Named Entity Disambiguation using Freebase and Syntactic Parsing

Kamel Nebhi

LATL, Department of linguistics
University of Geneva
Switzerland
kamel.nebhi@unige.ch

Abstract. Named Entity Disambiguation (NED) is a fundamental task of semantic annotation for the Semantic Web. The task of Word Sense Disambiguation (WSD) in Ontology-Based Information Extraction (OBIE) aims to establish a link between the textual entity mention and the corresponding class in the ontology. In this paper, we propose a NED process integrated in a rule-based OBIE system for French. We show that our SVM approach can improve disambiguation efficiency using syntactic features provided by the Fips parser and popularity score features extracted from the Freebase knowledge base.

Keywords: Named Entity Disambiguation, Syntactic Parsing, Linked Open Data

1 Introduction

The realization of the Web of data on a large scale implies the widespread annotation of Web documents with ontology base knowledge markup [20]. To encourage the emergence of solutions, OBIE seems to be a mature NLP technology to introduce supplementary information and knowledge into a document.

OBIE has been conceived only a few years ago and has recently emerged as a subfield of Information Extraction (IE). OBIE is different from traditional IE because it finds type of extracted entity by linking it to its semantic description in a formal ontology. The main difficulty of OBIE is the disambiguation process, which identifies the meaning of a word when it has multiple senses. For example, the string "Washington" is used to refer to more than 90 different NE in DBpedia database.

In this paper, we present a Named Entity Disambiguation (NED) process integrated in a rule-based OBIE system for French. We show that our SVM approach can improve disambiguation efficiency using syntactic features provided by the Fips parser and popularity score features extracted from the Freebase knowledge base.

This paper is divided as follows. In Section 2 we describe related work to Word Sense Disambiguation. Then, we present our approach in Section 3. Next, we show our experimental setup for testing in Section 4. Finally, we summarize the paper.

2 Related Work

In 1949, WSD was introduced as a fundamental task of machine translation [18]. WSD consists in determining which sense of a word is used when it appears in a particular context. In recent years, WSD is also an intermediate task in several NLP applications as information retrieval [17] or information extraction [5].

For OBIE, the task of WSD aims to establish a link between the textual entity mention and the corresponding class in the ontology. The researches on WSD for IE task are divided on two main approaches: unsupervised approach and supervised approach. Recent works generally use additional resources as knowledge bases to improve traditional methods.

WSD unsupervised approach for IE uses clustering word occurrences to induce word senses and do not exploit any manually sense-tagged corpus. One of the first works in the domain of disambiguation of named entity was in 1998 by [1]. They create context vectors for each occurrence of a name. Next, they compute the similarity among names using the cosine measure. The results were encouraging with an F-Measure of 84 percent. In 2003, the KIM semantic annotation platform [16] uses a basic rule-based approach without knowledge resources to disambiguate named entity. The IE system obtains an F-Measure of 84 percent. Recently, in 2011, [6] proposes an approach based on a resource of contextual words called *LinkedData Inferface* (LDI). The disambiguation task uses the *Semantic Disambiguation Algorithm* (SDA), which identify the item in the LDI that is most similar to the context of the named entity. The system shows very promising results with 90 percent in French and 86 percent in English.

WSD supervised approach for IE uses machine-learning techniques to learn a classifier from manually sense-annotated data sets. In 2006, [4] resolves entities using an SVM kernel trained on Wikipedia. This approach generates a ranked list of plausible entities. Experimental results show that the method improves accuracy. In 2007, [8] suggests a disambiguation process based on information extracted from Wikipedia and Web search results. They use a similarity measure and richer features for the similarity comparison. The IE system obtains a precision of 88 percent. A few years later, in 2009, [15] uses a kernel classifier to determine the right matching entity in Wikipedia, which is considered as an unambiguous reference. The model is trained on 100000 Wikipedia articles for German entity disambiguation. The F-Measure of the system was reported to be about 80 percent. In 2011, [11] presents a robust disambiguation approach using knowledge bases like YAGO and several measures, such as popularity prior and keyphrase based similarity, to build a weighted graph. To improve this method, they also consider the syntactic context of the mention using a large corpus for training. The approach obtains good results with a precision of 81 percent. Nevertheless, the more advanced configurations of the system did not use syntactic context.

3 Approach

The task is divided into three steps in our OBIE system. In this section, we describe the different steps of the proposed system.

