JOURNAL OF INFORMATION SCIENCE AND ENGINEERING 35, 597-610 (2019)
DOI: 10.6688/JISE.201905_35(3).0007

Combining Mutual Information and Entropy for Unknown Word Extraction from Multilingual Code-Switching Sentences*

CHENG-WEI LEE, YI-LUN WU AND LIANG-CHIH YU
Department of Information Management
Yuan-Ze University
Taoyuan, 320 Taiwan
E-mail: {s989206; s986301}@mail.yzu.edu.tw; lcyu@saturn.yzu.edu.tw

In multilingual environments, a single statement may include content from more than one language, a phenomenon known as code-switching. Among speakers of Mandarin Chinese, code switching is a frequent occurrence in daily life, and this mixing of different languages poses serious challenges for language processing. This paper collects text corpora including code switching between Mandarin and English and Mandarin and Taiwanese, where Mandarin is the dominant language. Mutual information and entropy are then used as a basis for an algorithm to identify unknown words from multilingual texts which are then automatically referenced for multilingual inclusions. Experimental results show that the proposed method effectively filters unrelated new words, thus improving the accuracy of extracting unknown words.

Keywords: code switching, unknown word extraction, mutual information, entropy, natural language processing

1. INTRODUCTION

Speech and language processing play a critical role in human-machine interface applications. In recent years, speech and language processing technologies have improved significantly, and are now incorporated into a wide range of applications, including speech recognition and synthesis [1, 2], spoken dialog systems [3-5], voice activity detection [6-8], information retrieval [9, 10], question answering [11, 12], and sentiment analysis [13-15]. However, there are over 6,900 human languages in use worldwide [16] and increasing trends towards globalization and international exchange are driving increased demand for multilingual services such as in the hospitality industry and in emergency/ medical services. In addition, international companies now use multi-language customer service systems to provide customers around the world with phone-based support. Therefore, how current systems support multiple languages has emerged as a key challenge for speech and language processing technologies.

In a multilingual environment, a sentence may include content from more than one language, a phenomenon known as code-switching or language mixing [17, 18]. Code-switching frequently occurs in bilingual or multilingual areas where cultural and educa-

Received December 5, 2017; revised April 4, 2018; accepted May 16, 2018.
Communicated by Berlin Chen.
* This paper has been presented at the 23rd Conference on Computational Linguistics and Speech Processing (ROCLING 2011) held in Taipei, Taiwan, September 8-9, 2011.
* This paper was supported by the Ministry of Science and Technology, Taiwan, under Grant No. MOST 105-2221-E-155-059-MY2 and MOST 105-2218-E-006-028.

tional processes have exposed speakers of the locally dominant language to the dominant languages of globalization. Specific customs, occasions and interlocutors may induce speakers to engage in code-switching. Code-switching utterances are typically based on the locally dominant language (primary language), inserting words or phrases borrowed from the secondary language. In Taiwan, Mandarin Chinese is the dominant language, while Taiwanese is also frequently spoken. In addition, a strong emphasis on English language learning has steadily improved English language skills. As a result, Mandarin speakers frequently code-switch in Taiwanese or English, and this appears not only in conversation but in media, including newspapers, magazines and online resources. The following sentences are based in Mandarin, but include language items in English (E) or Taiwanese (T).

(E1) 兩岸 ECFA 即將進入正式協商。
(Formal negotiations on the cross-strait ECFA are about to begin.)
(E2) 享受樂活舒壓的 SPA 活動。
(Enjoy the LOHAS spa activities.)
(T1) 選情緊繃，候選人四處趴趴走拜票。
(The election is tight and the candidates are pa-pa-tsáu pleading for support.)
(T2) 這裡有一家很傳統的柑仔店。
(This is a very traditional kam-á-tiàm.)

Research on code-switching is primarily the domain of linguistics or sociology, and few studies have examined code-switching from the perspective of automatic speech processing. Corpora include EAT (http://www.aclclp.org.tw) and the Formosa Speech Database (ForSDat) [19]. Studies of speech recognition have examined mixed Mandarin and English speech [20, 21], Mandarin and Taiwanese [22, 23], and Cantonese and English [24, 25], along with the development of a multiple-language acoustic model [26], and speech synthesis of mixed Mandarin and English utterances [27, 28].

The above-mentioned multilingual code-switching research focuses on corpus establishment, language modeling and language recognition, lexicon augmentation, language modeling, spoken language recognition and other language processing issues, but speech processing technology also plays an important role. For example, in lexicon augmentation, the "ECFA" in (E1) is not found in an English dictionary, while the phrases "趴趴走" (pa-pa-tsáu), which means hang out elsewhere in (T1), and "柑仔店" (kam-á-tiàm), which means a grocery store in (T2), are not found in a Taiwanese dictionary. The presence of these unknown expressions will impact the effectiveness of the language model and subsequent voice recognition.

