New word learning for spoken document processing through discovery of comparable texts from external resources

Kuan-Ting Chen[1,3], Shui-Lung Chuang[1], Frank Seide[2], Hsin-Min Wang[1], Lee-Feng Chien[1], and Eric Chang[2]

[1]Institute of Information Science, Academia Sinica, Taipei.
[2]Microsoft Research Asia, Beijing.
[3]Graduate Institute of Communication Engineering, National Taiwan University, Taipei.
Email: {kenneth,slchuang,whm,lfchien}@iis.sinica.edu.tw, {fseide,echang}@microsoft.com

ABSTRACT

This paper presents a new out-of-vocabulary (OOV) word learning approach that dynamically extends the pronunciation lexicon and the language model for large vocabulary continuous speech recognition (LVCSR) in spoken document retrieval (SDR) systems. Based on the assumption that the graphemes as well as the *n*-gram statistics of the OOV words can be effectively learned from other contemporary or in-domain text documents, the proposed approach suggests an iterative procedure of dynamic unsupervised new word learning, which makes uses of the relevant text documents (termed *comparable texts*) retrieved from the external resource, such as special-domain text databases or the Internet, as the lexicon/language model (LM) adaptation data. The preliminary experiments were conducted on Hub-4 '96 English broadcast news development set (F0 condition only), using TREC-2001 WebTrack data (WT10g) as the external resource. The results showed that, when neither any key term selection nor new word extraction/filtering techniques were applied, the proposed framework significantly reduced the OOV rates of various artificially created lexicons, from OOV rates 2.64%, 5.18%, 10.66%, to 1.83%, 2.93%, 4.58%, respectively.

Keywords: speech recognition, information retrieval, spoken document retrieval, out-of-vocabulary words.

1. INTRODUCTION

Spoken document retrieval (SDR) systems have attracted extensive research interests in the recent years [1-3]. Recently, real-world applications of such systems have also started to appear, such as SpeechBot, a web-based audio indexing system developed by Compaq [4]. In general, most of the SDR systems seek to apply speech recognition techniques generating either word-level or subword-level (e.g. phoneme) transcriptions for a large amount of spoken documents, and then use these transcriptions to index the spoken documents for retrieval purposes. It is clear that in such scenarios, speech recognition accuracy is critical to the retrieval performance.

The problem caused by out-of-vocabulary (OOV) words is one of the key issues in LVCSR systems and in SDR systems as well. It is well known that for certain languages, when spoken documents contain words that are not in the speech recognition dictionary (OOV words), there will be no chance for these words to be correctly recognized, since not only are the appropriate word-pronunciation mappings missed, but the language model (LM) also doesn't contain any information about these OOV words. Such problems will inevitably hurt speech recognition accuracy and lead to certain degradation of spoken document retrieval performance. The situation becomes even worse when the OOV words happen to be the keywords that constitute users' queries, which are actually very common cases in SDR applications. Moreover, when it comes to spoken document presentation, incorrect recognition caused by the OOV words will result in possibly a sequence of wrong words displayed, which is surely unsatisfactory to the users.

Conventionally, the pronunciation dictionary and the statistical language models for an LVCSR system are learned from a certain *external resource*, namely, *the training corpus*. The problem of OOV words somehow implies the deficiency of the training corpus with respect to the test data. For example, spoken documents such as medical lectures may contain a lot of proper nouns corresponding to some medicine names or syndromes, which are of little chance to be included in the general-domain lexicons for most LVCSR systems. Under such an assumption, a reasonable solution may be to better utilize other external resources in an efficient way; i.e., to learn the OOV words from the *comparable texts* (relevant text documents) that are more similar to the test data in the sense of topic or style. Such comparable texts can be retrieved and collected on-demand from various external text data sources, such as the Internet or certain special-domain text collections. For example, for the medical lecture transcription/ retrieval task, the database of medical literature may be utilized. It is possible to identify the topical characteristic of the test data based on the initial recognition output generated according to the

