Temporal and Lexical Context of Diachronic Text Documents for Automatic Out-Of-Vocabulary Proper Name Retrieval

Irina Illina, Dominique Fohr, Georges Linares, Imane Nkairi

Temporal and Lexical Context of Diachronic Text Documents for Automatic Out-Of-Vocabulary Proper Name Retrieval

Irina Illina[1], Dominique Fohr[1], Georges Linarès[2], Imane Nkairi[1]

[1]MultiSpeech team, LORIA-INRIA, 54602 Villers-les-Nancy, France
[2]LIA – University of Avignon, 84911 Avignon, France

Abstract. Proper name recognition is a challenging task in information retrieval from large audio/video databases. Proper names are semantically rich and are usually key to understanding the information contained in a document. Our work focuses on increasing the vocabulary coverage of a speech transcription system by automatically retrieving proper names from contemporary diachronic text documents. We proposed methods that dynamically augment the automatic speech recognition system vocabulary using lexical and temporal features in diachronic documents. We also studied different metrics for proper name selection in order to limit the vocabulary augmentation and therefore the impact on the ASR performances. Recognition results show a significant reduction of the proper name error rate using an augmented vocabulary.

Keywords: speech recognition, out-of-vocabulary words, proper names, vocabulary augmentation

1 Introduction

The technologies involved in information retrieval from large audio/video databases are often based on the analysis of large, but closed corpora. The effectiveness of these approaches is now acknowledged, but they nevertheless have major flaws, particularly for those which concern new words and proper names. In our work, we are particularly interested in *Automatic Speech Recognition* (ASR) applications.

Large vocabulary ASR systems are faced with the problem of *out-of-vocabulary* (OOV) words. This is especially true in new domains where named entities are frequently unexpected. OOV words are words which are in the input speech signal but not in the ASR system vocabulary. In this case, the ASR system fails to transcribe the OOV words and replaces them with one or several in-vocabulary words, impacting the transcript intelligibility and introducing the recognition errors.

Proper Name (PN) recognition is a complex task, because PNs are constantly evolving and no vocabulary will ever contain all existing PNs: for example, PNs represent about 10% of words of English and French newspaper articles and they are more important than other words in a text to characterize its content [7]. Bechet and Yvon [2] showed that 72% of OOV words in a 265K-words lexicon are potentially PNs.

Increasing the size of ASR vocabulary is a strategy to overcome the problems of OOV words. For this purpose, the Internet is a good source of information. Bertoldi and Federico [3] proposed a methodology for dynamically extending the ASR vocabulary by selecting new words daily from contemporary news available on the Internet: the most recently used new words and the most frequently used new words were added to the vocabulary. Access to large archives, as recently proposed by some institutions, can also be used for new word selection [1].

Different strategies for new word selection and vocabulary increasing were proposed recently. Oger *et al* [13] proposed and compared local approaches: using the local context of the OOV words, they build efficient requests for submitting it to a web search engine. The retrieved documents are used to find the targeted OOV words. Bigot *et al* [4] assumed that a person name is a latent variable produced by the lexical context it appears in, i.e. the sequence of words around the person name and that a spoken name could be derived from ASR outputs even if it has not been proposed by the speech recognition system.

Our work uses context modeling to capture the lexical information surrounding PNs so as to retrieve OOV proper names and increase the ASR vocabulary size. We focus on exploiting the lexical context based on *temporal* information from *diachronic documents* (documents that evolve through time): we assume that the time is an important feature for capturing name-to-context dependencies. Compared to approaches of Bigot *et al*. [4] and Oger *et al* [13], we also use the proper name context notion. However, our approaches focus on exploiting the documents' temporality using diachronic documents. Our assumption is that PNs are often related to an event that emerges in a specific time period in diachronic documents and evolve through time. For a given date, the same PNs would occur in documents that belong to the same period. Temporal context has been proposed before by Federico and Bertoldi [3] to cope with language and topic changes, typical to new domains, and by Parada *et al*. [14] for predict OOV in recognition outputs. Compared to these works, our work extends vocabulary using shorter and more precise time periods to reduce the excessive vocabulary growth. We are seeking a good trade-off between the lexical coverage and the increase of vocabulary size that can lead to dramatically increasing the resources required by an ASR system.