Previous studies on unknown word extraction can be generally divided into rule-based, statistical-based and learning-based methods. Rule-based methods typically use a set of hand-crafted morphological rules to detect unknown words and have been successfully applied for many languages such as Chinese [29] and Arabic [30] along with French medical words [31]. However, manually designing such morphological rules is labor- intensive and statistical-based and learning-based methods were proposed to automatically extract unknown words from text corpora. For example, pointwise mutual information (PMI) [32-34] and entropy [35] are commonly used statistical measures. PMI extracts unknown words based on co-occurrence frequency between them, while

entropy accomplishes this by considering the contextual distributions of words as an indicator of semantic integrity. The rule-based and statistical-based methods can further be combined to build a hybrid model [36, 37]. For learning-based methods, a series of conditional random fields (CRF) based [38, 39] and boosting-based [40] methods were proposed to detect unknown words. More recently, deep neural network models have been developed to deal with the unknown word issue in various tasks such as parsing [41], discourse [42] and question answering [43]. For our task, the supervised machine learning methods are not suitable because of the difficulty of collecting a large dataset of mixed Mandarin and Taiwanese sentences for model training. Therefore, this study uses a statistical-based method that combines the mutual information and entropy to extract unknown words from low-resource mixed Mandarin and Taiwanese sentences.

The remainder of this paper is organized as follows: Section 2 introduces the multilingual corpora and analysis results. Section 3 explains the algorithm used to extract unknown words. Section 4 provides experimental results, and Section 5 presents conclusions.

2. MULTILINGUAL CORPORA COLLECTION AND ANALYSIS

2.1 Multilingual Corpora Collection

This paper considers code-switching between three languages – Mandarin Chinese, Taiwanese and English. It summarizes the characteristics of the multilingual corpora including sentence or grammatical structures prone to inclusion, facilitating follow up on unknown terms, and describes the design of the language model and language processing module. A relatively large number of Mandarin-English and Mandarin-Taiwanese corpora have been developed from online BBS sources, blogs and discussion forums focusing on current affairs, tourism, cuisine and other topics.

2.2 Code-Switching Analysis

Taiwanese and English phrases can appear anywhere in a Mandarin code-switched sentence, and there are no actual rules regarding code-switched content positioning. However, several studies have found consistent structures and patterns in Mandarin-English code-switched sentences [17, 18]. Based on this pattern analysis, we collected Mandarin/Taiwanese/English corpora from online sources, with partial results shown in Table 1.

To further analyze parts of speech and patterns in these mixed phrases, we randomly selected 500 online news articles for manual segmentation of Taiwanese and English phrases, with results shown in Tables 2 and 3. Statistical results for Table 2 indicate that about 90% of the English content appearing in the Mandarin texts are nouns, and are typically personal or geographical names, while verbs account for only about 10% of such content. This indicates that verbs are typically expressed in L1, with speakers resorting to the original English names for companies, restaurants and geographical locations, as using these English terms will not result in communication problems. These terms are not frequently translated into Chinese, thus using the original English can actually improve comprehension. The distribution of Taiwanese phrases within Mandarin

sentences contrasts sharply with that of English, with the statistical results in Table 3 indicating that about 70% of Taiwanese code-switching consists of verbs, as opposed to about 24% for nouns. Aside from the speaker's personal habits, these phrases are popular semantic expressions. In addition, Taiwanese is used relatively infrequently for the formal names of places and organizations, on the other hand, Taiwanese noun and verb.

Table 1. Taiwanese and English patterns in the Mandarin corpora.

No.	Pattern	Example
1	Degree adverb + Adj. phrase	這道料理**非常 smooth** 入口即化 (The flavor of this dish is **very smooth**.) 這家餐廳的老闆**很阿莎力** (The owner of this restaurant is **very a-sha-li**.)
2	Adj. phrase + 的(--ê)	這裡都沒有可以 **shopping 的**地方 (There's no place for **shopping** around here.) 終於體驗到甚麼是**足感心的**服務了 W(e finally got some **tsiok-kám-sim --ê** service.)
3	的(--ê) + Noun phrase	紐約**的 pizza**，單片就幾乎比臉大 (Each slice of New York style **pizza** is about as big as your face.) 好想念劉文聰**的番仔火**跟雞蛋糕啊！ (I really miss Liu Wencong**'s huan-á-hué** egg cake!)
4	Pattern 1 + Pattern 2 Pattern 2 + Pattern 3 Pattern 1 + Pattern 2 + Pattern 3	介紹你一家我**很尬意的**火鍋店 (I'll take you to my **favorite** hot pot restaurant.) 哪一間 **hotel 的 view** 最棒？ (Which **hotel** has the best **view**?) 這附近有一些**很古早的柑仔店** (This area has some **very old kam-á-tiàm**.)
5	Quantity pronoun + Noun phrase	我們還有**一些 issue** 要解決 (We still have **a few issues** to resolve. 這夜市有賣**很多賊仔貨** (This night market sells **many tsha t-á-huè**.)
6	Noun phrase + Location noun	我們約在 **lobby 旁**的水池見面 (We'll meet **by** the fountain in the **lobby**.) **烘爐地上**有一尊超級大的土地公神像 (**Hang-lôo-tè** has a very large status of the God of Wealth.)

