Mining Broadcast News data: Robust Information Extraction from Word Lattices

Benoît Favre[1], *Frédéric Béchet*[2], *Pascal Nocéra*[2]

[1]Thales, MMP Laboratory, 160 Bd Valmy
92700 Colombes, France
benoit.favre@fr.thalesgroup.com

[2]LIA, 339 chemin des Meinajaries
84911 Avignon Cedex 9, France
{frederic.bechet,pascal.nocera}@univ-avignon.fr

Abstract

Fine-grained information extraction performance from spoken corpora is strongly correlated with the Word Error Rate (WER) of the automatic transcriptions processed. Despite the recent advances in Automatic Speech Recognition (ASR) methods, high WER transcriptions are common when dealing with unmatched conditions between the documents to process and those used to train the ASR models. Such mismatch is inevitable in the processing of large spoken archives containing documents related to a large number of time periods and topics. Moreover, from a text indexation point of view, rare events and entities are often the most interesting information to extract as well as the ones that are very likely to be poorly recognized. In order to deal with high WER transcriptions this paper proposes a robust Information Extraction method that mines the full ASR search space for specific entities thanks to a 3-steps process: firstly, adaptation of the extraction models thanks to metadata information linked to the documents to process; secondly transduction of a word lattice outputs by the ASR module into an entity lattice; thirdly a decision module that scores each entity hypothesis with different confidence scores. A first implementation of this model is proposed for the French Broadcast News Named Entity extraction task of the evaluation program ESTER.

1. Introduction

Speech Mining is a research field that aims to bridge the gap between Speech Processing and Text Data Mining methods. One of the main differences between Information Extraction from written text input and spoken transcription input is the uncertainty in the automatic transcriptions. This uncertainty is due on one hand to the lack of robustness of Automatic Speech Recognition (ASR) methods to new conditions (different from those used to train the acoustic and linguistic models) and on the other hand to the dysfluencies occurring in spontaneous speech. One way to deal with this uncertainty is to consider not only the best transcription output by the ASR module but the whole set of alternate transcriptions that can be found in the word lattices output by most of the ASR systems.

Following these remarks this paper presents an Information Extraction method that explicitly takes advantage of the generation of alternate transcriptions of a speech signal in order to increase its robustness to high Word Error Rate (WER) transcriptions. More precisely the model proposed in this paper focuses on:

- firstly defining what kind of information or entity is likely to occur in a given speech signal (thanks to *metadata* information);
- secondly looking explicitly in the ASR search space for occurrences of each of these entities, each of them scored by confidence scores.

A first implementation of this model is proposed within the framework of the French evaluation program ESTER [10]. The task consists in extracting named entities from radio broadcast news data.

2. Information extraction and Automatic Speech Recognition

The NIST Topic Detection and Tracking (TDT) and TExt Retrieval Conference (TREC) evaluation programs have studied the impact of recognition errors on the overall performance of Information Extraction systems for tasks like story segmentation, topic detection and text retrieval [9, 2]. The *redundancy effect* of salient words representing topics helped to limit the impact of recognition errors compared to performance obtained on written text. However, if this phenomenon is true for Information Extraction tasks that rely on long text segments, this is not true for tasks related to the extraction of fine grained entities, like Named Entities (NE), where the extraction performance is strongly correlated with the ASR Word Error Rate (WER).

Despite the continuous improvement of ASR techniques, high WER transcriptions are inevitable in difficult conditions like those found in large spoken archives like in the MALACH project [13] or in the speech archives of national archive institutions like the French *Institut National de l'Audiovisuel* (INA). Moreover, Named Entities extraction performance is greatly affected by a mismatch between training and testing data. This is due mainly because proper names, which represent most of the Named Entity items, are a very dynamic category of words, strongly related to the period of time representing the documents to process. This problem can be tackled by using *metadata* information on these documents, like news related to the same period of time or abstract describing the document content. This paper presents an Information Extraction method that uses such metadata information in order to update the NE models that are used in the extraction process from word lattices.

