Local methods for on-demand out-of-vocabulary word retrieval

Stanislas Oger, Georges Linarès, Frédéric Béchet

Laboratoire d'Informatique d'Avignon (LIA) - University of Avignon
339 ch. des Meinajaries, BP 1228
F-84911 Avignon Cedex 9 (France)

{stanislas.oger,georges.linares,frederic.bechet}@univ-avignon.fr

Abstract

Most of the Web-based methods for lexicon augmenting consist in capturing global semantic features of the targeted domain in order to collect relevant documents from the Web. We suggest that the local context of the out-of-vocabulary (OOV) words contains relevant information on the OOV words. With this information, we propose to use the Web to build locally-augmented lexicons which are used in a final local decoding pass. First, an automatic web based OOV word detection method is proposed. Then, we demonstrate the relevance of the Web for the OOV word retrieval. Different methods are proposed to retrieve the hypothesis words. We finally retrieve about 26% of the OOV words with a lexicon increase of less than 1000 words using the reference context.

1. Introduction

Statistical language models for natural language processing are generally estimated on static text corpora. In spite of the potentially large amount of training data, these models are subject to the problem of out-of-vocabulary (OOV) words when they are confronted to highly epoch-dependent data where topics and named entities are frequently unexpected.

The problem of OOV words remains a key point in large vocabulary continuous speech recognition, especially on transcription of contemporary documents like broadcast news.

OOV words are words which are in the speech signal but not in the automatic speech recognition (ASR) system lexicon, in this case the ASR system is unable to transcribe correctly the OOV words. These words are critical for sentence intelligibility since most of them are named entities and technical words, which implies that the performance of ASR-based dialog systems are also affected.

The extensive growing of dictionaries with little regards to the trade-off between the lexical coverage and the increase of lexicon size leads to dramatically increase the resources required by an ASR system. Moreover, the real world is an inexhaustible source of new words which can not be fully listed in any closed lexicon.

Substantial efforts have been recently produced in using external text sources for lexicon augmenting. Some papers report experiments on *a posteriori* search of new words in large external databases (Ohtsuki et al., 2005), but such static approaches fail in contemporary document transcription, where topics and named entities are frequently unexpected. Nevertheless, the web constitutes an immense and continuously updated source of language data, in which most of the *possible* word-sequences are stored. This idea has been largely developed in the field of large vocabulary continuous speech recognition. Generally, authors proposed to collect a large amount of documents that are supposed to be relatively close to the targeted linguistic and semantic context (Kajiura et al., 2006)(Allauzen and Gauvain, 2005)(Monroe et al., 2002)(Bertoldi and Federico, 2001). Unfortunately, web-data suffers from the lack of structuring information and large well-targeted corpora generally outperform web-based language models (Lapata and Keller, 2005).

Focusing on the problem of lexicon building, OOV word retrieval methods tackle the difficulty of how missing words could be automatically found on the web. These problems are traditionally addressed by capturing the global semantic features from the document and collecting relevant documents which are used for language modeling.

We hypothesize that the Web can be used as an unlimited collection of ngrams. Search engines can be considered as interfaces for these ngrams if efficient requests are built.

In this paper, we suggest that the local linguistic context might bring some characteristic information of the missing words, and that this information could allow to retrieve words in the unlimited collection of web-ngrams. Starting from this idea, we propose local methods for OOV word retrieval.

The next section presents the experimental framework in which our experiments are carried out. Then, a web based OOV word detection method is proposed. In the following section, we evaluate the hypothesis that the web is relevant for the task of unknown word retrieval and we propose strategies funded on word-template matching. These methods are both evaluated on exact transcription and on the outputs of the ASR system.

Lastly, we conclude on the interest of web-based local approaches for new word learning in the field of posterior correction of automatic transcriptions.

2. Experimental framework

Our general approach consists in using the local context of the OOV words to build efficient requests for submitting to the search engine. The retrieved documents are then parsed in order to find the targeted OOV words.

The experiments are carried out in the framework of ESTER evaluation campaign (Gravier et al., 2004) and the Google search engine is used to access Web data[1]. All the

This research is supported by the ANR (Agence Nationale de la Recherche), SIGMUND project ANR-05-RIAM-0903.

