



Continuous Word Representation using Neural Networks for Proper Name Retrieval from Diachronic Documents

Dominique Fohr, Irina Illina

MultiSpeech team
 Université de Lorraine, LORIA, UMR 7503, Vandoeuvre-lès-Nancy, F-54506, France
 Inria, Villers-lès-Nancy, F-54600, France
 CNRS, LORIA, UMR 7503, Vandoeuvre-lès-Nancy, F-54506, France

Abstract

Developing high-quality transcription systems for very large vocabulary corpora is a challenging task. Proper names are usually key to understanding the information contained in a document. One approach for increasing the vocabulary coverage of a speech transcription system is to automatically retrieve new proper names from contemporary diachronic text documents. In recent years, neural networks have been successfully applied to a variety of speech recognition tasks. In this paper, we investigate whether neural networks can enhance word representation in vector space for the vocabulary extension of a speech recognition system. This is achieved by using high-quality word vector representation of words from large amounts of unstructured text data proposed by Mikolov. This model allows to take into account lexical and semantic word relationships. Proposed methodology is evaluated in the context of broadcast news transcription. Obtained recall and ASR proper name error rate is compared to that obtained using cosine-based vector space methodology. Experimental results show a good ability of the proposed model to capture semantic and lexical information.

Index Terms: speech recognition, neural networks, vocabulary extension, out-of-vocabulary words, proper names.

1. Introduction

Large-vocabulary Automatic Speech Recognition (ASR) systems are faced with the problem of *out-of-vocabulary* (OOV) words (words that are not in ASR system vocabulary) when used for very large vocabulary corpora recognition or in new domains (voice search, spoken dialog systems, etc.). Among these OOVs, *Proper Names* (PNs) are very largely represented. Note that these PN words constantly evolve and no vocabulary will ever contain all existing PNs [8].

After some time of abandon [4], in recent years *Neural Networks* (NN) have been successfully applied to a variety of speech recognition tasks [6][18][22] and to several levels of speech recognition modelling [1][15][20]. Recently, Mikolov *et al.* [15][16][17] have proposed an original NN architecture for continuous word representations in vector space: *word2vec*. The goal is to capture a large number of semantic and syntactic word relationships using huge amounts of unstructured text data. Linguistic regularities and patterns are learned using continuous distributed context representation of words, maximizing accuracy and minimizing computational complexity. The results on word similarity task outperformed the previously proposed best techniques. Compared to other techniques of continuous word representation, like LDA (*Latent Dirichlet Analysis*) [3] or LSA (*Latent Semantic*

Analysis) [5], this NN-based representation gives distributed word representation and preserves linear regularity.

In our previous work, we investigated methods that augment the vocabulary of the ASR system with new PNs, using lexical and temporal features. Our context model was based on mutual information and vector space model [10]. Our hypothesis is that PNs evolve through time, and that for a given date, the same PNs would occur in documents that belong to the same time period. This is taken into account by the temporal context [13].

In this paper, we focus on the same problem of PN retrieval using lexical and temporal context for ASR vocabulary extension. The novelty of our paper is the use of the high-quality continuous space word representation proposed by Mikolov *et al.* Our work uses *word2vec* model architecture to capture the lexical and semantic relations so as to retrieve OOV PNs and increase the ASR vocabulary size. Compared to previous works on OOV retrieval [2][7], NN has never been used for this task.

Section 2 introduces the proposed approach. Sections 3 and 4 describe the experimental sets and the results of the evaluation of this model.

2. Proposed methodology

Our proposed methodology is based on temporal and lexical context proposed in [10]: we use text documents from a diachronic corpus that are contemporaneous with each test document to be transcribed. The diachronic documents allow building a specific 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 and to select relevant PNs. We assume that, for a certain date, a PN from the test corpus will co-occur with other PNs in diachronic documents of the same time period. These co-occurring PNs might contain the targeted OOV words. The idea is to exploit the relationships between PNs for a better lexical enrichment.

In [10], different PN selection strategies have been proposed to build this augmented PN vocabulary. In the present work, we propose to use high-quality vector representation of words from large amounts of unstructured text data proposed by Mikolov *et al.* [16]. We chose to use the *Skip-gram* model because it achieved very good results on semantic tasks. This *Skip-gram* model tries to predict surrounding words of one input word. This is performed by maximizing the classification rate of these nearby words given the input word. More formally, given a sequence of training words w_1, w_2, \dots, w_T , it maximizes the average *log* probability:

$$\frac{1}{T} \sum_{t=1}^T \sum_{-c \leq j \leq c, j \neq 0} \log p(w_{t+j} | w_t) \quad (1)$$

where c is the context size and T the number of training words. Compared to classical NN, the non-linear hidden layer is removed and the projection layer is shared for all words. This model assumes that semantically similar words will be projected in the same space region. An important property of this model is that the word representations learned by the *Skip-gram* model exhibit a linear structure.

