International Journal of Knowledge and Language Processing
Volume 5, Number 1, **2014**

www.ijklp.org
KLP International ©2014 ISSN 2191-2734
pp.27—40

Multiword Expression Extraction Based on Word Relativity

Liang Hu and Xuri Tang

Huazhong University of Science and Technology
No. 1037 Luoyu Road, Wuhan, China
husthuliang@gmail.com; xrtang@hust.edu.cn

Received March 2014; revised March 2014

ABSTRACT. *Multiword Expression has been an increasingly important issue in Natural Language Processing tasks. This paper proposes an algorithm for multiword expression extraction from bilingual corpus based on word relativity. The bilingual corpus is firstly aligned with GIZA++. Multiword expression candidates are then extracted on the basis of word relativity from the corpus and filtered by the use of word relativity and word alignment information. The results showed that our extraction system, combining linguistic knowledge with statistical information, performed better than purely statistical approach.*

Keywords: Multiword Expression Extraction; Word Relativity; Word Alignment

1. **Introduction.** A multiword expression (MWE) is a semantic unit consisted by several words and its syntactic or semantic properties cannot be derived from its parts [1]. According to [2], we can define MWEs roughly as "idiosyncratic interpretations that cross word boundaries (or spaces)". A MWE can be a compound, a fragment of a sentence, or a sentence. Examples for MWEs would be idioms as "kick the bucket", compound nouns as "telephone box" and "post office", verb phrases as "look sth. up" and proper names as "San Francisco". From the examples above, it can be known that a MWE may be more or less frozen. For example, the English MWE "kick the bucket" means to die rather than to hit a bucket with one's foot. In this example, the MWE is frozen, in the sense that no variation is possible. In another English MWE "throw somebody to the lions", the pattern "somebody" restricts the usage. The expression is half-frozen because a certain degree of variation is possible but everything is not possible. It is not possible for instance to say "to the three lions".

In practice, MWEs are commonly used in any filed of language. [3] estimates the number of MWEs in a speaker's lexicon as comparable to the number of single words. According to Fellbaum's statistics, 41% of vocabulary entries in WordNet v1.7 are MWEs [4]. Due to the idiosyncrasy, complexity and the high frequency of MWEs in natural languages, there is a growing awareness in the Natural Language Processing (NLP) community for the

problems they pose and MWE has become a research hotspot. In fact, in the NLP area, MWE serves as a basis for other NLP researches and applications, such as machine translation (MT), multilingual information retrieval, data mining and many others.

In recent years, seminars and workshops for MWE can been seen frequently in many large-scale academic conferences, which mainly focused on the basic issues such as definition, identification, disambiguation and application. Among all the tasks, identification and application of MWEs have got the main focus. For example, the identification of MWEs can greatly improve the efficiency and accuracy of many tasks like word segmentation, part-of-speech tagging, machine translation, and so on. In machine translation, accurate identification of MWEs from source language can help a lot in choosing the right translation equivalent in the target language, so as to avoid the unreadability or even ambiguity of translated sentences caused by separate translation of different single words. Therefore, identification and extraction of MWEs have played a vital role in NLP researches and applications. An efficient MWE extraction system will promote greatly many NLP researches like machine translation and computer aided translation, which are drawing more and more attention today. Both of them require the segmentation of sentences in order to obtain the sentence segments among which some are actually MWEs. So MWE extraction will help in segmenting sentences and in finding the translation equivalents of those segments. Because of the significance of MWE to NLP researches, a considerable amount of research has been devoted to this task by scholars working in this area.

2. **Literature Review.** Over the past years, a variety of methods for the automatic extraction of multiword expressions have been proposed and tested. Generally speaking, there are three categories of methods: (a) knowledge-based or symbolic approaches using parsers, lexicons and language filters; (b) statistical approaches based on frequency and co-occurrence affinity; (c) hybrid approaches combining different methods [5].

2.1. **Knowledge-based Approaches.** In practice, most knowledge-based or symbolic approaches use linguistic information to identify and extract MWEs. For example, in 2003, Piao et al. proposed an approach to MWE extraction using semantic field information. In their approach, multiword expressions depicting single semantic concepts are recognized using an English semantic tagger “UCREL Semantic Analysis System” (USAS) developed by Lancaster University [5]. Some other researchers also used semantic information to improve the performance of MWE extraction. Lexical resources and parsers are used to obtain better coverage of the lexicon in MWE extraction. For example, [6] used an English-Chinese bilingual parser based on random transduction grammars to identify terms, including MWEs. [7] and [8] employed vector space to calculate the semantic distance.

Though knowledge-based approaches have been tested successful to different extents in identifying and extracting MWEs, especially in contexts where MWEs have low frequency, the complexity and huge quantity of the real texts still pose great challenges to MWE extraction. The requirement for large-scale dictionary or rule base brings about extra

burden for the extraction systems. And the knowledge-based or symbolic algorithm cannot deal well with the exceptions which are quite common for human natural languages. Hence, many researches tended to search for different approaches like statistical approaches.

