A Language Modeling Approach to Identifying Code-Switched Sentences and Words

Liang-Chih Yu[1], Wei-Cheng He[1] and Wei-Nan Chien[1,2]
[1]Department of Information Management, Yuan Ze University, Taiwan, R.O.C.
[2]Information Technology Center, National Taiwan Normal University, Taiwan, R.O.C.
Contact: lcyu@saturn.yzu.edu.tw

Abstract

Globalization and multilingualism contribute to code-switching – the phenomenon in which speakers produce utterances containing words or expressions from a second language. Processing code-switched sentences is a significant challenge for multilingual intelligent systems. This study proposes a language modeling approach to the problem of code-switching language processing, dividing the problem into two subtasks: the detection of code-switched sentences and the identification of code-switched words in sentences. A code-switched sentence is detected on the basis of whether it contains words or phrases from another language. Once the code-switched sentences are identified, the positions of the code-switched words in the sentences are then identified. Experimental results on Mandarin-Taiwanese code-switching sentences show that the language modeling approach achieved a 79.52% F-measure and an accuracy of 80.23% for detecting code-switched sentences, and a 51.20% F-measure for the identification of code-switched words.

1 Introduction

Increasing globalism and multilingualism has significantly increased demand for multilingual services in current intelligent systems (Fung and Schultz, 2008). For example, an intelligent traveling system which supports multiple language inputs and outputs can assist travelers in booking hotels, ordering in restaurants, and navigating attractions. Multinational corporations would benefit from developing automatic multilingual call centers to address customer problems worldwide. In such multilingual environments, an input sentence may contain constituents from two or more languages, a phenomenon known as code-switching or language mixing (Hoffmann, 1991; Myers-Scotton, 1993; Ayeomoni, 2006; Liu, 2008). A code-switched sentence consists of a primary language and a secondary language, and the secondary language is usually manifested in the form of short expressions such as words and phrases. This phenomenon is increasingly common, with multilingual speakers often freely moving from their native dialect to subsidiary dialects to entirely foreign languages, and patterns of code-switching vary dynamically with different audiences in different situations. When dealing with code-switched input, intelligent systems such as dialog systems must be capable of identifying the various languages and recognize the speaker's intention embedded in the input (Ipsic, et al., 1999; Holzapfel, 2005). However, it is a significant challenge for intelligent systems to deal with multiple languages and unknown words from various languages.

In Taiwan, while Mandarin is the official language, Taiwanese and Hakka are used as a primary language by more than 75% and 10% populations, respectively (Lyu, et al., 2008). Moreover, English is the most popular foreign language and compulsory English instruction begins in elementary school. The constant mix of these languages result in various kinds of code-switching, such as Mandarin sentences mixed with words and phrases from Taiwanese, Hakka, and English. Such code-switching is not limited to everyday conversation, but can frequently be heard on television dramas and even current events commentary programs. This paper takes a linguistic view towards the problem of code-

Proceedings of the Second CIPS-SIGHAN Joint Conference on Chinese Language Processing, pages 3–8,
Tianjin, China, 20-21 DEC. 2012

switching language processing, focusing on code-switching between Mandarin and Taiwanese. We propose a language modeling approach which divides the problem into two subtasks: the detection of code-switched sentences followed by identification of code-switched words within the sentences. The first step detects whether or not a given Mandarin sentence contains Taiwanese words. Once a code-switched sentence is identified, the positions of the code-switched words are then identified within the sentence. These code-switched words can be used for lexicon augmentation to improve understanding of code-switched sentences.

The rest of this work is organized as follows. Section 2 presents related work. Section 3 describes the language modeling approach to the identification of code-switched sentences and words in the sentences. Section 4 summarizes the experimental results. Conclusions are finally drawn in Section 5, along with recommendations for future research.

2 Related Work

Research on code-switching speech processing mainly focuses on speech recognition and synthesis (Lyu, et al., 2008; Wu, et al., 2006; Hong, et al., 2009; Chan, et al., 2006; Qian, et al.,2009). Lyu et al. (2008) proposed a three-step data-driven phone clustering method to train an acoustic model for Mandarin, Taiwanese, and Hakka. They also discussed the issue of training with unbalanced data. Wu et al. (2006) proposed an approach to segmenting and identifying mixed-language speech utterances. They first segmented the input speech utterance into a sequence of language-dependent segments using acoustic features. The language-specific features were then integrated in the identification process. Hong et al. (2009) developed a Mandarin-English mixed-language speech recognition system in resource-constrained environments, which can be realized in embedded systems such as personal digital assistants (PDAs). Chan et al. (2006) developed a Cantonese-English mixed-language speech recognition system, including acoustic modeling, language modeling, and language identification algorithms. For speech synthesis, Qian et al. (2009) developed a text-to-speech system that can generate Mandarin-English mixed-language utterances.

