NLGbAse: a free linguistic resource for Natural Language Processing systems

Eric Charton(1), Juan-Manuel Torres-Moreno

École Polytechnique de Montréal / Université d'Avignon et des Pays de Vaucluse (LIA)
2900, boulevard Edouard-Montpetit, Montréal (QC) H3C 3A7 / 339, chemin des Meinajaries, 84911 Avignon Cedex 9
eric.charton@polymtl.ca, juan-manuel.torres@univ-avignon.fr

Abstract

Availability of labeled language resources, such as annotated corpora and domain dependent labeled language resources is crucial for experiments in the field of Natural Language Processing. Most often, due to lack of resources, manual verification and annotation of electronic text material is a prerequisite for the development of NLP tools. In the context of under-resourced language, the lack of copora becomes a crucial problem because most of the research efforts are supported by organizations with limited funds. Using free, multilingual and highly structured corpora like Wikipedia to produce automatically labeled language resources can be an answer to those needs. This paper introduces NLGbAse, a multilingual linguistic resource built from the Wikipedia encyclopedic content. This system produces structured *metadata* which make possible the automatic annotation of corpora with syntactical and semantical labels. A *metadata* contains semantical and statistical informations related to an encyclopedic document. To validate our approach, we built and evaluated a Named Entity Recognition tool, trained with Wikipedia corpora annotated by our system.

1. Introduction

Easy access to language resources is crucial for experiments in the field of Natural Language Processing (NLP), Information Extraction (IE) and Retrieval (IR) tasks. Such resources, like annotated corpora, domain dependent labeled corpora or lexicons, are an essential part of evaluation, software prototyping and design implementation. Most often, due to lack of resources, manual verification and annotation of electronic text material like corpora is a prerequisite for the development of NLP tools. In this paper, we present NLGbAse, a multilingual NLP resource built from the Wikipedia encyclopedic content. It produces structured *metadata* and annotated corpora with syntactical and semantical labels, from any language edition of Wikipedia. Our approach is validated by experiments on a Named Entity Recognition (NER) task. The NER tool used is trained on annotated corpora generated by our system. The results obtained by this NER system on NER evaluation campaign test sets confirm its reliability and the potential of the proposed method.

This paper is structured as follow. First we present existing structured resources extracted or derived from Wikipedia and describe our proposition. Then we explain how we extract the *metadata* from Wikipedia, and generate with those *metadata*, rich multilingual annotated corpora. Next, we illustrate one application of our system, with training of a multilingual NER application. We finally evaluate results obtained with this NER tool and conclude with description of our future development.

2. Wikipedia as NLP resource

With most of NLP applications, significant performance gains can be obtained with an increasing of available data and improvement of its quality, rather than algorithm complexity. This is illustrated by the increasingly popular use of the web as a very large and exhaustive corpus. It seems therefore natural to exploit the web knowledge to try to discover semantic informations, useful relationships between named entities mentioned in text documents, or to solve the name disambiguation problem. As it's an open, collaborative, and rapidly growing encyclopedia on the Web, Wikipedia is considered as one of the most promising resources to extract knowledge dedicated to NLP applications. Many methods have been proposed to transform Wikipedia into structured content like DBpedia (Auer et al., 2007) or Yago (Suchanek et al., 2007) improvement . Other proposition experiment the transformation of the encyclopedic content into various forms of lexicons. Those lexicons are used for identification and disambiguation tasks, and more specifically, to improve Named Entity Recognition (NER) systems. In (Bunescu and Pasca, 2006) a named entity disambiguation method is presented that is intrinsically linked to a dictionary mapping proper names to their possible named entity denotations. The use of Wikipedia as external knowledge to improve NER is also explored by (Kazama and Torisawa, 2007) . The described method retrieves the corresponding Wikipedia entry for each candidate word sequence and extracts a category label from the first sentence of the entry, which can be thought of as a definition part. These category labels are used as features in a CRF-based NE tagger.

