Online Vocabulary Adaptation using Limited Adaptation Data

C. E. Liu, K. Thambiratnam, F. Seide

Microsoft Research Asia, 5F Beijing Sigma Center,
No. 49 Zhi Chun Rd., 100080 Beijing, P.R.C.
f-celiu@msrchina.research.microsoft.com, [kit,fseide]@microsoft.com

Abstract

This paper presents a study of low-latency domain-independent online vocabulary adaptation using limited amounts of supporting text data. The target applications include blind indexing of Internet content, indexing of new content with low latency, and domains where Out-Of-Vocabulary (OOV) words are problematic. A number of methods to perform document-specific adaptation using a small amount of support metadata and the Internet are examined. It is shown that a combination of word feature fusion and cross-file statistics pooling provides robust adaptation. The best evaluated method achieved an absolute reduction of 27.6% in OOV detection false alarm rate over the baseline word feature thresholding methods.

Index Terms: Vocabulary Adaptation, Vocabulary Selection, Out-Of-Vocabulary Detection, Spoken Document Retrieval

1. Introduction

As the popularity of Internet video and audio continues to grow, so to does the need for more robust methods for indexing and retrieving such content. The majority of websites use textual sources such as content metadata and user-generated tags to facilitate browsing and retrieval of multimedia content. However, the ability to search *within* content continues to be a desired feature, particularly for long content such as lectures and news.

Spoken Document Retrieval (SDR) techniques seek to index the spoken content of multimedia documents to enable more convenient browsing and retrieval. Word-level indexing is the most common approach, and uses an Automatic Speech Recognizer (ASR) to generate word transcriptions or lattices that are then used for indexing. Unfortunately, such systems are susceptible to Out-Of-Vocabulary (OOV) query issues since the vocabulary of most ASR systems is fixed. Phonetic and sub-word indexing systems (eg. [1, 2]) provide unconstrained vocabulary indexing but are plagued by poor performance for massive databases. Thus, word-level indexing systems require considerable effort to select and/or adapt vocabulary and language models to minimize the OOV rate. Typically, this can be done using offline adaptation techniques for domain specific indexing such as news reports or scientific lectures.

However, there are many domains for which careful tuning of language models and vocabularies is not possible. For example, the diversity of content for the Internet domain prevents the use of targeted vocabulary and language models. The broadcast news domain is also problematic due to regular topic changes and frequent introduction of new terms.

Therefore, research has focused on developing online vocabulary and language model adaptation algorithms for SDR systems. For news content, a common approach is to use text news reports from preceding days as adaptation data for future content (eg. [3, 4, 5, 6]). However, this introduces a considerable amount of indexing latency since sufficient adaptation data must first be accumulated. Additionally, such an approach is difficult to use in non-news domains.

This paper examines online vocabulary adaptation methods for a single spoken document using a small amount of supporting text data. This text data would typically be sourced from metadata or from related text documents (for example, the surrounding text for a web page containing an audio file, or the abstract for an audio lecture). Vocabulary and language model adaptation can then performed in a *document-specific* fashion by sourcing adaptation text documents from the Internet. The support text data is employed as a query in an Internet search. The intention of such an approach is to enable low latency indexing and low dependency on external adaptation data sources. This is is particularly desirable for Internet indexing systems that perform blind crawling and indexing of spoken content.

The study focuses on vocabulary adaptation, and thus does not consider any subsequent language model adaptation. This intentional choice was driven by preliminary experiments that showed traditional online vocabulary adaptation methods suffer from large vocabulary growth. Typically, the majority of these words are irrelevant when adapting to a single audio document. Therefore, this research focuses on methods to better constrain vocabulary growth while preserving adaptation performance.

A number of methods are examined in this work, and the results of vocabulary adaptation on a technical lecture corpus are reported. The methods evaluated include a selection of standard word feature thresholding approaches, such as Term Frequency Inverses Document Frequency (TFIDF) thresholding. Additionally, a feature fusion approach is proposed that exploits word feature fusion for OOV detection. Finally, the use of pooled statistic estimation from multiple documents is examined.

This paper is organized as follows. Section 2 provides a brief introduction to single-document online vocabulary adaptation. This is followed by a description in Section 3 of the adaptation methods explored in this study. Section 4 provides results for adaptation experiments on a technical lecture corpus. Conclusions and future work are discussed in Section 5.

2. Online Vocabulary Adaptation

A typical system for single-document online vocabulary and language model adaptation is shown in Figure 1. The functionality of each block is as follows:

Source Audio File: This is the audio document that is to be indexed. A one-off online vocabulary adaptation will be performed for this document.