Research on code-switching and multilingual language processing included applications of unknown word extraction (Wu, et al., 2011), text mining (Yang, et al., 2011; Zhang, et al., 2011), and information retrieval (Tsai, et al., 2011). Wu et al. (2011) proposed the use of mutual information and entropy to extract unknown words from code-switched sentences. Yang et al. (2011) used self-organizing maps for multilingual document mining and navigation. Zhang et al. (2011) addressed the problem of multilingual sentence categorization and novelty mining on English, Malay, and Chinese sentences. Tsai et al. (2011) used the FRank ranking algorithm to build a merge model for multilingual information retrieval.

3 Language Modeling Approach

Language modeling approaches have been successfully used in many applications such as grammar error correction (Wu, et al., 2010) and lexical substitution (Yu, et al., 2010; 2011). For our task, a code-switched sentence generally has a higher probability of being found in a code-switching language model than in a non-code-switching one. Thus we built code-switching and non-code-switching language models to compare their respective probabilities of identifying code-switched sentences and code-switched words within the sentences. Fig. 1 shows the system framework. First, a corpus of code-switched and non-code-switched sentences are collected to build the respective code-switching and non-code-switching language models. To identify code-switched sentences, we compare the probability of each test sentence output by the code-switching language model against the output of the non-code-switching one to determine whether or not the test sentence is code-switched. To identify code-switched words within the sentences, we select the *n*-gram with the highest probability output by the code-switching language model, and then compare it against the output of the non-code-switching one to verify whether the n-*th* word in the given sentence is a code-switched word.

3.1 Corpus collection

A non-code-switching corpus refers to a set of sentences containing just one language. Because Mandarin is the primary language in this study, we used the Sinica corpus released by the Association

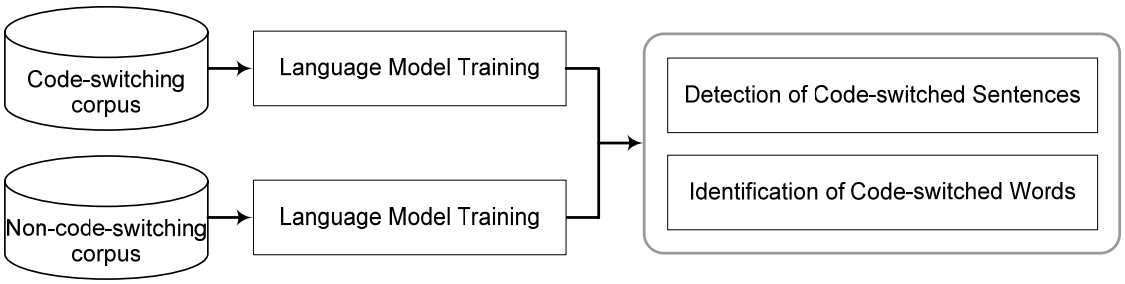

Figure 1. Framework of identification of code-switched sentences and words in the sentences.

for Computational Linguistics and Chinese Language Processing (ACLCLP) as the non-code-switching corpus. A code-switching corpus refers to a set of Mandarin sentences featuring Taiwanese words. However, it can be difficult to collect a large number of such sentences, and training a language model on insufficient data may incur the data sparseness problem. Therefore, we used more common Mandarin-English sentences as the code-switching corpus, based on the assumption that the code-switching phenomenon in Mandarin-English sentences has a certain degree of similarity to Mandarin-Taiwanese sentences because, in Taiwan, both English and Taiwanese are secondary languages with respect to Mandarin. The Mandarin-English sentences were collected from a large corpus of web-based news articles which were then segmented using the CKIP segmentation system developed by the Academia Sinica, Taiwan (http://ckipsvr.iis.sinica.edu.tw) (Ma and Chen, 2003). The sentences containing words with the part-of-speech (POS) tag "FW" (i.e., foreign word) were selected as code-switched sentences.