Table 2. Distribution of English expressions by part of speech and type in the Mandarin-English corpus.

Part of Speech		Num.	Proportion		Examples
Nouns	Personal Names	89	25.14%	**90.4%**	John Culver, Kobe, Paul Hertz
	Place Names	80	22.60%		Boston, London, Paris
	Organizations	70	19.77%		NASA, NIKE, NHK, LV, Sony
	Units of Measurement	14	3.95%		cm, GHz, kg
	Food Items	13	3.67%		bagel, coffee, salad
	Other	54	15.25%		cartoon, CPR, I-phone, MSN, MVP
Verbs	–	34	9.6%	**9.6%**	call-in, DIY, po, shopping
Total		**354**	**100%**		

Table 3. Distribution of Taiwanese expressions by part of speech and type in the Mandarin-Taiwanese corpus.

POS	Count	Proportion	Example
Noun	17	23.61%	運將 (ūn-tsiong), 囡仔 (gín-á) , 天公伯 (thinn-kong-peh)
Verb	50	69.44%	假仙 (ké-sian), 挫咧等 (tshò-leh-tán), 阿莎力(a-sha-li)
Adverb	2	2.78%	攏 (lóng), 嘛 (mā)
Interrogative	3	4.17%	按怎 (án-tsuánn), 啥 (siánn)
Total	**72**	**100%**	

3. EXTRACTING UNKNOWN WORDS

Many of the phrases found in the multilingual corpora are not found in current dictionaries, and thus may not be accurately segmented. For example, "趴趴走" (pa-pa-tsáu) is not identified in the CKIP system (http://ckipsvr.iis.sinica.edu.tw) [44], thus another task for the current year is to identify unknown words in the multilingual corpora, particularly phrases. We propose an algorithm that extracts unknown words and segments them against language corpora. These unknown words are then cut into individual words or shorter phrases, and the frequent repetition of two adjacent words can be used as an important basis for detecting new words. We use the frequently cited PMI approach [32-34] to assess the collocation of two words, and filter adjacent words with higher PMI values as candidates for new words. PMI only considers two directly adjacent words. However, combinations of adjacent words are typically not new words. Therefore, in addition to PMI, we also use contextual entropy to filter unrelated new words, thus improving accuracy [35]. The process is shown in Fig. 1.

3.1 Word Segmentation

The collected multilingual corpora were segmented using Mandarin, Taiwanese and English dictionaries, along with CKIP to identify unknown words and phrases. CKIP annotates English words as "FW".

3.2 Mutual-Information-Based Word Aggregation

Among complete sentences in the language corpus, PMI score for any two adjacent words are calculated from left to right as follows:

$$PMI(w_i, w_j) = \log \frac{P(w_i, w_j)}{P(w_i)P(w_j)} = \log \frac{C(w_i, w_j) \bullet N}{C(w_i)C(w_j)}, \quad (1)$$

where $C(w_i, w_j)$ is the number of times w_i and w_j appear in the corpus and N is a constant that represents the number of words in the corpus. Here we use Google to query the number of returned files as $C(w_i, w_j)$ and $C(w_i)$ and $C(w_j)$. Because we are unable to precisely determine the value of N, we use 10^{12} as a substitute (the size of the value of N does not influence the PMI ordering results). The PMI values for the collocation of all adjacent words are then arranged in descending order to filter for new word candidates.

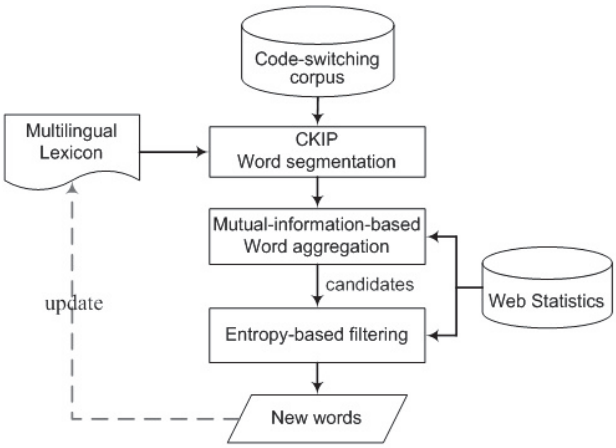


Fig. 1. Flowchart of the proposed unknown word extraction method.

