Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2013, Article ID 898714, 7 pages
http://dx.doi.org/10.1155/2013/898714

Research Article

Identification of Code-Switched Sentences and Words Using Language Modeling Approaches

Liang-Chih Yu,[1] Wei-Cheng He,[1] Wei-Nan Chien,[1] and Yuen-Hsien Tseng[2]

[1] *Department of Information Management, Yuan Ze University, 135 Yuan-Tung Road, Chung-Li 32003, Taiwan*
[2] *Information Technology Center, National Taiwan Normal University, 162 Heping East Road, Section 1, Taipei 10610, Taiwan*

Correspondence should be addressed to Liang-Chih Yu; lcyu@saturn.yzu.edu.tw

Received 3 May 2013; Revised 22 July 2013; Accepted 19 August 2013

Academic Editor: Orwa Jaber Housheya

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 show that the language modeling approach achieved an F-measure of 80.43% and an accuracy of 79.01% for detecting Mandarin-Taiwanese code-switched sentences. For the identification of code-switched words, the word-based and POS-based models, respectively, achieved F-measures of 41.09% and 53.08%.

1. Introduction

Increasing globalism and multilingualism has significantly increased demand for multilingual services in current intelligent systems [1]. 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 [2–6]. Table 1 lists several definitions of code-switching described in previous studies.

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 [7, 8]. 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% of the population, respectively [9]. 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-switching language processing, focusing on code-switching between Mandarin and Taiwanese. We propose a language modeling approach

Table 1: Definitions of code-switching.

Study	Definition
Hymes et al. [2]	A common term for alternative use of two or more languages, varieties of a language, or even speech styles
Hoffmann [3]	The alternate use of two languages or linguistic varieties within the same utterance or during the same conversation
Myers-Scotton [4]	The use of two or more languages in the same conversation, usually within the same conversational turn or even within the same sentence of that turn

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 [9–14], language identification [15, 16], text-to-speech synthesis [17], and code-switching speech database creation [18]. Lyu et al. proposed a three-step data-driven phone clustering method to train an acoustic model for Mandarin, Taiwanese, and Hakka [9]. They also discussed the issue of training with unbalanced data. Wu et al. proposed an approach to segmenting and identifying mixed-language speech utterances [10]. 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. Chan et al. developed a Cantonese-English mixed-language speech recognition system, including acoustic modeling, language modeling, and language identification algorithms [11]. Hong et al. 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) [12]. Ahmed and Tan proposed a two-pass code-switching speech recognition framework: automatic speech recognition and rescoring [13]. Vu et al. recently developed a speech recognition system for code-switching in conversational speech [14]. For language identification, Lyu et al. proposed a word-based lexical model integrating acoustic, phonetic, and lexical cues to build a language identification system [15]. Yeong and Tan proposed the use of morphological structures and sequence of the syllable for language identification from Malay-English code-switching sentences [16]. For speech synthesis, Qian et al. developed a text-to-speech system that can generate Mandarin-English mixed-language utterances [17].

Research on code-switching and multilingual language processing included applications of text mining [19–22], information retrieval [23–25], ontology-based knowledge management [26], and unknown word extraction [27]. For text mining, Seki et al. extracted opinion holders for discriminating opinions that are viewed from different perspectives (author and authority) in both Japanese and English [19]. Yang et al. used self-organizing maps to cluster multilingual documents [20]. A multilingual Web directory was then constructed to facilitate multilingual Web navigation. Zhang et al. addressed the problem of multilingual sentence categorization and novelty mining on English, Malay, and Chinese sentences [21]. They proposed to first categorize similar sentences and then identify new information from them. De Pablo-Sánchez et al. devised a bootstrapping algorithm to acquire named entities and linguistic patterns from English and Spanish news corpora [22]. This lightly supervised method can acquire useful information from unannotated corpora using a small set of seeds provided by human experts. For information retrieval, Gey et al. pointed out several directions for cross-lingual information retrieval (CLIR) research [23]. Tsai et al. used the FRank ranking algorithm to build a merge model for multilingual information retrieval [24]. Jung discovered useful multilingual tags annotated in social texts [25]. He then used these tags for query expansion to allow users to query in one language but obtain additional information in another language. For other application domains, Segev and Gal proposed an ontology-based knowledge management model to enhance portability and reduce costs in multilingual information systems deployment [26]. Wu et al. proposed the use of mutual information and entropy to extract unknown words from code-switched sentences [27].