2.1. Architecture of proposed system

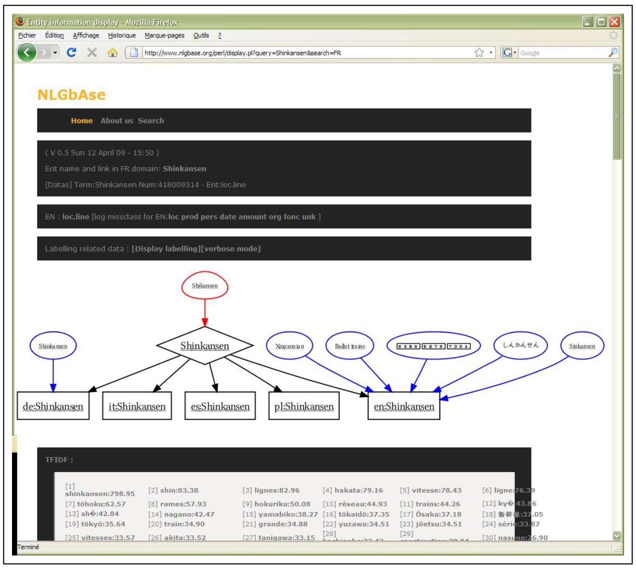
Our idea is to extract lexical resources from Wikipedia, then to use those resources to produce labeled corpus, usable to train NLP applications like NER tools.

Lexical resources are *metadata* extracted from the Wikipedia content. Those *metadata* represent each encyclopedic article as a concept, described by its possible surface writing forms, a set of contextual words and a category tag. **Labeled corpus** component of the system is the Wikipedia corpus augmented by *Part Of Speech* and *Named Entities* tags. Finally, we use this labeled corpus to train a **NER system**. .

[1] Contribution of the first author have been done has affiliate of Laboratoire Informatique d'Avignon

3. Extraction of *metadata* from Wikipedia

According to the Wikipedia DTD, each description of an encyclopedic object is associated with its name as a unique database key. For example, if two objects share the same name (i.e. *Martin Gray*), the unique key for each object is distinguished by a complementary information included in parentheses such as *Martin Gray (Photographer)* or *Martin Gray (Holocaust Survivor)*. Moreover, a disambiguation page *Martin Gray (disambiguation)* is instantiated and contains a link to the two different pages. The disambiguation page is the first displayed page in case the user submits an ambiguous request. Several words or locutions can denote an encyclopedic object. For example, *Paris* is also called *Ville Lumière* or *Métropole française*. This situation entails the instantiation of a special empty Wikipedia page, named *redirect page* for each alternate word or location. The redirect page includes a single redirection to the main page describing the encyclopedic object. We consider all the characteristics of the Wikipedia internal structure to extract the *metadata*. For each encyclopedic article, a corresponding *metadata* is extracted. It consists of a semantic graph of all writing forms, a set of weighted words extracted from the encyclopedic article description and a taxonomic class label (such as PERson, ORGanization, LOCation). The example of the *metadata* associated to the *"Shinkansen"* encyclopedic article is depicted in figure 1.


Entity information display - Mozilla Firefox
Fichier
Édition
Affichage
Historique
Marque-pages
Outils
?
http://www.nlgbase.org/perl/display.pl?query=Shinkansen&search=FR
Google
NLGbAse
Home
About us
Search

(V 0.5 Sun 12 April 09 - 15:50)

Ent name and link in FR domain: **Shinkansen**

[Datas] Term:Shinkansen Num:418009314 - Ent:loc.line

EN : **loc.line** [log missclass for EN:**loc prod pers date amount org fonc unk**]

Labelling related data : **[Display labelling][verbose mode]**

TFIDF :