Support Metadata: This is the supplementary text for the source audio file that will be used as the starting point for vo-

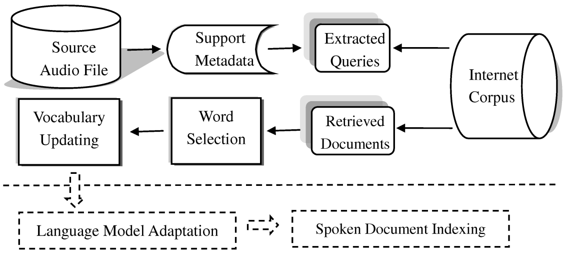

Figure 1: *A system for single-document online vocabulary and language model adaptation*

cabulary adaptation. For example, for an audio lecture, author, topic and description of lecture contents could be used. For an Internet page, this could be the surrounding text on the web page in which the audio file is embedded.

Extracted Queries: This is a set of queries that are automatically extracted from the metadata. Each query is used to retrieve related text documents using a search engine. Robust extraction of query terms is important to ensure that only relevant documents are retrieved. The approach used in this paper is to first run a Part-Of-Speech (POS) tagger over the metadata and then use the noun phrases as individual queries. Boolean combination of query terms was explored but is not reported in this work.

Internet Corpus: The Internet contains a broad range of content, and thus is ideal as a background corpus for vocabulary adaptation. Here, a search engine is used to retrieve documents that are relevant to an audio file.

Retrieved Documents: The retrieved documents are ranked according to relevance to the query terms. Preliminary experiments examined using various top-N documents for each query as adaptation data for word selection. It was expected that using a larger number of documents would give better OOV retrieval rates, but would be more susceptible to high vocabulary growth. Empirical experiments showed that using between 30 and 100 documents gave acceptable performance.

Word Selection: The retrieved document set typically contains a large number of unique words. Adding all unique words to the recognition vocabulary would thus result in a large adapted vocabulary. In turn, this would result in slower ASR speed, a poorly adapted language model, and almost certainly a much higher recognition word error rate. Thus, it is prudent to perform word selection to only include useful words in the adapted vocabulary. Careful word selection is paramount to make vocabulary adaptation actually improve SDR performance.

Vocabulary Updating: Once the appropriate list of words for adaptation is identified in the word selection stage, they can be added into the recognition vocabulary. Letter-to-sound rules may be applied to synthesize unknown pronunciations.

Language Model Adaptation: The language model must be adapted to incorporate the new words introduced by vocabulary adaptation. This is generally done by using the retrieved document set as adaptation data.

Spoken Document Indexing: Finally, indexing is performed using the adapted vocabulary and language model. The OOV and word error rates should be lower after adaptation.

3. Word Selection methods

Given a source audio document, A, the corresponding support metadata, $T(A)$, and an initial vocabulary set, $\mathbf{V}$, the task of word selection is to determine the set of auxiliary words, $\mathbf{V}^*$ that will reduce the OOV rate for ASR of audio document A while constraining the vocabulary growth $|\mathbf{V}^*|$. The adapted vocabulary set is then given by $\mathbf{V}' = \mathbf{V} \cup \mathbf{V}^*$.

First, a set of K query terms, $\mathbf{Q} = (q_1, q_2, \ldots, q_K)$ is automatically extracted using an appropriate query extraction technique. In this paper, a POS tagger is used to automatically select noun phrases from the metadata. Each noun phrase is then used as an independent query. An Internet search engine is queried to retrieve the top N relevant documents, $\mathbf{D^k} = \left(d_1^k, d_2^k, \ldots, d_N^k\right)$ for each query term q_k. Each document, d_j^k, is represented by the set of words it contains, $W(d_j^k) = (w_1, \ldots, w_M)$ and the corresponding word counts, $\mathbf{F}(d_j^k) = (f_1, \ldots, f_M)$. An appropriate word selection algorithm is then used to select the auxiliary word set, $\mathbf{V}^*$, for subsequent vocabulary updating.

A number of methods are explored in this paper. Detailed descriptions of these techniques provided below.

3.1. Null Background Corpus (NULL)

In this approach, only words from the support metadata, $T(A)$ are used. The auxiliary vocabulary, $\mathbf{V}^*$ is constructed by using all words in $T(A)$ that are not in the initial vocabulary, $\mathbf{V}$.

