Automatic Semantic Web annotation of named entities

Eric Charton[1], Michel Gagnon[1], and Benoit Ozell[1]

École Polytechnique de Montréal, Montréal, H3T 1J4 , Québec (Canada),
{eric.charton, michel.gagnon, benoit.ozell}@polymtl.ca

Abstract. This paper describes a method to perform automated semantic annotation of named entities contained in large corpora. The semantic annotation is made in the context of the Semantic Web. The method is based on an algorithm that compares the set of words that appear before and after the name entity with the content of Wikipedia articles, and identifies the more relevant one by means of a similarity measure. It then uses the link that exists between the selected Wikipedia entry and the corresponding RDF description in the Linked Data project to establish a connection between the named entity and some URI in the Semantic Web. We present our system, discuss its architecture, and describe an algorithm dedicated to ontological disambiguation of named entities contained in large-scale corpora. We evaluate the algorithm, and present our results.

1 Introduction

Semantic Web is a web of data. This web of data is constructed with documents that are, unlike HTML files, RDF[1] assertions establishing links between facts and things. RDF documents, like HTML documents, are accessible through URI[2]. A set of best practices for publishing and connecting RDF semantic data on the Web is referred by the term *Linked Data*. An increasing number of data providers have delivered *Linked Data* documents over the last three years, leading to the creation of a global data space containing billions of RDF assertions. For the usability of the Semantic Web, a new breed of smarter applications must become available. To encourage the emergence of such innovative softwares, we need NLP solutions that can effectively establish a link between documents and Semantic Web data. The 20 billions RDF triples currently available on the Semantic Web [3] makes this problem both formidable and acute. In this paper, we propose a general schema of automatic annotation, using disambiguation resources and algorithms, to establish relations between named entities in a text

[1] Resource Description Framework, is an official W3C Semantic Web specification for metadata models.

[2] Uniform Resource Identifier (URI) is the name of the string of characters used to identify a resource on the Internet.

[3] According to http://esw.w3.org/TaskForces/CommunityProjects/LinkingOpenData/DataSets.

and the ontological standardized semantic content of the *Linked Data* network.

This article is structured as follows: section 2 investigates the annotation task problem from a broad perspective and describes the features of semantic annotation task in the context of Semantic Web; section 3 describes the proposed system architecture and its implementation. In section 4 we present the experiment and corpora on which the evaluation has been done. Finally, section 5 comments the results obtained by our system. We conclude and give some perspectives in section 6.

2 Problem description

The basic principle of annotations is to add information to a source text. In a computer perspective, annotations can take various forms, but their function is always the same: to introduce complementary information and knowledge into a document. Two main kinds of information can be attributed to a word or a group of words by an annotation process : a fixed class label defined by a taxonomy standard or a link to some external knowledge.

A class description can be assigned to a word or a group of words called a *Named Entity (NE)*. By class, we mean a label describing the nature of the object expressed by the words. This object can be, for example, a person, an organization, a product, or a location. Attribution of such class is the *Named Entity Recognition (NER)* task, widely investigated ([2, 1, 11]).The granularity of classes contained in a NE taxonomy can be highly variable ([14]) but strictly, NER task is a classification task, whose purpose is to assign to a sequence of words a unique class label. Label will be for example *PERS* to describe a person, or *ORG* for an organization, and so on. This means that NER task is unable to introduce any more complementary information into the text. It is possible to introduce an upper level of granularity in the NE taxonomy model (for example, we can distinguish two kinds of places, *LOC.ADMI* for a city and *LOC.GEO* for a National Park) but with strong limitations. Thus, there is no way to introduce data like birth date of a person or ground surface of a city.

To achieve this task of associating properties to NE, an upper level of annotation is needed, expressed by a relation between NE and an external knowledge. It consists in assigning to an identified NE a link to a structured external knowledge base, like the one delivered on the Semantic Web. This is the *Semantic Annotation (SA)* task, previously investigated by ([10, 7]).

2.1 Entity labeling versus Semantic labeling

The example in Figures 2 and Table 2 illustrates the difference between SA and NER and its implication on knowledge management. Let's consider a sample text to annotate, as presented in Table 1.