[1] **shinkansen:798.95**	[2] **shin:83.38**	[3] **lignes:82.96**	[4] **hakata:79.16**	[5] **vitesse:78.43**	[6] **ligne:76.39**
[7] **tōhoku:62.57**	[8] **rames:57.93**	[9] **hokuriku:50.08**	[10] **réseau:44.93**	[11] **trains:44.26**	[12] **ky�:43.86**
[13] **sh�:42.84**	[14] **nagano:42.47**	[15] **yamabiko:38.27**	[16] **tōkaidō:37.35**	[17] **Ōsaka:37.18**	[18] **新幹線:37.05**
[19] **tōkyō:35.64**	[20] **train:34.90**	[21] **grande:34.88**	[22] **yuzawa:34.51**	[23] **jōetsu:34.51**	[24] **série:33.87**
[25] **vitesses:33.57**	[26] **akita:33.52**	[27] **tanigawa:33.15**	[28] [illegible]	[29] [illegible]	[30] **nasuno:26.90**

Figure 1: The NLGbAse *metadata* representation with surface forms, class of entity and contextual words with their $tf.idf$ weight.

A sequential process provides the *metadata* set as follow:

1. For each article of Wikipedia we create a *metadata* instance.

2. We explore all *redirection* and *disambiguation* pages of Wikipedia linked to this article to include in the *metadata* instance a graph representation of all its possible writing forms.

3. We collect all the terms from the textual description of the Wikipedia article and calculate their $tf.idf$ weight. We include all pairs of terms and weight in *metadata* instance

4. We use a classification system to associate a class label to the *metadata* instance.

3.1. Metadata extraction process

Formalization of steps 1,2 and 3 can be expressed as follow. Considering C, the Wikipedia corpus. Inside C, exists some C^l representing a linguistic edition of Wikipedia (i.e *fr.wikipedia.org* or *en.wikipedia.org* are independent language sub-corpus of the whole Wikipedia). Each Wikipedia article $D \in C^l$ is required to be associated with properties $D = (D.t, D.w, D.l)$. Property $D.t$ is the title of an article, made of words, $D.w$ is a collection of words contained in the Wikipedia article, $D.l$ is a set of links between D and other Wikipedia pages of C. The $D.l$ links can be internal redirection inside C^l (a link from a redirection page or a disambiguation page) or C (in this case, a link to the same article in an other language). The *Metadata* E have properties $E = (E.t, E.w, E.r, E.k)$. We consider that E and D are in relation if and only if, $E.t = D.t$. This means $E \rightarrow D$. $E.w$ contain twins built with all words of $D.w$ associated with their $Tf.Idf$ value calculated from C^l. All possible writing surface forms for a *metadata* are collected trough heuristics in Wikipedia articles linked in $D.l$, and then stored in $E.r$.

3.2. Classification process

Step 4 is performed as follow. The property $E.k$ is a class label, according to taxonomy model of a NER tasks (i.e PERS for a person or LOC.GEO for a geographic place). The classification process is based on a combination of **SVM**, **Boostexter**, **naive bayes** classifiers and heuristics applied on the text content of the Wikipedia article. It has been described in (Charton and Torres-Moreno, 2009). The final accuracy of the classification process is around 0.92% using the ESTER 2 taxonomy model (i.e PROD ase general product tag and subclasses like *PROD.VEHICLE* for a vehicle description). The final accuracy of the classification task is around 0.92% for a ESTER 2 like taxonomy model[1] (PERS, ORG, LOC, PROD, FONC and 32 subclasses). The amount of entity descriptions presented in table 3.2. are obtained from three language editions of Wikipedia, transformed in *metadata*.

4. Rich labeled corpus generation

Next, we use the generated *metadata* to automatically build rich labeled corpora. Those corpora are then used to train a NER system. Considering that we hold now a *metadata* representation for each Wikipedia encyclopedic article. Each *metadata* representation includes a graph of surface form describing how the entity corresponding to the article can be written and a class label describing the NE category of the entity. The basic principle of this phase is to use the internal links between pages of Wikipedia in conjunction with the *metadata* to automatically label Wikipedia texts corpora.

