ON-DEMAND NEW WORD LEARNING USING WORLD WIDE WEB

Stanislas Oger, Georges Linarès, Frédéric Béchet, Pascal Nocera

LIA - University of Avignon, BP1228 84911 Avignon Cedex 09 - France
{*stanislas.oger,georges.linares,frederic.bechet,pascal.nocera*}@*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. Our experiments confirm the relevance of the Web for the OOV word retrieval. Different methods are proposed to retrieve the hypothesis words. Finally we present the integration of new words in the transcription process based on part-of-speech models. This technique allows to recover 7.6% of the significant OOV words and the accuracy of the system is improved.

***Index Terms*—** Lexical modeling, Speech recognition, Information retrieval, Natural languages

1. INTRODUCTION

In spite of the huge amount of data used for language model training, the problem of out of vocabulary (OOV) words remains a key point in large vocabulary continuous speech recognition (LVCSR), especially on transcription of highly epoch-dependent documents.

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 automatic speech recognition (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 [1], 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 LVCSR. Generally, authors proposed to collect a large amount of documents that are supposed to be relatively close to the targeted linguistic and semantic context [2][3][4][5]. Unfortunately, web-data suffers from the lack of structuring information and large well-targeted corpora generally outperform web-based language models [6].

Focusing on the problem of lexicon building, OOV word retrieval methods tackle two main difficulties : how missing words could be automatically found on the web and how these new words could be integrated in the ASR system. These problems are traditionally addressed by capturing the global semantic features from the document and collecting relevant documents which are used for language modeling.

In this paper, we suggest that the local linguistic context might bring some characteristic information of the OOV words and we propose to use this information to retrieve the OOV words in the unlimited collection of web documents. 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, we evaluate the hypothesis that the web is relevant for the task of unknown word retrieval and we propose word retrieval strategies funded on word-template matching. These methods are both evaluated on exact transcripts and on ASR system outputs. The third section of the paper addresses the problem of integrating locally augmented lexicons in the speech recognition process. We propose a method based on part-of-speech (POS) models. Lastly, we conclude on the interest of web-based local approaches for posterior correcting of automatic transcriptions.

2. EXPERIMENTAL FRAMEWORK

Our general approach consists in correcting *a posteriori* automatic transcriptions produced by a LVCSR system. Here, we use the LIA broadcast news system, SPEERAL [7]. This system is an A * decoder based on state-dependent HMM for acoustic modeling. The language models are trigrams estimated on about 200M words from the French newspaper Le Monde and from the ESTER [8] broadcast news corpus (about 1M words). We use a lexicon made of the 65000 most frequent words of these corpora.

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

The experiments are carried out in the framework of ESTER evaluation campaign and the Google search engine is used to access Web data [1]. All the tests are performed on about 6 hours of French broadcast news from the test corpus of ESTER 2005. The transcription word error rate (WER) of this corpus with the previously described system is 24.5% after the first pass (without acoustic adaptation, including OOV words). There are 645 OOV words on the 62024 words of these 6 hours, the OOV 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. It is important to notice that named entities and technical words are critical for the sentence intelligibility and represent 97% of the OOV words.

3. 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 study the impact of automatic transcript on retrieval method performance.

3.1. 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 that contains all possible n-grams, comprising the ones which contain 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 transcripts. Requests like $w_{t-n-1}...w_{t-1}w_t$, where $w_{t-n-1}...w_{t-1}$ is the history of w_t, are submitted to Google. The compound OOV words, like compound nouns, first name/last name, *etc.* are merged as single OOV word. The results presented in table 1 show that 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	100.0	88.2	50.5	27.3	16.1

Table 1. Recall (in percentage) of *n*-grams containing OOV words retrieval on Google depending on the size *n*.

3.2. Collecting augmented lexicons

Here we evaluate the performance of various methods for OOV word retrieval using both the ASR system outputs and the exact transcripts. The segments containing OOV words are manually pointed out and we propose and compare some retrieval strategies.

