Named Entity Disambiguation for German News Articles

Andreas Lommatzsch, Danuta Ploch, Ernesto William De Luca, Sahin Albayrak
DAI-Labor, TU Berlin, 10587 Berlin, Ernst-Reuter-Platz 7
{andreas.lommatzsch, danuta.ploch, ernesto.deluca, sahin.albayrak}@dai-labor.de

Abstract

Named entity disambiguation has become an important research area providing the basis for improving search engine precision and for enabling semantic search. Current approaches for the named entity disambiguation are usually based on exploiting structured semantic and lingual resources (e.g. WordNet, DBpedia). Unfortunately, each of these resources cover independently from each other insufficient information for the task of named entity disambiguation. On the one hand WordNet comprises a relative small number of named entities while on the other hand DBpedia provides only little context for named entities. Our approach is based on the use of multi-lingual Wikipedia data. We show how the combination of multi-lingual resources can be used for named entity disambiguation. Based on a German and an English document corpus, we evaluate various similarity measures and algorithms for extracting data for named entity disambiguation. We show that the intelligent filtering of context data and the combination of multi-lingual information provides high quality named entity disambiguation results.

1 Introduction

Named entity recognition (NER) and named entity disambiguation (NED) are usually seen as a subtask of information extraction aiming to identify and classify text elements into predefined categories, such as name of persons, organizations, or locations. Ambiguous names are resolved by determining the correct referent. Data provided by NER and NED is required for various applications reaching from semantic search and clustering to automatic summarization and translation. The process of named entity recognition and disambiguation usually consists of three steps:

1. Identify words (or groups of words) representing an entity in a given text
2. Collect data for describing the identified entity in detail (entity properties, the relationship to other entities)
3. Classify the identified entity and calculate which candidate entity (from an external knowledge base) matches best (for resolving ambiguity).

The most complex step in the disambiguation process is the collection of metadata for the identified entity describing the entity context. Many systems use linguistic grammar-based techniques as well as statistical models for NED. Manually created grammar-based systems often obtain a high precision, but show a lower recall and require months of work by linguists. Other systems are based on deploying dictionaries and lexical resources, such as WordNet [Fellbaum1998] or DBpedia [Lehmann *et al.*2009].

In this paper we evaluate how comprehensive, multilingual resources (such as Wikipedia) can be deployed for NED. Based on the context of identified entities and information retrieved from Wikipedia we disambiguate entities and show which parameter configurations provide the best accuracy.

The paper is organized as follows. The next section describes the current state of the art and presents the most popular methods for NED. Subsequently, we introduce our approach. Based on the context of identified entities and data retrieved from Wikipedia we use text similarity measures to perform NED. The approach is analyzed on a collection of German news articles and on the Kulkarni name corpus[1]. The evaluation shows that our approach provides high quality results. Finally we draw a conclusion and give an outlook to future work.

2 Fundamentals

The concept of named entities (NE) was first introduced by the Message Understanding Conferences (MUC) evaluations [Sundheim1996]. The entities were limited to numerical expressions and a few classes of names. However, along with the development of information extraction and search technologies, the categories used for NEs were extended. Nowadays, complex ontologies (such as the Suggested Upper Merged Ontology, containing about 1,100 most general concepts [Pease and Niles2002]) are considered as the basis for named entities.

Current strategies for named entity recognition and disambiguation are mostly based on the use of ontological data and on context analysis. Since corpus based approaches are likely to suffer from the data sparseness problem, large lexical data collections, such as Wikipedia are a good choice to mine concepts and to enrich sparse corpora. Wikipedia is the largest online encyclopedia which provides linked and partially annotated data and descriptive information. Based on the evaluation of articles types (e.g. disambiguation pages, category articles) and the analysis of templates (e.g. info boxes) even semantic knowledge can be extracted. A popular project aiming to provide structured information from Wikipedia is DBpedia [Lehmann *et al.*2009]. Some authors [Cui *et al.*2009] built domain specific taxonomies from Wikipedia by analyzing the URLs and internal links

[1] http://www.d.umn.edu/~pederse/namedata.html

in Wikipedia pages, such as category labels or info boxes. For extracting new facts from Wikipedia Wu & Weld [Wu and Weld2007] suggest a cascade of conditional random field models.

