Exposé for Student Research Project

Entity Linking - A Survey of Recent Approaches

Torsten Huber[1]

March 30, 2012

supervised by:
Prof. Dr. Ulf Leser[1]
Prof. Dr. Hans Uszkoreit[2]
Dipl. Inf. Peter Adolphs[2]

[1] Department of Computer Science, Humboldt-Universität zu Berlin [2] DFKI GmbH, project office Berlin

1 Problem Definition

Entity linking describes the task of matching references to named entities found in natural language texts to a unique identifier, denoting a specific entity (Larson, 2010). Typically the identifier is part of a knowledge base, which provides the background knowledge for the task. An entity mention may be ambiguous, i.e. it can refer to different entities which share the same name. Consider the following text from a newspaper article:

> "*David Cameron will today tell Angela Merkel, German chancellor, that he will back her plans to strengthen economic union in the eurozone, but only on condition that he wins safeguards to protect the City of London from unwelcome European legislation.*"

The ambiguous term `David Cameron` may refer to David Cameron, the British politician or David Cameron, the English actor. An entity linking system must use available information, such as the context of the entity mention or information from the knowledge base, to decide which entity is being referred to in the text. This task becomes increasingly difficult when an entity is not referred to by its full name, but by partial names, acronyms or other name variations (e.g. D. Cameron, David W. D. Cameron, Cameron, D. C., etc.).

2 Motivation

The entity linking problem can be considered to be a part of the named entity recognition (NER) and identification task. An NER Tagger typically finds mentions of named entities and asserts a type to them (e.g. `PERSON`, `ORGANISATION`, etc.). Determining a unique database identifier provides access to more structured contextual information, which is useful in several information extraction applications (Larson, 2010, p. 217). For example, it can be used to enrich texts with semantic information, as proposed by the Semantic Web community, to provide useful meta information about unstructured texts. One application is the automatic link generation for entity references in news articles. For example, Mihalcea and Csomai (2007) created a software to parse new Wikipedia articles and automatically create tags and links to other articles about relevant entities. Furthermore it can be used in e-mail clients to process messages and identify references to people in the contact list, current tasks or upcoming events in the calendar.

Linking entities is also necessary for most knowledge discovery tasks focusing on real-life entities, such as companies. For example, to monitor events like merges of companies or new product releases as done by Saggion et al. (2007), a system must be able to accurately identify references to companies, even if they are being referred to by highly ambiguous acronyms like ABC with roughly 100 different word senses in the English Wikipedia [1]. Moreover entity linking can be useful in automating corporate customer care, like complaint filing systems (Chakaravarthy et al., 2006). It can be utilized to process inquiries or complaints and identify products or orders by examining the information provided by the customer (like product names, ids, order numbers, etc.). The message can then be automatically routed to the according support staff member.

3 State of Research

Much research has been performed in the field of entity linking. Due to the diversity of applications, the actual task and available background information for entity disambiguation often varies. However, there are two common basic approaches to this problem. First is the use of similarity measures to other texts with unstructured knowledge about entities, such as Wikipedia articles. The second approach utilizes semi-structured data such as YAGO or Freebase for disambiguation.

Bunescu and Pasca (2006) use cosine similarity to rank candidate entities based on the relatedness of the context of an entity mention to a Wikipedia article. Cucerzan (2007) utilized

[1] http://en.wikipedia.org/wiki/Abc, accessed 13th February 2012

Wikipedia articles, disambiguation pages, redirects and categories to extract semantically enriched data and compare it to the text with a vector-based comparison model. In a similar fashion Han and Zhao (2009) parsed the Wikipedia to extract a concept graph, measuring the similarity by means of the distance of co-occurring terms to candidate concepts.

A notable contribution to research in this field was made by the participants of the Text Analytics Conference (TAC). In 2009 and 2010 entity linking was part of the Knowledge Base Population (KBP) track. Given a rudimentary knowledge base extracted from Wikipedia infoboxes and access to Wikipedia articles, the participants were requested to develop a system which can reliably link mentions of persons, organizations and geopolitical entities to their respective knowledge base identifiers. Several different approaches have been implemented and tested, ranging from simple matches of noun phrases and knowledge base nodes to support-vector-machine-based learning algorithms (Chang et al., 2010; Lehmann et al., 2010; Zhang et al., 2010a).

4 Goals

The goal of this research project is to give a detailed overview of the current state of research about entity linking. A survey of existing approaches and techniques shall be conducted to identify common approaches and solutions to the problem. The focus will be on disambiguating references to persons, organizations and geopolitical entities. The system shall be able to disambiguate references to entities of these types, regardless of the label that is being used to refer to it in the text.

As a part of this study, an implementation of an entity linking system shall be developed. Next to a simple baseline algorithm for comparison purposes, one of the existing approaches shall be re-implemented and examined in more detail. In order to test the algorithm, an evaluation setup has to be devised. This step includes finding a suitable evaluation corpus, as described in section 4.1.