[1] http://www.google.fr

tests are performed on about 6 hours of French broadcast news from the test corpus of ESTER 2005. The table 1 provides details on the amount of audio data used, according to the radio station.

The figure 1 shows a graphic representation of the experimental framework, from the speech signal to the final transcription.

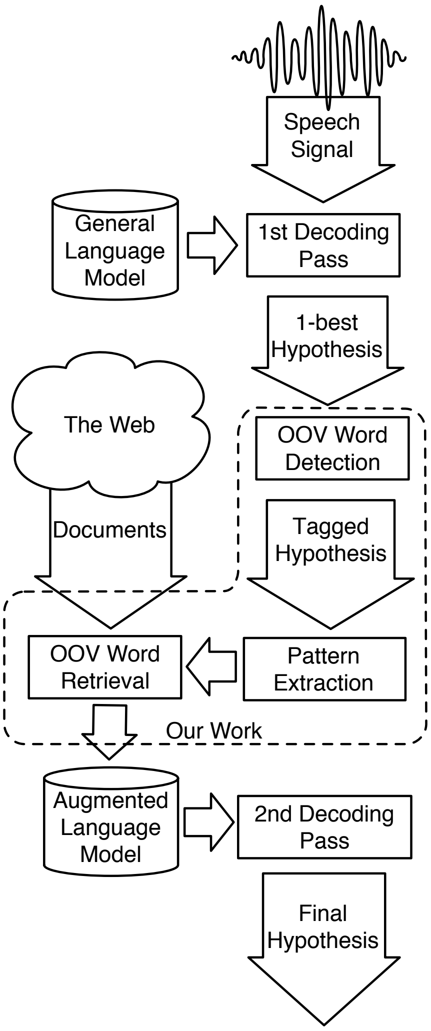

Figure 1: General view of the experimental framework

The speech signal is transcribed using the LIA broadcast news system, SPEERAL (Nocera et al., 2004). This system is an A * decoder based on state-dependent HMM for acoustic modeling. The language model used for this first pass is a 3-gram estimated on about 200M words from the French newspaper Le Monde and from the ESTER broadcast news corpus (about 1M words). We use a lexicon made of the 65000 most frequent words of these corpora.

The test corpus we used contains 5.5 hours of French broadcast news from the ESTER database. On this test set, the first decoding pass obtain a WER of 24.5%. The OOV word rate is about 1,03%. 73% of the OOV words are named entities, 24% are technical and domain-specific words, and the other 3% are infrequent verb forms or misspelled words in the reference transcription. It is important to notice that named entities and technical words are critical for the sentence intelligibility and represent 97% of the OOV words.

OOV words are then automatically detected in the automatic transcription. In this work, we present automatic OOV words detection methods using the Web and the local context.

Local patterns and keywords are then extracted from the tagged transcription and are used to retrieve web documents. These documents are parsed taking account of the pattern in order to build lists of OOV word candidates.

The OOV word candidates are then integrated in the general language model to form the augmented language model which is used for the final transcription pass. This part of the system is not studied in this paper.

Radio station	amount of audio data
France Classique	1.00
France Inter	1.00
France Info	2.00
Radio France Internationale	1.50
Total	5.50

Table 1: The amount of audio data in hours used in this work, sorted by radio station.

3. OOV Words detection

Many works have been made on the OOV word detection, for example (Tanigaki et al., 2000; Hazen and Bazzi, 2001; Bazzi and Glass, 2002; Yazgan and Saraclar, 2004). These methods are applied during the transcription and are parts of the speech recognition engine. We propose here an *a posteriori* OOV word detection.

3.1. The basic hypothesis

When an OOV word occurs in the speech signal, we observe a strong decrease of the system accuracy. The reason of this phenomenon is that the system try to find words which fit to the acoustic signal but they don't fit to the linguistic context, which strongly perturbs the language model and so degrades the transcription around the OOV word. We assume that OOV words impact dramatically the system accuracy. Therefore by finding the worst parts of hypothesis, we expect to focus on areas which probably contain OOV words. In order to detect low quality zones which potentially contain OOV words, metrics are computed on each word.

3.2. A web-based approach

First, a boolean parameter inform about the presence of the 3-gram on the general language model used for the first pass. The 3-gram is composed of the word and the two words before. A second boolean metric indicates if the 3-gram is on the Web. A search engine with constrained queries is used.