[1] see http://www.afcp-parole.org/ester/ for naming convention

	PERS	ORG	LOC	PROD	FONC	TIME	UNK
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: Named entity classes of objects contained in three Wikipedia encyclopedic editions

Basically the process can be described with the following example: inside the Wikipedia article related to *Victor Hugo*, there is some internal links to Wikipedia objects like *Panthéon*, *Besançon*, the *Napoléon III* Emperor, etc. With the *metadata* class label $E.k$, we know that *Panthéon* is a Location (LOC.FAC class label), *Besançon* is a city (LOC.ADMI class label) and *Napoléon III* is a person (PERS.HUM class label). According to this, we can produce a version of the *Victor Hugo* Wikipedia article labeled with NE tags. This is possible because each internal links contained in the encyclopedic text description is related to an encyclopedic object and we know the class of each encyclopedic object by its *metadata* representation.

Consequently, the labeled generated file, derived from a Wikipedia encyclopedic corpus, is built as follows. We explore sequentially each word of the text description of the encyclopedic object. We encounter two cases: either a word or a group of words is linked to another encyclopedic object of Wikipedia; or words are just part of text and are not associated with any internal links. If a word or group of words is linked, we search the class label $E.k$ of the *metadata* E related to the encyclopedic article linked, and we associate this label to the word sequence as a NE tag. To illustrate the process, let's consider the following sentence, extracted with its *wiki syntax*[2] from the English Wikipedia corpus:

- *The national team of [[Kenya national football team|Kenya]] is controlled by the [[Kenya Football Federation]].*

As we know that in their *metadata* representations, *Kenya* encyclopedic article is referenced as LOC object and *Kenya Football Federation* encyclopedic article is referenced as ORG object, we obtain the following annotated text.

- *The national team of <ent begin> Kenya <ent=kenya><tag=LOC> is controlled by the <ent begin> Kenya Football Federation <ent=kenya football federation><tag=ORG>.*

We call this Wikipedia corpora labeled representation C_{d1}. We apply to C_{d1} a part of speech tagger[3] to align text content, POS label, and semantic label (NE) in the perspective of training a NER system, and obtain C_{d2}:

```
The           DT
national      JJ
team          NN
of            IN
<ent begin>
Kenya         NAM
<ent=kenya><tag=loc>
is            VBZ
controlled    VVN
by            IN
the           DT
<ent begin>
Kenya         NAM
Football      NAM
Federation    NAM
<ent=kenya football federation>
<tag=org>
.             SENT
```

The C_{d2} representation of Wikipedia content is not sufficient to train a statistical NER system because internal links of the encyclopedia are not exhaustive, and consequently, a lot of unlabeled potential NE remains in corpus. The corpus C_{d3} is obtained by rejection of sentences from C_{d2} with probable missing labels, determined by logical rules. Unlabeled NE should introduce a bias in the inference learning process. To avoid this, we select from C_{d2} the sentences containing probably labeled NE, with a limited set (less than a dozen) of logical rules. For example, we reject sentences containing proper names (NP), numbers (NUM) and words with capital letters, not associated with NE label. Finally, the corpus C_{d3} contains sentences that can be used as it for NER system training :

```
Gregory          NAM      PERS
of               IN       PERS
Tours            NAM      PERS
accused          VVD      UNK
Mummolus         NAM      PERS
of               IN       UNK
subjecting       VVG      UNK
many             JJ       UNK
Franks           NNS      ORG
who              WP       UNK
had              VHD      UNK
...
```

5. Training of NER system

An interesting application of the system is its ability to train quickly and automatically multilingual NER tools, avoiding the cost of manually annotation of training corpora. Extraction of *metadata* of a linguistic edition of Wikipedia is the only step needed to generate a labeled corpus, using internal links of Wikipedia. To demonstrate this possibility, we have deployed a French, Spanish and English NER

[2] The specific syntax used by the Mediawiki software

[3] The corpus described in this paper has been tagged with **Treetagger** in French, English, and Spanish. There is no constraint related to the POS tagger tool. Experiments have also be done with the **LIA_Tagg tagger**.