2.2. **Statistical Approaches.** With the rapid development of corpus linguistics and many foundations of large-scale corpora, a growing number of statistical approaches have been suggested and have achieved success to various extents.

Statistical systems usually extract MWEs from corpora by means of association measure. As they use plain text corpora and only require the information appearing in texts, such systems are highly flexible and extract relevant units independently from the domain and the language of the input text. For example, [9] utilized Log Likelihood Ratio (LLR), X2, DICE and MI method to extract translation pairs. [1] introduced their HRA algorithm to extract Chinese MWEs on the MWE workshop in 2009. In order to effectively extract domain MWEs, a procedure including extracting, filtering, evaluating has been explored. In the extracting step, according to the features of Chinese domain MWEs, they proposed a hierarchical reducing algorithm (HRA) based on LLR. Compared with previous work, the algorithm can not only extract MWE candidates gradually, but also have the advantages of avoiding meaningless MWEs and setting threshold conveniently. [10] used Log Likelihood Ratio (LLR) and X2 to extract Chinese MWEs from the Chinese corpus of CCID (China Center for Information Industry Development, Beijing, China) for the purpose of improving a machine translation system. They used an existing statistical tool built for English and extended it to Chinese. The tool exploits statistical collocational information between near-context words. It first collects collocates within a given scanning window, and then searches for MWEs using the collocational information as a statistical dictionary.

Among all those researches above, different kinds of association measures have been employed, either separately or collaboratively. The paper by [11] summarized altogether 55 associations measures (AM) which are commonly used in statistical approaches and showed that "different measures give different results for different tasks (data)". [12] was the first to exploit statistical approach and association measure to extract MWEs. He used point-wise mutual information to calculate the relativity between two words. With regard to two-word expressions, [13] proposed a boundary-extended method: first of all, he extracted all the two-word expressions with high relativity; then extended the boundaries of the two-word expressions, thus obtaining all the two-word, three-word, four-word and even K-word expressions; at last he picked out those invalid expressions according to a group of filter rules. [10] used another approach based on the association measure of LLR: he searched a sentence from left to right and obtained sequences of words; then any sequence of words within which the LLR between two words was higher than a set threshold would be considered as a MWE candidate. To some extent, statistical approaches have showed their great potential in the work of automatic extraction of MWEs and have been tested quite successful, especially when being provided with large-scale corpus.

2.3. **Hybrid Approaches.** One of the main problems facing statistical approaches, however,

is that they are unable to deal with low frequency multiword expressions. In fact, the majority of the words in most corpora have low frequencies, occurring only once or twice, especially for those corpora which are not big enough. Like pure statistical approaches, purely knowledge-based or symbolic approaches also face problems. They are language dependent and not flexible enough to cope with complex structures of MWEs. That's why a lot of hybrid systems, which usually combines knowledge-based or semantic-based approaches with the statistical approaches, have been proposed. For example, [14] proposed a hybrid system called HELAS that extracts MWE candidates from part-of-speech tagged corpora. Unlike classical hybrid systems that manually pre-define local part-of-speech patterns of interest, his solution automatically identifies relevant syntactical patterns from the corpus. Word statistics are then combined with the endogenously acquired linguistic information in order to extract the most relevant sequences of words i.e. MWE candidates. Meanwhile, some Chinese researchers also have devoted their attention to the hybrid systems. For example, in the paper by [15], a method of combining semantic template and statistical tool was proposed for automatically extracting native English MWE from three-tuple comparable corpus. Another example is the dissertation by [16], which focuses on the MWE extraction and its applications. Aiming at the features of monolingual and bilingual MWEs, the author proposes a set of approaches to extract flexible MWEs. They are inspired by gene sequence alignment in bioinformatics. These models combine the characteristics of natural language and some machine learning methods.

Although the automatic extraction of Multiword Expression (MWE) has been explored and discussed by a lot of people, it still presents a tough challenge for the NLP community and corpus linguistics. As for knowledge-based or rule-based approaches, dealing with a large number of exceptions and the complexity of the real texts often make it quite complicated and ill-robust. The performance of MWE extraction system depends heavily on the performance of the parser and the knowledge analysis. On the other hand, statistical approaches do not require the complicated rules and knowledge analysis, and large-scale parallel corpora nowadays are readily available to researchers, while they are unable to deal well with low frequency multiword expressions. Hence, both purely knowledge-based approaches and statistical approaches have their advantages and disadvantages. A possible better method is to combine the advantages of both knowledge-based approaches and statistical approaches and to find the right balance between the two types of approaches. For instance, statistical approaches can be used to extract high frequency MWE candidates whereas linguist information or language knowledge can be applied to extract low frequency MWE candidates. Afterwards, a model which is able to filter and reevaluate the MWE candidates will be needed so as to improve the performance and accuracy of the MWE extraction system.

