An Approach to Named Entity Disambiguation Based on Explicit Semantics

Martin Jačala*
Institute of Informatics and Software Engineering
Faculty of Informatics and Information Technologies
Slovak University of Technology in Bratislava
Ilkovičova 3, 842 16 Bratislava, Slovakia
jacala06@student.fiit.stuba.sk

Abstract

Identification of the named entities in an unstructured, human written text is a well established subtask of Natural Language Processing. However, marking each occurrence of a named entity in the text with a class label is usually not sufficient, as the same word often describe several different entities. A dedicated area of NLP, Named Entity Disambiguation, has been devised to solve this problem. In our work we present an approach to the problem of Named Entity Disambiguation based on the Explicit Semantic Analysis. We use a semantic similarity measure based on the similarity between context of the entity and the documents describing the possible meanings. We use an additional semantics provided by Wikipedia, such as disambiguation and redirect pages or links between the documents. Evaluation of the proposed method shows an improvement over the traditionally used Latent Semantic Analysis.

Categories and Subject Descriptors

H.3.3 [**Information Systems**]: Information storage and retrieval

Keywords

Named Entity Disambiguation, Explicit Semantic Analysis, Semantic Similarity

*Master degree study programme in field Information Systems. Supervisor: Dr. Jozef Tvarožek, Institute of Informatics and Software Engineering, Faculty of Informatics and Information Technologies, STU in Bratislava. Work described in this paper was presented at the 7th Student Research Conference in Informatics and Information Technologies IIT.SRC 2011.

Jačala, M. An Approach to Named Entity Disambiguation Based on Explicit Semantics. Information Sciences and Technologies Bulletin of the ACM Slovakia, Special Section on Student Research in Informatics and Information Technologies, Vol. 3, No. 2 (2011) 55-58

1. Introduction

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 the 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 [5] and many of them are ambiguous. In our work we propose an approach to answer such questions by disambiguating the named entities using explicit semantics extracted from a web-based corpora used as the background knowledge. We follow the Miller and Charles distributional hypothesis [12] stating that similar entities appear in similar contexts even across multiple documents.

This paper reports on proposed method based on Explicit Semantic Analysis with data from Wikipedia as the source of background knowledge. Additionally, we use Wikipedia to disambiguate meaning of the entities by creating link (relation) with the appropriate page. The paper is organised as follows. In the Section 2 we discuss various approaches to the named entity disambiguation and present naive problem formalisation. The Section 3 explains the proposed method, results of experimental evaluation are presented in section 4. The last Section concludes and presents future work.

2. Related Work

The named entity disambiguation problem is a well established task in the NLP community. The task of named entity recognition was introduced at the 6th Message Understanding Conference [8] and is defined as identification of proper names, temporal and numeric expressions with appropriate labels. Such tasks are referred as structural disambiguation. The task has been further refined as grouping all mentions of an entity within given document (within-document coreference, WDC) [4].

Breaking the document boundary and moving to open, web based corpora is an natural step, mainly because of the constantly growing amount of such data and affordable computing power. The approach proposed by [1] used within-document coreference to find all entity references and constructed a vectorial representation of the sentences that contain the individual entities. The vec-

tors from the constructed set are compared with other documents using an support vector model. The method has been further improved by eliminating WDC [14], evaluation on 197 news articles containing 35 different *John Smiths* showed precision of 90%.

The presented methods are designed rather to group together entities or their context based on similarity measure than find relations of an entity group with objects in the real world. The approaches based on a list of reference entities, such as geographical names disambiguation [15] or disambiguation of scientific paper authors based on DBLP data [9] achieved precision of approximately 85%. The remaining problem is that such approaches are domain specific. This makes the disambiguiation of an open text such as news articles problematic.

Using the data from Wikipedia as a background knowledge for the disambiguation has been proved successful by mapping the entities to the appropriate Wikipedia articles [3]. For each string that contains an ambiguous entity they extract all of the articles that can be referred with the entity. Then, an tf-idf cosine similarity measure is computed for the ambiguous string for each retrieved article. The documents are then further extended with term vectors created from the documents in the same category. The evaluation of the system on various Wikipedia articles gives precision of approximately 80%. Similar approach by [6] use different context generation method together with a secondary measure based on the Wikipedia's category taxonomy, improving the precision up to 88% on selected Wikipedia articles.

2.1 Formal Definition

The majority of the presented approaches share similar definition of the disambiguation task as a ranking problem. Given a document fragment q (query) containing the ambiguous entity and a set of documents D, where each $d \in D$ is a document describing one of the candidate meanings for the ambiguous entity, we compute the maximal similarity given as

$$sim = \arg\max_{d} \ rank(q, d)$$

where rank is a ranking function. The definition of the ranking function varies, as varies the choice of the entity query (q). Additional measures are combined to further improve and refine the results.

