UNSUPERVISED KNOWLEDGE ACQUISITION FOR EXTRACTING NAMED ENTITIES FROM SPEECH

Frederic Bechet[*]

Aix-Marseille Université
Marseille, France
frederic.bechet@lif.univ-mrs.fr

Eric Charton

Universite d'Avignon
Avignon, France
eric.charton@univ-avignon.fr

ABSTRACT

This paper presents a Named Entity Recognition (NER) method dedicated to process speech transcriptions. The main principle behind this method is to collect in an unsupervised way lexical knowledge for all entries in the ASR lexicon. This knowledge is gathered with two methods: by automatically extracting NEs on a very large set of textual corpora and by exploiting directly the structure contained in the Wikipedia resource. This lexical knowledge is used to update the statistical models of our NER module based on a mixed approach with generative models (Hidden Markov Models - HMM) and discriminative models (Conditional Random Field - CRF). This approach has been evaluated within the French ESTER 2 evaluation program and obtained the best results at the NER task on ASR transcripts.

Index Terms— Speech recognition, Information retrieval, Named Entity, Statistical Tagging Models

1. INTRODUCTION

Extracting Named Entities (NEs) is one of the main tasks performed in any shallow parsing process of a text document. These entities correspond to all the basic concepts that can be found in a document: persons, locations, products, numerical entities, ... Since the MUC evaluation program, numerous methods have been proposed and evaluated for extracting NEs from corpora containing manually annotated entities. These evaluation corpora differ according to the ontology of NE used and the kind of document to process. Very good results can be achieved when using a limited NE set on written documents such as newspaper corpora. However the size of the ontology defining the NEs and the *noise* occurring in the documents to process have a very strong impact on the performance that can be reached. Extracting NEs from speech faces other challenges such as speech disfluencies, Automatic Speech Recognition errors and the lack (or the unreliability) of the punctuations and capital letters. Moreover, the speech transcriptions obtained automatically contain only words belonging to the ASR lexicon, although NEs are essentially made of proper names for which it is very difficult to have an exhaustive list. This paper presents a Named Entity Recognition (NER) method dedicated to process speech transcriptions. The main principle behind this method is to collect in an unsupervised way lexical knowledge for all entries of the ASR lexicon. This knowledge is gathered with two methods: by automatically extracting NEs on a very large set of textual corpora and by exploiting directly the structure contained in the Wikipedia resource. This lexical knowledge is used to update the statistical models of our NER module based on an approach mixing generative (Hidden Markov Models - HMM) and discriminative models (Conditional Random Field - CRF). This approach has been evaluated within the French ESTER 2 evaluation program and obtained the best results at the NER task on ASR transcripts.

This paper is structured as follows: section 2 presents the ESTER NE task, the ontology chosen and the corpora on which the evaluation has been done; section 3 describes the NER system developed at the Universite d'Avignon for participating to ESTER 2; section 4 presents the use of very large unlabelled corpora for acquiring lexical knowledge for the NER models; section 5 shows how the structure of a textual resource such as Wikipedia can be used to update NER models in an unsupervised way and finally section 6 presents the results obtained by our system at the ESTER NE evaluation task.

2. THE ESTER 2 EVALUATION PROGRAM

The French ESTER 2 program [1] was jointly organized by the French-speaking Speech Communication Association (AFCP, French-speaking ISCA Regional Branch) and the French Defense expertise and test center for speech and language processing (DGA/CEP), with the collaboration of the Evaluation and Language resources Distribution Agency (ELDA). This evaluation program was made of three categories of tasks, namely segmentation, transcription and information extraction.

The audio training data given to the participants consisted of about 300 hours of radio broadcast recorded from various French speaking radios: France Inter, Radio France International, France Culture, Radio Classique, Africa number one, Radio Congo and Radio Television du Maroc. The test set, recorded from January to February 2008, consisted of 7 hours of radio broadcast shows taken from the same radios. Most of the data contains broadcast news however talk shows with a lot of spontaneous speech are also in the corpus. Another difficulty is the different French accents that can be found in the French speaking African radios.