3.3 Entropy-Based Filtering

The candidate words generated in the previous stage may include irrelevant terms, mainly because PMI only considers collocation between individual words and does not consider whether the term is a complete semantic unit. In general, it is more difficult to use statistical methods to directly assess the completeness of a denotation, but a semantically complete word has specific characteristics in pragmatic performance. That is, it can be used with many other words to form a larger unit. Therefore, if a small number of words are typically found adjacent to a given word, then the word in question is highly dependent on these words and thus provides more opportunities for combination into a new word. Given these characteristics, the semantic completeness of a word can be measured indirectly by the number and degree of dispersion of adjacent words as follows:

- Large number of adjacent words in an even distribution→semantically complete→standalone use
- Small number of adjacent words in a concentrated distribution→semantically incomplete→suitable for merging with adjacent words to create new words

For example, the CKIP segmentation result for "趴趴走" (pa-pa-tsáu) is "趴趴" (pa-pa) "走" (tsáu). Therefore we queried Google for "趴趴" (pa-pa), extracting 999 titles featuring "趴趴" (pa-pa) from the results to analyze the distribution of the left and right adjacent words. The results in Table 4 shows 212 and 58 different words respectively collocate to the left and right of "趴趴" (pa-pa), and "走" (tsáu) was found to follow "趴趴" (pa-pa) 336 times, while the distribution of the left adjacent words was considerably more even, indicating that "趴趴" (pa-pa) and "走" (tsáu) are highly correlated. Thus "趴趴" (pa-pa) is more likely to merge right to form a new word.

To systematically analyze the contextual distribution of each word, we convert the word frequency into a probabilistic representation by dividing the left and right word frequency by the total frequency. The entropy measure can then be used to indicate the degree of concentration in the left and right contextual distributions. Assume $RC(w_t)$ =

Table 4. Five most common collocations for "趴趴" (pa-pa).

趴趴 (pa-pa)			
Frequency	Prior	Following	Frequency
8	愛 (ài)	走 (tsáu)	336
5	軟 (nńg)	GO	42
5	Taiwan	造 (tsáu)	23
2	【	熊 (hîm)	21
1	「	照 (tsáu)	19

$\{w_1, \ldots, w_n\}$ indicates the right context of a word w_t which is combined with the word to its right, thus the entropy of the right context of w_t can be defined as:

$$H_{RC}(w_t) = -\sum_{w_i \in RC(w_t)} P(w_i)\log_2 P(w_i), \tag{2}$$

where $H_{RC}(w_t)$ is the right context entropy of w_t and $P(w_i) = C(w_i)/N$ is the probability of a given word appearing in the right context of w_t, where $C(w_i)$ is the number of times w_i appears to the right of w_t in the corpus, and N is total number of times that all adjacent words appear to the right of w_t. Similarly, $H_{LC}(w_t)$ is the left context entropy of w_t, and is calculated in the same way. Using the abovementioned calculation for entropy, the greater the concentration of contextual distribution, the smaller the entropy, while a more even distribution corresponds with greater entropy. Therefore, using entropy as an indicator of the semantic integrity of a word, we find:

- Large entropy→high semantic integrity→standalone use
- Small entropy→low semantic integrity→suitable for merging with adjacent words to create new words

Table 5 shows the right and left context entropy for 趴趴走 (pa-pa-tsáu). It shows that the right context entropy for 趴趴走 (pa-pa-tsáu) is relatively lower, indicating low semantic integrity, thus making it well suited for combination to the right adjacent走 (tsáu) to create a new word. The left and right entropy for 是 (sī) and 的 (--ê) are very high, indicating that these words have a high degree of semantic completeness and do not need to be merged with other words. We can see that, for two words to be merged, if one word to the right or left is incomplete, it is a good candidate for merging. Therefore, we define the pre-merging entropy of two words w_i and w_j as the minimum entropy value for the right context of w_i and the left context of w_j as follows.

$$H_{before}(w_i, w_j) = min(H_{RC}(w_i), H_{LC}(w_j)) \tag{3}$$

Table 5. Pre-merging left and right context entropy, using 趴趴走 (pa-pa-tsáu) and 是的 (sī--ê) as an example.

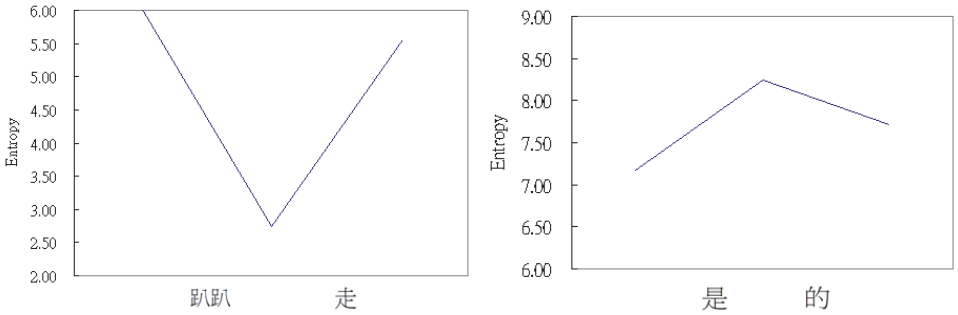
	$H_{LC}(w_t)$	$H_{RC}(w_t)$		$H_{LC}(w_t)$	$H_{RC}(w_t)$
趴趴 (pa-pa)	6.07	**2.75**	是 (sī)	7.17	8.24
走 (tsáu)	4.85	5.56	的 (--ê)	8.34	7.72
趴趴走 (pa-pa-tsáu)	6.07	5.56	是的 (sī--ê)	7.17	7.72

