Exposé for Diplom Thesis

Entity Linking by Means of Explicit Semantic Contexts

Torsten Huber[1]

June 26, 2013

supervised by:
Prof. Dr. Ulf Leser[1]
Dr.-Ing. Gjergji Kasneci[2]

[1]Knowledge Management in Bioinformatics, Department of Computer Science, Humboldt-Universität zu Berlin

[2]Institut für Softwaresystemtechnik, Hasso-Plattner-Institut an der Universität Potsdam

1 Motivation and Background

1.1 Problem Definition

"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."

(a) *David Cameron* could refer to the British politician, the English actor or any other namesake.

"The Ford Motor Company, founded in 1903 by Henry Ford, is one of the largest auto makers in the world. During a time of crisis throughout the auto industry in recent years, Ford emerged as the sole American automaker in a position to survive the steepest sales downturn in decades without a government bailout."

(b) The term *Ford* could refer to the *Ford Motor Company* or its founder *Henry Ford*.

Table 1: Variants of the entity linking problem.

Entity linking refers to the task of determining the correct database identifier for a mention of a named entity in a natural language text. A *mention* of a named entity – for example, a name of a person being referred to in a newspaper article – may be ambiguous, since a

number of entities may share that name. Consider the text in Figure 1a. The ambiguous term "*David Cameron*" may refer to David Cameron, the British politician or David Cameron, the English actor. The goal of entity linking is to determine the sense or *meaning* of a named entity reference. As such, it can be considered to be a special case of the word sense disambiguation (WSD) problem and name disambiguation in particular. As with the more general WSD problem, in order to assign a meaning, the entity mention must be associated with or *linked to* an external knowledge source, which contains such meanings. It is typically a knowledge base of some sort (e.g. a database or ontology) with unique identifiers for each entity.

1.2 Relevance for NER applications

Entity linking is not a trivial problem. A mention may have large number of meanings in the knowledge source, such as highly ambiguous acronyms like ABC (with roughly 100 different senses in the English Wikipedia[1]). Moreover, natural language texts may also use name variations (e.g. omitting the first or middle name of a person). Consider the text in Figure 1b. The term "*Ford*" may refer to the "*Ford Motor Company*" or the founder "*Henry Ford*". An entity linking system must be able to reliably link these mentions to the correct identifier, even if "*Ford*" is used to refer to the person instead of the company later on in the text.

Determining a unique database identifier provides access to more (structured) information, which is useful in several information retrieval (IR) and information extraction (IE) applications (Larson, 2010, p. 217). For example, it can be used to enrich texts with semantic information to provide useful meta information about the discussed entities. One application is the automatic link generation for entity references in news articles. Mihalcea and Csomai (2007) developed a software to parse new Wikipedia articles and automatically create tags and links to other articles about relevant entities. Entity linking can also 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 (Bekkerman and McCallum, 2005).

Linking entities is also required for most knowledge discovery tasks focusing on real-life entities. For example, to monitor events like product releases or company mergers as done by Saggion et al. (2007), a linker system must be able to accurately identify references to companies. Moreover entity linking can be useful in automating corporate customer care, such as complaint filing systems (Chakaravarthy et al., 2006). It can be used 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.), allowing the message to be automatically routed to the respective support staff member.

[1] `http://en.wikipedia.org/wiki/Abc`, accessed 26th February 2013

2 Goal

The goal of this Diplom thesis is to explore the feasibility of linking persons, organizations and geopolitical entities (GPEs) to database identifiers based on explicitly represented semantic constraints. The pivotal assumption is that, in order to successfully link entities, only a few clues are necessary to determine the *semantic context* of a natural language text. Given this context, a linker system only has to consider entities that fit into these semantic constraints while ruling out others that do not. For example, if a newspaper article mentions a person named "Michael Jackson" in the context of a football game, an entity linking system should not always return a link to the late American pop singer, but rather consider the American football player as a possible answer as well and rank the candidates according to the context.

In order to explore this approach, an entity linking system shall be developed, which utilizes semantic clues to determine contexts and then attempts to link every mention of a named entity in a newspaper article to identifiers of entities that fit into the determined context(s). It should then determine the set of entities with the highest semantic agreement.

