Incorporating Named Entity Recognition into the Speech Transcription Process

Mohamed Hatmi, Christine Jacquin, Emmanuel Morin, Sylvain Meignier

Incorporating Named Entity Recognition into the Speech Transcription Process

Mohamed Hatmi[1], *Christine Jacquin*[1], *Emmanuel Morin*[1], *Sylvain Meignier*[2]

[1] LINA, University of Nantes, France,
[2]LIUM, University of Le Mans, France
{mohamed.hatmi, christine.jacquin, emmanuel.morin}@univ-nantes.fr
sylvain.meignier@lium.univ-lemans.fr

Abstract

Named Entity Recognition (NER) from speech usually involves two sequential steps: transcribing the speech using Automatic Speech Recognition (ASR) and annotating the outputs of the ASR process using NER techniques. Recognizing named entities in automatic transcripts is difficult due to the presence of transcription errors and the absence of some important NER clues, such as capitalization and punctuation. In this paper, we describe a methodology for speech NER which consists of incorporating NER into the ASR process so that the ASR system generates transcripts annotated with named entities. The combination is achieved by adapting ASR language models and pre-annotating the pronunciation dictionary. We evaluate this method on ESTER 2 corpus, and show significant improvements over traditional approaches.

Index Terms: Named Entity Recognition, Automatic Speech Recognition, language modeling, ASR vocabulary

1. Introduction

Named Entity Recognition (NER) from speech is mainly performed by transcribing speech and then applying NER techniques to transcripts. The Person, Organization and Location names are the main lexical units to be located and classified. NER systems are generally categorized into whether they are based on symbolic or learning methods [1]. Both types of systems are adapted to fit in with the characteristics of automatic speech transcripts. NER systems have to face the problems of graphic ambiguity (lack of capitalization), segmentation ambiguity (lack of punctuation), and speech disfluencies [2, 3]. This deprives the exploitation of some vital NER features. Moreover, automatic speech transcripts are noisy due to Automatic Speech Recognition (ASR) errors and out-of-vocabulary (OOV) problems. ASR errors occurring in words constituting the named entities or in their word context have a direct impact on the NER performance [4]. Previous work to improve speech NER has focused on ASR outputs. It has included restoring punctuation and capitalization in transcripts [5], incorporating indicative OOV words and ASR confidence features [6, 7, 8], or using intermediate ASR outputs such as N-best hypothesis [9] and word lattices [10] instead of only relying on 1-best hypothesis. Few studies have focused on NER at the ASR level [11].

In this work, we propose to go upstream into the speech transcription process and directly integrate the NER task so that the ASR system generates transcripts annotated with named entities. Our hypothesis is that we can assign *a priori* named-entity tags to certain words at the ASR level since ASR vocabulary is closed. The words that are not in this closed vocabulary will not appear in transcripts. In fact, ASR vocabulary words are selected from the corpora used to train ASR language models. These corpora are composed of small quantities of manual transcriptions of speech and relatively larger quantities of newspaper archives. The content has to be comparable to the domain targeted by the ASR process. Therefore, named entities encountered in automatic transcripts should have the same tags and limits as in the training corpora of language models. Thus, annotating these corpora using a state-of-the-art named-entity recognizer will allow to determine the candidate named-entity tags for certain ASR vocabulary entries (for example, pre-tagging "Obama" as a person). Named-entity recognizers give good performance on well-written texts. Retraining ASR language models using the annotated corpora will constrain the ASR system to generate syntactically correct output annotated with named entities.

This paper is organized as follows: Section 2 briefly discusses the prior work in the field of speech NER. Section 3 describes the LIUM speech transcription system used in this work. Section 4 presents the corpus used for evaluation. Section 5 presents the method to integrate NER into the ASR process. Section 6 reports experimental results, while section 7 concludes and presents future work.