3. The Named Entity extraction task of the ESTER program

The ESTER program is the first French Rich Transcription of Broadcast News program organized by *l'Association Francophone de la Communication Parlée* (AFCP), *la Délégation Générale pour l'Armement* (DGA) and the Evaluation Language resources Distribution Agency (ELDA). One of the task is Named Entity extraction over a tag set of eight main cate-

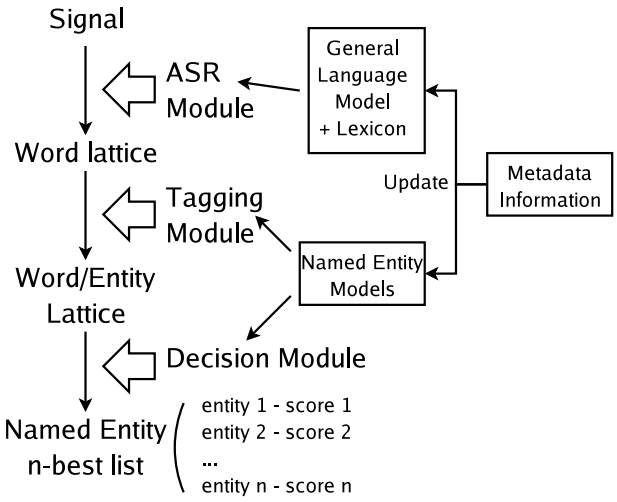

Figure 1: *Architecture of the Named Entity extraction system*

gories and more than thirty sub categories organized as below:

- persons (**pers**): human, fiction characters, animals;
- locations (**loc**): geographical, traffic lines, electronic and real addresses, dial numbers;
- organizations (**org**): political, business, non profit;
- socio-political groups (**gsp**): clans, families, nations, administrative regions;
- amounts (**amount**): durations, money, lengths, temperature, age, weight and speed;
- time (**time**): relative and absolute time expressions, hours;
- products (**prod**): art, printings, awards and vehicles;
- facilities (**fac**): buildings, monuments.

Let's point out that this tag set is significantly larger that the one used in the MUC-7 and DARPA HUB-5 NE extraction task, where only three categories were considered, and therefore the results obtained can't be directly compared. Indeed, several categories are highly ambiguous (e.g. administrative regions and nations in **gsp** and facilities overlap with locations, durations overlap with time entities, ...).

The ESTER corpora are made of manual transcriptions of various radio broadcast shows of 6 French speaking radios. The training corpus contains 1M words and 52K NE occurrences. The test corpus contains 100K words with 6K NEs and has a 6 months time mismatch with any training data that could be used to train the models, as specified by the ESTER program rules. This will allow us to see the impact of time mismatch on the NE extraction results.

4. Architecture of the system

The general architecture of the Named Entity extraction method is presented in figure 1. The Named Entity models are context-free grammars (hand written and data induced) coded by Finite State Machines (FSM) thanks to the FSM AT&T toolkit [11]. There are three main processes in this method which are going to be discussed in the next sections:

1. updating the ASR models (language model + lexicon) and the Named Entity models (grammars + word lists) with *metadata information* related to the documents to process;
2. composing the Named Entity models with the word lattice in order to obtain an entity lattice, with all the entities that can be extracted from it;
3. choosing the most reliable entities detected in the word lattice and sorting them according to confidence scores.

4.1. Updating Language and NE models with metadata information

Broadcast News corpora are very time dependent and therefore it is crucial to train the ASR models on data temporally very close to the documents to process. However this can be difficult when dealing with large spoken archives spanning over a long period of time. But if in one hand it is often not possible to collect enough data for training new models for each time period considered, on the other hand some metadata information related to these periods is very likely to be available. Metadata information can be a description or an abstract of the spoken document, text corpus containing newspaper data from the same date of the document, or documents related to the same topic as the targeted documents.

Several studies [8, 5] propose adaptation methods of a general language model to the possibly small corpora extracted from these kinds of metadata information (an overview of all these methods can be found in [4]). It appears from these studies that the choice of the metadata and the size of the adaptation corpus collected are critical in this adaptation process: if the adaptation corpus is not exactly related to the topics of the document to process, no real gains are obtained.

