Evaluating DBpedia Spotlight for the TAC-KBP Entity Linking Task

Pablo N. Mendes, Joachim Daiber, Max Jakob, Christian Bizer

Web-based Systems Group, Freie Universität Berlin, Germany
first.last@fu-berlin.de

Abstract

This paper reports the participation of team WBSG in the English entity-linking task at TAC KBP 2011. As first-timers in TAC, we used the opportunity to test the entity linking accuracy of our general-purpose entity and concept annotation system. Our system, DBpedia Spotlight, does not specialize on named entities of certain types. Its aim is to annotate any of the ≈3.5M entities and other concepts in DBpedia, a knowledge base extracted from Wikipedia. We applied the disambiguation step of our algorithm as is, without benefiting from the a priori knowledge of target entity types, and without applying specialized NIL detection or clustering approaches. We simply mapped our KB to that of TAC KBP's, and labelled as NIL the entities from the former that were absent from the latter. Our simple approach worked surprisingly well, achieving 0.703 B^3 F1. In future work we plan to specialize our algorithm to the entity types in TAC KBP, and investigate in which cases a larger KB provided us an advantage or disadvantage in the context of this evaluation.

1 Introduction

The DBpedia project (Bizer et al., 2009) extracts structured information from Wikipedia editions in 97 different languages and combines this information into a large multi-lingual knowledge base. The knowledge base contains textual descriptions (titles and abstracts) of concepts in up to 97 languages. Furthermore, it contains structured knowledge that has been extracted from the infobox templates of Wikipedias in 15 different languages. This information is mapped onto a single consistent ontology. The DBpedia Ontology organizes the knowledge on Wikipedia according to a hierarchy of 320 classes and 1,650 different properties. Mappings between Wikipedia infoboxes and the DBpedia Ontology are community-generated, and allow a more homogenized view of the knowledge – e.g. mapping multiple spellings of properties to one canonical name, or recognizing that many infobox types describe the same entity type.

In order to enable the linkage of text corpora with this knowledge base, we developed DBpedia Spotlight (Mendes et al., 2011), a system to recognize and link DBpedia entities and concepts mentioned in text. In comparison with previous work, DBpedia Spotlight aims at a more comprehensive and adaptable solution. First, while other semantic annotation systems are often restricted to a small number of entity types, such as people, organisations and places, our system attempts to annotate DBpedia entities of any of the 320 classes (more than 30 top level) in the DBpedia Ontology. Second, since a single generic solution is unlikely to fit all task-specific requirements, our system enables user-provided configurations for different use cases with different needs. Users can flexibly specify the domain of interest, as well as the desired coverage and error tolerance for each of their specific annotation tasks.

In this paper we report the evaluation of DBpedia Spotlight on the TAC KBP Entity Linking Task [1]. The objective of this task is to find the correct identifier from a knowledge base (*KB*) of persons, loca-

[1] http://nlp.cs.qc.cuny.edu/kbp/2011/

tions and organizations, given a surface form (e.g. an entity name) and the text in which this name occurred. In case there is no such identifier (the entity is not in the *KB*), the system should return NIL. Moreover, the system is required to cluster entities referring to the same NIL entity.

In the following we describe how we prepared the data and interlinked DBpedia with TAC KBP's knowledge base (Section 2), our Entity Linking approach (Section 3) as well as our results (Section 4).

2 Data Preparation

We built DBpedia Spotlight from DBpedia and Wikipedia. In this section we describe how we obtained the necessary datasets and how we mapped our knowledge base to that of TAC KBP's.

2.1 DBpedia Spotlight Index

We build G, a graph of page titles, redirects and disambiguations in DBpedia to extract a mapping that associates multiple surface forms to an entity and interconnects multiple entities to an ambiguous surface form. Every page title is a node in G. An edge $red(u, v)$ is in G if a redirect page u links to a page v. Similarly, an edge $dis(u, v)$ is in G if a disambiguation page u links to a page v. A surface form s is considered a name variation for an entity r if $dis(s, r) \in G$ or if there is a path $red(s, ..., r)$.

We pre-processed Wikipedia articles, extracting every page link $l = (s, e)$ with surface form s as anchor text and entity e as link target, along with the paragraph representing the context of that wikilink occurrence. Each wikilink was stored as an evidence of occurrence $o = (e, s, C)$. Each occurrence o records the fact that the DBpedia entity e represented by the link target has been mentioned in the context of the paragraph C through the use of the surface form s. Before storage, the context paragraph was tokenized, stopworded and stemmed, generating a vector of terms $C = \langle w_1, ..., w_n \rangle$. The collection of occurrences for each entity was then stored as a document in a Lucene index[2] for retrieval in the disambiguation stage.

