Adapted Language Modeling for Recognition of Retelling Story in Language Learning

Meng Chen, Yang Song, Lan Wang
Shenzhen Institutes of Advanced Technology,Chinese Academy of Sciences/The Chinese University of Hong Kong
{chenmeng, yang.song, lan.wang}@siat.ac.cn

Abstract

N-gram language modeling typically requires large quantities of in-domain training data, i.e., data that matches the task in both topic and style. For the task of retelling stories, obtaining large volumes of speech transcriptions is often unrealistic. In this paper, we propose a novel method of language modeling using mixture models with very limited text data in the task of retelling stories. We modeled topic-specific, spoken-style, and document-style language models separately and interpolated them. We also interpolated the class-based language model with the N-gram models. Experimental results show that up to 61.6% reduction of perplexity and 20.7% reduction of word error rate (WER) have been obtained by our best performing model.

1. Introduction

With the development of computer technology, Computer Assisted Language Learning (CALL) system has offered great advantages over traditional language learning methods. Retelling stories has been presented to the language learner to evaluate his/her oral proficiency. Automatic scoring based on Automatic Speech Recognition (ASR) to evaluate the speaking ability in the task of retelling stories has been studied recently. In the task of retelling, students listen to a monologue of story (200~300 words) spoken by a native speaker, and then retell the story with their own words. The audios of students are non-native spontaneous speech with specific spoken style, which is not only different with the original story but also contains lexical and syntactic errors.

For spontaneous speech recognition, researchers have made numerous efforts to increase the ASR accuracy by employing a variety of improved language modeling techniques. In the study of [1], the authors constructed the language model for spontaneous speech by combination of written text from textbooks and transcripts of conversational telephone speech of Switchboard and Fisher corpora. Another work of [2] presented a method of generating simulated spoken-style text by randomly inserting fillers into written-style text. However, this approach handles only fillers, and doesn't consider features like repeat and self-repair. G. Moore and S. Young [3] used class-based language models for robust estimation of N-gram probabilities with limited or unmatched data. Akita and Kawahara [4] proposed the other approach using a probabilistic transformation model trained from a parallel aligned corpus of the faithful transcripts and their written-style texts. However, it is quite difficult to obtain such aligned corpus. All the above efforts mainly focused on spontaneous speech of native speakers, few researchers have explored the language modeling for the task of non-native spontaneous speech recognition.

Although these language model improvement techniques are undoubtedly helpful, they either need large amounts of closely matched data, or can only cover limited features of spontaneous speech style. For a task of retelling stories, the students are required to repeat the story based on what they heard, and they would organize the sentences with their own words when they can't remember the exact words used by the native speakers. Therefore, the speech of students is non-native spontaneous speech with three specific features. Firstly, the speech is closely related to the original story in topic, but not restricted with the vocabulary of original story. Secondly, there are lots of disfluencies, such as filled pauses, hesitation, repeated words and self-repaired words. Thirdly, it contains various lexical and syntactic errors since the speakers are non-native and their oral abilities are far from that of native speakers. Due to these specific features of retelling speech, transcripts of telephone conversations or newswire text are obviously not suitable.

In this paper we proposed an effective method to

improve language modeling in the task of retelling stories by using mixture models with very limited text sources. First, we collected the topic-related, spoken-style related and document-style related text sources to cover the specific features of retelling story, and trained the N-gram and class-based language models separately. Then, we mixed the topic-related, spoken-style and document-style language models for both N-gram and class-based language models by dynamic linear interpolation. Finally, we interpolated the mixture of N-gram language model with mixture of class-based language model to generate a final adapted language model. We evaluated our mixture model on the transcriptions of retelling speech. Experiments showed a significant reduction of perplexity, and the WER of ASR had been improved.

This paper is organized as follows. Section 2 introduces the text data we collected for language modeling and describes the diagram of language modeling using mixture language models. The experiments and results will be presented in Section 3. Section 4 is the conclusions.

2. Generating the LM for non-native spontaneous speech

2.1 Text sources for LM

As we mentioned in the above, we need to collect text data to match the specific features of non-native spontaneous speech. However, it's not realistic to collect large matched corpus for the task of retelling stories by transcribing from the audios directly, which is costly of man power and time consuming. As a result, the language model must be trained using the combination of alternative text sources. Considering the retelling speech is closely related to the original story, the prompts of original story can provide lots of the topic-related words. However, as the students can also express with their own words, the original story text data is too limited in the vocabulary. Thus, document style text sources, such as English textbooks and oral English tutorials etc., should be included in order to cover the vocabulary of high school students. Furthermore, non-native spoken-style related text sources would be necessary to contain the disfluencies and possible lexical and syntactic errors of non-native spontaneous speech.