Language	Original corpus C_{d2}	Final training corpus C_{d3}
English	74 711 331	170 804
French	10 008 100	167 707
Spanish	4 900 862	152 432

Table 2: Amount of sentences in original corpora and NER training corpora

system and evaluated it. The NLGbAse corpus C_{d3}, derived from Wikipedia, contains enough information to train a conditional random field (CRF) based NER system. A CRF is a discriminative probabilistic model often used for the labeling or parsing of sequential data.

For our experiments, we used the CRF++ implementation[4]. CRF++ is designed for generic purpose and had been applied to a variety of NLP tasks, such as Named Entity Recognition, Information Extraction and Text Chunking. This software is natively compatible with C_{d3} corpus for learning and labeling. CRF++ uses tokens. A token consists of multiple (but fixed-numbers) columns. The definition of tokens depends on tasks, in our application case, they simply correspond to words, POS and NE tags for the learning process, and POS and NE tag for the tagging process. In the context of this article, we train with CRF++ three models for French, English and Spanish language. Training time is about 16 hours for each language with 3 CPU (I7) and 5 gb of memory. Labeling process is in real time.

6. Evaluation and results

We applied CRF NER system to three evaluation test sets covering French, English and Spanish language. Our performance results are given as micro-averaged precision, recall and F-measure both in terms of ESTER 2 and ACE-2 style scoring. We complete evaluation with the slot error rate (SER) error measure (used as reference in the ESTER 2 campaign). The NE tag set consists of 7 main categories (persons, locations, organizations, products, amounts, time and functions). The tag set considered is therefore more complex than the one used in the NE extraction tasks of the MUC 7 and DARPA HUB 5 programs, where only 3 categories were considered, or the CoNLL tagset (4 tags PER,ORG, LOC, MISC). We evaluated all experiments with the original ESTER 2 categories (PROD, ORG, PERS, LOC, FONC, TIME, AMOUNT) when available in test corpus, or with equivalent categories (i.e GPE from the ACE test corpus is adapted to LOC).

6.1. French test set

The French test set come from ESTER 2 evaluation campaign. This named entity (NE) detection task on French was first implemented in ESTER 1 as a prospective task in order to define the first annotation guideline, corpus and scoring tools. The corpus come from broadcast news transcript. In ESTER 2, the task was proposed as a standard one. Two subtasks were defined: detection on the reference transcriptions and detection on automatic transcriptions. For the automatic transcript subtask, in order to precisely measure the impact of the WER on named entity detection, three automatic transcripts with different WER were given to the participants. Original results obtained during this campaign are described in (Galliano et al., 2009).

6.2. English test set

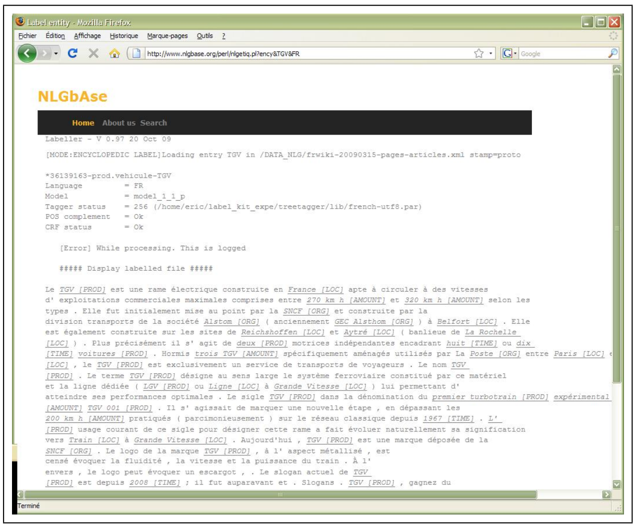
The English test set corpora come from ACE-2 Version 1.0 evaluation campaign (Doddington et al., 2004). Data sources include audio and image data in addition to pure text, and Arabic and Chinese in addition to English. The entity attributes are the type (person, organization, geo-political, location, facility, vehicle, weapon) and subtype of the entity, the entity class (specific, generic), and the name(s) of the entity that appear in the source data. Corpora include broadcasts transcripts (comparable to ESTER 2 ne-ref), and labeled documents from newspaper and newswire. We used for our experiments the broadcast transcript test corpus.