The first level of ambiguity encountered by the NER task is related to the words polysemy. To illustrate this we show in Figure 1 the numerous possible

Paris is a town in Oneida County, New York, USA. The town is in the southeast part of the county and is south of Utica. The population was 4,609 at the 2000 census. The town was named after an early benefactor, Colonel Isaac Paris.

Table 1. A sample document to label with various named entities contained in.

Fig. 1. Ambiguity of a class label for a named entity like **Paris**. It can be a city, and asteroid, a movie, a music album or a boat.

concept-class values available for the *Paris* word. The main objective of the NER task is to manage this first level of disambiguation, generally through statistical methods ([2], [8], [12]). The NER task results in a text where NE are labeled by classes, as presented in Table 2. But despite the NE labeling process, we can show that a level of ambiguity is still present. *Paris* is correctly annotated with the *LOC* (locality) class label, but this class is not sufficient to determine precisely which locality it is, according to the numerous existing cities that are also named Paris (Figure 2).

Paris{LOC} is a town in **Oneida County**{LOC}, **New York**{LOC}, **USA**{LOC}. The town is in the southeast part of the county and is south of **Utica**{LOC}. The population was **4,609**{AMOUNT} at the **2000 census**{DATE}. The town was named after an early benefactor, **Colonel Isaac Paris**{PERS}.

Table 2. Sample of word with standard NE labels in the document.

2.2 Previous Semantic labeling propositions

The task of SA has received an increasing attention in the last few years. A general survey of all the semantic annotation techniques have been proposed by ([16]). None of the described systems have been integrated in the general schema of Semantic Web. They are all related to specific and proprietary or non-standard ontological representations. The KIM platform ([10]) provides a two-step labeling process including a NER step to attribute NE labels to words before establishing the semantic link. The semantic descriptions of entities and relations between them are kept in a knowledge base encoded in the KIM ontology and resides in the same "semantic repository". SemTag ([5]) is another example of a tool that focuses only on automatic mark-up. It is based on IBM's text analysis platform

Fig. 2. Ambiguity of entity for a same NE class label: the *Paris* word, even with its Location class, is still ambiguous.

Seeker and uses similarity functions to recognize entities that occur in contexts similar to marked up examples. The key problem with large-scale automatic mark-up is ambiguity. A Taxonomy Based Disambiguation (TBD) algorithm is proposed to tackle this problem. SemTag can be viewed as a bootstrapping solution to get a semantically tagged collection off the ground. Recently, ([9]) presented Moat, a proposition to bridge the gap between tagging and *Linked Data*. Its goal is to provide a simple and collaborative way to annotate content thanks to existing URI with as little effort as possible and by keeping free-tagging habits. However, Moat does not provide an automatic generic solution to establish a link between text and an entry point in the *Linked Data* Network.

2.3 The Word sense disambiguation problem

The problem with those previous propositions is related to the *Word Sense Disambiguation (WSD)*. WSD consists in determining which sense of a word is used when it appears in a particular context. KIM and Semtag, when they establish a link between a labeled NE and an ontology instance, need a complementary knowledge resource to deal with the homonymic NEs of a same class. For the NER task, this resource can be generic and generative: a labeled corpus used to train a statistical labeling tool (CRF, SVM, HMM). This statistical NER tool will be able to infer a class proposition through its training from a limited set of contexts. But this generative approach is not applicable to the SA task, as each NE to link to a semantic description has a specific word context, marker of its exact identity. Many propositions have been done to solve this problem. Recently, ([17]) suggest to use the LSA[4] techniques mixed with cosine similarity measure to disambiguate terms in the perspective of establishing a semantic link. The Kim system ([10]) re-uses the Gate platform and its NLP components and apply rules to establish a disambiguated link. Semtag uses two kinds of similarity functions: bayesian, and cosinus. But the remaining problem for all those propositions is the lack of access to an exhaustive and wide knowledge of contextual information related to the identity of the NE. For our previous *Paris* example, those systems could establish a disambiguated link between any *Paris* NE and its exact *Linked Data* representation only if they have access to

[4] Latent Semantic Analysis is a technique of analyzing relationships between a set of documents and terms using term-document matrix built from Singular Value Decomposition.