Moreover, the approaches presented in [3][14] are a priori approaches that increase the lexicon before the speech recognition of test documents and are based on the dates of documents. This type of techniques has the disadvantage of a very large increase of the lexicon, which ignores the context of appearance of missing words in the documents to recognize. Our proposal uses a first pass of decoding to extract information relating to the lexical context of OOV words, which should lead to more accurate model of the context of OOV words and avoid excessive increase of vocabulary size.

This paper is organized as follows. The next section of this paper provides the proposed methodology for new PN retrieval from diachronic documents. Section 3 describes preliminary experiments and results. The discussion and conclusion are presented in the last section.

2 Methodology

Our general idea consists in using lexical and temporal context of diachronic documents to derive OOV proper names automatically from diachronic documents. We assume that missing proper names can be automatically found in contemporary documents, that is to say corresponding to the same time period as the document we want to transcribe. We hypothesize that proper names evolve through time, and that for a given date, the same proper names would occur in documents that belong to the same period. Our assumption is that the linguistic context might contain relevant OOV proper names or to allow to add some specific information about the missing proper names.

We propose to use text documents from the diachronic corpus that are contemporaneous with each *test* document. We want to build a locally augmented vocabulary. So, we have a test audio document (to be transcribed) which contains OOV words, and we have a diachronic text corpus, used to retrieve OOV proper names. An augmented vocabulary is built for each test document.

We assume that, for a certain date, if proper names co-occur in diachronic documents, it is very likely that they co-occur in the test document corresponding to the same time period. These co-occurring PNs might contain the targeted OOV words. The idea is to exploit the relationship between PNs for a better lexical enrichment.

To reduce the OOV proper name rate, we suggest building a PN vocabulary that will be added to the large vocabulary of our ASR system. In this article, different PN selection strategies will be proposed to build this proper name vocabulary:

- **Baseline method:** Selecting the diachronic documents only using a time period corresponding to the test document.
- **Local-window-based method:** same strategy as the baseline method but a co-occurrence criterion is added to exploit the relationship between proper names for a given period of time.
- **Mutual-information-based method** same strategy as the baseline method but mutual information metric is used to better choose OOV proper names.
- **Cosine-similarity-based method** same strategy as the baseline method but the documents are represented by word vector models.

In all proposed methods, documents of the diachronic corpus have been processed by removing punctuation marks and by turning texts to lower case, like in ASR outputs.

2.1 Baseline method

This method consists in extracting a list (collection) of all the OOV proper names occurring in a diachronic corpus, using a time period corresponding to the test document. Proper names are extracted from diachronic corpus using *Treetagger*, a tool for annotating text with part-of-speech and lemma information [15]. Only the new proper names (compared to standard vocabulary of our recognition system) are kept. Then, our vocabulary is augmented with the collection of extracted OOV proper names. Augmented

lexicon is built for each test file and for each time period. This period can be, for example, a day, a week or a month. The OOV PN pronunciations are generated using a phonetic dictionary or a grapheme-to-phoneme tool [10].

This method will result in recalling a large number of OOV proper names from the diachronic corpus. Therefore, we consider this method as our baseline. The problem of this approach is if the diachronic corpus is large, we can have a bad tradeoff between the lexical coverage and the increase of lexicon size. Moreover, only temporal information about document to transcribe is used. In the methods, presented in the following, the lexical context of PN will be taken into account to better select OOV proper names.

2.2 Local-window-based method

To have a better tradeoff between the lexical coverage and the increase of lexicon size, we will use a local lexical context to filter the selected PNs.

We assume that a context is a sequence of (2N+1) words centered on one proper name. Each PN can have as many contexts as occurrences.

In this method, the goal is to use the in-vocabulary proper names of the test document as an anchor to collect linked new proper names from the diachronic corpus. The OOV proper names that we need to find might be among the collected names. This method consists of several steps:

A) In-vocabulary PN extraction from each test document:

For each test document from the test corpus, we perform an automatic speech recognition with our standard vocabulary. From obtained test file transcription (that can contain some recognition errors) and we extract all PNs (in-vocabulary PNs).