The third goal is to perform a brief error analysis in order to identify the most common problems of linking entities to database identifiers. The data collected in the experiments and the subsequent error analysis shall serve as a basis for further research.

4.1 Evaluation

A preliminary examination of the different options for performing an evaluation has shown that there is a lack of easily accessible evaluation corpora suitable for the entity linking task.

The most commonly used corpus from the TAC 2009 KBP challenge is not publicly available. In order to use the latest TAC corpus, participation in the TAC 2012 KBP challenge would be required.

As a second option, the Web People Search (WePS)[2] data can be used for evaluation. This task focuses on clustering documents with references to specific people and is hence a subtask of the general entity linking task. Nevertheless, the freely available data can be used for the evaluation of the more general task proposed for this work.

The third alternative is to use the Wikipedia for evaluation, as done by Cucerzan (2007). Every Wikipedia article contains links to other articles which are relevant for the respective topic. These links usually refer to a specific entity rather than a disambiguation page and hence provide useful information for the entity linking task. An evaluation setup for an entity linking system would be to remove all links from an article text and let the system re-generate them. These generated links can be compared to the original links chosen by the editors of the article.

References

Bunescu, R. and Pasca, M. (2006). Using encyclopedic knowledge for named entity disambiguation. In *Proceesings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL-06)*, pages 9–16, Trento, Italy.

Chakaravarthy, V. T., Gupta, H., Roy, P., and Mohania, M. (2006). Efficiently linking text documents with relevant structured information. In *VLDB '06: Proceedings of the 32nd international conference on Very large data bases*, pages 667–678. VLDB Endowment.

Chang, A. X., Spitkovsky, V. I., Yeh, E., Agirre, E., and Manning, C. D. (2010). Stanford-UBC entity linking at TAC-KBP. In *Proceedings of the Third Text Analysis Conference (TAC 2010)*, Gaithersburg, Maryland, USA.

Cucerzan, S. (2007). Large-scale named entity disambiguation based on wikipedia data. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Prague, Czech Republic, June 28-30, 2007*, pages 708–716.

Han, X. and Zhao, J. (2009). Named entity disambiguation by leveraging wikipedia semantic knowledge. In *Proceedings of the 18th ACM Conference on Information and*

[2] http://nlp.uned.es/weps/

Knowledge Management, CIKM 2009, Hong Kong, China, November 2-6, 2009, pages 215–224.

Larson, R. R. (2010). Information retrieval: Searching in the 21st century; human information retrieval. *J. Am. Soc. Inf. Sci. Technol.*, 61(11):2370–2372.

Lehmann, J., Monahan, S., Nezda, L., Jung, A., and Shi, Y. (2010). LCC approaches to knowledge base population at TAC 2010. In *Proceedings of the Third Text Analysis Conference, TAC 2010, National Institute of Standards and Technology, Gaithersburg, Maryland, USA, November 15-16, 2010*.

Mihalcea, R. and Csomai, A. (2007). Wikify!: linking documents to encyclopedic knowledge. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, CIKM '07, pages 233–242, New York, NY, USA. ACM.

Radfordyz, W., Hacheyz, B., Nothmanyz, J., Honnibalyz, M., and Curranyz, J. R. (2010). Document-level entity linking: CMCRC at TAC 2010. In *Proceedings of the Third Text Analysis Conference, TAC 2010, National Institute of Standards and Technology, Gaithersburg, Maryland, USA, November 15-16, 2010*.

Saggion, H., Funk, A., Maynard, D., and Bontcheva, K. (2007). Ontology-based information extraction for business intelligence. In Aberer, K., Choi, K.-S., Noy, N. F., Allemang, D., Lee, K.-I., Nixon, L. J. B., Golbeck, J., Mika, P., Maynard, D., Mizoguchi, R., Schreiber, G., and CudrÃ©-Mauroux, P., editors, *ISWC/ASWC*, volume 4825 of *Lecture Notes in Computer Science*, pages 843–856. Springer.

Tamilin, A., Magnini, B., and Serafini, L. (2010). Leveraging entity linking by contextualized background knowledge: A case study for news domain in italian. In *6th Workshop on Semantic Web Applications and Perspectives, SWAP 2010, Bressanone, Italy, September 21-22, 2010*.

Zhang, W., Chuan, Y., Sim, Su, J., and Tan, C. L. (2010a). NUS-I2R: Learning a combined system for entity linking. In *Proceedings of the Third Text Analysis Conference, TAC 2010, National Institute of Standards and Technology, Gaithersburg, Maryland, USA, November 15-16, 2010*.

Zhang, W., Su, J., Tan, C. L., and Wang, W. T. (2010b). Entity linking leveraging: automatically generated annotation. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 1290–1298, Stroudsburg, PA, USA. Association for Computational Linguistics.