Table 5 also shows that the right and left contextual entropy of two merged words will increase because, after merging, the new word has higher semantic integrity. For example, the entropy of 趴趴走 (pa-pa-tsáu) is 2.75 before merging. After merging, if we extract the left context of 趴趴 (pa-pa) and the right context of 走 (tsáu), then the post-merging right and left contextual entropy are respectively 6.07 and 5.56, as shown in Fig. 2 (a). On the other hand, if two words with low semantic integrity are merged, the post-merging entropy does not increase significantly, and may even decrease, as shown in the example of 是 (sī) and 的 (--ê) in Fig. 2 (b). The change in entropy following the merger is an important indicator for determining whether the word is new or not. Therefore, we define the ratio of pre- and post-merging entropy as follows:

$$\lambda_{w_i w_j}^{LC} = \frac{H_{after}^{LC}(w_i, w_j)}{H_{before}(w_i, w_j)}, \tag{4}$$

$$\lambda_{w_i w_j}^{RC} \frac{H_{after}^{RC}(w_i, w_j)}{H_{before}(w_i, w_j)}, \tag{5}$$

where $\lambda_{w_i w_j}^{LC}$ and $\lambda_{w_i w_j}^{RC}$ are respectively the ratios of pre- and post-merger right and left contextual entropy for merging w_i and w_j, and $H_{after}^{LC}(w_i, w_j) = H_{LC}(w_i)$. That is, the left context of the extracted w_i is the post-merging left context of w_i and w_j. Similarly, $H_{after}^{RC}(w_i, w_j) = H_{RC}(w_j)$. Thus, when $\lambda_{w_i w_j}^{LC}$ and $\lambda_{w_i w_j}^{RC}$ are both greater than 1, the post-merging entropy is greater than the pre-merging entropy, and the two words can be considered for joining as a new word.



(a) (b)

Fig. 2. Pre- and post-merging entropy changes for two words.

4. EXPERIMENTAL RESULTS

4.1 Experimental Design

We randomly selected 500 news stories from Yahoo News, and then manually identified sentences with Taiwanese code-switching. We then used CKIP for word segmentation. If a code-switching sentence could not be successfully segmented, it indicated that

the phrase is a new Taiwanese word. Conversely, sentences in which the Taiwanese text was successfully segmented were not included in the experiment. According to this principle, we selected a total of 40 sentences including new Taiwanese words as the Mandarin-Taiwanese test sentence set. These sentences included a total of 200 adjacent words as candidates for merging, of which 41 were Taiwanese new words and were used as the ground truth in the experiment. Given the segmentation results for a test sentence, we first use the Google query results to calculate the PMI score for the left and right contextual entropy for any two adjacent words. The pre- and post-merging entropy ratio for adjacent words with higher PMI scores were then calculated, and $\lambda_{w_i w_j}^{LC}$ and $\lambda_{w_i w_j}^{RC}$ were taken as a threshold to determine whether or not the merged word was new. Finally, recall, precision and the F-measure were used to assess the effectiveness of the algorithm used to obtain the frequency of unknown words. The recall rate is used to determine how many of the 41 new words were correctly identified by the evaluation system, while precision was used to determine system accuracy in recommending new words, and F-measure provided a total assessment of the recall and precision rates, where 2 * recall * precision / (recall+precision). Generally speaking, raising PMI ensured the precision of the thresholds $\lambda_{w_i w_j}^{LC}$ and $\lambda_{w_i w_j}^{RC}$, but the recall rate may fall. On the other hand, setting the thresholds are set too low may result in additional new words being identified, which will reduce system precision.

4.2 Results

Table 6 shows the PMI scores and pre- and post-merger left and right contextual entropy ratios for selected candidate words. The results show that raising the PMI threshold filters out many non-new words, but that this results in new words such as 破糊糊 (phuà-kôo-kôo) (PMI too low) and 皮皮挫 (phî-phî-tshò) ($\lambda_{w_i w_j}^{LC}$ too low) being missed. This experiment finds that parameter settings of PMI = 3, $\lambda_{w_i w_j}^{LC}$ = 1.15, and $\lambda_{w_i w_j}^{RC}$ = 1.05 obtain optimal results of F-measure = 68.49%, Recall = 60.98%, and Precision = 78.13%. Table 7 shows the experimental results for Fig. 3 for simultaneously adjusting $\lambda_{w_i w_j}^{LC}$ and $\lambda_{w_i w_j}^{RC}$ (0.75~1.25).

Table 6. PMI scores and pre- and post-merger left and right contextual entropy ratios for candidate words.