B) Context extraction from diachronic documents:

After extracting the list of the in-vocabulary proper names from the test document transcription, we can start extracting their "contexts" in the diachronic set. Only documents that correspond to the same time period as the test document are considered. In this method, a context refers to a window of (*2N+1*) words centred on one proper name. We tag all diachronic documents that belong to the same time period as our test document. Words that have been tagged as proper names by *Treetagger* are kept, and all the others are replaced by "*X*". In this step, the substitution with "*X*" aims to save the absolute positions of the words composing the context. We go through all tagged contemporary documents from the diachronic corpus and we extract all contexts corresponding to all occurrences of in-vocabulary PNs of the test document: in the (*2N+1*) window centred on in-vocabulary proper name, we select all words that are not labelled as "*X*" and that are new proper names (that are not already in our vocabulary). The idea behind using a centered window is that the short-term local context may contain missing proper names.

C) Vocabulary Augmentation:

From the extracted new PNs obtained in step B, we keep only the new PNs whose number of occurrences is greater than a given threshold. Then we add them to our vocabulary. Their pronunciations are generated using a phonetic dictionary or an automatic phonetic transcription tool.

Using this methodology, we expect to extract a reduced list (compared to the baseline) of all the potentially missing PNs.

2.3 Mutual-information-based method

In order to reduce the vocabulary growth, we propose to add a metric to our methodology: the *mutual information* (MI) [5]. The MI-based method consists in computing the mutual information between the in-vocabulary PNs found in the test document and other PNs that have appeared in contemporary documents from the diachronic set. If two PNs have high mutual information, it would increase the probability that they occur together in the test document.

In probability theory and information theory, the mutual information of two random variables is a quantity that measures the mutual dependence of the two random variables. Formally, the mutual information of two discrete random variables *X* and *Y* can be defined as:

$$I(X;Y) = \sum\sum p(x,y)\log(\frac{p(x,y)}{p(x)p(y)}) \qquad (1)$$

In our case, *X* and *Y* represent proper names and *x=1* if it is present in the document and *x=0* otherwise. For example:

$$P(x = 1, y = 1) = \frac{number\ of\ documents\ containing\ x\ and\ y}{total\ number\ of\ documents} \qquad (2)$$

The higher the probability of the co-occurrence of two proper names in the diachronic corpus, the higher the probability of their co-occurrence in a test document.

Finally we compute the mutual information between all the combinations of the variable *X* (*X* is the in-vocabulary PN extracted from the test document) and the variable *Y* (*Y* is the OOV proper name extracted from the contemporary documents from the diachronic corpus).

Compared to local-window-based method, only the step B is modified.

B) Context extraction from diachronic documents:

After extracting the list of the in-vocabulary proper names from the test document transcription, we can start extracting their "contexts" in the diachronic set. Only documents that correspond to the same time period as the test document are considered. The list of in-vocabulary words from test document transcriptions is extracted like in local-window-based method. As previously, we tag all diachronic documents that belong to the same time period as our test document. Words that have been tagged as proper names by *Treetagger* are kept. Finally, the mutual information between each word from

in-vocabulary PN list and each extracted new PNs from diachronic document is calculated. If two PNs have high mutual information, this increases the probability that they both appear in the document to transcribe.

Using this methodology, we expect to extract a shortlist (compared to the reference method) potentially missing PNs.

2.4 Cosine-similarity-based method

In this method we want to consider additional lexical information to model the context: we will use not only proper names (as in the previous methods) but also verbs, adjectives and nouns. These words are extracted using *Treetagger*. They are lemmatized because we are interested in the semantic information. The other words are removed.

In this method we propose to use the term vector model [16]. This model is an algebraic representation of the content of a text document in which the documents are generally represented by the word vectors. The proximity between the documents is often calculated using the cosine similarity. So we will represent diachronic documents and documents to transcribe as a word vectors and use the cosine similarity between vector models of these documents to extract relevant PNs.

We use the same steps as in the previous methods, with the following modifications:

A) In-vocabulary PN extraction from each *test* document:

As in the previous methods, each test document is transcribed using the speech recognition system and the standard vocabulary. Then, each document to be transcribed is shown by the histogram of occurrences of component words: vector of words (*bag of words*, BOW). As stated above, only the verbs, adjectives, proper names and common names are lemmatized and considered.

B) Context extraction from diachronic documents:

Each selected (according to the time period) diachronic document is also represented by the BOW in the same way as above. Then, we build the list of new PNs by choosing from selected diachronic documents the words that have been labeled as "proper name" and which are not in the standard vocabulary. This list is built in the same manner as in the previous methods. For each PN of this list, we calculate a *PNvector*. For this, first of all, a common lexicon is built: it contains the list of words (verbs, adjectives, nouns and proper names) that appear at least once in the selected diachronic document or in the document to transcribe. Then, every BOW are projected on the common lexicon. Finally, the *PNvector* is calculated as the sum of the BOW of the selected diachronic documents in which these new PNs appear.