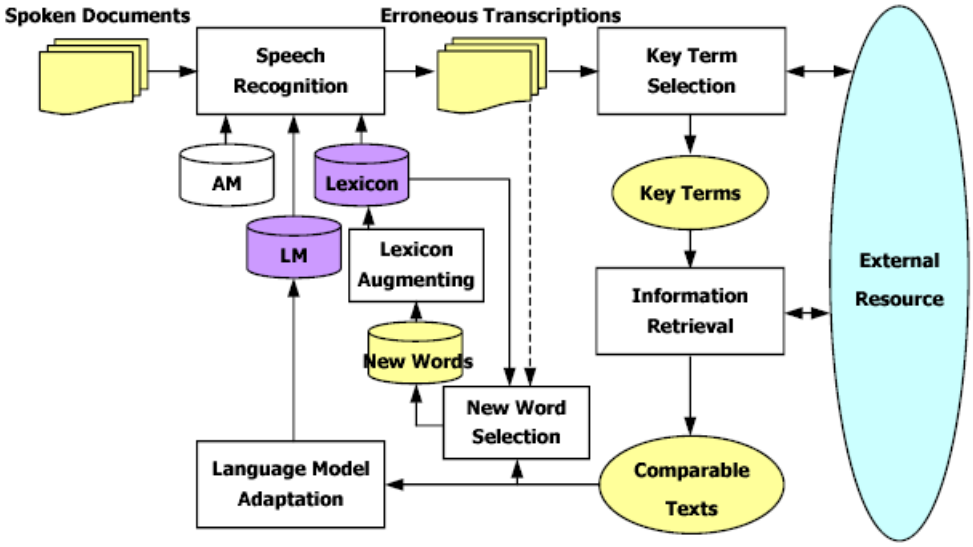

Figure 1: The proposed OOV word learning and language model adaptation framework.

baseline lexicon/LM, and then to collect adaptation data for lexicon/LM augmentation. In other words, information retrieval (IR) techniques may be helpful in solving the OOV problem.

Several studies have been conducted along the line of applying IR techniques to improve the LVCSR performance. Mahajan *et al* suggested improving topic-dependent language modeling by explicitly applying IR techniques to dynamically identify the topic according to the contextual development [5]. Chen *et al* utilized various IR methods to pre-cluster the language model training texts for topic-adaptive language model building, as well as to retrieve in-topic adaptation data on-demand through measuring the similarity among the test data and the training text clusters [6]. Though having proved the effectiveness of applying IR techniques to improve language modeling, none of the above approaches makes use of the external resources other than the training data. On solving the OOV problem, Bertoldi and Federico proposed a lexicon adaptation scheme for broadcast news transcription, which dynamically extends the lexicon by selecting new words day-by-day from contemporary news available on the Internet, according to a strategy that tries to minimize the OOV rate of the language model [7]. In their approach, pronunciations for the new words are automatically generated, and the OOV word class of the trigram language model is linked to a new unigram LM updated on a daily basis. Such an idea is novel, but the nature of learning on a per period basis somehow makes the approach limited to broadcast news applications.

Our idea is similar to that of Bertoldi and Federico [7], in the sense that the external resource is explicitly utilized as the basis for new word learning and language model adaptation. However, instead of fetching the text resources on a per period basis, we proposed to retrieve the relevant text documents *on-demand*. Some IR techniques are applied to identify a set of informative terms from the initial recognition outputs of the spoken documents. Using these terms to retrieve the adaptation data from the external resources makes the framework capable of on-demand as well as in-topic new word learning.

2. NEW WORD LEARNING FROM EXTERNAL RESOURCES

The proposed framework is depicted in Figure 1. The framework is applied on a per topical unit basis. For each topical unit in the spoken document collection, the baseline pronunciation lexicon and language model are first used to generate the initial speech recognition outputs. The resulting transcriptions will inevitably contain recognition errors, and the OOV words in the spoken documents will have no chance to be transcribed correctly. Certain key term selection scheme is then applied to extract a set of informative terms from the erroneous transcriptions so as to effectively identify the characteristics of the spoken document, e.g. topic, style, …etc. The resulting key terms are then used as queries to retrieve the comparable texts from the external resource, which can be the Internet or other special domain text sources available, such as technical literature databases. Graphemes of possible OOV word candidates are extracted and learned from the comparable texts retrieved. A grapheme-to-phoneme transcription tool is then used to automatically generate the pronunciations for the OOV words, which are finally augmented to the initial lexicon. On the other hand, *n*-gram statistics for the learned OOV words can be computed according to the comparable texts retrieved, so as to adapt the baseline language model and to further

improve speech recognition accuracy for these OOV words. Such an OOV word learning procedure can be iteratively carried out, so that the improved spoken document transcriptions can lead to better quality of the comparable texts and, as a result, better improved lexicon and language modeling.

