A disambiguation resource extracted from Wikipedia for semantic annotation

Eric Charton, Michel Gagnon

École Polytechnique de Montréal
2900 Edouard Montpetit, Montreal, QC H3T 1J4, Canada
eric.charton@polymtl.ca, michel.gagnon@polymtl.ca

Abstract

The Semantic Annotation (SA) task consists in establishing the relation between a textual entity (word or group of words designating a named entity of the real world or a concept) and its corresponding entity in an ontology. The main difficulty of this task is that a textual entity might be highly polysemic and potentially related to many different ontological representations. To solve this specific problem, various Information Retrieval techniques can be used. Most of those involves contextual words to estimate wich exact textual entity have to be recognized. In this paper, we present a resource of contextual words that can be used by IR algorithms to establish a link between a named entity (NE) in a text and an entry point to its semantic description in the LinkedData Network.

Keywords: Semantic Annotation, Ontology, Disambiguation

1. Introduction

The Semantic Web is a vision of a future web of structured data (Berners-Lee et al., 2001). Ontological support for the semantic web is an active area of both research and business development. In the recent time, as part of the Semantic Web activity, a number of different systems have been built to mainly perform two tasks: (i) create ontologies, and (ii) annotate web pages with ontology derived semantic tags. The Semantic Annotation (SA) task is the particular field of investigation related to the ability of an information system to establish automatically links between a document and knowledge of the semantic web. The main difficulty of the SA task is that a textual sequence might be highly polysemic and potentially related to many different ontological representations (e.g., *New-York* can be a song, a city, a state, a movie, and many other things).

One possible answer to this question is to compare the textual sequence with a set of possible contextual words related to each potential matching semantic concept, using an Information Retrieval (IR) algorithm like a cosine similarity or a distance measure. However, the specific difficulties of the SA task is that each unique semantic concept is related to a different set of potential contextual words. Another difficulty is to find a way to establish a standardized description of all the semantic concepts used for an annotation task, as only an association between the textual sequence in a text and an exact description of some of its properties establish clearly that its semantic identity have been determined with accuracy by a SA system.

In this paper, we present an ontological resource that can be used by classical IR algorithms to establish a disambiguated link between a textual sequence in a text and an *entry point* to its semantic description in the LinkedData Network[1]. By *entry point*, we mean an URI pointing on a RDF description of a semantic property, according to the semantic web standard, that can be used to develop a RDF graph semantically related to the annotated entity. The presented resource is derived from Wikipedia encyclopedic content. It contains for each encyclopedic document contained in Wikipedia, a representation composed by a bag of potential contextual word and one ore more links to the LinkedData Network.

This communication is organized as follow. In section 2. we present the SA task and existing algorithms and disambiguation strategy deployed. In section 3. we explain how a universal and standardized disambiguation resource called LinkedData Interface (LDI) can be built for SA and we describe our proposed one. Then, in section 3.3., we briefly describe an algorithm of disambiguation that can be used in conjunction with the LDI to manage the SA task. We evaluate the complete solution in section 4.2. and then conclude.

2. The semantic annotation problem

Associating a semantic information to a textual sequence can be done with various level of knowledge.

- The basic level consist in a class label associated to the textual sequence. This is the Named Entity Recognition (NER) task. The most common NER task consists in association of class label like Person names, Organizations names, Product names (Ng and Lee, 1996; Nadeau and Sekine, 2009). The NER task can be extended to various classes of annotation like for example Biomedical related labels (Settles, 2004). Many approaches have been proposed to solve the NER task(Lafferty et al., 2001; Kazama and Torisawa, 2007; Béchet and Charton, 2010).
- The second level of semantic information that can be attributed to a textual sequences consists in its asso-

[1]The term Linked Data refers to a set of best practices for publishing and connecting structured data on the Web. Key technologies that support Linked Data are URIs (a generic means to identify entities or concepts in the world), HTTP (a simple yet universal mechanism for retrieving resources, or descriptions of resources), and RDF (a generic graph-based data model used to structure and link data that describes things in the world).

ciation with a link to a formal representations of its characteristics.