For each PN from this list, we calculate the cosine similarity between the BOW and its *PNvector* of the document to transcribe. The new OOV PNs whose cosine similarity is greater than a threshold are selected.

C) Vocabulary Augmentation:
This step is the same as step C of the local-window-based method.

Compared with previous methods, cosine method takes into account broader contextual information using not only proper names but also verbs, adjectives and nouns present in the selected diachronic documents and in the test document.

3 Experiments

3.1 Test corpus

To validate the proposed methodology, we used as test corpus five audio documents extracted from the ESTER2 corpora [8] (see Table 1). The objective of the ESTER2 campaign was to assess the automatic transcription of broadcast news in French. The campaign targeted a wide variety of programs: news, debates, interviews, etc.

Doc1	Doc2	Doc3	Doc4	Doc5
2007/12/20	2007/12/21	2008/01/17	2008/01/18	2008/01/24

Table 1. Date of test documents.

Table 2 presents the occurrences of all PNs (in-vocabulary and out-of-vocabulary) in each test document with respect to our 97k ASR vocabulary. To artificially increase OOV rate, we have randomly removed 75 proper names occurring in the test set from our 97k ASR vocabulary. We call this vocabulary a *standard vocabulary.* Finally, the OOV proper name rate is about 1% (404/38525).

	Number of diff. words	Number of occur.	In-vocab PNs	OOV PNs	OOV PN occur.
Doc1	1350	4099	86	44	93
Doc2	1446	4604	89	39	70
Doc3	1958	11803	43	25	63
Doc4	2107	10152	90	39	71
Doc5	1432	7867	48	27	107
All	-	38525	-	-	404

Table 2. Proper name coverage in test documents.

In this preliminary experiment (Table 2), in-vocabulary PN extraction is performed from manual transcription of test documents instead of automatic speech transcription. The goal of this preliminary study is to validate our proposed approaches.

3.2 Diachronic corpus

As diachronic corpus, we have used the *Gigaword* corpora: *Agence France Presse* (AFP) and *Associated Press Worldstream* (APW). French Gigaword is an archive of newswire text data and the timespans of collections covered for each are as follows: for AFP May 1994 - Dec 2008, for APW Nov 1994 - Dec 2008. The choice of Gigaword and ESTER corpora was driven by the fact that one is contemporary to the other, their temporal granularity is the day and they have the same textual genre (journalistic) and domain (politics, sports, etc.).

Using *Treetagger*, we have extracted 45981 OOV PNs from 6 months of the diachronic corpus. From these OOV PNs, only 103 are present in the test corpus, which corresponds to 71% of recall. It shows that it is necessary to filter this list of PNs to have a better tradeoff between the PN lexical coverage and the increase of lexicon size.

3.3 Transcription system

ANTS (Automatic News Transcription System) [9] used for these experiments is based on Context Dependent HMM phone models trained on 200-hour broadcast news audio files. The recognition engine is Julius [12]. Using SRILM toolkit [17], the language model is estimated on text corpora of about 1800 million words. The corpus of texts comes from newspaper articles (*Le Monde*), broadcast transcriptions and data collected on the Internet. The language model is re-estimated for each augmented vocabulary.

The baseline phonetic lexicon contains 218k pronunciations for the 97k words.

4 Experimental results

4.1 Baseline results

We call ***selected PNs*** the new proper names that we were able to retrieve from diachronic documents using our methods. We call ***retrieved OOV PNs*** the OOV PNs that belong to the *selected PN* list and that are present in the test documents.

We build a specific augmented vocabulary for each test document, each chosen period and each method. The augmented vocabulary contains all words of standard vocabulary and the *selected PNs* given by the chosen method and corresponding to the chosen period. So, we need to estimate the language model (n-gram probabilities) for these retrieved OOV PNs. For this we have chosen to completely re-estimate the language model for each augmented vocabulary using the entire text corpus (see section 3.3). The best way to incorporate the new PNs in the language model is beyond the scope of this paper.