We propose to train Mikolov’s *Skip-gram* on a large text corpus.

2.1. OOV retrieval method

Our OOV retrieval method consists of 5 steps:

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

For each test document, we extract IV PNs from the automatic transcription performed using our standard vocabulary. The goal is to use these PNs as anchors to collect linked new proper names from the diachronic corpus.

B) Selection of diachronic documents and extraction of new PNs from them: only diachronic documents (DDs) that correspond to the same time period as the test document are considered. After POS-tagging of these DDs, meaningful words are kept: verbs, adjectives, nouns and PNs. Among these PNs, we create a list of those that do not belong to our standard vocabulary (OOV PN).

C) Temporal and lexical context extraction from diachronic documents (DD): The goal is to extract the most relevant OOV PNs. After extracting the list of the IV PNs from the test document (step A), and the list of the new PNs from DDs (step B), we build their *temporal and lexical* contexts. For this, a high-dimensionality word representation space is used (see description below). We hope that in this space semantically and lexically related words will be in the same region of the space.

D) Ranking of new PNs: The cosine-similarity metric is calculated between the projected vector of IV PNs found in the test document and the projected vector of each OOV PN occurring in the selected diachronic documents.

E) Vocabulary augmentation: To reduce the vocabulary growth, only the top-N OOV PNs according to the cosine-similarity metric are added to our vocabulary. OOV 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, cf. Section 4.1) of all the potentially missing PNs.

2.2. Continuous word representation using neural networks

At step *C*, we propose to model the word space using Mikolov’s model: each word in this high-dimensional space is represented by a continuous vector that takes into account semantic and lexical relationships.

In the task of OOV retrieval it is important to use semantic and lexical word context. During the training, Mikolov’s *Skip-gram* model takes into account linear relationships between the word vectors by using local context window (given by the parameter c of eq. (1)). During the step *C* of our proposed method, for a PN only the NN-projection of this PN is used. However, the context of this PN can be different compared to the training step. To better take into account the local context of the PN, we propose to use a *local-window context* NN-projection (LWCP) for each occurrence of the PN (IV or OOV)

in test or in diachronic documents. LWCP can be computed by vector addition because of the linearity of the *Skip-gram* model. So, in step *C*, we propose to use the sum of NN-projections of context words as LWCP, instead of using the NN-projection of each word. After this, in step *D*, cosine similarity can be calculated between the LWCP of IV PNs and the LWCP of OOV PNs. Using LWCP for each occurrence of IV PN and for each occurrence of new OOV PN from diachronic documents is time consuming, thus we propose to use LWCP only for each occurrence of IV PN in the test document (cf. Figure 1). The complexity of the NN-method depends only on the number of PNs.

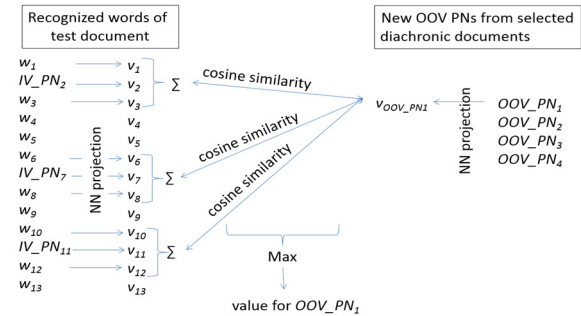


Figure 1. Local-window context NN-projection for each occurrence of IV PN of the test document.

3. Experiments

We call *selected PNs* the new proper names that we were able to retrieve from DDs using our methods. We call *retrieved OOV PNs* the OOV PNs that we were able to retrieve from DDs using our method and that are present in the test documents. Using the DDs, we build a specific augmented lexicon for each test document according to the chosen period.

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

3.1. Development and test corpora

As development corpus, seven audio documents of development part of ESTER2 (between 07/07/2007 and 07/23/2007) are used. For the test corpus, 13 audio documents from RFI (*Radio France International*) and *France-Inter* (test part of ESTER2) (between 12/18/2007 and 01/28/2008) [9] are used. Table 1 gives the average occurrences of all PNs (IV 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	IV PNs	IV 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

French *GigaWord* corpus is used as diachronic corpus: newswire text data from *Agence France Presse* (AFP) and *Associated Press Worldstream* (APW) from 1994 to 2008. The choice of GigaWord and ESTER corpora was driven by the fact that one is contemporary with the other, their temporal granularity is the day and they have the same textual genre (journalistic) and domain (politics, sports, etc.).