With the boolean metrics, the words are considered as OOV if the 3-gram is absent. The metrics combination is performed by a simple vote. A word is detected as OOV if all the metrics detect it as OOV.

3.3. Experiments

In order to evaluate these metrics, the false positive (FP) rate (see formula 1) and the false negative (FN) rate (see formula 2) are measured for the detection of OOV words. The Google search engine is used to test the presence of the 3-grams on the web.

$$\mathrm{FP} = \frac{\text{number of correct words considered as OOV}}{\text{total number of correct words}} \quad (1)$$

$$\mathrm{FN} = \frac{\text{number of OOV words considered as correct}}{\text{total number of OOV words}} \quad (2)$$

The table 2 represents the FP and FN rates of the Google metric and the language model metric separately and a combination of the two for the detection of OOV words. The performance of the Google metric is the same as the combination of the two, which indicates that the 3-grams which are not on the Web is a sub-group of the 3-grams which are not in the language model, so the combination is not useful. The Google metric has a good FP rate, which indicates that this metric is powerful for the OOV word detection. Moreover, the FR rate of the language model metric is low, which indicates that this metric is useful for the detection of the in-vocabulary words.

	Google 3-gram	Language Model 3-gram	Google+LM 3-gram
False Posituve	17.89 %	66.96 %	17.70 %
False Negative	50.94 %	11.32 %	50.94 %

Table 2: FP and FN rate of the proposed metrics for the detection of the OOV words in the automatic transcription.

3.4. Conclusion

These results indicates that the Google metric and the language model metric are complementary. The Googe metric allows to isolate low confidence segments with a good precision whereas high confidence segments are found by the language model metric with a better precision.

4. New word learning

In this part, we first evaluate the hypothesis that the web is an exhaustive source of words and is relevant for the task of unknown word retrieval. Then we present some pattern extraction methods and study the impact of automatic transcription on retrieval performance.

4.1. The web as an unlimited source of words

In the purpose of retrieving unknown words by using the local context, hypothesizing the web as an exhaustive source of words leads to consider it as an infinite n-gram model. Such model contains all possible n-grams, comprising the ones with the targeted OOV word w_t.

In order to evaluate this assumption, we measure the rate on the web of the n-grams containing OOV words extracted from the exact transcriptions. Requests are submitted to Google. The results presented in table 3 show that all the targeted words and most of the 2-grams containing w_t can be found on the web. As expected, the recall decreases when n increases. These results indicate that the web has an interesting potential in the task of OOV word retrieval, but under the condition we know how to formulate relevant requests. This key point is tackled in the next section.

n-gram	1	2	3	4	5
Recall	100.00 %	88.22 %	50.54 %	27.29 %	16.12 %

Table 3: *n*-grams containing OOV words retrieval on Google depending on the size *n*.

4.2. Collecting augmented lexicons

Here we evaluate the performance of various methods for OOV word retrieval using both the ASR system outputs and the exact transcription. For each OOV word in the transcription a list of candidate words is built using each methods. The recall rate and the average size of these candidate lists is measured and compared. The segments containing OOV words are manually pointed out in order to avoid the automatic OOV detection errors.

4.2.1. N-gram strategy

The first technique consists in building requests by taking the *n*-gram containing w_t and substituting it by a wild-card character. For example, for the sentence :

> Les otages Christian *Chesnot* et Georges [...]

with *Chesnot* the OOV word, the corresponding 3-gram pattern is :

> “otages Christian (*)”

The hypothesis word sets are built with the words which replace the wild-card character *(*)* in the retrieved documents.

This strategy was tested with both the correct word utterance and the outputs of the ASR system in order to measure the impact of the ASR errors on the retrieval performance. Results are reported in table 4 for the recall and 5 for the hypothesis word set size.

Right, left and both contexts was tested and the experiments shown that these configurations obtain about the same performance. The results presented in this paper use the left context.

It is obvious that the recall rate with the ASR system outputs is generally worse than with the reference context. The cause is that the automatic transcription contains errors around the OOV words. Moreover, it’s commonly admitted that the presence of OOV words locally perturb the ASR system around it.

We observe that the use of short sequences leads to low discriminative requests. When sequence size increases, precision improves very quickly but the recall drops and requests become useless.

