Language Modeling for Mixed Language Speech Recognition using Weighted Phrase Extraction

Ying Li, Pascale Fung

Department of Electronic and Computer Engineering
The Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong
eewing@ust.hk, pascale@ece.ust.hk

Abstract

To train a code switching language model for mixed language speech recognition, we propose to assign weights to the sentence pairs in the parallel text data. The code switching language model which is composed of the code switching boundary prediction model, code switching translation model and reconstruction model is incorporated with a language for mixed language speech recognition. The code switching translation model which is trained using selected subsets of the sentence pairs in the parallel text data allows the decoder to make the decision whether a phrase is in the matrix language or in the embedded language. Moreover, we propose a weighting procedure while training the code switching translation model. We evaluate our methods on Mandarin-English code switching lecture speech and lunch conversations. Our proposed method reduces word error rate by a statistically significant 1.74% on the lecture speech, and by 1.29% on the lunch conversation over the conventional interpolated language model.

Index Terms: mixed language, language model, code switching

1. Introduction

Large scale social changes lead to a considerable increase of the population that speak more than on language. Multilingual people often code switch, or mix words or phrases in the other language when they speak. The principal language in the mixed language sentence is the matrix language (ML), and the secondary, foreign language of the mixed words or phrases is the embedded language (EL) [2].

Speech recognition for such significant mixed language phenomenon is becoming important. There are two approaches of mixed language speech recognition. One is to segment the mixed language speech into segments, each segment is in either the matrix language or the embedded language. After the languages of the speech segments are identified, each speech segment is recognized by the speech recognition system of the corresponding language [3, 4, 5, 6]. In these works, the accuracies of the speech recognition systems depend on the accuracies of the segmentation and the language identification. The performance of the method is affected since there is no pause to indicate the code switching boundaries of the speech segments and the duration of the embedded word or phrase is too short for the language of the speech segments to be identified.

The other approach is to train a universal set of acoustic models and a language model which can handle both the matrix and embedded languages [7, 8, 9]. The main challenge of this approach is the lack of mixed language text data for training the language models. It have been found that code switching does not occur in positions where the order of the words is inverted between the matrix language and the embedded language [10, 11, 12], which corresponds to an inversion constraint in statistical machine translation (SMT)[13, 14]. Based on this, we incorporate a monolingual language model and a code switching language model composed of a boundary prediction model, a translation model and a reconstruction model. The monolingual language model is trained using text data in the matrix language. Subject to the inversion constrain, the code switching language model is trained using parallel text in the matrix and embedded languages. We also generate in-domain parallel text data using a statistical machine translation system. This can solve the sparse data problem of the language model for mixed language speech recognition. However, the use of parallel text is at the cost of specificity: either the data is unrelated to the task, or the data is small and generated by a machine. To better make use of the parallel text, sample weighting for language model training and translation model training [15, 16, 17].

In this paper, we propose to incorporate weights in the phrase training procedure of the code switching translation model. By weighting the sentence pairs in the parallel text data, the code switching translation model strikes a balance between the small in-domain machine generated data and the out-of-domain data.

2. Code Switching Language Modeling with Syntactic Constraint

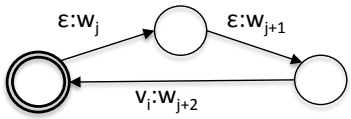
Instead of using the code switching text data which is rare to train a language model, we use a monolingual language model in the matrix language together with the code switching language model as follows:

$$P(W_1^M) = \sum_{w_1^m} P(w_1^m) P(W_1^M | w_1^m) \tag{1}$$

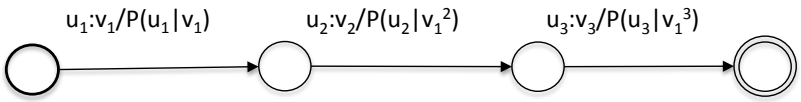
where W_1^M is in mixed language, and w_1^m is in the matrix language. The code switching language model $P(W_1^M|w_1^m)$ is the composition of a monolingual language model in the matrix language, a code switching boundary prediction model, a code switching translation model and a reconstruction model to avoid propagated error and to incorporate syntactic constraint of code switching speech.

