

IMPROVING LANGUAGE MODELS FOR MANDARIN CONVERSATIONAL SPEECH RECOGNITION WITH WEB DATA

Tim Ng^{†‡}, Mari Ostendorf[†], Mei-Yuh Hwang[†], Manhung Siu[‡], Ivan Bulyko[†], Xin Lei[†]

[†]University of Washington
Electrical Engineering Department
Seattle, Washington, USA

[‡]Hong Kong University of Science and Technology
Department of Electrical and Electronic Engineering
Hong Kong, China

ABSTRACT

Lack of data is a problem in training language models for conversational speech recognition, particularly for languages other than English. Experiments in English have successfully used web-based text collection targeted for a conversational style to augment small sets of transcribed speech; here we look at extending these techniques to Mandarin. In addition, we investigate different techniques for topic adaptation. Experiments in recognizing Mandarin telephone conversations show that use of filtered web data leads to a 28% reduction in perplexity and 7% reduction in character error rate, with most of the gain due to the general filtered web data.

1. INTRODUCTION

Language models (LMs) constitute one of the key components in modern speech recognition systems. Training an N-gram language model, the most commonly used type of model, requires large quantities of text that is matched to the target recognition task both in terms of style and topic. In tasks involving conversational speech, the ideal training material, i.e. transcripts of conversational speech, is costly to produce, which limits the amount of training data currently available. This is particularly an issue in moving to new languages, as is our goal in this work with Mandarin.

Methods have been developed for topic-dependent language model adaptation, but it is critical that the base model be of reasonable quality. For conversational speech applications, this means that it reflects an informal speaking style. Recently researchers have turned to the World Wide Web as an additional source of training data for language modeling, e.g. [1, 2] which aim to adjust counts of N-grams occurring in the baseline model. However, since most web text tends to be of a more formal written style, the resulting distributions are typically not representative of conversational speech. Hence, we follow the basic approach proposed in [3] to extract additional training data from the web, where the authors search the web for documents containing conversational phrases, i.e. frequent n-grams from conversational transcripts. In experiments in English, this approach was shown to provide text that is better matched to a conversational speaking style. Some of the more conversational

data is obtained from chatrooms, which are commonly used in China as well as in the United States, so we expect that the techniques developed on English will be useful for language modeling in Mandarin, with some extensions to address language-specific issues.

Updating language model training data is important in any language, but China in particular has been changing rapidly in recent years. Lexical items have been invented or borrowed from other languages into Mandarin. For example, the word ‘买单’ (Can I have the check?), which is originally Cantonese and came with Hong Kong movies, is now used across China. The newly defined word ‘非典’ (SARS) has been frequently used since the epidemic broke out. Hence, in addition to capturing conversational style, we are also concerned with covering such recent words, which may be sparsely represented in the training data. Recent words sometimes cannot be found at all in older data collections, even if they are frequently seen in current data. Gathering text from more recent sources, such as from the web, can help LMs cover higher-order n-grams with these words, so we have developed a topic-oriented web data collection protocol, still tailored to a conversational style. Because we have information on the possible topics in our test data, we also investigate alternatives for explicit topic modeling.

In this paper, we will first describe our task briefly, including an overview of our baseline system, and then outline the procedures for general conversational and topic-based web data collections. Techniques for constructing the general language model and alternatives for topic adaptation will be discussed next. Finally, we will present the experiments followed by a summary of results.

