Recognize Foreign Low-Frequency Words with Similar Pairs

Xi Ma[1], *Xiaoxi Wang*[1], *Dong Wang*[*1,2], *Zhiyong Zhang*[1]

[1]Center for Speech and Language Technology (CSLT),
Research Institute of Information Technology, Tsinghua University
[2]Tsinghua National Lab for Information Science and Technology

{mx,wxx,zhangzy}@cslt.riit.tsinghua.edu.cn
wangdong99@mails.tsinghua.edu.cn

Abstract

Low-frequency words place a major challenge for automatic speech recognition (ASR). The probabilities of these words, which are often important name entities, are generally underestimated by the language model (LM) due to their limited occurrences in the training data. Recently, we proposed a word-pair approach to deal with the problem, which borrows information of frequent words to enhance the probabilities of low-frequency words. This paper presents an extension to the word-pair method by involving multiple 'predicting words' to produce better estimation for low-frequency words. We also employ this approach to deal with out-of-language words in the task of multi-lingual speech recognition.

Index Terms: speech recognition, language model, multilingual

1. Introduction

The language model (LM) is an important module in automatic speech recognition (ASR). The most well-known language modelling approach is based upon word n-grams, which relies on statistics of n-gram counts to predict the probability of a word given its past n-1 words. In spite of the wide usage, the n-gram LM possesses an obvious limitation in estimating probabilities of words that are with low frequencies and the words that are absent in the training data. For low-frequency words, the probabilities tend to be under-estimated due to the lack of occurrences of their n-grams in the training data; for words that are absent in training, estimating the probabilities is simply impossible. Ironically, these words are often important entity names that should be emphasized in decoding, which means the probability under-estimation for them is a serious problem for ASR systems in practical usage.

A well-known approach to dealing with low-frequency and absent words is various smoothing techniques such as backoff [1] and discount [2, 3]. Another famous approach is to train an LM with some structures that can be dynamically changed, e.g., the class-based LM with classes that are adaptable online [4]. These dynamic structures, however, need to be pre-defined and can not handle words that are not in the structure. For example, words that are not in the pre-defined classes cannot be handled by class-based LMs. Additionally, involving such dynamic structures often requires to modify the decoder, which is not ideal to our opinion.

Recently, we proposed a similar-pair approach to deal with the problem [5]. The basic idea is to borrow some information from high-frequency words to be enhanced low-frequency words. More specifically, we seek for a high-frequency word that is similar to the word to enhance, and then re-weight the probability of the low-frequency word by adding a proportion of the probability of the high-frequency words to the probability of the low-frequency words. This approach has been implemented with the LM FST graph [6]. Compared to the traditional class-based LM approach, the new approach is flexible to enhance any words and does not need to change the decoder. It has been shown that this approach can provide significant performance gains for low-frequency words and words that are totally absent in the training data.

This paper is a following work of [5]. We first present an extension that allows multiple high-frequency words ('indicating words') to be used when enhancing a low-frequency word. This extension helps to involve multi-source information in the word enhancement, and is particularly important for words with multiple senses. Secondly, the similar-pair approach is applied to deal with a particular kind of low-frequency words: out-of-language (OOL) words that are from another language but embedded in utterances of the host language, for example English words appearing in Chinese utterances. These words are totally new for the host language and no context information can be employed to estimate the probabilities for them. The similar-pair approach can deal with the situation, by assuming that words in different languages share the same semantic space and hence similar pairs can be essentially across languages. The experimental results in Section 5 demonstrated the capability of this approach in dealing with OOL words.

The remainder of this paper is structured as follows. Section 2 discusses relevant works, and the similar-pair method is described in section 3. The two new extensions are presented in Section 4, which is followed by Section 5 where the experiments are presented. Finally, the entire paper is concluded by Section 6.

2. Related works

This work is related to dynamic language modeling that adds new words and re-weighting word probabilities, particulary the approaches that are based on FSTs. This section reviews some typical techniques of this approach, and primarily focuses on the class-based LM that deals with dynamic vocabularies and low-frequency words.

