Collocation Extraction Using Monolingual Word Alignment Method

Zhanyi Liu[1,2], Haifeng Wang[2], Hua Wu[2], Sheng Li[1]

[1]Harbin Institute of Technology, Harbin, China
[2]Toshiba (China) Research and Development Center, Beijing, China
{liuzhanyi,wanghaifeng,wuhua}@rdc.toshiba.com.cn
lisheng@hit.edu.cn

Abstract

Statistical bilingual word alignment has been well studied in the context of machine translation. This paper adapts the bilingual word alignment algorithm to monolingual scenario to extract collocations from monolingual corpus. The monolingual corpus is first replicated to generate a parallel corpus, where each sentence pair consists of two identical sentences in the same language. Then the monolingual word alignment algorithm is employed to align the potentially collocated words in the monolingual sentences. Finally the aligned word pairs are ranked according to refined alignment probabilities and those with higher scores are extracted as collocations. We conducted experiments using Chinese and English corpora individually. Compared with previous approaches, which use association measures to extract collocations from the co-occurring word pairs within a given window, our method achieves higher precision and recall. According to human evaluation in terms of precision, our method achieves absolute improvements of 27.9% on the Chinese corpus and 23.6% on the English corpus, respectively. Especially, we can extract collocations with longer spans, achieving a high precision of 69% on the long-span (>6) Chinese collocations.

1 Introduction

Collocation is generally defined as a group of words that occur together more often than by chance (McKeown and Radev, 2000). In this paper, a collocation is composed of two words occurring as either a consecutive word sequence or an interrupted word sequence in sentences, such as "by accident" or "take … advice". The collocations in this paper include phrasal verbs (e.g. "put on"), proper nouns (e.g. "New York"), idioms (e.g. "dry run"), compound nouns (e.g. "ice cream"), correlative conjunctions (e.g. "either … or"), and the other commonly used combinations in following types: verb+noun, adjective+noun, adverb+verb, adverb+adjective and adjective+preposition (e.g. "break rules", "strong tea", "softly whisper", "fully aware", and "fond of").

Many studies on collocation extraction are carried out based on co-occurring frequencies of the word pairs in texts (Choueka et al., 1983; Church and Hanks, 1990; Smadja, 1993; Dunning, 1993; Pearce, 2002; Evert, 2004). These approaches use association measures to discover collocations from the word pairs in a given window. To avoid explosion, these approaches generally limit the window size to a small number. As a result, long-span collocations can not be extracted[1]. In addition, since the word pairs in the given window are regarded as potential collocations, lots of false collocations exist. Although these approaches used different association measures to filter those false collocations, the precision of the extracted collocations is not high. The above problems could be partially solved by introducing more resources into collocation extraction, such as chunker (Wermter and Hahn, 2004), parser (Lin, 1998; Seretan and Wehrli, 2006) and WordNet (Pearce, 2001).

This paper proposes a novel *monolingual word alignment* (MWA) method to extract collocation of higher quality and with longer spans only from monolingual corpus, without using any additional resources. The difference between MWA and bilingual word alignment (Brown et al., 1993) is that the MWA method works on monolingual parallel corpus instead of bilingual corpus used by bilingual word alignment. The

[1] Here, "span of collocation" means the distance of two words in a collocation. For example, if the span of the collocation (w_1, w_2) is 6, it means there are 5 words interrupting between w_1 and w_2 in a sentence.

Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pages 487–495,
Singapore, 6-7 August 2009.

monolingual corpus is replicated to generate a parallel corpus, where each sentence pair consists of two identical sentences in the same language, instead of a sentence in one language and its translation in another language. We adapt the bilingual word alignment algorithm to the monolingual scenario to align the potentially collocated word pairs in the monolingual sentences, with the constraint that a word is not allowed to be aligned with itself in a sentence. In addition, we propose a ranking method to finally extract the collocations from the aligned word pairs. This method assigns scores to the aligned word pairs by using alignment probabilities multiplied by a factor derived from the exponential function on the frequencies of the aligned word pairs. The pairs with higher scores are selected as collocations.