3.1 Ontology-based named entity recognition

In the first step, we use an OBIE system for French using a rule-based approach to recognize entities in text. Our application [14] is built on GATE [9] to annotate entities and relate them to the DBpedia ontology[1]. The GATE application consists of a set of processing resources, executed in a pipeline over a corpus of news articles. The pipeline consists of 5 parts: linguistic pre-processing, gazetteers, rule-based semantic annotation[2] and final RDF output.

3.2 Entity Candidate

LOD refers to data published with a number of best practices based on W3C standards for publishing and connecting structured data on the Web [2]. In the past few years, we have assisted to a growth in LOD publishing on the Web, leading to 31 billion RDF triples published online. In this context, using these resources as complementary information can enhance IE [7].

In the second step, we use Freebase[3] [3] to obtain *entity candidate*. For each entity detected by our OBIE system, we request the Freebase Suggest API to have all possible entities for a surface form.

3.3 Named Entity Disambiguation

The NED task can be modeled as a classification problem. In the third step, a named entity disambiguation SVM classifier is trained using syntactic and popularity score features.

Popularity score features. This score is provided by Freebase Search API. Popularity score of entities can be seen as a probabilistic estimation based on Wikipedia frequencies in link anchor. The Freebase Search API allows access to Freebase data given a text query. A large number of filter constraints are supported to better aim the search at the entities being looked for. For example, the highest ranked result for the query “Washington” is the city “Washington,D.C.”.

Syntactic features. Syntactic Features (SF) are commonly used in WSD classification [10] because they are much richer in information than bag-of-words features[4]. The syntactic parser Fips [19] was used to annotate the training data.

[1] We used DBpedia interlinking to resolve ontological differences between the schema provided by Freebase and the DBpedia ontology.

[2] The grammar rules are written in a language called JAPE which is a finite state transducer.

[3] Freebase is a collaborative knowledge base that contains structured data of almost 23 million entities.

[4] Bag-of-words features only specify words in symmetric window around target ignoring position.

Fips was developed over the last decade in our laboratory, LATL. It is a deep symbolic parser based on generative grammar concepts for its linguistic component and object-oriented design for its implementation. The parser uses a bottom up parsing algorithm with parallel treatment of alternatives, as well as heuristics to rank alternatives.

As remarked by [12], incorporating syntactic information is a strong clue about the interpretation of a word. Our algorithm uses the context around the surface forms. For this, we construct a context from all target entities in news articles using syntactic information based on dependency structures provided by Fips. From the output of the parser, we extract syntactic (binary) relations between constituents (subject-verb, verb-object, verb-complement, etc.). For example, from "Hollande souhaite diminuer le nombre de députés" a subject-verb relation is extracted between "Hollande" and "souhaite". The SF also includes information on each of its words such as part-of-speech and lemma.

Algorithm. The input of the algorithm is a set of ambiguous entities. For each ambiguous entity we find a set of candidate entities. Then, we use popularity scores and syntactic information as features to train an SVM classifier. Table 1 shows examples of NED for a sentence of *LeMonde.fr*.

News	John Kerry a déclaré que Washington prévoyait de fournir une aide à l'opposition syrienne.[5]
Extracted Mentions	John Kerry (senator), Washington (city)
Candidate Entities	{John Kerry (senator), John Kerry (author)}, {Georges Washington (president), Washington (city), Denzel Washington (actor))}

Table 1: Example of Extraction

4 Experiments

4.1 Data set

The data set we use for our experiments is a collection of 1100 news articles[6] extracted from *LeMonde.fr*. We designate 2/3 of the data as the training set and the remaining 1/3 as the test set. For the experiments, we trained our model with an SVM classifier on 735 news articles.

[5] John Kerry said Washington planned to provide assistance to the Syrian opposition.

[6] This set contains approximately 20 named entities/news article on average.

To build the training set, we used a semi-automatic method. To start, the NE detection task was automatically done by our ontology-based named entity recognition system (cf. section 3.1). In addition, we removed or corrected manually the wrong semantic links to the DBpedia ontology classes. Finally, we used the parser Fips and the Freebase Search API to produce syntactic and popularity score features.

4.2 Results

We evaluate our system on a set of 365 test articles[7]. Traditional measures like Precision, Recall and F-Measure are inadequate when dealing with ontologies, thus we used the Balanced Distance Metric (BDM) which is useful to measure performance of OBIE systems taking into account ontological similarity [13].

	OBIE without disambiguation	Popularity Score	Popularity Score + Syntactic Parsing
F_1	0.65	0.70	0.86
BDM_F_1	0.69	0.73	0.90

Table 2: Experimental results

In table 2, the OBIE system without disambiguation process achieved a traditional F-Measure of 65% and an augmented F-Measure of 69%. Adding a disambiguation process based on popularity score does not improve a lot F-Measure with 70% and 73%. Adding the entire disambiguation layer improves extraction effectiveness, traditional F-Measure rises to 86% and augmented F-Measure rises to 90%.

