DATA-DRIVEN LEXICON EXPANSION FOR MANDARIN BROADCAST NEWS AND CONVERSATION SPEECH RECOGNITION

Xin Lei   Wen Wang   Andreas Stolcke

Speech Technology and Research Laboratory
SRI International, Menlo Park, CA 94025 USA
{leixin,wwang,stolcke}@speech.sri.com

ABSTRACT

We present a data-driven framework for expanding the lexicon to improve Mandarin broadcast news and conversation speech recognition. The lexicon expansion includes the generation of pronunciation variants for frequent words and vocabulary augmentation with new words and phrases derived from the training data. To learn multiple pronunciations, we first generate all possible pronunciation candidates for a word from its character pronunciation network. The top pronunciation variants are then selected from forced alignment statistics. To augment the acoustic vocabulary, we propose an efficient algorithm that derives new words based on N-gram statistics. Experiments show that a dictionary expanded in this manner yields significant improvements on a Mandarin broadcast speech recognition task.

Index Terms— Pronunciation learning, vocabulary expansion, Mandarin speech recognition.

1. INTRODUCTION

A well-designed lexicon, both in terms of vocabulary selection and pronunciation dictionary, is an essential component for good automatic speech recognition (ASR) performance. Typically, the ASR lexicon is constructed manually using linguistic resources. The lexicon often needs to be updated or expanded over time. There are two main motivations to update and expand the lexicon for Mandarin broadcast news (BN) and broadcast conversation (BC) speech recognition. First, some frequent Mandarin (especially single-character) words have multiple pronunciations, some of which are easily omitted when generating pronunciations by hand or by rule. Furthermore, pronunciation errors are unavoidable in preexisting pronunciation dictionaries. Our approach will be to generate common pronunciation variants automatically from training corpora and incorporate them into the lexicon, or use them to identify potential pronunciation errors.

Second, since new words and phrases are created constantly in broadcast speech, the vocabulary needs to be expanded with new words. As discussed in [1], the vocabulary expansion problem is more serious in Chinese than in some western languages. Unlike in English and most other languages, the writing systems employed by Chinese is not alphabetic and does not mark word boundaries; consequently the notion of a word in Chinese is not conventionalized. For practical purposes (including speech recognition) a word is a commonly used string of one or more characters. New words such as personal and event names can be created easily by combining several distinctive characters. These new words may well be important keywords in broadcast audio sources. Since so many new words are created in Chinese speech over time, it is important to continually include these new words in the recognition lexicon and to generate pronunciations for them.

In this paper we propose a data-driven framework to expand the lexicon for Mandarin speech recognition. We first present a method for learning word pronunciations and pronunciation variants from training data. A character pronunciation dictionary is used to build a character pronunciation network (CPN) for each word. From it, all possible phone strings are generated and given a uniform weight, to be used in aligning the training data. The most appropriate pronunciation variants can be inferred from the forced alignment statistics and used as multiple pronunciations or to fix pronunciation errors. We also propose an efficient algorithm to learn new vocabulary words from N-gram statistics obtained on training text corpora. The pronunciation learning method can then be used to generate pronunciations for these new vocabulary words.

In Section 2, we describe a method to learn pronunciation variants for Chinese words from training data. In Section 3, an algorithm is proposed to extract new Chinese words from text corpora. Section 4 describes experiments in Mandarin broadcast recognition to evaluate the efficacy of the proposed lexicon expansion approach. Finally, we summarize the key findings and propose future work in Section 5.

2. PRONUNCIATION LEARNING

A Mandarin word may have different pronunciations depending on context. The pronunciation differences can be due to surface-form variation, such as neutral tones and tone sandhi changes, or because of baseform variation that indicates different meanings. For example, the word "同行" can have the meaning of "walk together" when its tonal Pinyin pronunciation is "*tong*2 *xing*2", and can also mean "person in the same occupation" when it is pronounced as "*tong*2 *hang*2". Both pronunciation variants are frequent for this word and should be explicitly incorporated in the dictionary as baseforms.

