Named Entity Disambiguation using Wikipedia

Martin JAČALA*

Slovak University of Technology
Faculty of Informatics and Information Technologies
Ilkovičova 3, 842 16 Bratislava, Slovakia
jacala06@student.fiit.stuba.sk

The constantly growing amount of human written textual content available on the Web is a source of interesting and actual information about persons, organisations or places. One of the problems we face when analysing or querying in such content is name ambiguity. Does the word jaguar mean the sports car, the jungle animal or something different? Which Michael Jordan does the text refer to?

The proper names in news articles comprise approximately 10% of text and many of such proper names are ambiguous. In our work we propose an approach to answer these questions by disambiguating named entities using explicit semantics extracted from web-based corpora serving as background knowledge. We follow Miller and Charles distributional hypothesis [4] stating that similar entities appear in similar contexts even across multiple documents.

The problem of disambiguating named entities found in common, human written textual resources (usually referred as Named Entity Disambiguation) is a well established task in the natural language processing community. This task originated at the 6th Message Understanding Conference and has come a long way since then. Various approaches has been proposed over time, such as creating clusters of similar entities within set of given documents, or approaches specific to problem domain (e.g., geographical names or persons). However, these methods are rather limited when dealing with constantly changing open web data.

Using Wikipedia data as background knowledge for disambiguation has been proved successful by mapping entities on Wikipedia articles [1]. For each string containing an ambiguous entity they extract all articles that can be referred with the entity. We compute tf-idf cosine similarity measure with the ambiguous string for each retrieved article. The documents are further extended with term vectors from documents belonging to the same category. Evaluation of the system on various Wikipedia articles gives precision of approximately 80%. Similar approach use different context generation method together with secondary measure based on Wikipedia's category taxonomy, improving the precision to about 88% on Wikipedia articles.

---

* Supervisor: Jozef Tvarožek, Institute of Informatics and Software Engineering

Spring 2011 PeWe Workshop, April 1, 2011, pp. 53–54.

In our work we use the explicit semantics already present in Wikipedia data to extract all possible meanings of currently analysed entity. The network of redirects and page links helps us to resolve possible synonyms. We further use disambiguation pages to extract articles corresponding to various meanings of this entity (and any other extracted via redirects).

Instead of directly comparing the fragments of documents using similarity measure such as cosine similarity, we build a 'semantic space' using Explicit Semantic Analysis [3]. This method is similar to the Latent Analysis, however it does have certain assumptions on processed data. The most notable difference is that in ESA each document corresponds to an 'explicit concept' instead of inferring latent concepts from large, un-labelled dataset with LSA. The created semantic space is in fact a term-document matrix containing weighted list of Wikipedia concepts with respect to the individual keywords.

Further in the analysis process, we transform the analysed document and each of Wikipedia article retrieved in the previous step using disambiguation pages into the concept space. The document vector is computed as the running total of intermediate vectors computed while creating the semantic space. With vectors transformed, we compare them using cosine similarity metric normalised with Euclidean distance.

Finally, we assume that most similar article will describe the correct meaning of analysed entity. In our preliminary evaluation, we were able to obtain precision up to 74% on news article texts. We experienced similar behaviour as [2] when the method failed to rank the correct meaning first, even when using semantic spaces during the comparison. In most cases the correct meaning was ranked among the top three entities from whole list of possible meanings, containing usually 20 to 50 'candidates'.

We plan to further improve the results with additional classifier based on Wikipedia category structure. Our final goal is to prepare working web-based demonstration of proposed method, which require further optimisations and design decisions to speed up the current experimental implementation.

Acknowledgement. This work was partially supported by the Scientific Grant Agency of Slovak Republic, grant No. VG1/0508/09.

References

[1] Bunescu, R., Pasca, M.: Using encyclopedic knowledge for named entity disambiguation. In *Proc. of EACL*, Vol. 6, pp. 9–16, 2006.
[2] Cucerzan, S.: Large-scale named entity disambiguation based on Wikipedia data. In *Proc. of EMNLP-CoNLL*, pp. 708–716, 2007.
[3] Gabrilovich, E., Markovitch, S.: Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis. Evaluation, pp. 1606–1611, 2006.
[4] Miller, G.A., Charles, W.G.: Contextual correlates of semantic similarity. *Language and Cognitive Processes*, Vol. 6, No. 1, 1–28, 1991.