6.3. Spanish test set

As there is no NER reference evaluation corpus for Spanish, we semi-automatically annotated a set of articles from Spanish newspapers to elaborate one. Results on the mains NE labels are given on table 4.

Language	Precision	Recall	F-measure	SER
French	0,87	0,73	0,80	19
English	0,90	0,83	0,86	21
Spanish	0,89	0,81	0,84	34

Table 4: Evaluation results


Label entity - Mozilla Firefox
Fichier Édition Affichage Historique Marque-pages Outils ?
http://www.nlgbase.org/perl/nlgetiq.pl?ency&TGV&FR
Google
NLGbAse
Home About us Search
Terminé

```
Labeller - V 0.97 20 Oct 09

[MODE:ENCYCLOPEDIC LABEL]Loading entry TGV in /DATA_NLG/frwiki-20090315-pages-articles.xml stamp=proto

*36139163-prod.vehicule-TGV
Language         = FR
Model            = model_1_1_p
Tagger status    = 256 (/home/eric/label_kit_expe/treetagger/lib/french-utf8.par)
POS complement   = Ok
CRF status       = Ok

   [Error] While processing. This is logged

   ##### Display labelled file #####

Le TGV [PROD] est une rame électrique construite en France [LOC] apte à circuler à des vitesses
d' exploitations commerciales maximales comprises entre 270 km h [AMOUNT] et 320 km h [AMOUNT] selon les
types . Elle fut initialement mise au point par la SNCF [ORG] et construite par la
division transports de la société Alstom [ORG] ( anciennement GEC Alsthom [ORG] ) à Belfort [LOC] . Elle
est également construite sur les sites de Reichshoffen [LOC] et Aytré [LOC] ( banlieue de La Rochelle
[LOC] ) . Plus précisément il s' agit de deux [PROD] motrices indépendantes encadrant huit [TIME] ou dix
[TIME] voitures [PROD] . Hormis trois TGV [AMOUNT] spécifiquement aménagés utilisés par La Poste [ORG] entre Paris [LOC] e
[LOC] , le TGV [PROD] est exclusivement un service de transports de voyageurs . Le nom TGV
[PROD] . Le terme TGV [PROD] désigne au sens large le système ferroviaire constitué par ce matériel
et la ligne dédiée ( LGV [PROD] ou Ligne [LOC] à Grande Vitesse [LOC] ) lui permettant d'
atteindre ses performances optimales . Le sigle TGV [PROD] dans la dénomination du premier turbotrain [PROD] expérimental
[AMOUNT] TGV 001 [PROD] . Il s' agissait de marquer une nouvelle étape , en dépassant les
200 km h [AMOUNT] pratiqués ( parcimonieusement ) sur le réseau classique depuis 1967 [TIME] . L'
[PROD] usage courant de ce sigle pour désigner cette rame a fait évoluer naturellement sa signification
vers Train [LOC] à Grande Vitesse [LOC] . Aujourd'hui , TGV [PROD] est une marque déposée de la
SNCF [ORG] . Le logo de la marque TGV [PROD] , à l' aspect métallisé , est
censé évoquer la fluidité , la vitesse et la puissance du train . À l'
envers , le logo peut évoquer un escargot , . Le slogan actuel de TGV
[PROD] est depuis 2008 [TIME] ; il fut auparavant et . Slogans . TGV [PROD] , gagnez du
```

Figure 2: The NLGbAse NER tool.

