Recognition of Collocation Frames from Sentences

Xiaoxia LIU\(^{a}\), Degen HUANG\(^{b}\), Zhangzhi YIN\(^{c}\), Nonmembers, and Fuji REN\(^{d}\), Member

SUMMARY Collocation is a ubiquitous phenomenon in languages and accurate collocation recognition and extraction is of great significance to many natural language processing tasks. Collocations can be differentiated from simple bigram collocations to collocation frames (referring to distant multi-gram collocations). So far little focus is put on collocation frames. Oriented to translation and parsing, this study aims to recognize and extract the longest possible collocation frames from given sentences. We first extract bigram collocations with distributional semantics based method by introducing collocation patterns and integrating some state-of-the-art association measures. Based on bigram collocations extracted by the proposed method, we get the longest collocation frames according to recursive nature and linguistic rules of collocations. Compared with the baseline systems, the proposed method performs significantly better in bigram collocation extraction both in precision and recall. And in extracting collocation frames, the proposed method performs even better with the precision similar to its bigram collocation extraction results.

key words: collocation patterns, collocation frames, recursive nature of collocations, collocation rules

1. Introduction

Collocation is differently understood and defined in related literature, thus so far, no universal definition exists. In both fields of theoretical linguistics and computational linguistics exist different definitions of collocation, among which, two are representatives. Firth thinks “collocations of a given word are statements of habitual and customary places of that word.”\(^{[1]}\) In his opinion, a word can be defined and determined by its collocations. Manning believes that “a collocation is an expression consisting of two or more words that correspond to some conventional way of saying things.”\(^{[2]}\) By jointly considering the significance of collocations in language learning and natural language processing and the representative definitions, we define collocation as a phraseological unit consisting of two or more words that are syntactically related and corresponding to some habitual and conventional patterns. Although there is no universal definition for collocations, they do have some unanimous characteristics. First, collocations are prefabricated phraseological units. Whether collocations are correctly used directly determines whether the language is authentic or not. Second, collocations are arbitrary. They are not freely combined according to grammatical rules. Third, collocations are not predictable. One cannot predict a collocate for a word simply according to syntactic and semantic rules and the meaning of a collocation by simply putting the meaning of its collocating words together. Fourth, collocations are recurrent. They frequently appear in languages. Fifth, collocations are recursive. Long collocations may contain short collocations. Sixth, collocations are a universal linguistic phenomenon. They are pervasive in texts of all genres and domains and are said to represent the highest proportion of phraseological units. Seventh, collocations are culture dependent. Different languages have different collocating habits. The universality gets collocation a very important status in language related activities. But arbitrariness, unpredictability and culture-dependence make it a difficulty in those activities. Fortunately, the recurrence and recursive nature of collocations make it possible for researchers to study on them.

Research on collocation in natural language processing focuses on two related tasks: collocation extraction and collocation identification. Collocation identification is to discover the collocation tokens in the corpus while collocation extraction is to find the collocation types. Most collocation identification work is based on the collocation extraction results or the existing collocation resources (such as collocation dictionary). High quality collocation identification and extraction is crucial to many natural language processing tasks such as syntactic parsing, machine translation, text generation. In this study, we can extract and identify collocations in a given sentence simultaneously, so we use the verb recognize.

According to different criteria, collocations can be classified differently. According to the distance between collocating words, they can be classified into adjacent collocations and distant collocations; according to the word numbers that constitute collocations, they can be classified into bigram collocations and multi-gram collocations. Most collocation extraction and identification work so far focuses on bigram and adjacent multi-gram collocations. But researches seldom focus on distant multi-gram collocations.

But in application, especially in translation, no matter whether it is human translation or machine translation, accurate identification of collocations, especially multi-gram
distant collocations is of great significance. Why is it so? We know that in translation the first important step is to decide translation unit. A translation unit is a segment of a text which the translator (human or machine) treats as a single cognitive unit in the source language for the purpose of finding an equivalence in the target language. It may be a single word, a phrase, one or more sentences, or even a larger unit. And it is said that when a translator segments a text into translation units, the larger the units are, the better chance there is to obtain an idiomatic translation. This is true not only in human translation, but also in cases where computer-assisted translation, such as translation memories, is used, and where translations are performed by automatic machine translation systems. Because of this theory, statistical machine translation is developed from word based to phrase based (continuous word combinations, not grammatical phrases) and to syntactic based method. This development witnesses changes of translation unit in different phases. Collocations can be regarded as a special translation unit. Bigram and adjacent collocations are easy to be learned no matter which machine translation model is used. The most collocation related problems are incurred by distant multi-gram collocations.