3.3. Transcription system

ANTS (*Automatic News Transcription System*) [12] is based on Context Dependent HMM phone models trained on 200-hour broadcast news audio files. The recognition engine is Julius [14]. The baseline phonetic lexicon contains 260k pronunciations for the 122k words. Using the SRILM toolkit [21], 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 beyond 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 method. In a second step, we will evaluate the proposed approach on the test set.

4.1. Baseline results

The baseline method consists in extracting a list of all the OOV PNs occurring in the selected diachronic documents corresponding to the time period of the document to be transcribed. This period can be, for example, a day, a week or a month. Then, our vocabulary is augmented with the list of all extracted OOV PNs. The problem with this approach is that if the diachronic corpus is large, a bad tradeoff between the lexical coverage and the increase of the lexicon size is obtained.

Using TreeTagger [19], we extracted 160k PNs from 1 year of the diachronic corpus. Among these 160k PNs, 119k are not in our lexicon. Among these 119k, only 151 PNs 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 the lexicon size.

Time period	Average of selected PNs per file	Average of retrieved OOV PNs per file	Recal 1 (%)
1 day	532.9	10.0	32.6
1 week	2928.4	11.4	37.2
1 month	13131.0	17.6	57.2
1 year	118797.0	24.0	78.1

Table 2. Baseline results for *development* corpus according to time periods. Values averaged on the 7 development files.

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

4.2. NN-based results

We used Mikolov’s open-source NN project available on the web. The NN is trained on the diachronic corpus described in Section 3.2 using only meaningful words. For this network, the important parameters are: the model architecture, the size of hidden layer (which is also the dimension of the projection space) and the size of the context window. After several

experiments, we defined the best parameter set that will be used here: 400 for the size of the hidden layer, 20 for the context size. We observed that the *Skip-gram* works better than Mikolov’s *bag-of-word model*. A preliminary study of the extension from word-based to expression-based representation (like “New York” and “Times” versus “New_York_Times”, called *phrases* according to Mikolov) has shown a limited improvement. We performed 5 training epochs. Moreover, for the month time period, in order to select more relevant PNs and to remove spelling errors, we introduced a frequency threshold: the OOV PN occurring less than 3 times in the selected diachronic documents are excluded

Figure 2 shows the results for the proposed NN-based method using different time periods for the development corpus. The recall increases rapidly at the beginning and then a plateau is reached. This shows the good quality of NN word representation: many relevant PNs are very well classified by the neural networks. The plateau is reached for about 400 PNs for the week period and for about 2000 PNs for the month period.

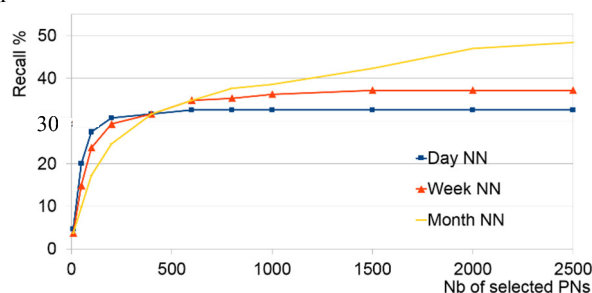


Figure 2. NN-based results according to time duration period depending on number of selected PNs. *Development* corpus.

To evaluate the importance of selecting documents of the same period than the document to transcribe, we set up a *temporal mismatch* experiment (called “*mism*” in Table): we extract OOV PN, from DDs occurring 10 months after the date of the document to be transcribed.

Table 3 represents some results of Figure 1 with a fixed number of selected PNs corresponding to an operating point of 15% of the average number of selected PNs per development file: 80 for day, 440 for week and 2000 for month (cf. Table 2, about 15% of 532, 2928 and 13131). This operating point is chosen to obtain a good recall with a reasonable number of *selected PNs*.

Time period	Method	Selected PNs	Retrieved OOV PNs	Recall (%)
1 day	NN	80	7.4	24.2
	NN mism	80	2.3	7.4
1 week	NN	440	9.9	32.1
	NN mism	440	5.6	18.1
1 month	NN	2000	14.4	47.0
	NN mism	2000	11.1	36.3

Table 3. NN-based results according to time duration period for *development* corpus, with and without temporal mismatch. Values averaged on the 7 development files.

