Proper Name Retrieval from Diachronic Documents for Automatic Speech Transcription using Lexical and Temporal Context

Irina Illina, Dominique Fohr, Georges Linarès

Proper Name Retrieval from Diachronic Documents for Automatic Speech Transcription using Lexical and Temporal Context

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

[1]Speech Group, LORIA-INRIA, 54602 Villers-les-Nancy, France

[2]LIA – University of Avignon, 84911 Avignon, France

Abstract

Proper names 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 new proper names from contemporary diachronic text documents. The idea is to use in-vocabulary proper names as an anchor to collect new linked proper names from the diachronic corpus. Our assumption is that time is an important feature for capturing name-to-context dependencies, that was confirmed by temporal mismatch experiments. We studied a method based on Mutual Information and proposed a new method based on cosine-similarity measure that dynamically augment the automatic speech recognition system vocabulary. Recognition results show a significant reduction of the word error rate using augmented vocabulary for broadcast news transcription.

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

1. Introduction

Even with a large vocabulary, Automatic Speech Recognition (ASR) systems are faced with the problem of out-of-vocabulary (OOV) words, especially in new domains. OOV words are words which are in the input speech signal but not in the ASR system vocabulary. 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 [6]. Bechet and Yvon [2] showed that 72% of OOV words in a 265K-word lexicon are potentially PNs.

Our work uses temporal 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) [1]. Our assumption is that time is an important feature for capturing name-to-context dependencies [10]. Our approach was inspired by [4] and [13]: we also use the proper name context notion. However, our approaches focus on exploiting the documents' temporality using diachronic documents. We assume that PNs are often related to an event that emerges in a specific time period in diachronic documents. We hypothesize that PNs evolve through time, and that for a given date, the same PNs would occur in documents that belong to the same period. Temporal contexts have been proposed before by Federico and Bertoldi [3] to cope with language and topic changes, typical to new domains, and by [14] for OOV prediction in recognition outputs. In contrast to these works, our work extends vocabulary using shorter time periods to reduce the excessive vocabulary growth.

This paper is organized as follows. The next section of this paper provides the proposed methodology. Section 3 describes experiments and results. The discussion and conclusion are presented in the last section.

2. Methodology

Our idea consists in deriving OOV PNs automatically from diachronic text documents, using their lexical and temporal context. Our OOV proper name retrieval methods are based on the idea 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 the same proper names would occur in documents that belong to the same period.

We propose to use text documents from the diachronic corpus that are contemporaneous with each test document from the test corpus, 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 dynamically built for each test document to avoid an excessive increase of vocabulary size.

We assume that, for a certain date, a proper name from the test corpus will co-occur with other PNs in diachronic documents 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.

In this article, different PN selection strategies will be proposed to build this augmented PN vocabulary:

- Baseline method: Selecting the diachronic text documents only using a time period corresponding to the test document and select all new PNs from these diachronic documents.
- Mutual-information-based (MI) method: same strategy as the baseline method but mutual information metric is used to better choose new proper names.
- Cosine-similarity-based method: same strategy as the baseline method but the test and diachronic documents are represented by vector space model and cosine similarity is calculated between two vectors.

In a previous study [12], we presented results for local-window-based and MI-based methods. Here, a new method, cosine-similarity based method, is proposed. The parameter estimation is done on a development corpus and the final evaluation is performed on a test corpus.

2.1. Baseline method

This method consists in extracting a list (collection) of all the new proper names occurring in a diachronic corpus, using a time period corresponding to the test document. This period can be, for example, a day, a week or a month. Then, our vocabulary is augmented with the collection of extracted OOV PNs. This method will result in recalling a large number of OOV PNs. We consider this method as our baseline. The problem of this approach is that if the diachronic corpus is

large, we can have a bad tradeoff between the lexical coverage and the increase of the lexicon size.