Beside the approaches based on structured, ontological knowledge, there are projects that use unstructured Wikipedia data for NED. Cucerzan [Cucerzan2007] retrieves for words ("surface forms") identified to represent relevant entities all potentially matching Wikipedia articles. The Wikipedia contexts that occur in the document and the category tags are aggregated into a string vector, which is subsequently compared with the Wikipedia entity vector (of categories and contexts) of each possible entity. Then the assignment of entities to surface forms is chosen that maximizes the similarity between the document vector and the entity vectors. Alternative approaches for named entity disambiguation are based on the co-occurring analysis of named entities [Nguyen and Cao2008], the analysis of word based features (e.g. part-of speech, pattern) [Mann and Yarowsky2003] or document meta-information [Nadeau and Sekine2007, Bunescu and Pasca2006].

3 Approach

In our approach we focus on multi-lingual data retrieved from Wikipedia. Our intention is to have a robust approach providing a constantly high disambiguation precision. Thus, we do not rely on semantic annotations in Wikipedia (e.g. info boxes) or DBpedia content due to the fact that this data is not available for all entities.

The developed NED component is part of a project[2] for the semantic clustering of news, implemented using the UIMA framework[3]. The identification of words representing potentially relevant entities is done using DBpedia data. The component searches for surnames in news articles (present in the DBpedia person dataset). The found surnames and the assigned DBpedia entities are used as input data for the disambiguation task.

For performing NED, we analyze various methods to retrieve context data for the potentially matching entities. Due to the fact that we perform the named entity recognition based on DBpedia we focus on Wikipedia as data source. We analyze the following four content extraction strategies:

- We extract the first section of the Wikipedia article and remove the stop words. Usually the first section contains the most important information of the respective article.
- We extract the first section of the Wikipedia article and remove the stop words. The whole text is converted to lower case characters when calculating the string similarity.
- We extract all words with capital letters from the first section of the Wikipedia article. This is done to restrict the content to proper nouns (stop words are removed).
- We extract all words of the (complete) Wikipedia article that link to Wikipedia articles. The idea behind this method is, that the linked words contain the data suitable for the NED.

[2]http://www.dai-lab.de/competence_centers/irml/projekte/spiga/
[3]http://uima.apache.org/

We extract these data from the German as well as from the English Wikipedia. We use language-specific stop word lists and rule based stemmers.

In the next step we calculate the similarity between the retrieved Wikipedia content and the analyzed news article. The similarity calculation is done with the following algorithms:

1. **Jaccard-Similarity:** The Jaccard Similarity calculates the similarity of two word vectors (X, Y) as follows:
$$\text{Jaccard}(X,Y) = \frac{X * Y}{|X|\,|Y| - (X * Y)}$$
where $(X * Y)$ is the inner product of X and Y, and $|X| = \sqrt{X * X}$ the Euclidean norm of X.
2. **Dice-Similarity:** The Dice coefficient is a term based similarity measure whereby the similarity measure is defined as twice the number of terms common to compared entities divided by the total number of terms in both tested entities. The coefficient result of 1 indicates identical vectors as where a 0 equals orthogonal vectors.
$$\text{Dices coefficient}(X,Y) = \frac{2 * \#\text{common terms}}{\#\text{terms(X)} + \#\text{terms(Y)}}$$
3. **Overlap-Coefficient:** The Overlap-Coefficient calculates the similarity based on the common terms:
$$\text{Overlap}(X,Y) = \frac{\#\text{common terms}(X,Y)}{min\left(\#\text{terms}(X), \#\text{terms}(Y)\right)}$$
4. **Weighted Term-Similarity**: Similarity based on the number of matched terms weighted by the term length and the percentage of matched terms.
$$\text{wTSim}(X,Y) = \sqrt{|T|} * |X| * \sum_{t \in T} \left(\#(t \text{in} Y) \left(1 + \log\left(\frac{1}{1 + \#t\,(\text{in}Y)} \right) \right)^2 \text{len}(t) \right)$$
where T is the set of common terms in X and Y and $\text{len}(t)$ the function that calculates the length of term t. The weighted term similarity is often used by search engines for calculating a relevance score.