Table 3 shows that the operating point of 15% gives good performance, indeed using only 15% of all *selected PNs* from diachronic documents, we lost only about 20% of recall compared to the Table 2. NN-based method with temporal context performs significantly better than NN with mismatch time period. This is true for all period durations. For example,

if we use the diachronic documents for the same week that development document, the obtained recall is 32.1%. If we use the DDs for the week period but 10 months later than the document to be transcribed, the performance is almost divided by 2 (18.1% versus 32.1%). These results demonstrate the importance of temporal context.

4.3. Comparison of NN-based method with previously proposed cosine-based method

4.3.1 Cosine-based method

We have compared the NN-based method proposed in this paper with the previously proposed cosine-based method. We chose this method because it based on the classical bag-of-words (BOW) paradigm and it gave good results [10]. We remind briefly this method here.

Only the step *C* (cf. Section 2) of our method is modified, other steps are kept unchanged. During the step *C*, instead of using the NN-space, each document (diachronic or test) is represented as a bag-of-words vector of meaningful words. So, to every document is associated a BOW vector. Moreover, we associate one BOW vector to each OOV PN as the sum of all BOW vector documents in which this OOV PN occurred. During the step *D* of our method, to select relevant PNs, the cosine similarity between the BOW vector of the test document and the BOW vector of each new OOV PN occurring in selected document of diachronic set is computed.

The complexity of this method is higher than that of the NN method. Indeed, the computation of the BOW vector depends on the number of selected DDs. For the NN method, the computation of the projection is independent of the number of selected DDs and can be precomputed and stored in a table after the training phase.

4.3.2 Experimental comparison

Table 4 illustrates the benefit of using the continuous word representation given by NN instead of the cosine-based one. For all studied time periods, NN-based system achieved better performance than the cosine-based system. This can be explained by the fact that NN possibly preserves better the semantic relations [15] between the words than the cosine-based method.

Increasing the duration of time periods, the performance gap between two methods increases. For example, for a day period the absolute difference is 0.9%, for a month it is 7.5%.

Time period	Method	Selected PNs	Retrieved OOV PNs	Recall (%)
1 day	NN	80	7.4	24.2
	Cos	80	7.1	23.3
1 week	NN	440	9.9	32.1
	Cos	440	9.3	30.2
1 month	NN	2000	14.4	47.0
	Cos	2000	12.1	39.5

Table 4. NN-based and cosine-based results according to time duration period for **development** corpus. Values averaged on the 7 development files.

For the cosine-based method, the average rank of retrieved OOV PNs is higher than for the NN-based method: for the week period, 243 versus 177 and for the month period 923 versus 560. For the day period the average rank is similar for two methods. This confirm the better performance of NN-based method.

4.4. Automatic speech recognition results for the development corpus

We performed automatic transcription of the 7 development documents using augmented lexicons for the proposed method (generating one lexicon per development file and per period). For generating the pronunciations of the added PNs, we used the G2P CRF approach [11], trained on phonetic lexicon containing about 12000 PNs. In order to incorporate the new PNs in the language model, we re-estimated it for each augmented vocabulary using the large text corpus described in Section 3.3. The number of selected PNs per period is the same as in Table 3 and 4: 80 for a day, 440 for a week and 2000 for a month. Table 5 shows that, compared to standard lexicon, a significant improvement is obtained for the NN system in terms of PN error rate (38.3% versus 43.6%). However, there is no significant difference between NN and cosine-based results.

4.5. Automatic speech recognition results for test corpus

For validating the proposed approach, we performed automatic transcription of the 13 test documents using augmented lexicons of the proposed method (cf. Table 5) and the parameters chosen for the development corpus. As for the development corpus, in order to incorporate the new PNs in the language model, we re-estimated it for each augmented vocabulary. Similar performance is observed for both methods. PN error rate is significantly improved when an augmented lexicon is used. In terms of word error rate, the improvement is not significant.

Stand. lexicon	Method	Augmented lexicon		
		1 day	1week	1month
Development corpus				
43.6	PNER NN	40.9	40.4	38.3
	PNER Cos	40.4	40.2	38.6
Test corpus				
46.4	PNER NN	43.7	42.6	41.0
	PNER Cos	43.8	42.3	41.0

Table 5. PNER (%) for NN and cosine methods according to time duration period for **development** and **test** corpus.