3.2.1. N-gram strategy

The first technique consists in building requests by taking the *n*-gram containing w_t and substituting w_t by a wildcard character. This strategy was tested with the reference context and the ASR transcript context in order to measure the impact of the ASR errors on the retrieval performance. Results are reported in table 2 for the recall and 3 for the hypothesis word set size.

It is obvious that the recall rate with the ASR outputs is generally worse than with the reference context. The cause is that the ASR transcription contains errors around the OOV words because they perturb the ASR system.

Moreover, we observe that the use of short sequences leads to low discriminative request. 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 word retrieval. In the next section, we propose to relax word utterance sequentiality constraints by defining patterns rather than hard-fixed n-grams.

3.2.2. Pattern strategy

Here, we build soft requests by extracting word templates from the context. This is achieved by replacing the most frequent French words by wildcard characters which will be automatically substituted by one to five words by the search engine. For example, given the sentence *Cette région a été touchée par le tsunami du 26 décembre 2004* with *trunami* as OOV word, the corresponding pattern is *région * touchée (*) 26 décembre 2004* with *(*)* the place to collect OOV word candidates.

The results presented in tables 2 and 3 show that recall rates are better than the n-gram based strategy with both the ASR system output and exact transcript 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. In addition, the recall with small patterns is better than with large ones, indicating that this strategy allows the search engine to better rank the documents.

Moreover, we notice that the 2-gram recall decreases less than with the previous strategy when the ASR context is used, indicating a better robustness in this specific configuration.

3.2.3. 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

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

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

We can see in tables 2 and 3 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 one with the best configuration of the previous strategies, but the precision decreases a lot. This last point is a major drawback for augmented lexicon integration 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	14.0	4.7	20.0	7.3	32.6	18.5
3	18.1	5.1	20.3	5.0	39.7	27.8
4	16.4	2.3	17.5	2.0	45.9	35.2
5	13.8	1.9	12.3	1.2	50.2	40.9

Table 2. Recall (in percentage) of OOV words retrieval on the best 100 Google ranked documents depending on the size n with the reference and 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 3. Average hypothesis-set size of OOV words retrieval on the best 100 Google ranked documents depending on the size n with the reference context and ASR outputs.

3.2.4. *N-gram semantics driven based word retrieval*

As shown in the section 3.2.1, only 14% of the 2-grams containing the targeted word can be retrieved with the reference context whereas almost 88% are on the Web. We presume that all 2-grams containing the targeted word are in the documents returned by the search engine but not well ranked. We assume that adding relevant context words in the n-gram request strategy may help the search engine to better rank documents which are relevant for the context and, we hope, contain the targeted 2-gram. These additional words are called here drive-words (DW).

The DW are selected in a fixed-size window around the OOV word. A list of the words located in the window is built and sorted by decreasing language frequency and the top n words are selected as DW. These words are added to the request as keywords without ordering constraint. The hypothesis-sets are built by selecting all potential n-grams, like in the standard n-gram strategy. The results using the reference and ASR context are reported in table 4.

Results significantly outperform previous ones with the reference context (tables 2 and 3); we obtain good recall rates 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 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	24.0	268	8.7	292
2/2	26.1	789	8.1	306
2/3	27.0	1.3k	6.5	295
3/1	19.1	16	4.0	87
3/2	15.0	15	3.9	79
3/3	13.3	19	3.1	98

Table 4. Recall (in percentage) 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 reference and ASR context.

4. DECODING WITH AUGMENTED LEXICON

Here we evaluate the performance of the proposed methods in correcting ASR system outputs. This is achieved by performing, on each segment, a new decoding pass based on the segment-specific augmented lexicon.

