Vocabulary and Language Model Adaptation using Information Retrieval

Brigitte Bigi, Yan Huang, Renato De Mori

ICSI
University of California, Berkeley
{brigitte,yan}@icsi.berkeley.edu

LIA-CNRS
Université d'Avignon
renato.demori@lia.univ-avignon.fr

Abstract

The goal of vocabulary optimization is to construct a vocabulary with exactly those words that are the most likely to appear in the test data. We will present a new approach to reduce the out-of-vocabulary (OOV) rate by adapting the vocabulary model during the ASR process. This method can also be used for the statistical language model (SLM) adaptation. An information retrieval system is used after the first pass of the ASR system to obtain a set of relevant documents. These documents are then used to generate the new vocabulary and/or corpus. In this paper, we propose a new retrieving method well-adapted for this purpose. Experiments were carried out on French with a 28% OOV rate reduction. Experiments were also carried out on English for the SLM adaptation, with 7.9% perplexity reduction, and minor WER improvement.

1. Introduction

Statistical Language Models (SLMs) have been successfully applied to a lot of problems, including automatic speech recognition (ASR), handwriting, automatic translation, etc. In ASR, SLMs, such as trigram, are used to provide adequate information to predict the probabilities of hypothetized word sequences. Enormous effort has been spent on building and improving language models; this effort follows two directions. The first one is to apply increasingly sophisticated estimation methods to a fixed training data set to achieve better estimation. The second one is to acquire more training data, because lack of training data will cause SLMs to be sub-optimal. However, automatically collecting and incorporating new training data is a non-trivial dynamic process.

Out-Of-Vocabulary (OOV) is a long existing problem coming from the fact that recognizers can recognize only a fixed vocabulary. [1] established that even in systems with a very large dictionary vocabulary (more than 100,000 words), the OOV rate can exceed 1%.

Lexical coverage of a vocabulary should be as high as possible to minimize out-of-vocabulary words. The general principles for vocabulary optimization are 1. it is inherently task-dependent, 2. the coverage is strongly affected by the amount of training data used, 3. source and recency of the training data is very important. Finally, the trade-off is that reducing the OOV rate increases the lexical coverage and so the acoustic confusability.

Fixing the task domain and matching training corpus is hard and fastidious. Actually, assuming an open domain, the task domain changes dynamically during the ASR process. By using a general vocabulary, OOV words therefore pose a problem to the ASR system, and the vocabulary of the speech recognizer should be as large as possible to ensure low OOV rates. The problem is that adding words in the vocabulary increases the acoustic confusions and does not always increase the ASR results.

This paper proposes an algorithm following 5 steps:

- use the ASR system to recognize the document,
- apply a retrieving system to obtain a set of documents related to the recognized sentences,
- learn a new vocabulary and train a new language model using the retrieved documents,
- combine this language model with the general one,
- use the new vocabulary and the interpolated language model in a ASR 2nd pass.

A new solution to retrieve documents which does not need an indexing phase is presented. This method can be used in *a dynamic database*, without normalizing the documents. This will also allow us to obtain recent vocabulary, etc. However, classical information retrieval methods are not well-adapted to deal with a dynamic database. Experiments show that this flexible retrieving method reduces the OOV words (French) and reduces the perplexity in the language model adaptation (English).

2. Retrieving documents using the Kullback-Leibler Distance

Kullback and Leiber in 1951 [2] introduced a measure of divergence between two probability distributions associated with the same experiment. Such a measure is also called cross entropy. Relative information depends on the order in which the probability distributions are considered. A symmetric version of the Kullback-Leibler divergence of probability distributions P, Q on a finite set

χ is defined as:

$$KLD(P \parallel Q) = \sum_{x \in \chi} \{P(x) - Q(x)\} \log \frac{P(x)}{Q(x)} \quad (1)$$

Besides being symmetric, KLD is zero between a distribution and itself, alway positive.

KL or KLD have been used in many natural language applications such as for query expansion [3]. They have also been used, for example, in natural language and speech processing applications based on statistical language modeling [4], and in information retrieval, for topic identification [5], for choosing among distributed collections [6]. Here, the idea is that documents to be considered as relevant are those which mostly contribute to the distance defined in the equation 1.