There are several important components in the framework mentioned above. Firstly, since the speech recognition outputs are word-level sequences, a term generation approach should be applied to automatically generate term-level information from them. Secondly, given the fact that the recognition outputs may be severely erroneous, a robust key term selection method is required so as to extract only a few informative and representative terms, so that later on the information retrieval component can return a set of high quality comparable texts for use. This is especially important when the external resources are general-domain text collections in nature, such as the Internet. In such cases, inaccurate topical characterization of the erroneous recognition outputs, no better than query-by-document, will possibly result in topically scattered texts retrieved, which may break the proposed framework and even degrade the recognition accuracy iteratively. Thirdly, since the comparable texts may contain a huge number of new words unseen in the initial lexicon, a robust new word selection mechanism is required to filter out the irrelevant new word candidates so as to control the size of the new lexicon. It is of the same importance for the following language model adaptation by limiting the size of the resulting language model. It also makes the resulting language model not to include too many noises. It is suggested that the phoneme-level transcriptions of the spoken documents may be useful to new word selection.

The appropriate solutions to all the key components mentioned above actually depend on the nature of the OOV problem, the characteristics of the external resources, and the possible quality of the retrieved texts. In order to gain some idea, some preliminary experiments were first conducted to reveal the possible problems, and the more specific solutions are being investigated.

3. EXPERIMENTAL DESIGN AND SETUP

3.1 Task design and database selection

In order to investigate the effectiveness of the proposed framework, and also to see the impacts of both varying the level of OOV rates and initial speech recognition accuracy on the performance, it is required to *simulate* the OOV problem by creating baseline lexicons from the LM training data and then artificially removing a certain number of lexical words, so as to achieve various predefined levels of OOV rates. The proposed method was then applied to reveal the OOV word *recovery* performance as well as the improvements of recognition accuracy and retrieval performance.

The most suitable application of the proposed framework would be to improve the recognition accuracy as well as the retrieval performance for the spoken documents which are very special in terms of *topics*, such as technical meetings or lectures, while using the general-domain dictionary and LM for the initial recognition. We believe that it would be very interesting to conduct experiments in such a scenario; to be more specific, taking certain technical lectures along with their manual transcriptions as the test set, and employing the corresponding knowledge base (such as collections of the presentation slides) or the Internet resource (through efficient Web search engine such as *Google*) as the external resource.

On the other hand, considering the reproducibility and benchmarking of the experiments, in the current study we chose the English broadcast news transcription task (1996 Hub-4) for word error rate (WER) evaluation, and the broadcast news retrieval task (1997 TREC-7 SDR Track) for SDR performance evaluation. As for the external resource, we used WT10g used in TREC-2001 WebTrack as a "frozen Internet". WT10g contains about 1.7 millions of pages and is a 10 gigabytes subset of a 1997 crawl of the WWW (Internet Archive).

As mentioned earlier, the key point of the experimental design is a reasonable OOV problem simulation method. At the beginning, we believed that the phenomenon of OOV words could be manipulated to some extent by utilizing the *temporal differences* among the LM training data, the WER/SDR test data, and the external resources. Careful considerations have thus been made to magnify the temporal differences when deciding the data set usage. Table 1 summarizes the temporal information of the selected data set combination. Although it was found later that merely utilizing the temporal differences is not enough to meet the demand of OOV problem simulation, such a task definition is nevertheless reasonable in the sense that it reflects well the situations of possible practical applications of the proposed approach.

3.2 Baseline lexicon generation

The 1996 CSR Hub-4 language model corpus was used as the basis for baseline lexicon generation and language model training. The corpus consists of approximately 130 million words of loosely transcribed broadcast shows. In addition, we further used the acoustic model training material, which consists of 380,000 words of manual transcriptions to augment the lexicon/LM corpus. We chose to merge the two sets of training texts, while multiplying the word counts of the latter set by 10, to make both sets more comparable in size.

Usage		Data set	Time period
Training	Acoustic model training	Hub-4 '96 Training set	96/05/10 – 96/07/07
	Lexicon determination Language model training	96CSR Hub-4 LM training data	92/01 – 96/06
Testing	WER evaluation	Hub-4 '96 Developing set	96/07
	SDR evaluation	Hub-4 '97 Training set (*TREC-7 SDR track*)	97/06 – 98/02
External resource		WT10g (*TREC-2001 WebTrack data*)	– 97 Internet Archive

Table 1: Data set information and usage.

Setup (% OOV rate)	#stems remained	#words in the lex	% OOV rate of the test set					
			DEV96(F0)	DEV96(F1)	EVAL96(F0)	EVAL96(F1)	TREC-7	96 CSR LM test
10.0	2,500	**15,234**	**10.66**	7.09	12.21	6.06	**9.85**	8.93
5.0	5,001	**26,536**	**5.18**	4.17	7.10	3.78	**4.82**	4.47
2.5	9,504	**42,817**	**2.64**	2.88	4.40	2.42	**2.49**	2.14
0.0	54,322	125,675	0.79	1.70	1.33	1.33	0.45	0.34
		127,625	**0.00**	0.00	0.00	0.00	**0.00**	0.00

Table 2: Details of the finalized lexicons.