3.2 Detection of code-switched sentences

Generally, an *n*-gram language model is used to predict the n-*th* word based on the previous n-1 words using a probability function $P(w_n|w_1...w_{n-1})$. Given a sentence $S=w_1 \ldots w_k$, the non-code-switching *n*-gram language model is defined as

$$\begin{aligned} P_{\overline{CS}}(S) &= P(w_1)P(w_2 \mid w_1)\cdots P(w_k \mid w_1...w_{k-1}) \\ &= \prod_{i=1}^{k} P(w_i \mid w_1...w_{i-1}) \\ &\approx \prod_{i=1}^{k} P(w_i \mid w_{i-1}...w_{i-n+1}) \end{aligned} \quad (1)$$

where $P(w_i|w_{i-1}...w_{i-n+1})$ is estimated by

$$P(w_i|w_{i-1}...w_{i-n+1}) = \frac{C(w_i...w_{i-n+1})}{C(w_{i-1}...w_{i-n+1})}, \quad (2)$$

where $C(\bullet)$ denotes the frequency counts of the *n*-grams retrieved from the non-code-switching corpus (i.e., Sinica corpus). Instead of estimating the surface form of the next word, the code-switching *n*-gram language model estimates the probability that the next word is a code-switched word, i.e., $P(cs_n|w_1...w_{n-1})$, defined as

$$\begin{aligned} P_{cs}(S) &= P(w_1)P(cs_2 \mid w_1)\cdots P(cs_k \mid w_1...w_{k-1}) \\ &= \prod_{i=1}^{k} P(cs_i \mid w_1...w_{i-1}) \\ &\approx \prod_{i=1}^{k} P(cs_i \mid w_{i-1}...w_{i-n+1}) \end{aligned} \quad (3)$$

where $P(w_i|w_{i-1}...w_{i-n+1})$ is estimated by

$$P(cs_i|w_{i-1}...w_{i-n+1}) = \frac{C(cs_i...w_{i-n+1})}{C(cs_{i-1}...w_{i-n+1})}, \quad (4)$$

To estimate $P(cs_n \mid w_1...w_{n-1})$, the code-switching corpus is processed by replacing the code-switched words (i.e., the words with the POS tag "FW") in the Mandarin-English sentences with a special character *cs*. The frequency counts of $C(cs_i...w_{i-n+1})$ can then be retrieved from the code-switching corpus. This processing may also reduce the effect of the data sparseness problem in language model training.

Once the two language models are built, they can be compared to detect whether a given sentence contains code-switching. That is,

$$c = \frac{P_{CS}(S)}{P_{\overline{CS}}(S)}. \quad (5)$$

The sentence S is predicted to be a code-switched sentence if the probability of the sentence output by the code-switching language model is greater than that output by the non- code-switching one (i.e., $c \geq 1$).

3.3 Identification of code-switched words

This step identifies the positions of the code-switched words within the sentences. To this end, the code-switching n-gram language model (Eq. (3)) is applied to each test sentence and the probability of being a code-switched word is assigned to every next word (position) in the sentence. Among all the n-grams in the sentence, the one with the highest probability indicates the most likely position of a code-switched word. That is,

$$cs^* = \arg\max_i P(cs_i \,|\, w_{i-1}...w_{i-n+1}), \quad (6)$$

where cs^* denotes the best hypothesis of the code-switched word in the sentence. However, not all n-grams with the highest probability suggest correct positions. Therefore, we further propose a verification mechanism to determine whether to accept the best hypothesis. That is,

$$cs = \begin{cases} cs^* & P^*(cs_i \,|\, w_{i-1}...w_{i-n+1}) \geq P(w_i \,|\, w_{i-1}...w_{i-n+1}) \\ \phi & P^*(cs_i \,|\, w_{i-1}...w_{i-n+1}) < P(w_i \,|\, w_{i-1}...w_{i-n+1}) \end{cases} \quad (7)$$

where $P^*(cs_i \,|\, w_{i-1}...w_{i-n+1})$ represents the probability of the best hypothesis in the code-switching corpus, and $P(w_i \,|\, w_{i-1}...w_{i-n+1})$ represents its probability in the non-code-switching corpus. The best hypothesis cs^* is accepted if its probability in the code-switching corpus is greater than that in the non-code-switching corpus.

4 Experimental Results

This section first explains the experimental setup, including experiment data, implementation of language modeling, and evaluation metrics. We then present experimental results for the identification of code-switched sentences and words within the sentences.