3.2. Greedy Selection (GREEDY)

In the GREEDY approach, all words in the retrieved document set, $\left(\mathbf{D^1}, \mathbf{D^2} \ldots \mathbf{D^K}\right)$, are included in the adapted vocabulary as follows: $\mathbf{V}^* = W(\mathbf{D^1}) \cup W(\mathbf{D^2}) \cup \ldots W(\mathbf{D^K})$

3.3. Word Feature Thresholding

A number of useful features can be extracted for each word in the retrieved document sets. Thresholding can then be used for word selection. The features examined in this paper are:

1. **WFREQ:** The word frequency, f_m, for each word, w_m
2. **TF:** The Term Frequency of w_m, $TF(w_m) = \frac{f_m}{\sum_i f_i}$
3. **TFIDF:** The Term Frequency times the Inverse Document Frequency (IDF). IDF is a global term importance weight and is given by $IDF(w_m) = \log \frac{N_D}{N_D(w_m)}$, where N_D is the number of documents in the dataset, and $N_D(w_m)$ is the number of documents that have at least 1 occurrence of w_m. Thus, $TFIDF(w_m) = TF(w_m) \times IDF(w_m)$.
4. **TTF:** The tapered term frequency of each word, given by $TTF(w_m) = \log\left(1 + TF(w_m)\right)$. Tapering reduces sensitivity to large word frequencies and is popular in traditional text-based Information Retrieval.
5. **TTFIDF:** This is the tapered equivalent of TFIDF: $TTFIDF(w_m) = TTF(w_m) \times IDF(w_m)$
6. **POS:** The coded part of speech tag for w_m.

3.4. Feature Fusion

The goal for vocabulary adaptation is to select all words in $W(\mathbf{D}^1) \cup W(\mathbf{D}^2) \cup \ldots W(\mathbf{D}^K)$ that are candidate OOVs for source document A, while excluding all other words. This is a classification task, and thus an OOV classification approach is proposed here for constructing the auxiliary vocabulary, $\mathbf{V}^*$.

A classifier is built as follows. A feature vector is first constructed for each word in each retrieved document, d_j^k, where features are sourced from the list of word features listed in Section 3.3. A classifier is then built that classifies a given feature vector as either a Candidate OOV word or a Non-Candidate OOV word. A word is considered a Candidate if it appears in the transcript of the source audio document, A and is therefore an OOV word that needs to be recovered. Thus, it is desired that all Candidate words are included in the auxiliary word set, $\mathbf{V}^*$, while all Non-Candidate words in $\mathbf{V}^*$ are considered as false alarms. Note that a Candidate word for one source document could possibly be a Non-Candidate word for another document, since the reference classification of candidate words is based on the reference transcript of a given source audio document.

A number of classifiers were explored in early experiments, including decision trees, neural networks, and support vector machines. It was found that neural networks provided a good trade-off between simplicity and generalization.

3.5. Pooled Frequency Statistics

The word feature thresholding and OOV classifier approaches described above operate on statistics derived from the retrieved document set $D(A) = \left(\mathbf{D^1} \ldots \mathbf{D^K}\right)$ for a given source audio file A. No information from retrieved document sets for other source audio documents is used in the estimation of statistics. However, in offline vocabulary adaptation, a large corpus is frequently used, which allows for more robust estimation of word frequency statistics, such as TF, and TFIDF. To study the effects of using a larger corpus, a pooled statistics estimation approach is proposed. For a given word, statistics are thus evaluated using the pooled document set $(D(A_1) \cup D(A_2) \cup \ldots)$. The statistics are retrieved for all source audio documents in the training or evaluation set $(A_1, A_2, \ldots)$. The intention is that using pooled statistics will result in more robust statistics estimation and thus more robust word selection.

It should be noted that the proposed Pooled approach is *not* suitable for single document online vocabulary adaptation. This is because pooling would impose a latency in indexing; a large number of source audio documents would first need to be accumulated before adaptation could be performed. Nevertheless, the method is included in this study to gain a better understanding of the ramifications of adaptation from a single document.

3.6. Pooled-P Frequency Statistics

An extension of the Pooled Frequency Statistics method is to only pool statistics from a small set of source documents. Thus, rather than using the retrieved document set for the entire training/evaluation set, it is proposed only subsets of this be used. Specifically, audio documents are formed into groups of P documents. Pooled estimation is then done only using the retrieved document sets from all other documents in the same group.

The motivation here is that low indexing latency can still be maintained, particularly if a low value of P is used. Vocabulary adaptation would be performed every P documents, thus maintaining a low indexing latency but preserving any performance gains from using pooled statistics.