3 Related Work

Much research has been performed in the field of entity linking. A notable contribution was made by the participants of the Text Analytics Conference (TAC). Since 2009, entity linking is 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 required to develop an entity linking 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 evaluated, ranging from simple matches of noun phrases to SVM-based learning algorithms (Chang et al., 2010; Lehmann et al., 2010; Zhang et al., 2010).

Bunescu and Pasca (2006) use cosine similarity to rank candidate entities based on the relatedness of the terms near an entity mention to a Wikipedia article. Cucerzan (2007) utilized 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 link distance of co-occurring terms to candidate concepts. While these methods also utilize semantic information from Wikipedia articles, their use of it is rather narrow, as they only consider information from one article or those that are directly linked to it. The proposed approach makes use of more general semantic information contained in a broad semantic context (e.g. domains like sports, music, politics).

Recently, research focused on entity linking systems that utilize machine learning algorithms. In the three KBP tracks of the TAC conferences since 2009, the best results were achieved by systems that use machine learning (Varma et al., 2009; Lehmann et al., 2010; Monahan et al., 2011), with accuracies of 81.4 %, 86.8 % and 84.6 % respectively. Many methods define a set of features in order to measure the similarity of documents. These features are aggregated into a vector representation, upon which a machine learning algorithm can learn to decide whether a mention refers to an entity or not. These features encompass:

- surface features: the similarity of the mention to the Wikipedia article title in terms of dice coefficient, substring match, acronym test, etc.,
- contextual features: terms in the context of the mention that are related to the entity (e.g. appear in the Wikipedia article),
- semantic features: e.g. testing whether entity mention and candidate are of the same type like `PERSON`,
- source features: the number of sources that generated this entity during candidate selection step and
- other features like measures for the popularity and polysemy of a candidate.

Tamilin et al. (2010) devised a method that is similar in spirit to our proposed approach, in that it is based on the observation that a human reader makes use of semantic background knowledge to disambiguate between namesakes. The *database context* they propose is defined by partially ordered dimensions that create a generalization/specialization hierarchy of the information in the database. It consists of a set of logical rules, such as `Player_Of(Zvonimir Boban, AC_Milan)` and `Has_Role(Zvonimir Boban, midfielder)`. The context ontology has been created from various sources, such as database tables, excel sheets, HTML pages and structured date like Freebase and Wikipedia. However, these information were manually extracted and converted into the target database format and only applied on Italian newspapers. Thus, this approach lacks feasibility for larger applications, since the effort of manually annotating all relevant information for a large number of entities would be considerable.

4 Approach

The basic idea of the proposed approach is to link a mention of an entity in a natural language text based on semantic contexts, which are in essence a set of related terms, as we will describe later in this section. We do not consider NER as part of the entity linking problem, but assume that the mentions that are to be linked are provided as input as well. Thus, for a given input document (e.g. a newspaper article) and the entity mentions, the linking process will then consist of three steps:

1. Assign semantic contexts to the text.

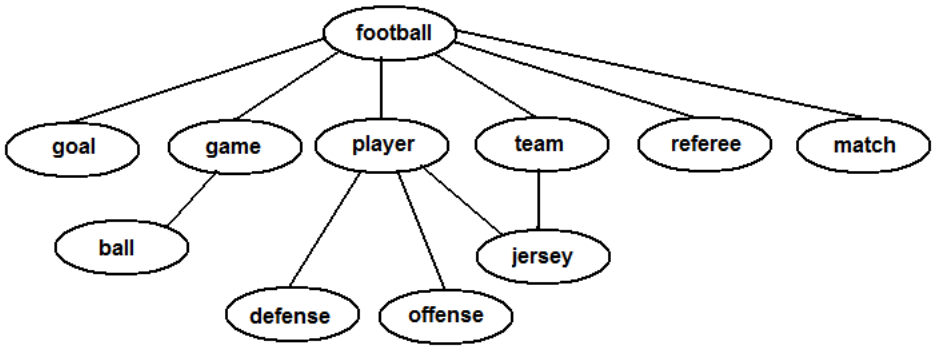

Figure 1: A semantic network of terms for the example context "`Football`"

2. For all mentions, determine suitable candidate entities that fit into either of the determined contexts of the previous step.
3. Determine the combination of contexts and entities with the highest semantic agreement.