3. Language Modeling Approach

Language modeling approaches have been successfully used in many applications, such as grammar error correction [28], code-switching language processing [29], and lexical substitution [30–32]. For our task, a code-switched sentence generally has a higher probability of being found in a code-switching language model than in a noncode-switching one. Thus, we built code-switching and noncode-switching language models to compare their respective probabilities of identifying code-switched sentences and code-switched words within the sentences. Figure 1 shows the system framework. First, a corpus of code-switched and noncode-switched sentences is collected to build the respective code-switching and noncode-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 noncode-switching one to determine whether or not the test sentence is code-switched. To identify code-switched words within the sentences,

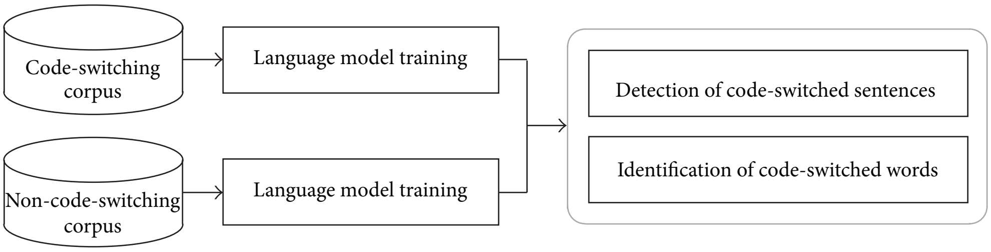

Code-switching corpus
Language model training
Non-code-switching corpus
Language model training
Detection of code-switched sentences
Identification of code-switched words

Figure 1: Framework of identification of code-switched sentences and words in 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 noncode-switching one to verify whether the nth word in the given sentence is a code-switched word.

3.1. Corpus Collection. A noncode-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 for Computational Linguistics and Chinese Language Processing (ACLCLP) as the noncode-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 word segmentation system developed by the Academia Sinica, Taiwan (http://ckipsvr.iis.sinica.edu.tw/) [33, 34]. 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 nth word based on the previous $n-1$ words using a probability function $P(w_n \mid w_1 \cdots w_{n-1})$. Given a sentence $S = w_1 \cdots w_k$, the noncode-switching n-gram language model is defined as

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

where $P(w_i \mid w_{i-1} \cdots w_{i-n+1})$ is estimated by

$$P(w_i \mid w_{i-1} \cdots w_{i-n+1}) = \frac{C(w_i \cdots w_{i-n+1})}{C(w_{i-1} \cdots w_{i-n+1})}, \tag{2}$$

where $C(\cdot)$ denotes the frequency counts of the n-grams retrieved from the noncode-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, that is, $P(cs_n \mid w_1 \cdots w_{n-1})$, defined as

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

where $P(w_i \mid w_{i-1} \cdots w_{i-n+1})$ is estimated by

$$P(cs_i \mid w_{i-1} \cdots w_{i-n+1}) = \frac{C(cs_i \cdots w_{i-n+1})}{C(cs_{i-1} \cdots w_{i-n+1})}. \tag{4}$$

To estimate $P(cs_n \mid w_1 \cdots w_{n-1})$, 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 $C(cs_i \cdots w_{i-n+1})$ 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,

$$c = \frac{P_{cs}(S)}{P_{\overline{cs}}(S)}. \tag{5}$$

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 noncode-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 (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,

$$cs^{*} = \underset{i}{\operatorname{argmax}}\, P(cs_i \mid w_{i-1} \cdots w_{i-n+1}), \tag{6}$$

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,

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

where $P^*(cs_i \mid w_{i-1} \cdots w_{i-n+1})$ represents the probability of the best hypothesis in the code-switching corpus and $P(w_i \mid w_{i-1} \cdots w_{i-n+1})$ represents its probability in the noncode-switching corpus. The best hypothesis cs^* is accepted if its probability in the code-switching corpus is greater than that in the noncode-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 both Mandarin-Taiwanese and Mandarin-English code-switched sentences and words within the sentences.