The code switching language model can be modeled as

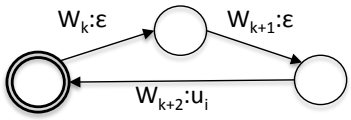
$$P(W_1^M|w_1^m) \cong \sum_{v_1^n, u_1^n, W_1^M} \{P(v_1^n, n|w_1^m) \cdot P(u_1^n|v_1^n, w_1^m) \cdot P(W_1^M|u_1^n, v_1^n, w_1^m)\} \tag{2}$$



A portion of the code-switch boundary prediction model



A portion of the code-switch translation model



A portion of the code-switch reconstruction model

Figure 1: Weighted finite state transducers of the code switching language model

where $P(v_1^n, n|w_1^m)$ is the code switching boundary prediction model, $P(u_1^n|v_1^n, w_1^m)$ is the code switching translation model, and $P(W_1^M|u_1^n, v_1^n, w_1^m)$ is the reconstruction model. A word sequence in the matrix language w_1^m is segmented into phrases, v_1^n; and u_1^n is a phrase sequence in mixed language.

2.1. Code Switching Boundary Prediction Model Training

According to linguistic findings [10, 11, 12], the code switching can only occur at points where the word order requirements of both the matrix and embedded languages are satisfied. Figure 2 shows an example of a Mandarin-English mixed language sentence. For example, code switching is not permissible between the first three words with syntactic inversions.

The code switching boundary prediction model is trained on the word-aligned parallel sentences in the matrix and embedded languages. The code switching boundary prediction model is the probabilities of a sequence of words segmented into a sequence of phrases. We define a phrase as a word or a concatenation of words in which there are one or more inversions of a word-aligned sentence pair in the matrix language and the embedded language.

$$P(v_1^n, n|w_1^m) = \frac{1}{Z_n} \prod_{i=1}^{n} P(v_i) \tag{3}$$

$$Z_n = \sum_{v_1^n} \prod_{k=1}^{m} P(v_i) \tag{4}$$

where $P(v_i)$ can be approximated by the relative frequency of the i-th phrase.

2.2. Code Switching Translation Model Training

The code switching translation model trains the probability of code switching from the matrix language to the embedded language at hypothesis code switching points given by the boundary prediction model. We assume the code switching translation probability $P(u_1^i|v_1^i)$ depends on the previous phrase v_{i-1}. The code switching translation probability distribution is specified by probabilities $\pi(\mathbf{x})$ trained from word-aligned bilingual sentences.$\mathbf{x}$ is an n-tuple which includes the word code switching probability $P(e|w)$, the reordering probability $\prod_{j=1}^{k} P(r_j|j,k,l)$, the phrase translation probability $Pr(u|v)$ and the phrase penalty $Pen(v)$, where w is an ML word, e is an EL word, k, l are the lengths of phrases in the matrix language and the embedded language, r_j denotes that the j-th word is aligned to the r_j-th EL word, v is an ML phrase, and u is an EL phrase.

The code-switch translation probability changes dramatically near the code switching threshold. Thus we use a logit regression model to describe the code switching translation probability

$$\text{logit}[\pi(\mathbf{x})] = \log(\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})}) = \alpha + \sum \beta_j x_j \tag{5}$$

where β_j is the effect of the j-th item in the n-tuple $\mathbf{x}$ on the logit of the code-switch translation probabilities, controlling