2.1. The probability distributions

Let p_t be an element of a series of documents to be analyzed and q_j a document of a database collections C. Let $P(p_t)$ and $Q(q_j)$ be their probability distributions on the finite vocabulary V. The term-probability distribution of a document is compared with each document probability distribution of the collection. A *back-off model* is proposed in which term frequencies appearing in the document are discounted and all the terms which are not in the document are given an ε-probability equal to the probability of unknown words. The reason is that in practice, often not all the terms in V appear in the document represented in q_j. Let $V(q_j) \subset V$ and $V(p_t) \subset V$ be the vocabulary of the terms which do appear in the document q_j and p_t respectively. For the terms not in $V(q_j)$, it is useful to introduce a back-off probability for $Q(w, q_j)$ when w does not occur in $V(q_j)$; otherwise the distance measure will be infinite. For the terms not in $V(p_t)$, it is also useful to introduce the **same** back-off probability for $P(w, p_t)$ when w does not occur in $V(p_t)$. The resulting definition of document probability $P(w, p_t)$ is:

$$P(w, p_t) = \begin{cases} \alpha \frac{f(w,p_t)}{\sum_{x \in p_t} f(x,p_t)} & if\ w \in p_t \\ \varepsilon & else \end{cases} \quad (2)$$

where $f(w, p_t)$ is the number of occurrences of w in p_t. The definition of a document probability $Q(w, q_j)$ is:

$$Q(w, q_j) = \begin{cases} \beta \frac{f(w,q_j)}{\sum_{x \in q_j} f(x,q_j)} & if\ w \in q_j \\ \varepsilon & else \end{cases} \quad (3)$$

where $f(w, q_j)$ is the number of occurrences of w in q_j.

α, β and the ε value have to be chosen in order that the corresponding probabilities sum to 1.

The ε-probability must be smaller than the minimum probability of a term in each document. Moreover, due to the fact that the documents in the database are not normalized, it is important that ε take into account the size of the document q_j. This leads to the following definition:

$$\varepsilon = \frac{1}{|V| \times |q_j|}$$

Equations 2 and 3 must respect the following property: $\sum_{w \in V} P(w, p_t) = 1$ and $\sum_{w \in V} Q(w, q_j) = 1$ leading to:

$$\sum_{w \in V(p_t)} \alpha \frac{f(w, p_t)}{\sum_{x \in p_t} f(x, p_t)} + \sum_{w \in V, w \notin V(p_t)} \varepsilon = 1$$

and

$$\sum_{w \in V(q_j)} \beta \frac{f(w, q_j)}{\sum_{x \in q_j} f(x, q_j)} + \sum_{w \in V, w \notin V(q_j)} \varepsilon = 1$$

These formule imply:

$$\alpha = 1 - ((|V| - |p_t|) \times \varepsilon)$$
$$\beta = 1 - ((|V| - |q_j|) \times \varepsilon)$$

2.2. Using KLD to retrieve documents

The method to retrieve documents q_j is based on the estimation of the Kullback-Leibler symmetric divergence $KLD(p_t \parallel q_j) =$

$$\sum_{w \in V} \{P(w, p_t) - Q(w, q_j)\} \log \frac{P(w, p_t)}{Q(w, q_j)} \quad (4)$$

This computation involves four cases:

1. $w \in p_t$ and $w \in q_j$, i.e. the term is in the two documents,
2. $w \in p_t$ and $w \notin q_j$, i.e. the term is in the reference and not in the document of the database,
3. $w \notin p_t$ and $w \in q_j$, i.e. the term is not in the reference but it is in the document of the database,
4. $w \notin p_t$ and $w \notin q_j$, i.e. the term is not in the two documents. In this case, the contribution to the distance is null. Consequently, this case is not taken into account in the KLD estimation, which means a faster estimation.

The documents q_j retrieved are those with the smallest $KLD(p_t, q_j)$ values.