Our results are presented in terms of *Recall* (%): number of retrieved OOV PN versus the number of OOV PNs contained in the document to transcribe. We place ourselves in the context of speech recognition. In this context, the fact that PN present in the document to recognize is not in the vocabulary of the recognition system will produce a significant error because the PN cannot be recognized. However, adding to the vocabulary of the recognition system a PN that is not present (pronounced) in the test

file, will have little influence on the recognized sentence (if we add too many words, there may increase the confusion between words and thus cause errors). So, in our case, the recall is more important than precision. Thus, we will present the results in term of recall.

For the recognition experiments, *Word Error Rate* (WER) is given.

In order to investigate whether time is a significant feature, we studied 3 time intervals in the diachronic documents:

- 1 day: using the same day as the test document;
- 1 week: using 3 days before until 3 days after the test document date;
- 1 month: using the current month of the test document.

Time period	Selected PNs	Retrieved OOV PNs	Recall (%)
1 day	925	16	44.0
1 week	4305	21	58.6
1 month	13069	24	**67.6**

Table 3. Coverage in test documents of retrieved OOV PNs. Average over all test files.

As we build an augmented vocabulary for each test file, the results presented in Table 3 are averaged over all test files. Table 3 shows that the use of diachronic documents whose date is closest to that of the test document (document to transcribe) allows greatly reduce the number of added new proper names while maintaining an attractive recall. For example, limiting the time interval to 1 month reduces the set of PN candidates to 13069 (Table 3) while still retrieving 67.6% of the missing OOVs, compared to 45 981 candidates (6 months) for almost the same recall (67.6%, cf. section 3.2). Moving from a one month time period to one day, we reduced the number of selected PNs by a factor of 14 (13069/925) while the recall is reduced by a factor of 1.5. This result confirms the idea that the use of the temporal information reduces the list of selected new PN for augmented vocabulary while maintaining a good recall. In the rest of this article, we will study three time periods (one day, one week and one month).

4.2 Local-window-based results

Table 4 presents the results for the local-window-based method on the test corpus for the three studied time periods and for different window sizes. The threshold *occ* is used to keep only the selected PNs whose number of occurrences are greater than occ. Compared to the baseline method, the local-window-based method reduces significantly the number of selected proper names and only slightly decreases the recall. For example, for a period of one day and window size of 100, the number of selected PNs is divided by 3.6 compared to baseline method, while the recall drops to 6% (253.4 versus 925 and 37.9% versus 44.0%). For the period of one month, the filter allows to divide the number of selected PNs by 11, losing only 9% of recall compared to the baseline method. This shows the effectiveness of the proposed local-window-based method.

Time period	Window size	Selected PNs	Retrieved OOV PNs	Recall (%)
1 day (occ>0)	50	164.6	11.8	33.9
	100	253.4	13.2	37.9
1 week (occ>1)	50	344.0	16.4	47.1
	100	596.4	17.4	50.0
1 month (occ>2)	50	589.8	19.0	54.6
	100	1137.2	20.4	**58.6**

Table 4. Local-window-based results according to window size and time period.

We notice that we were not able to recall 68% of the missing PNs as we did using the baseline. 58.6% of recall is the maximal value that we obtain using this methodology.

4.3 Mutual-information-based results

Table 5 shows the results for the method based on mutual information using different time periods and thresholds. Two PNs having a mutual information greater than this threshold will be added to selected PN list. The best recall is obtained using a time period of one week.

Time period	Threshold	Selected PNs	Retrieved OOV PNs	Recall (%)
1 day (occ>0)	0.05	10.6	5.0	14.4
	0.01	295.0	12.8	36.8
	0.005	421.2	14.2	40.8
	0.001	531.2	15.4	44.25
1 week (occ>1)	0.05	3.8	3.0	8.6
	0.01	50.8	8.8	25.3
	0.005	228.6	12.0	34.5
	0.001	947.8	18.4	**52.9**
1 month (occ>2)	0.05	2.6	1.6	4.6
	0.01	21.2	7.2	20.7
	0.005	56.4	9.4	27.0
	0.001	806.4	17.2	49.4

Table 5. Mutual-information-based results according to threshold and time period.

As for local-window-based method, the use of the diachronic documents from one day period is sufficient to obtain a recall of over 30%. For the recognition experiments (see section 4.4) we will set the threshold to 0.001 for all time periods.

4.4 Cosine-similarity-based results

The results for the method based on cosine similarity are shown in Table 6. In order to further reduce the number of selected PNs, we keep only the PNs whose number of occurrences is greater than a threshold depending on the time period.