The major problem faced by any attempt to associate a semantic information to a textual sequence is related to the *Word Sense Disambiguation (WSD)*. WSD consists in determining which sense of a textual sequence is used when it appears in a particular context. It is necessary to include a disambiguation resource in an annotation system to handle the WSD problem.
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, (Zelaia et al., 2009) suggested to use the LSA[2] techniques mixed with cosine similarity measure to disambiguate terms in the perspective of establishing a semantic link. The Kim system (Popov et al., 2003) re-uses the Gate platform and its NLP components and apply rules to establish a disambiguated link. Semtag (Dill et al., 2003) uses two kinds of similarity functions: bayesian, and cosine. 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 a city name example, like *Paris*, those systems could establish a disambiguated link between any *Paris* NE and its exact *Linked Data* representation only if they have access to 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. Finally, the way used to detect a NE with a NER system is not sufficient to establish a link between this NE and it's exact ontological representation. Any entity has a specific probable word context, related to its identity: for example, *la Seine* or *Tour Eiffel* are potentially contextual for NE *Paris* as a city but not for the Ocean Liner *Paris*, whose word context will be mostly composed by words like *ocean* or *docked*.

2.1. Semantic labeling task according to Semantic Web standards

The emergence of semantic web and LinkedData network, with billions of RDF triples representing virtually any knowledge, transforms the principal objectives of semantic annotation. Using Semantic Web content as complementary information extracted from semantic ontologies can enhance information retrieval, enable faceted document browsing and analytics based on semantics. Some new tools tries today to establish a link between entities in documents and LinkedData Network. This appear to be a formal and normalized application of the SA task.

[2]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.

Recently, various systems have been launched as web service dedicated to SA task respecting the new emergent semantic web standards like LinkedData network. DBPedia Spotlight[3] (Mendes et al., 2011) is an adaptable system to find and disambiguate natural language mentions of DBpedia resources. The disambiguisation strategy of DBPedia Spotlight appear to be similar that the previously proposed on KIM system. First a spotting stage recognizes in a sentence the phrases that may indicate a mention of a DBpedia resource. Candidate selection is subsequently employed to map the spotted phrase to resources that are candidate disambiguations for that phrase. Then at a disambiguation stage, the context around the spotted phrase is used to decide for the best choice amongst the candidates. DBpedia Spotlight uses a vector space model algotihm using TF.IDF weights. Contextual words are collected on Wikipedia corpus. Wikimeta[4] is another system. It uses a set of bag of words according to a cosine similarity algorithm to disambiguate semantic entities (Charton et al., 2011). This system uses the surface form of a NE to extract a set of candidate contained in *metadata*, each one corresponding to an entity that possesses at least a label matching the surface form. A cosine similarity measure between the word context of the NE and the *metadata* bag of words is calculated. Finally, the *metadata* that gets the highest score is considered as potentially reflecting the exact identity of the NE.

3. Universal resource for semantic tagging

The lack of access to exhaustive contextual information related to entity denotated by a NE is an open and crucial problem. SA systems need access to a normalized resource of the usual contextual words of any NE to establish precisely their semantic identity. Some authors suggest that the standardization of the SA task using Semantic Web resources (Uren et al., 2006) associated with exhaustive word context knowledge, could be a solution to the WSD problem of SA.
We propose in this paper to supply IR algorithm dedicated to SA task with an exhaustive knowledge, derived from Wikipedia content, including for all conceptual entities one or more URI, compatible with the Semantic Web network LinkedData. Each *metadata* associated to a conceptual entity, contains also one ore more URI link to the LinkedData Network. This would allow, for each NE connected to one of these *metadata*, to establish a link between this NE and the LinkedData Network.
Our resource contains about 4 millions *metadata* (2,5 Million in English, 860 K in French and 694 K in spanish), each describing a unique concept. For each concept, a metadata unit contains word-context information and possible writings surface forms of the concept (ie *New-York, NYC, Big Apple*). This resource 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

[3]spotlight.dbpedia.org

[4]www.wikimeta.org

	Persons	Org's	Locations	Products	Functions	Time	Encyclopedic
FR	232027	87052	183729	96571	1588	18871	130530
EN	754586	305706	565941	326155	3783	13575	468829
ES	84623	58600	93030	51427	41	2048	92462

Table 1: Metadatas available for each language, with their classification groups.

pointer to the Linked Data document that provides an RDF description of the entity.

3.1. General Metadatas description and use

The structure of Wikipedia and the sequential process to build *metadata* like ours, has been described in (Bunescu and Pasca, 2006) and is applied in
(Charton and Torres-Moreno, 2010). For each document in Wikipedia, we build one *metadata*, composed of two elements: (i) a set of *surface forms*, (ii) all the words contained in the document, where each word is assigned to a $tf.idf$ weight (Salton and Buckley, 1988).