Similarly, adapting dynamically the ASR lexicon is a much more practical approach than increasing indefinitely the lexicon size as a lot of words are strongly linked to a given time period or topic and covering all time periods and all topics is unrealistic. A recent study [1] proposes to add new words to an existing lexicon without modifying the language model by using open classes, integrated in the LM, corresponding to the Part-Of-Speech labels of the new words to add. This approach has the advantage to be more robust to a mismatch between the adaptation corpus and the document to process, as the main LM remains intact.

The adaptation process presented in this paper is done for both the LM and lexicon models and the NE models, as follows:

- a text corpus corresponding to the same time period as the documents to process is extracted from two French newspapers (*Le Monde* and *Le Monde Diplomatique*);
- the ASR lexicon is enriched with all the new proper names detected in this collected corpus and a linear interpolation between this general LM and the one obtained on this adaptation corpus is done during the ASR process;
- a NE extraction system (based on the same tools as presented in this paper), working on text input, is applied to the adaptation corpus in order to collect all the entities that are likely to occur in the documents targeted. The entities detected are added to the NE grammars and wordlists used to extract all the possible entities from the word lattices.

4.2. From a word lattice to an entity lattice

The ASR module outputs a word lattice ($W_{lattice}$) coded as a FSM. The words belong to the ASR vocabulary V and the scores are the likelihood scores combining the acoustic and linguistic model scores. Let G be a context-free right regular grammar, coded as a transducer FSM, made of the union of local Named Entity grammars and a background filler model. The input symbols of this transducer are words from V and the output symbols are words and tags from $V \cup T_{open} \cup T_{close}$ (T_{open} is an opening tag list of Named Entity labels and T_{close} the respective closing tag list). Any word string (with words belonging to V) can be accepted by G, thanks to the filler model, and any substring that matches a NE grammar outputs a beginning and ending tag as well as the words part of the entity detected. Words outside entities are deleted. The composition of the word lattice and the entity grammar results in a word+tag transducer called $E_{lattice}$:

$$E_{lattice} = W_{lattice} \oplus G$$

By enumerating the n-best best paths *on the output symbols* of $E_{lattice}$ one obtains the n-best list of NE occurring in the word lattice, and by enumerating the n-best paths *on the input symbols* for a given NE detected, one obtains the n-best support word strings containing the entity.

4.3. Named Entity n-best list generation

The last step in the NE extraction process is the output of a NE n-best list sorted according to a confidence score. This is done by generating the n-best paths on the output symbols of $E_{lattice}$ and scoring them according to their posterior probabilities obtained on the word lattice $E_{lattice}$. Posterior probabilities from word graphs have proven to be efficient confidence measures ([7]), they are estimated in this study as follows: let e be an entity, W_i a word string accepted by $E_{lattice}$, W_i^e a word string accepted by $E_{lattice}$ and emitting the entity e, and Y a string of acoustic vectors,

$$P(e|E_{lattice}) = \frac{\sum_{W_i^e \in E_{lattice}} P(Y|W_i^e) \times P(W_i^e)}{\sum_{W_i \in E_{lattice}} P(Y|W_i) \times P(W_i)}$$

The scores involved in this confidence measure calculation rely only on the ASR models as the NE grammars used are non-weighted. We chose not to merge ASR scores and Named Entity model scores into the same search process: the n-best list of NE is obtained only with ASR confidence scores, then a decision module evaluate each hypothesis thanks to models trained to reject, re-label or accept any entity previously detected. This decision module is presented in the next section.

5. Decision module

The decision module is the last step in the NE extraction process, as presented in figure 1. Its goal is firstly to choose among all the possible segmentations of a NE string the one that is more likely to be correct, thanks to a NE tagger; secondly to apply a text classifier to the NE strings obtained, with their occurrence contexts, in order to give to each entity its final label, thanks to a discriminative process. These two steps are briefly described in the next sections.