4.1. Decoding with the unknown word model

We first studied a way to incorporate the sets of hypothesis words without modifying the language model. The competing words are integrated in the decoding lexicon as pronunciation variants of the unknown word. Phonetization of new words is automatically performed by LIA_PHON[2] [9]. Therefore, the probability of the unknown word is assigned to each hypothesis word. The results on the test corpus shows that 5 % of the total number of OOV words are correctly transcribed with the augmented lexicon (called recall in the table 5). 8.7% of the total number of OOV words was in the hypothesis-sets (see the table 4). The low precision indicates that many wrong new words are introduced, it increases a bit the WER which grew from 24.5% to 24.6% because the amount of introduced wrong words is greater than the amount of successfully recovered OOV words. However 71.9% of correctly introduced words are named entities (NE), 25.0% are technical words and

[2] http://www.lia.univ-avignon.fr/chercheurs/bechet/download_fred.html

	Recall	Precision	WER
baseline	0.0	0.0	24.5
unknown word	5.0	22.0	24.6
POS classes	6.1	55.1	24.3

Table 5. Recognizer precision, recall and WER (in percentage) with the unknown word and POS class methods for the augmented lexicon integration.

the other 3.1% are infrequent verb forms and common nouns. The WER increase is balanced by the number of significant words retrieved which increases the document intelligibility.

Moreover, these results indicate that when the OOV word is in the hypothesis-sets, the probability that decoding with the augmented lexicon allows to retrieve correctly the word is about 57%. These results are summarized in the table 5.

Nevertheless, the negative impact of this method on system accuracy suggests that the system should take advantage of a more precise linguistic modeling.

4.2. Decoding with the part-of-speech model

Some recent works study the use of morphosyntactic classes for capturing the linguistic profile of infrequent words [3]. We propose to estimate the probabilities of the retrieved words from their POS class. This method relies on the following n-gram approximation :

$$P(w_u|w_i, ..., w_{i-n}) \approx \alpha * P(POS_w|w_{i-1}, ..., w_{i-n})$$

where w_u is the word added to the lexicon and POS_w the POS class of w_u. α is a scale factor which reduces the POS class probability. The POS class of w_u is identified by using an HMM-based tagger[3]. The POS n-gram probabilities $P(POS_w|w_{i-1}, ..., w_{i-n})$ can be estimated directly on the training corpus. The α factor is obtained experimentally by testing several values on the development corpus.

Results are compared to the ones obtained by the previous method in table 5. We observe that the POS-based method provides a gain in terms of WER, absolute OOV word recall and decoding precision. The global WER decreases by 0.2% and 92.3% of correctly recovered OOV words are NE. Therefore, 7.7% of the total missing NE are successfully recovered with a good precision.

5. CONCLUSION AND PERSPECTIVES

We proposed a method for improving lexical coverage by using locally augmented lexicon. This method relies on a two-pass decoding strategy where the first pass is used to build Google requests. Augmented lexicons are built with the returned documents. We presented several strategies for request formulation. 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. Lastly, we propose to integrate POS-based n-gram probabilities in augmented language models which constitutes the second pass of the decoding. This method brings an absolute WER decrease of about 0.2% in comparison of the first pass. Moreover, most of the successfully recovered OOV words consist in named entities which are crucial for understanding tasks. Globally, our approach allows to recover 7.7% of significant OOV words while slightly increases the system accuracy.

6. REFERENCES

[1] K. Ohtsuki, N. Hiroshima, M. Oku, and A. Imamura, “Unsupervised vocabulary expansion for automatic transcription of broadcast news,” in *Proceedings of the ICASSP*, 2005, pp. 1021–1024.

[2] Y. Kajiura, M. Suzuki, A. Ito, and S. Makino, “Generating search query in unsupervised language model adaptaion using www,” *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 3043–3044, 2006.

[3] A. Allauzen and J. Gauvain, “Open Vocabulary ASR for Audiovisual Document Indexation,” in *Proceedings of the ICASSP*, 2005, vol. 1, pp. 1013–1016.

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

[5] N. Bertoldi and M. Federico, “Lexicon adaptation for broadcast news transcription,” in *Proceedings of ISCA ITRW workshop on AMSR*, 2001, pp. 187–190.

[6] M. Lapata and F. Keller, “Web-based models for natural language processing,” *ACM Trans. Speech Lang. Process.*, vol. 2, no. 1, pp. 1–30, 2005.

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

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

[9] F. Béchet, “LIA_PHON: Un système complet de phonétisation de textes,” in *Proceedings of Traitement Automatique des Langues*, 2001, vol. 42, pp. 47–67.

[3] http://www.lia.univ-avignon.fr/chercheurs/bechet/download_fred.html