2. Related work

Three main approaches exist in the literature to improve speech NER. The first is to incorporate ASR features into the NER tagger. In [7], an ASR confidence feature is employed to indicate whether each word has been correctly recognized. Automatic transcriptions tagged with named entities are used to model ASR errors. The goal is to reject named entities with ASR errors thereby increasing NER precision. Experiments show a gain in precision of 7.46 %. Recent work [8] has proposed to include features indicative OOV words. A CRF-based tagger exploits the output of an OOV detector in order to identify and ignore regions containing incorrectly transcribed named entities. This allows an improvement in F-measure from 58.5 to 60.7 %. The second approach consists of exploiting intermediate ASR outputs in order to broaden the search space. In [9], an NER system based on maximum entropy is used to annotate the N-best ASR hypothesis. Then a weighted voting based on ASR and NER scores is made to select the most probable named entities, even if they do not occur in the 1-best ASR hypothesis. Experimental results show an improvement of 1.7 % in F-measure. Other work [10] has proposed directly to recognize named entities in the word lattice. The used named entity grammars integrate the words belonging to the ASR lexicon and

exploit the whole ASR word lattice in order to extract the N-best list of named entity hypotheses. The ASR and NER scores are attached to each named entity hypothesis. Experimental results show an improvement of 1 % in F-measure. The third approach consists of annotating named entities at the ASR level by using an extremely large vocabulary lexicon [11]. Named entities are incorporated as compound words into the lexicon and the language model. This considerably increases the size of the vocabulary (1.8 million words). A one-pass ASR system is used to transcribe the annotated named entities. 500 Japanese spoken queries for a question-answering system are used for evaluation. Results shows an improvement of 2.4 % in F-measure.
As in [11], we propose to integrate the NER task directly into the speech transcription process instead of dealing with ASR outputs. However, the fundamental difference in our approach is that the NER task is performed at the word level. This results a wider coverage of named entities mainly for entities composed of common nouns like amount and time names and a better control of the size of the vocabulary. In addition, we work with a multi-pass ASR system using limited vocabulary size. The results we have obtained are compared with those of a state-of-the-art NER system.

3. The LIUM speech transcription system

The LIUM speech transcription system for the French news [12, 13] is based on the CMU Sphinx system. Many distributed tools in the CMU Sphinx open-source package have been supplemented and adapted to improve the transcription performance. The transcription process is based on multi-pass decoding involving five passes:

- The first pass uses a trigram language model and an acoustic model corresponding to the gender and the bandwidth.
- The second pass applies a Constrained Maximum-Likelihood Linear Regression (CMLLR) transformation for each speaker based on the best hypotheses generated by the first pass, and word-graphs are generated using SAT and Minimum Phone Error (MPE) acoustic models and CMLLR transformations.
- The third pass rescores the word-graphs of the second pass using a full triphone context with a better acoustic precision, particularly in inter-word areas. New word-graphs are generated.
- The fourth pass updates the linguistic scores of the new word-graphs using a quadrigram language model.
- The last pass transforms the word-graphs of the fourth pass to a confusion network, and generates the *1*-best hypothesis.

3.1. Acoustic models

The acoustic models for 35 phonemes and 5 kinds of fillers are trained using 240 hours of transcribed French news from ESTER 1 & 2 campaigns [14] [15]. Models for the first pass are composed of 6,500 tied states. Models for other passes are composed of 7,500 tied states.

3.2. Vocabulary

The vocabulary is built by generating a unigram model as a linear interpolation of unigram models trained on corpora presented in Table 1. The linear interpolation was optimized on the ESTER 2 development corpus in order to minimize the perplexity of the interpolated unigram model. Then, the first 122,981 probable words from this language model were extracted.

Table 1: *Training corpora used to create ASR langage models*

	Period	**No. of words**
AFP corpus	1994-2006	488,929,004
APW corpus	1994-2006	173,598,873
Le Monde corpus	1994-2004	335,446,061
Afrik corpus	2007	6,319,708
l'Humanité corpus	1990-2007	63,624,367
Web corpus	2007	9,617,468
Ester corpus	2007	3,249,228

3.3. Language models

The trigram and quadrigram backoff language models are trained on corpora presented in Table 1 with modified Kneser-Ney smoothing using SRILM toolkit [16]. No cut-off is applied on trigrams and quadrigrams. The linear interpolation is optimized on the ESTER 2 development corpus. The models are composed of 122,981 unigrams, 29,042,901 bigrams, 162,038,928 trigrams and 376,037,558 quadrigrams.