2. TASK DESCRIPTION AND BASELINE LM

2.1. Training Corpora

Our baseline training data involved 35 hours of Mandarin conversational telephone speech (CTS) from the CallHome (CH) and CallFriend (CF) corpora, and 58 hours of a new corpus that was collected by HKUST, denoted here as RT04. CH and CF were collected in the US and the speakers were

mostly students from China who talked to their friends or families in China over long distance calls. They were free to choose what to talk about. RT04 was collected in China and the speakers were mostly college students in mainland China or Hong Kong. In addition to the difference in telephone channels, the subjects in RT04 were constrained to select from a list of 40 conversation topics. RT04 contains 251 topic-labeled phone calls (conversations), which we divided into two subsets: Heldout and Train04. Heldout comprised 25 conversations, randomly selected from RT04, for tuning language modeling parameters. The remaining 226 conversations, Train04, were used in LM training. Acoustic model training included the full set of RT04, plus CH+CF. The test set (Dev04) was also collected by HKUST and included 24 phone calls (2.5 hrs of speech). In all the experiments reported here, Dev04 audio was automatically segmented into sentence-size utterances by a silence/speech Gaussian mixture model prior to recognition.

2.2. Word Segmentation and Vocabulary Selection

Unlike English, words in Chinese are not demarcated by blank spaces, and characters are the basic units of transcription. While words in Chinese are not well-defined, human annotators will often agree on a word segmentation, and recognition accuracy tends to be higher when such words are used as the modeling units rather than characters. The RT04 and Dev04 transcripts did not have word segmentation marked by hand, and so they were automatically segmented using a tool from NMSU [4]. Also, the definition of hesitations in Chinese is complicated and there are some mismatches between CH+CF and RT04. In CH+CF, character ‘啊’ is a hesitation and marked as ‘%啊%’, however, in RT04, it is just a normal character. In order to keep the training data consistent, we also auto-segmented transcriptions of CH and CF. The vocabulary of the system included 11K words that had at least two occurrences in CH+CF+Train04, plus words of the single characters in the auto-segmented training texts. Note that the vocabulary included some English words, such as ‘ok’, ‘yeah’, ‘computer’, and ‘Christmas’.

There are a lot of compound words in Chinese and they need to be taken into consideration in language modeling. For example, ‘湖南省’ (Hunan Province) could be one of the OOVs, however, ‘湖南’ (Hunan) and ‘省’ (Province) might be able to be found in the 11K vocabulary list. Therefore, the Chinese OOVs in all of the data sets, including the Heldout set and the Dev04, were further segmented into some of the words in the 11K vocabularies by using the longest string match.

2.3. Baseline LM

We trained two separate LMs using CH+CF (P_{CC}) and Train04 (P_{04}). Our baseline LM was the interpolation of the two:

$$P_B = 0.87P_{04} + 0.13P_{CC},$$

where the weights were obtained by maximizing the Held-out likelihood. The low weight of CH+CF was consistent with our observation that it was very different from RT04 in topic and style.

2.4. Recognition System

We used SRI’s 5-times real time system as our speech recognition engine [5]. The acoustic models were gender-independent maximal mutual information estimated models that used tree-based state clustering [6, 7]. The front-end was the same as that described in [8], excluding speaker clustering because there are no longer multiple speakers in a single conversation side in the test set. The system uses a 2-stage search, in which a lattice is first generated by a system with an MFCC front-end, within-word triphones, phone-loop adaptation and a bigram LM and then rescored by a second system based on a PLP front-end, unsupervised MLLR-adapted cross-word triphones and a trigram LM. The PLP triphone model was adapted based on the recognition output of the MFCC first pass.

3. WEB DATA COLLECTION

3.1. General collection

The amount of text available on the web is enormous (over 3 billion web pages indexed by Google alone) and continues to grow. Most of the text on the web is not conversational, but there is a fair amount of chat-like material that is conversational-style (though without disfluencies). This was our primary target when extracting data from the web. Queries submitted to Google were composed of the 8800 4-grams that occur most frequently in the training corpora, e.g. ‘对对对’ (yes, yes, yes), ‘呃我觉得’ (uh, I think), etc. We searched for the exact match to one or more of these N-grams within the text of the web pages in GB encoding only. Web pages returned by Google mostly consisted of conversational-style phrases like ‘说实话’ (to be honest) and ‘你在说什么’ (What are you talking about?)