2.2. Mutual-information-based method

To have a better tradeoff between the lexical coverage and the increase of lexicon size, we will filter the selected PNs using Mutual-information (MI). This method consists in 3 steps:

A) In-vocabulary PN extraction from each test document: For each test document, we extract PNs. The goal is to use those in-vocabulary PNs as an anchor to collect linked new proper names from the diachronic corpus.

B) Temporal context extraction from diachronic documents: After extracting the list of the in-vocabulary proper names from the test document, 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. We tag all diachronic documents that belong to the same time period as our test document. Words that have been tagged as proper names are kept, and all the others are discarded.

In order to reduce the vocabulary growth, we propose to add a metric to our methodology, the mutual information: computing the MI 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 is defined as:

$$I(X;Y) = \sum_{} \sum_{} p(x,y)\, log\left(\frac{p(x,y)}{p(x)p(y)}\right)$$

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

We compute the MI 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 of the diachronic corpus).

C) Vocabulary Augmentation: From the extracted contexts of each PN of the test document, we collect the new PNs (that are not already in our vocabulary) and then we add them to our vocabulary. In order to reduce the number of added PNs, we kept only the most frequent PNs. PN 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. Cosine-similarity-based method

Compared to MI-based method, only step B is modified. We represent each document of the diachronic corpus as a word vector (*Bag of Words*, *BOW*). Only meaningful words are kept: verbs, adjectives, nouns and PNs. For each PN, a word vector is computed as the sum of all vector documents in which this PN occurred. The test document is represented by its BOW vector. Then cosine similarity between the test BOW vector and all the PN vectors is computed. The PNs which cosine similarity is greater than a threshold are selected and added to the vocabulary.

3. Experiments

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

We extract in-vocabulary PN words from the automatic transcription generated by our transcription system ANTS. For that, the lexicon used for recognition was tagged in terms of PNs and not-PNs. In-vocabulary PNs will be used as an anchor to collect linked new proper names from the diachronic corpus. Using the diachronic documents, we build a specific augmented lexicon for each document according to the chosen period.

Results are presented in terms of Recall (%): number of retrieved OOV PNs versus the number of OOV PNs. For the recognition experiments, *Word Error Rate* (WER) and of *PN Error Rate* (PNER) are given. PNER is calculated like WER but taking into account only proper names.

3.1. Development and test corpora

To validate the proposed methodology, we used as development corpus, seven audio documents of development part of ESTER2[1] (between 2007/07/07 and 2007/07/23). For the test corpus, we used 13 audio documents from RFI (*Radio France International*) and *France-Inter* (test part of ESTER2) (between 2007/12/18 and 2008/01/28) [7].

Table 1 presents the average of occurrences of all PNs (in-vocabulary and OOV) in development and test documents with respect to 122k-word ASR vocabulary. To artificially increase OOV rate, we have randomly removed 223 PNs occurring in the development and test set from our 122k ASR vocabulary. Finally, the OOV PN rate is about 1.2%.

File	Word occ	In-vocab PNs	In-vocab PN occ	OOV PNs	OOV PN occ
Dev	4525.9	99.1	164.0	30.7	57.3
Test	4024.7	89.6	179.7	26	46.6

Table 1. *Average proper name coverage for **development** and **test** corpora per file*

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.).

[1] The aim of the ESTER2 evaluation campaign (2007 to 2009) was to evaluate automatic radio broadcasts rich transcription systems for the French language. The campaign targets a wide variety of speaking styles and accents, broadcast show and news, entertainment shows and debates.

3.3. Transcription system

ANTS (*Automatic News Transcription System*) [8] 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 [11]. The baseline phonetic lexicon contains 260k pronunciations for the 122k words. Using SRILM toolkit [16], the language model is estimated on text corpora of about 1800 million words. The language model is re-estimated for each augmented vocabulary using the whole text corpora. The best way to incorporate the new PNs in the language model is not the scope of this paper.

4. Experimental results

In a first step, we will use the development corpus to set the parameters of the proposed methods. In a second step, we will evaluate the proposed approaches on the test set.