As for the mutual-information-based method, considering only one day time period to retrieve new PNs seems unsatisfactory. The best compromise between the recall and the number of selected PNs is obtained for the period of one month and a threshold of 0.05 (59.8% of recall).

Time period	Threshold	Selected PNs	Retrieved OOV PNs	Recall (%)
1 day (occ>0)	0.025	813.4	15.4	44.3
	0.05	437.6	14.4	41.4
	0.075	131.4	11.2	32.2
	0.1	51.8	8.4	24.1
1 week (occ>1)	0.025	1880.0	19.4	55.8
	0.05	1127.6	18.8	54.0
	0.075	431.6	17.0	48.9
	0.1	152.0	13.4	38.5
1 month (occ>2)	0.025	3795.6	21.4	**61.5**
	0.05	2473.8	20.8	59.8
	0.075	1010.2	19.4	55.8
	0.1	334.4	17.0	48.9

Table 6. Cosine-similarity-based results according to threshold and time period.

4.5 Automatic speech recognition results

For validating the proposed approaches, we performed the automatic transcription of the 5 test documents using an augmented lexicon generated by the three proposed methods.

We generate an augmented vocabulary for each test file, for each time period and for each PN selection method. To generate the pronunciations of added PNs, we use an automatic approach based on CRF (*Conditional Random Fields*). We chose this approach because it has shown very good results compared to the best approaches of the state-of-art [10]. The CRF [11] is a probabilistic model for labeling or segmentation of structured data such as sequences, trees or trellises. The CRF allows to take into account long-term relations, achieve a discriminating training and converge to a global optimum. Using this approach, we obtained a precision and recall of more than 98% for phonetization of common names in French (*BDLex*) [10]. In the context of this article, we have trained our CRF with 12,000 phonetized proper names.

Table 7 presents the recognition results for five test documents using the local- window-based method for two time periods (one week and one month) and 100 for the

window size. Compared to our standard lexicon, the augmented lexicon gives a significant decrease of the word error rate for a week period (confidence interval ± 0.4%). We recall that the OOV rate of OOV PNs in the test corpus is about 1% (cf. section 3.1). So, the expected improvement will hardly exceed that rate.

	Standard lexicon	**Augmented lexicon (local-window size 100)**	
		1 week	**1 month**
Doc 1	19.7	17.7	17.6
Doc 2	20.9	19.9	20.1
Doc 3	28.3	28.2	28.7
Doc 4	24.5	23.9	24.5
Doc 5	36.5	36.0	36.6
All	27.1	**26.4**	26.9

Table 7. Word Error Rate (%) for local-window-based method according to time period (local window size 100).

The results for the mutual information method (threshold 0.001) are presented in Table 8. On average, the augmented lexicon reduces slightly the Word Error Rate.

The results for the cosine-similarity-based method (threshold 0.025 for a week and 0.05 for a month) are presented in Table 9. For both time periods, the WER is significantly reduced.

	Standard lexicon	**Augmented lexicon (MI threshold 0.001)**	
		1 week	**1 month**
Doc 1	19.7	18.0	18.4
Doc 2	20.9	19.7	20.0
Doc 3	28.3	28.1	28.2
Doc 4	24.5	24.2	24.2
Doc 5	36.5	36.1	36.0
All	27.1	**26.6**	26.7

Table 8. Word Error Rate (%) for mutual-information-based method according to time period (threshold 0.001)

For the three methods, performance depends on the test document. For documents 1 and 2, regardless the time period used to create the augmented lexicon, the word error rate improvement is significant. But for document 3, no improvement or degradation is observed. This can be due to the fact that the OOV proper names in the test document are not observed in the corresponding diachronic documents. The detailed results show that the performance in terms of WER depends on the type of documents: for some broadcast programs we do not observe any improvement (for example, a debate on nuclear), for others a strong improvement is reached (news).

Finally, the three proposed methods give about the same performance.

	Standard lexicon	Augmented lexicon	
		1 week thr 0.025 occ>1	1 month thr 0.05 occ>2
Doc 1	19.7	18.0	18.2
Doc 2	20.9	19.5	19.4
Doc 3	28.3	27.9	28.0
Doc 4	24.5	23.8	23.7
Doc 5	36.5	35.8	35.8
All	27.1	**26.4**	**26.4**

Table 9. Word Error Rate (%) for cosine-similarity-based method according to time period.