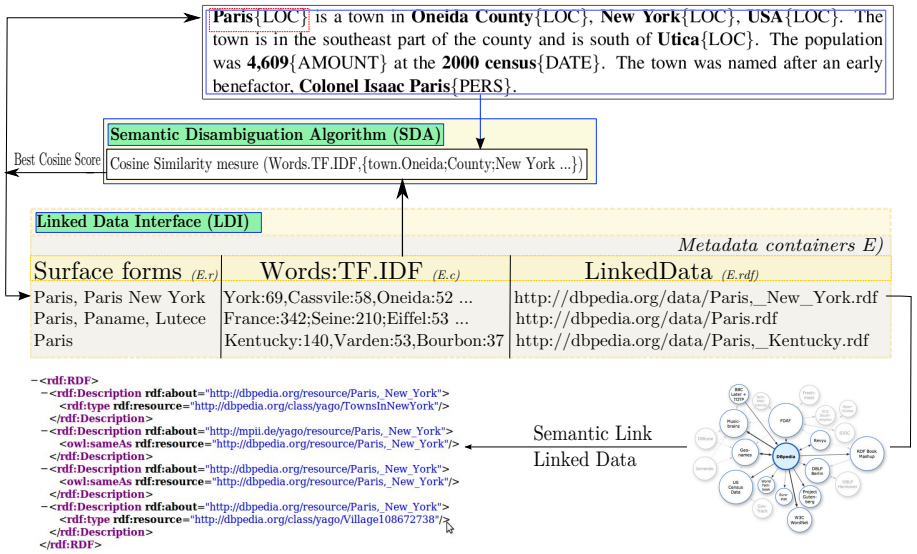

Fig. 3. Architecture of the system with *metadata* used as Linked Data Interface (LDI) and Semantic Disambiguation Algorithm (SDI).

an individual usual word contextual modelized resource. Unfortunately, such a knowledge is not present in RDF triples of the LinkedData network, neither in standard exhaustive ontologies like DBPedia.

3 Our proposition: a Linked Data Interface

To solve this problem, we propose a SA system that uses an intermediate *structure* to determine the exact semantic relation between a NE and its ontological representation on the *Linked Data* network. In this structure, called *Linked Data Interface* (LDI), there is an abstract representation for every Wikipedia article. Each one of these abstract representations contains a pointer to the Linked Data document that provides an RDF description of the entity. The disambiguation task is achieved by identifying the item in the LDI that is most similiar to the context of the named entity (the context is represented by the set of words that appear before and after the NE). This algorithm is called *Semantic Disambiguation Algorithm* (SDA). The architecture of this semantic labeling system is presented in Figure 3.

3.1 The Linked Data Interface (LDI)

To each entity that is described by an entry in Wikipedia, we associate some *metadata*, composed of three elements: (i) a set of *surface forms*, (ii) the set of

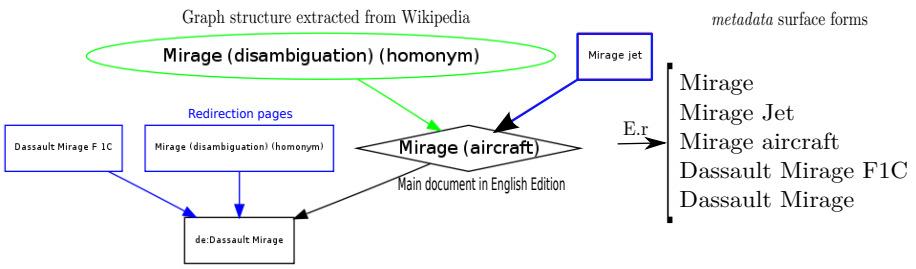

Fig. 4. All possible surface forms are collected from multiple linguistic editions of Wikipedia and transferred into a set $E.r$. Here two complementary surface forms for a plane name are collected from the German edition.

words that are contained in the entity description, where each word is accompanied by its $tf.idf$ weight ([13]), and (iii) an URI that points to some entity in the *Linked Data* Network. The $tf.idf$ value associated to a word is its frequency in the Wikipedia document, multiplied by a factor that is inversely proportional to the number of Wikipedia documents in which the word occurs (the exact formula is given below).