3. **Word Relativity Based Multiword Expression Extraction.** Based on the above analysis of different extraction approaches, this research will adopt a hybrid system combining statistical approaches and knowledge-based approaches in order to explore the automatic extraction of MWEs based on word relativity. To be more specific, it will use

statistical association measure of LLR to calculate the word relativity between two words, on the basis of which an algorithm of MWE extraction will be provided. First of all, a Chinese-English parallel corpus will be created and annotated manually. Second, the word alignment information generated by GIZA++ will be improved by the use of a machine-readable bilingual dictionary of Hownet (2008). Third, an algorithm based on word relativity will be employed to extract MWE candidates from the source language (Chinese in this case). In the end, two kinds of filter methods based on association measure of LLR and word alignment information respectively will be used to filter out those invalid MWE candidates, thus obtaining the remained MWE candidates which we consider as the final results of the MWE extraction. Figure 1 shows the organization structure of the research.

3.1. **Chinese-English Parallel Corpus.** A Chinese-English parallel corpus will be created for the purpose of obtaining the word alignment information and evaluating the performance of the MWE extraction approach proposed in this research.

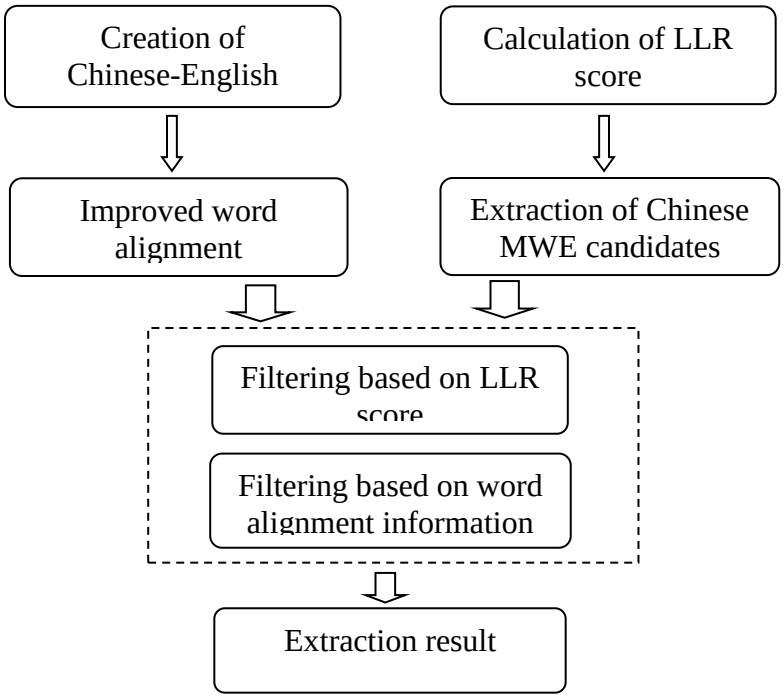

FIGURE 2. ORGANIZATION STRUCTURE OF THE RESEARCH

3.1.1. **Bilingual Parallel Corpus.** In linguistics, a corpus is a large and structured set of texts which are usually electronically stored and processed. They are used to do statistical analysis and hypothesis testing, checking occurrences or validating linguistic rules. A corpus may contain texts in a single language (monolingual corpus) or texts in multiple languages (multilingual corpus). Multilingual corpora that have been specially formatted for side-by-side comparison are called aligned parallel corpora. In this research, the corpus

containing two languages of Chinese and English is called Chinese-English parallel corpus, which is aligned manually on the sentence level. We select altogether 1051 sentence pairs from both the Chinese and English version of Chinese government work report in 2010 and 2011. In each sentence pair, either of the sentence is the translation equivalent of the other.

3.1.2. **Processing of Bilingual Parallel Corpus.** Based on the parallel corpus obtained in the last part, we will do some extra processing in this part, including word segmentation on Chinese text and tokenization on English text. For word segmentation, we use the Chinese Lexical Analysis System (ICTCLAS) by Institute of Computing Technology, Chinese Academy of Sciences. In tokenization, the tool of Tokenizer from NLTK (Natural Language Toolkit) is employed.

3.1.3. **Annotation of Bilingual Parallel Corpus.** In order to make the corpus more useful for doing linguistic research, they are often subjected to a process known as annotation. An example of annotating a corpus is part-of-speech tagging, in which information about each word's part of speech (verb, noun, adjective, etc.) is added to the corpus in the form of tags. In this corpus, true MWEs will be annotated manually so as to compare them with the MWEs extracted by the hybrid system and then evaluate its performance. Example 1 gives part of the corpus which has been annotated (The numbers in the example refer to the word ID).

Example 3.1.

C: 覆盖 1 城乡 2 的 3 社会 4 保障 5 体系 6 逐步 7 健全 8 。9

E: The1 social2 security3 system4 covering5 both6 urban7 and8 rural9 areas10 was11 progressively12 refined13 .14

MWE: C: 4 5 [T]; C: 4 5 6 [T]; E: 2 3 [T]; E: 2 3 4 [T]; E: 7 8 9 10 [C]

3.2. **Word Alignment.** This part will explore the word alignment approach based on the tool of GIZA++ in order to gain better word alignment information, which will be used for extracting Chinese MWEs.