We are interested in this paper in the Named Entity Recognition (NER) evaluation of the information extraction task of ESTER 2. Two subtasks were defined: detection on the reference transcriptions and detection on several automatic transcriptions with different word error rates. The NE tag set consists of 7 main categories: persons, locations, organizations, human products, amounts, time and functions. Although not used in this evaluation, 38 sub-categories have been also defined and annotated in the corpus. This tag set is rather complex, more than those used in previous NER evaluations such as MUC7, DARPA HUB5 and CoNLL 2003 shared task. Moreover it has been decided to label each NE according to its use in an utterance. For example the NE "*University of Avignon*" can be considered

[*] The first author performed the work while at the Universite d'Avignon

either as an organization in the sentence: ”*The University of Avignon is delivering a new diploma.*” or as a location in the sentence ”*Let's meet near the University of Avignon*”.

The ESTER 2 corpus has been manually annotated with this tag set. Because of adjustments in the annotation guide, only the test set is now fully compliant with the last version of the annotation guide, the training data contains some mismatch due to a previous version of the guide.

3. MIXING GENERATIVE AND DISCRIMINATIVE METHODS FOR EXTRACTING NES

A lot of methods have been proposed for extracting NEs from texts from rule-based methods to many corpus-based ones. Among these latter, two main approaches have been followed: generative methods such as Hidden Markov Models (HMM) [2] and discriminant methods like MaxEnt [3] or Conditional Random Field (CRF) [4].

For corpus based approaches, NER is seen as a tagging process where a label is given to each word of a sentence for being inside or outside a given entity. By adding position information to these labels (like *Begin*, *Inside*, *Outside* labels in the BIO model), it is possible to retrieve entities spanning over several words. Several studies [5] have shown that CRF outperforms HMM or MaxEnt models for this kind of tagging task. However, as importantly as the tagging method, the choice of the features used to learn the translation between words and labels is crucial.

Two kinds of features can be used to predict a NE label for a given word w_i:

- contextual features on the surface form of the utterance, such as the *preceding word:* w_{i-1} or the *following word:* w_{i+1}, w_i *starts with a capital letter*, w_{i-1} *is a punctuation symbol*, ...
- *a priori* knowledge on w_i, such as w_i *is a city name*, w_i *is a first name*, ...

When dealing with ASR transcriptions, contextual features can be unreliable because of both ASR errors and the lack of formatting of the transcripts: ASR systems output a stream of word with no punctuation and often no word capitalization[1]. To deal with ASR transcripts, several methods have been proposed such as taking into account not only the ASR 1-best but an ASR word lattice [6] or also explicitly encode as features the confidence scores given to each word by the ASR process [7]. Another method proposed in this paper is to increase the weight of the *a priori* knowledge over the contextual features in the NER tagging process.

This *a priori* knowledge can be obtained in dictionaries or gathered from textual resources such as Wikipedia [8], unlabelled textual corpora [9] or even directly from the WEB like in [4]. To each word will be associated one or several semantic labels corresponding to the kind of entity it belongs to in the textual resources used. However to the same word can be associated several semantic label (*Paris* can be a town, an organization, a perfume, a first name, ...) and some errors can occur in the knowledge automatically acquired. Therefore it is important to take into account these ambiguities and errors in the features used to predict a NE label to a word.

We propose in this paper a mixed approach for NER based firstly on a generative process (HMM) to predict semantic and syntactic labels for each word of a sentence; secondly a discriminative process (CRF) is used to effectively retrieve the NEs by using contextual features on the words and their labels. There are two reasons for using first an HMM to predict some of the features used by the CRF: the first one is to simplify the CRF training process by limiting the amount of features (the ambiguities in the word semantic labels are removed by the HMM tagger); the second reason is the simplicity in the integration of various knowledge sources in the probabilistic estimation of a semantic label to a given word with the HMM model.

Word	*POS+sem. label*	*NE+position*
bonjour	NMS	O
investiture	NFS	O
aujourd'hui	ADV	B-TIME
à	PREPADE	O
bamako	XLOC	B-LOC
mali	XLOC	B-LOC
du	PREPDU	O
président	NMS	B-FONC
amadou	XPERS	B-PERS
toumani	XPERS	I-PERS
touré	XPERS	I-PERS
réélu	VPPMS	O
en	PREP	B-TIME
avril	NMS	I-TIME
dernier	AMS	I-TIME

Table 1. Example of the CRF training corpus with the POS and semantic label given by the HMM tagger and the NE+BIO position labels to predict

