepts in DBpedia. We employ several types of features, each associated with either the current query n-gram, the current concept, their combination, or the session history.
We define semantic query suggestion as a ranking problem; the system has to return five concepts for a given input query and the assessments described above are used to determine the relevance status of these concepts. We employ several measures which are well-known in the field of information retrieval [3].
Using Support Vector Machines and features extracted from the full input queries yields optimal results. The best performing run is able to locate almost 90% of the relevant concepts on average. Moreover, this particular run achieves a precision@1 of 89% which means that for this percentage of queries a relevant concept is returned as the first suggestion. In sum, we have shown that the semantic query suggestion problem can be successfully cast as a ranking problem. The best way of handling query terms is not as separate n-grams, but as a single unit—a finding also interesting from an efficiency viewpoint, since the number of n-grams is quadratic with respect to the length of the query. All types of feature were found to be helpful and, besides document and term features, we found that concept features were also important in achieving our best performance.

Acknowledgments

This research was carried out in the context of the Virtual Laboratory for e-Science project and supported by the DuOMAn project carried out within the STEVIN programme which is funded by the Dutch and Flemish Governments under project number STE-09-12 and the DAESO project carried out within the STEVIN programme which is funded by the Dutch and Flemish Governments (http://www.stevin-tst.org) under project number STE-05-2 and the Netherlands Organisation for Scientific Research (NWO) under project numbers 017.001.190, 640.001.501, 640.002.501, 612.-066.512, 612.061.814, 612.061.815, 640.004.802.

References

[1] S. Auer and J. Lehmann. What have Innsbruck and Leipzig in common? Extracting semantics from wiki content. In *The Semantic Web: Research and Applications*, 2007.

[2] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives. Dbpedia: A nucleus for a web of open data. *ISWC '07*, 2007.

[3] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison Wesley, 1999.

[4] C. Bizer, R. Cyganiak, S. Auer, and G. Kobilarov. DBpedia–querying Wikipedia like a database. In *WWW '07*, 2007.

[5] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[6] D. L. Mcguinness. Ontologies come of age. In D. Fensel, J. Hendler, H. Lieberman, and W. Wahlster, editors, *Spinning the Semantic Web: Bringing the World Wide Web to Its Full Potential*. MIT Press, 2003.

[7] E. J. Meij, M. Bron, B. Huurnink, L. Hollink, and M. de Rijke. Learning semantic query suggestions. In *8th International Semantic Web Conference (ISWC 2009)*, 2009.

[8] R. Mihalcea and A. Csomai. Wikify!: Linking documents to encyclopedic knowledge. In *CIKM '07*, 2007.

[9] D. Milne and I. H. Witten. Learning to link with wikipedia. In *CIKM '08*, 2008.

[10] F. M. Suchanek, G. Kasneci, and G. Weikum. Yago: a core of semantic knowledge. In *WWW '07*, 2007.

[11] W. R. van Hage, M. de Rijke, and M. Marx. Information retrieval support for ontology construction and use. In *ISWC '04*, 2004.