[4] Toolkit CRF++:http://crfpp.sourceforge.net/

Language	NE test reference	NE label	Source
French	ne-ref_primaire_test_ester2	PROD, ORG, PERS, LOC, FONC, TIME, AMOUNT	Ester 2
English	Broadcast corpus	PERS, ORG, LOC, GPE	ACE phase 2
Spanish	Newspaper semi-manually annotated	PROD, ORG, PERS, LOC, FONC, TIME, AMOUNT	Original

Table 3: Characteristics of test sets

6.4. Discussion

Performances of the NER system have to be examined with consideration to it's nature : it's an unsupervised system (the CRF training is done with Wikipedia anotated corpora and no particular adaptation to the test conditions), and it is strictly not rule based (the presented results are obtained directly from the output of CRF++ labeling tool). Best systems ranked on reference transcriptions part of evaluation campaign uses both a rule-based approach where thousands of manually written rules are applied in conjunction with very large entity dictionaries. Although these carefully handcrafted knowledge models give excellent performance on a reference transcriptions, there is a clear lack of robustness of these models when applied to speech transcripts or when operate in unsupervised context.

Primary objective of our work is to generate automatically robust NER system, using the diversity of languages available in Wikipedia Corpus. Performances of the French NER system are directly comparable to those obtained by systems presented in ESTER 2 evaluation campaign, and specially with the CRF based system (described in (Béchet and Charton, 2010)) that obtained best results on the 3 robustness tasks, and first rank of non-rule based systems on ref task. Our French NER system F-measure performances are similar to this system (+0,02%) and minimize its SER rate (19 vs 23.91). Performances of English NER system are not directly comparable to official results of ACE-2 campaign [5] as they are obtained on Dev-Test corpus. However, we obtain a good precision and recall rate regarding to usual performances obtained on broadcast reference transcriptions. Performance of Spanish NER tool is for reference only and to demonstrate the ability of our system to train rapidly and automatically a new linguistic edition of our NER tool.

7. Conclusion

We have presented NLGbAse, a set of *metadata* representing encyclopedic concepts contained in Wikipedia. We have shown that, those *metadata* allow to automatically produce annotated corpora. Those corpora can be used to train statistical annotation tools. We have tested a CRF NER tool trained with the annotated corpora generated by our system, in the context of NER evaluation campaign and obtained interesting results.

For evaluation purposes, the taxonomy model of *metadata* is inspired from NER tasks. But the classification process used to apply this model can be modified for specific applications. We plan to exploit yet unclassified Wikipedia content (like medicine terms, biological terms, animal names) to generate new subject oriented *metadata* classes and labeled corpora. Next, the language coverage of this system will be extended by the progressive inclusion of new linguistic editions, using the 267 available language versions of Wikipedia[6]. We plan to update the NER system to Polish, German and Italian languages.

The *metadata*, annotated corpora, and CRF NER models are free to use. They can be downloaded or used on line on their dedicated website[7]. They are available in the English, French and Spanish languages. Metadata representation can be browsed on line and a demonstration tool of the NER application is also available.

8. References

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, and Zachary Ives. 2007. DBpedia: A Nucleus for a Web of Open Data. In *In 6th Intl Semantic Web Conference, Busan, Korea*, pages 11–15. Springer.

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

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. 2009. Classification dun contenu encyclopédique en vue dun étiquetage par entités nommées. In *Taln 2009*, volume 1, pages 24–26. TALN.

George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. 2004. The automatic content extraction (ACE) programtasks, data, and evaluation. In *Proceedings of LREC*, volume 4, page 837840. Citeseer.

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)*, page 698707.

F.M. Suchanek, G. Kasneci, and G. Weikum. 2007. Yago: a core of semantic knowledge. In *Proceedings of the 16th international conference on World Wide Web*, page 706. ACM.

[5] see http://www.itl.nist.gov/iad/mig//tests/ace/

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

[7] www.nlgbase.org