The HMM tagger used is a simple POS tagger enrich with semantic labels for proper names. Four semantic labels are used: *person*, *organization*, *location* and *product*. To predict the best sequence of labels $t_{1,n}$ on the sequence of n words $w_{1,n}$ (referred as $\tau(w_{1,n})$), we use the following equation:

$$\tau(w_{1,n}) = arg \max_{t_{1,n}} P(t_{1,n}, w_{1,n}) \tag{1}$$

By defining terms such as $t_{1,0}$ and their probabilities, we obtain the general POS equation 2.

$$\tau(w_{1,n}) = arg \max_{t_{1,n}} \prod_{i=1}^{n} P(t_i|t_{i-2,i-1}) P(w_i|t_i) \tag{2}$$

The term $P(w_i|t_i)$ is directly obtained through the maximum likelihood criteria by computing: $C(w_i, t_i)/C(t_i)$ where $C(w_i, t_i)$ is the count of the number of times w_i has been associated to t_i in a training corpus and $C(t_i)$ is the counts of all the words labelled with t_i on the training corpus. By collecting counts on various textual resources, as presented in sections 4 and 5, we can easily model the semantic ambiguity of a given proper names. For example, from all our collected counts, we obtained the following distribution for the proper-name *Marseille*: *LOCATION=32973 ORGANISATION=15731 PERSON=1140 PRODUCT=317.*

The term $P(t_i|t_{i-2,i-1})$ is obtained thanks to a 3-gram language models on the POS and semantic labels, trained on an automatically annotated corpus as explained in section 4.

The CRF NER module is applied on the output of this HMM tagger. The goal of the CRF is to predict a NE label as well as a position to each word of a sentence. We use the BIO (*Begin*, *Inside*, *Outside*) position model. An example of the CRF training corpus can be found in table 1. The CRF toolkit used is *CRF++*[2].

[1] even if the capitalization is provided by the ASR module, it is prone to be erroneous

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

4. UPDATING NER MODELS WITH VERY LARGE UNLABELLED CORPORA

We propose here to use huge collection of textual corpora in order to extract pairs *(proper name, semantic label)* for estimating $P(w_i|t_i)$ as presented in the previous section. To bootstrap the process we need first a lexicon of frequent proper names with semantic labels and a training corpus manually labelled with NE tags to train the CRF NER model. This process is as follows:

1. A POS tagger (including a proper name lexicon) is applied to the NE training corpus;
2. All the proper names not included in the lexicon are labelled with the tag *unknown*
3. The CRF NER model is trained on this corpus with POS and semantic labels.
4. This first NER system is applied to a huge textual corpus (1.3G);
5. In this corpus automatically tagged with NE labels, each proper name w_i belonging to a named entity of type π is labelled with the tag π;
6. We collect the counts of pairs (w_i, π), keep only those above a threshold α[3] and use these counts to update the model $P(w_i|t_i)$ of the HMM tagger. Then the process is iterated at step 1.

5. UNSUPERVISED EXTRACTION OF LEXICAL KNOWLEDGE WITH WIKIPEDIA

Following previous studies on the same topic [8], we propose a method that extracts semantic data from the multilingual encyclopaedic web-resource Wikipedia[4]. We call *metadata* all the information related to an entity, extracted from the encyclopaedic content, represented as the terminology graphs associated with all the entities selected in Wikipedia. These graphs are extracted from five linguistic editions of Wikipedia (English, German, Italian, Spanish and French). All the internal links of Wikipedia (titles, redirect pages, disambiguations pages) are used to generate graphs of surface forms that can be used as features in our NER system.

Each document contained in a linguistic version of Wikipedia can include links to related documents contained in other linguistic editions of Wikipedia. Such link is called *interwiki*, a redirection in Wikipedia, linking an encyclopaedic entry to its equivalent documents in other language corpora of Wikipedia. We use this *interwiki* relation to aggregate in one graph all the possible writing form collected from the five linguistic editions of Wikipedia.

As an example, the graph set for the name *Paris* contains 39 surface forms, (eg. *Ville Lumiere*, *Ville de Paris*, *Paname*, *Capitale de la France*, *Departement de Paris*).