4. Experiments and Results

A number of experiments were performed on the MITWorld technical lecture corpus to evaluate vocabulary adaptation performance. A technical lecture corpus was specifically used in this case as the OOV rate was considerably higher than other corpora such as broadcast news.

Part-of-Speech	# OOV	%
Noun	6916	93.6
Verb	140	1.9
Adjective	133	1.8
Other	200	2.7

Table 1: OOV Part-Of-Speech breakdown

Method	#C	#NC	#Final
ORACLE	103.1	0.0k	56.5k
NULL	0.2	0.06k	56.5k
GREEDY	71.9	111.6k	168.0k

Table 2: Baseline word selection results. #C/#NC words are the number of Candidate/Non-Candidate words in the adapted vocabulary. #Final is total size of the adapted vocabulary

4.1. Experiment setup

Experiments were performed on a subset of the MITWorld technical lecture corpus. The MITWorld corpus consists of a large number of recorded lecture videos with range of technical topics. Each lecture has an associated textual abstract and these abstracts were used as the support metadata for adaptation in the reported experiments. Training, development, and evaluation sets were constructed using 35, 15, and 20 hours of lectures.

The initial ASR vocabulary was constrained to 56.5k to maintain an OOV rate of 5%. There was limited overlap between the OOV sets of different lectures. A POS analysis of the lecture data is shown in Table 1. In total, there were 7,389 unique OOV words across all lectures. Among these, the majority of OOVs were nouns (93.6%). Thus, it was expected that the POS feature for individual words in the retrieved document sets would only be of limited benefit for OOV classification.

Vocabulary adaptation was performed using the procedure described in Section 2. First, each abstract was labeled with a POS tagger, and then all noun phrases were extracted as queries. A popular search engine was queried using these extracted queries, and the top 100 documents for each query were downloaded. Word features were then computed using either per-source-document estimation, Pooled estimation, or Pooled-5 estimation (Pooled-P estimation using 5 documents per group).

Word selection was performed using a classification framework. A score was generated for each word using an appropriate word selection method. False alarm rate (the number of Non-Candidate words in words in V^*) was then computed at various miss rates (the number of Candidate OOVs that were missed).

The feature fusion experiments used a three-layer neural network with 15 hidden neurons and two output nodes (Candidate and Non-Candidate). Mean and variance normalization was applied to all features. Additionally, class prior balancing was used as the number of Non-Candidate words in the training data was orders of magnitude greater than the number of Candidate words. The neural network was trained on a held-out training set and tuned on a development set. During evaluation, each word, w_m was classified using the neural network, and included in V^* if it was classified as a Candidate word.

4.2. Results

Baseline experiments were performed using the NULL and GREEDY methods as well as an ORACLE system. The results are shown in Table 2. It is clear that both the NULL and GREEDY methods yield unacceptable performance.

Feature	@ 10% Miss		@ 20% Miss	
	%FA	#NC	%FA	#NC
Single Feature Thresholding				
TF	60.5	67.5k	49.0	54.7k
TFIDF	63.6	71.0k	52.6	58.7k
TTF	60.5	67.5k	49.0	54.7k
TTFIDF	63.6	71.0k	52.7	58.8k
WFREQ	100.0	111.6k	100.0	111.6k
Feature Fusion				
TF+TFIDF+POS	43.4	48.4k	21.7	24.2k
TF+TFIDF+WFREQ	40.8	45.5k	21.9	24.4k
TF+TFIDF+TTF+TTFIDF	40.4	45.0k	22.8	25.4k

Table 3: Results for thresholding and feature fusion experiments. Miss/FA rates are relative to the GREEDY experiment. #NC is the number of erroneous Non-Candidate words.

Subsequent experiments were performed to evaluate the thresholded word feature methods and the results are shown in Table 3. Since the GREEDY method is equivalent to a thresholded system with an infinite threshold, the miss and false alarm performance for the GREEDY method is a hard limit for the thresholded methods (and all subsequent methods). Thus all reported miss and false alarm rates are relative to the $30.3\% = (103.1 - 71.9)/103.1$ miss rate and 111.6k Non-Candidate count of the GREEDY system.

Most notable was that WFREQ threshold was an ineffective feature. This is a result of the lack of document length normalization for this feature. Performance for the remaining features was still quite poor, with the best method, TF, having a #NC size of 54.7k at 20% miss rate.