4.1. Experimental Setup. The test set included 393 sentences of which 131 were Mandarin only (i.e., noncode-switched), while another 131 were Mandarin sentences containing Taiwanese words, and the remaining 131 were Mandarin sentences containing English words. For the evaluation of Mandarin-Taiwanese sentences, n-gram models for both code-switching and noncode-switching were trained using the SRILM toolkit [35] with $n = 2$ and 3 (i.e., bigram and trigram). For the evaluation of Mandarin-English sentences, the CKIP word segmentation system [33, 34] was used because it can associate a POS tag "FW" to English words/characters within the sentences. 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 as $(2 \times \textit{recall} \times \textit{precision})/(\textit{recall} + \textit{precision})$. 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

4.2.1. Evaluation on Mandarin-Taiwanese Code-Switched Sentences. To identify Mandarin-Taiwanese code-switched sentences, the code-switching and noncode-switching bigram/trigram language models were used to determine whether or not each test sentence features code-switching (5), with results presented in Table 2. The bigram language model correctly identified 113 code-switched sentences and 94 noncode-switched sentences, thus yielding 86.26% (113/131) recall, 75.33% (113/150) precision, 80.43% F-measure, and 79.01% (207/262) accuracy. The trigram language model, however, did not outperform the bigram model, possibly due to the data sparseness problem caused by a lack of sufficient training data for building the trigram language model.

Table 2: Results of the identification of Mandarin-Taiwanese code-switched sentence.

Methods	Recall	Precision	F-measure	Accuracy
Bi-gram	86.26%	75.33%	80.43%	79.01%
Tri-gram	77.86%	62.20%	69.15%	65.27%

Table 3: Results of code-switched word identification in Mandarin-Taiwanese code-switched sentences.

Methods		Recall	Precision	F-measure
Random	Top 1	17.65%	18.32%	17.98%
	Top 2	31.62%	16.41%	21.61%
	Top 3	49.26%	17.05%	25.33%
Word bigram	Top 1	40.46%	41.73%	41.09%
	Top 2	60.31%	31.85%	41.69%
	Top 3	74.81%	27.53%	40.25%
Word trigram	Top 1	14.50%	15.83%	15.14%
	Top 2	35.88%	20.35%	25.97%
	Top 3	55.73%	22.53%	32.09%
POS bigram	Top 1	42.75%	43.08%	42.91%
	Top 2	67.94%	35.18%	46.35%
	Top 3	82.44%	29.35%	43.29%
POS trigram	Top 1	52.67%	53.49%	53.08%
	Top 2	73.28%	39.34%	51.20%
	Top 3	83.97%	31.98%	46.32%

To identify code-switched words in Mandarin-Taiwanese code-switched sentences, all word bigrams and trigrams in each test sentence were first ranked according to their probabilities. The top N word bigrams/trigrams were then selected as candidates for further verification using (7). For instance, top 1 means that the bigram/trigram with the highest probability in a given test sentence is considered a candidate. If the candidate n-gram is accepted by the verification method, then the position indicated by the n-gram will be considered a foreign word. Similarly, top 2 means that the method can propose two candidates for verification. To examine the effect of the data sparseness problem, we used the part-of-speech (POS) tags of words to build additional POS bigram/trigram models from the code-switching corpus. In addition to the word/POS n-gram models, we also implemented a baseline system to randomly guess the positions of code-switched words in the sentences, and the top N, herein, means that the system can randomly propose N candidate positions. Table 3 shows the results for the identification of code-switched words.

Table 4: Results of code-switched word identification in Mandarin-English code-switched sentences.

Methods	Recall	Precision	F-measure
CKIP (FW)	94.70%	95.33%	95.02%
Random (top 1)	17.88%	20.61%	19.15%
Random (top 2)	37.09%	21.37%	27.12%
Random (top 3)	48.34%	18.58%	26.84%

The results show that the F-measure of the baseline system (Random) was only around 18~25%, indicating that identifying code-switched words is more difficult than identifying code-switched sentences. In addition, the proposed word/POS n-gram models significantly outperformed Random. For the word-based n-gram models, the word bigram model achieved an F-measure of around 41%, which was much better than that of both the word trigram model and Random. Once the POS tags were used to build the language models, both the POS bigram and trigram models outperformed their corresponding word-based models in terms of F-measure, as well as for recall and precision. This finding indicates that training with the POS tags can reduce the impact of the data sparseness problem. In addition, as shown in Figure 2, the accuracy improvement derived from the trigram model was significantly greater than that from the bigram model, because the trigram model tends to suffer from a more serious data sparseness problem than the bigram model when training data is insufficient. Overall, the best performance of the POS n-gram models was achieved at an F-measure of 53.08% (POS trigram, top 1).