Therefore, we collected three different types of text sources which are topic-related, spoken-style related and document style related text sources. The topic-related (TR) text source consists of the prompts of original story of retelling task. The spoken-style related (SS) text source is from SWECCL 2.0 [5], which consists of the spoken and written English corpus of Chinese learners. We only used the transcriptions of retelling speech in the Test for English Majors, Band 4 (TEM-4) from 2003 to 2006, containing 713 speakers and 4 topics in total. The document style (DS) related text source consists of English textbooks, oral English tutorials and materials from exam papers etc. Text conditioning is used for all the three text data to convert numerals to words and expand abbreviations. Table 1 lists the details of each type of text data.

Table 1 Details of text data for language modeling

Corpus	TR	SS	DS
#Words	537	159.1K	275.6K
#Uniq. words	212	2616	11712
#Sentences	42	8268	20095

2.2 Mixture of language models

A common technique for combining several language models is using a linear interpolation of two or more component models. The component models can be several N-gram language models, or N-gram language model and class-based language model. In the trigram case, each probability $P(w_i|w_{i-1}, w_{i-2})$ is replaced with a weighted sum of probabilities from $|S|$ individual models computed as follows:

$$P(w_i|w_{i-1}, w_{i-2}) = \sum_{s \in S} \lambda_s P_s(w_i|w_{i-1}, w_{i-2}) \quad (1)$$

The interpolation weights λ_s are estimated automatically using the Expectation-Maximization algorithm to maximize likelihood on a small held-out data set (or, equivalently, minimize perplexity) with the constraint that $\sum_{s \in S} \lambda_s = 1$. For the N-gram and class-based language model, the interpolation is similar as Equation (2), in which C_i represents word class.

$$P(w_i|w_{i-1}, w_{i-2}) = \sum_{s \in S} \lambda_s P_s(C_i|C_{i-1}, C_{i-2}) P(w_i|C_i) \quad (2)$$

The traditional method of language modeling using mixture models is by simply adding all the text sources to the available pool of training data, training the N-gram and class-based language models separately, and then interpolating the N-gram language model with class-based language model. In this paper, we proposed a novel language modeling method using mixture models. Instead of adding all the text sources to one corpus, both N-gram and class-based language model were trained separately with the TR corpus, SS corpus

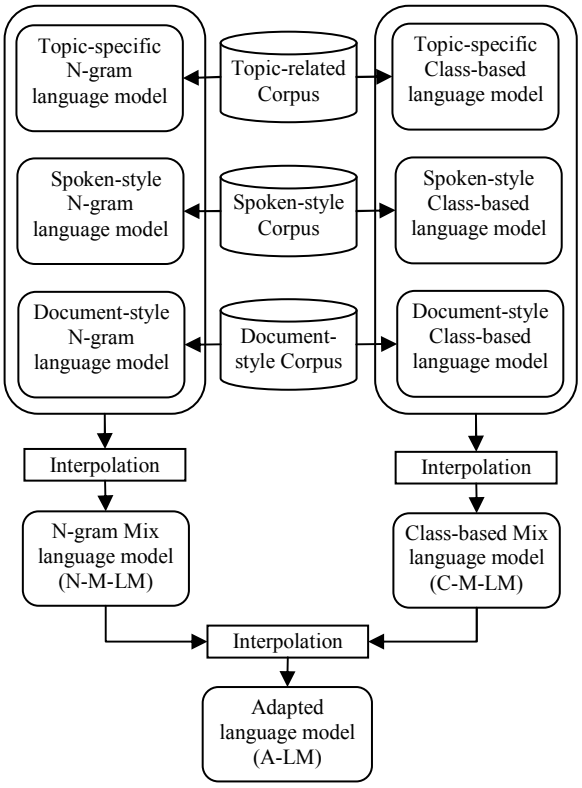

Figure 1. Flowchart of proposed method

and DS corpus firstly. Then interpolation were conducted both on the N-gram and class-based language models to generate the mixture of N-gram language model and mixture of class-based language model. Finally, we interpolated the two mixture language models to generate the adapted language model. Figure 1 shows the overview of our method.