The main contribution of this paper is that the well studied bilingual statistical word alignment method is successfully adapted to monolingual scenario for collocation extraction. Compared with the previous approaches, which use association measures to extract collocations, our method achieves much higher precision and slightly higher recall. The MWA method has the following three advantages. First, it explicitly models the co-occurring frequencies and position information of word pairs, which are integrated into a model to search for the potentially collocated word pairs in a sentence. Second, a new feature, *fertility*, is employed to model the number of words that a word can collocate with in a sentence. Finally, our method can obtain the long-span collocations. Human evaluations on the extracted Chinese collocations show that 69% of the long-span (>6) collocations are correct. Although the previous methods could also extract long-span collocations by setting the larger window size, the precision is very low.

In the remainder of this paper, Section 2 describes the MWA model for collocation extraction. Section 3 describes the initial experimental results. In Section 4, we propose a method to improve the MWA models. Further experiments are shown in Sections 5 and 6, followed by a discussion in Section 7. Finally, the conclusions are presented in Section 8.

2 Collocation Extraction With Monolingual Word Alignment Method

2.1 Monolingual Word Alignment

Given a bilingual sentence pair, a source language word can be aligned with its corresponding target language word. Figure 1 shows an example of Chinese-to-English word alignment.

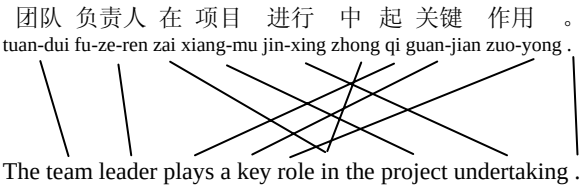
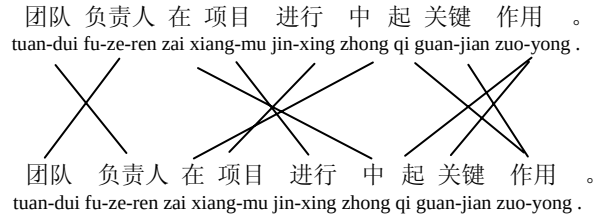


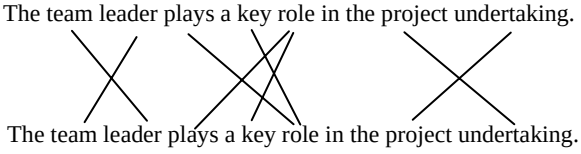
Figure 1. Bilingual word alignment

In Figure 1, a word in one language is aligned with its counterpart in the other language. For examples, the Chinese word "团队/tuan-dui" is aligned with its English translation "team", while the Chinese word "负责人/fu-ze-ren" is aligned with its English translation "leader".

In the Chinese sentence in Figure 1, there are some Chinese collocations, such as (团队/tuan-dui, 负责人/fu-ze-ren). There are also some English collocations in the English sentence, such as (team, leader). We separately illustrate the collocations in the Chinese sentence and the English sentence in Figure 2, where the collocated words are aligned with each other.



(a) Collocations in the Chinese sentence



(b) Collocations in the English sentence

Figure 2. Word alignments of collocations in sentence

Comparing the alignments in Figures 1 and 2, we can see that the task of monolingual collocations construction is similar to that of bilingual word alignment. In a bilingual sentence pair, a source word is aligned with its corresponding target word, while in a monolingual sentence, a word is aligned with its collocates. Therefore, it is reasonable to regard collocation construction as a task of aligning the collocated words in monolingual sentences.

Statistical bilingual word alignment method, which has been well studied in the context of machine translation, can extract the aligned bilingual word pairs from a bilingual corpus. This paper adapts the bilingual word alignment algorithm to monolingual scenario to align the collocated words in a monolingual corpus.

Given a sentence with l words $S = \{w_1, ..., w_l\}$, the word alignments $A = \{(i, a_i) \mid i \in [1, l]\}$ can be obtained by maximizing the word alignment probability of the sentence, according to Eq. (1).

$$A = \arg\max_{\forall A'} p(A' \mid S) \quad (1)$$

Where $(i, a_i) \in A$ means that the word w_i is aligned with the word w_{a_i}.

In a monolingual sentence, a word never collocates with itself. Thus the alignment set is denoted as $A = \{(i, a_i) \mid i \in [1, l] \,\&\, a_i \neq i\}$.