3. Vocabulary adaptation in French

3.1. Corpus description

We carried out our experiments on the ESTER corpus. The ESTER[1] project aims to evaluate French broadcast news transcription systems and to establish a reference of the current performance levels of each system components. The ESTER corpus is made up of 3 data collections:

[1] http://www.afcp-parole.org/ester/

- 20h "France Inter" (7h-9h), December 1998
- 5h "France Inter" 19h broadcast news, May/June 1999
- 15h "RFI international" news and chronicle "Accents d'Europe" (9h30-10h30) or chronicle "Media d'Afrique" (11h30-12h30), April/May/September 2000.

This corpus is divided in 3 parts: training (30h40), development (4h40) and test (4h40). Moreover, the project includes a corpus of the French newspaper "Le Monde", from 1987 to 2002 (2.3 Gb of filtered corpus).

Pre-processing modules were built to enable the normalization of the documents to a version more suitable for language modeling purposes. They remove punctuation, expand numbers, do case processing for ignoring case distinction, etc.

3.2. Baseline vocabulary selection

Table 1 shows some OOV rates depending on the vocabulary V selection. We choose the vocabulary made of the 15,000 most frequent words in the newspaper Le Monde and all the words of the ESTER training corpus. After phonetized, the vocabulary is $V = 26121$.

Table 1: OOV depending on the vocabulary selection, on the developement corpus

	$\|V\|$	# OOV	% OOV
ESTER	23k	2193	4.78
Le Monde	10k	4952	10.79
Le Monde	15k	3561	7.76
Le Monde	20k	2864	6.24
Le Monde	40k	1769	3.85
ESTER + Le Monde 10k	24k	1846	4.02
ESTER + Le Monde 15k	26k	1550	*3.38*
ESTER + Le Monde 20k	29k	1321	2.88
ESTER + Le Monde 40k	45k	766	1.67

3.3. Experiments using Random Sampling

To estimate the KLD performance, we will compare its results with what can be obtained with a random retrieval system. All the different documents in the database have equal probabilities of being chosen. We randomly made 5 sets of 1000 documents and estimate the number of OOV by combining the resulting vocabulary and the ESTER vocabulary. The vocabulary sizes ranged between 43k and 45k with a OOV rate between 2.35% and 2.51%. This result can be compared with the last line of the table 1.

3.4. Experiments using KLD

Each document with more than 20 words of the development corpus is given as an input of the retrieving system. Then, we use the KLD method to retrieve the 1000 documents with the smaller distances (as equation 4). All words of the retrieved documents are used to learn the new vocabulary.

Table 2: OOV rate vs vocabulary selection, on documents more than 20 words of the development corpus

	$\|V\|$	% OOV
ESTER	23k	4.83
ESTER + 15,000 Le Monde	26k	3.42
ESTER + 1000 doc retrieved randomly	av. 44k	2.45
ESTER + 1000 doc retrieved with KLD	av. 26k	2.50

The results are presented in table 2. The first column indicates the method/corpus used to learn the vocabulary, the second column is the vocabulary size (av. means average when the size is variable).

It reduces significantly the OOV rate by using the KLD retrieved documents to learn a local vocabulary instead of the entire corpus to learn a general vocabulary. Compared to the vocabulary deduced from the ESTER training corpus, the OOV rate is reduced by 48%. Compared to the vocabulary from the ESTER training corpus and the newspaper corpus, the OOV rate is reduced by 28%. Moreover, this table shows that the random sampling method gives us the same OOV rate as KLD but the vocabulary is made of 44k words instead of 26k words by using the KLD method. This means that the random sampling method retrieves documents related to a large variety of different topics comparing to the KLD method which retrieves *relevant documents* related to a restricted number of topics. This experiment shows that using an information retrieval system like KLD is a good way to reduce the OOV rate.

4. SLM Adaptation in English

4.1. Corpus description

We did experiments on the English BNews task investigating how retrieving helps in language model adaptation. The training data consisted of 4 sources:

1. Hub4 SLM training corpus,
2. Hub4 acoustic model transcription,
3. the North American Buisiness News (NABN) corpus, and
4. the Switchboard-I corpus

A word-based 5-gram language model was estimated for each of the above 4 sources by the SRILM toolkit using the modified Kneser-Ney smoothing. Our baseline language model was constructed by interpolating these SLMs with weights optimized on some held-out data. The vocabulary size is 48k, constructed based on the frequency of words appearing in the taining data. The OOV rate is 1% on 1998 Hub4 test set, with only a quarter part related to real words and the others are word fragments. Consequently, it is less interesting to apply KLD for OOV problem in this case. However, we want to see how KLD can be used in language model adaptation. The basic idea is to shift the background language model to the relevant topic [7, 8, 5].