4. Corpus description

To carry out the experiments, we used the ESTER 2 test corpus available in two modalities:

- The audio resources containing 26 French broadcasts, recorded from January to February 2008. Most of these are the news from four different sources: France Inter, Radio France International (RFI), Africa 1 and TVME.
- The textual resources consisting of manual transcriptions of audio resources (72,534 words). Named entities were annotated manually according to a taxonomy consisting of 7 main categories: Person, Location, Organization, Human Product, Amount, Time and Function. There are 5,123 named entities in these manual transcriptions.

This corpus is divided into two parts, the development part (*DevPart corpus*) which is used to adjust some parameters (10 broadcasts) and the test part (*TestPart corpus*) which is used to evaluate our approach (16 broadcasts).

5. Integrating NER into the ASR process

The proposed method relies on the fact that ASR vocabulary is known and closed. The ASR language models trained with this vocabulary represent a mirror of which can appear in automatic transcripts. Thus, named entities encountered in transcripts should keep the same tags as those encountered in the data used to train language models. When dealing with ASR outputs, NER performance is greatly affected by both ASR errors and the lack of punctuation and capitalization. To avoid these problems, we propose to annotate named entities at ASR level. Integrating ASR and NER processes allows the ASR system to generate transcripts annotated with named entities. Then the basic recognition problem becomes to find the most likely sequence of words tagged with named entities $((\hat{W}, \hat{E}) = (w_1, e_1), (w_2, e_2), (w_3, e_3), ..., (w_k, e_k))$ given a sequence of sounds $(X = x_1, x_2, x_3, ..., x_p)$:

$$(\widehat{W}, \widehat{E}) = \arg\max_{w} \mathbb{P}(W, E|X) \quad (1)$$

To achieve this, we have relied on the LIUM speech transcription system described in Section 2. This system shows 19.2 % of Word Error Rate (WER) on the Ester 2 test corpus. We used the same acoustic models as the base system.

5.1. Corpora annotation

We first annotated automatically the corpora used to create ASR language models and presented in Table 1 with named entities. For that, we used the named entity tagger LIANE [3]. This tagger is based on a combination of a generative and a discriminative model. At first, a Hidden Markov Model (HMM) based model is used to predict part-of-speech tags. Then a Conditional Random Field (CRF) based model is used to effectively annotate named entities. Graphical features have been exploited to boost LIANE performance. The reason we chose this system is because it obtained the best results during the ESTER 2 French evaluation campaign on automatic transcriptions. LIANE obtained 23.9 % of Slot Error Rate (SER) [17] on manual transcriptions and 51.6 % of SER on automatic transcriptions (17.83 % of WER) [14].
We then encoded named entities in BI notation. BI notation identifies the boundaries and the category tags of phrases that make up the named entities: words outside of named entities are not tagged, while the first word in a named entity is tagged with "entity-tag-B" for the beginning, and further named entity words are tagged with "entity-tag-I" for the inside. The position allows to distinguish boundaries when several named entities of the same category are listed side by side.
Here is an example of the application of the BI notation:

*Il est vingt-**time-B** heures-**time-I** à Paris-**location-B**.*
*Le journal, Denis-**person-B** Astagneau-**person-I**.*
(*It is twenty-**time-B** hours-**time-I** in Paris-**location-B**.*
*The journal, Denis-**person-B** Astagneau-**person-I**.*)