In conclusion, it seems clear that since targeted words appear on the web, strict n-gram strategy is not discriminative enough for OOV word retrieval. In the next section, we propose to relax word utterance sequentiality constraints by defining patterns rather than hard-fixed n-grams.

4.2.2. Pattern strategy

Here, we build soft requests by extracting word templates from the context. This is achieved by taking out the most frequent French words (the stop-words) and inserting wild-card characters between each words. This wild-card will be automatically substituted by one to five words by the search engine, which allows to retrieve variants of the context. For example, for the sentence :

> Les otages Christian *Chesnot* et Georges [...]

with *Chesnot* the OOV word, the corresponding pattern with 3 keywords is :

> “otages * Christian (*)”

The candidates are extracted by taking the words which replace the wildcard between brackets in the retrieved documents.

The results presented in tables 4 and 5 show that recall rates are better than when using the n-gram based strategy with both the ASR system outputs and the exact transcription context. In addition, the average size of hypothesis word sets increases a bit. It indicates that relaxing constraints on stop-words allows to retrieve variants of the original context, which introduces noise in hypothesis-sets but enables to increase recall. Moreover, we notice that the 2-gram recall decreases less than with the previous strategy when the ASR system outputs context is used, indicating a better robustness in this specific configuration.

4.2.3. Short-term semantics-based strategy

In order to reduce the impact of the ASR errors, only the relevant words are used without ordering constraints. Words in a short temporal window around the OOV word are sorted by decreasing language frequency and only the top *n* words are used as keywords and submitted to Google. For example, for the sentence :

> Les otages Christian *Chesnot* et Georges [...]

with *Chesnot* the OOV word, the corresponding 3 keywords are :

> otages Christian Georges

The sets of hypothesis words are built by taking the lexicons of the whole best ranked documents.

We can see in tables 4 and 5 that the recall rate increases strongly with the number of keywords, which vary from 2 to 5. With the best configuration, the recall rate is more than twice better than the recall with the best configuration of the previous strategies. However the precision decreases a lot, and this last point is a major drawback for the integration of the augmented lexicon in the speech recognition engine. Nevertheless, considering the complementarity of this approach with the previous ones, a composite strategy combining patter-matching and semantics-based word retrieval could be efficient.

	n-gram strategy		pattern strategy		semantics strategy	
n	REF	ASR	REF	ASR	REF	ASR
2	13.95 %	4.65 %	20.00 %	7.29 %	32.56 %	18.45 %
3	18.14 %	5.12 %	20.31 %	4.96 %	39.69 %	27.75 %
4	16.43 %	2.33 %	17.52 %	2.02 %	45.89 %	35.19 %
5	13.80 %	1.86 %	12.25 %	1.24 %	50.23 %	40.93 %

Table 4: Recall of OOV word retrieval on the best 100 Google ranked documents depending on the size *n*, with the exact transcription (REF) and the ASR output context.

	n-gram strategy		pattern strategy		semantics strategy	
n	REF	ASR	REF	ASR	REF	ASR
2	145	322	411	475	16.0k	13.7k
3	49	207	139	166	19.0k	38.1k
4	13	34	34	21	37.9k	42.6k
5	4	9	15	8	44.9k	45.0k

Table 5: Average hypothesis-set size of OOV words retrieval on the best 100 Google ranked documents depending on the size *n* with the exact transcription context (REF) and the ASR system outputs.

4.2.4. N-gram semantics-driven strategy

As shown in the section 4.2.1., only 13.95% of the 2-grams containing the targeted word can be retrieved whereas almost 88 % are on the Web. We presume that all 2-grams containing the targeted words are in the documents returned by the search engine but not well ranked. We assume that adding relevant context words in the n-gram requests may help the search engine to better rank documents which are relevant for the context and, we hope, contain the targeted n-gram. These additional words are called here drive-words.

The drive-words are selected in a fixed-size window around the OOV word, sorted by decreasing language frequency and the top *n* words are kept as drive-words. These words are added to the request as keywords without ordering constraints. For example, for the sentence :

> Les otages Christian *Chesnot* et Georges [...]

with *Chesnot* the OOV word, the corresponding 3-gram/1-drive-word is :

> “otages Christian (*)” +Georges

The hypothesis-sets are built by selecting all potential n-grams, like in the previous n-gram strategy. The results using the reference context and ASR system outputs are reported in table 6.