The set of surface forms for an entity is obtained by taking every Wikipedia entry that points to it by a redirection link, every entry that corresponds to its description in another language and, finally, in every disambiguation page that points to this entity, the term in the page that is associated to this pointer. As an example, the surface form set for the NE Paris (France) contains 39 elements, (eg. *Ville Lumière, Ville de Paris, Paname, Capitale de la France, Département de Paris*).

In our application, the surface forms are collected from five linguistic editions of Wikipedia (English, German, Italian, Spanish and French). We use such cross-linguistic resource because in some cases, a surface form may appear only in a language edition of Wikipedia that is not the one of the source text. A good example of this is given by the Figure 4. In this example, we see that the surface form *Dassaut Mirage* is not available in the English Wikipedia but can be collected from the German edition of Wikipedia.

The structure of Wikipedia and the sequential process to build *metadata* like ours, has been described previously ([3, 4]).

We will now define more formally the LDI.

Let C be the Wikipedia corpus. C is partitioned into subsets C^l representing linguistic editions of Wikipedia (i.e *fr.wikipedia.org* or *en.wikipedia.org*, which are independent language sub-corpus of the whole Wikipedia).

Let D be a Wikipedia article. Each $D \in C^l$ is represented by a triple $(D.t, D.c, D.l)$, where $D.t$ is the title of the article, made of a unique word sequence, $D.c$ is a collection of terms w contained in the article, $D.l$ is a set of links between D and other Wikipedia pages of C. Any link in $D.l$ can be an internal redirection inside C^l (a link from a redirection page or a disambiguation

page) or in another document in C (in this case, a link to the same article in another language).

The LDI may now be described the following way. **Let $E \in LDI$ be a metadata container that corresponds to some $D \in C$.** E is a tuple $(E.t, E.c, E.r, E.rdf)$. We consider that E and D are in relation if and only if $E.t = D.t$. We say that E represents D, which will be noted $E \rightarrow D$. $E.c$ contains pairs built with all words w of $D.c$ associated with their $tf.idf$ value calculated from C^l.

The $tf.idf$ weight for a term w_i that appears in document d_j is the product of the two values tf and idf which are calculated as shown in equations 1 and 2. In the definition of idf, the denominator $|\{d : d \in C^l, w_i \in d\}|$ is the number of documents where the term w_i appears. tf is expressed by equation 2, where $w_{i,j}$ is the number of occurrences of the term w_i in document d_j, and the denominator is the sum of number of occurrences of all terms in document d_j.

$$\mathrm{idf_i} = \log \frac{|C^l|}{|\{d : d \in C^l, w_i \in d\}|} \tag{1}$$

$$\mathrm{tf_{i,j}} = \frac{w_{i,j}}{\sum_k w_{k,j}} \tag{2}$$

The $E.c$ part of a metadata container must be trained for each language. In our LDI the three following langages have been considered: English, French and Spanish. The amount of representations collected can potentially elaborate semantic links for 745 k different persons or 305 k organizations in English, 232 k persons, and 183 k products in French.

The set of all surface forms related to a document D is built by taking all the titles of special documents (i.e redirection or disambiguation pages) targeted by the links contained in $D.l$, and stored in $E.r$.

The $E.rdf$ part of the metadata container must contain a link to one or more entry points of the *Linked Data* network. An entry point is an URI, pointing to an RDF document that describes the entity represented by E. As an example, *http://dbpedia.org/data/Spain.rdf* is the entry point of the DBpedia instance related to Spain inside the *Linked Data* network. The special interest of DBpedia for our application is that the ontology is a mirror of Wikipedia. Any English article of Wikipedia (and most French and Spanish ones) is supposed to have an entry in DBpedia. DBpedia delivers also correspondence files between others entry point in the *Linked Data* Network and Wikipedia records[5]: for example, another entry point for Spain in the *Linked Data* Network is on the CIA Factbook RDF collection[6]. We use those table files to create $E.rdf$. For our experiments, we included in $E.rdf$ only the link to the DBPedia entry point in the *Linked Data* Network.