Several prior studies have dealt with automatic learning of multiple pronunciations from speech corpora [2, 3, 4]. Most approaches perform free phone recognition or phone-level forced alignment with pronunciation variants generated by linguistic rules. In this study, we explore learning of baseform variants as well as surface-form variants for Chinese words. Since around 6500 commonly used characters can combine to form hundreds of thousands of words, we propose to generate pronunciation variants by exploiting this special structure of Chinese words.

As illustrated in Figure 1, for a Chinese word the pronunciation variants of each constituent character are enumerated from a character pronunciation dictionary and combined to construct a character

pronunciation network (CPN). Note that "*tong*4" is a rare pronunciation of the character "同". The CPN generates all possible pronunciation variants of the word, including neutral tone variants for unstressed characters. After the pronunciation variants are enumerated, tone sandhi rules may be applied to explicitly encode tone changes within the word.
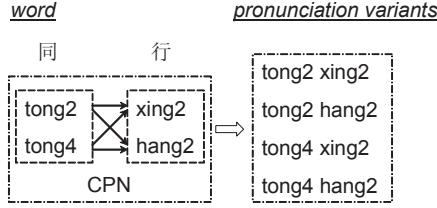
Fig. 1. Generation of pronunciation variants for a Chinese word from its character pronunciation network (CPN).

All the variants are then added to the dictionary with equal weight and used to perform forced alignment of the acoustic training data. The recognizer will choose the best pronunciation variant for each word in the sentence. Using the forced alignment statistics, we can select the top pronunciation candidates. If several variants have significant probabilities, these can be selected as multiple pronunciations. Existing dictionary pronunciations can be checked against the estimated pronunciation probabilities to spot dictionary errors.

3. VOCABULARY EXPANSION

Extraction of new words or phrases in Chinese has been studied recently in [5, 1]. The task is similar to the extraction of multiwords [2] or compound words [6] for English ASR. In [6, 1], an iterative approach is adopted to learn compound words. In each iteration, the top word pairs are merged according to a language model based measure. The measure is defined as the geometric mean of the *forward bigram* probability $P_f(w_j|w_i)$ between two words w_i and w_j and the *reverse bigram* probability $P_r(w_i|w_j)$

$$\mathcal{LM}_2(w_i, w_j) = \sqrt{P_f(w_j|w_i)P_r(w_i|w_j)}$$
$$= \frac{N(w_i\_w_j)}{\sqrt{N(w_i)N(w_j)}}, \tag{1}$$

where $N(\cdot)$ represents the frequency count of a word in the training data and $w_i\_w_j$ denotes the new compound word formed by concatenating two words w_i and w_j.

The iterative procedure for new word extraction can be time-consuming for large corpora, since it requires processing the entire corpus and updating of bigram counts or probabilities at every iteration. For Chinese, it additionally involves word segmentation, and may therefore require retraining of the word segmenter. As an efficient shortcut, we can approximate counts for iterated word formations directly from higher-order N-gram counts, as follows.

To combine three words w_i, w_j and w_k, we can either combine w_i and w_j first and then combine the new word $w_i\_w_j$ with w_k, or combine w_j and w_k first and then combine w_i and $w_j\_w_k$. We use the geometric mean of the merging measures of the two options to

create three-word compound words:

$$\mathcal{LM}_3(w_i, w_j, w_k)$$
$$= \sqrt{\mathcal{LM}_2(w_i\_w_j, w_k)\mathcal{LM}_2(w_i, w_j\_w_k)}$$
$$= \frac{N(w_i\_w_j\_w_k)}{\sqrt[4]{N(w_i\_w_j)N(w_k)N(w_i)N(w_j\_w_k)}}. \tag{2}$$