If we consider only the recognition of proper names, using the standard lexicon, the *Proper Name Error Rate* (PNER) is 47.7%. However, using augmented lexicon obtained by cos method (one month time period), PNER dropped to 37.4%. Therefore, we observe a huge decrease of PNER, which shows the effectiveness of our proposed methods.

5 Conclusion and discussion

This article has focused on the problem of out-of-vocabulary proper name retrieval for vocabulary extension using diachronic text documents (which change over time). This work is performed in the framework of automatic speech recognition. We investigated methods that augment the vocabulary with new proper names. We propose to use the lexical and temporal features. The idea is to use in-vocabulary proper names as an anchor to collect new linked proper names from the diachronic corpus. Our context model is based on the co-occurrences, mutual information and cosine similarity.

Experiments have been conducted on broadcast news audio documents (ESTER2 corpus) using AFP and AWP text data as a diachronic corpus. The results validate the hypothesis that exploiting time and the lexical context could help to retrieve the missing proper names without excessive growth of the vocabulary size. The recognition results show a significant reduction of the word error rate using the augmented vocabulary and a huge reduction of the proper name error rate.

An interesting perspective could be to exploit “semantic” information contained in the test document: when a precise date is recognized in a test document, the diachronic document around this date could be used to bring new proper names. Our future work will also focus on investigating the use of several Internet sources (Wiki, texts, videos, etc.).

Acknowledgements

The authors would like to thank the ANR *ContNomina* SIMI-2 of the French National Research Agency (ANR) for funding.

References

1. Allauzen, A., Gauvain, J.-L.: Diachronic vocabulary adaptation for broadcast news transcription. In Proc. of Interspeech (2005)
2. Bechet, F., Yvon, F.: Les Noms Propres en Traitement Automatique de la Parole. In Revue Traitement Automatique des Langues, vol. 41, num. 3, pp. 672-708 (2000)
3. Bertoldi, N., Federico, M.: Lexicon adaptation for broadcast news transcription. In Adaptation-2001, pp. 187- 190 (2001)
4. Bigot, B., Senay, G., Linares, G., Fredouille, C., Dufour, R.: Person name recognition in ASR outputs using continous context models. In Proc. of ICASSP (2013)
5. Church, K., Hanks, P.: Word association norms, mutual information, and lexicography. In Proc. of the 27th Annual Meeting of the Association for Computational Linguistics (1989)
6. Federico, M., Bertoldi, N.: Broadcast news LM adaptation using contemporary texts. In Proc. of Interspeech, pp. 239-242 (2001)
7. Friburger, N., Maurel, D.: Textual Similarity Based on Proper Names. In Proc. of the workshop Mathematical/Formal Methods in Information Retrieval, pp. 155-167 (2002)
8. Galliano, S., Geoffrois, E., Mostefa, D., Choukri, K., Bonastre, J.-F., Gravier, G.: The ESTER Phase II Evaluation Campaign for the Rich Transcription of French Broadcast News. In Proc. of Interspeech (2005)
9. Illina I., Fohr, D., Mella, O., Cerisara, C.: The Automatic News Transcription System: ANTS, some Real Time Experiments. In Proc. ICSLP (2004)
10. Illina I., Fohr D., Jouvet, D.: Grapheme-to-Phoneme Conversion using Conditional Random Fields. In Proc. of Interspeech (2011)
11. Lafferty, J., McCallum, A., Pereira, F.: Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In Proc. of International Conference on Machine Learning, p. 282-289 (2001)
12. Lee, A., Kawahara, T.: Recent Development of Open-Source Speech Recognition Engine Julius. In Asia-Pacific Signal and Information Processing Association Annual Summit and Conference APSIPA ASC (2009)
13. Oger, S., Linarès, G., Béchet, F.: Local methods for on-demand out-of-vocabulary word retrieval. In Proc. of the Language Resources and Evaluation Conference LREC (2008)
14. Parada, C., Dredze, M., Filimonov, F., Jelinek, F.: Contextual Information Improves OOV Detection in Speech.In Proc. of NAACL (2010)
15. Schmid, H.: Probabilistic part-of-speech tagging using decision trees. In Proc. of ICNMLP (1994)
16. Singhal, A.: Modern Information Retrieval: A Brief Overview. In Bulletin of the IEEE Computer Society Technical Committee on Data Engineering 24 (4): p. 35-43 (2001)
17. Stolcke, A.: SRILM - An Extensible Language Modeling Toolkit. In Proc. of ICSLP (2002)