5.2. Vocabulary annotation

In order to annotate the ASR vocabulary of the baseline LIUM system, we assigned to each vocabulary word all the tags it appears with in annotated corpora. Tags include the category tag and the location of the word within the named entity. For instance, the tags "washington-**location-B**", "washington-**location-I**", "washington-**organization-B**", "washington-**organization-I**", "washington-**person-B**", "washington-**person-I**" are associated to "washington". Words not belonging to any named entity are not tagged. In the final, the vocabulary size has increased from 122,981 to 503,192 words.
However, the NER system produces some annotation errors. Words constituting the erroneous-tagged named entities are incorporated into the vocabulary.
In order to select the optimal annotated vocabulary, we relied on the hypothesis that the erroneous tags occur much less frequently than the correct tags of a target word in annotated corpora. For example, the adjective "footballistique" (footballing) appears 88 times without any tag and just once as a person (footbalistique-**person-B**) because of one annotation error. Therefore, we retrained a unigram model as a linear interpolation of unigram models using the annotated vocabulary (503,192 words). Interpolation weights were optimized by minimizing the perplexity on the manually-annotated ESTER 2 development corpus encoded with BI notation. The first N most likely annotated words appearing in the corpora are retained. In the experiments we carried out, N has been varied according to the word probabilities.

5.3. Pronunciation dictionary adaptation

Pronunciation dictionary adaptation consists of assigning one or more pronunciation variants to the selected annotated words. Adding tags has no effect on the pronunciation. Annotated words retain the same pronunciations without tags in the base pronunciation dictionary. For instance, "nn an tt" for "nantes-**location-I**" and "nantes-**organization-B**".

5.4. Language models adaptation

Once the annotated vocabulary selected, we retrained the language models in order to constrain the ASR system to generate syntactically correct transcripts annotated with named entities. Language models also serves to select the appropriate tagged word when the word has several potential tags. We therefore recreated the trigram and the quadrigram backoff language models for each corpus augmented with named entity tags (the same corpora presented in Table 1). No cut-off was applied. The linear interpolation was optimized on the ESTER 2 manually-annotated development corpus encoded in BI notation.

5.5. Annotated vocabulary selection

We used the *DevPart corpus* in order to adjust the size of the annotated vocabulary and to select the most likely annotated words. The size N of the vocabulary is selected according to the word probabilities.
The WER has a direct impact on the NER performance [4]. So we started by evaluating the effect of integrating NER into the ASR process on the ASR performance. In order to precisely evaluate the ASR performance, named-entity tags have been removed from the transcriptions provided by the LIUM system integrating NER.

Figure 1: *Effect of the choice of the annotated vocabulary on the transcription quality computed on the DevPart corpus*

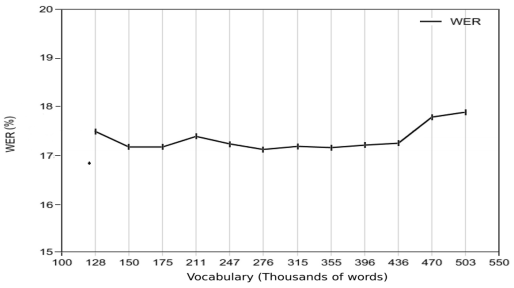
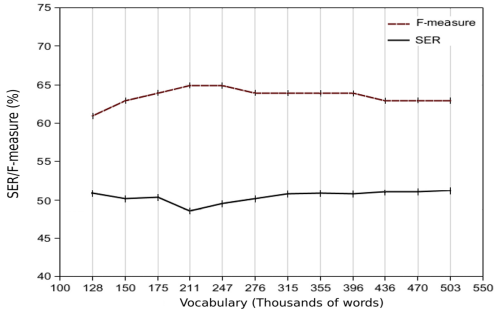


Figure 1 shows the WER obtained for different sizes N of the annotated vocabulary. Using the baseline ASR vocabulary (N = 122,981 words), the LIUM system shows a WER of 16.88 %. With all the annotated vocabulary (N = 503,192 words), the WER increases by 1 %. Reducing the size of the annotated vocabulary allows to decrease slightly the WER. We conclude that incorporating the NER task into the speech transcription process does not affect the transcription quality in a significant way.
The evaluation of the NER performance is performed using the SER and the F-measure. The SER combines the different types of error: substitutions, deletions, and insertions. The F-measure combines precision and recall.
Figure 2 shows the SER and the F-measure for different thresh-

Figure 2: *Effect of the choice of the annotated vocabulary on NER quality computed on the DevPart corpus*



olds of the annotated vocabulary size. In any case, the annotated vocabulary covers all the words appearing in the base vocabulary. Using all the annotated vocabulary, the system shows 51.23 % of SER and 63 % of F-measure. Filtering the erroneous-tagged words enables a continual improvement of NER performance.
The optimal annotated vocabulary is composed of 211,576 words. The models obtained using this vocabulary are composed of 211,576 unigrams, 16,3047,041 bigrams, 163,047,041 trigrams and 377,272,219 quadrigrams. The system shows 17.38 % of WER, 48.56 % of SER and 65 % of F-measure.