5 Conclusions

In this paper, we presented a NED process integrated in an OBIE system for French using a rule-based approach. Our named entity disambiguation SVM classifier is trained using syntactic and popularity score features. As our evaluation shows, this method of extraction performed significantly better using the entire disambiguation features.

In our future work, we aim to incorporate more knowledge from LOD and we'll integrate the application into a complete annotation pipeline for the Semantic Web.

[7] For the evaluation, we only use Person, Organization and Location named entity categories.

References

1. Bagga, A., Baldwin, B.: Entity-based cross-document coreferencing using the vector space model. In Proceedings of the 36th ACL and 17th ICCL, Montreal, Canada, 1998.
2. Bizer, C., Heath, T., Berners-Lee, T.: Linked Data - The Story So Far. International Journal on Semantic Web and Information Systems, 2009.
3. Bollacker, K., Evans, C., Paritosh, P., Sturge, T., Taylor, J.: Freebase: a collaboratively created graph database for structuring human knowledge. In Proceedings of the 2008 ACM SIGMOD, Vancouver, Canada, 2008.
4. Bunescu, R. C., Pasca, M.: Using encyclopedic knowledge for named entity disambiguation. In EACL. ACL, 2006.
5. Chai, J. Y., Biermann, A. W.: The use of word sense disambiguation in an information extraction system. In Proceedings of the 16th National Conference on Artificial Intelligence, Orlando, USA, 1999.
6. Charton, E., Gagnon, M., Ozell., B.: Automatic semantic web annotation of named entities. In Canadian Conference on AI, St. John's, Canada, 2011.
7. Ciravegna, F., Gentile, A. L., Zhang, Z.: LODIE: Linked Open Data for Web-scale Information Extraction. SWAIE: 11-22, 2012.
8. Cucerzan, S.: Large-scale named entity disambiguation based on Wikipedia data. In Proceedings of the 2007 EMNLP and CoNLL, Prague, Czech Republic, 2007.
9. Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V., Aswani, N., Roberts, I., Gorrell, G., Funk, A., Angus, R., Damljanovic, D., Heitz, T., Greenwood, M. A., Saggion, H., Petrak, J., Li, Y., Wim, P.: Text Processing with GATE (Version 6). University of Sheffield, 2011.
10. Lin, D.: Using syntactic dependency as local context to resolve word sense ambiguity. In Proceedings of ACL '98, Madrid, Spain, 1998.
11. Hoffart, J., Yosef, M. A., Bordino, I., Fürstenau, H., Pinkal, M., Spaniol, M., Taneva, B., Thater, S., Weikum, G.: Robust disambiguation of named entities in text. In Proceedings of EMNLP '11, Edinburgh, United Kingdom, 2011.
12. Levin, B.: Lexical Semantics in Review. Lexicon Project Working Papers. MIT Press, 1985.
13. Maynard, D., Peters, W., Li, Y.: Evaluating evaluation metrics for ontology-based applications: Infinite reflection. In LREC, Marrakech, Morocco, 2008.
14. Nebhi, K.: Ontology-Based Information Extraction for French Newspaper Articles. In KI 2012, Saarbrücken, Germany, 2012.
15. Pilz, A., Paaß, G.: Named entity resolution using automatically extracted semantic information. In KDML 2009 : Workshop on KDML, pages 84 – 91, 2009.
16. Popov, B., Kiryakov, A., Kirilov, A., Manov, D., Goranov, O. M.: Semantic annotation platform. In Proceedings of the 2nd Intl. Semantic Web Conf. 2003, pages 834–849. Springer, 2003.
17. Schütze, H., Pedersen, J. O.: Information retrieval based on word senses. In Proceedings of the 4th Annual Symposium on Document Analysis and Information Retrieval, pages 161–175, 1995.
18. Weaver, W.: Translation. In W. N. Locke and A. D. Boothe, editors, Machine Translation of Languages, pages 15–23. MIT Press, Cambridge, MA, 1949/1955.
19. Wehrli, E.: Fips, a deep linguistic multilingual parser: In ACL 2007 Workshop on deep Linguistic Processing, Prague, Czech Republic, 2007.
20. Wilks, Y., Brewster, C.: Natural language processing as a foundation of the semantic web. Foundations and Trends in Web Science, 1(3-4):199–327, 2009.