There are many more orderings when combining four words. Again, we could use the geometric mean of merging measures of all possible combination orders. However, since typical Chinese phrases are of length four characters or less while each word has at least one character, four-word compound words do not appear very often. Hence, we have used the following simplified measure:

$$\mathcal{LM}_4(w_i, w_j, w_k, w_l) = \mathcal{LM}_2(w_i\_w_j, w_k\_w_l)$$
$$= \frac{N(w_i\_w_j\_w_k\_w_l)}{\sqrt{N(w_i\_w_j)N(w_k\_w_l)}}. \tag{3}$$

Based on these merging measures, we can create new compound words with different numbers of constituent words. However, there are cases in which both a longer word $w_i\_w_j\_w_k\_w_l$ and its subword $w_i\_w_j$ are extracted as new words, while the shorter word does not appear by itself and should not be a separate new word. Therefore, after the new words are learned, we need to remove these subset words if their counts are almost the same as those of the corresponding longer words. This is similar to using the context variation criterion in [1]. The complete proposed Chinese new word extraction process is shown in Algorithm 1.

Algorithm 1: Extraction of new Chinese new words

Data: text training data
Threshold: M, Θ_n, β
Result: new vocabulary words
$N \longleftarrow 4$
for $n \leftarrow 1$ **to** N **do**
  dump n-gram counts
  sort n-grams by counts
  if $n \geq 2$ **then**
    foreach *element W_n of the top M n-grams* **do**
      compute the $\mathcal{LM}_n(\cdot)$ measure
      if $\mathcal{LM}_n(\cdot) > \Theta_n$ **then** save W_n

for $n \leftarrow (N-1)$ **to** 2 **do**
  foreach *element W_n of the saved n-grams* **do**
    foreach *element W_m of the saved higher-order m-grams $(m > n)$* **do**
      if W_m *contains* W_n **then**
        compute counts ratio r of W_n and W_m
        if $r < \beta$ **then**
          remove W_n from saved n-grams

output new words from all saved n-grams

Note that the algorithm may be sped up by learning word items with different numbers of constituents in parallel.

4. EXPERIMENTS AND RESULTS

All experiments were performed on a Mandarin broadcast speech recognition task. The SRI-UW Mandarin ASR system [7, 8] developed under the DARPA Global Autonomous Language Exploitation

(GALE) program was used. All results reported are from the first-pass decoding stage in our December 2007 evaluation system.

4.1. Baseline system

Acoustic models were trained on about 1440 hours of BN and BC speech data collected by LDC, including Mandarin Hub4 (30 hours), GALE Year 1 (747 hours), GALE Year 2 (394 hours) and GALE Year 3 Release 1 (269 hours) corpora. Training data were collected from broadcast stations such as CCTV, RFA, NTDTV and PHOENIX, spanning 1997 through 2007. Comparing to the December 2007 evaluation, we have removed 89 hours of TDT4 data that was relatively noisy and hurt performance slightly. Standard Mel frequency cepstral coefficients (MFCCs) plus smoothed pitch features [9] were used in the front end. Vocal tract length normalization and cepstral normalization were applied. The acoustic models were maximum-likelihood trained within-word models, with 3500 states clustered by decision tree and 128 Gaussian components per state.

The used text corpora comprised 1.4 billion words, including transcriptions of the acoustic training data, the LDC-released Mandarin Gigaword corpus, GALE-related web text data, and conversational telephone speech text data. A maximum likelihood word segmentation algorithm [8] was used to segment the training texts into multi-character words. BN and BC language models were trained with modified Kneser-Ney smoothing on a 60K word vocabulary, and interpolated to optimize the perplexity on a held-out set. A pruned trigram version of the final LM was used in first-pass decoding.

Two test sets were used for evaluation: DARPA GALE 2007 development set (dev07) and 2008 development set (dev08). The dev07 test set contains 2.4 hours of speech data from 74 broadcast shows. The dev08 test set has 1.0 hour of speech from 44 shows.