We adapt the bilingual word alignment model, IBM Model 3 (Brown et al., 1993), to monolingual word alignment. The probability of the alignment sequence is calculated using Eq. (2).

$$p(A \mid S) \propto \prod_{i=1}^{l} n(\phi_i \mid w_i) \prod_{j=1}^{l} t(w_j \mid w_{a_j}) d(j \mid a_j, l) \quad (2)$$

Where ϕ_i denotes the number of words that are aligned with w_i. Three kinds of probabilities are involved:

- Word collocation probability $t(w_j \mid w_{a_j})$, which describes the possibility of w_j collocating with w_{a_j};
- Position collocation probability $d(j, a_j, l)$, which describes the probability of a word in position a_j collocating with another word in position j;
- Fertility probability $n(\phi_i \mid w_i)$, which describes the probability of the number of words that a word w_i can collocate with (refer to subsection 7.1 for further discussion).

Figure 3 shows an example of word alignment on the English sentence in Figure 2 (b) with the MWA method. In the sentence, the 7th word "role" collocates with both the 4th word "play" and the 6th word "key". Thus, $t(w_4 \mid w_7)$ and $t(w_6 \mid w_7)$ describe the probabilities that the word "role" collocates with "play" and "key",

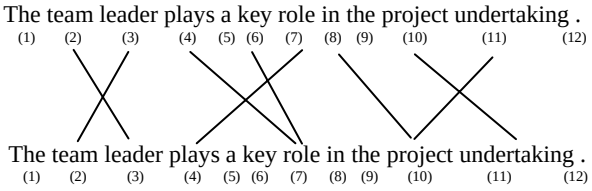


Figure 3. Results of MWA method

respectively. $d(4 \mid 7,12)$ and $d(6 \mid 7,12)$ describe the probabilities that the word in position 7 collocates with the words in position 4 and 6 in a sentence with 12 words. For the word "role", ϕ_7 is 2, which indicates that the word "role" collocates with two words in the sentence.

To train the MWA model, we implement a MWA tool for collocation extraction, which uses similar training methods for bilingual word alignment, except that a word can not be aligned to itself.

2.2 Collocation Extraction

Given a monolingual corpus, we use the trained MWA model to align the collocated words in each sentence. As a result, we can generate a set of aligned word pairs on the corpus. According to the alignment results, we calculate the frequency for two words aligned in the corpus, denoted as $freq(w_i, w_j)$. In our method, we filtered those aligned word pairs whose frequencies are lower than 5. Based on the alignment frequency, we estimate the alignment probabilities for each aligned word pair as shown in Eq. (3) and (4).

$$p(w_i \mid w_j) = \frac{freq(w_i, w_j)}{\sum_{w'} freq(w', w_j)} \quad (3)$$

$$p(w_j \mid w_i) = \frac{freq(w_i, w_j)}{\sum_{w'} freq(w_i, w')} \quad (4)$$

With alignment probabilities, we assign scores to the aligned word pairs and those with higher scores are selected as collocations, which are estimated as shown in Eq. (5).

$$\bar{p}(w_i, w_j) = \frac{p(w_i \mid w_j) + p(w_j \mid w_i)}{2} \quad (5)$$

3 Initial Experiments

In this experiment, we used the method as described in Section 2 for collocation extraction. Since our method does not use any linguistic information, we compared our method with the