6. Results

We evaluated our approach on the *TestPart corpus* using the optimal annotated vocabulary found.

Table 2: *Word error rates of LIUM system before and after integrating NER computed on the TestPart corpus*

	WER (%)	**NE WER (%)**
Baseline LIUM	20.23	23.12
LIUM with NER	21.17	25.98

Table 2 shows the WER before and after integrating the NER and ASR processes. We can observe that by adapting the language models and pronunciation dictionary, the overall WER increases by 0.94 %. For named entities, the WER increases by 2,86 %. Of the 6,114 words constituting the named entities, 4,525 words were correctly transcribed by the two LIUM system versions. Many of the named-entity transcription errors concern names of people (around 48 % of WER for both systems).
To evaluate the contribution of integrating NER directly into the ASR process, we have, on the one hand, decoded the test data using LIUM base system and annotated the obtained transcriptions using LIANE NER tagger. On the other hand, we have decoded the test data using LIUM system augmented with named entity information. The transcriptions obtained are directly annotated with named entities. Table 3 shows the NER performance obtained using the classical approach (LIUM then LIANE) and the proposed approach (LIUM with NER). We can see that by integrating NER and ASR processes, there is an improvement of about 5 % in terms of SER and F-measure over the baseline system (LIANE). We can also observe an improvement in NER precision with a gain of about 6 %. We notice that

Table 3: *NER results before and after integrating NER computed on the TestPart corpus (F: F-measure, P: precision, R: recall).*

	SER (%)	**F (%)**	**P (%)**	**R (%)**
LIUM then LIANE	54.01	58	64.5	52.76
LIUM with NER	**49.22**	**63**	**70.32**	**57.59**

Table 4: *NER results by category.*

	LIUM then LIANE	**LIUM with NER**
	SER (%)	**SER (%)**
Person	69.90	**64.76**
Organization	76.64	**71.56**
Location	60.19	**54.46**
Function	65.33	**50.17**
Product	112.5	**97.92**
Amount	53.21	**51.28**
Time	57.44	**55.38**

LIANE shows 54.12 % of SER and 58 % of F-measure on the output of LIUM-with-NER system after removing the named-entity tags. Table 3 indicates NER results by category. The results show an improvement for different categories.
Although the transcription quality decreases slightly, integrating NER into the speech transcription process allows to perform better recognition of named entities. We attribute this gain to the fact that assigning *a priori* tags to named entities enables the labelling of named entities to be controlled even if there are ASR errors in context words. For example, the tag "UNESCO**-organization-B**" is associated to "UNESCO" because it always appears with this tag in annotated corpora. "UNESCO" will appear as an organization (UNESCO**-organization-B**) in automatic transcripts, whatever its context.
Aside from named entity transcription errors that we do not deal with in this work, we distinguish two main reasons of annotation errors:

- Duplication of some errors committed by the NER system in the training corpora used to create ASR language models, in spite of filtering a large number of erroneous-tagged words. This concerns, in particular, the names of product and organization.
- Presence of some named entities that are infrequent or do not appear in the training corpora. The constituting words of these named entities are not annotated in the annotated vocabulary. This represents a major disadvantage of this approach, mainly when it comes to dealing with a dynamic domain that requires frequent updating of the annotation of named entities.

7. Conclusions

We have proposed a method for speech NER that integrates ASR and NER processes by pre-annotating ASR language models and pronunciation dictionary. We have shown that this adaptation does not greatly affect the ASR performance and provides improvements in NER performance. Future work will concentrate on improving ASR language models and filtering erroneous named entity tags in ASR vocabulary. We also intend to exploit the word lattice in order to reduce transcription errors concerning named entities.