4.1. Baseline results

Using Treetagger [15], we have extracted 160k PNs from 1 year of the diachronic corpus. From these 160k PNs, 119k are not in our lexicon. From these 119k, only 151 PN are present in the development corpus (193 in the test corpus). 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.

Time period	Average of selected PNs per dev file	Average of retrieved OOV PNs per dev file	Recall (%)
1 day	532.9	9.9	32.1
1 week	2928.4	11.3	36.7
1 month	13131.0	17.4	56.7
1 year	118797.0	24.0	**78.1**

Table 2. *Baseline results for **development** corpus according to time periods*

Table 2 shows that using the diachronic documents of 1 year, in average we retrieve 118797.0 PNs per test file. Among these PNs, we retrieve in average 24.0 OOV PNs per development file (compared to 30.7 of Table 1). This represents the recall of 78.1%.

In this experiment we found that limiting the time interval reduces dramatically the set of PN candidates (Table 2) while still retrieving more than 32.1% of the missing OOV PNs. This result supports the idea that using temporal information may help to reduce the list of new PN candidates for the vocabulary enrichment.

In order to investigate whether time is a significant feature, we studied 3 time intervals in the diachronic documents using the same day as the test document (*1 day*); using 3 days before until 3 days after the test document date (*1 week*); using the current month of test document (*1 month*). One year duration seems to be not interesting in the framework of broadcast news.

4.2. Mutual-information-based method results

Table 3 shows the results for the method based on MI using different time periods and thresholds (MI value) for development corpus. As in our methodology we build an augmented lexicon for each file, the results presented in the Table 3 are given by averaging the values computed on the 7 development files.

In order to select more relevant PNs, we introduced a frequency threshold (occ in Table 3) of each PN in the diachronic corpus of the corresponding period.

With the 0.05 threshold, very few PNs are retrieved. For instance, for one week period, among 13.1 selected PNs, 2.4 are OOV PNs.

The best recall is obtained using a time period of one month and the threshold of 0.001 (47.0%). Compared to Table 2, MI method allows to reduce the number of selected PNs by almost three, keeping about the same recall.

We set the threshold to 0.001 for all periods for recognition results (cf. section 4.4, 4.5).

To confirm our hypothesis that to use the documents from diachronic corpus corresponding to same time period that development files is important, we set up a temporal mismatch experiment: to select new PN candidates, we use the diachronic documents (for one day, for one week and for one month) 10 months after period of development documents (cf. Table 4, called "mism"). We set the threshold to 0.001.

We observe that the number of retrieved OOV PNs is much lower compared to those of Table 3. For instance, for one day period, we retrieve only 3.0 OOV PNs using temporal mismatch documents compared to 10.0 retrieved OOV PNs using the diachronic documents of the same time period. Using only one day data, already a good recall is obtained.

Time period	Threshold	Selected PNs	Retrieved OOV PNs	Recall (%)
1 day (occ>0)	0.01	244.7	9.7	31.6
	0.005	335.6	9.9	32.1
	0.001	438.4	10.0	32.6
1 week (occ>1)	0.01	261.4	8.1	26.5
	0.005	581.4	9.9	32.1
	0.001	1196.9	10.6	34.4
1 month (occ>2)	0.01	98.3	5.3	17.2
	0.005	170.4	7.4	24.2
	0.001	1843.9	**14.4**	**47.0**

Table 3. *Mutual-information-based results according to threshold and time duration period for **development** corpus.*

Time period	Method	Selected PNs	Retrieved OOV PNs	Recall (%)
1 day (occ>0)	MI	438.4	10.0	32.6
	MI mism	738.9	3.0	9.8
1 week (occ>1)	MI	1196.9	10.6	34.4
	MI mism	1623.4	6.1	20.0
1month (occ>2)	MI	1843.9	**14.4**	**47.0**
	MI mism	1707.0	7.9	25.6

Table 4. *Mutual-information-based results according to time duration period for **development** corpus, with and without **temporal mismatch,** threshold 0.001.*

4.3. Cosine-similarity-based results

The results for the method based on cosine similarity are presented in Table 5. In order to further reduce the number of selected PNs, we keep only selected PNs occurring more than a threshold in the diachronic corpus.