Code-switched word identification can also be evaluated by allowing the methods to propose more than one candidate, that is, top 1 to top 3. Table 3 shows that, with more candidates included for verification, more code-switched words were correctly identified, thus dramatically increasing the recall of all methods, but at the cost of reduced precision. Overall, the F-measure of top 2 was increased for all methods except for the POS trigram, but for top 3, increasing the number of candidates only increased the F-measure of Random and word trigram.

4.2.2. Evaluation on Mandarin-English Code-Switched Sentences. To identify code-switched words in Mandarin-English code-switched sentences, the words associated with the POS tag "FW" (representing a foreign word) by the CKIP word segmentation system were proposed as the answers. The Random system was also implemented to guess the English words in the test sentences. Table 4 shows the comparative results. As expected, the CKIP word segmentation system can provide very precise information for identifying English words in sentences, thus yielding very good performance. Actually, the CKIP system has been under development for over ten years and is still updated periodically. For the Random system, the F-measure was around 19~27% which was similar to that (18~25%, Table 3) for code-switched word identification in Mandarin-Taiwanese code-switched sentences.

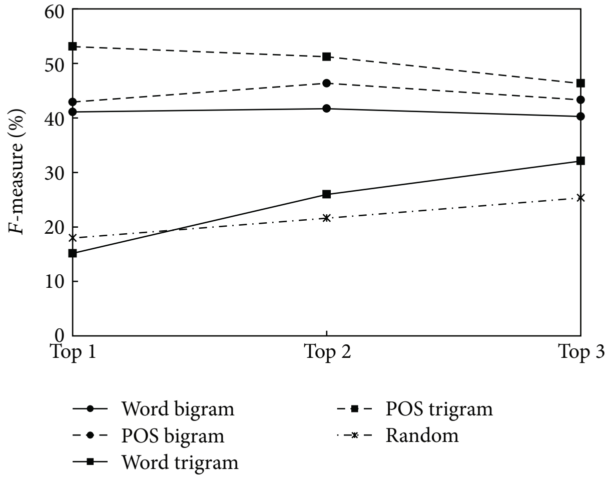


Figure 2: Comparative results of top N performance on code-switched word identification in Mandarin-Taiwanese code-switched sentences.

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 an F-measure of 80.43% and an accuracy of 79.01% for the detection of Mandarin-Taiwanese code-switched sentences. For the identification of code-switched words in Mandarin-Taiwanese code-switched sentences, the POS n-gram models outperformed the word n-gram models, mainly because of the reduced impact of the data sparseness problem. The highest F-measures (top 1) for the word-based and POS-based models were 41.09% and 53.08%, respectively. For code-switched word identification in Mandarin-English code-switched sentences, the CKIP word segmentation system achieved very high performance (95.02% F-measure).

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.

Acknowledgments

This research was partially supported by the National Science Council (NSC), Taiwan, under Grant nos. NSC 99-2221-E-155-036-MY3, NSC 100-2511-S-003-053-MY2, and NSC 102-2221-E-155-029-MY3, and the "Aim for the Top University Project" of National Taiwan Normal University (NTNU), sponsored by the Ministry of Education, Taiwan. The authors are also grateful to the support of the International Research-Intensive Center of Excellence Program of NTNU and NSC, Taiwan, under Grant no. NSC 102-2911-I-003-301.

References

[1] P. Fung and T. Schultz, "Multilingual spoken language processing," *IEEE Signal Processing Magazine*, vol. 25, no. 3, pp. 89–97, 2008.

[2] D. Hymes, "The ethnography in speaking," in *Anthropology and Man Behaviour*, T. Gladwin and W. C. Sturtevant, Eds., Anthropology Society of Washington, Washington, DC, USA, 1962.