4.2. Learning pronunciations

The baseline training and decoding dictionaries used in our December 2007 evaluation have a 71K word and a 60K word vocabulary, respectively. We first experimented with adding multiple pronunciations to the baseline decoding lexicon, which has several hundred multiple pronunciations. Pronunciation variants for the 10K most frequent words in the decoding vocabulary were extracted from the expanded CPNs. A single-character dictionary was used in CPN construction, containing 6765 characters and 8095 multiple pronunciations (about 1.2 pronunciation variants per character). After expansion, we replaced the pronunciations of these top 10K words in the training dictionary with around 19K generated pronunciation variants. The minimum phone error (MPE) trained within-word acoustic model from the December 2007 evaluation was used to force-align the training data. (Note the acoustic models used for alignment were different from those trained based on the revised dictionary.)

We collected the counts of each pronunciation variant and computed their probabilities for each word. If a variant had a probability exceeding 0.9 it was chosen as a single pronunciation. It was then compared to the pronunciation in the original dictionary to check for a potential error. Otherwise, if a variant had a probability larger than 0.25, it was included as one of multiple pronunciations to be added to the lexicon. The expanded pronunciation lexicon was used to train new acoustic models. The pronunciation probabilities estimated from acoustic training were used to improve recognition by combining them with language model probabilities during decoding. We performed experiments with two versions of the expanded

Table 1. *CER (%) results with expanded lexicons. Dict v1 is expanded with 400 multiple pronunciations. Dict v2 is expanded by 800 multiple pronunciations over baseline. Dict v3 is augmented by 4245 new words over Dict v2. Dict v4 has the additional words filtered by human inspections, retaining about 3500 new words over Dict v2.*

Lexicon	dev07			dev08		
	BN	BC	all	BN	BC	all
Baseline	4.9	20.6	13.9	6.2	19.2	12.9
Dict v1	4.7	20.1	13.5	5.9	19.0	12.7
Dict v2	4.7	20.1	13.5	5.7	19.0	12.6
Dict v3	4.8	19.4	13.2	5.7	18.3	12.2
Dict v4	4.8	19.3	13.1	5.5	18.3	12.1

lexicon: "dict v1" with 400 pronunciation variants selected by word frequency, and "dict v2" with 800 pronunciation variants.

4.3. Expanding the vocabulary

New vocabulary words were extracted from BN and BC transcriptions separately, since the two genres are quite different lexically. In addition, the topics of BC genre are more diversified. In Algorithm 1, the threshold M is set to 50000, β is set to 1.2, and the Θ_n's are tuned to obtain the desired number of new words in each genre. About 1600 two-component, 800 three-component, and 100 four-component new words were extracted from both BN and BC data, resulting in a total of 4245 unique new vocabulary items (with less than 20% overlap between BN and BC new words). To generate pronunciations for these new words, the CPN-based pronunciation learning procedure described above was used. This expanded lexicon is referred to as "dict v3". Finally, we had native speakers of Mandarin inspect the added vocabulary items in dict v3, and eliminate those that were not semantically plausible as word units. The result of this hybrid method was "dict v4", which had about 3500 new words compared to dict v2.

For these experiments, we did not retrain the acoustic model (retraining the acoustic model did not result in further improvements). As in the multiword handling in [10], the new compound words were used to replace the corresponding constituent words in the text data. The new words were then added to the original 60K decoding vocabulary, yielding in a new vocabulary of about 64K. The language model was retrained with the expanded vocabulary.