[5] See on `http://wiki.DBpedia.org/Downloads34` files named *Links to Wikipedia articles*

[6] http://www4.wiwiss.fu-berlin.de/factbook/resource/Spain

3.2 Semantic disambiguation algorithm (SDA)

To identify a named entity, we compare it with every metadata container $E_i \in LDI$. Each E_i that contains at least one surface form that corresponds to the named entity surface form in the text is added into the candidate set. Now, for each candidate, its set of words $E_i.c$ is used to calculate a similarity measure with the set of words that forms the context of the named entity in the text. In our application, the context consists of the n words that come immediately before and after the NE. The $tf.idf$ is used to calculate this similarity measure. The E_i that gets the higher similarity score is selected and its URI pointer $E_i.rdf$ is used to identify the entity in Linked Data that corresponds to the NE in the text.

Regarding the candidate set CS that has been found for the NE to be disambiguated, three situations can occur:

1. $CS = \emptyset$: there is no metadata container for NE.
2. $|CS| = 1$: there is only one metadata container available to establish a semantic link between EN and an entity in the Linked Data Network.
3. $|CS| > 1$: there are more than one possible relevant metadata container, among which at most one must be selected.

Case 1 is trivial (no semantic link available). For cases 2 and 3, a cosine similarity measure (see equation 3) is applied to NE context $\mathbf{S.w}$ and $\mathbf{E.c}_{\boldsymbol{tf.idf}}$ for every metadata container $E \in CS$. As usual, the vectors are formed by considering each word as a dimension. If a word appears in the NE context, we put the value 1 in its position in the vector space, 0 otherwise. For $E.c$, we put in the vector the $tf.idf$ values. The similarity values are used to rank every $E \in CS$.

$$cosinus(S, E) = \frac{\mathbf{S.w} \cdot \mathbf{E.c}_{\boldsymbol{tf.idf}}}{\|\mathbf{S.w}\| \, \|\mathbf{E.c}_{\boldsymbol{tf.idf}}\|} \tag{3}$$

Finally the best candidate E_Ω according to the similarity ranking is chosen if its similarity value is higher than the threshold value α, as described in 4. The algorithm derived from this method is presented in Table 3.

$$\begin{gathered} \forall E_i \in CS \, \{E_\omega = argmax(cosinus(S, E_i))\} \\ E_\Omega = \begin{cases} \emptyset & \text{if score}(E_\omega) \leq \alpha \\ E_\omega & \text{otherwise} \end{cases} \end{gathered} \tag{4}$$

4 Experiments

There is no standard evaluation schema for applications like the one described in this paper. There are many metrics (precision, recall, word error rates) and annotated corpus for NER task, but none of them includes a Gold Standard for

SDA **Function**: rdf = SDA(sf , S[])	SDA
Input: sf = surface form of detected NE to link S[] = contextual words of EN **Output**: rdf = *uri* link between EN and *Linked Data* entry point	**Local variables**: E[]=metadata CS[]=Candidate Set of metadata α=threshold value **Algorithm**: (1) CS[]=search all E[] where E[].c match sf (2) if (CS[] == null) return null (3) for x = all CS[] (3.1) $CS[x]$.score=cosinus($CS[x].w$: $TF.idf$[], S[]) (4) order CS[] by descending CS[].score (5) if ($CS[0].score > \alpha$) return $CS[0].rdf$ (5.1) else return null

Table 3. Pseudo code of Semantic Disambiguation Algorithm (SDA).

Semantic Web annotation. We evaluated our system with an improved standard NER test corpus. We associate to each NE of such corpus a standard *Linked Data* URI coming from DBpedia. This proposal has the following advantage. DBpedia is now one of the most known and accurate RDF resource. Because of this, DBpedia evolved as a reference interlinking resource[7] to the *Linked Data* semantic network[8]. The NER corpora used to build semantically annotated corpora are described below.

Test corpora

The base corpus for French semantic annotation evaluation is derived from the French ESTER 2 Corpus ([6]). The named entity (NE) detection task on French in ESTER 2 was proposed as a standard one. The original NE tag set consists of 7 main categories (persons, locations, organizations, human products, amounts, time and functions) and 38 sub-categories. We only use PERS, ORG, LOC, and PROD tags for our experiments.

The English evaluation corpus is the *Wall Street Journal (WSJ)* version from the CoNLL *Shared Task* 2008 ([15]). NE categories of WSJ corpus include: Person, Organization, Location, GPE, Facility, Money, Percent, Time and Date, based on the definitions of these categories in MUC and ACE7 tasks. Sub- categories are included as well. We only use PERS, ORG, LOC, and PROD tags and convert most of the GPE in ORG for our experiments. Some NE tags assigned to common names in WSJ (like *plane* as PROD) had been removed.