We calculate for each entity the similarity between the retrieved content and the news article. For the disambiguation task we determine the entity with the highest similarity. Overall, we test 32 different variants (4 content extraction strategies $\times$ 4 similarity measures $\times$ 2 languages). Additionally, we create ensembles combining the results of different languages (based on the CombSUM algorithm [Hull *et al.*1996]).

4 Experiments and evaluation

We evaluate the performance of our approach on a corpus of German news documents as well as on a collection of web search results in English.

4.1 Evaluation on the Kulkarni name corpus

The Kulkarni name corpus was created to evaluate NED algorithms. The corpus contains a set of queries where each query consists of a query string, and a set of approximately 200 documents (retrieved from a search engine for the query string). The query strings represent ambiguous person names. The documents in the result set are manually

annotated with the DBpedia URL of the entity related to the document. We performed the NED as described in section 3 and calculate the accuracy[4] over all queries. We consider only the English Wikipedia as content source since the relevant entities in the Kulkarni name corpus are not present in the German Wikipedia.

The accuracy for the analyzed content extraction strategies and the considered similarity measures are shown in Figure 1.

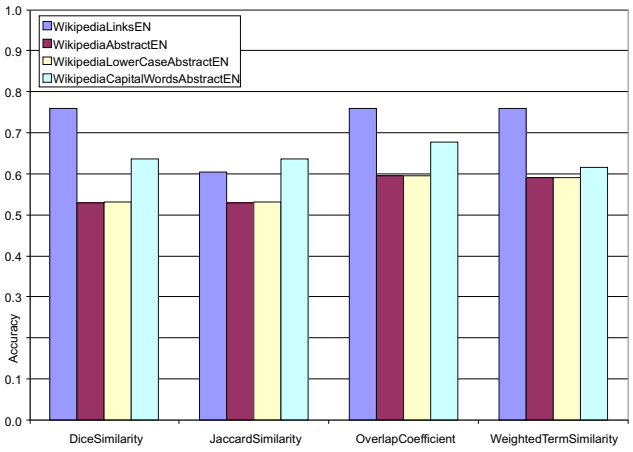

Figure 1. The named entity disambiguation accuracy on the Kulkarni name corpus

The results show that the disambiguation accuracy highly depends on the retrieved Wikipedia content. The best results are obtained if the disambiguation is based on the linked words in the Wikipedia article. The applied similarity measure has only a small influence on the result. Overall, the overlap coefficient and the weighted term similarity provide the best result. Thus, for a further improvement of the disambiguation accuracy the content extraction strategies should be optimized e.g. by considering the different types of Wikipedia links.

4.2 Evaluation on German news articles

Due to the fact, that the focus of our project is to cluster German news documents semantically, we create a new corpus optimized on our scenario. We randomly selected 65 news articles (from January 2009 crawled from various internet news sources, e.g. `Netzzeitung` and `DiePresse.com`) covering the topics politics and sports. Based on the German DBpedia person corpus[5] we identified potentially relevant people (having a surname present in the news document). We manually annotated which of the identified entities are related to the news article. Thus, each corpus element consists of the following data:

- The news document as plain text (⌀390 words)
- The search string (surname of the person), e.g. `Schumacher`
- A list of DBpedia entities having the search string as surname (limited to 10 entities), e.g. `[Michael Schumacher, Brad Schumacher, Anton Schumacher, Heinrich Christian ...]`

For the disambiguation, we considered the German and the English Wikipedia. The disambiguation accuracy for the considered content extraction strategies and the similarity measures are shown in Figure 2.

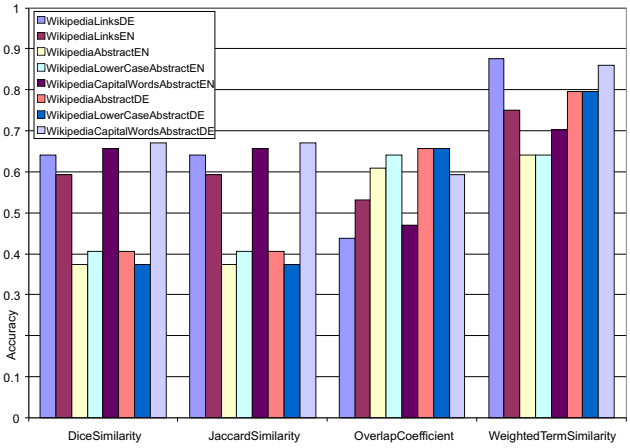

Figure 2. The named entity disambiguation accuracy on the German news corpus.