Candidate word	PMI	λ^{LC}	λ^{RC}
瞭解(liáu-kái) 甚麼(siánn-mih)	12.69	1.11	1.00
碎碎(tshuì-tshuì) 念(liām)	12.33	4.14	5.29
皮皮(phî-phî) 挫(tshò)	11.42	0.92	1.07
囝(kiánn) 囡(gín)	11.07	1.29	1.38
才(tsâi) 發生(huat-sing)	3.90	0.79	0.96
好(hó) 山(suann)	0.80	1.13	1.18
破(phuà) 糊糊(kôo-kôo)	–2.14	1.33	1.14
就(tō) 是(sī)	–2.87	1.13	1.27

Table 7. The impact of different threshold values on the identification of unknown words (PMI = 3).

$\lambda_{w_i w_j}^{LC}$	$\lambda_{w_i w_j}^{RC}$	Recall	Precision	F-measure
1.15	0.75	0.6585	0.6000	0.6279
1.15	0.80	0.6585	0.6136	0.6353
1.15	0.85	0.6585	0.6136	0.6353
1.15	0.90	0.6585	0.6136	0.6353
1.15	0.95	0.6341	0.6341	0.6341
1.15	1.00	0.6098	0.6757	0.6410
1.15	**1.05**	**0.6098**	**0.7813**	**0.6849**
1.15	1.10	0.5610	0.7667	0.6479
1.15	1.15	0.5366	0.8148	0.6471
1.15	1.20	0.4390	0.9000	0.5902
1.15	1.25	0.4390	1.0000	0.6102

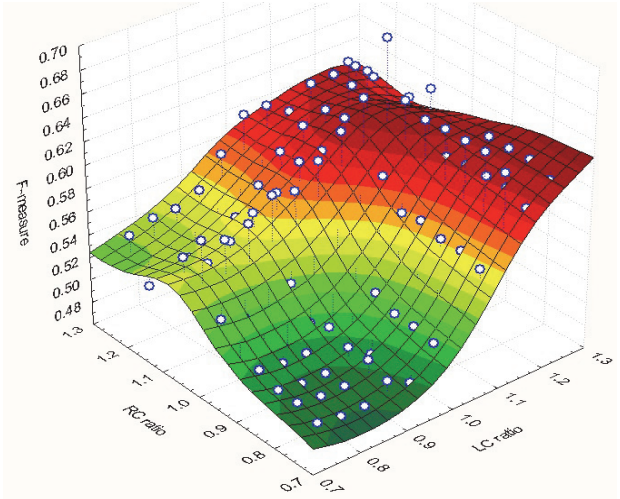


Fig. 3. Changes to entropy before and after merging two words.

5. CONCLUSIONS

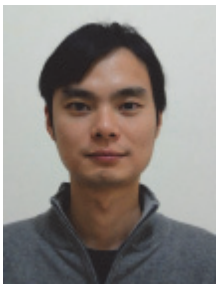
An algorithm is proposed to identify unknown words based on mutual information and entropy. The algorithm automatically identifies unknown words in multilingual corpora, using mutual information to calculate the cohesion between pairs of words, and selecting pairs with higher cohesion as candidates. Then, the entropy-based filtering mechanism filters irrelevant new words based on the distribution of the right and left context. Experimental results show that the proposed entropy-based method can improve the accuracy of filtering of unknown words. Future work will focus on further improving accuracy through the use of machine learning methods.