3.2. Topic-based collection

We used a slightly different search strategy when collecting topic-specific data. Instead of selecting keywords, we chose a set of ‘key phrases’ as queries for each topic. Key phrases embed keywords that provide some indicator of the topic. For topics that had sufficient data in Train04, key phrases were selected as follows. The word sequence, $w_i w_{i+1} w_{i+2}$,

is considered as a key phrase for topic t if

$$\frac{C(w_i, w_{i+1}, w_{i+2} | \text{topic } t) / \alpha_t}{\sum_{j=1}^{40} C(w_i, w_{i+1}, w_{i+2} | \text{topic } j) / \alpha_j} > \theta$$

where $C(w_i, w_{i+1}, w_{i+2} | \text{topic } t)$ is the trigram count in topic t , θ is an empirically chosen constant ($\theta = 0.3$), and α_t is the total number of word tokens in topic t . Using this criterion helps obtain topic-discriminant key phrases. For example, the above formula yields a small value for phrases common to many topics, such as ‘对对对’ (yes, yes, yes). For a topic-related phrase, such as ‘抽烟我觉得’ (Smoking, I think) the formula yields a value closer to 1. This example illustrates that, because the key phrases are learned from conversational speech, they also reflect that speaking style.

Some of the 40 specified topics were actually not covered in Train04. For those topics, we manually designed key phrases based on the brief descriptions that were provided to the subjects as part of the data collection protocol. After the key phrases were defined for all 40 topics, we then queried Google for 40 collections of web pages.

3.3. Text Normalization and Filtering

The retrieved web pages required several steps of processing before they could be used in language model training. As in [3], headers were stripped, duplicate pages (from different queries) were discarded, sentences were isolated based on punctuation, and the text was normalized to correspond to spoken (vs. written) form. Text normalization in Mandarin is much simpler than in English, and in this work it only involved transforming numerical digit strings to the words corresponding to their natural spoken form. In addition, because our LM is word-based, we also ran the automatic word segmentation algorithm on the resulting data.

After word segmentation, perplexity filtering was applied to ensure that the data had more of a conversational style, similar to the document pruning strategy used in [9]. Perplexity of each page was measured using the baseline LM; then we kept the pages with the lowest perplexity so that 60% of the total number of words was retained. Many pages containing heavily corrupted Chinese characters were removed automatically by the perplexity filtering. (Although most of the broken characters were removed in the text normalization procedure, there were still some corrupted characters that look like valid characters individually.)

The different amounts of data resulting from the final stage of the collection are listed in Table 1 together with amounts of data from the two conversational speech sources. As we have found in English, the web data is easily more than 100 times the in-domain data, even after filtering.

Sources	Number of Words
CH+CF	479K
Train04	398K
conversational web data	100M
topic-based web data	244M

Table 1. Data sources for LM training

4. LANGUAGE MODEL CONSTRUCTION

4.1. Static General Models

The most common method for using additional text sources is to train separate LMs for the different sources using a unified vocabulary and then to combine them by interpolation, also referred to as mixtures of language models. The mixture weights are optimized to maximize the likelihood of a heldout set by using the EM algorithm. The technique was reported by IBM in 1995 [10], and has been used by many sites since then. Interpolation tends to be much more effective than pooling data, because the better matched in-domain sources are usually much smaller in size than the out-of-domain sources. We built two static general models using component LMs trained on the conversational web data P_{conv} and optionally the topic web data P_{top}

$$\begin{aligned} P_{G3} &= 0.65P_{04} + 0.05P_{CC} + 0.30P_{conv} \\ P_{G4} &= 0.64P_{04} + 0.04P_{CC} + 0.16P_{conv} + 0.16P_{top} \end{aligned}$$

where the mixture weights were again tuned by maximizing the Heldout likelihood. Note that after web data was added, CH+CF received an even smaller weight.

4.2. Topic-Based Models

Researchers have developed a variety of methods for incorporating topic information in LMs. In this work, we consider three alternatives: a static topic mixture model, a dynamic topic mixture model, and a simple maximum entropy estimation technique for marginal adaptation [11].