The feature fusion methods achieved considerably better performance. A number of different feature combinations were evaluated. The best three methods are reported in Table 3. TF+TFIDF+POS achieved a #NC size of 24.2k at 20% miss rate while the TF+TFIDF+TTF+TTFIDF method yielded a #NC size of 45.0k at 10% miss rate. Fusion clearly provided a notable benefit over single-feature thresholding.

Finally, the experiments using the Pooled and Pooled-5 methods are reported in Table 4. The Pooled method improved performance at 10% miss rate but degraded performance at 20% miss rate. For example, the Pooled TF+TFIDF+TTF+TTFIDF method yielded a 4.2% absolute FA gain at 10% miss rate but increased FA by 1.3% at 20% miss rate.

In contrast, the Pooled-5 method fared considerably better, achieving similar gains to the pooled method at 10% miss rate, but in some cases gains at 20% miss rate also. In particular, the TF+TFIDF+TTF+TTFIDF+POS method achieved the best performance, with absolute gains of 3.4%/0.3% at 10/20% miss rate over the best fused feature false alarm rates.

The pooled experiments clearly demonstrate the benefits of pooled statistics. Particularly pleasing though is that only a relatively small degree of pooling (five documents per group) is required to improve performance over unpooled methods. It is suggested that the Pooled method did not perform as well due to a loss in the ability to perform *document-specific* adaptation since statistics were pooled across the entire database.

Overall, the experiments demonstrate that both feature fusion and statistics pooling are beneficial for single document online vocabulary adaptation. Feature fusion achieved dramatic improvements over unfused methods, while Pooled-5 pooling provided some smaller though notable gains in exchange for a small impact on indexing latency.

Feature	@ 10% Miss		@ 20% Miss	
	%FA	#NC	%FA	#NC
Pooled				
TF+TFIDF+...				
POS	37.3	41.6k	23.3	26.0k
WFREQ	38.7	43.2k	26.0	29.0k
TTF+TTFIDF	36.2	40.4k	24.1	26.9k
Pooled-5				
TF+TFIDF+...				
POS	36.8	41.1k	22.3	24.9k
WFREQ	39.5	44.1k	23.2	25.9k
TTF+TTFIDF	36.7	41.0k	24.1	26.9k
TTF+TTFIDF+POS	37.0	41.3k	21.4	23.9k
TTF+TTFIDF+WFREQ+POS	36.1	40.3k	22.3	24.9k

Table 4: Word selection with Pooled and Pooled-5 Frequency Statistics methods

5. Conclusions

This paper has examined the task of low latency and domain-independent single-document online vocabulary adaptation using a limited set of adaptation data. A number of word feature thresholding methods were evaluated on the MITWorld technical lecture corpus and were shown to yield poor word selection performance. However, the proposed feature fusion approach achieved a considerable improvement in word selection, with an absolute reduction of 27.3% in false alarm rate at 20% miss rate. Experiments using adaptation data pooling were also conducted and decribed. It was shown that large-scale pooling was only beneficial at low miss rates. However, small-scale pooling achieved false alarm improvements across a greater range of miss rates, with the best system yielding absolute false alarm gains of 3.4% at 10% miss rate. The best performance was achieved by combining feature fusion and small-scale pooling.

6. Acknowledgments

The authors would like to thank Jim Glass for providing access to the MITWorld lecture corpus.

7. References

[1] S. J. Young, M. G. Brown, J. T. Foote, G. J. F. Jones, and K. Sparck Jones, “Acoustic indexing for multimedia retrieval and browsing,” in *Acoustics, Speech and Signal Processing. Proceedings. International Conference on*, 1997.

[2] K. Thambiratnam and S. Sridharan, “Rapid yet accurate speech indexing using dynamic match lattice spotting,” *Audio, Speech and Language Processing, IEEE Transactions on*, 2007.

[3] S. Renals, D. Abberley, D. Kirby, and T. Robinson, “Indexing and retrieval of broadcast news,” *Speech Communication*, vol. 32, pp. 5–20, 2000.

[4] B. Bigi, Y. Huang, and R. De Mori, “Vocabulary and language model adaptation using information retrieval,” in *Spoken Language Processing, International Conference on*, 2004.

[5] A. Allauzen and J. L. Gauvain, “Diachronic vocabulary adaptation for broadcast news transcription,” in *Interspeech*, Lisbon, 2005.

[6] T. Kemp and A. Waibel, “Reducing the oov rate in broadcast news speech recognition,” in *Spoken Language Processing, International Conference on*, 1998.