REFERENCES

1. M. Ostendorf and I. Bulyko, "The impact of speech recognition on speech synthesis," in *Proceedings of IEEE Workshop on Speech Synthesis*, 2002, pp. 99-106.
2. G. Saon and M. Picheny, "Recent advances in conversational speech recognition using convolutional and recurrent neural networks," *IBM Journal of Research and Development*, Vol. 61, 2017, pp. 1:1-1:10.
3. M. Korpusik and J. Glass, "Spoken language understanding in a nutrition dialogue system," *IEEE/ACM Transactions on Audio*, *Speech*, *and Language Processing*, Vol. 25, 2017, pp. 1450-1461.
4. G. Mesnil, Y. Dauphin, K. Yao, Y. Bengio, *et al.*, "Using recurrent neural networks for slot filling in spoken language understanding," *IEEE/ACM Transactions on Audio*, *Speech and Language Processing*, Vol. 23, 2015, pp. 530-539.
5. S. Young, M. Gašić, B. Thomson, and J. D. Williams, "Pomdp-based statistical spoken dialog systems: A review," in *Proceedings of the IEEE*, Vol. 101, 2013, pp. 1160-1179.
6. D. Dov, R. Talmon, and I. Cohen, "Kernel-based sensor fusion with application to audio-visual voice activity detection," *IEEE Transactions on Signal Processing*, Vol. 64, 2016, pp. 6406-6416.
7. R. J. Elton, P. Vasuki, and J. Mohanalin, "Voice activity detection using fuzzy entropy and support vector machine," *Entropy*, Vol. 18, 2016, pp. 298.
8. W. Q. Ong, A. W. C. Tan, V. V. Vengadasalam, C. H. Tan, and T. H. Ooi, "Real-time robust voice activity detection using the upper envelope weighted entropy measure and the dual-rate adaptive nonlinear filter," *Entropy*, Vol. 19, 2017, p. 487.
9. M. Larson and G. J. Jones, "Spoken content retrieval: A survey of techniques and technologies," *Foundations and Trends® in Information Retrieval*, Vol. 5, 2012, pp. 235-422.
10. L.-C. Yu, C.-H. Wu, and F.-L. Jang, "Psychiatric document retrieval using a discourse-aware model," *Artificial Intelligence*, Vol. 173, 2009, pp. 817-829.
11. R. Bakis, D. Connors, P. Dube, P. Kapanipathi, *et al.*, "Performance of natural language classifiers in a question-answering system," *IBM Journal of Research and Development*, Vol. 61, 2017, pp. 14:11-14:10.
12. Y. Liu, L. Wang, R. Chen, Y. Song, and Y. Cai, "A PUT-based approach to automatically extracting quantities and generating final answers for numerical attributes," *Entropy*, Vol. 18, 2016, p. 235.
13. S. M. Mohammad, "Sentiment analysis: Detecting valence, emotions, and other affectual states from text," *Emotion Measurement*, 2015, pp. 201-238.
14. X. Wang, Y. Liu, M. Liu, C. Sun, and X. Wang, "Understanding gating operations in recurrent neural networks through opinion expression extraction," *Entropy*, Vol. 18, 2016, p. 294.
15. J. Wang, L.-C. Yu, K. R. Lai, and X. Zhang, "Community-based weighted graph model for valence-arousal prediction of affective words," *IEEE/ACM Transactions on Audio*, *Speech*, *and Language Processing*, Vol. 24, 2016, pp. 1957-1968.
16. P. Fung and T. Schultz, "Multilingual spoken language processing," *IEEE Signal Processing Magazine*, Vol. 25, 2008, pp. 89-97.
17. L. Ge, "An investigation on English/Chinese code-switching in BBS in Chinese

alumni's community," MA Dissertation, Department of Applied Linguistics, University of Edinburgh, 2007.
18. 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, 2008, pp. 75-92.
19. R.-Y. Lyu, M.-S. Liang, and Y.-C. Chiang, "Toward constructing a multilingual speech corpus for Taiwanese (Minnan), Hakka, and Mandarin," *International Journal of Computational Linguistics and Chinese Language Processing*, Vol. 9, 2004, pp. 1-12.
20. C.-L. Huang and C.-H. Wu, "Generation of phonetic units for mixed-language speech recognition based on acoustic and contextual analysis," *IEEE Transactions on Computers*, Vol. 56, 2007, pp. 1225-1233.
21. 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, 2006, pp. 266-276.
22. W. Hong, H. Chen, I. Liao, and W. Wang, "Mandarin/English mixed-lingual speech recognition system on resource-constrained platforms," in *Proceedings of the 21st Conference on Computational Linguistics and Speech Processing*, 2009, pp. 237-250.
23. D.-C. Lyu, R.-Y. Lyu, Y.-C. Chiang, and C.-N. Hsu, "Speech recognition on code-switching among the Chinese dialects," in *Proceedings of IEEE International Conference on Acoustics*, *Speech and Signal Processing*, 2006, pp. 1105-1108.
24. J. Y. Chan, P. C. Ching, T. Lee, and H. Cao, "Automatic speech recognition of Cantonese-English code-mixing utterances," in *Proceedings of the 9th International Conference on Spoken Language Processing*, 2006, pp. 113-116.
25. J. Y. C. Chan, P. C. Ching, T. Lee, and H. M. Meng, "Detection of language boundary in code-switching utterances by bi-phone probabilities," in *Proceedings of the 4th International Symposium on Chinese Spoken Language Processing*, 2004, pp. 293-296.
26. C. M. White, S. Khudanpur, and J. K. Baker, "An investigation of acoustic models for multilingual code-switching," in *Proceedings of the 9th Annual Conference of the International Speech Communication Association*, 2008, pp. 2691-2694.
27. Y. Qian, H. Liang, and F. K. Soong, "A cross-language state sharing and mapping approach to bilingual (Mandarin-English) TTS," *IEEE Transactions on Audio*, *Speech*, *and Language Processing*, Vol. 17, 2009, pp. 1231-1239.
28. Y. Zhang and J. Tao, "Prosody modification on mixed-language speech synthesis," in *Proceedings of the 6th International Symposium on Chinese Spoken Language Processing*, 2008, pp. 1-4.
29. W.-Y. Ma and K.-J. Chen, "A bottom-up merging algorithm for Chinese unknown word extraction," in *Proceedings of the 2nd SIGHAN Workshop on Chinese Language Processing*, 2003, pp. 31-38.
30. L. Cahill, "A rule-based approach to unknown word recognition in Arabic," in *Proceedings of the 12th Meeting of the Special Interest Group on Computational Morphology and Phonology*, 2012, pp. 35-41.
31. N. Grabar and T. Hamon, "Understanding of unknown medical words," in *Proceed-*

