WIKIPEDIA EXPLICIT SEMANTICS BASED ENTITY DISAMBIGUATION

Jozef TVAROŽEK

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 2, 842 16 Bratislava, Slovakia
jtvarozek@fiit.stuba.sk

Abstract. We present an approach for employing explicit semantics in large knowledge bases such as Wikipedia to construct a disambiguation dictionary and vector-based word model for named entity disambiguation task. The relatedness of surface forms is computed as cosine similarity between the explicit semantic vectors. The proposed approach outperforms traditional approaches such as latent semantic analysis.

1. Introduction and related work

Named entity disambiguation is a problem of identifying and resolving references to entities in natural text. The task is crucial for building entity links for knowledge graphs in digital spaces, and also plays an important part in users' queries [5] in navigating large information spaces. Named entities can describe user interests and therefore are also very useful both in user modelling and personalization scenarios.

The goal of named entity disambiguation is to label surface forms describing entities in natural text with an appropriate label, such as person, organisation, location, etc. Most existing approaches attempt to solve this problem by first extracting various textual features from large knowledge bases (e.g. Wikipedia, Open Directory). Various methods have been proposed: knowledge base building using a heuristic analysis of letter capitalisation in Wikipedia articles together with context enhanced with SVMs with tree kernels [2], disambiguation using an approach to maximize similarity between entity's context and context extracted from Wikipedia's articles [3] and the WikiRelate project [6].

Even the largest knowledge bases provide limited information for not so well-known contexts, therefore methods for augmenting the knowledge base with additional information [7] from outside and inside sources are also been researched, improving the disambiguation accuracy from 43% to 86%.

2. Named entity disambiguation based on explicit semantics

We briefly present our approach described in [1]. The main idea is to leverage the semantics already present in Wikipedia – i.e. hyperlinks, redirects, disambiguation pages, and markup from human editors.

In the first step, we construct a disambiguation dictionary that contains a set of candidate meanings for each entity's surface forms we find in Wikipedia. Surface forms are extracted from MediaWiki markup format [SurfaceForm|ArticleName]. Similarly for redirect pages and disambiguation pages that provide us with various misspellings, informal names, and possible meanings. Besides the structural information, we aggregate contexts in which the surface forms appear. We employ explicit semantic analysis [4] to build vector semantic space – that is each dimension in the semantic space corresponds to one concept defined in the training set (Wikipedia articles).

The disambiguation process then follows four stages:

1. Entity identification and boundary detection – we employ a standard entity recogniser, Stanford NER, providing us with baseline performance.
2. Transformation of the input document – the input document is transformed into explicit semantic vector, in respective vector's values correspond to the relatedness of the document to respective explicit concepts (Wikipedia articles).
3. Lookup of candidate meanings – query disambiguation dictionary to retrieve possible candidate meanings for each surface form retrieved from the NER system. If no matches are found we employ approximate string matching algorithm to handle misspellings. For each meaning found, the corresponding Wikipedia article is passed to the next stage.
4. Ranking of candidate meanings – we rank the possible meanings according to the cosine similarity between the input document vector and the vector of the Wikipedia article corresponding the meaning.

In case of short input document or incomplete Wikipedia article, this approach fails to select the correct meaning as the best ranked. We, therefore, construct additional contextual vectors as semantic vectors for sliding window around the surface forms found in the documents.

3. Evaluation and conclusions

Three separate datasets each containing 20 manually disambiguated news articles we used in evaluation: one during development, and two for evaluation only. During dataset coding, around 10% of surface forms were discarded due to not having a corresponding Wikipedia entity. On average, named entities in the datasets have 18 meanings, and about 78% of the entities have more than one meaning.

We compared named entity disambiguation of a human annotator to the disambiguation of the proposed method. We achieved 85.06% to 87.84% accuracy with our single vector approach, and 86.56% to 90.25% accuracy with the additional contextual vectors. The accuracy was improved in cases where surface forms referred to topics significantly different from the rest of the input document.

Comparing to the traditional methods the explicit semantic analysis brings improvement. Latent semantic analysis with 250 latent dimensions achieved accuracy of 80.34% to 82.33%. Increasing or decreasing the number of latent dimensions did not bring any significant improvement.

In this paper we briefly summarized our approach to named entity disambiguation using explicit semantics found in large knowledge bases such as Wikipedia. Our approach outperforms traditional approaches such as latent semantic analysis. The approach can be extended with additional contextual features to better improve on various in-document entity links or other linguistic phenomena.

Acknowledgement: This work was partially supported by the grants VG1/0971/11/2011-2014, KEGA 028-025STU-4/2010, APVV-0208-10 and it is the partial result of the Research & Development Operational Programme for the project Research of methods for acquisition, analysis and personalized conveying of information and knowledge, ITMS 26240220039, co-funded by the ERDF.

References

to other papers publishing the results that are summarized here

[1] Jačala, M., Tvarožek, J.: Named entity disambiguation based on explicit semantics. In: *Proceedings of SOFSEM 2012: Theory and Practice of Computer Science*. Springer Berlin Heidelberg, (2012), pp. 456–466.

Other references

[2] Bunescu, R., Pasca, M.: *Using encyclopedic knowledge for named entity disambiguation*. In: *Proceedings of EACL*, (2006), pp. 9–16.

[3] Cucerzan, S.: *Large-scale named entity disambiguation based on Wikipedia data*. In: *Proceedings of EMNLP-CoNLL*, (2007), no.6, pp. 708–716.

[4] Gabrilovich, E., Markovitch, S.: *Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis*. In: *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, (2007), pp. 1606–1611.

[5] Guo, J., Xu, G., Cheng, X., Li, H.: *Named entity recognition in query*. In: *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval - SIGIR '09*. ACM Press, New York, (2009), pp. 267–274.

[6] Strube, M., Ponzetto, S.: *WikiRelate! Computing semantic relatedness using Wikipedia*. In: *Proceedings of AAAI-06*. AAAI Press, MIT Press, vol. 21, (2006), pp. 1419–1424.

[7] Yang, L., et al.: *Mining evidences for named entity disambiguation.* In: Proceedings of the 19th ACM SIGKDD. ACM, (2013), pp. 1070–1078.