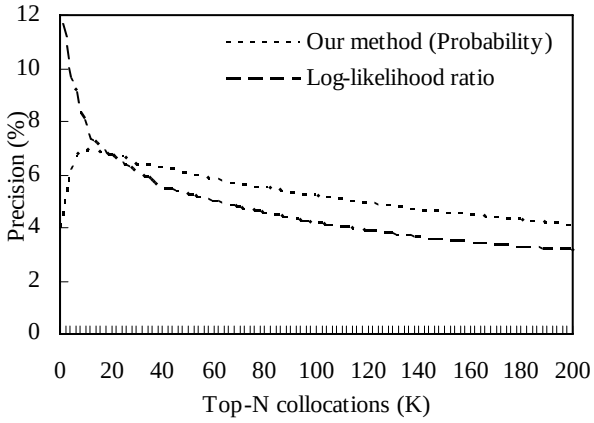


Figure 4. Precision of collocations

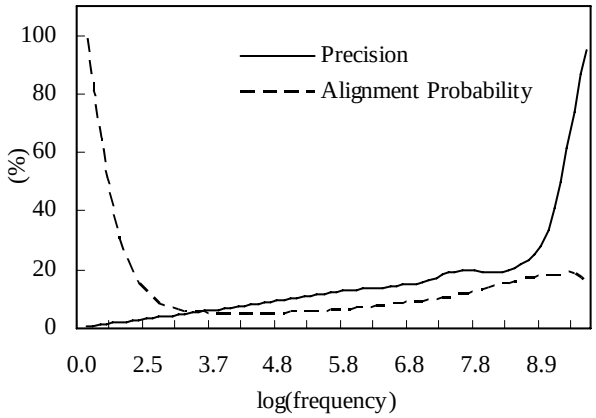


Figure 5. Frequency vs. precision/alignment probability

baseline methods without using linguistic knowledge. These baseline methods take all co-occurring word pairs within a given window as collocation candidates, and then use association measures to rank the candidates. Those candidates with higher association scores are extracted as collocations. In this paper, the window size is set to [-6, +6].

3.1 Data

The experiments were carried out on a Chinese corpus, which consists of one year (2004) of the Xinhua news corpus from LDC[2], containing about 28 millions of Chinese words. Since punctuations are rarely used to construct collocations, they were removed from the corpora. To automatically estimate the precision of extracted collocations on the Chinese corpus, we built a gold set by collecting Chinese collocations from handcrafted collocation dictionaries, containing 56,888 collocations.

3.2 Results

The precision is automatically calculated against the gold set according to Eq. (6).

$$precision = \frac{\#(C_{\text{Top}-N} \bigcap C_{\text{gold}})}{\#(C_{\text{Top}-N})} \qquad (6)$$

Where $C_{\text{Top-}N}$ and C_{gold} denote the top collocations in the N-best list and the collocations in the gold set, respectively.

We compared our method with several baseline methods using different association measures[3]: co-occurring frequency, log-likelihood ratio, chi-square test, mutual information, and t-test. Among them, the log-likelihood ratio measure achieves the best performance. Thus, in this paper, we only show the performance of the log-likelihood ratio measure.

Figure 4 shows the precisions of the top N collocations as N steadily increases with an increment of 1K, which are extracted by our method and the baseline method using log-likelihood ratio as the association measure.

The absolute precision of collocations is not high in the figure. For example, among the top 200K collocations, about 4% of the collocations are correct. This is because our gold set contains only about 57K collocations. Even if all collocations in the gold set are included in the 200K-best list, the precision is only 28%. Thus, it is more useful to compare precision curves for collocations in the N-best lists extracted by different methods. In addition, since this gold set only includes a small number of collocations, the precision curves of our method and the baseline method are getting closer, as N increases. For example, when N is set to 200K, our method and the baseline method achieved precisions of 4.09% and 3.12%, respectively. And when N is set to 400K, they achieved 2.78% and 2.26%, respectively. For convenience of comparison, we set N up to 200K in the experiments.

From the results, it can also be seen that, among the N-best lists with N less than 20K, the precision of the collocations extracted by our method is lower than that of the collocations extracted by the baseline, and became higher when N is larger than 20K.

In order to analyze the possible reasons, we investigated the relationships among the frequencies of the aligned word pairs, the alignment

[2] Available at: http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2007T03

[3] The definitions of these measures can be found in Manning and Schütze (1999).

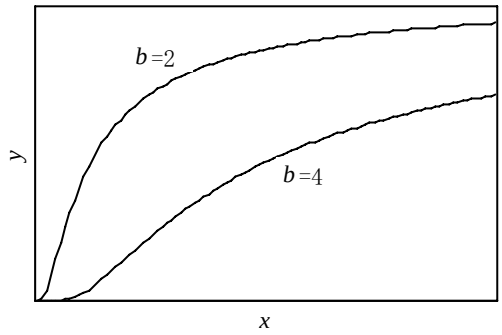


Figure 6. $y = e^{-b/x}$

probabilities, and precisions of collocations, which are shown in Figure 5. From the figure, we can see (1) that the lower the frequencies of the aligned word pairs are, the higher the alignment probabilities are; and (2) that the precisions of the aligned word pairs with lower frequencies is lower. According to the above observations, we conclude that it is the word pairs with lower frequencies but higher probabilities that caused the lower precision of the top 20K collocations extracted by our method.

4 Improved MWA Method

According to the analysis in subsection 3.2, we need to penalize the aligned word pairs with lower frequencies. In order to achieve the above goal, we need to refine the alignment probabilities by using a penalization factor derived from a function on the frequencies of the aligned word pairs. This function $y = f(x)$ should satisfy the following two conditions, where x represents the log function of frequencies.

