A Mandarin-English Code-Switching Corpus

Ying Li, Yue Yu, Pascale Fung

Human Language Technology Center
Department of Electronic & Computer Engineering, HKUST
eewing@ust.hk, eeyyxaa@ee.ust.hk, pascale@ee.ust.hk

Abstract

Generally the existing monolingual corpora are not suitable for large vocabulary continuous speech recognition (LVCSR) of code-switching speech. The motivation of this paper is to study the rules and constraints code-switching follows and design a corpus for code-switching LVCSR task. This paper presents the development of a Mandarin-English code-switching corpus. This corpus consists of four parts: 1) conversational meeting speech and its data; 2) project meeting speech data; 3) student interviews speech; 4) text data of on-line news. The speech was transcribed by an annotator and verified by Mandarin-English bilingual speakers manually. We propose an approach for automatically downloading from the web text data that contains code-switching. The corpus includes both intra-sentential code-switching (switch in the middle of a sentence) and inter-sentential code-switching (switch at the end of the sentence). The distribution of part-of-speech (POS) tags and code-switching reasons are reported.

Keywords: code-switch, code-mix, mixed language

1. Introduction

International business, communication and immigration have a great effect on increasing the multilingual population. Many countries and regions, such as Hong Kong or the United States, are officially or unofficially multilingual. There are more bilingual or multilingual people than monolingual ones in the world.

Some multilingual people code switch when they speak. When multilingual speakers hold a conversation, their utterances are often mixed with words or phrases in other languages. Sometimes the sentences within the same speech are in different languages. Mixed language is a significant phenomenon wherever two languages are in contact. Such a code-switching phenomenon is an aspect of bilingualism.

Code-switching is defined as the juxtaposition within the speech of words or phrases belonging to two or more grammatical systems or subsystems (Gumperz, 1982). It should be distinguished from loanword, which is a word borrowed from one language and incorporated into another language to become part of the lexicon. The integration of a loanword into the local language results in morphological and phonological modification of the foreign word, while code-switching is the point at which people actually try to speak another language. The matrix language frame model (MLF), proposed by C. and C. (1993), is one of the theoretical approaches to code-switching. In the MLF model, the matrix language is the 'principal' language in code-switching, the one 'around which something develops'; the 'embedded' language is the one 'fixed firmly in a surrounding mass', in this case the matrix language.

There are two types of code-switching: inter-sentential code-switching and intra-sentential code-switching. Inter-sentential code-switching is at utterance or sentence boundaries, while intra-sentential code-switching is within an utterance or a sentence.

The phenomenon of code-switching has been increasingly reported in linguistic studies in the past years. More and more people code-switch when they speak. Since the late 1970s, studies on Cantonese-English code-switching in Hong Kong have been published (D., 2000). Nowadays code-switching can be found in many countries, such as Indonesian-English-Chinese in Indonesia and German-Turkish in Germany (I.T. et al., 2010). The distribution of code-switching is not random but obeys certain rules and constraints which are still being researched. The reasons for code-switch can be divided into several categories including message qualification, quotation and addressee specification.

There are two approaches to recognizing code-switching speech. One of the approaches is to detect the boundaries at which the speaker code-switches, then identify the language in the speech segments between the boundaries, and decode the speech segments by the acoustic and language models in the corresponding language. Chan et al. (2005b) presented a method to detect the language boundary of Cantonese-English code-switching utterances by the bi-phone probabilities calculated using a Cantonese lexicon database. A universal phone set was defined as all the Cantonese phone models and English phone set, which is selected using either knowledge-based or data-driven methods. Shia et al. (2004) proposed a maximum a posteriori (MAP)-based framework for boundary detection and language identification of code-switching utterances. The latent semantic analysis formalism is adopted to reduce the number of dimensions of the space. A maximum likelihood ratio scheme is adopted to optimize the boundary number; then, a hypothesized language sequence is determined by maximizing the log-likelihood with respect to speech utterance. Dynamic programming is adopted to search the best boundary positions. Lyu and Lyu (2008) reported results on the language identification (LID) system using acoustic, phonetic, and prosodic cues and MAP-framework to identify Mandarin-Taiwanese utterance. Lyu et al. (2006) proposed to use triphone acoustic models and a word-based bi-gram language dependent model to recognize the content of the utterances based on the LID results.