Surface forms

The set of surface forms is obtained trough the collection of every Wikipedia internal links that points to an encyclopedic document. This can be a *redirection link*, an *interwiki link* (directing to the same document in another language edition) and, finally, every *disambiguation page* that points to the encyclopedic document.

As an example, the surface form set for the NE *Paris (France)*[5] contains 39 elements, (eg. *Ville Lumière, Ville de Paris, Paname, Capitale de la France, Département de Paris*). In our resource, 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 but should be used in other language. A good example of this are the surface forms *Renault-Dacia* or *RNUR* that can be found in a French text. They are not available in the French Wikipedia but can be collected from the Polish edition of Wikipedia. The exhaustive nature of surface forms sets allows in this situation to maximize the research of candidates *metadata* units related to a NE.

Bag of words and weights

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 bags of words and $tf.idf$ contained in *metadata* are specific to a linguistic edition of Wikipedia: for example an encyclopedic description of the *automobile* concept will exists as a distinct French, English and Spanish *metadata* description, with different bags of words - according to the language used. This allows word context similarity measurement with a language related bag of words.

[5] `http://www.nlgbase.org/perl/display.pl?query=Paris&search=FR`

3.2. Semantic Links

In addition to surface forms and weighted bag of word, a metadata unit contains semantic links. The Semantic Link section must contain one or more link to an entry points of the *Linked Data* network.
For instance, *http://dbpedia.org/data/Spain.rdf* is the entry point of the DBpedia RDF set related to Spain inside the LinkedData network. More than one entry point can exists for a unique concept in the LinkedData Network: for example, a country concept like **Turkey** have an entry point in the Dbpedia Data Set [6] and another one in the CIA World Factbook[7].

LinkedData Dataset	FR	EN	ES
DBPedia	447253	2063687	109502
Geonames	27836	47592	11203
CIA World Factbook	229	229	191
Wikicompany	150	200	10
Geodata	17236	34527	-
US Census	-	540	-

Table 2: Available entry points in various RDF Data sets of the LinkedData Network in our metadata

To collect those references to URI entry points in the LinkedData network we use the following techniques and resources:

1. DBPedia provides for each Wikipedia concept a unique RDF page. This page is described by a unique description key, corresponding to the English Wikipedia name of the document describing the concept. So we know that for each *metadata*, there is a RDF description page in the DBPedia Resource: e.g the Wiki page `http://en.wikipedia.org/wiki/Istanbul`, which has been used to produce the metadata unit for Istanbul, has a corresponding page `http://dbpedia.org/data/Istanbul.rdf` in DBpedia, which provides RDF triples describing this city. .

2. Frequently, numerous entry points in LinkedData corresponding to a unique concept will be described trough a *owl:sameAs*[8] RDF tag, according to the

[6] `http://dbpedia.org/page/Turkey` and `http://dbpedia.org/data/Turkey.rdf`

[7] `http://www4.wiwiss.fu-berlin.de/factbook/page/Turkey` and `http://www4.wiwiss.fu-berlin.de/factbook/resource/Turkey`

[8] see `http://www.w3.org/TR/owl-ref/#sameAs-def`

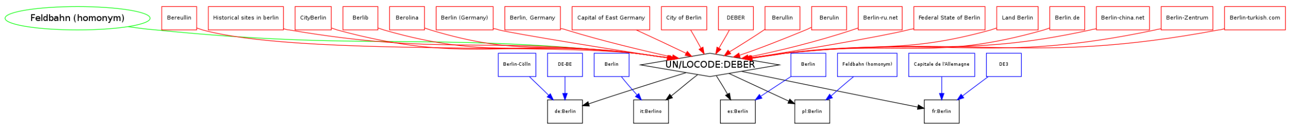

Figure 1: All possible surface forms are collected from multiple linguistic editions of Wikipedia. In this example, multiple complementary surface forms for a city name are collected from various language editions of Wikipedia.

OWL[9] definition. For example the RDF set for **Istambul** in DBPedia is connected to the CIA World Factbook RDF and to Geoname RDF Set[10] with an *owl:sameAs* definition.

3. When a new RDF set related to a concept have been found, a new mining can been done on it, to search again a *owl:sameAs* and find a new entry point, and so on. We obtain with this technique some LinkedData descriptions of *metadata* concepts in various semantic data sets like *Geonames, Wikicompany, Geodata* and *US Census*. The amount of entry points found in each RDF data set is given in table 2.