(1) The function is monotonic. When x is set to a smaller number, y is also small. This results in the penalization on the aligned word pairs with lower frequencies.
(2) When $x \to \infty$, y is set to 1. This means that we don't penalize the aligned word pairs with higher frequencies.

According to the above descriptions, we propose to use the exponential function in Eq. (7).

$$y = e^{-b/x} \tag{7}$$

Figure 6 describes this function. The constant b in the function is used to adjust the shape of the line. The line is sharp with b set to a small number, while the line is flat with b set to a larger number. In our case, if b is set to a larger number, we assign a larger penalization weight to those aligned word pairs with lower frequencies.

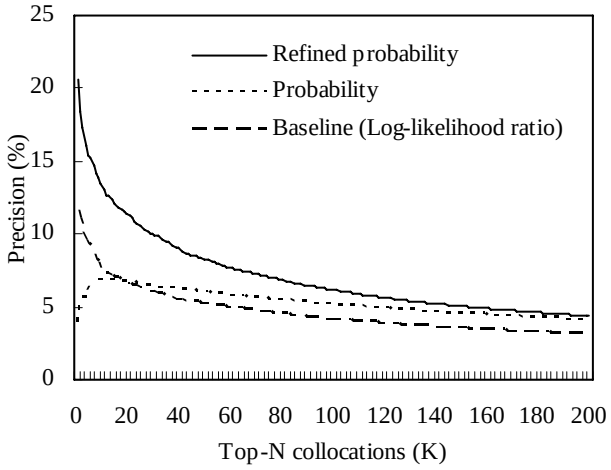


Figure 7. Precision of collocations extracted by the improved method

According to the above discussion, we can use the following measure to assign scores to the aligned words pairs generated by the MWA method.

$$\overline{p}_r(w_i, w_j) = \frac{p(w_i \mid w_j) + p(w_j \mid w_i)}{2} \times e^{-\frac{b}{\log(freq(w_i, w_j))}} \tag{8}$$

Where w_i and w_j are two aligned words. $p(w_i|w_j)$ and $p(w_j|w_i)$ are alignment probabilities as shown in Eq. (3) and (4). $\log(freq(w_i, w_j))$ is the log function of the frequencies of the aligned word pairs (w_i, w_j).

5 Evaluation on Chinese corpus

We used the same Chinese corpus described in Section 3 to evaluate the improved method as shown in Section 4. In the experiments, b was tuned by using a development set and set to 25.

5.1 Precision

In this section, we evaluated the extracted collocations in terms of precision using both automatic evaluation and human evaluation.

Automatic Evaluation

Figure 7 shows the precisions of the collocations in the N-best lists extracted by our method and the baseline method against the gold set in Section 3. For our methods, we used two different measures to rank the aligned word pairs: alignment probabilities in Eq. (5) and refined

		Our method	Baseline
True		569	290
False	A	25	16
	B	5	4
	C	240	251
	D	161	439

Table 1. Manual evaluation of the top 1K Chinese collocations. The precisions of our method and the baseline method are 56.9% and 29.0%, respectively.

alignment probabilities in Eq. (8). From the results, it can be seen that with the refined alignment probabilities, our method achieved the highest precision on the N-best lists, which greatly outperforms the best baseline method. For example, in the top 1K list, our method achieves a precision of 20.6%, which is much higher than the precision of the baseline method (11.7%). This indicates that the exponential function used to penalize the alignment probabilities plays a key role in demoting most of the aligned word pairs with low frequencies.

Human Evaluation

In automatic evaluation, the gold set only contains collocations in the existing dictionaries. Some collocations related to specific corpora are not included in the set. Therefore, we selected the top 1K collocations extracted by our improved method to manually estimate the precision. During human evaluation, the true collocations are denoted as "**True**" in our experiments. The false collocations were further classified into the following classes.

A: The candidate consists of two words that are semantically related, such as (医生 doctor, 护士 nurse).

B: The candidate is a part of the multi-word (≥3) collocation. For example, (自我 self, 机制 mechanism) is a part of the three-word collocation (自我 self, 约束 regulating, 机制 mechanism).

C: The candidates consist of the adjacent words that frequently occur together, such as (他 he, 说 say) and (很 very, 好 good).