Test set	LM Training (inside)	DEV96_F0	DEV96_F1	EVAL96_F0	EVAL96_F1
Perplexity	50.20	**294.64**	172.27	318.98	159.62

Table 3: Perplexity test results

In order to effectively generate several baseline lexicons with respect to various predefined levels of OOV rates (0.0%, 2.5%, 5.0%, 10.0%), we decided to apply frequency-based cutting off on the word list. Moreover, we decided to carry out unigram count calculation and list cutting off at the *stem* level of the words. The reason for doing so is explained as follows. The information retrieval engine applied in our experiments constantly conducts word stemming prior to indexing/retrieval, which is actually, to some extent, a standard procedure for most of the IR systems. Therefore, since one of the goals of experimental design is to see how well the proposed framework improves SDR performance, it is reasonable to require that for a specific word w_i to be defined as the OOV word, all other words that have the identical stem to that of w_i (denoted as W_i^C) should also be dropped out of the lexicon. Otherwise, the SDR performance may undesirably benefit from the remaining lexical words W_i^C, even when the spoken word w_i is incorrectly recognized.

The baseline lexicons were created as follows. A complete list of words appearing in the normalized LM training texts was first generated, and the unigram counts of these words were calculated. Then, a word stemming technique was applied to find the stem for each word, and the unigram counts of all the words having identical stems were merged, to form a frequency-ordered list of the stems. Next, a predefined threshold value of the stem counts was applied to cut those infrequent stems off. The words corresponding to the remaining stems constituted the finalized baseline lexicon, and an automatic grapheme-to-phoneme transcription tool was used to generate the necessary pronunciations. The procedure was repeated to construct four lexicons corresponding to the desired OOV rates (0.0%, 2.5%, 5.0%, 10.0%), respectively, by setting different cut-off thresholds. As for the lexicon corresponding to 0.0% OOV rate, a fairly large word list (125,675 lexical words) was first picked, then all the unseen words in the test sets were manually added to the word list. Table 2 summarizes the statistics of the finalized lexicons. The OOV rate measurements made on the sets other than the WER and SDR test sets were for reference purposes. However, for the construction of the 0.0% OOV rate lexicon, the unseen words appearing in *all* of the mentioned sets were added.

3.3 Baseline language model training

Baseline trigram language model was trained using the training texts described in the previous subsection with the 0.0% OOV rate lexicon (127,625 words) as the training vocabulary. The LM training process applied modified absolute discounting and an entropy-based pruning technique. Table 3 summarizes the results of perplexity test with respect to various test sets.

4. PRELIMINARY EXPERIMENTAL RESULTS

Preliminary experiments were conducted to investigate the OOV word recovery rates of the proposed framework *without* applying any key term selection method and new word extraction process. From the baseline lexicon of each OOV rate level, initial erroneous recognition outputs for each *section*[1] of the Hub-4 '96 dev set (F0 condition) were taken directly as the query to retrieve a number of comparable texts from the WT10g data. According to the retrieved text documents, all the new words unseen in the baseline lexicon were blindly added to form a new section-dependent lexicon. We investigated the resulting lexicons section-by-section to see how well the OOV words could be recovered, as well as to see how large the finalized lexicon would grow. Table 4 summarizes the results with the number of retrieved documents for each section set at 50 and 100.

In another series of experiments, the manual transcriptions of each section were taken directly as the query for comparable text retrieval, except that the OOV words in the transcriptions were manually "hidden" by a special tagging. For example, if the word "CLINTON" which appears in a section was not in the lexicon, it was manually tagged as "oovCLINTONvoo", which is supposed to be a non-existing word. The purpose is to simulate an optimistic situation that only OOV words themselves are incorrectly recognized, while no other recognition errors exist, including those due to the lack of sufficient *n*-gram statistics regarding these OOV words. In other words, in such a simulation, the word error rate is equal to the frequency-weighted OOV rate. After comparable text retrieval, the section-dependent lexicons were formed according to the same procedure as above, and the OOV rate improvements were investigated. Table 5 summarizes the results with the number of retrieved documents for each section set at 100.

It can be observed from Table 4 that, even without a key term selection scheme to better identify the topics of the spoken documents so as to improve the qualities of the comparable texts, a significant portion of the OOV words could have already been learned by blindly adding all new words to the lexicon. It is thus believed that the OOV recovery rate may be even better if an appropriate key term selection method is incorporated. On the other hand, compared to the number of the OOV words in the test set, the comparable texts actually contained an amazingly large number of unseen words, which implies the inefficiency of new word learning. In addition to key term selection, we believe that the problem may be further alleviated by incorporating a good new word extraction/filtering method. Finally, comparing Table 4 and 5 shows no major performance difference between using recognition outputs and using processed manual transcriptions for comparable text retrieval. We think this might be also due to the lack of an appropriate key term selection method, since query-by-document would somehow minimize the impact of recognition errors.