With those techniques, we associated to each *metadata* unit from table 1, one ore more links to various data sets in the LinkedData Network according to table 2.

3.3. Building algorithm

The complete metadata set builded for disambiguation is called the **LinkedData Interface** (LDI). 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 $E \to D$. $E.c$ 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[11]: for example, another entry point for Spain in the *Linked Data* Network is on the CIA Factbook RDF collection[12]. 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.

[9] The Web Ontology Language (OWL) is a family of knowledge representation languages for authoring ontologies. The languages are characterised by formal semantics and RDF/XML-based serializations for the Semantic Web. OWL is endorsed by the World Wide Web Consortium.

[10] `http://sws.geonames.org/745044/`

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

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

Word	POS	NE	Semantic Link
il	PRO:PER	UNK	
est	VER:pres	UNK	
20	NUM	TIME	
heures	NOM	TIME	
a	PRP	UNK	
Johannesburg	NAM	LOC.ADMI	`http://dbpedia.org/data/Johannesburg.rdf`

Table 3: Sample annotation of the French ESTER 2 NE test corpus. A link to DBpedia resource corresponding to the NE is added in a new column.

Word	POS	NE	Semantic Link
Laura	NNP	PERS.HUM	*NORDF*
Colby	NNP	PERS.HUM	
in	IN	UNK	
Milan	NNP	LOC.ADMI	`http://dbpedia.org/data/Milan.rdf`

Table 4: Sample annotation of the English CoNLL 2008 test corpus. The special semantic annotation *NORDF* is used when no RDF link is available

4. Example of disambiguation application

The disambiguation task can be 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).

4.1. 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 $\mathbf{S}.\vec{\mathbf{w}}$ and $\mathbf{E}.\vec{\mathbf{c}}_{tf.idf}$ 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 $E \in CS$.

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

Finally the best candidate E_Ω according to the similarity ranking is chosen if its similarity value is higher than the threshold value α.

4.2. Disambiguation evaluation

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 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. An example of this annotation is given in Table 3.
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[13] to the *Linked Data* semantic network[14]. 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 ((Galliano et al., 2009)). 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 ((Surdeanu et al., 2008)). NE categories of WSJ corpus include:

[13] See `http://wiki.dbpedia.org/Interlinking`.
[14] DBpedia is now an *rdf* interlinking resource for CIA World Fact Book, US Census, Wikicompany, RDF Wordnet and more.

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.

4.2.1. 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 5.

4.3. 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 5) 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 6. 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* (see example in Table 3) label when no entry is available in LDI.

We use the threshold value[15] as a confidence weight score to assign as annotation an URI link or a *NORDF* label. Those results are given in Table 6 in the section [α]. We used recall measure (as in 4) to evaluate the amount of correctly annotated NEs according to the Gold Standard.

$$Recall = \frac{Total\ of\ correct\ annotations \rightarrow NE}{NE\ total} \quad (4)$$

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.

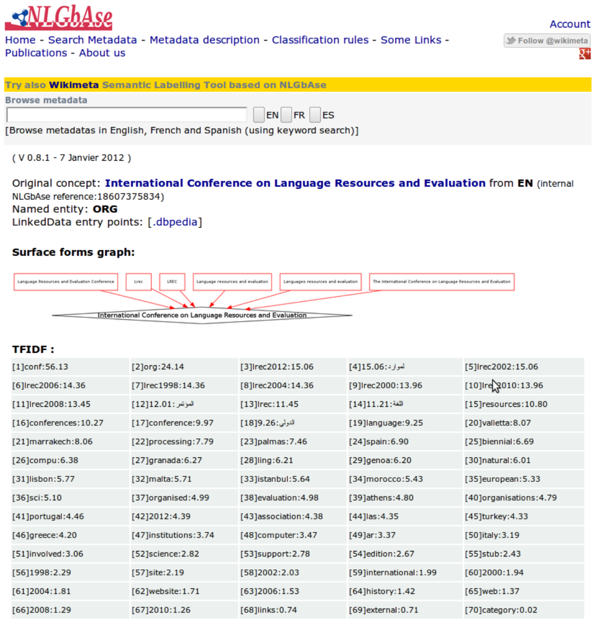

Figure 2: LDI can be downloaded or explored on line on the NLGbAse website.