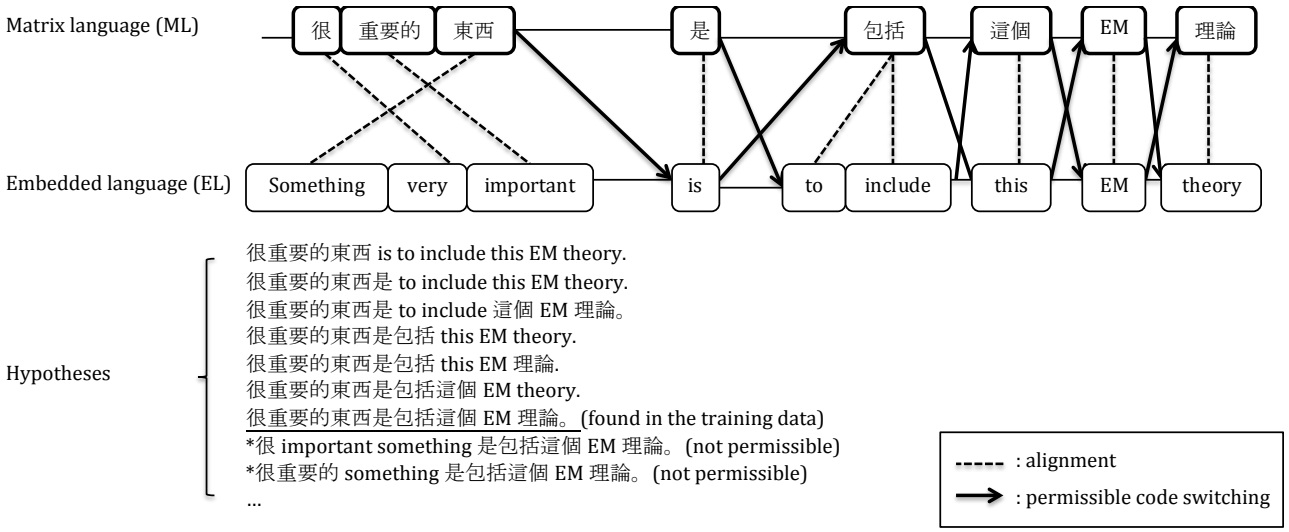


Figure 2: An example of permissible code switching points

the other items of $\mathbf{x}$. The code-switch translation probability

$$\pi(\mathbf{x}) = \frac{\exp(\alpha + \sum \beta_j x_j)}{1 + \exp(\alpha + \sum \beta_j x_j)} \tag{6}$$

$$P(u_i|v_1^i) = \begin{cases} 1 - \pi(\mathbf{x_{i-1}^i}) & u_i = v_i \\ \pi(\mathbf{x_{i-1}^i}) & otherwise \end{cases} \tag{7}$$

where $\mathbf{x_{i-1}^i}$is the n-tuple of the word alignment probabilities, reordering probability and the phrase penalty of the $(i-1)$th and ith phrases.

2.3. Code-switch Reconstruction Model Training

The reconstruction model assigns probabilities to a sequence of mixed language words, W_1^M, given that the words in the sequence are the same as the words of the phrases, u_1^n

$$P(W_1^M|u_1^n, v_1^n, n, w_1^m) = \prod_{i=1}^{n} P(W_{S_i}^{E_i}|u_i) \tag{8}$$

$$P(W_{S_i}^{E_i}|u_i) = \begin{cases} \frac{1}{Z_i} \prod_{j=S_i}^{E_i} q(W_j) & W_{S_i}^{E_i} = u_i \\ 0 & otherwise \end{cases} \tag{9}$$

where $p(W_j)$ is the frequency of occurrences of word W_j obtained from the bilingual sentences. $W_{S_i}^{E_i} = u_i$ indicates that the word sequence $W_{S_i}^{E_i}$ is exactly the same as the phrase u_i, S_i is the start of phrase u_i, and E_i is the end of phrase u_i. Z_i is set so that the probabilities sum to unity over possible values of u_i.

3. Weighted Phrase Extraction

In this paper, we propose to train the translation model $P(u|v)$ by assign a weight ω_r to each sentence pair.

$$P(u|v) = \frac{\sum_r \omega_r \cdot c_r(u,v)}{\sum_u \sum_r \omega_r \cdot c_r(u,v)} \tag{10}$$

where v is a phrase in the matrix language, u is a phrase in the embedded language, $c_r(u,v)$ is the count of v and u being a translation of each other in sentence pairs. Assigning higher weights to the sentence pairs will increase the corresponding probabilities. Therefore, the sentence pairs will have more effect of computing the code switching translation model.

Suppose I denote the in-domain corpus and O denotes the out-of-domain corpus, we randomly partition O in to subsets $\hat{O}$. The language models LM_I and $LM_{\hat{O}}$ are trained using I and $\hat{O}$ respectively.

The perplexity of sentence o is:

$$2^{-\sum_x LM_{\hat{O}}(x) log LM_I(x)} = 2^{H_I(o)} \tag{11}$$

where $H_I(o)$ is the cross-entropy according to the language model LM_I. Assume LM_I is a trigram language model,

$$H_I(o) = -\frac{1}{n} \sum_{n=1}^{n} logP(w_i|w_{i-1}, w_{i-2}) \tag{12}$$

where $o = w_1, w_2, ..., w_n$.