5.1. Named Entity tagger

To each NE hypothesis e output from the transducer $E_{lattice}$ is associated the best support word string W_{best}^e, which is the best path on the input symbols (words) of $E_{lattice}$ that emits e. All the entities e that share the same word string W_{best} can be seen as different NE segmentations of W_{best}. For example, to the word string `near the Eiffel tower`, two segmentations can be made with the NE grammars: `near the <person> Eiffel </person> tower` and `near the <fac> Eiffel tower </fac>`. In order to choose the most probable one, a NE tagger based on an n-gram language model following the method presented in [3] has been implemented. All the different segmentations of W_{best} are then processed by the tagger in order to give to each of them another confidence score based this time only on NE models.

5.2. Decision tree based text classifier

It has been shown that text classification methods can be very efficient tools for labeling NE contexts [6]. The method consists in building a training corpus where each example is made of two word strings, the left and right word contexts of an entity, some information about the entity itself, and the NE label of the entity according to a reference corpus. By training a classifier to discriminate the entities according to their labels, one can build a classification tool that can process unseen entities and be more accurate than the NE grammars alone as they don't incorporate any contextual information. This is particularly true for very ambiguous NE like gsp, facilities or locations. A boosting based classifier called *BoosTexter* [14] has been used in this last step. All the NEs of the n-best list are labeled according to the classifier and a third confidence measure is attached to each of them: the confidence score given by the classifier to the NE label to be supported by the entity and its context.

6. Experiments

The experiments have been carried out on the corpora of the French Broadcast News transcription program ESTER. The ASR engine used is the toolkit SPEERAL [12] developed at the computer lab (LIA) of the University of Avignon. To measure the NE extraction performance, the ESTER organizers use an error measure called the Slot Error Rate (SER). This measure is defined as follows:

$$SER = \frac{\sum_{e \in \mathcal{E}} \alpha_e |e|}{|Refslots|}$$

where $e \in \mathcal{E}$ is one of the following error types: inserted slot, deleted slot, type error, extent error, type+extent error, multiple errors. Each error type e has a particular weight noted α_e. Every reference slot can only belong to one error category and ASR errors occurring in a NE slot are counted as extent errors.

For the evaluation purpose we added a simple decision rule to the NE extraction strategy proposed in this paper: from the NE n-best list presented in section 4.3, only the NEs that have the highest posterior probabilities are kept. Then the corresponding support word string is processed by the NE tagger and the final label for each remaining NE is given by the text classifier BoosTexter.

Four contrastive results are presented in table 6 corresponding to the following test conditions:

- **ASR** for experiments done on automatic transcription;
- **REF** for reference transcriptions;
- **Corpus**$_1$ that corresponds to the development corpus provided during the ESTER program;

- **Corpus**$_2$ that is the test corpus of the ESTER program.

The WER of the automatic transcriptions on the two corpora is also displayed.

The main difference between **Corpus**$_1$ and **Corpus**$_2$ is their temporal mismatch. **Corpus**$_1$ is very close, from a temporal point of view, to the ESTER training corpus used to adapt the ASR and NE models to the task. For **Corpus**$_2$ there is a 6 months temporal mismatch between this corpus and any training or adaptation data used, as specified by the ESTER evaluation rules.

Corpus	**SER(ASR)**	**SER(REF)**	**WER(ASR)**
Corpus$_1$	39.9	22.7	21.2
Corpus$_2$	57.4	34.1	26.4

Table 1: Slot Error Rate (SER) and Word Error Rate (WER) performance on two ESTER corpora

These are the first results presented for fine-grained NE extraction for French. They still have to be compared, at the F-measure level, with other studies using also an extended NE tag set for other languages. As we can see in table 6, the impact of both ASR errors and temporal mismatch is very significant for SER performance. The degradation observed on the automatic transcriptions is due to ASR errors on the entities alone but also on their immediate context, preventing the decision module to give the correct tag to the entities detected. The impact of the temporal mismatch between the training and the test corpora is even bigger. As we can see the relative increase in WER between Corpus$_1$ and Corpus$_2$ is 24.5% while the relative increase in SER is 43.8%. This confirms our point that Information Extraction performance is much more affected by temporal mismatch than word transcription performance and it really points out the importance to carefully adapt the NE models to the task targeted. Even a relatively small temporal mismatch (6 months) can have dramatic effects on the performance.