D: Two words in the candidates have no relationship with each other, but occur together frequently, such as (北京 Beijing, 月 month) and (和 and, 为 for).

Table 1 shows the evaluation results. Our method extracted 569 true collocations, which

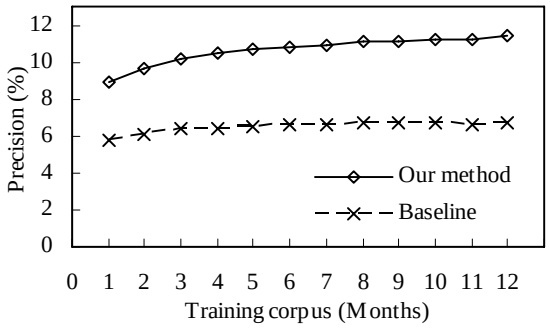


Figure 8. Corpus size vs. precision

are much more than those extracted by the baseline method. Further analysis shows that, in addition to extracting short-span collocations, our method extracted collocations with longer spans as compared with the baseline method. For example, (处于 in, 状态 state) and (由于 because, 因此 so) are two long-span collocations. Among the 1K collocations, there are 48 collocation candidates whose spans are larger than 6, which are not covered by the baseline method since the window size is set to 6. And 33 of them are true collocations, with a higher precision of 69%.

Classes C and D account for the most part of the false collocations. Although the words in these two classes co-occur frequently, they can not be regarded as collocations. And we also found out that the errors in class D produced by the baseline method are much more than that of those produced by our method. This indicates that our MWA method can remove much more noise from the frequently occurring word pairs.

In Class A, the two words are semantically related and occur together in the corpus. These kinds of collocations can not be distinguished from the true collocations by our method without additional resources.

Since only bigram collocations were extracted by our method, the multi-word (≥3) collocations were split into bigram collocations, which caused the error collocations in Class B[4].

Corpus size vs. precision

Here, we investigated the effect of the corpus size on the precision of the extracted collocations. We evaluated the precision against the gold set as shown in the automatic evaluation. First, the whole corpus (one year of newspaper) was split into 12 parts according to the published months. Then we calculated the precisions as the training

[4] Since only a very small faction of collocations contain more than two words, a few error collocations belong to Class B.

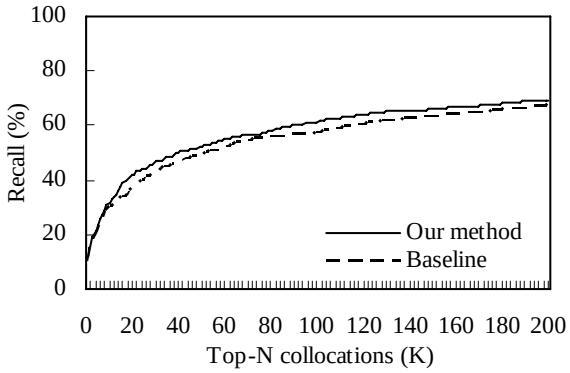


Figure 9. Recall on the Chinese corpus

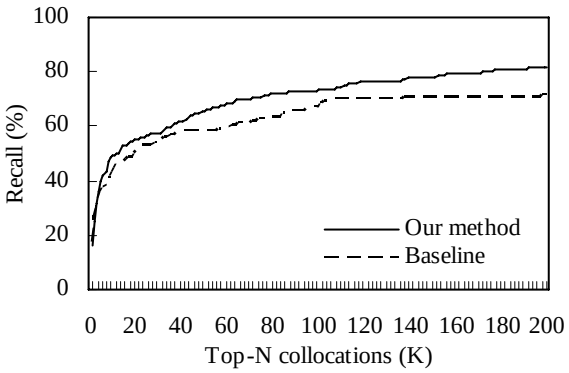


Figure 10. Recall on the English corpus

corpus increases part by part. The top 20K collocations were selected for evaluation.

Figure 8 shows the experimental results. The precision of collocations extracted by our method is obviously higher than that of collocations extracted by the baseline method. When the size of the training corpus became larger, the difference between our method and the baseline method also became bigger. When the training corpus contains more than 9 months of corpora, the precision of collocations extracted by the baseline method did not increase anymore. However, the precision of collocations extracted by our method kept on increasing. This indicates the MWA method can extract more true collocations of higher quality when it is trained with larger size of training data.