The evaluation shows that the highest accuracy is achieved when using the weighed term similarity and the linked words in the German Wikipedia page as reference content. In the analyzed scenario the similarity measure has a higher impact on the accuracy than the content extraction strategy. As expected, the accuracy achieved using the German Wikipedia content is always above the accuracy achieved based on the English content (since we analyzed German news). Nevertheless, the accuracy based on the analysis of the linked words in the English Wikipedia provided relatively good results. Comparing the dependencies between the accuracy from the content extraction strategies, the evaluation shows that good results are obtained if only relevant terms from the content source are filtered.

4.3 Multi-lingual ensembles

We analyze how the disambiguation accuracy can be improved by combining German and English content extraction strategies. Based on the strategy CombSUM we build ensembles combining the strategies discussed in section 4.2. The ensembles are created incrementally adding in each step the strategy to the ensemble that enables the best accuracy improvement. The weights for each strategy in the ensemble are calculated using an optimization algorithm based on genetic algorithms. The evaluation (Figure 3) shows that the weighted combination of strategies extracting data from the German and the English Wikipedia can improve the disambiguation accuracy by 13%. If English Wikipedia articles are used for the NED in German news articles the similarity is implicitly restricted to proper nouns (such as person or location names) that are not translated. Even though the strategies based only on the English Wikipedia content do not show a high disambiguation accuracy, in a combination with other strategies they improve the accuracy.

5 Conclusion and future work

We analyzed how multi-lingual content retrieved from Wikipedia can be used for NED. For evaluating the algorithms on German and English documents, we created a corpus of German news documents. Based on an English and a German corpus we analyzed various string similarity

[4] $\text{accuracy} = \frac{\text{number of correctly assigned entities}}{\text{total number of entities}}$

[5] http://downloads.dbpedia.org/3.5.1/de/persondata_de.nq.bz2

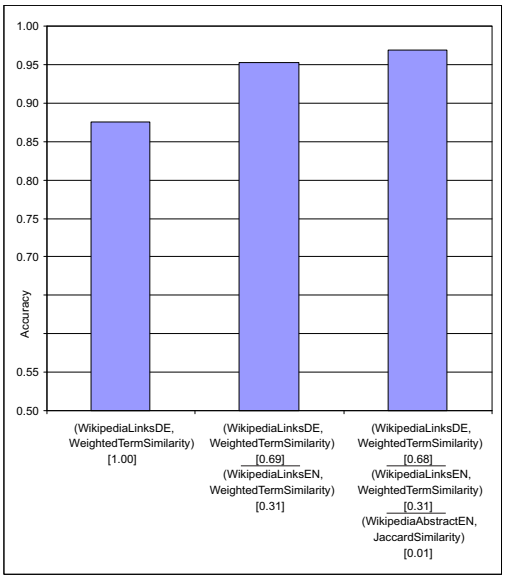

Figure 3. The disambiguation accuracy of the analyzed ensembles on the German news corpus.

measures and several strategies for content extraction. The evaluation results show, that in the chosen scenarios a high disambiguation accuracy can be obtained. The accuracy highly depends on the content extraction strategy. In general, the more complex extraction strategies provide better results.

Moreover, we showed that the combination of multilingual data improves the accuracy. The combination of German and English content extraction strategies improves the accuracy up to 13%. The combination of German and English content gives proper nouns a higher weight what results in a better disambiguation accuracy. A deeper analysis how the used parameter settings and the article language influence the NED accuracy will be done in the near future.

As future work we will take a deeper look on the relationships of the relevant named entities. The goal is to integrate semantic data sources (such as Freebase[6], or YAGO[7]) and to combine the data with multi-lingual Wikipedia data. We want to analyze the semantic relationships between the entities and learn weights for all relevant types of connections between entities. The learned weighs are used to adapt the similarity measures to the context and the respective domains. Moreover, in the project we will focus on ensemble learning algorithms [Polikar2006] (such as preference learning [Tsai *et al.*2007] and boosting strategies [Freund and Schapire1996]) to combine multi-lingual data and various features.