[7] See `http://wiki.dbpedia.org/Interlinking`.

[8] DBpedia is now an *rdf*interlinking resource for CIA World Fact Book, US Census, Wikicompany, RDF Wordnet and more.

	ESTER 2 2009 (French)			WSJ CoNLL 2008 (English)		
Labels	Entities in test corpus	Equivalent entities in LDI	Coverage (%)	Entities in test corpus	Equivalent entities in LDI	Coverage (%)
PERS	1096	483	44%	612	380	62%
ORG	1204	764	63%	1698	1129	66%
LOC	1218	1017	83%	739	709	96 %
PROD/GPE	59	23	39%	61	60	98 %
Total	3577	2287	64%	3110	2278	73%

Table 4. All NE contained in a text document does not have necessarily a corresponding representation in LDI. This Table shows the coverage of built metadata contained in LDI, regarding NE contained in the French ESTER 2 test corpus and in the English WSJ CoNLL 2008 test corpus.

4.1 Gold standard annotation method

To build test corpora, we used a semi-automatic method. We first applied our semantic annotator and then removed or corrected manually the wrong semantic links. For some NE, the Linked Data Interface does not provide semantic links. This is the problem of coverage, managed by the use of the α threshold value. Level of coverage for the two test corpus in French and English is given in Table 4.

5 Results

To evaluate the performances of SA we applied it to the evaluation corpora with only Word, POS and NE. Two experiments have been done. First, we verify the annotation process under the scope of quality of disambiguation: we apply SA only to NEs which have their corresponding entries in LDI. This means we do not consider uncovered NE (as presented in Table 4) in the labeling experiment. We only try to label the 2287 French and 2278 English covered NEs. Those results are given in the section [no α] of Table 5. Then, we verify the capacity of SA to annotate a text, with potentially no entry in LDI for a given NE. This means we try to label the full set of NEs (3577 French and 3110 in English) and to assign the *NORDF* label when no entry is available in LDI. We use the threshold value[9] as a confidence weight score to assign as annotation an URI link or a *NORDF* label. Those results are given in Table 5 in the section [α]. We used recall measure (as in 5) to evaluate the amount of correctly annotated NEs according to the Gold Standard.

$$Recall = \frac{Total\ of\ correct\ annotations \rightarrow NE}{NE\ total} \tag{5}$$

[9] α value is a cosine threshold selected empirically and is positioned for this experiment on 0.10 in French and 0.25 in English.

	French tests				English tests			
NE	[no α]	Recall	[α]	Recall	[no α]	Recall	[α]	Recall
PERS	483	0.96	1096	0.91	380	0.93	612	0.94
ORG	764	0.91	1204	0.90	1129	0.85	1608	0.86
LOC	1017	0.94	1218	0.92	709	0.84	739	0.82
PROD	23	0.60	59	0.50	60	0.85	61	0.85
Total	2287	0.93	3577	0.90	2278	0.86	3020	0.86

Table 5. Results of the semantic labeler applied on the ESTER 2 and WSJ CoNLL 2008 test corpus.

Our results indicate a good level of performance for our system, in both language with over .90 of recall in French and .86 in English. The lower performances in English task can be explained by the structural difference of metadata in the two languages: near 0.7 million metadata containers are available in French and more than 3 millions in English (according to each local Wikipedia size). A biggest amount of metadata containers means also more propositions of synonymic words for a specific NE and a higher risk of bad disambiguation by the cosine algorithm. A way to solve this specific problem could be to weight the $tf.idf$ according to the amount of available metadata containers. The slight improvement of recall on English [α] experiment is attributed to the better detection of *NORDF* NEs, due to the difference of NE classes representation between the French and the English Corpora.