4.1 Experimental setup

The test set included 86 sentences where 43 sentences were Mandarin only (i.e., non-code-switched) and another 43 Mandarin sentences containing Taiwanese words (i.e., code-switched). N-gram models for both code-switching and non-code-switching were trained using the SRILM toolkit (Stolcke, 2002) with n=2 (i.e., bigram). The evaluations metrics included recall, precision, F-measure, and accuracy. The recall was defined as the number of code-switched sentences correctly identified by the method divided by the total number of code-switched sentences in the test set. The precision was defined as the number of code-switched sentences correctly identified by the method divided by the number of code-switched sentences identified by the method. The F-measure was defined by defined as $\frac{2 \times recall \times precision}{recall + precision}$. The accuracy was defined as the number of sentences correctly identified by the method divided by the total number of sentences in the test set.

4.2 Results

To identify code-switched sentences, the code-switching and non-code-switching bigram models were used to determine whether or not each test sentence features code-switching (Eq. (5)), with results presented in Table 1. The language modeling approach correctly identified 33 code-switched sentences and 36 non-code-switched sentences, thus yielding 76.74% (33/43) recall, 82.50% (33/40) precision, 79.52% F-measure, and 80.23% (69/86) accuracy.

To identify code-switched words in the sentences, all word bi-grams in each test sentence were first ranked according to their probabilities. The top N word bi-grams were then selected as candidates for further verification using Eq. (7). To examine the effect of the data sparseness problem, we built an additional POS bi-gram model from the code-switching corpus. Table 2 shows the results for the identification of code-switched words using the word and POS bi-gram models. With more candidates included for verification (i.e., Top 1 to Top 3), more code-switched words were correctly identified, thus dramatically increasing the recall of both word and POS bi-gram models, while slightly decreasing the precision of both models.

	Recall	Precision	F-measure	Accuracy
Bi-gram	76.74%	82.50%	79.52%	80.23%

Table 1. Results of the identification of code-switched sentence.

Word Bi-gram	Recall	Precision	F-measure
Top1	39.53%	42.50%	40.96%
Top2	62.79%	32.93%	43.20%
Top3	88.37%	33.33%	48.41%
POS Bi-gram	Recall	Precision	F-measure
Top1	41.86%	42.86%	42.35%
Top2	74.42%	39.02%	51.20%
Top3	93.02%	34.78%	50.63%

Table 2. Results of the identification of code-switched words.

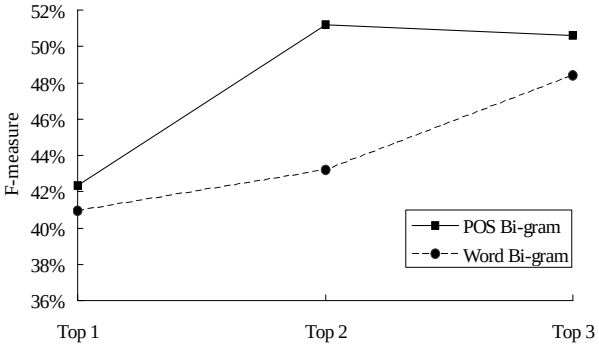

Figure 2. Comparative results of word and POS bi-gram language models.

Overall, the F-measure of both models increased as the number of candidates included increased. Figure 2 compares the word and POS bi-gram models, showing that the POS bi-gram model outperformed the word bi-gram model in terms of F-measure, as well as for recall and precision (see Table 2). This finding indicates that training with the POS tags can reduce the impact of the data sparseness problem, thus improving the identification performance.

5 Conclusions

This work presents a language modeling method for identifying sentences featuring code-switching, and for identifying the code-switched words within those sentences. Experimental results show that the language modeling approach achieved a 79.52% F-measure and 80.23% accuracy for the detection of code-switched sentences. For the identification of code-switched words within sentences, the POS bi-gram model outperformed the word bi-gram model, mainly because of the reduced impact of the data sparseness problem. The highest F-measure for this task was 51.20%. Future work will focus on improving system performance by incorporating other effective machine learning algorithms and features such as sentence structure analysis. The proposed method could also be integrated into practical applications such as a multilingual dialog system to improve effectiveness in dealing with the code-switching problem.

Acknowledgement

This work was supported by National Science Council, Taiwan, R.O.C (NSC99-2221-E-155-036-MY3), and Aim for the Top University Plan, Ministry of Education, Taiwan, R.O.C. The authors would like to thank the anonymous reviewers and the area chairs for their constructive comments.