3. Explicit Semantic Analysis

Traditionally, the approaches such as Latent Semantic Analysis [11] are used to cluster documents into the latent concepts from a large, text based corpora. Such corpora can be uncategorized, lacking any added semantics. The advantage of such approach is inexpensive data acquisition and fairly good discovery of the hidden concepts [13].

With this method we can construct a term-concept matrix (*semantic space*) where the matrix values are the tf-idf frequencies of the words extracted from given corpora. Usually, we decrease the number of dimensions using mathematical models, e.g. singular value decomposition, retaining only the 'notable' latent concepts.

The recently proposed method in [7] represent the meaning of the text in a high dimensional semantic space of natural concepts derived from a human generated dataset, the Wikipedia. The evaluation of this method shows that the classification of the text using the natural concepts extracted from Wikipedia yield better results than the traditional LSA method.

In our work we extend the approach that use vector representation of contexts by transforming them into a high-dimensional space of Wikipedia's concepts. According to the formal definition explained in the Section 2.1 we use the cosine-based similarity measure computed from the transformed vector representations of the ambiguous document fragment and each document from the set of possible meanings. The documents which contain possible meanings are then sorted according to the similarity with the source document. We assume that the most suitable meaning will yield the highest similarity score.

3.1 Semantic Space Construction

The construction of ESA based semantic space is possible from the Wikipedia dumps available for free download. For the processing we need full text representation of the articles, thus the original Wikipedia dump is preferred over it's derivatives (e.g. dbpedia). We preprocess the document, remove all non-textual markup, stub articles, stopwords and meta articles, leaving only raw text. This result is then used to create a term-document matrix, counting the weighted inverted index for each term found in the dataset and building weighted list of Wikipedia concepts. We use airhead-research Java package [10] along with Apache Lucene to compute and represent the semantic space model. Additioanlly, we exclude any articles shorter than 250 terms in order to remove any incomplete or non-informative articles.

3.2 Extracting Possible Meanings

We use the rich semantics already present in the Wikipedia in order to retrieve all pages with possible meaning for the currently processed entity. Each ambiguous entity in the Wikipedia have designated *Disambiguation page*, where each meaning for the given entity does have a link to the full article. For instance, the entity *Jaguar* have more than 25 meanings defined in the disambiguation page to date. While disambiguating the entity *Jaguar*, we retrieve the Wikipedia articles such as *Jaguar_(car)*, *Jaguar*, *Jaguar_(band)*, etc.

Additionally, we use the link structure to obtain all additional redirects to any other possible meaning, extending the set of documents with many additional meanings. The link structure helps us to solve the problem, where the same entity can be referred with various *surface forms*. For instance, the firm *IBM* can be referred in text as *International Business Machines*, *IBM* or even *Big Blue*.

The link structure and the set of internal redirects help us find all additional meanings, even if the text uses alternative surface forms. Many of the Wikipedia links are in the bracket format ([[Article Name |SurfaceForm]]) which allows us easy extraction of the Surface Form - Article Name pairs and use them later in the process. Finally, we use the *redirect pages* to extract the additional mappings and to further extend the extracted list.

3.3 Computing similarity

First, we transform each document (source q, and each possible meaning $d \in D$) into a high-dimensional space extracted from the Wikipedia and represent them as document vectors Q and C_d for the source and extracted concepts, respectively.

At this point, we describe how the documents are transformed into the concept space. Firstly, we use common text pruning techniques to remove any stopwords, markup and other elements. Then, for each term in each document we look up the appropriate vector from the semantic space. The final document vector is then computed as a running total of intermediate vectors computed while creating the semantic space.

Each vector then holds the information about *relatedness* of the document sith each of the concepts. Then, we use the cosine similarity to compute the similarity score between the source vector and each of the meanings. The cosine similarity is computed as follows.

$$rank(q,d) = \frac{Q \cdot C_d}{\|Q\|\|C_d\|} = \frac{\sum_{i=1}^{n} Q_i \times C_{di}}{\sqrt{\sum_{i=1}^{n} (Q_i)^2} \times \sqrt{\sum_{i=1}^{n} (C_{di})^2}}$$

Finally, we sort the documents according to the score given by ranking function, the most similar article should be the correct meaning for disambiguated entity.

4. Evaluation and Results

The evaluation of our proposed method requires an appropriate dataset to test the disambiguation component. However, most of the authors of similar papers use custom evaluation corpora, compiled from various sources and varying in size. There is common practice to use news articles, as those contain large amount of ambiguous proper names. We compiled a development and two test corpora. Each corpus consist of 20 randomly chosen news articles varying in topic. For each document an human annotator checked the results of the NER component and manually disambiguated all the entities. The results of the proposed method were then compared with the human annotator, the precision mentioned later in this section is thereby defined as the "Correlation with human annotator".

In order to compare the results we implemented the baseline system described in [3] and compared results of the proposed method with the baseline score. We also compared the results of LSA in comparison with our method. During the error analysis of preliminal evaluation we discovered that correct meanings are ranked high, but fail to rank first because of other dominant topic present in the analysed document.