3.2.1. Word Alignment and GIZA++. Corpora are the main knowledge base in corpus linguistics. Through statistical analysis of corpora, people can get different kinds of language knowledge or information, among which word alignment is of great significance. For example, word alignment is an important supporting task for most methods of statistical machine translation. Nowadays, the most commonly used approach for word alignment is to exploit bilingual parallel corpus. A representative example is the tool of GIZA++. However, the word alignment generated by GIZA++ is not as satisfactory as this research require. Example 2 shows the word alignment information of a sentence pair.

Example 3.2.

Sentence pair (1) source length 18 target length 26 alignment score: 3.46947e-29

E: On1 behalf2 of3 the4 state5 council 6 ,7 i8 now9 present10 to11 you12 my13 report14 on15 the16 work17 of18 the19 government20 for21 your22 deliberation23 and24

approval25 .26
C: NULL ({ 3 4 16 18 19 24 }) 现在 ({ 1 2 }) ， ({ }) 我 ({ }) 代表 ({ 15 }) 国务院 ({ 5 6 }) ， ({ 7 }) 向 ({ 11 }) 大会 ({ 12 }) 作 ({ }) 政府 ({ 20 }) 工作 ({ 17 }) 报告 ({ }) ， ({ }) 请 ({ 8 }) 各位 ({ 9 10 13 14 }) 代表 ({ }) 审议 ({ 21 22 23 25 }) 。 ({ 26 })

3.2.2. **Improved Word Alignment Based on GIZA++.** From Example 2, we see a lot of errors in the GIZA++ word alignment. However, GIZA++ has still a certain degree of precision and is commonly used in word alignment work. This part will explore a method based on GIZA++ to get the improved word alignment information by the use of a machine-readable bilingual dictionary of Hownet [17]. To use the dictionary successfully, we have to solving the problem of getting the stem of each word. Here, we use the existing online tool of CST's Lemmatiser by University of Copenhagen.

Based on the word alignment generated by GIZA++, the improved alignment algorithm will first of all search the GIZA++ alignment file and then leave the alignment as it is for the correct one and revise it for the incorrect one. The algorithm is shown as below:

Step 1: Search the Chinese sentence in the alignment file (as shown by Example 3.2) generated by GIZA++, look up each word in the dictionary. If English meaning of the Chinese word is the same as the alignment information shows, leave the information as it is. Otherwise add the Chinese word to a non-aligned list.

Step 2: Search the non-aligned list, look up each word in the dictionary. If English meaning of the Chinese word can have correspondence in English sentence, record the correspondence as alighment information. Otherwise no action is taken, and the Chinese word has no correspondence.

Step 3: Sort the alignment information according to the order of Chinese word

Step 4: Filter out those Chinese words which should not be aligned.

In Step 4, we take some commonly used Chinese function words into consideration. Those words are used frequently in practice and usually do not have their English equivalents, which would case noise in the word alignment. So we create a "non-align" word list like "的" and "和". Words in the "non-align" list are aligned to nothing.

3.3. **Extraction of MWE Candidates.** In this part, an algorithm based on word relativity will be proposed to extract Chinese MWE candidates, which will be filtered in the next part.

3.3.1. **Word Relativity.** Due to MWE's idiosyncrasy of high word relativity within itself, we usually consider a word group or a word sequence within which words co-occur frequently as a possible MWE. Relativity between words is the degree of interdependence of words on each other by the calculation of the frequency of word group and its components (i.e. words). To be more specific, if the occurrence number of a word group as a whole is much larger than that of its components, it's highly possible that the word group is a MWE. In NLP, we use association measure (AM) to calculate the word relativity. There are many AMs which people commonly use in NLP research such as point-wise mutual

information, DICE coefficient and Log Likelihood Ratio (LLR). [11] have done researches on the comparison and evaluation of different AMs. In this research, we used LLR as the AM to calculate the word relativity.

To calculate the word relativity (i.e. LLR) between word W1 and word W2, we assume that 1) a is the number of sentences in the corpus that contain both word W1 and word W2; 2) b is the number of sentences that contain word W1 but not word W2; 3) c is the number of sentences that contain word W2 but not word W1; 4) d is the number of sentences that contain neither word W1 nor word W2. Formula 1 gives the equation about how to calculate the LLR score.

$$\begin{aligned} LLR(W1, W2) = 2(&a \log a + b \log b + c \log c + d \log d \\ &+ (a + b + c + d) \log (a + b + c + d) \\ &- (a + b) \log (a + b) - (a + c) \log (a + c) \\ &- (b + d) \log (b + d) - (c + d) \log (c + d)) \end{aligned} \quad (1)$$

For example, in the word group “政府 工作”, both the word “政府” and “工作” occur in 9 sentences; 41 sentences contain “政府” but not “工作” while 47 sentences contain “工作” but not “政府”; 958 sentences contain neither of the two words. Then the LLR score of the word group is 11.01. Another word group “工作 报告” get the score of 6.48, which means the words “政府” and “工作” is more closely connected than the words “工作” and “报告”. So “政府 工作” is more likely to be a MWE candidate. (To ensure the validity of the above equation, the four parameters, i.e. a, b, c and d, at least get the value of 1 respectively.)