Static Topic Mixture Model

The distribution of the 40 topics in Train04 was not balanced, and some topics had few or no examples represented in RT04. Hence, it was necessary to first cluster the 40 topics into a small set. We used an agglomerative clustering strategy with a modified Kullback-Liebler (KL) distance (made symmetric) between topic-dependent unigrams, where we chose to merge the clusters with the minimum max KL distance between individual topic models in the two clusters. Automatic clustering resulted in 15 clusters, which we merged (and in one case regrouped) by hand into 10 clusters to ensure that all clusters had some representation in both Train04 and Heldout. We then trained one LM (T_t)

for each cluster of topic web data. The static topic mixture model was a mixture of 13 components – the 10 web-only topic models $T_{1...10}$, P_{04} , P_{CC} , and P_{conv} , with weights estimated on Heldout.

Dynamic Topic Mixture Model

The idea behind dynamic mixture models was to estimate which topic among the 10 was most likely spoken in the conversation, and to adapt the decoding LM accordingly. To estimate the likelihood of each topic t , we computed topic posterior λ_t based on the hypothesis, $w_1w_2\dots$, from the first recognition pass, assuming equally likely topics:

$$\lambda_t = \frac{\prod_{i:w_i \in V^*} p_t(w_i|h_i)}{\sum_s \prod_{i:w_i \in V^*} p_s(w_i|h_i)}$$

where p_t is the N-gram trained on the t -th topic cluster of Train04 and V^* is the set of discriminative vocabulary items. (Experiments here used only unigrams, i.e. h_i is null.) The words in V^* are those with the maximum mutual information with the clustered topic labels, or equivalently the minimum topic entropy, according to Train04 unigrams. In English, this approach gave more accurate topic prediction than TF-IDF selection.

The dynamic topic model used the same 13 mixture components as the static model, but the mixture weights were a combination of static weights for the 4-component general model and the topic posteriors:

$$P_D = 0.64P_{04} + 0.04P_{CC} + 0.16P_{conv} + 0.16 \sum_{t=1}^{10} \lambda_t T_t$$

That is, the total contribution from topic web data was fixed (0.16) as in P_{G4} , but the individual contribution ($0.16 * \lambda_t$) from each topic t varied depending on what was spoken.

Marginal Adaptation Model

To obtain the marginal adaptation model, we first selected the most likely topic $t^* = \operatorname{argmax}_t \lambda_t$, according to the topic posterior above. Then we adapted the 4-component general LM P_{G4} to p_{t^*} using a maximum entropy approach [11] that constrains the model to match the topic-dependent unigram estimated from the Train04 clusters.

5. EXPERIMENTS

Table 2 gives recognition performance in terms of character error rate (CER) and the word perplexity on Dev04 for the different language models explored here. The biggest improvement in performance, both in perplexity and CER, came from adding the conversational web data. As mentioned earlier, this simple 3-component model had a very small weight (0.05) for CH+CF, which confirmed once more that RT04 data was very different from CH+CF.

web	Model	PPL	CER
none	Baseline LM	269.3	38.8%
C	3-comp general LM	202.2	36.4%
C,T	4-comp general LM	192.6	36.1%
C,T	Static topic mixture	196.6	36.3%
C,T	Dynamic topic mixture	-	36.2%
C,T	Marginal adaptation	-	36.4%

Table 2. Word perplexity and CER on Dev04 for different methods of using web data, where the ‘web’ column indicates the type of web data used: none, C (conversational), and/or T (topic).

In all models using topic data except for marginal adaptation, there was a small improvement over the 3-component general model. The gain from pooling all the topic web data to build a single component for the 4-component general model is significant ($p < .03$, using a matched pair test), but the gain associated with the various 10-component topic models was not statistically significant. The improvement of the simple 4-component model over the baseline is quite significant ($p < .001$). Note that the added topic web data had little impact on the mixture weights of Train04 and CH+CF. Instead, the total contribution from all web data remained about the same (0.30-0.32) as shown earlier.