One of the crucial tasks will be devising appropriate and well-defined semantic contexts. These contexts must be chosen such that they can be accurately identified in natural language texts while being unambiguous enough to distinguish entities from different contexts. Before we can think of defining the contexts, a suitable knowledge source has to be chosen first though. Since we intend to focus on linking persons, organizations and geopolitical entities, the English Wikipedia[2] or the Wikipedia-based ontology YAGO[3] appear suitable for the task. Both provide access to rich information about a large number of entities. Also, they introduce a set of categories that are assigned to each entity, which can be used to determine semantic contexts. An appropriate level of detail for this categorization has to be determined before it can be applied (e.g. `Athletes`, `Basketball_players`, `Basketball_players_from_Virginia`, `Basketball_players_at_the_1984_NCAA_Men's_Division`[4].

In detail, a context consists of at least a set of related terms, which together describe its semantic meaning. In its simplest form, this can be realized with a flat list of common terms for the context or a semantic network as depicted in Figure 1. In order to determine common terms for a particular context, the Wikipedia articles can again be used to, for example, calculate the tf-idf scores for words occurring in articles about *all* football players

[2] `http://en.wikipedia.org`

[3] `http://www.mpi-inf.mpg.de/yago-naga/yago/`

[4] Wikipedia categories that are assigned to articles are typically rather specific like the last example, as opposed to the more general parent categories that only contain sub-categories and no articles. Thus, a suitable level of abstraction in the category hierarchy is necessary and will have to be determined.

as well as articles related to football (including the article about the game itself) to determine terms that appear to be relevant for the context "football". Moreover, these terms have to be filtered, such that only those that do not occur in too many other contexts are retained (while also considering different meanings of, for example, the word "goal" or "offense" in different contexts, which technically poses a WSD problem in itself). In order to create a semantic network from terms that were determined in this step, their relationship in the Wikipedia can be examined, e.g. whether they directly or indirectly link to each other. Other sources such as WordNet[5] which is also integrated into YAGO could be utilized as well.

The main focus of the thesis will be linking mentions of persons to Wikipedia articles. Additionally, we also want to examine the applicability of our approach for organizations and geopolitical entities. Wikipedia categories like `Chip_Manufacturer` or `Environmental_Organization` fit into the concept of a category or context hierarchy. For geopolitical entities such as countries, lakes or cities, we must consider geological constraints such as a city being located in a state which in turn is located in a country and a continent etc., which can be realized with the Wikipedia category hierarchy as well. Other information such as temporal constraints (e.g. a person lived or a country existed for a certain period of time) and other semi-structured information from the Wikipedia info-boxes could be used as well. However, the use of factual knowledge is not part of our research goal, since we intend to focus mainly on the usability of semantic contexts.

After a set of suitable contexts has been defined in the first phase of the development, an algorithm for identifying them in natural language texts has to be devised. The goal is to develop a strategy for determining a number of contexts that could apply to a natural language text, e.g. it is about the topic sports, politics or music. For a given text, it is conceivable that several contexts apply, such that a probabilistic approach will be necessary. Moreover, for each mention there could be several candidates, leading to a number of possible combinations of contexts and entities. A strategy has to be devised to find the best combination, such as only considering the best-matching context or attempting to calculate the combination of contexts and candidate entities with the highest semantic agreement (in terms of their semantic relationships, e.g. the distance in the Wikipedia or them belonging to sub-categories of the same parent category). Considering these relationships of entities that are mentioned in the same text, it appears sensible to link all mentions at once in order to determine and establish semantic relationships. The algorithm also has to deal with inconsistencies of all sorts, such as entities that "fall out of context" in some way, e.g. an article mentions a singer that appears in the half-time show of a football game. Thus, the possibility of assigning contexts on a per-entity or per-sentence basis (additionally or exclusively) rather than the document-wide approach described above should be explored as well.