5 Conclusion

We have investigated the problem of OOV words and the vocabulary extension of a speech recognition system. DDs contemporary to test documents were used to retrieve proper names for enriching the vocabulary. We focused on the continuous space word representation using neural networks. We learned this representation from large amounts of unstructured text data.

Experimental results suggest the proposed method can model the semantic and lexical context of proper names and can be useful for PN retrieval from DDs. Compared to the cosine-based method, NN gives better recall and has lower complexity. In terms of PNER, the results of both methods are comparable and show a significant decrease of the proper name error rate. Future research will study how to optimize both the OOV retrieval and the DD selection in NN space representation.

6 Acknowledgements

This work is funded by the *ContNomina* project supported by the French national Research Agency (ANR) under the contract ANR-12-BS02-0009.

7 References

- [1] Bengio, Y., Ducharme, R., Vincent, P. "A neural probabilistic language model" *Journal of Machine Learning Research*, 3: 1137-1155, 2003.
- [2] Bigot, B., Senay, G., Linares, G., Fredouille, C., and Dufour, R. "Person name recognition in ASR outputs using continuous context models", *Proceedings of ICASSP*, 2013.
- [3] Blei, D., Ng, A.-Y., Jordan, M.I. "Latent Dirichlet Allocation", *Journal of Machine Learning Research*, 3: 9953-1022, 2003.
- [4] Bourlard, H., Morgan, N. "Connectionist Speech Recognition: a hybrid approach", Kluwer, 1994.
- [5] Deerwester, S., Dumais, S.T., Fumas, G.W., Landauer, T.K. and Harshman, R. "Indexing by latent semantic analysis", *Journal Assoc. Inf. Science Techn.*, vol. 41, n.6, 391-407, 1990
- [6] Deng, L., Li, J., Huang, J.-T., Yao, K., Yu, D., Seide, F., Seltzer, M., Zweig, G., He, X., Williams, J., Gong, Y. and Acero A. "Recent Advances in Deep Learning for Speech Research at Microsoft", *Proceedings of ICASSP*, 2013
- [7] Federico, M. and Bertoldi, N. "Broadcast news LM adaptation using contemporary texts", *Proceeding. of Interspeech*, pp. 239-242, 2001
- [8] Friburger, N. and Maurel, D. "Textual Similarity Based on Proper Names", *Proceedings of the workshop Mathematical/Formal Methods in Information Retrieval*, pp. 155-167, 2002
- [9] 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", *Proceedings of Interspeech*, 2005.
- [10] Illina, I., Fohr, D. and Linares, G. "Proper Name Retrieval from Diachronic Documents for Automatic Transcription using Lexical and Temporal Context", *Proceedings of SLAM*, 2014.
- [11] Illina I., Fohr D., Jouvét D. "Grapheme-to-Phoneme Conversion using Conditional Random Fields", *Proceedings of Interspeech*, 2011.
- [12] Illina I., Fohr D., Mella O., Cerisara C. "The Automatic News Transcription System: ANTS, some Real Time experiments", *Proceedings of ICSLP*, 2004.
- [13] Kobayashi A., Onoe K., Imai T., Ando A. "Time dependent language model for broadcast news transcription and its post-correction", *Proceedings of ICSPL*, 1998
- [14] 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.
- [15] Mikolov, T., Chen, K., Corrado, G. and Dean, J. "Efficient Estimation of Word Representations in Vector Space", *Proceedings of Workshop at ICLR*, 2013.
- [16] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. and Dean, J. "Distributed Representations of Words and Phrases and their Compositionality", *Proceedings of NIPS*, 2013.
- [17] Mikolov, T., Yih, W. and Zweig, G. "Linguistic Regularities in Continuous Space Word Representations", *Proceedings of NAACL HLT*, 2013.
- [18] Ng, T., Zang, B., Nguyen, L., Matsoukas, S., Vesely, K., Matejka, P., Zhu, X. and Mesgarani, N. "Developing speech activity detection system for the DARPA RATS program", *Proceedings of Interspeech*, 2012.
- [19] Schmid, H. "Probabilistic part-of-speech tagging using decision trees", *Proceedings of ICNMLP*, 1994.
- [20] Seide, F., Li, G., Yu, D. "Conversation speech transcription using context-dependent deep neural network", *Proceedings of Interspeech*, 2011.
- [21] Stolcke, A. "SRILM - An Extensible Language Modeling Toolkit", *Proceedings of ICSLP*, 2002
- [22] Vinyals, O., Ravuri, S.V. And Povey, D. "Revisiting recurrent neural network for robust ASR", *Proceedings of ICASSP*, 2012