The sentences which are more like to the in-domain corpus and unlike the generalized model of the out-of-domain corpus will have a higher chance of being used to train the code switching translation model, the sentence pairs s_o, t_o in $\hat{O}$ is scored

using the summation of the cross-entropy differences between the in-domain and out-of-domain corpora of the matrix land embedded languages:

$$d_o = [H_{I_{matrix}}(s_o) - H_{\hat{O}_{matrix}}(s_o)] + [H_{I_{embedded}}(t_o) - H_{\hat{O}_{embedded}}(t_o)] \quad (13)$$

where I_{matrix} and $I_{embedded}$ denotes the in-domain corpus in the matrix and embedded language, $\hat{O}_{matrix}$ and $\hat{O}_{embedded}$ denotes a partition of the out-of-domain corpus in the matrix and embedded language. By assigning

$$\omega_o = e^{-d_o} \quad (14)$$

to the sentence o for weighted phrase extraction, the training of the code switching translation model biases to the sentences which are as close as possible to the in-domain corpus and different from the average of the out-of-domain corpus.

This term is proportional to perplexities and inverse perplexities, as the exponent of entropy is perplexity by definition.

4. Experiments

The bilingual acoustic model is trained from 160 hours of speech from GALE Chinese broadcast conversation, 40 hours of speech from GALE English broadcast conversation, and 3 hours of in-house nonnative English data. The acoustic features used in our experiments consist of 39 components (13MFCC, 13ΔMFCC, 13$\Delta\Delta$MFCC using subtraction of the cepstral mean), which are analyzed at a 10msec frame rate with a 25msec window size. The acoustic models used throughout our paper are state-clustered crossword tri-phone HMMs with 16 Gaussian mixture output densities per state. We use the phone set consists of 21 Mandarin standard initials, 37 Mandarin finals, 6 zero initials and 6 extended English phones. The pronunciation dictionary is obtained by modifying Mandarin and English dictionaries using the phone set. The acoustic models are reconstructed by decision tree tying [18, 19]. A WFST decoder is used for decoding .

4.1. Corpora

We compare our proposed method to the baseline interpolated language model on two mixed language speech corpora of different speaking styles, namely a lecture speech corpus and a lunch conversation corpus and a lunch conversation corpus recorded at the Hong Kong University of Science and Technology. The lecture speech corpus recorded at National Taiwan University contains about 20 hours of lecture speech of a digital speech processing course and are separated into three sets. 18 hours of the lecture speech is used to adapt the acoustic models, 0.9 hours of the speech is used as a development set, and one hour of 1037 utterances are used as test set. The lecture is given in Mandarin by a single speaker with 16% embedded English words.

The lunch conversation speech recorded at the Hong Kong University of Science and Technology contains 163 minutes of conversation speech. The speech is highly spontaneous and the topics are wide ranging. 127 minutes of the conversation speech is used to adapt the acoustic models, 26 minutes of the speech is used as a development set, and 280 utterances are used as test set. The percentage of the embedded English words is 22%.

250,000 sentences from digital speech processing conference papers, power point slides and web data and 250,000 sentences of the Gale conversational speech transcription are used for language model training and parallel sentence generation. The sentences generated by the SMT system and the GALE Phase 1 Chinese broadcast conversation parallel text are used for training the code switching translation model.

4.2. Baseline Language Models

250,000 sentences from digital speech processing conference papers, power point slides and web data are used for language model training and parallel sentence generation for the lecture speech recognition task (LM data 1). 250,000 sentences of the Gale conversational speech transcriptions are used for language model training and parallel sentence generation for the lunch conversion speech recognition (LM data 2). The baseline language model for the lecture speech recognition is an interpolation of the language model trained from LM data 1 and the language model trained on the transcriptions of the mixed language lecture speech. Another baseline model of the lunch conversations recognition is trained from LM data 2 and interpolated with the language model trained from the transcriptions of the mixed language lunch conversations.

4.3. Experimental Results

Table 1 shows the word error rates (WER) of experiments on the mixed language lecture speech and lunch conversations. The code switching language model outperforms the baseline interpolated language models by 0.84% on the lecture speech data and 1.15% on the lunch conversation data. Compare to the baseline language models, the code switching language model with weighted phrase extraction gives about 1.29% word error rate reduction on the lecture speech data and 1.74% word error rate reduction on the lunch conversation data. All the WER reductions are statistically significant at 99%.