References

Ayeomoni, M. O. 2006. Code-Switching and Code-Mixing: Style of Language Use in Childhood in Yoruba Speech Community. Nordic Journal of African Studies, 15(1): 90–99.

Chan, J. Y. C., Ching, P. C., Lee T. and Cao, H. 2006. Automatic Speech Recognition of Cantonese-English Code-mixing Utterance. In Proc. of Interspeech, pages 113-116.

Fung, P., and Schultz, T. 2008. Multilingual Spoken Language Processing. IEEE Signal Processing Magazine, 25(3): 89-97.

Hoffmann, C. 1991. An Introduction to Bilingualism. London. New York: Longman.

Holzapfel, H. 2005. Building Multilingual Spoken Dialogue Systems. Archives of Control Sciences, 15(4): 555-566.

Hong, W. T., Chen, H. C., Liao, I. B. and Wang W. J. 2009. Mandarin/English Mixed-Lingual Speech Recognition System on Resource-Constrained Platforms. In Proc. of the 21st Conference on Computational Linguistics and Speech Processing (ROCLING-09), pages 237-250.

Ipsic, I., Pavesic, N., Mihelic, F. and Noth, E. 1999. Multilingual Spoken Dialog System. In Proc. of the IEEE International Symposium on Industrial Electronics, pages 183-187.

Liu, Y. 2008. Evaluation of the Matrix Language Hypothesis: Evidence from Chinese-English Code-switching Phenomena in Blogs. Journal of Chinese Language and Computing, 18(2): 75-92.

Lyu, D. C., Hsu, C. N., Chiang, Y. C. and Lyu R. Y. 2008. Acoustic Model Optimization for Multilingual Speech Recognition. International Journal of Computational Linguistics and Chinese Language Processing, 13(3): 363-386.

Ma, W. Y. and K. Chen J. 2003. A Bottom-up Merging Algorithm for Chinese Unknown Word Extraction. In Proc. of the ACL Workshop on Chinese Language Processing, pages 31-38.

Myers-Scotton, C. 1993. Social Motivations for Code Switching: Evidence from Africa. Oxford University Press, New York.

Qian, Y., Liang, H. and Soong F. 2009. A Cross-Language State Sharing and Mapping Approach to Bilingual (Mandarin–English) TTS. IEEE Trans. on Audio, Speech, and Language Processing, 17(6): 1231-1239.

Stolcke, A. 2002. SRILM — An Extensible Language Modeling Toolkit. In Proc. of the 7th International Conference on Spoken Language Processing (ICSLP-02), pages 901-904.

Tsai, M. F., Chen, H. H. and Wang Y. T. 2011. Learning a Merge Model for Multilingual Information Retrieval. Information Processing and Management, 47(5): 635-646.

Wu, C. H., Chiu, Y. H., Shia, C. J. and Lin C. Y. 2006. Automatic Segmentation and Identification of Mixed-language Speech using Delta-BIC and LSA-based GMMs. IEEE Trans. Audio, Speech, and Language Processing, 14(1): 266-276.

Wu, Y. L., Hsieh, C. W., Lin, W. H., Liu, C. Y. and Yu L. C. 2011. Unknown Word Extraction from Multilingual Code-Switching Sentences In Proc. of the 23rd Conference on Computational Linguistics and Speech Processing (ROCLING-11), pages 349-360.

Wu, C. H., Liu, C. H., Matthew, H. and Yu L. C. 2010. Sentence Correction Incorporating Relative Position and Parse Template Language Models. IEEE Trans. on Audio, Speech, and Language Processing, 18(6): 1170-1181.

Yang, H. C., Hsiao, H. W., and Lee, C. H. 2011. Multilingual Document Mining and Navigation Using Self-organizing Maps. Information Processing and Management, 47(5): 647-666.

Yu, L. C., Wu, C. H., Chang, R. Y., Liu, C. H. and Hovy, E. H. 2010. Annotation and Verification of Sense Pools in OntoNotes. Information Processing and Management, 46(4): 436-447.

Yu, L. C., Chien, W. N. and Chen, S. T. 2011. A Baseline System for Chinese Near-Synonym Choice. In Proc. of the 5th International Joint Conference on Natural Language Processing (IJCNLP-11), pages 1366-1370.

Zhang, Y., Tsai, F. S. and Kwee A. T. 2011. Multilingual Sentence Categorization and Novelty Mining. Information Processing and Management, 47(5): 667-675.