Compared to MI results (cf. Table 3), cosine-similarity method gives about the same recall for each time period. For the same recall and the month time period, MI method generates half as many selected PNs than cosine method.

Table 6 show the temporal mismatch experiment results for cosine-similarity method. We observe that, like for MI method, to use the same date for diachronic documents and for development corpus is important.
We set a threshold of 0.025 for the one-day and one-week period and for a threshold of 0.05 for the one-month period for the recognition experiments (cf section 4.4, 4.5).

Time period	Threshold	Selected PNs	Retrieved OOV PNs	Recall (%)
1 day (occ>0)	0.025	358.7	9.9	32.1
	0.05	208.1	9.7	31.6
	0.075	107.9	8.6	27.9
1 week (occ>1)	0.025	1188.3	10.0	32.6
	0.05	746.0	9.7	31.6
	0.075	351.3	8.6	29.7
1 month (occ>2)	0.025	3939.7	**14.7**	**47.9**
	0.05	2781.7	13.1	42.8
	0.075	1289.9	11.1	36.3

Table 5. *Cosine-similarity-based results according to threshold and time duration period for **development** corpus.*

Time period	Method	Selected PNs	Retrieved OOV PNs	Recall (%)
1 day (occ>0)	Cos	358.7	9.9	32.1
	Cos mism	1017.3	3.3	10.7
1 week (occ>1)	Cos	1188.3	10.0	32.6
	Cos mism	2457.4	6.9	22.3
1month (occ>2)	Cos	2781.7	**13.1**	**42.8**
	Cos mism	3242.1	9.1	29.8

Table 6. *Cosine-similarity-based results according to time duration period **for development** corpus, with and without **temporal mismatch***

4.4. Automatic speech recognition results for development corpus

We performed automatic transcription of the 7 development documents using augmented lexicons for the proposed methods. We generate one lexicon per development file. For generating the pronunciations of the added PNs, we use G2P CRF approach [9]. This CRF was trained on phonetic lexicon containing about 12000 PNs.

Stand. lexicon	Method	Augmented lexicon		
		1 day	1 week	1 month
30.2	**WER MI**	29.9	29.9	29.8
	WER MI mism	30.3	30.0	30.1
	WER Cos	29.8	30.0	**29.7**
	WER Cos Mism	30.1	30.0	29.9
40.7	**PNER MI**	36.8	37.0	35.6
	PNER MI mism	40.3	38.3	37.8
	PNER Cos	36.8	36.9	**35.4**
	PNER Cos mism	40.3	38.1	36.9

Table 7. *WER (%) and PNER (%) for MI and cosine-based methods according to time duration period for **development** corpus, with and without **temporal mismatch***

The results for MI-based and cosine-similarity-based methods are given in Table 7. On average, the augmented lexicon using diachronic data slightly reduces the WER (confidence interval ±0.4%). Detailed results show that the WER performance depends of the broadcast type: for some broadcast shows we have no WER improvement.
In term of PNER, substantial improvement is observed for all duration periods compared to standard lexicon. Best results are obtained for month period (35.4% PNER compared to 40.7% for cosine method), but the number of added PNs is also greatest.
Results of temporal mismatch experiments confirm our hypothesis that temporal correspondence between development and diachronic documents dates is important.

4.5. Automatic speech recognition results for test corpus

For validating the proposed approaches, we performed automatic transcription of the 13 test documents using augmented lexicons for the proposed methods (cf.Table 8). We use the parameters that we have chosen for the development corpus.
In order to incorporate the new PNs in the language model, we re-estimated it for each augmented vocabulary using the whole text corpora (1.8 Giga words).
Compared to standard lexicon, both methods give a significant improvement of WER (confidence interval ±0.4%). But there is no significant difference between the two methods. Results obtained for Mismatch experiments are, as expected, much less good.
In term of PNER, the relative improvement of about 15% is achieved for both methods.