The class-based language modeling [7] is an approach that clusters similar words into classes and the probabilities of words in each class are re-distributed, for instance according to their unigram statistics. Typically, the class-based LM delivers better representations than the word-based LM for low-frequency words [4], since the class-based structure factorizes probabilities of low-frequency words into class probabilities and class member probabilities, and so increases robustness of the probability estimation. Moreover, new words can be easily added into classes with the class-based LM, leading to a dynamic vocabulary. Additionally, [8] and [9] introduced two techniques to build both the class-based LMs and the class words into FST graphs and embed class FSTs into the class-based LM FST. This embedding can be done on-the-fly, thus offering a flexible dynamic decoding that supports instant introduction of new words. Similar approaches have been proposed in [10, 11, 12], where various dynamic embedding methods are introduced, and the classes are extended to complex grammars.

The work is an extension of the similar-pair method proposed in [5]. In this approach, the probabilities of low-frequency words are enhanced and new words are supported by adding new FST transitions, both referring to the transitions of the similar and high-frequency words. Compared to the other approaches mentioned above, this method is more flexible, which supports any words instead of words limited in some pre-defined classes.

The extensions we made in this paper for the work in [5] are two-fold: firstly, the similar-pair algorithm is extended to allow multiple predicting words, which enables multiple information engaged; second, the similar-pair approach is employed to deal with OOL words, which demonstrated that similar pairs can be cross-lingual.

3. Word enhancement by similar-pairs

3.1. FST-based speech recognition

Most of current large-vocabulary speech recognition systems are based on statistical models, including hidden Markov models (HMMs), lexicons, decision trees and n-gram LMs. All these models can be converted into FSTs. For an FST, the correlation between the input and output symbols represent the mapping from a low-level sequence (e.g., phones) to a high-level sequence (e.g., words), and the weights encode the probability distribution of the mapping. More importantly, FSTs that represent different levels of statistical models can be composed together to form a unified mapping function that associates the primary inputs to high-level outputs. The composed FST can be further optimized by standard FST operations, including determinization, minimization and weight pushing. This produces very compact and efficient graphs that represent the knowledge of all the statistical models involved in the composition. In speech recognition, the composition can be used to produce a very efficient graph that maps HMM states to word sequences. The graph building process can be represented as follows:

$$HCLG = min(det(H \circ C \circ L \circ G)) \qquad (1)$$

where H, C, L and G represent the HMM, the decision tree, the lexicon and the LM (or grammar in grammar-based recognition) respectively, and $\circ$, 'det' and 'min' denote the FST operations of composition, determinization and minimization respectively.

3.2. Low-frequency word enhancement with similar pairs

The similar pairs method proposed in [5] is based on the FST architecture. In order to enhance low-frequency words, and for conducting the enhancement on the LM FST, or the G graph, a list of manually defined similar pairs are provided with corresponding frequency information obtained from training data. The low-frequency words are selected to be enhanced and the high-frequency words are chosen to provide the enhancement information. Each similar pair in the list includes a high-frequency word and a low-frequency words. Given a set of similar words, the low-frequency words are enhanced by looking at the information of the high-frequency word, including its transitions in the G graph and the associated weights. The high-frequency words are preserved since they have been well represented by the n-gram model already.

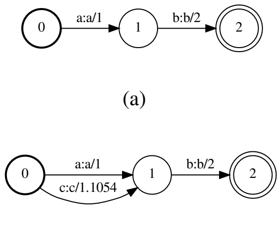
4. The Method

Based upon the similar pair method, the probability of low-frequency or new words are enhanced by looking at the information of high frequency words. Given a set of low-frequency words $W = \{x_1, x_2, ..., x_m\}$ to be enhanced, for each word $x_i \in W$, a set of words $S_i = \{y_{i,1}, y_{i,2}, ..., y_{i,n}\}$ that are similar to x_i are manually selected. The similarity can be defined in terms of either syntactic roles or semantic meanings. We assume that, for each $y_{i,j} \in S_i$, if there exists an n-gram of $y_{i,j}$ in the training corpora, the corresponding n-gram of x_i should also have a relative higher probability of appearance. As the probabilities are represented as the weights in the G graph in FST, according to this assumption, for any word x_i that to be enhanced, search all the appearances of a word $y_{i,j} \in S_i$ within the G FST. Let $A(y_{i,j})$ denote the set of all the transitions of the word $y_{i,j}$. Denote a particular transition by $(s, t, y_{i,j} : y_{i,j}/w_{y_{i,j}} \in A(y_{i,j})$, where s and t are the entering and existing states respectively, $w_{y_{i,j}}$ is the weight of this transition. Check if a transition $(s, t, x_i : x_i/w_{x_i})$ exists in G for x_i. If it exists, the wight w_{x_i} is adjusted to a new weight given by:

$$w_{x_i} = w_{y_{i,j}} + ln(\frac{f_{x_i}}{\sum\limits_{x_l \in W} f_{x_l}}) + \theta \qquad (2)$$

where θ is a parameter that tunes the enhancement scale and f_x represents the word frequency of x. Note that according to (2), a larger f_{x_i} leads to a higher w_{x_i}, which means that a more frequent word (still low-frequency) is assigned a larger weights after the enhancement, and so the rank of the low-frequency words in probabilities is preserved. If the transition $(s, t, x_i : x_i/w_{x_i})$ does not exist, then it is created and the weight is set to be $w_{y_{i,j}} + \theta$.

An example of the enhancement process is illustrated in Fig. 1, where Fig. 1(a) shows the G graph before the enhancement, and Fig. 1(b) shows the G graph after the enhancement. Note that 'a' is the high-frequency word, and (a,c) forms a similar pair. A new transition has been added in Fig. 1(b) for the low-frequency word c.



(a)

(b)

Figure 1: An example of low-frequency word enhancement based on similar pairs. (a,c) is a similar pair, where 'a' is the high-frequency word, and 'c' is a low-frequency word. A new transition is added in (b).

Compared to the original similar pair method proposed in [5], each low-frequency or new words x_i in the new method is enhanced by multiple high-frequency words (S_i) rather than only one word. This will add more transitions for the low-frequency word in the G graph, thus covering more desirable contexts.

Foreign word embedding is often observed in modern languages. For example, English words are often seen in Chinese sentences. These words are often name entities and are very rare in the training data. The word pair method provides a simple approach to enhancing these foreign words, by translating them into Chinese and use the Chinese words as referrals. Particularly, foreign words are often translated into multiple words in the host language, and so the multiple word pair method is more appropriate.

5. Experiment

The bilingual ASR tasks in the telecom domain is chosen to evaluate the proposed approach. We first introduce the experimental configurations, and then present the performance with the proposed low-frequency English words enhancement based on similar pairs.

5.1. Database

Our ASR task aims to transcribe conversations recorded from online service calls. The domain is the telecom service and the main language is Chinese, with some English word embedded. The acoustic model (AM) is trained on an 1400-hour online speech recording which is manually transcribed from a large call center service provider. The Chinese LM is trained on a corpus including the transcription of the AM training speech and some logs of web-based customer service systems in the domain of telecom service.

We selected 22 similar pairs to evaluate the performance of the similar-pair method. Each similar pair contains one low-frequency English and 1$\sim$5 high-frequency Chinese words as referrals. The task is to use the referral Chinese words to enhance the English words. A 'FOREIGN' test set was deliberately designed to test the enhancement with these similar pairs, which consists of 42 sentences from online speech recording. For each sentence, some English words that appear in the similar pairs are embedded among the Chinese words.

Additionally, a 'GENERAL' test set that involves 2608 utterances is selected to test the generalizability of the proposed method. Each utterance in this set contains words in various frequencies and therefore it can be used to examine if the proposed method impacts general performance of ASR systems at the time of enhancing low-frequency and new words.

5.2. Acoustic model training

The ASR system is based on the state-of-the-art HMM-DNN acoustic modeling approach, which represents the dynamic properties of speech signals using the hidden Markov model (HMM), and represents the state-dependent signal distribution by the deep neural network (DNN) model. The feature used is the 40-dimensional FBank power spectra. An 11-frame splice window is used to concatenate neighboring frames to capture long temporal dependency of speech signals. The linear discriminative analysis (LDA) is applied to reduce the dimension of the concatenated feature to 200.

The Kaldi toolkit [13] is used to train the HMM and DNN models. The training process largely follows the WSJ s5 GPU recipe published with Kaldi. Specifically, a pre-DNN system is first constructed based on Gaussian mixture models (GMM), and this system is then used to produce phone alignments of the training data. The alignments are employed to train the DNN-based system.

5.3. Language model training

The training text is normalized before training. The normalization includes removing unrecognized characters, unifying different encoding schemes and normalizing the spelling form of numbers and letters. Then the training text is segmented into word sequences. A word segmentation tool provided by Google is used in this study. The lexicon involves $150,000$ words in total. The SRILM toolkit [1] is then used to train a 3-gram LM with the Kneser-Ney discounting. The Kaldi toolkit is used to convert n-gram LMs to G graphs, and the openFST toolkit[2] is used to manipulate FSTs.