Results significantly outperform previous ones with the reference context (tables 4 and 5). Good recall rates are obtained by using slightly augmented lexicon. For example the $2/2$ configuration obtains about 26% of recall for a dictionary increase of less than 1000 words. However, using the ASR system outputs degrades the recall even if the method remains the best of all in terms of recall with low lexicon increase.

	Semantics driven *n*-gram strategy			
	REF		ASR	
n/m	Recall	sets size	Recall	sets size
2/1	24.03 %	268	8.68 %	292
2/2	26.05 %	789	8.06 %	306
2/3	26.98 %	1.3k	6.51 %	295
3/1	19.07 %	16	4.03 %	87
3/2	15.04 %	15	3.88 %	79
3/3	13.33 %	19	3.10 %	98

Table 6: Recall and average hypothesis-set size of OOV words retrieval on the best 100 Google ranked documents, using the semantics-driven n-gram strategy depending on the *n* value and the number of drive-words *m*, using the reference context (REF) and the ASR system outputs.

5. Conclusion and perspectives

First, a web-based automatic OOV word detection method is proposed. Then the huge potential of the World Wide Web for the task of OOV word retrieval using the local context is demonstrated. We proposed and compared local approaches for lexical coverage improvement. In our strategies, search engine requests are dynamically made using the local context and augmented lexicons are built with the returned documents.

We presented several ways for request formulation and our results validate the initial idea that the short-term context holds some characteristic information about missing words. The best performance is obtained by combining local word templates and semantics-driven requests. Moreover, this strategy allows to retrieve OOV words with a limited lexicon augmentation and most of the successfully retrieved words consist in named entities which are crucial for understanding tasks.

6. References

A. Allauzen and J.L. Gauvain. 2005. Open Vocabulary ASR for Audiovisual Document Indexation. In *Proceedings of the ICASSP*, volume 1.

I. Bazzi and J. Glass. 2002. A multi-class approach for modeling out-of-vocabulary words. In *Proceedings of the ICSLP*, pages 1613–1616.

N. Bertoldi and M. Federico. 2001. Lexicon adaptation for broadcast news transcription. In *Proceedings of ISCA ITRW workshop on AMSR*, pages 187–190.

G. Gravier, J.F. Bonastre, S. Galliano, E. Geoffrois, K. Mc Tait, and K. Choukri. 2004. The ESTER evaluation campaign of rich transcription of french broadcast news. In *Proceedings of Language Resources and Evaluation Conference*.

T.J. Hazen and I. Bazzi. 2001. A comparison and combination of methods for oov word detection and word confidence scoring. *Proceedings of the ICASSP*, 1:397–400.

Y. Kajiura, M. Suzuki, A. Ito, and S. Makino. 2006. Generating search query in unsupervised language model adaptaion using www. *The Journal of the Acoustical Society of America*, 120(5):3043–3044.

M. Lapata and F. Keller. 2005. Web-based models for natural language processing. *ACM Trans. Speech Lang. Process.*, 2(1):3.

G.A. Monroe, J.C. French, and A.L. Powell. 2002. Obtaining language models of web collections using query-based sampling techniques. In *Proceedings of the 35th Annual Hawaii International Conference on*, pages 1241–1247.

P. Nocera, C. Fredouille, G. Linares, D. Matrouf, S. Meignier, JF Bonastre, D. Massonié, and F. Béchet. 2004. The LIA's French Broadcast News Transcription System. In *SWIM: Lectures by Masters in Speech Processing*.

K. Ohtsuki, N. Hiroshima, M. Oku, and A. Imamura. 2005. Unsupervised vocabulary expansion for automatic transcription of broadcast news. In *Proceedings of the ICASSP*, pages 1021–1024.

K. Tanigaki, H. Yamamoto, and Y. Sagisaka. 2000. A hierarchical language model incorporating class-dependent word models for oov words recognition. *Proceedings of the ICSLP*, 3:123–126.

A. Yazgan and M. Saraclar. 2004. Hybrid language models for out of vocabulary word detection in large vocabulary conversational speech recognition. *Proceedings of the ICASSP*, 1.