5.2 Recall

Recall was evaluated on a manually labeled subset of the training corpus. The subset contains 100 sentences that were randomly selected from the whole corpus. The sentence average length is 24. All true collocations (660) were labeled manually. The recall was calculated according to Eq. (9).

$$recall = \frac{\#(C_{\text{Top}-N} \bigcap C_{\text{subset}})}{\#(C_{\text{subset}})} \qquad (9)$$

Here, $C_{\text{Top-}N}$ denotes the top collocations in the N-best list and C_{subset} denotes the true collocations in the subset.

Figure 9 shows the recalls of collocations extracted by our method and the baseline method on the labeled subset. The results show that our method can extract more true collocations than the baseline method.

		Our method	Baseline
True		591	355
False	A	11	4
	B	19	20
	C	200	136
	D	179	485

Table 2. Manual evaluation of the top 1K English collocations. The precisions of our method and the baseline method are 59.1% and 35.5%, respectively.

In our experiments, the baseline method extracts about 20 millions of collocation candidates, while our method only extracts about 3 millions of collocation candidates[5]. Although the collocations of our method are much less than that of the baseline, the experiments show that the recall of our method is higher. This again proved that our method has the stronger ability to distinguish true collocations from false collocations.

6 Evaluation on English corpus

We also manually evaluated the proposed method on an English corpus, which is a subset randomly extracted from the British National Corpus[6]. The English corpus contains about 20 millions of words.

6.1 Precision

We estimated the precision of the top 1K collocations. Table 2 shows the results. The classification of the false collocations is the same as that in Table 1. The results show that our methods outperformed the baseline method using log-

[5] We set the threshold to 7.88 with a confidence level of $\alpha = 0.005$ (cf. page 174 of Chapter 5 in (McKeown and Radev, 2000) for more details).

[6] Available at: http://www.hcu.ox.ac.uk/BNC/

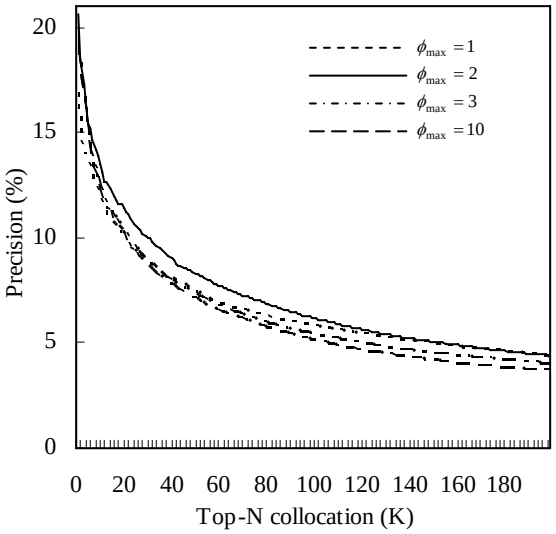


Figure 11. Fertility vs. precision

likelihood ratio. And the distribution of the false collocations is similar to that on the Chinese corpus.

6.2 Recall

We used the method described in subsection 5.2 to calculate the recall. 100 English sentences were labeled manually, obtaining 205 true collocations. Figure 10 shows the recall of the collocations in the N-best lists. From the figure, it can be seen that the trend on the English corpus is similar to that on the Chinese corpus, which indicates that our method is language-independent.

7 Discussion

7.1 The Effect of Fertility

In the MWA model as described in subsection 2.1, ϕ_i denotes the number of words that can align with w_i. Since a word only collocates with a few other words in a sentence, we should set a maximum number for ϕ, denote as $\phi_{\max}$.

In order to set $\phi_{\max}$, we examined the true collocations in the manually labeled set described in subsection 5.2. We found that 78% of words collocate with only one word, and 17% of words collocate with two words. In sum, 95% of words in the corpus can only collocate with at most two words. According to the above observation, we set $\phi_{\max}$ to 2.

In order to further examine the effect of $\phi_{\max}$ on collocation extraction, we used several different $\phi_{\max}$ in our experiments. The comparison

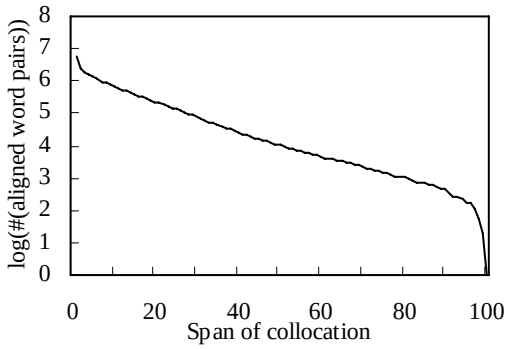


Figure 12. Distribution of spans

results are shown in Figure 11. The highest precision is achieved when $\phi_{\max}$ is set to 2. This result verifies our observation on the corpus.