[3] C. Hoffmann, *An Introduction to Bilingualism*, Longman, London, UK, 1991.

[4] C. Myers-Scotton, *Social Motivations for Code Switching: Evidence from Africa*, Oxford University Press, New York, NY, USA, 1993.

[5] M. O. Ayeomoni, "Code-switching and code-mixing: style of language use in childhood in yoruba speech community," *Nordic Journal of African Studies*, vol. 15, no. 1, pp. 90–99, 2006.

[6] Y. Liu, "Evaluation of the matrix language hypothesis: evidence from Chinese-English code-switching phenomena in blogs," *Journal of Chinese Language and Computing*, vol. 18, no. 2, pp. 75–92, 2008.

[7] I. Ipsic, N. Pavesic, F. Mihelic, and E. Noth, "Multilingual spoken dialog system," in *Proceedings of the IEEE International Symposium on Industrial Electronics (ISIE '99)*, pp. 183–187, Bled, Slovenia, July 1999.

[8] H. Holzapfel, "Building multilingual spoken dialogue systems," *Archives of Control Sciences*, vol. 15, no. 4, pp. 555–566, 2005.

[9] D. C. Lyu, C. N. Hsu, Y. C. Chiang, and R. Y. Lyu, "Acoustic model optimization for multilingual speech recognition," *International Journal of Computational Linguistics and Chinese Language Processing*, vol. 13, no. 3, pp. 363–386, 2008.

[10] C.-H. Wu, Y.-H. Chiu, C.-J. Shia, and C.-Y. Lin, "Automatic segmentation and identification of mixed-language speech using delta-BIC and LSA-based GMMs," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 1, pp. 266–275, 2006.

[11] J. Y. C. Chan, P. C. Ching, T. Lee, and H. Cao, "Automatic speech recognition of Cantonese-English code-mixing utterances," in *Proceedings of the INTERSPEECH and 9th International Conference on Spoken Language Processing (INTERSPEECH—ICSLP '06)*, pp. 113–116, Pittsburgh, Pa, USA, September 2006.

[12] W. T. Hong, H. C. Chen, I. B. Liao, and W. J. Wang, "Mandarin/English mixed-lingual speech recognition system on resource-constrained platforms," in *Proceedings of the 21st Conference on Computational Linguistics and Speech Processing (ROCLING '09)*, pp. 237–250, Taichung, Taiwan, September 2009.

[13] B. H. A. Ahmed and T. P. Tan, "Automatic speech recognition of code switching speech using 1-best rescoring," in *Proceedings of the International Conference on Asian Language Processing (IALP '12)*, pp. 137–140, Hanoi, Vietnam, November 2012.

[14] N. T. Vu, D. C. Lyu, J. Weiner et al., "A first speech recognition system for Mandarin-English code-switch conversational speech," in *Proceedings of the 37th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '12)*, pp. 4889–4892, Kyoto, Japan, March 2012.

[15] D.-C. Lyu, C.-L. Zhu, R.-Y. Lyu, and M.-T. Ko, "Language identification in code-switching speech usingword-based lexical model," in *Proceedings of the 7th International Symposium on Chinese Spoken Language Processing (ISCSLP '10)*, pp. 460–464, Tainan, Taiwan, December 2010.

[16] Y.-L. Yeong and T.-P. Tan, "Applying grapheme, word, and syllable information for language identification in code switching sentences," in *Proceedings of the International Conference on Asian Language Processing (IALP '11)*, pp. 111–114, Penang, Malaysia, November 2011.

[17] Y. Qian, H. Liang, and F. Soong, "A cross-language state sharing and mapping approach to bilingual (Mandarin-English) TTS," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 6, pp. 1231–1239, 2009.

[18] H.-P. Shen, C.-H. Wu, Y.-T. Yang, and C.-S. Hsu, "CECOS: a Chinese-English code-switching speech database," in *Proceedings of the International Conference on Speech Database and Assessments (Oriental COCOSDA '11)*, pp. 120–123, Hsinchu, Taiwan, October 2011.

[19] Y. Seki, N. Kando, and M. Aono, "Multilingual opinion holder identification using author and authority viewpoints," *Information Processing and Management*, vol. 45, no. 2, pp. 189–199, 2009.