3. Experiments

3.1 Training and test sets

The training data set consists of the TR corpus, SS corpus and DS corpus, which is very limited. We evaluated our models on two test sets, XIXIANG and YUCAI, which are the transcriptions from real audios in two tasks of retelling stories with different topics. XIXIANG consists of 2750 sentences and 28150 words, which is transcribed from real audios of 280 students. YUCAI consists of 207 sentences and 2257 words, which is transcribed from real audios of 15 students.

3.2 Adapted language models

We conducted the language modeling according the flowchart of our proposed method. All language models were trained using the SRILM toolkit [6]. The topic-related N-gram language model and all the class-based language models used interpolated Witten-Bell smoothing method [7]. Spoken-style N-gram language model and Document-style N-gram language model used interpolated modified Kneser-Ney smoothing method [8]. For each type of language model, we trained 2-gram, 3-gram and 4-gram models with cutoffs of 1 for all 2-gram, 3-gram and 4-gram. All interpolation weights were optimized with the tool of compute-best-mix of SRILM.

Considering the vocabulary size of high school students (below 5000) and the size of training set, we chose different vocabulary size for different training text sources. Given all the speech is derived from the original story, we took all words occurring in TR corpus when training topic-specific language models. Empirically we took top 2500 words in SS corpus when training spoken-style language models and a top of 3500 words in DS corpus when training document-style language models by word used frequency. Thus, by interpolation, there are totally 4373 words in N-gram Mix language model (N-M-LM), Class-based Mix language model (C-M-LM) and Adapted language model (A-LM). For the class-based language modeling, we clustered the word classes of 50, 200 and 300 for the TR corpus, SS corpus and DS corpus empirically, with the full greedy merging algorithm [9].

In order to compare with the traditional language modeling method, we trained the N-gram language model (N-LM) and Class-based language model (C-LM) by mixing all text sources together into a pool of training set. We restricted the vocabulary with the same 4373 words above. And for the class-based language model, we clustered 500 word classes. We also conducted the interpolation of N-LM and C-LM to generate the mixture of language model (M-LM).

3.3 Experimental results

We evaluated all the interpolation language models on test sets of XIXIANG and YUCAI. As the vocabulary is the same, the OOV (Out Of Vocabulary) rate is 1.18% in XIXIANG and 1.24% in YUCAI in all models. Table 2 is the perplexity of different language models in XIXIANG, and Table 3 is the perplexity of language models in YUCAI. Figure 2 and Figure 3 compares the perplexity of different models in a more direct way. From Figure 2, we can see that the A-LM reduced perplexity by relative 47% in average compared to the M-LM. And the A-LM reduced perplexity by relative 61.6% in average compared to M-LM from Figure 3.

Table 2 Perplexity of language models in XIXIANG

LM	2-gram	3-gram	4-gram
N-LM	153.586	138.968	134.119
N-M-LM	82.735	71.961	71.543
C-LM	158.655	162.104	167.211
C-M-LM	79.783	77.199	77.199
M-LM	146.506	132.831	129.529
A-LM	77.334	69.684	69.336

Table 3 Perplexity of language models in YUCAI

LM	2-gram	3-gram	4-gram
N-LM	154.888	111.921	106.872
N-M-LM	60.776	43.508	43.025
C-LM	176.971	150.592	158.884
C-M-LM	52.891	49.477	49.036
M-LM	149.186	107.461	102.981
A-LM	51.593	42.785	42.152