5.4. Experiment result and analysis

The ASR performance is evaluated in terms of the word error rate (WER). The results with the basic similar pair method (only one referral Chinese word) are presented in Table 1. Table 2 and Table 3 report detailed results where the number of referral Chinese words increases from 1 to 5. We report the results on two test sets: 'GENERAL' and 'FOREIGN', and the results with different values of the enhancement scale θ are presented.

It can be seen that with the similar-pair-based enhancement, the ASR performance on utterances embedded with English words is significantly improved. In addition, compared with the basic similar pair approach, the multiple similar pair approach delivers better results. Interestingly, the enhancement on the infrequent words does not cause degradation on other words, as shown by the results on the 'GENERAL' test set. This indicates that the proposed approach does not impact general performance of ASR systems, and thus is safe to employ. For a more clear presentation, the trends of WERs on the two test sets with different numbers of referral Chinese words are presented in Figure 2, where θ has fixed to -4.

		WER%	
	θ	GENERAL	FOREIGN
Baseline	-	33.75	77.64
+ SP	-4	33.76	66.95
	-2	33.76	64.39
	0	33.77	62.11
	2	33.8	62.96
	4	33.83	69.8

Table 1: WERs with and without the similar-pair-based enhancement. 'SP' stands for enhancement with similar pairs, which uses one referral Chinese word. θ is the enhancement scale in equation (2).

	WER%				
N θ	1	2	3	4	5
-4	33.76	33.77	33.77	33.78	33.78
-2	33.76	33.77	33.77	33.77	33.78
0	33.77	33.79	33.79	33.78	33.79
2	33.8	33.83	33.82	33.81	33.82
4	33.83	33.96	33.96	33.97	33.96

Table 2: WERs on the 'GENERAL' test set. N is the number of referral high-frequency Chinese words.

[1] http://www.speech.sri.com/projects/srilm/

[2] http://www.openfst.org

	WER%				
N θ	1	2	3	4	5
-4	66.95	62.68	60.4	60.68	61.82
-2	64.39	63.53	61.54	61.54	62.68
0	62.11	64.96	66.95	66.95	65.53
2	62.96	66.95	65.53	65.53	65.53
4	69.8	73.5	77.49	77.49	77.78

Table 3: WERs on the 'FOREIGN' test set. N is the number of referral high-frequency Chinese words.

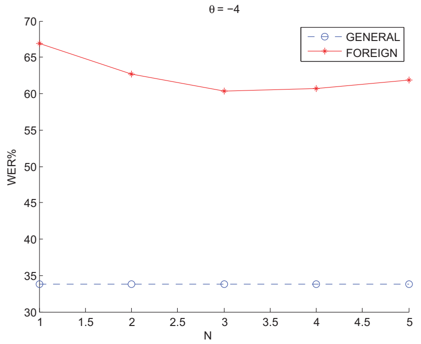


Figure 2: WERs on the two test sets. N is the number of referral high-frequency Chinese words. The value of θ is -4.

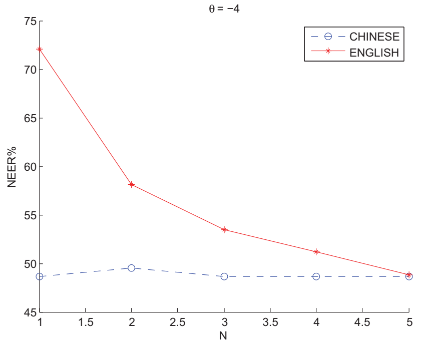


Figure 3: NEERs of English and Chinese words on the 'FOREIGN' test set. N is the number of referral high-frequency Chinese words. The value of θ is -4

To further examine the gains offered by the proposed approach, the name entity error rate (NEER) is used. In contrast to the WER that measures the accuracy on all words, the NEER evaluates the accuracy on focused words, e.g., the embedded English words. Table 4 present the results on the 'FOREIGN' test set with the basic similar pair method (one referral word). The NEER results on both the embedded English words and the rest Chinese words are reported. Table 5 and Table 6 present more details with multiple referral Chinese words. It can be seen that the similar-pair-based enhancement does deliver a much better accuracy on the embedded English words. Importantly, the improvement on the English words does not impact the performance on the Chinese words, which confirms the effectiveness and safety of the proposed method. A more clear presentation is given in Figure 3, where θ is set to be -4, and the number of referral words increases from 1 to 5.