The more straightforward way to decode code-switching speech is by using a set of universal acoustic models for both matrix and embedded languages and a language model that permits code-switching. Chan et al. (2006) used the monolingual Cantonese corpus and the Cantonese-English code-mixing corpus for training language independent models. Code-mixing text data was collected for language modeling. Zhang et al. (2008) presented a grammar-constrained, Mandarin-English bilingual speech recognition system. A singer's name and the title of a song in mono-Mandarin, mono-English and mixed language can be recognized using one set of acoustic models by clustering Mandarin and English phones. Y. et al. (2011) propose to improve the accuracy of speech recognition on mixed language speech by asymmetric acoustic modeling. The proposed system using selective decision tree merging of bilingual and accented embedded acoustic models improved recognition on embedded foreign speech without degrading the recognition on the matrix language speech.

Since code-switching involves at least two languages, the size of a vocabulary becomes larger and the combination of possible words increases exponentially. Therefore, the recognition of code-switching speech is problematic to state of the art-automatic speech recognition systems. Moreover, the large vocabulary speech recognition of code switching speech requires intensive resources.

This paper is motivated by two considerations. One of the goals is to collect data for the study of internal rules which code-switching should follow so that the event of code-switching can be predicted. The second goal is that the corpus can be used for training acoustic and language models to recognize Mandarin-English code switching speech.

Lyu et al. (2010) presented a Mandarin-English code-switching corpus. In this corpus, 30 hours of code-switching speech is recorded. Chan et al. (2005a) developed a Cantonese-English code-mixing speech corpus, and Solorio and Liu (2008) recorded about 40 minutes of English-Spanish code-switching conversation and proposed a primary experiment on predicting the code-switching points.

In this paper, we propose a Mandarin-English code-switching corpus. The design of the speech part of the corpus was done under the consideration that code-switching mostly occurs in spontaneous speech. Thus, it consists of conversational meeting, project meeting and student interviews speech with no prompts or scripts. The transcriptions were annotated manually. We also collected articles and sentences containing code-switching from on-line news automatically.

The structure of the paper is as below: Related works are introduced in Section 2. Section 3, 4 and 5 present the conversational meeting speech, project meeting speech and student interview corpora, respectively. The text data from the web and the approach to collecting these articles automatically are described in Section 6.

2. Conversational Meetings

The conversational meeting speech was recorded in a closed meeting room at the Hong Kong University of Science and Technology. The environment was quiet to minimize noise disturbance. The data was recorded and digitalized at 16kHz sampling rate, and 16-bit pulse-code modulation (PCM) with mono channel.

Four meetings were recorded during several weeks. Between seven and nine speakers from different backgrounds participated in the meetings. Except for one meeting, all meetings consisted of Chinese-English bilingual speakers of varying degrees of fluency. Only one participant in one meeting could not understand Chinese. The use of English in this meeting was noticeably more than in the other three meetings. The meetings were in Mandarin with code-switching to English. A few words and phrases form could be found in Cantonese, French and German. The meeting leader was a female Mandarin-English bilingual speaker. The topics of the meetings included university administration, historical events and politics.

Three microphones were used to record the conversational data. The speaker who led the group meetings was equipped with a close-talking microphone. Speech of the other speakers were captured using two table mounted microphones. The total length of the meetings was 163 minutes.

The recorded speech was divided by languages. Speech in Chinese was transcribed by an annotator into Chinese characters. Other speech, expect for a few words in French or German, was transcribed by an English speaker.