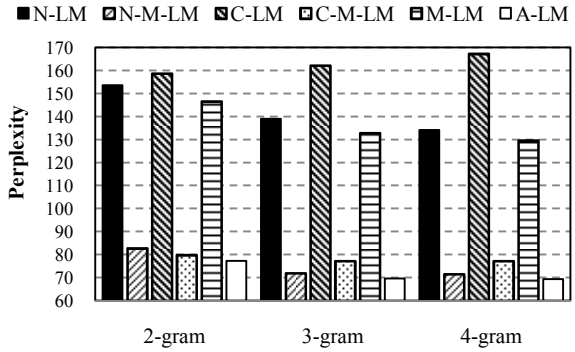

Figure 2. Perplexity of language models in XIXIANG

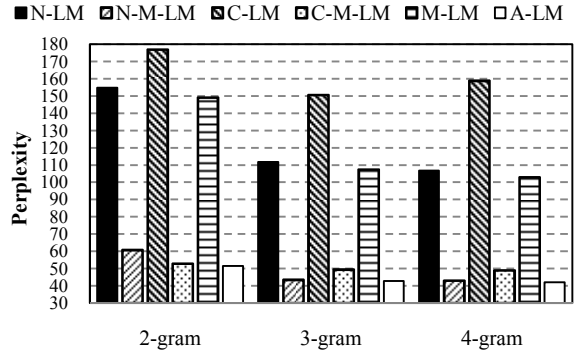

Figure 3. Perplexity of language models in YUCAI

We also evaluated the use of the adapted language models in ASR by the Word Error Rate (WER). We selected 11 corresponding audios of students from XIXIANG and 15 corresponding audios from YUCAI. The acoustic models are trained with the TIMIT/WSJ databases [10] [11]. The acoustic features include 12 MFCCs (Mel Frequency Cepstrum Coefficient), energy, delta and acceleration features. The HMMs are composed of three emitting states, each of which has 16 mixed Gaussian distributions. There are 8000 tied states in total. All decoding experiments were using the HTK toolkit [12]. We chose the 3-gram A-LM and M-LM language models for comparison. And we obtained an absolute 16.9% reduction of WER in XIXIANG test set and absolute 20.7% in YUCAI test set by A-LM.

4. Conclusions

We presented a novel approach to language modeling using mixture models in a task of retelling stories with limited text sources. By training topic-specific, spoken-style and document-style language models separately, interpolating the N-gram and class-based language models separately, and interpolating the N-gram language model with class-based language model, we obtained a reduction up to 61.6% in perplexity and 20.7% in WER compared to the traditional language modeling method.

An important point of future work will focus on class-based language model adaptation and acoustic modeling for the task of non-native spontaneous speech recognition. We plan to improve the clustering algorithm to cover the possible errors in vocabulary during retelling. And more features of retelling should be considered during acoustic modeling.

Acknowledgement

This work is supported by National Nature Science Foundation of China (NSFC 61135003, NSFC 90920002), The Knowledge Innovation Program of the Chinese Academy of Sciences (KJCXZ-YW-617), and Introduced Innovative R&D Team of Guangdong Province Robot and Intelligent Information Technology (201001D0104648280).

References

[1] A. Park, T. Hazen, and J. Glass, "Automatic Processing of Audio Lectures for Information Retrieval: Vocabulary Selection and Language Modeling", in *Proc. ICASSP*, 2005, vol. 1, pp.497-500.

[2] H. Schramm, X. L. Aubert, C. Meyer, and J. Peters, "Filled-Pause Modeling for Medical Transcriptions", in *Proc. SSPR*, 2003, pp. 143-146.

[3] G. Moore and S. Young, "Class-based Language Model Adaptation using Mixtures of word-class Weights", in *Proc. ICSLP*, 2000, pp. 512-515.

[4] Y. Akita and T. Kawahara, "Efficient Estimation of Language Model Statistics of Spontaneous Speech via Statistical Transformation Model", in *Proc. ICASSP*, Toulouse, France, 2006.

[5] Q. F. Wen, M. C. Liang, and X. Q. Yan, *Spoken and Written English Corpus of Chinese Learners (SWECCL) 2.0*,

Foreign Language Teaching and Research Press, Beijing, China, 2008.
[6] A. Stolcke, "SRILM – an extensible language modeling toolkit", in *Proc. ICSLP*, Denver, 2002, pp. 901-904.
[7] T. C. Bell, J. G. Cleary, and I. H. Witten, *Text Compression*. Prentice Hall, Englewood Cliffs, N. J.
[8] R. Kneser and H. Ney, "Improved Backing-off for m-gram Language Modeling", in *Prof. ICASSP*, 1995, pp.181-184.
[9] P. F. Brown, V. J. Della Pietra, P. V. deSouza, J. C. Lai, and R. L. Mercer, "Class-based n-gram Models of Natural Language", *Computational Linguistics*, vol. 18, no. 4, 1992, pp.467-479.
[10] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, *The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CDROM*. NTIS order number PB91-100354, 1993.
[11] D. Paul and J. Baker, "The Design for the Wall Street Journal-based CSR Corpus", *DARPA Speech & Nat. Lang. Workshop*, Arden House, NY, 1992.
[12] S. Young, G. Evermann, M. J. F. Gales, T. Hain, D. Kershaw, G. Moore, J. J. Odell, D. G. Ollason, D. Povey, V. Valtchev, and P. C. Woodland, *The HTK Book Version 3.4 Manual*, Cambridge, 2006.