3.3.2. **Extraction of MWE Candidates.** The LLR calculates the word relativity between only two words, while the extraction of MWEs need word relativity between two or more than two words. Therefore, we need a method to extend the word numbers when using the AM of LLR. The algorithm in this part will first of all search the LLR score between every two adjacent words in a sentence and obtain several word sequences based on a setting threshold. Within the word sequence, LLR score between every two adjacent words is higher than the threshold. Next, the algorithm will extract every possible sub-sequence from the word sequences and take those sub-sequences as MWE candidates. The following example shows how the algorithm works:

(1) To begin with, we have a Chinese sentence as shown by Example 3.3.

Example 3.3. 现在 ， 我 代表 国务院 ， 向 大会 作 政府 工作 报告 ， 请 各位 代表 审议 。

(2) After calculation, we get LLR score between every two adjacent words. (The numbers between words present the LLR scores)

Example 3.4. 现在 0 ， 0 我 0 代表 35.4 国务院 0 ， 0 向 0 大会 3.9 作 9.9 政府 11.0 工作 6.5 报告 0 ， 0 请 0 各位 0 代表 15.2 审议 0 。

(3) If the threshold is 5, we can get three word sequences, within which very LLR score is higher than the threshold: a) 代表 国务院 b) 作 政府 工作 报告 c) 代表 审议

(4) Based on the word sequences in Step (3), we can get every possible sub-sequence, as shown by Example 3.5.

Example 3.5. 代表 国务院; 作 政府; 作 政府 工作; 作 政府 工作 报告; 政府 工作; 政府 工作 报告; 工作 报告; 代表 审议

In the calculation of LLR, we use a stop-word list so as to avoid the noise that functions words would bring about. LLR score in any two-word sequence containing stop word is zero.

3.4. **Filtering of MWE Candidates.** This part will filter all the MWE candidates extracted in the last part. Two kinds of filter methods based on association measure of LLR and word alignment information respectively will be used to filter out those invalid MWE candidates. After the two stages of filtering, the remained MWE candidates will be considered as the final results of MWE extraction.

3.4.1. **Filtering Based on LLR.** According to the analysis in part 3.3, MWEs are those word groups or word sequences within which words co-occur frequently, which means the internal relativity within the MWE should be higher than the relativity between MWE itself and its neighbors. In this research, we take the boundaries of MWE into consideration. For those boundary words which are less closely connected with the internal words of the MWE than the external words, the MWEs will be filtered out and the rest of the MWEs will be remained for the next stage of filtering.

To be more specific, for the left boundary word of a MWE, if the LLR score between it and its left adjacent word is higher than that between it and its right adjacent word, the MWE is filtered out. Similarly, for the right boundary word of a MWE, if the LLR score between it and its right adjacent word is higher than that between it and its left adjacent word, the MWE is filtered out. If and only if a MWE has passed both the two procedures above, it will be remained. The following shows how this filtering stage works:

(1) We have a Chinese sentence "并 请 全国政协 委员 提出 意见 。" and the MWE candidates after the processing in part 3.3.

Example 3.6. 请 全国政协; 请 全国政协 委员; 请 全国政协 委员 提出;
请 全国政协 委员 提出 意见;
全国政协 委员; 全国政协 委员 提出; 全国政协 委员 提出 意见;
委员 提出; 委员 提出 意见; 提出 意见;

(2) From part 3.3, we have already got LLR score between every two adjacent words. (The numbers between words present the LLR scores)

Example 3.7. 并 0 请 19.7 全国政协 21.4 委员 16.9 提出 13.9 意见 0 。

(3) For each MWE candidate, we examine its boundary words. Here two example candidates are shown. a) "请 全国政协": Two boundary words are "请" and "全国政协". Since 0 is smaller than 19.7 but 21.4 is larger than 19.7, the MWE is invalid. b) "请 全国政协 委员": Two boundary words are "请" and "委员". Since 0 is smaller than 19.7 and 16.9 is also smaller than 21.4, the MWE is remained.