Wikilinks can also be used to estimate prior probability of observing a given entity in text, or the likelihood of a surface form s referring to a specific candidate entity $e \in E_s$. We consider each wikilink as evidence that the anchor text is a commonly used surface form for the DBpedia entity represented by the link target. By counting the number of times a surface form occurred with and without a DBpedia entity $n(s, e)$, we can empirically estimate a prior probability $P(e)$ and a conditional probability of seeing an entity e given that surface form s was used $P(e|s) = n(s, e)/n(s)$ using a maximum likelihood estimate. These and other estimates are shared in our *lexicalizations dataset* (Mendes et al., 2012).

2.2 Knowledge Base Mapping

The TAC KBP knowledge base uses custom entity identifiers (e.g. `E0456437`). The definition for each entity in the knowledge base contains the URL segment of the corresponding article in the English Wikipedia (e.g `A._K._Antony`). As DBpedia also derives its URIs from the Wikipedia pages, we used the page titles to map the TAC KBP knowledge base to DBpedia. However, the versions of Wikipedia used by the TAC KBP knowledge base and DBpedia are different. This means that some articles may have become redirect pages, requiring a preprocessing step to map these two KBs. We perform this mapping using the graph of redirects from DBpedia 3.6 (based on the Wikipedia dump from 11/10/2010).

3 Approach

Our Entity Linking approach works in three stages. *Candidate selection* is employed to map a surface form to entities that are candidate disambiguations for that surface form. The *Disambiguation* stage uses the context around the surface form in the source text to decide for the best choice amongst the candidates. Finally, the *Linking* stage decides if the chosen candidate should be linked, or if it should be assigned to a NIL cluster.

3.1 Candidate Selection

The candidate selection phase uses G – the graph defined in Section 2.1 – to choose the possible disambiguations for a surface form and returns them as a list. The probability estimates of the lexicalizations dataset, for example, can be utilized to pre-rank the candidates for disambiguation before observing a

[2] `http://lucene.apache.org`

	(McNamee et al., 2010)	(Bagga and Baldwin, 1998)			(Ji et al., 2011)		
Data Set	μAVG	B^3P	B^3R	B^3F_1	$B^{3+}P$	$B^{3+}R$	$B^{3+}F_1$
TAC-KBP 2010 EL-1.0	0.827	0.904	0.958	0.930	0.773	0.805	0.789
TAC-KBP 2011 EL-1.1	0.727	0.920	0.971	0.945	0.693	0.713	0.703

Table 1: Evaluation results for TAC KBP English Entity Linking Gold Standards.

surface form in the context of a paragraph. Choosing the DBpedia entity with highest prior probability for a surface form is the equivalent of selecting the "default sense" of some phrase according to its usage in Wikipedia.

3.2 Disambiguation

After selecting a set of candidate entities for each surface form, our system uses the context around the names, e.g. paragraphs, as information to find the most likely disambiguations.

We modeled DBpedia entity occurrences in a Vector Space Model (VSM) (Salton et al., 1975) where each DBpedia entity is represented by a point in a multidimensional space of words. We use the Term Frequency (TF) to measure the relevance of a word for a given entity. The TF is further combined with an entropy-inspired score, the Inverse Candidate Frequency (ICF) (Mendes et al., 2011) that weighs the importance of a word on their ability to distinguish between candidates for a given surface form.

The intuition behind ICF is that the discriminative power of a word is inversely proportional to the number of DBpedia entities it is associated with. Formally, let E_s be the set of candidate entities for a surface form s. Let $n(w_j)$ be the total number of entities in E_s that are associated with the word w_j. Then we define:

$$ICF(w_j) = \log \frac{|E_s|}{n(w_j)} = \log |E_s| - \log n(w_j) \qquad (1)$$

Given the VSM representation of DBpedia entities with TF*ICF weights, the disambiguation task can be cast as a ranking problem where the objective is to rank the correct DBpedia entities at position 1. Our approach is to rank candidate entities according to the similarity score between their context vectors and the context surrounding the surface form. In this work we use cosine as the similarity measure.

We also attempted a simple combination of the prior and TF*ICF scores. The mixing weights were estimated through a small experiment using linear regression over held out training data. The results reported in this work used mixed scores computed through the formula:
$Mixed(r, s, C) = 1234.3989 * P(r) + 0.9968 * contextualScore(r, s, C) - 0.0275$

Further research is needed to carefully examine the contribution of each component to the final score. The difference in magnitude of the weights is due to lack of normalization in our scores. We are currently working on another formulation with normalized scores.

3.3 Linking

As DBpedia contains more entities than the TAC KBP knowledge base (*KB*), it is possible that the highest ranking entity after the disambiguation stage is not present in the *KB*. If that is the case, or if no disambiguation was found, this entity is considered NIL. NIL clustering is performed as follows: i) if the entity is in DBpedia, but not in TAC KBP, we use the DBpedia URI to place all references in to the same NIL cluster, and ii) if the entity is not on DBpedia, we use the surface form to perform the clustering.