4.4. Results and discussion

Recognition results in terms of character error rate (CER) are listed in Table 1. As we can see, significant improvements are obtained using the expanded lexicons. By adding 400 frequent multiple pronunciations (dict v1), we are able to achieve consistent improvements on both test sets and both genres. No significant improvement is obtained by adding more pronunciation variants (dict v2). The biggest improvement on conversational speech is obtained by expanding the vocabulary (dict v3), which yields 0.7% absolute improvement on the BC subset of both dev07 and dev08 test sets. We may attribute this to the fact that conversational speaking style renders word more confusable, and that concatenating words into longer units makes these less confusable again [6]. There is no improvement or a small degradation on the BN subsets after vocabulary expansion. This means that vocabulary expansion is more beneficial for the BC genre, possibly because pronunciations are more affected

by co-articulation in that genre. Inspecting the automatically generated words showed that some of them were not semantically plausible, yet experiments (dict v4) showed that using this human knowledge to filter the vocabulary gave only very small or no improvements in recognition accuracy.

5. CONCLUSIONS AND FUTURE WORK

We have investigated data-driven methods for expanding a Mandarin ASR lexicon with new multiple pronunciations and new vocabulary items. Based on Chinese word structure, a character pronunciation network is used to generate exhaustive pronunciation variants which are then pruned by forced alignments. We also propose a simple Chinese word extraction algorithm based on N-gram statistics. Experimental results show that both pronunciation learning and vocabulary expansion can improve recognition accuracy. Vocabulary expansion is particularly effective for conversational speech.

Future work includes learning new words with discriminative methods, such as conditional random fields. We also plan to study the temporal characteristics of the vocabulary words in broadcast speech recognition. The outdated words may be removed from the decoding vocabulary.

6. ACKNOWLEDGMENTS

7. REFERENCES

[1] P.-C. Chang, S.-P. Liao, and L.-S. Lee, "Improved Chinese broadcast news transcription by language modeling with temporally consistent training corpora and iterative phrase extraction," in *Proc. Eurospeech*, 2003, pp. 421–424.

[2] M. Finke and A. Waibel, "Speaking mode dependent pronunciation modeling in large vocabulary conversational speech recognition," in *Proc. Eurospeech*, 1997, vol. 5, pp. 2379–2382.

[3] J.M. Kessens, C. Cucchiarini, and H. Strik, "A data-driven method for modeling pronunciation variation," *Speech Commun.*, vol. 40, no. 4, pp. 517–534, 2003.

[4] M.-Y. Tsai, F.-C. Chou, and L.-S. Lee, "Pronunciation modeling with reduced confusion for Mandarin Chinese using a three-stage framework," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 15, no. 2, pp. 661–675, Feb. 2007.

[5] J. Zhang, J. Gao, and M. Zhou, "Extraction of Chinese compound words - an experimental study on a very large corpus," in *The Second Chinese Language Processing Workshop attached to ACL2000*, 2000.

[6] G. Saon and M. Padmanabhan, "Data-driven approach to designing compound words for continuous speech recognition," *IEEE Trans. on Speech and Audio Processing*, vol. 9, no. 4, pp. 327–332, May 2001.

[7] M.Y. Hwang, X. Lei, W. Wang, and T. Shinozaki, "Investigation on Mandarin broadcast news speech recognition," in *Interspeech*, 2006, pp. 1233–1236.

[8] M.Y. Hwang, G. Peng, W. Wang, A. Faria, A. Heidel, and M. Ostendorf, "Building a highly accurate Mandarin speech recognizer," in *Proc. Automatic Speech Recognition and Understanding Workshop*, 2007, pp. 490–495.

[9] X. Lei, M. Siu, M.Y. Hwang, M. Ostendorf, and T. Lee, "Improved tone modeling for Mandarin broadcast news speech recognition," in *Proc. Interspeech*, 2006, pp. 1237–1240.

[10] A. Stolcke, H. Bratt, J. Butzberger, H. Franco, V.R. Rao Gadde, M. Plauché, C. Richey, E. Shriberg, K. Sönmez, F. Weng, and J. Zheng, "The SRI March 2000 Hub-5 conversational speech transcription system," in *Proc. NIST Speech Transcription Workshop*, College Park, MD, May 2000.