6 Conclusions and perspectives

In this paper, we presented a system to semantically annotate any named entity contained in a text, using a URI link. The URI resource used is a standard one, compatible with the Semantic Web network *Linked Data*. We have introduced the concept of *Linked Data Interface*, an exhaustive statistical resource containing contextual and nature description of potential semantic objects to label. The Linked Data Interface gives a possible answer to solve the problem of ambiguity resolution for an exhaustive semantic annotation process. This system is a functional proposition, available now, to establish automatically a relation between the vast amount of entry points available on the *Linked Data* network and named entities contained in an open text. We have shown that a large and expandable *Link Data Interface* of high quality containing millions of contextual descriptions for potential semantic entities, available in various languages, can be derived from Wikipedia and DBpedia. We proposed an evaluation schema of semantic annotators, using standard corpora, improved with DBpedia URI annotations. As our evaluation shows, our system can establish semantic relations automatically, and can be introduced in a complete annotation pipeline behind a NER tools.

References

1. Bikel, D., Schwartz, R., Weischedel, R.: An algorithm that learns whats in a name. Machine learning 7 (1999)
2. Borthwick, A., Sterling, J., Agichtein, E., R: Exploiting diverse knowledge sources via maximum entropy in named entity. Proc. of the Sixth pp. 152–160 (1998)
3. Bunescu, R., Pasca, M.: Using encyclopedic knowledge for named entity disambiguation. In: Proceedings of EACL. vol. 6 (2006)
4. Charton, E., Torres-Moreno, J.: NLGbAse: a free linguistic resource for Natural Language Processing systems. In: LREC (ed.) LREC 2010. No. 1, Proceedings of LREC 2010, Matla (2010)
5. Dill, S., Eiron, N., Gibson, D., Gruhl, D., Guha, R., Jhingran, A., Kanungo, T., Rajagopalan, S., Tomkins, A., Tomlin, J., Others: SemTag and Seeker: Bootstrapping the semantic web via automated semantic annotation. In: Proceedings of the 12th international conference on World Wide Web. p. 186. ACM (2003)
6. Galliano, S., Gravier, G., Chaubard, L.: The ESTER 2 Evaluation Campaign for the Rich Transcription of French Radio Broadcasts. In: International Speech Communication Association conference 2009. pp. 2583–2586. Interspeech 2010 (2009)
7. Kiryakov, A., Popov, B., Terziev, I., Manov, D., Ognyanoff, D.: Semantic annotation, indexing, and retrieval. Web Semantics: Science, Services and Agents on the World Wide Web 2(1), 49–79 (2004)
8. Lafferty, J., McCallum, A., Pereira, F.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proceedings of the Eighteenth International Conference on Machine Learning. pp. 282–289. Citeseer (2001)
9. Passant, A., Laublet, P.: Meaning of a tag: A collaborative approach to bridge the gap between tagging and linked data. WWW 2008 Workshop Linked Data on the Web (2008)
10. Popov, B., Kiryakov, A., Kirilov, A., Manov, D., Ognyanoff, D., Goranov, M.: Kim-semantic annotation platform. Lecture Notes in Computer Science pp. 834–849 (2003)
11. Ratinov, L., Roth, D.: Design Challenges and Misconceptions in Named Entity Recognition. In: Proceedings of the Thirteenth Conference on Computational Natural Language Learning. International Conference On Computational Linguistics (2009)
12. Raymond, C., Riccardi, G.: Generative and discriminative algorithms for spoken language understanding. In: Proceedings of Interspeech2007, Antwerp, Belgium. p. 2. Citeseer (2007)
13. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval* 1. Information processing & management (1988)
14. Sekine, S., Sudo, K., Nobata, C.: Extended named entity hierarchy. In: Proceedings of the LREC-2002 Conference. pp. 1818–1824. Citeseer (2002)
15. Surdeanu, M., Johansson, R., Meyers, A., L: The CoNLL-2008 shared task on joint parsing of syntactic and semantic dependencies. Proceedings of the p. 159 (2008)
16. Uren, V., Cimiano, P., Iria, J., Handschuh, S., Vargasvera, M., Motta, E., Ciravegna, F.: Semantic annotation for knowledge management: Requirements and a survey of the state of the art. Web Semantics: Science, Services and Agents on the World Wide Web 4(1), 14–28 (Jan 2006)
17. Zelaia, A., Arregi, O., Sierra, B.: A multiclassifier based approach for word sense disambiguation using Singular Value Decomposition. Proceedings of the Eighth International Conference on Computational Semantics - IWCS-8 '09 (January 2009), 248 (2009)