8. References

[1] D. Nadeau and S. Sekine, "A survey of named entity recognition and classification," *Linguisticae Investigationes*, vol. 30, pp. 3–26, January 2007.

[2] J.-h. Kim and P. Woodland, "A Rule-Based Named Entity Recognition System for Speech Input," in *Proceedings of ICSLP '00*, Beijing, China, 2000, pp. 521–524.

[3] F. Béchet and E. Charton, "Unsupervised knowledge acquisition for Extracting Named Entities from speech," in *Proceedings of ICASSP '10*, Dallas, Texas, USA, 2010, pp. 5338–5341.

[4] D. Miller, S. Boisen, R. Schwartz, R. Stone, and R. Weischedel, "Named entity extraction from noisy input: speech and OCR," in *Proceedings of ANLC '00*, Seattle, Washington, USA, 2000, pp. 316–324.

[5] A. Gravano, M. Jansche, and M. Bacchiani, "Restoring punctuation and capitalization in transcribed speech," in *Proceedings of ICASSP '09*, Taipei, Taiwan, 2009, pp. 4741–4744.

[6] D. D. Palmer and M. Ostendorf, "Improving information extraction by modeling errors in speech recognizer output," in *Proceedings of HLT '01*, San Diego, California, USA, 2001, pp. 1–5.

[7] K. Sudoh, H. Tsukada, and H. Isozaki, "Incorporating speech recognition confidence into discriminative named entity recognition of speech data," in *Proceedings of ACL '06*, Sydney, Australia, 2006, pp. 617–624.

[8] C. Parada, M. Dredze, and F. Jelinek, "OOV Sensitive Named-Entity Recognition in Speech." in *Proceedings of INTERSPEECH '11*, Florence, Italy, 2011, pp. 2085–2088.

[9] L. Zhai, P. Fung, R. Schwartz, M. Carpuat, and D. Wu, "Using N-best lists for named entity recognition from Chinese speech," in *Proceedings of HLT-NAACL '04*, Boston, Massachusetts, USA, 2004, pp. 37–40.

[10] B. Favre, F. Béchet, and P. Nocéra, "Robust named entity extraction from large spoken archives," in *Proceedings of the HLT-EMNLP '05*, Vancouver, British Columbia, Canada, 2005, pp. 491–498.

[11] T. Hori and A. Nakamura, "An extremely large vocabulary approach to named entity extraction from speech," in *Proceedings of ICASSP '06*, Toulouse, France, 2006, pp. 973–976.

[12] P. Deléglise, Y. Estève, S. Meignier, and T. Merlin, "The LIUM speech transcription system: a CMU Sphinx III-based system for French broadcast news," in *Proceedings of INTERSPEECH '05*, Lisbon, Portugal, 2005, pp. 1653–1656.

[13] ——, "Improvements to the LIUM french ASR system based on CMU Sphinx: what helps to significantly reduce the word error rate?" in *Proceedings of INTERSPEECH '09*, Brighton, United Kingdom, 2009, pp. 2123–2126.

[14] S. Galliano, G. Gravier, and L. Chaubard, "The ester 2 evaluation campaign for the rich transcription of French radio broadcasts," in *Proceedings of INTERSPEECH '09*, Brighton, UK, 2009, pp. 2583–2586.

[15] A. Zidouni, S. Rosset, and H. Glotin, "Efficient combined approach for named entity recognition in spoken language," in *Proceedings of INTERSPEECH '10*, Makuhari, Japan, 2010, pp. 1293–1296.

[16] A. Stolcke, "Srilm - an extensible language modeling toolkit," in *Proceedings of ICSLP '02*, Denver, Colorado, USA, 2002, pp. 901–904.

[17] J. Makhoul, F. Kubala, R. Schwartz, and R. Weischedel, "Performance measures for information extraction," in *Proceedings of DARPA Broadcast News Workshop*, Herndon, Virginia, USA, 1999, pp. 249–252.