If we consider an article about the band Texas, the vectorial representation favors concepts about music, entertainment, television etc. However, if an another entity (i.e. London) is present in the text, the representation assumes the music-based context for this entity as well. We partially overcame this problem with construction of two separate vectors. The first vector is constructed as described earlier, while the second contain only the text within a sliding window around occurences of disambiguated entity in the text. We evaluate the impact if this modification and present the results in tables 1 and 2.

Table 1: Evaluation results - ESA

dataset	ESA - Article	ESA - Combined
devel	89,31	**91,93**
eval-1	87,84	**90,25**
eval-2	85,06	**86,56**

Table 2: LSA and Baseline

dataset	LSA	Baseline
devel	81,41	76,28
eval-1	81,05	85,36
eval-2	82,33	82,37

5. Conclusion and Future Work

In our paper we present an approach to automated named entity disambiguation based on the Wikipedia data. We use explicit semantics already defined in the Wikipedia to retrieve all possible disambiguations for the entities. Additionally, we leverage the semantics to create a high-dimensional word model to compute the similarity between the documents based on human created concepts defined in the Encyclopedia. Our method use an existing named entity recogniser as preprocessor, therefore no human annotation of unknown text is necessary.

Our proposed method can be used as a part of an automatic machine aided document analysis platform such as [2] for the purpose of knowledge gathering, personalization and user modeling. The disambiguated named entities in an unstructured text can be also used as a metadata for recommenders, to increase text search relevance or many others.

The evaluation of our method shows slightly better results than the traditionally used Latent Semantic Analysis as well as the baseline system. As the future work we plan to extend the method with an contextual classifier.

Such classifier will take into acount already disambiguated entities in the document as opposed to individual disambiguiation of the entites. Such contextual awareness can be of great help to resolve cases when an entity has been successfully disambiguated earlier in the text (expressed with different surface form) but fails to rank appropriately later on.

Additionally, we experiment with various scenarios of the context generation to solve the outstanding issue when the occurrences of the same surface form in the given document have two distinct meanings. Currently, all such occurrences are merged into one, most probable meaning. Our final goal is to prepare a web-based service to allow integration of the proposed method with other projects.

Acknowledgements. This work was partially supported by the Scientific Grant Agency of Slovak Republic, grant No. VG1/0675/11.

References

[1] A. Bagga and B. Baldwin. Entity-based cross-document coreferencing using the Vector Space Model. In *Proceedings of the 36th annual meeting on Association for Computational Linguistics*, pages 79–85, Morristown, NJ, USA, 1998. Association for Computational Linguistics.

[2] M. Barla. Towards Social-based User Modeling and Personalization. *Information Sciences and Technologies Bulletin of the ACM Slovakia*, 3(1):52–60, March 2011.

[3] R. Bunescu and M. Pasca. Using encyclopedic knowledge for named entity disambiguation. In *Proceedings of EACL*, pages 9–16, 2006.

[4] N. Chinchor and L. Hirschman. MUC-7 Coreference Task Definition (version 3.0). In *Proceedings of Message Understanding Conference MUC7*, 1997.

[5] S. Coates-Stephens. The analysis and acquisition of proper names for the understanding of free text. *Computers and the Humanities*, 26:441–456, 1993.

[6] S. Cucerzan. Large-scale named entity disambiguation based on Wikipedia data. In *Proceedings of EMNLP-CoNLL*, number 6, pages 708–716, 2007.

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

[8] R. Grishman and B. Sundheim. Message understanding conference-6: A brief history. In *Proceedings of the 16th conference on Computational linguistics (COLING96)*, volume 1, pages 466–471, Copenhagen, Denmark, 1996.

[9] J. Hassell and B. Aleman-Meza. Ontology-driven automatic entity disambiguation in unstructured text. *The Semantic Web-ISWC 2006*, 2006.

[10] D. Jurgens and K. Stevens. The S-Space Package: An Open Source Package for Word Space Models. In *Proceedings of the ACL 2010 System Demonstrations*, pages 30–35. Association for Computational Linguistics, 2010.

[11] T. K. Landauer and P. W. Foltz. An Introduction to Latent Semantic Analysis. *Discourse Processes*, 1(25):259–284, 1998.

[12] G. A. Miller and W. G. Charles. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1–28, 1991.

[13] D. Milne, O. Medelyan, and I. Witten. Mining domain-specific thesauri from wikipedia: A case study. In *IEEE/WIC/ACM International Conference on Web Intelligence, 2006.*, pages 442–448, Hong Kong, Dec. 2006. IEEE Computer Society.

[14] Y. Ravin and Z. Kazi. Is Hillary Rodham Clinton the president?: disambiguating names across documents. In *Proceedings of the Workshop on Coreference and its Applications*, pages 9–16, Maryland, USA, 1999. Association for Computational Linguistics.

[15] A. Woodruff and C. Plaunt. GIPSY: Automated geographic indexing of text documents. *Journal of the American Society for Information Science*, 45(9):645–655, Oct. 1994.