[5] `http://wordnet.princeton.edu/`

Lastly, it may happen that an entity mentioned in a text is missing in the knowledge source. In such cases, the linker system is supposed to return `NIL` to denote that the mentioned entity is not contained in the knowledge source. In theory, this *`NIL` detection* is trivial for our approach, assuming that the assigned context were always the correct one. In this case, the linker system would only need to determine whether there is a candidate entity that fits into the determined context or return `NIL` if there is no such one. Since the context detection will most likely produce errors in a number of cases though, strategies will have to be developed to make `NIL` detection more accurate.

4.1 Evaluation

Although entity linking is a well-researched area, the amount of suitable publicly available evaluation corpora is sparse. Thus, a common practice used by several researchers (Cucerzan, 2007; Mihalcea and Csomai, 2007; Bunescu and Pasca, 2006; Milne and Witten, 2008) is to use the Wikipedia for evaluation. By removing all inter-Wikipedia links that were manually created by human editors from the articles and then tasking the entity linking with reconstructing them, the links that were removed before can be used as the gold standard annotation. Since our main intended application is linking named entities mentioned in newspaper articles, we do not consider this an appropriate evaluation method for our purpose due to the difference in style of writing and information provided in Wikipedia articles as opposed to newspaper articles (which we assume will generally expect more background knowledge from the reader). However, it may provide a measure for the general practicability of the proposed approach under arguably favorable circumstances.

The few available corpora such as the newspaper corpus developed by Cucerzan (2007)[6] or the IITB dataset[7] used by Han et al. (2011) can be used for evaluation and comparison. A different evaluation corpus could be generated automatically by utilizing the *external links* and *references* in Wikipedia articles. These links point to websites that are in some way related to the entity being discussed in the article. Thus, we can safely assume that if a mention occurs, it will most likely refer to that entity and not any other namesake. The reference links in particular often point to online newspaper articles and as such might prove particularly useful for our evaluation requirements.

Recently, another large corpus has been created by Google Inc., which can be used for the purpose of entity linking. The Wikilinks Corpus[8] contains more than 40 million entities from about 10 million websites mapped to their respective Wikipedia articles. Although

[6] Cucerzan (2007) created two evaluation corpora; one based on 350 Wikipedia articles and the other based on 20 news articles with 756 entity mentions. The latter we are referring to here.

[7] `http://www.cse.iitb.ac.in/~soumen/doc/QCQ/`

[8] `http://googleresearch.blogspot.de/2013/03/learning-from-big-data-40-million.html`

we cannot confirm this with absolute certainty at this point, the corpus appears to have been extracted from the links and references in Wikipedia articles as described above, thus making their extraction superfluous. Nevertheless, due to its size and presumably high coverage, the corpus should be suitable for the evaluation of the entity linking system.

References

Bekkerman, R. and McCallum, A. (2005). Disambiguating web appearances of people in a social network. In *Proceedings of the 14th International World Wide Web Conference*.

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.

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*, 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*, pages 708–716.

Han, X., Sun, L., and Zhao, J. (2011). Collective entity linking in web text: a graph-based method. In *Proceedings of the 34th international ACM conference on Research and development in Information Retrieval*, SIGIR '11, pages 765–774, New York, NY, USA.

Han, X. and Zhao, J. (2009). Named entity disambiguation by leveraging wikipedia semantic knowledge. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management, CIKM 2009, Hong Kong, China*, 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, Gaithersburg, Maryland, USA*.

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.

Milne, D. and Witten, I. H. (2008). Learning to link with wikipedia. In *Proceedings of the 17th ACM conference on Information and knowledge management*, CIKM '08, pages 509–518, New York, NY, USA. ACM.

Saggion, H., Funk, A., Maynard, D., and Bontcheva, K. (2007). Ontology-based information extraction for business intelligence. In *Proceedings of the 6th international The semantic web and 2nd Asian conference on Asian semantic web conference*, ISWC'07/ASWC'07, pages 843–856, Berlin, Heidelberg.

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*.

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