Each graph corresponds to one entity. In order to use these graphs in the NER system developed for the ESTER evaluation we have associated each entity with one of the NE category used in ESTER: person, location, organization and product. To obtain these NE labels for each graph we first have selected a set of entities for which we had the correct NE label according to the ESTER training corpus and other NE lexicon already collected at the lab. This *bootstrap* set of entities contained 413 persons, 741 products, 463 locations and 794 organizations. Then we trained a set of classifiers (SVM, Boostexter and nave bayes classifier) in order to associate the correct NE label to every Wikipedia document related to a given entity. Finally we used the combination of all classifiers in order to label all the Wikipedia documents used to build our entity graphs. To check the reliability of this classification method we extracted a random set of 1000 documents from Wikipedia and manually labelled them. Our classification process obtained an average F-measure of about 90% on this corpus.

NE	from test set	Equ in metadata	Coverage (%)
Pers	1096	483	44%
Org	1204	764	63%
Loc	1218	1017	83%
Prod	59	23	39%

Table 2. Coverage of the entities obtained from Wikipedia on the entities occurring in the ESTER evaluation test corpus

This classification process was applied on 600839 entities[5] extracted from the Wikipedia French corpus[6]. By associating each entity graph with the NE label predicted on the Wikipedia documents that were used to built it, we obtain a NE label for every surface form representing a given entity. For example, the 39 surface forms of the entity *Paris* are associated with the label *location*. Finally we estimate the counts (w_i, π) for all proper name words w_i in the surface forms of an entity of label π and use these counts to update the model $P(w_i|t_i)$ of the HMM tagger.

We have checked the coverage of the entities obtained on Wikipedia on the entities of the test set of the ESTER evaluation used in the experiment section. The results are given in table 2.

6. EXPERIMENTS

All the results presented in this section have been obtained on the ESTER 2 test corpus, manually annotated with NE. There are 5123 NE occurrences in this test corpus.

The official results of the ESTER evaluation program are given in table 3 (see [1] for a complete overview of the evaluation for all tasks). The system presented in this paper is referred as **LIA**. The table reports results in term of Slot Error Rate (SER): the best system is the one obtaining the lowest value of SER. As we can see our system achieves the best results for all tasks on ASR transcriptions with different Word Error Rate (WER). It is interesting to see that on the reference transcriptions (*WER=0*) the participants *part6* and *part7* are far ahead the other participants: less than 10% SER instead of 23.9% for the third system (LIA). These two participants used 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 the reference transcriptions, there is a clear lack of robustness of these models when applied to speech transcripts.

In adjunction to high WER, the two last corpora with *WER=17.8* and *WER=26.1* have no capitalization. This explains the very bad results obtained by most of the systems on them. On the other hand,

[3] We used $\alpha = 5$ in our experiments

[4] The whole metadata set generated can be downloaded and viewed on www.nlgbase.org

[5] The amount of entities differs than the total number of pages in the original Wikipedia corpus, because special pages from Wikipedia like redirections or disambiguation are included in a unique *metadata* entity

[6] Dump reference frwiki-20080323-pages-articles.xml from download.wikipedia.org

system/WER	*WER=0*	*WER=12.1*	*WER=17.8*	*WER=26.1*
LIA	23.9	**43.4**	**51.6**	**56.8**
part2	30.9	45.3	55.5	61.2
part3	37.1	54.0	60.4	65.2
part4	33.7	50.7	80.8	82.9
part5	35.0	55.3	86.5	88.6
part6	9.9	44.9	60.7	66.2
part7	**9.8**	44.6	X	X

Table 3. Official results of the ESTER evaluation program on NE. Slot Error Rate (SER) measures according to the WER in the transcriptions for the 7 participants to the evaluation. The best result is indicated in bold.

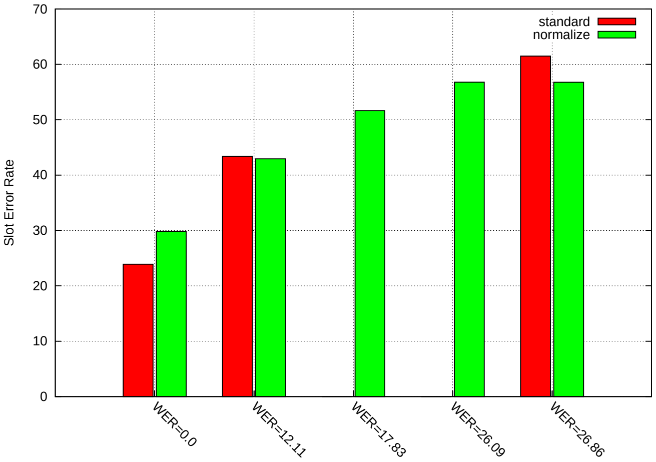

Fig. 1. Slot Error Rate (SER) as a function of the Word Error Rate (WER) for the LIA system. Two conditions are examined: with capitalization (*standard*) and without capitalization (*normalize*)