References

[Bunescu and Pasca, 2006] Razvan C. Bunescu and Marius Pasca. Using encyclopedic knowledge for named entity disambiguation. In *Proc. of the 11th Conference of the EACL)*. The Assn. for Computer Linguistics, 2006.

[Cucerzan, 2007] Silviu Cucerzan. Large-scale named entity disambiguation based on Wikipedia data. In *Proc. of the 2007 Joint Conf. on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 708–716, Prague, Czech Republic, June 2007. Assn. for Computational Linguistics.

[Cui *et al.*, 2009] Gaoying Cui, Qin Lu, Wenjie Li, and Yirong Chen. Mining concepts from wikipedia for ontology construction. In *Proc. of the 2009 Intl. Joint Conf. on Web Intelligence and Intelligent Agent Technology WI-IAT '09*, pages 287–290, Washington, DC, USA, 2009. IEEE Computer Society.

[Fellbaum, 1998] C. Fellbaum. *WordNet: An Electronical Lexical Database*. The MIT Press, Cambridge, MA, 1998.

[Freund and Schapire, 1996] Yoav Freund and Robert E. Schapire. Experiments with a new boosting algorithm. In *Intl. Conference on Machine Learning*, pages 148–156, 1996.

[Hull *et al.*, 1996] David A. Hull, Jan O. Pedersen, and Hinrich Schütze. Method combination for document filtering. In *SIGIR '96: Proc. of the 19th ACM SIGIR conf. on Research and development in information retrieval*, pages 279–287, New York, USA, 1996. ACM Press.

[Lehmann *et al.*, 2009] Jens Lehmann, Chris Bizer, Georgi Kobilarov, Sren Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. DBpedia - a crystallization point for the web of data. *Journal of Web Semantics*, 7(3):154–165, 2009.

[Mann and Yarowsky, 2003] Gideon S. Mann and David Yarowsky. Unsupervised personal name disambiguation. In *Proc. of the 7th conference on Natural language learning at HLT-NAACL 2003*, volume Volume 4, pages 33 – 40, Edmonton, Canada, 2003. Assn. for Computational Linguistics Morristown, NJ, USA.

[Nadeau and Sekine, 2007] David Nadeau and Satoshi Sekine. A survey of named entity recognition and classification. *Linguisticae Investigationes*, 30(1):3–26, January 2007. Publisher: John Benjamins Publishing Co.

[Nguyen and Cao, 2008] H.T. Nguyen and T.H. Cao. Named entity disambiguation on an ontology enriched by wikipedia. In *Research, Innovation and Vision for the Future*, pages 247 –254, 2008.

[Pease and Niles, 2002] Adam Pease and Ian Niles. Ieee standard upper ontology: a progress report. *Knowl. Eng. Rev.*, 17(1):pages 65–70, 2002.

[Polikar, 2006] Robi Polikar. Ensemble based systems in decision making. *Circuits and Systems Magazine, IEEE*, 6(3):21 – 45, 9 2006. 1531-636X.

[Sundheim, 1996] Beth M. Sundheim. The message understanding conferences. In *Proc. of the TIPSTER Text Program: Phase II*, pages 35–37, Vienna, VA, USA, May 1996. Assn. for Computational Linguistics.

[Tsai *et al.*, 2007] Ming-Feng Tsai, Tie-Yan Liu, Tao Qin, Hsin-Hsi Chen, and Wei-Ying Ma. FRank: a ranking method with fidelity loss. In *SIGIR '07: Proc. of the 30th ACM SIGIR conf. on Research and development in information retrieval*, pages 383–390, New York, USA, 2007. ACM.

[Wu and Weld, 2007] Fei Wu and Daniel S. Weld. Autonomously semantifying wikipedia. In *CIKM '07: Proc. of the 16th ACM conf. on information and knowledge management*, pages 41–50, New York, USA, 2007. ACM.

[6] http://www.freebase.com/

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