The retrieving experiments were carried out on Hub4 SLM training corpus only, which contains 130M words and 125k documents. It is tested on 1998 Hub4 test set.

4.2. Experiments using KLD

Each document of the true transcription of the test set is given as a query input to the retrieving system. We use the KLD method to retrieve the English BNews text corpus. Then the top 1000 documents ranked by distances (as equation 4), which is roughly 1% of the whole training corpus, are selected as the retrieved text corpus. A new topic language model is trained on the retrieved data, which is interpolated with the baseline language model.

The perplexity of the interpolated retrieved language model is 111.0, compared with 120.5 for the baseline language model (7.9% improvement). Moreover, we also did recognition experiments by using the subsystem of the SRI recognition system described in [9]. The baseline WER is 16.3%, and the new result obtained by the SLM interpolation between the baseline SLM and the retreived SLM is 16.0%.

KLD selects topic related documents from the training corpus, which boosts the topic homogeneous n-gram probability. This shows that KLD is efficient and robust in selecting a topic related pool from background data. The resulting interpolated model is a better estimator in the context of a topic related test set.

5. Conclusion and Perspectives

In this paper, we have validated the use of an information retrieval system in a 2-pass ASR process. By using this method to select a dynamic vocabulary, instead of a static vocabulary, the OOV rate is reduced by 28%, with the same vocabulary size. Moreover, the perplexity is significantly reduced by using this method to adapt the SLM. The WER reduction confirms this result.

The information retrieval method proposed in this paper uses the Kullback-Leibler distance and can deal with a dynamic database because it does not need any kind of pre-indexation. This will be particularly relevant for a database coming from the web, or for broadcast news, where the database can be updated daily. In future experiments, the use of KLD could be combined with the cache of the ASR system, and so integrated in a dynamic process. In this case, the content of the cache could be used, for example at the end of each sentence, to retrieve relevant documents of the history.

6. Acknowledgments

Thanks to CNRS, France.
Thanks Andreas Stolcke and Barbara Peskin for stimulating discussion and valuable suggestion. This work was funded in part by DARPA under contract No. MDA972-02-C-0038. Distribution is unlimited.

7. References

[1] I. L. Hetherington, "The problem of new, out-of-vocabulary words in spoken language systems," Ph.D. dissertation, Department of Electrical Engineering and Computer Science, Massachussets Institute of Technology, Cambridge, MA, 1994.

[2] S. Kullback and R. Leibler, "On information and sufficiency," vol. 22, pp. 79–86, 1951.

[3] C. Carpineto, R. De Mori, G. Romano, and B. Bigi, "An information theoretic approach to automatic query expansion," *ACM Transactions On Information Systems*, vol. 19, no. 1, pp. 1–27, 2001.

[4] I. Dagan, L. Lee, and F. Pereira, "Similarity-based models of word cooccurrence probabilities," *Machine Learning*, vol. 34, no. 5-9, pp. 43–69, 1999.

[5] B. Bigi, R. De Mori, M. El-Bèze, and T. Spriet, "A fuzzy decision strategy for topic identification and dynamic selection of language models," *Special Issue on Fuzzy Logic in Signal Processing, Signal Processing Journal*, vol. 80, no. 6, 2000.

[6] J. Xu and B. Croft, "Cluster-based language models for distributed retrieval," in *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Berkeley, CA, 1999, pp. 254–261.

[7] P.R. Clarkson and A.J. Robinson, "The applicability of language model adaptation for the broadcast news task," in *Proceedings 5th International Conference on Spoken Language Processing*, Sydney, Australia, 1998.

[8] R. Iyer and M. Ostendorf, "Relevance weighting for combining multi-domain data for n-gram language modeling," *Computer Speech and Language*, vol. 13, no. 3, pp. 267–282, 1999.

[9] A. Stolcke et al., "Speech-to-text research at sri-icsi-uw," in *DARPA Rich Transcription Workshop*, Vienna, VA, 2003.