3.4.2. **Filtering Based on Word Alignment.** In real-life human communication, meaning is often conveyed by word groups rather than single words. And word groups or MWEs in

this case usually have a particular but not uncertain meaning. Consequently, when translated into another language, MWEs have certain translation equivalents which convey the same independent meaning. In this research, we assume that the English equivalents of Chinese MWEs are consecutive word sequences rather than discontinuous ones. Specifically speaking, we will employ the word alignment information which we get in above parts to obtain the English equivalents of Chinese MWE candidates. Those MWE candidates whose English equivalents are consecutive are remained as our final results of the automatic extraction of Chinese MWEs. The algorithm shows how this stage of filtering works:

Input: word alignment file F; MWE candidate M= Mst ; Word list W=W1n;
Step 1: Read F, get the list NULL="1 2 4 5 6 7 9 10 11 18 20 21 22"
Step 2: for i=s to t do
Get alignment information of word i, such as list Li="12 13 15 16 17"
Fill in numbers to make list Li consecutive, then Li="12 13 14 15 16 17"
Step 3: L=Ls+Ls+1+…+Lt
Step 4: If L is empty set
Then: M is valid, exit
Step 5: If intersection(Ls, Ls+1, …, Lt) is not empty set
Then: M is invalid, exit
Step 6: Sort(L)
Step 7: for i=L[1] to L[length(L)] do
If (i in Null) and (i not in L) Then: L=L+i
Step 8: if L is consecutive
Then: M is valid
Else: M is invalid

At the beginning, we have word alignment file as shown by Example 3.8. (Numbers in the braces refer to the word ID in English sentence.)
Example 3.8.
Sentence pair (2) source length 7 target length 26 alignment score: 5.86324e-37
E: I also invite the members of the national committee of the Chinese people's political consultative conference (cppcc) to submit comments and suggestions.
C: NULL ({ 1 2 4 5 6 7 9 10 11 18 20 21 22 }) 并 ({ 24 }) 请 ({ 3 }) 全国政协 ({ 19 }) 委员 ({ 12 13 15 16 17 }) 提出 ({ 8 14 25 }) 意见 ({ 23 }) 。 ({ 26 })
In the algorithm, the index likes in $\mathbf{M}_s^t$ refers to the word ID in the Chinese sentence. So $\mathbf{M}_s^t$ means word sequence ranging from word s to word t.

After the processing of the MWE candidates obtained in part 3.4.1, we get the final result of the automatic extraction of Chinese MWEs "全国政协 委员".

4. **Evaluation and Analysis.** In order to test our approach of extracting MWEs, we first conduct experiment on the corpus which has been processed in part 3.1.2 and get the extracted MWEs. Then, based on the corpus which has been annotated manually in part 3.1.3, we calculate the overall precision and recall. Finally, by comparing our extraction

system with the approach proposed by other researchers, we analyze the performance of our approach.

The extraction of MWEs in our experiment requires a setting threshold, and different thresholds lead to different extraction results. 错误!未找到引用源。 shows our extraction results under different thresholds.

TABLE 1. RESULT OF MWE EXTRACTION

LLR threshold	Candidates	True MWEs	Precision	Recall
5	3630	718	19.78%	41.72%
10	2389	586	24.53%	34.05%
15	1424	445	31.25%	25.86%
20	729	282	38.68%	16.39%
30	269	136	50.56%	7.90%
40	135	78	57.78%	4.53%
50	81	47	58.02%	2.73%
60	54	31	57.41%	1.80%

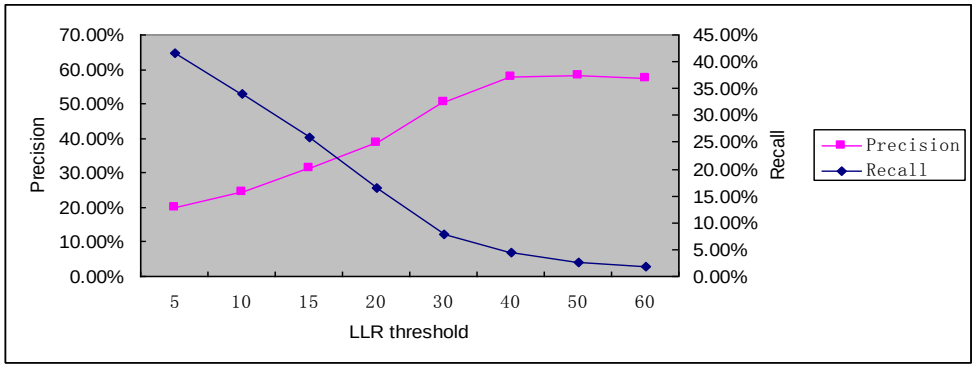

FIGURE 3. OVERALL PRECISION AND RECALL

From the table above, we know that the threshold setting plays a vital role in MWE extraction. In the following figures (Figure 3 and Figure 4), we show the relationship between threshold and the performance of our extraction system.

Furthermore, to analyze the performance of out extraction system, we will compare our approach with another approach proposed by other researches which is also based on the AM of LLR. Here, we experiment the HRA algorithm proposed by REN Zhixiang et al. (2009) on the same corpus, and get the extraction result as shown by 错误!未找到引用源。.