4 Evaluation

4.1 Results

The best results achieved by DBpedia Spotlight are presented in Table 1. The scores reported are the KBP2010 micro-average (μAVG) (McNamee et al., 2010), the B-cubed cluster scoring (B^3) (Bagga and Baldwin, 1998) and the B-Cubed+ modification (B^{3+}) (Ji et al., 2011). We report all three scoring metrics in order to allow comparisons with systems from other years.

Figure 1 compares the F1 by entity type for NIL and Non-NIL annotations. A NIL annotation means that an entity was marked in text, but it is not present

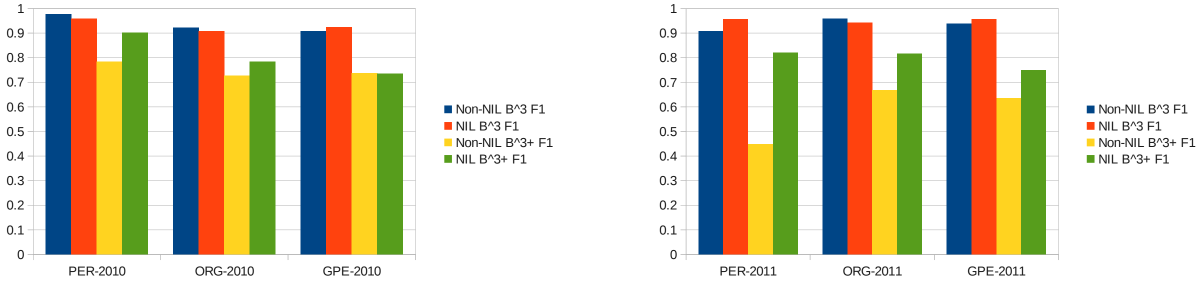

Figure 1: Comparison of B^3 and B^{3+} F_1 scores for NIL and Non-NIL annotations, for each entity type for 2010 and 2011 data sets.

in the *KB*. The entity types in the *KB* are PER (Person), ORG (Organizations) and GPE (Geopolitical Entity).

On the one hand, the fact that DBpedia is a superset of TAC KBP gave us an advantage of providing a rudimentary form of NIL detection. Every disambiguation that contained a DBpedia entity not in TAC KBP was considered a NIL. On the other hand, our disambiguation algorithm had to deal with many more candidates. For highly ambiguous entities (e.g. some person names) with low contextual support from the text in which it occurs (e.g. some geopolitical entities), this could be a problem.

5 Conclusion

We have tested DBpedia Spotlight, a general-purpose entity/concept annotation system for natural language text, on the task of English entity linking in TAC KBP 2011. The system was employed without customizations for this task and has obtained an encouraging near-median B^3 F1 score when compared to other participants that used Wikipedia text (but no Web). In future work we plan to implement specializations for the entity types in this dataset, as well as investigate the influence of our particular KB in the results obtained in TAC KBP.

DBpedia Spotlight is deployed as a freely available Web service, and includes a user interface for demonstration purposes. The source code is publicly available under the Apache license V2, and the documentation as well as supporting datasets are available from `http://dbpedia.org/spotlight`.

Acknowledgements

This work was partially supported by the European Commission through the project *LOD2 – Creating Knowledge out of Linked Data*.

References

Amit Bagga and Breck Baldwin. 1998. Algorithms for scoring coreference chains. In *LREC Workshop on Linguistics Coreference*, pages 563–566.

Christian Bizer, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. 2009. DBpedia - A crystallization point for the Web of Data. *Web Semantics: Science, Services and Agents on the World Wide Web*, 7:154–165, September.

Heng Ji, Ralph Grishman, and Hoa Trang Dang. 2011. Overview of the TAC2011 Knowledge Base Population Track. In *Text Analysis Conference*.

Paul McNamee, Hoa Trang Dang, Heather Simpson, Patrick Schone, and Stephanie Strassel. 2010. An evaluation of technologies for knowledge base population. In *LREC*.

Pablo N. Mendes, Max Jakob, Andrés García-Silva, and Christian Bizer. 2011. DBpedia Spotlight: Shedding Light on the Web of Documents. In *Proceedings of the 7th International Conference on Semantic Systems (I-Semantics)*.

Pablo Mendes, Max Jakob, and Christian Bizer. 2012. DBpedia for NLP: A Multilingual Cross-domain Knowledge Base. In *LREC - to appear*.

Gerard M Salton, Andrew K C Wong, and Chung-Shu Yang. 1975. A vector space model for automatic indexing. *Commun. ACM*, 18:613–620, November.