[20] H.-C. Yang, H.-W. Hsiao, and C.-H. Lee, "Multilingual document mining and navigation using self-organizing maps," *Information Processing and Management*, vol. 47, no. 5, pp. 647–666, 2011.

[21] Y. Zhang, F. S. Tsai, and A. T. Kwee, "Multilingual sentence categorization and novelty mining," *Information Processing and Management*, vol. 47, no. 5, pp. 667–675, 2011.

[22] C. de Pablo-Sánchez, I. Segura-Bedmar, P. Martínez, and A. Iglesias-Maqueda, "Lightly supervised acquisition of named entities and linguistic patterns for multilingual text mining," *Knowledge and Information Systems*, vol. 35, no. 1, pp. 87–109, 2013.

[23] F. C. Gey, N. Kando, and C. Peters, "Cross-language information retrieval: the way ahead," *Information Processing and Management*, vol. 41, no. 3, pp. 415–431, 2005.

[24] M.-F. Tsai, H.-H. Chen, and Y.-T. Wang, "Learning a merge model for multilingual information retrieval," *Information Processing and Management*, vol. 47, no. 5, pp. 635–646, 2011.

[25] J. J. Jung, "Cross-lingual query expansion in multilingual folksonomies: a case study on flickr," *Knowledge-Based Systems*, vol. 42, pp. 60–67, 2013.

[26] A. Segev and A. Gal, "Enhancing portability with multilingual ontology-based knowledge management," *Decision Support Systems*, vol. 45, no. 3, pp. 567–584, 2008.

[27] Y. L. Wu, C. W. Hsieh, W. H. Lin, C. Y. Liu, and L. C. Yu, "Unknown word extraction from multilingual code-switching sentences," in *Proceedings of the 23rd Conference on Computational Linguistics and Speech Processing (ROCLING '11)*, pp. 349–360, Taipei, Taiwan, September 2011.

[28] C.-H. Wu, C.-H. Liu, M. Harris, and L.-C. Yu, "Sentence correction incorporating relative position and parse template language models," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 18, no. 6, pp. 1170–1181, 2010.

[29] L. C. Yu, W. C. He, and W. N. Chien, "A language modeling approach to identifying code-switched sentences and words," in *Proceedings of the 2nd CIPS-SIGHAN Joint Conference on Chinese Language Processing (CLP '12)*, pp. 3–8, Tianjin, China, December 2012.

[30] L.-C. Yu, C.-H. Wu, R.-Y. Chang, C.-H. Liu, and E. Hovy, "Annotation and verification of sense pools in OntoNotes," *Information Processing and Management*, vol. 46, no. 4, pp. 436–447, 2010.

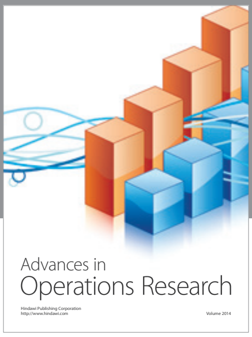
[31] A. Islam and D. Inkpen, “Near-synonym choice using a 5-gram language model,” *Research in Computing Science:*, vol. 46, pp. 41–52, 2010.

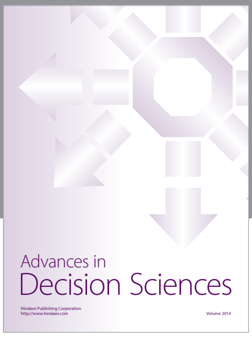
[32] L. C. Yu and W. N. Chien, “Independent component analysis for near-synonym choice,” *Decision Support Systems*, vol. 55, no. 1, pp. 146–155, 2013.

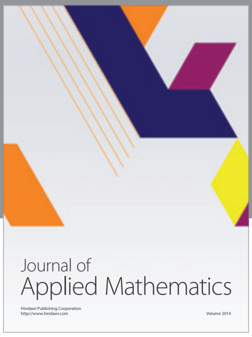
[33] W. Y. Ma and K. J. Chen, “Introduction to CKIP Chinese word segmentation system for the first international Chinese word segmentation bakeoff,” in *Proceedings of the 2nd ACL SIGHAN Workshop on Chinese Language Processing (CLP ’03)*, pp. 168–171, Sapporo, Japan, July 2003.