According to the results given by the Stanford Chinese word segmenter, there are 14762 Chinese words and 4280 English words. The percentage of the embedded language is 22.48%.

Table 1 shows the statistics of code-switching type for the conversational meeting speech.

	Code-switching from English	Code-Switching from Chinese
Inter-sentential	106	104
Intra-sentential	184	195

Table 1: Code-switching type for the conversational meetings

The Penn Treebank tag set and Chinese Treebank set were used for POS annotation. Figure 1 shows the distribution of POS tags of the conversational meeting data. Figure 2 shows the distribution of code-switching reasons.

3. Project Meetings

We recorded nine group meetings for discussing projects in a silent environment over six months. The project meeting data was collected with a mono channel; the sampling rate was 16000 Hz, and the resolution was 16 bit. The project meetings consisted of a group of six participants. Other than one English only speaker, all participants speak Chinese and English. The meetings were conducted in Mandarin and code-switched to English.

Two close-talking microphones were used to record the speech of participants. One desk mounted microphone was used to capture the speech of the other speakers. The duration of the nine meetings totalled 8 hours.

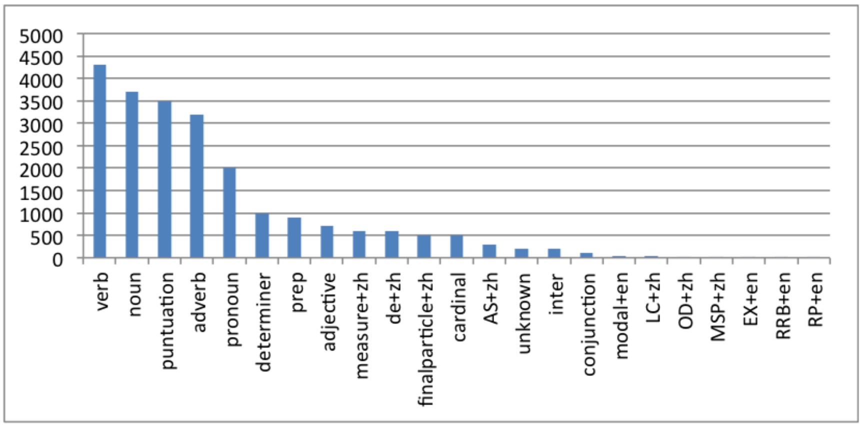

Figure 1: Distribution of POS tags of the conversational meetings.

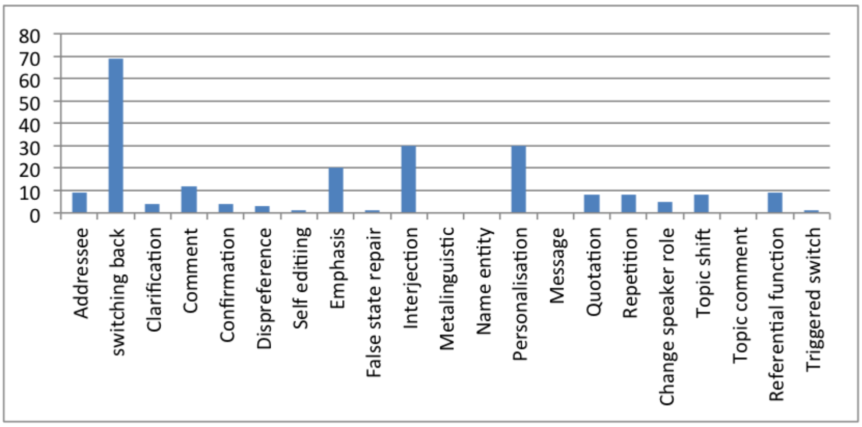

Figure 2: Distribution of POS tags for the conversational meetings.

The project meeting speech from three meetings was transcribed and annotated by a Mandarin-English bilingual speaker. The duration of the transcribed speech is 2.3 hours.