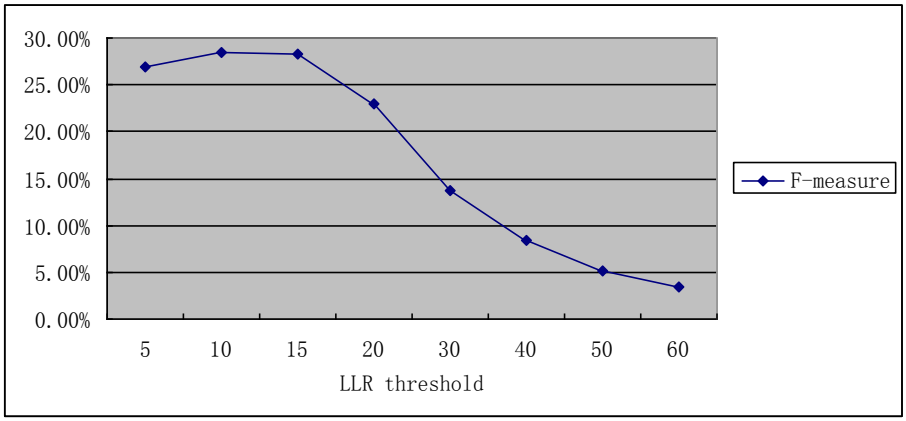

FIGURE 4. F-MEASURE OF THE EXTRACTION SYSTEM

TABLE 2. RESULT OF MWE EXTRACTION

LLR threshold	Candidates	True MWEs	Precision	Recall
5	5379	844	15.69%	49.04%
10	3425	672	19.62%	39.05%
15	1929	500	25.92%	29.05%
20	902	316	35.03%	18.36%
30	297	139	46.80%	8.08%
40	141	79	56.03%	4.59%
50	83	47	56.63%	2.73%
60	56	31	55.36%	1.80%

Figure 5 shows the different performance of two extraction approaches which have been tested. Precision 1 refers to the approach we proposed in our research and precision 2 refer to the system based on the HRA algorithm.

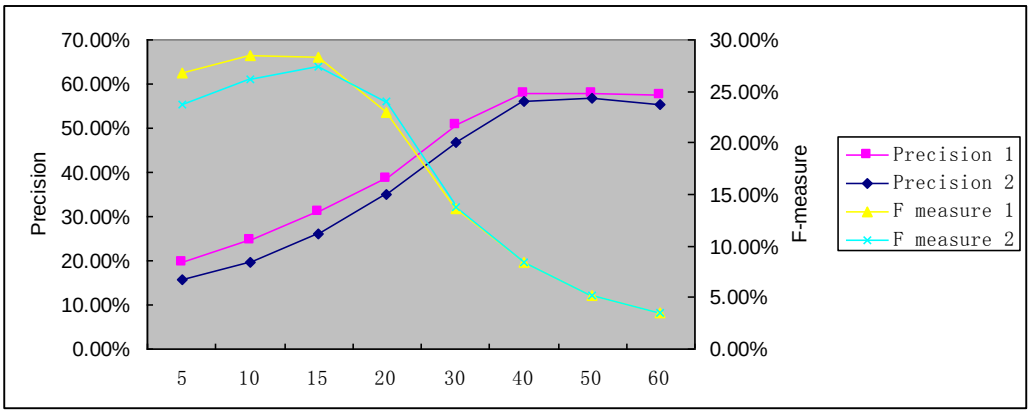

FIGURE 5. DIFFERENT PERFORMANCE OF THE TWO EXTRACTION SYSTEMS

From the figure, we can know that precision of our extraction system turned out to be larger on different thresholds. When the threshold is getting smaller, the difference between precisions is getting bigger, which means that our system, by employing word alignment

information, has a much more obvious advantage when the word relativity is low. That is in accordance with the analysis in above parts, which shows that statistical approaches are usually unable to deal well with low frequency MWEs while linguist information or language knowledge can be applied to extract low frequency MWE candidates. However, when threshold is getting bigger, meaning that the word relativity has carried more weights in extraction, the effect of linguist information becomes reduced. In addition, due to the complexity and flexibility of real texts as well as the limitations of a certain corpus, a word group co-occurring more frequently than another word group does not necessarily mean that the word group has inevitably higher word relativity in real-life human communication. Therefore, the LLR score calculated by the use of co-occurrence of words should be used as a reference rather than the only parameter. In our method, every word in the word sequence extracted according to the threshold of LLR is of equal significance, so every possible sub-sequence is considered as a MWE candidate, which makes the automatic extraction of MWEs an issue of judging and filtering.

5. **Conclusions.** In this thesis, we have proposed a hybrid system combining statistical and linguistic information to extract Chinese MWEs automatically. Since the relative frozen form of MWEs, it is reasonable to make use of word relativity to identify the boundaries of MWEs and extract MWE candidate. Besides, to further improve our extraction system, we also take linguistic information into consideration so as to address the problem of extracting low frequency MWEs. Specifically speaking, we first created a Chinese-English bilingual parallel corpus and explored an algorithm to improve the word alignment information. Next, we proposed our MWE extraction algorithm, which includes extracting the MWE candidates based on word relativity and filtering the candidates by the use of word relativity and word alignment information.