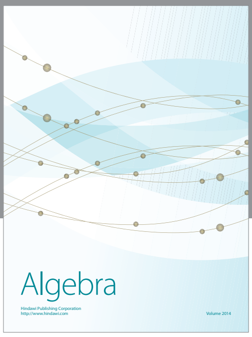
[34] W. Y. Ma and K. J. Chen, “A Bottom-up merging algorithm for Chinese unknown word extraction,” in *Proceedings of the 2nd ACL SIGHAN Workshop on Chinese Language Processing (CLP ’03)*, pp. 31–38, Sapporo, Japan, July 2003.

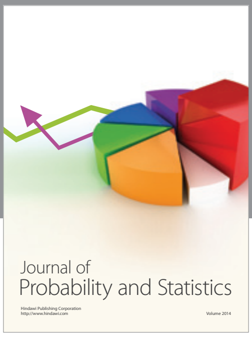
[35] A. Stolcke, “SRILM—an extensible language modeling toolkit,” in *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP ’02)*, pp. 901–904, Denver, Colo, USA, September 2002.


Advances in
Operations Research

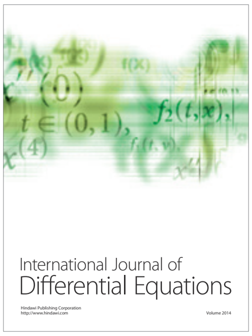

Advances in
Decision Sciences


Journal of
Applied Mathematics

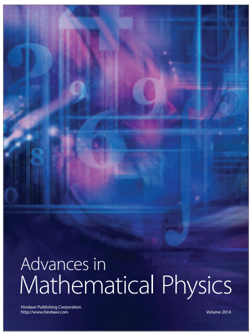

Algebra

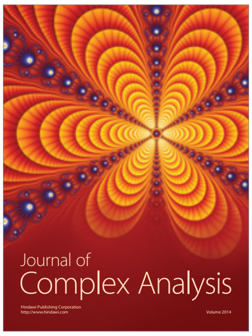

Journal of
Probability and Statistics

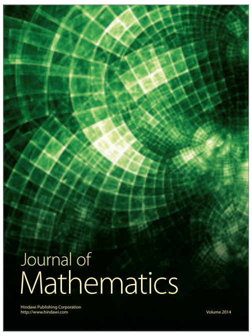
The Scientific
World Journal

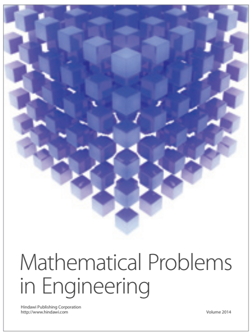

International Journal of
Differential Equations

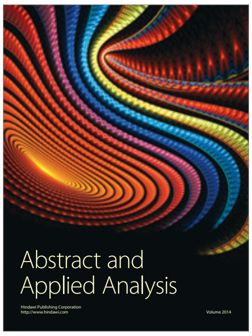
International Journal of
Combinatorics

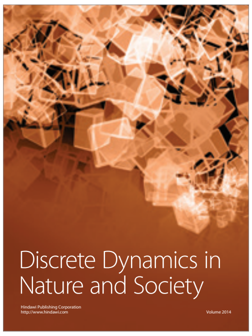

Advances in
Mathematical Physics

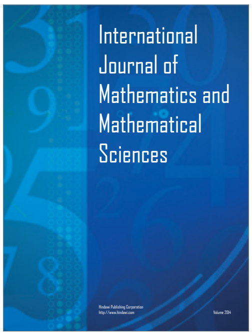

Journal of
Complex Analysis


Journal of
Mathematics


Mathematical Problems
in Engineering

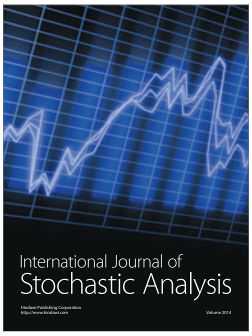

Abstract and
Applied Analysis


Discrete Dynamics in
Nature and Society


International
Journal of
Mathematics and
Mathematical
Sciences

Journal of
Discrete Mathematics

Journal of
Function Spaces


International Journal of
Stochastic Analysis

Journal of
Optimization