by retraining all our models on the training corpus without capitalisation, our system remains particularly robust to these conditions. This is illustrated by figure 1: the SER/WER histogram presents the results of our system with two conditions: *standard* where the capitalization produced by the ASR module is kept[7] and *normalize* when the corpora have been processed by the *normalize* tool which remove all punctuation and capitalization. For the *standard* condition we use a version of our NER system trained on a corpus with capitalization and punctuation; for the *normalize* condition we use another version trained on corpora processed by the *normalize* tool. As we can see removing punctuation and normalization to the reference transcription have a strong negative impact on the performance of our system. However, for all the ASR transcriptions, our system performs better if we remove the capital letters predicted by the ASR module. This gain increases with higher WER.

We have also evaluated the gain obtained by our two unsupervised lexical knowledge acquisition methods. The baseline system using only the ESTER training corpus as well as the proper names lexicon already collected at the lab obtained 28.2% of SER on the reference transcriptions. By adding to the models the knowledge collected in Wikipedia we obtained a 2% absolute gain leading to a SER of 26.2%. Another absolute 1% reduction was further reached by using 1.3G of unlabelled text, as presented in section 4. The final result of 23.9 was obtained by applying post-processing rules on the span of the entities detected to be compliant with the last changes in the ESTER annotation guidelines.

7. CONCLUSION

We have presented a Named Entity Recognition (NER) system dedicated to process speech transcriptions. This system mixes both generative and discriminative classification methods to increase its robustness to ASR errors. The use of unlabelled data and the Wikipedia resource for automatically updating the HMM models proved to reduce the Slot Error Rate of our system. Finally our approach obtained the best results at the NER task on ASR transcripts of the ESTER 2 evaluation program.

8. REFERENCES

[1] Sylvain Galliano, Guillaume Gravier, and Laura Chaubard, “The Ester 2 Evaluation Campaign for the Rich Transcription of French Radio Broadcasts,” in *Interspeech 2009*.

[2] Daniel M. Bikel, Richard L. Schwartz, and Ralph M. Weischedel, “An algorithm that learns what’s in a name,” 1999, vol. 24, pp. 211–231.

[3] Andrew Brothwick, John Sterling, Eugene Agichtein, and Ralph Grishman, “Exploiting diverse knowledge sources via maximum entropy in named entity recognition,” in *6th Workshop on Very Large Corpora (ACL ’98)*, Montréal, 1998.

[4] A. McCallum and W. Li, “Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons,” in *Seventh Conference on Natural Language Learning (CoNLL)*, 2003.

[5] Christian Raymond and Giuseppe Riccardi, “Generative and discriminative algorithms for spoken language understanding,” in *International Conference on Speech Communication and Technology (Interspeech)*, 2007, vol. 2.

[6] B. Favre, F. Béchet, and P. Nocéra, “Robust named entity extraction from large spoken archives,” in *Conference on Human Language Technology (HLT) and Empirical Methods in Natural Language Processing (EMNLP)*. 2005, pp. 491–498, Association for Computational Linguistics Morristown, NJ, USA.

[7] K. Sudoh, H. Tsukada, and H. Isozaki, “Incorporating speech recognition confidence into discriminative named entity recognition of speech data,” in *44th annual meeting of the Association for Computational Linguistics*. Association for Computational Linguistics Morristown, NJ, USA, 2006, pp. 617–624.

[8] J. Kazama and K. Torisawa, “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)*. Association for Computational Linguistics, 2007, pp. 698–707.

[9] T. Hori and A. Nakamura, “An Extremely Large Vocabulary Approach to Named Entity Extraction from Speech,” in *2006 IEEE International Conference on Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings*, 2006, vol. 1.

[7] except for the *WER=12.1* and *WER=17.8* corpora which don’t have punctuation