Stand. lexicon	Method	Augmented lexicon		
		1 day	1 week	1 month
31.8	**WER MI**	31.3	**31.2**	**31.2**
	WER MI mism	31.9	31.6	31.5
	WER Cos	31.3	**31.2**	**31.2**
	WER Cos mism	31.8	31.7	31.6
44.0	**PNER MI**	38.0	37.6	37.6
	PNER MI mism	44.1	41.8	41.0
	PNER Cos	38.1	37.6	**37.4**
	PNER Cos mism	43.8	41.3	40.2

Table 8. *WER (%) and PNER (%) for MI-based and cosine-based methods according to time period for **test** corpus, with and without **temporal mismatch***

5. Conclusion and discussion

In the framework of automatic speech recognition, this work has focused on the problem of OOV PN retrieval for vocabulary extension using diachronic text documents. We investigated methods that augment the vocabulary with new PNs, using 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 mutual information and vector space model.

Experiments on broadcast news audio documents using newswire text data as a diachronic corpus validate the hypothesis that exploiting time and the lexical context could help retrieve the missing proper names without excessive growth of vocabulary size. Recognition results show a significant reduction of the word and proper names error rate using the augmented vocabularies. Temporal mismatch results confirm our hypothesis about the importance of temporal correspondence between test and diachronic documents.

6. Acknowledgements

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

7. References

[1] Allauzen, A. and Gauvain, J.-L. "Diachronic vocabulary adaptation for broadcast news transcription", Proc. of Interspeech, 2005.

[2] Bechet, F. and Yvon, F. "Les Noms Propres en Traitement Automatique de la Parole », Revue Traitement Automatique des Langues, vol. 41, num. 3 pp. 672-708, 2000.

[3] Bertoldi, N. and Federico, M. "Lexicon adaptation for broadcast news transcription", In Adaptation-2001, pp. 187- 190, 2001

[4] Bigot, B., Senay, G., Linares, G., Fredouille, C., and Dufour, R. "Person name recognition in ASR outputs using continous context models", Proc. of ICASSP, 2013.

[5] Federico, M. and Bertoldi, N. "Broadcast news LM adaptation using contemporary texts", Proc. of Interspeech, pp. 239-242, 2001

[6] Friburger, N. and Maurel, D. "Textual Similarity Based on Proper Names", Proc. of the workshop Mathematical/Formal Methods in Information Retrieval, pp. 155-167, 2002

[7] 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", Proc. of Interspeech, 2005.

[8] Illina I., Fohr D., Mella O., Cerisara C "The Automatic News Transcription System: ANTS, some Real Time experiments", In Proc. ICSLP, 2004.

[9] Illina I., Fohr D., Jouvet D. "Grapheme-to-Phoneme Conversion using Conditional Random Fields", In Proc. Of Interspeech, 2011.

[10] Kobayashi A., Onoe K., Imai T., Ando A. "Time dependent language model for broadcast news transcription and its post-correction", In Proc. Of ICSPL, 1998

[11] Lee, A. and Kawahara, T. "Recent Development of Open-Source Speech Recognition Engine Julius", Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2009.

[12] Nkairi, I., Illina, I., Linarès, G., Fohr, D. « Exploring temporal context in diachronic text documents for automatic OOV proper name retrieval", In Proc. of LTC, 2013.

[13] Oger, S., Linarès, G. and Béchet, F. "Local methods for on-demand out-of-vocabulary word retrieval", Proc. of the Language Resources and Evaluation Conference (LREC), 2008.

[14] Parada, C., Dredze, M., Filimonov, F. and Jelinek, F. "Contextual Information Improves OOV Detection in Speech", Proc. of NAACL, 2010.

[15] Schmid, H. "Probabilistic part-of-speech tagging using decision trees", Proc. of ICNMLP, 1994.

[16] Stolcke, A. "SRILM - An Extensible Language Modeling Toolkit", Proc. of ICSLP, 2002.