7.2 Span of Collocation

One of the advantages of our method is that long-span collocations can be reliably extracted. In this subsection, we investigate the distribution of the span of the aligned word pairs. For the aligned word pairs occurring more than once, we calculated the average span as shown in Eq. (10).

$$AveSpan(w_i, w_j) = \frac{\sum_{s \in corpus} Span(w_i, w_j; s)}{freq(w_i, w_j)} \quad (10)$$

Where, $Span(w_i, w_j; s)$ is the span of the words w_i and w_j in the sentence s; $AveSpan(w_i, w_j)$ is the average span.

The distribution is shown in Figure 12. It can be seen that the number of the aligned word pairs decreased exponentially as the average span increased. About 17% of the aligned word pairs have spans longer than 6. According to the human evaluation result for precision in subsection 5.1, the precision of the long-span collocations is even higher than that of the short-span collocations. This indicates that our method can extract reliable collocations with long spans.

8 Conclusion

We have presented a monolingual word alignment method to extract collocations from monolingual corpus. We first replicated the monolingual corpus to generate a parallel corpus, in which each sentence pair consists of the two identical sentences in the same language. Then we adapted the bilingual word alignment algorithm to the monolingual scenario to align the

potentially collocated word pairs in the monolingual sentences. In addition, a ranking method was proposed to finally extract the collocations from the aligned word pairs. It scores collocation candidates by using alignment probabilities multiplied by a factor derived from the exponential function on the frequencies. Those with higher scores are selected as collocations. Both Chinese and English collocation extraction experiments indicate that our method outperforms previous approaches in terms of both precision and recall. For example, according to the human evaluations on the Chinese corpus, our method achieved a precision of 56.9%, which is much higher than that of the baseline method (29.0%). Moreover, we can extract collocations with longer span. Human evaluation on the extracted Chinese collocations shows that 69% of the long-span (>6) collocations are correct.

References

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, 19(2): 263-311.

Yaacov Choueka, S.T. Klein, and E. Neuwitz. 1983. Automatic Retrieval of Frequent Idiomatic and Collocational Expressions in a Large Corpus. *Journal for Literary and Linguistic computing*, 4(1):34-38.

Kenneth Church and Patrick Hanks. 1990. Word Association Norms, Mutual Information, and Lexicography. *Computational Linguistics*, 16(1):22-29.

Ted Dunning. 1993. Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*, 19(1): 61-74.

Stefan Evert. 2004. The Statistics of Word Cooccurrences: Word Pairs and Collocations. *Ph.D. thesis*, University of Stuttgart.

Dekang Lin. 1998. Extracting Collocations from Text Corpora. In *Proceedings of the 1st Workshop on Computational Terminology*, pp. 57-63.

Christopher D. Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*, Cambridge, MA; London, U.K.: Bradford Book & MIT Press.

Kathleen R. McKeown and Dragomir R. Radev. 2000. Collocations. In Robert Dale, Hermann Moisl, and Harold Somers (Ed.), *A Handbook of Natural Language Processing*, pp. 507-523.

Darren Pearce. 2001. Synonymy in Collocation Extraction. In *Proceedings of NAACL-2001 Workshop on Wordnet and Other Lexical Resources: Applications, Extensions and Customizations*, pp. 41-46.

Darren Pearce. 2002. A Comparative Evaluation of Collocation Extraction Techniques. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation*, pp. 651-658.

Violeta Seretan and Eric Wehrli. 2006. Accurate Collocation Extraction Using a Multilingual Parser. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics* (*COLING/ACL-2006*), pp. 953-960

Frank Smadja. 1993. Retrieving Collocations from Text: Xtract. *Computational Linguistics*, 19(1): 143-177.

Joachim Wermter and Udo Hahn. 2004. Collocation Extraction Based on Modifiability Statistics. In *Proceedings of the 20th International Conference on Computational Linguistics* (*COLING-2004*), pp. 980-986.