7. Conclusion

This paper presents a Named Entity extraction method applied to Broadcast News data within the framework of the French ESTER evaluation program. The main features of this method are: the explicit use of alternate transcriptions by extracting directly the NEs from word lattices; the adaptation of the NE models to the document targeted thanks to metadata information; the decision module that combines a generative approach based on a statistical NE tagger and a discriminative approach based on a text classifier. The first results obtained at the ESTER evaluation program validate this approach. They have pointed out the absolute necessity of adapting the NE models whenever a mismatch in the conditions is observed between the training data and the documents to process. Improving this adaptation process and the gathering of metadata information are the immediate perspectives of this work.

8. Acknowledgments

This work is supported by the French *Ministère de la Recherche et de l'Industrie* under the CIFRE grant number 692/2003.

9. References

[1] Alexandre Allauzen and Jean-Luc Gauvain. Open vocabulary ASR for audiovisual document indexation. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, pages 1013–1016, 2005.

[2] J. Barnett, S. Anderson, J. Broglio, M. Singh, R. Hudson, and S. Kuo. Experiments in spoken queries for document retrieval. In *In Eurospeech 97*, pages 1323–1326, 1997.

[3] F. Bechet, A. Gorin, J. Wright, and D. Hakkani-Tur. Detecting and extracting named entities from spontaneous speech in a mixed initiative spoken dialogue context: How May I Help Y. *Speech Communication*, 42:207–225, 2004.

[4] Jerome R. Bellegarda. Statistical language model adaptation: review and perspectives. *Speech Communication*, 42 Issue 1:93–108, 2004.

[5] Langzhou Chen, Jean-Luc Gauvain, Lori Lamel, and Gilles Adda. Dynamic language modeling for broadcast news. In *In International Conference on Speech and Language Processing*, pages 1281–1284, 2004.

[6] Michael Collins and Yoram Singer. Unsupervised models for named entity classification. In *Empirical Methods in NLP processing and Very Large Corpora - EMNLP-VLC'99*, University of Maryland, 1999.

[7] D. Falavigna, R. Gretter, and G. Riccardi. Acoustic and word lattice based algorithms for confidence scores. In *Proceedings of International Conference on Spoken Language Processing (ICSLP)*, pages 1621–1624, 2002.

[8] M. Federico and N. Bertoldi. Broadcast news LM adaptation using contemporary texts. In *Proceedings of European Conference on Speech Communication and Technology (Eurospeech)*, pages 239–242, Aalborg, Denmark, 2001.

[9] Jonathan G. Fiscus and George R. Doddington. Topic detection and tracking evaluation overview. *Topic detection and tracking: event-based information organization*, pages 17–31, 2002.

[10] G. Gravier, J.F. Bonastre, E. Geoffrois, S. Galliano, K. McTait, and K. Choukri. ESTER, une campagne d'évaluation des systèmes d'indexation automatique d'émissions radiophoniques en fixme: franais. In *Proc. Journées d'Etude sur la Parole (JEP)*, 2004.

[11] Mehryar Mohri, Fernando Pereira, and Michael Riley. Weighted finite-state transducers in speech recognition. *Computer, Speech and Language*, 16(1):69–88, 2002.

[12] P. Nocera, G. Linares, and D. Massonie. Principes et performances du décodeur parole continue Speeral. In *Proc. Journées d'Etude sur la Parole (JEP)*, 2002.

[13] Bhuvana Ramabhadran, Jing Huang, and Michael Picheny. Towards automatic transcription of large spoken archives - english ASR for the MALACH project. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, pages 216–219, 2003.

[14] Robert. E. Schapire and Yoram. Singer. BoosTexter: A boosting-based system for text categorization. *Machine Learning*, 39:135–168, 2000.