The experiments on the bilingual parallel corpus have shown its improvement and advantages over the purely statistical approach. The word alignment information has been improved greatly compared with the alignment generated by GIZA++. We have also dealt with the problem of extracting low frequency MWEs by making use of the word alignment information. When the word relativity is getting lower, the influence of word alignment information on extracting MWEs has increased.

The hybrid system proposed in this research extended the association measure to measure the relativity within multiple words rather than only two words. Moreover, the thresholds can be set conveniently. In addition, both the statistical and linguistic information used in the research come from the same parallel corpus, which makes the hybrid system easily conducted because of the easy availability of bilingual parallel corpus nowadays. Nevertheless, our hybrid system does have room of improvement. Since the word relativity is based on the co-occurrences of words in a certain corpus, the parallel corpus should be large enough so as to be more similar with the real-life human communication. Besides, because our word alignment algorithm used a bilingual dictionary and a dictionary usually doesn’t cover all the words in real-texts, the word alignment algorithm should be improved further.

Acknowledgment. This work is partially supported by National Social Science Fund of China (11CYY030), National Natural Science Fund of China (61272221), Jiangsu Province Fund of Social Science under Grant 12YYA002, and the Innovative Research Fund of Huazhong University of Science and Technology (2012WQN018). Heart-felt gratitude also goes to anonymous reviewers for their helpful comments on the paper.

REFERENCES

[1] Z. Ren, *et al.*, "Improving Statistical Machine Translation Using Domain Bilingual Multiword Expressions," in *Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications*, ed. Suntec, Singapore: Association for Computational Linguistics, 2009, pp. 47-54.

[2] I. A. Sag, *et al.*, "Multiword Expressions: A Pain in the Neck for NLP," in *Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing*, ed. London, UK, UK: Springer-Verlag, 2002, pp. 1-15.

[3] R. S. Jackendoff, *The architecture of the language faculty*. Cambridge, Mass.: MIT Press, 1997.

[4] C. Fellbaum, *WordNet : an electronic lexical database*. Cambridge, Mass: MIT Press, 1998.

[5] S. S. L. Piao, *et al.*, "Extracting Multiword Expressions with a Semantic Tagger," in *Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment - Volume 18*, ed. Sapporo, Japan: Association for Computational Linguistics, 2003, pp. 49-56.

[6] D. Wu, "Stochastic Inversion Transduction Grammars and Bilingual Parsing of Parallel Corpora," *Comput. Linguist.*, vol. 23, pp. 377-403, September 1997.

[7] T. Baldwin, *et al.*, "An Empirical Model of Multiword Expression Decomposability," in *The ACL-SIGLEX Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, 2003, pp. 89-96.

[8] G. Katz and E. Giesbrecht, "Automatic Identification of Non-compositional Multi-word Expressions Using Latent Semantic Analysis," in *Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, ed. Sydney, Australia: Association for Computational Linguistics, 2006, pp. 12-19.

[9] B. Chang, *et al.*, "Extraction of Translation Unit from Chinese-English Parallel Corpora," in *Proceedings of the First SIGHAN Workshop on Chinese Language Processing - Volume 18*, ed. Stroudsburg, PA, USA: Association for Computational Linguistics, 2002, pp. 1-5.

[10] S. Piao, *et al.*, "Automatic extraction of Chinese multiword expressions with a statistical tool," in *Workshop on Multi-word-expressions in a Multilingual Context held in conjunction with the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2006)*, Trento, Italy, 2006.

[11] P. Pecina, "A Machine Learning Approach to Multiword Expression Extraction," in *the LREC 2008 Workshop Towards a Shared Task for Multiword Expressions*, 2008, pp. 54-57.

[12] K. W. Church and P. Hanks, "Word Association Norms, Mutual Information, and Lexicography," *Comput. Linguist.*, vol. 16, pp. 22-29, March 1990.

[13] P. Pantel and D. Lin, "A Statistical Corpus-Based Term Extractor," in *Proceedings of the 14th Biennial Conference of the Canadian Society on Computational Studies of Intelligence: Advances in Artificial Intelligence*, ed. London, UK, UK: Springer-Verlag, 2001, pp. 36-46.

[14] G. e. Dias, "Multiword Unit Hybrid Extraction," in *Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment - Volume 18*, ed. Sapporo, Japan: Association for Computational Linguistics, 2003, pp. 41-48.

[15] J. Xiao, *et al.*, "Multiword Expression Extraction and Alignment in English-Chinese Comparable Corpora," *Computer Engineering and Application*, vol. 46, pp. 130-134, 2010.

[16] J. Duan, "Extraction and Application of Multiword Expression," PhD, Shanghai Jiaotong University, Shanghai, 2007.

[17] Z. Dong and Q. Dong, *HowNet and the computation of meaning*. Hackensack, NJ: World Scientific, 2006.