There are 10030 Chinese words and 2688 English words. The percentage of the embedded language is 21.15%.

Table 2 shows the statistics of code-switching type fore the project meeting speech.

	Code-switching from English	Code-Switching from Chinese
Inter-sentential	85	95
Intra-sentential	117	182

Table 2: Code-switching type of the project meetings

4. Interviews of Students

The interview speech data was collected from university students during the examination period for stress emotion detection in different languages. The recordings took place in a quiet conference room with high-quality equipment (Creativer Labs, Model No. SB0490). Speech was recorded in a lossless format with a sampling rate of 16,000Hz, using a single channel 16-bit digitization.

61 university students were asked to contribute to the Mandarin database, 42 university students to the English database and 69 university students to the Cantonese database. The duration of the Mandarin database is 9 hours, the duration of the English database is 4 hours and that of the Cantonese database is 12 hours.

Although only part of the English speech is transcribed, we found code-switching occurred during the interviews in the Mandarin and English databases.

In each interview setup, there was an interviewer and an interviewee. Only the interviewee's answer was recorded using a close-talking microphone.

5. Text Data from the Web

The text data containing code-switching was automatically downloaded from the web. The size of the database is 100 million characters.

There are constraints on data collection, because, although code-switching is commonly used in newsgroups and online diaries, it is quite different in style from either spoken or written Mandarin, and therefore, is not suitable for code-switching text data.

Inspired by the fact that newswire text can be used for estimating the language model parameters, we designed an algorithm to collect Chinese language news in which code-

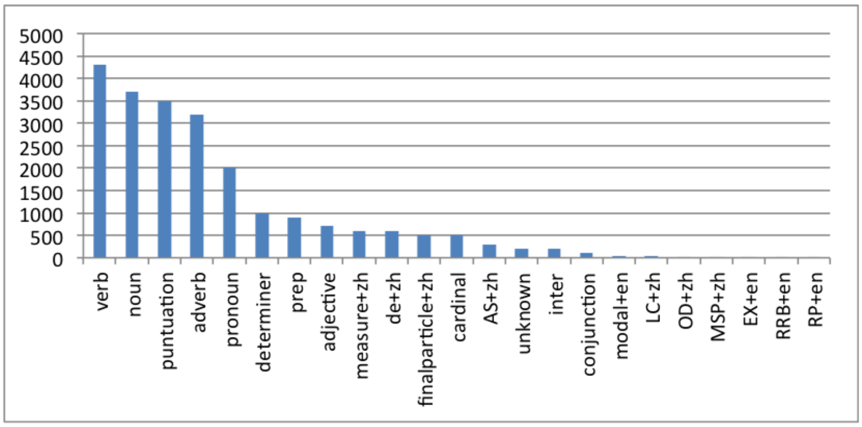

Figure 3: Distribution of POS tags of the conversational meetings.

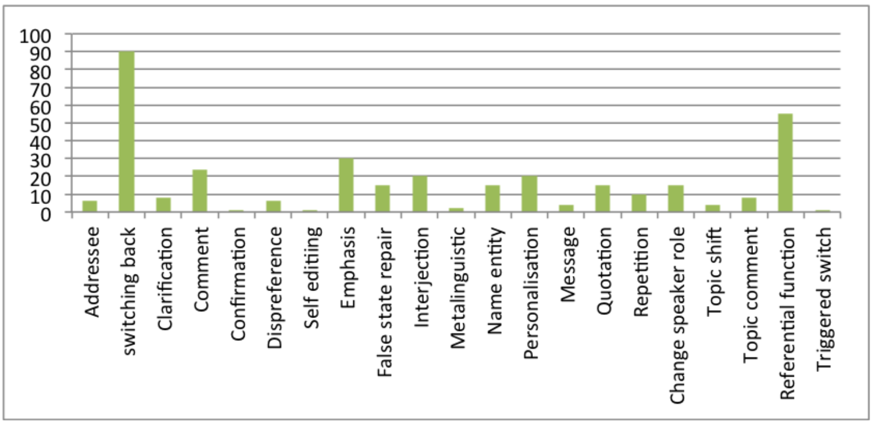

Figure 4: Distribution of POS tags of the conversational meetings.

switching occurs, using Google news API. The algorithm is described below:

1. Initialize the seed list of commonly used code-switching words and phrases. The seed terms are the most frequently used phrases in the embedded language of the collected database.