ings of Biomedical NLP Workshop Associated with RANLP, 2017, pp. 32-41.
32. K. W. Church and P. Hanks, "Word association norms, mutual information, and lexicography," *Computational Linguistics*, Vol. 16, 1990, pp. 22-29.
33. J. Sourati, M. Akcakaya, J. G. Dy, T. K. Leen, and D. Erdogmus, "Classification active learning based on mutual information," *Entropy*, Vol. 18, 2016, pp. 51.
34. L.-C. Yu, J.-L. Wu, P.-C. Chang, and H.-S. Chu, "Using a contextual entropy model to expand emotion words and their intensity for the sentiment classification of stock market news," *Knowledge-Based Systems*, Vol. 41, 2013, pp. 89-97.
35. Z. Luo and R. Song, "An integrated method for Chinese unknown word extraction," in *Proceedings of the 3rd SIGHAN Workshop on Chinese Language Processing*, 2004, pp. 148-154.
36. M. Nuo, H. Liu, C. Long, and J. Wu, "Tibetan unknown word identification from news corpora for supporting lexicon-based Tibetan word segmentation," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, 2015, pp. 451-457.
37. K. Zhang, R. Wang, P. Xue, and M. Sun, "Extract Chinese unknown words from a large-scale corpus using morphological and distributional evidences," in *Proceedings of the 5th International Joint Conference on Natural Language Processing*, 2011, pp. 837-845.
38. F. Peng, F. Feng, and A. McCallum, "Chinese segmentation and new word detection using conditional random fields," in *Proceedings of the 20th International Conference on Computational Linguistics*, 2004, pp. 562-569.
39. X. Sun, D. Huang, and F. Ren, "Detecting new words from Chinese text using latent semi-CRF models," *IEICE Transactions on Information and Systems*, 2010, Vol. E93-D, 2010, pp. 1386-1393.
40. J. TeCho, C. Nattee, and T. Theeramunkong, "Boosting-based ensemble learning with penalty profiles for automatic Thai unknown word recognition," *Computers and Mathematics with Applications*, Vol. 63, 2012, pp. 1117-1134.
41. B. Do, I. Rehbein, and A. Frank, "What do we need to know about an unknown word when parsing German," in *Proceedings of the 1st Workshop on Subword and Character Level Models in NLP*, 2017, pp. 117-123.
42. S. Kobayashi, N. Okazaki, and K. Inui, "A neural language model for dynamically representing the meanings of unknown words and entities in a discourse," in *Proceedings of the 8th International Joint Conference on Natural Language Processing*, 2017, pp. 473-483.
43. J. Xu, J. Shi, Y. Yao, S. Zheng, B. Xu, and B. Xu, "Hierarchical memory networks for answer selection on unknown words," in *Proceedings of the 26th International Conference on Computational Linguistics*, 2016, pp. 2290-2299.
44. 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 SIGHAN Workshop on Chinese Language Processing*, 2003, pp. 168-171.



Cheng-Wei Lee (李振瑋) is currently a Ph.D. candidate in the Department of Information Management at Yuan Ze University in Taiwan. He received his M.S. degree in the Graduate Program of Social Informatics at Yuan Ze University. His research interests include natural language processing, educational data mining, computer-assisted language learning, and risk management.

Yi-Lun Wu (吳依倫) received her M.S. degree in the Department of Information Management at Yuan Ze University in Taiwan. Her research interests include natural language processing, text mining, and computer-assisted language learning.

Liang-Chih Yu (禹良治) is a Professor in the Department of Information Management at Yuan Ze University in Taiwan. He received his Ph.D. in Computer Science and Information Engineering from National Cheng Kung University in Taiwan. He was a visiting scholar at the Natural Language Group, Information Sciences Institute, University of Southern California (USC/ISI) from 2007 to 2008, and at DOCOMO Innovations for three months in 2018. He is currently Board Member and Convener of SIGCALL of the Association for Computational Linguistics and Chinese Language Processing (ACLCLP), and serves as an editorial board member of *International Journal of Computational Linguistics and Chinese Language Processing*. His research interests include natural language processing, sentiment analysis, computer-assisted language learning. His team has developed systems that ranked first in *IJCNLP 2017 Task 4: Customer Feedback Analysis*, and second in the recent *SemEval* and *BEA* shared task competitions. His research has appeared in leading journals including *Artificial Intelligence*, *ACM/IEEE Transactions*, *Decision Support Systems*, and *Knowledge-based Systems*, and has presented at leading conferences including *ACL*, *EMNLP*, *COLING*, *NAACL*, and elsewhere.