5. Conclusions and perspectives

We have presented a set of *metadata* representing encyclopedic concepts contained in Wikipedia with weighted contextual words and writable surface forms. We have included in each *metadata* a standardized semantic link to the LinkedData network and its RDF data sets. Those *metadata* can be used as a resource to establish a semantic relation between a Named Entity in a text and its semantic representation on the LinkedData Network, using various Information Retrieval algorithms. The *metadata* are free to use. They can be downloaded or used on line on their dedicated website (see figure 2)[16]. Metadata representation can be browsed on line.

5.1. Perspectives

These *metadata* are available yet in three languages and can be extended by the same way to the 267 available language versions of Wikipedia[17]. We plan in the future to introduce new linguistic editions of the LDI in the future. We also work on a Semantic Annotation engine that will use various disambiguation algorithms to determine the performances of IR techniques with such resource.

6. Acknowledgements

The resource presented in this paper his hosted by courtesy of Wikimeta Technologies inc[18].

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

[16] www.nlgbase.org

[17] see meta.wikimedia.org/wiki/List_of_Wikipedias for a full description

[18] www.wikimeta.com

	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-GSP	59	23	39%	61	60	98 %
Total	3577	2287	64%	3110	2278	73%

Table 5: 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 test corpora.

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

7. References

Frédéric Béchet and Eric Charton. 2010. Unsupervised knowledge acquisition for extracting named entities from speech. In *ICASSP 2010*, Dallas. ICASSP.

T Berners-Lee, J Hendler, O R A Lassila, E Meaning, and IKWY Mean. 2001. The Semantic Web. *Scientific American*, 284(5):34–43.

R. Bunescu and M. Pasca. 2006. Using encyclopedic knowledge for named entity disambiguation. In *Proceedings of EACL*, volume 6.

Eric Charton and J.M. Torres-Moreno. 2010. NLGbAse: a free linguistic resource for Natural Language Processing systems. In LREC, editor, *LREC 2010*, number 1, Matla. Proceedings of LREC 2010.

Eric Charton, Michel Gagnon, and Benoit Ozell. 2011. Automatic Semantic Web annotation of named entities. In *Canadian AI*.

S. Dill, N. Eiron, D. Gibson, D. Gruhl, R. Guha, A. Jhingran, T. Kanungo, S. Rajagopalan, A. Tomkins, J.A. Tomlin, and Others. 2003. SemTag and Seeker: Bootstrapping the semantic web via automated semantic annotation. In *Proceedings of the 12th international conference on World Wide Web*, page 186. ACM.

S. Galliano, G. Gravier, and L. Chaubard. 2009. The ESTER 2 Evaluation Campaign for the Rich Transcription of French Radio Broadcasts. In *International Speech Communication Association conference 2009*, pages 2583–2586. Interspeech 2010.

J. Kazama and K. Torisawa. 2007. Exploiting Wikipedia as external knowledge for named entity recognition. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 698–707.

John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pages 282–289. Citeseer.

P Mendes, Max Jakob, A. Garcia-Silva, and C. Bizer. 2011. DBpedia Spotlight: Shedding Light on the Web of Documents. *Text*, pages 1–8.

D. Nadeau and S. Sekine. 2009. A survey of named entity recognition and classification. *NAMED ENTITIES: RECOGNITION, CLASSIFICATION AND USE*, (1991):3.

Hwee Tou Ng and Hian Beng Lee. 1996. Integrating multiple knowledge sources to disambiguate word sense: an exemplar-based approach. *Annual Meeting of the ACL*.

B. Popov, A. Kiryakov, A. Kirilov, D. Manov, D. Ognyanoff, and M. Goranov. 2003. Kim-semantic annotation platform. *Lecture Notes in Computer Science*, pages 834–849.

G Salton and C Buckley. 1988. Term-weighting approaches in automatic text retrieval* 1. *Information processing & management*.

Burr Settles. 2004. Biomedical named entity recognition using conditional random fields and rich feature sets. *Proceedings of the International Joint Workshop on*, page 104.

Mihai Surdeanu, Richard Johansson, Adam Meyers, and L. 2008. The CoNLL-2008 shared task on joint parsing of syntactic and semantic dependencies. *Proceedings of the*, page 159.

V Uren, P Cimiano, J Iria, S Handschuh, M Vargasvera, E Motta, and F Ciravegna. 2006. 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, January.

Ana Zelaia, Olatz Arregi, and Basilio Sierra. 2009. 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.