2. Use each word or phrase in the pool as a keyword to search for news in the matrix language. Detect if there is code-switching to the embedded language in the resultant news.

3. Update the list of words and phrases.

78k English words are mixed in the text corpus. The percentage of the embedded language is 7.8%. Most of the code-switching type is intra-sentential code-switching.

6. Conclusion

A straightforward approach to recognizing code-switching speech is to train acoustic and language models using code-switching data. However, there have not been many code-switching corpora. In this paper, the collection and annotation of a Mandarin-English code-switching corpus is described. The corpus includes conversational meeting speech and transcriptions, project meeting speech and transcriptions, a student interview corpus, as well as code-switching text from the web. We have provided an algorithm for downloading text data which contains code-switching from the internet. Our collected data can be used for the study of internal rules which code-switching should follow. The corpus can also be used for training acoustic and language models to recognize Mandarin-English code switching speech.

7. References

Myers-Scotton C. and Myers C. 1993. *Duelling languages: Grammatical structure in codeswitching*. Clarendon Press Oxford.

J.Y.C. Chan, PC Ching, and T. Lee. 2005a. Development of a cantonese-english code-mixing speech corpus. In *Ninth European Conference on Speech Communication and Technology*.

J.Y.C. Chan, PC Ching, T. Lee, and H.M. Meng. 2005b. Detection of Language Boundary in Code-switching utterances by Bi-phone Probabilities. In *Chinese Spoken Language Processing, 2004 International Symposium on*, pages 293–296. IEEE.

J.Y.C. Chan, PC Ching, T. Lee, and H. Cao. 2006. Automatic speech recognition of Cantonese-English code-mixing utterances. In *Ninth International Conference on Spoken Language Processing*. ISCA.

Li D. 2000. Cantonese-english code-switching research in hong kong: a y2k review. *World Englishes*, 19(3):305–322.

J.J. Gumperz. 1982. *Discourse strategies*, volume 1. Cambridge Univ Pr.

Schultz I.T., Fung P., and Burgmer C. 2010. Detecting code-switch events based on textual features.

D.C. Lyu and R.Y. Lyu. 2008. Language identification on code-switching utterances using multiple cues. In *Ninth Annual Conference of the International Speech Communication Association*.

D. Lyu, R. Lyu, Y. Chiang, and C. Hsu. 2006. Speech recognition on code-switching among the Chinese Dialects. *ICASSP, I (1105-1108)*.

D.C. Lyu, T.P. Tan, E.S. Chng, and H. Li. 2010. Seame: A mandarin-english code-switching speech corpus in south-east asia. In *Eleventh Annual Conference of the International Speech Communication Association*.

C.J. Shia, Y.H. Chiu, J.H. Hsieh, and C.H. Wu. 2004. Language boundary detection and identification of mixed-language speech based on map estimation. In *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*, volume 1, pages I–381. IEEE.

T. Solorio and Y. Liu. 2008. Learning to predict code-switching points. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 973–981. Association for Computational Linguistics.

Li Y., Fung P., Xu P., and Liu Y. 2011. Asymmetric acoustic modeling of mixed language speech. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 5004–5007. IEEE.

Q. Zhang, J. Pan, and Y. Yan. 2008. Mandarin-English bilingual speech recognition for real world music retrieval. In *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pages 4253–4256. IEEE.