Table 1: *Our proposed system outperforms the baselines in terms of WER*

	Lecture speech	Lunch conversations
InterpolatedLM	34.73%	46.20%
CodeSwitchingLM	33.89%	45.06%%
CodeSwitchingLM +WeightedPhraseExtraction	33.44%	44.47%

5. Conclusions

In this paper, we propose to assign weights to the sentence pairs in the parallel text data to train a code switching translation model. The code switching translation model, code switching boundary prediction model and reconstruction model are incorporated with a language model for mixed language speech recognition. We tested our system on two tasks, in mixed language lecture speech recognition and in mixed language lunch conversation. Our system reduces word error rate in a baseline of the interpolated language model by 1.29% in the first task, and by 1.74% in the second task. Our model also outperforms another baseline, that of code switching language model by 0.45% in the first task, and by 0.59% in the second task. All results are statistically significant. In addition, our method reduces word error rates for both the matrix language and the embedded language.

6. References

[1] J.J. Gumperz, "Discourse strategies", Cambridge University Press, 1, 1982.

[2] Coulmas, F., "The handbook of sociolinguistics", Wiley-Blackwell, 1998.

[3] Vu, N.T. and Lyu, D.C. and Weiner, J. and Telaar, D. and Schlippe, T. and Blaicher, F. and Chng, E.S. and Schultz, T. and Li, H., "A first speech recognition system for Mandarin-English code-switch conversational speech", ICASSP, 2012

[4] , J.Y.C. Chan and PC Ching and T. Lee and H.M. Meng "Detection of language boundary in code-switching utterances by bi-phone probabilities" Chinese Spoken Language Processing, 2004 International Symposium on, 293–296.

[5] C.J. Shia and Y.H. Chiu and J.H. Hsieh and C.H. Wu "Language boundary detection and identification of mixed-language speech based on MAP estimation"", ICASSP 2004.

[6] D.C. Lyu and R.Y. Lyu "Language identification on code-switching utterances using multiple cues" Ninth Annual Conference of the International Speech Communication Association, 2008.

[7] Tsai, T.L. and Chiang, C.Y. and Yu, H.M. and Lo, L.S. and Wang, Y.R. and Chen, S.H. "A study on Hakka and mixed Hakka-Mandarin speech recognition" Chinese Spoken Language Processing (ISCSLP), 2010 7th International Symposium on, 199–204

[8] Yeh, C.F. and Huang, C.Y. and Sun, L.C. and Lee, L.S. "An integrated framework for transcribing Mandarin-English code-mixed lectures with improved acoustic and language modeling" Chinese Spoken Language Processing (ISCSLP), 2010 7th International Symposium on, 214–219

[9] Cao, H. and Ching, PC and Lee, T. and Yeung, Y.T. "Semantics-based language modeling for Cantonese-English code-mixing speech recognition Chinese Spoken Language Processing (ISCSLP), 2010 7th International Symposium on,246–250

[10] Woolford, E., "Bilingual code-switching and syntactic theory", in Linguistic Inquiry, 14(3):520–536, JSTOR, 1983.

[11] MacSwan, J., "13 Code-switching and grammatical theory", in The Handbook of Bilingualism and Multilingualism, 323 Wiley-Blackwell, 2012.

[12] Poplack, S. and Sankoff, D. "A formal grammar for code-switching", in Papers in Linguistics: International Journal of Human Communication, 3–45, 1980.

[13] Wu, D. "Stochastic inversion transduction grammars and bilingual parsing of parallel corpora", in Computational linguistics, 23(3):377–403, MIT Press, 1997.

[14] Wu, D. and Fung, P. "Inversion transduction grammar constraints for mining parallel sentences from quasi-comparable corpora" in Natural Language Processing–IJCNLP 2005, 257–268 Springer.

[15] Moore, Robert C and Lewis, William "Intelligent selection of language model training data" Proceedings of the ACL 2010 Conference Short Papers, 220–224.

[16] Axelrod, Amittai and He, Xiaodong and Gao, Jianfeng "Domain adaptation via pseudo in-domain data selection" Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, 355–362.

[17] Mansour, Saab and Ney, Hermann "A simple and effective weighted phrase extraction for machine translation adaptation" Proceedings IWSLT 2012.

[18] Li, Y. and Fung, P. and Xu, P. and Liu, Y., "Asymmetric acoustic modeling of mixed language speech", ICASSP, 2011.

[19] Li, Y. and Fung, P., "Improved mixed language speech recognition using asymmetric acoustic model and language model with code-switch inversion constraints ", ICASSP, 2013.