The fact that the marginal adaptation did not give any gain was somewhat disappointing. It may be due to the fact that topic detection accuracy was only 75%, but we have observed similar results in experiments on English CTS with 95% topic detection accuracy. Topic adaptation has also been explored for English CTS by other sites without success [12], so it may be simply that the much more frequent general words dominate WER in CTS.

We investigated the effect of adding the Broadcast News text, because it has been found useful in experiments on English CTS. The Chinese BN corpus consists of three sources: China Radio, People’s Daily and Xinhua. We constructed 3 LMs for these 3 sources individually. The numbers of words for these 3 sources are 54.6M, 58.6M and 10.7M respectively. Then, they were interpolated with the 4 LM components. The optimized mixture weights for these 3 newly added components are 0.01, 0.002 and 0.02. The perplexity and CER for this LM are 192.5 and 36.2% respectively. In other words, they made no significant difference in system performance.

6. CONCLUSIONS

In this paper we have described methods for retrieving conversational and topic-based web data for use in augmenting training data for conversational speech language models. We obtain significant gains in performance with very simple mixture models: 28.5% reduction in word perplexity and 7% relative reduction in CER. In contrast, adding a similar amount of broadcast news text gave no benefit in

performance, which is somewhat in contrast with experiments in English except for the trend that the BN text is of diminishing value with the addition of web data.

Little additional benefit is obtained from explicit modeling of topics. This may be because first pass recognition errors limit the reliability of topic posteriors, or the fact that the small number of actual conversations for many of the topics made it difficult to estimate some of the parameters of the various models. In any case, there remain opportunities for further exploring the representation of topic in conversational language modeling.

Acknowledgments

This work was supported by the Defense Advanced Research Projects Agency grant MDA 972-02-C0038. The opinions, interpretations, conclusions and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

7. REFERENCES

- [1] A. Berger and R. Miller, "Just-in-time language modeling," in *Proc. ICASSP*, 1998, pp. II:705–708.
- [2] X. Zhu and R. Rosenfeld, "Improving trigram language modeling with the World Wide Web," in *Proc. ICASSP*, 2001, pp. I:533–536.
- [3] I. Bulyko, M. Ostendorf, and A. Stolcke, "Getting more mileage from web text sources for conversational speech language modeling using class-dependent mixtures," in *Proc. HLT/NAACL*, 2003, pp. 7–9.
- [4] W. Jin, "A case study: Chinese segmentation and its disambiguation," Tech. Rep. MCCS-92-227, New Mexico State University, Las Cruces, 1992.
- [5] Stolcke *et al.*, A., "Speech-to-text research at SRI-ICSI-UW," in *NIST RT-03 Workshop*, 2003.
- [6] I. P. C. Woodland and S. J. Young, "The HTK tied-state continuous speech recogniser," in *Proc. Eurospeech*, 1993, vol. 3, pp. 2207–2210.
- [7] Hwang *et al.*, M.-Y., "Predicting unseen triphones with senones," in *IEEE Trans. on Speech and Audio Processing*, 1996, vol. 4, pp. 412–419.
- [8] Hwang *et al.*, M., "Progress on Mandarin conversational telephone speech recognition," to appear in *ISCSLP*, 2004.
- [9] D. Klakow, "Selecting articles from the language model training corpus," in *Proc. ICASSP*, 2000, pp. III:1695–1698.
- [10] Liu *et al.*, F., "IBM Switchboard progress and evaluation site report," in *LVCSR Workshop*, Gaithersburg, MD, 1995, NIST.
- [11] Kneser *et al.*, R., "Language model adaptation using dynamic marginals," in *Proc. Eurospeech*, 1997, pp. 4:1971 – 1974.
- [12] L. Mangu, Personal Communication.