[1] In the 1996 Hub-4 data transcription convention, a *section* implies a topical unit (story, etc.)

5. CONCLUDING REMARKS AND FUTURE WORKS

At present stage of our work, the major achievements made are the experimental design and setup, as well as some preliminary experimental results that justified the effectiveness and validated the practicality of the proposed idea of learning from the external resource on-demand. The appropriate key term selection methods as well as the applicable new word extraction/filtering techniques are still being studied, while the idea was initiatively justified by applying the proposed framework with the above techniques skipped. Nevertheless, we believe that an appropriate key term selection method is necessary for efficient and precise topic identification under the presence of recognition errors, which is critical for the quality of the comparable texts retrieved. On the other hand, preliminary studies showed that the number of unseen words in the comparable texts can be amazingly huge, which implies the importance of a good OOV word candidate extraction/filtering method.

The results showed in this paper include only the OOV rate improvements, and we are working to see how the speech recognition would benefit from the new lexicon as well as the adapted language model. Finally, we will conduct similar experiments on the TREC-7 SDR task to further investigate the retrieval performance improvements by applying the proposed framework for OOV word learning and language model adaptation.

REFERENCES

[1] K. Spärck Jones, G. J. F. Jones, J. T. Foote and S. J. Young, "Experiments on Spoken Document Retrieval," in *Information Processing & Management*, 32(4), pp. 399-417, 1996.

[2] M. Wechsler, "Spoken Document Retrieval Based on Phoneme Recognition," Ph.D. thesis, Swiss Federal Institute of Technology (ETH), Zurich, 1998.

[3] B. Chen, H. M. Wang, and L. S. Lee, "Discriminating Capabilities of Syllable-based Features and Approaches of Utilizing Them for Voice Retrieval of Speech Information in Mandarin Chinese," in *IEEE Trans. on Speech and Audio Processing*, 10(5), pp. 303-314, July 2002.

[4] B. Logan, P. Moreno, J. M. Van Thong, and E.

Whittaker, "An Experimental Study of An Audio Indexing System for the Web," in *Proc. ICSLP2000*.

[5] M. Mahajan, D. Beeferman, and X. D. Huang, "Improved Topic-Dependent Language Modeling Using Information Retrieval Techniques," in *Proc. ICASSP1999*.

[6] L. Chen, J.-L. Gauvain, L. Lamel, G. Adda, and M. Adda, "Using Information Retrieval Methods for Language Model Adaptation," in *Proc. Eurospeech2001*.

[7] N. Bertoldi, and M. Federico, "Lexicon Adaptation for Broadcast News Transcription," in *Proc. ICSA Research And Technical Workshop on Adaptation Methods for Speech Recognition*, 2001.

OOV rate level	# lexical words	% OOV rate	% WER	# comparable text docs	% OOV rate resulted	% OOV rate improvement	# distinct OOV	# OOV recovered	# unseen words
0.0	127,625	**0.0**	**16.5**	–	–	–	–	–	–
2.5	42,817	**2.64**	**19.4**	50 100	**1.98** **1.83**	**25.00** **30.68**	3.4 3.4	0.8 1.0	9,762 16,099
5.0	26,536	**5.18**	**23.0**	50 100	**3.43** **2.93**	**33.78** **43.44**	7.1 7.1	2.4 3.1	13,183 20,860
10.0	15,234	**10.66**	**31.3**	50 100	**5.39** **4.58**	**49.44** **57.04**	14.6 14.6	7.2 8.3	17,841 27,028

Table 4: Preliminary experimental results using initial recognition outputs for comparable text retrieval. The OOV rates are frequency-weighted and based on all words in the test set. The numbers in the last 3 columns are document-averaged.

OOV rate level	# lexical words	% OOV rate	% WER	# comparable text docs	% OOV rate resulted	% OOV recovery rate
0.0	127,625	**0.0**	**0.0**	–	–	–
2.5	42,817	**2.64**	**2.64**	100	**1.94**	**26.52**
5.0	26,536	**5.18**	**5.18**	100	**3.02**	**41.70**
10.0	15,234	**10.66**	**10.66**	100	**4.42**	**58.54**

Table 5: Preliminary experimental results using tagged manual transcriptions for comparable text retrieval. The rates are frequency-weighted and based on all words in the test set.