		NEER%	
	θ	CHINESE	ENGLISH
Baseline	-	46.85	100
+ SP	-4	48.65	72.09
	-2	50.45	48.84
	0	50.45	32.56
	2	54.05	0
	4	59.46	0

Table 4: NEERs with and without the similar-pair-based enhancement. 'SP' stands for enhancement with similar pairs, which uses one referral Chinese word. θ is the enhancement scale in equation (2).

	NEER%				
N θ	1	2	3	4	5
-4	48.65	49.55	48.65	48.65	48.65
-2	50.45	49.55	51.35	51.35	51.35
0	50.45	51.35	56.76	56.76	56.76
2	54.05	58.56	57.66	57.66	58.56
4	59.46	63.96	63.06	62.16	62.16

Table 5: NEERs of Chinese words on the 'FOREIGN' test set. N is the number of referral high-frequency Chinese words.

	NEER%				
N θ	1	2	3	4	5
-4	72.09	58.13	53.49	51.16	48.84
-2	48.84	30.23	34.88	32.56	32.56
0	32.56	20.93	13.95	13.95	9.3
2	0	0	0	0	0
4	0	0	0	0	0

Table 6: NEERs of English words on the 'FOREIGN' test set. N is the number of referral high-frequency Chinese words.

6. Conclusion

In this paper, we proposed a similar-pair-based approach to enhance speech recognition accuracies on low-frequency and new words. Multiple referral words were explored, and the technique was applied to enhance foreign words. The experimental results demonstrated that the proposed method can significantly improve performance of speech recognition on low-frequency and new words and does not impact the ASR performance in general. Future work involves constructing word pairs using an automatic procedure, and combining this method with other dynamic LM approaches such as the class-based LM.

7. Acknowledgements

This research was supported by the National Science Foundation of China (NSFC) under the project No. 61371136, and the MESTDC PhD Foundation Project No. 20130002120011. It was also supported by Sinovoice and Huilan Ltd.

8. References

[1] S. Katz, "Estimation of probabilities from sparse data for the language model component of a speech recognizer," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 35, no. 3, pp. 400–401, 1987.

[2] R. Kneser and H. Ney, "Improved backing-off for m-gram language modeling," in *Proceedings of ICASSP*, 1995, pp. 181–184.

[3] S. F. Chen and J. Goodman, "An empirical study of smoothing techniques for language modeling," *Computer Speech & Language*, vol. 13, no. 4, pp. 359–393, 1999.

[4] W. Ward and S. Issar, "A class based language model for speech recognition," in *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*, vol. 1. IEEE, 1996, pp. 416–418.

[5] X. Ma, X. Wang, D. Wang, and R. Liu, "Low-frequency word enhancement with similar pairs in speech recognition," in *ChinaSIP15*, 2015.

[6] M. Mohri, F. Pereira, and M. Riley, "Weighted finite-state transducers in speech recognition," *Computer Speech & Language*, vol. 16, no. 1, pp. 69–88, 2002.

[7] P. F. Brown, P. V. Desouza, R. L. Mercer, V. J. D. Pietra, and J. C. Lai, "Class-based n-gram models of natural language," *Computational linguistics*, vol. 18, no. 4, pp. 467–479, 1992.

[8] M. Georges, S. Kanthak, and D. Klakow, "Transducer-based speech recognition with dynamic language models." in *INTERSPEECH*. Citeseer, 2013, pp. 642–646.

[9] J. Schalkwyk, I. L. Hetherington, and E. Story, "Speech recognition with dynamic grammars using finite-state transducers." in *INTERSPEECH*, 2003.

[10] P. R. Dixon, C. Hori, and H. Kashioka, "A specialized wfst approach for class models and dynamic vocabulary," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[11] W. Naptali, M. Tsuchiya, and S. Nakagawa, "Topic-dependent-class-based-gram language model," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 5, pp. 1513–1525, 2012.

[12] C. Samuelsson and W. Reichl, "A class-based language model for large-vocabulary speech recognition extracted from part-of-speech statistics," in *Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on*, vol. 1. IEEE, 1999, pp. 537–540.

[13] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," in *Proceedings of ASRU*, 2011, pp. 1–4.