This study focuses on the identification of these multi-gram distant collocations, which we call collocation frames. The second part includes some related work on collocation extraction. The third part defines important terms in our study and the fourth part explains the proposed collocation frame identification method in detail. In the fifth part, we detail the experiments from preparation, setup, evaluation, results and result analysis. Finally, conclusions and further work are drawn in part six.

2. Related Work

Collocation extraction so far adopts a two phase method, with the first phase to discover the candidates, the second to filter the noises. To discover collocation candidates, there are two main methods in the literature. One is window based and the other is syntactic based. In window based method, a fixed window is set for the word in study, and only those inside the window are regarded as possible candidates [4]–[6]. But obvious problem exists in window based method, that is, candidates outside the window cannot be extracted. To solve this problem, a new method emerged, which is based on syntactic parsing. The syntax used in collocation extraction is dependency. There are two ways to use dependency in collocation extraction. Some extract dependency parsing results directly as candidates [7], [8], [11]–[15], others transform dependency parsing results into dependency tree and extract high frequent sub-trees as candidates [9], [10]. The advantage of dependency tree based method is language independent and with no distance and word number limits.

No matter which method is used in discovering collocation candidates, the method used for candidate filtering is by using association measures. The only difference for different work lies in the choice of association measures. There are altogether more than eighty association measures reported. The differences between different measures are not obvious in collocation extraction. The only difference lies in that some method is a little bit better in extracting certain type of collocations. So in order to improve collocation extraction results researchers tend to change from only using one association measure to combining several association measures. They test different association measures and select and combine some most appropriate ones in their collocation extraction work [9], [16]–[21]. From the literature review, we find several association measures that are most appropriate to collocation extraction task. They are t-test, mutual information and log likelihood ratio. We call them traditional association measures.

Almost all the state-of-the-art collocation extraction methods are based on dependency, and the collocation types they focus are bigram collocations and adjacent multi-gram collocations. The problems are: first, according to our analysis, in the dependency relations of Chinese, only one fifth of them are collocations. That is to say, dependency based collocation candidate discovering introduces a great ratio of noises which makes the filtering very challenging. Secondly, traditional association measures are based on the co-occurrence frequency, which cannot depict the syntactic and semantic relations between words, so no matter how you change and combine them, you cannot solve the filtering problem. Thirdly, the existing collocation extraction methods focus mainly on bigram and adjacent multi-gram collocations, but seldom take collocation frames into consideration. But in application, especially in machine translation and parsing, bigram and adjacent multi-gram collocations can be easily learned, whereas collocation frames are the reason for problems. But collocation frame related researches are a few. Seretan et al. [8] try to extend the extracted bigram collocations into multi-gram collocations by using the recursive characteristic of collocations. Martens and Vandeghinste [9] use dependency tree to extract collocations that include collocation frames. Kato et al. [22] can also extract collocation frames out. But extraction of collocation frames is seldom a study focus and there is no specialized evaluation of methods on collocation frames.

To solve the three problems, we propose a collocation frame discovering method at sentence level. To solve the first problem, we introduce collocation patterns in dependency based collocation candidates discovering phase. Thus, a great number of candidates are filtered by collocation patterns. In the filtering phase, we propose to use word embedding based similarity to compute the association between words. Because word embedding, which uses low dimensional vector to denote a word and can grasp the semantic and syntactic meaning of the words, can be a better perspective to measure the relations between words. Liu and Huang [15] have proved the practicability of word embedding based similarity to be an association measure and tested it in a pilot collocation extraction and proved its effectiveness and complementariness to other association measures. So, in the proposed method, we use word embedding based
similarity ranking combining the three best traditional association measures to measure the relations between words. Thirdly, from extracted bigrams, by considering both the recursive nature of collocations and linguistic rules, we recursively extend bigrams to collocation frames.

3. Definition

In our proposed method, we introduce collocation patterns in bigram collocation extraction, and put forward a concept of collocation frames. We aim to extract collocation frames from a given sentence by using the recursive nature and rules of collocations. So in this part, we will make clear these concepts one by one.

3.1 Collocation Patterns

A collocation is a group of syntactically related words. This group of words should conform to a certain combination pattern according to Sinclair [23]. The combination pattern of collocations can be represented by the part-of-speech of the words in collocation. And these combination patterns are designated as collocation patterns in this study. For instance, in the collocation “提高能力 (to improve the ability)”, the part-of-speech for “提高 (improve)” is verb, for “能力 (ability)” is noun, so the collocation pattern for this collocation is “verb-noun”. Modern Chinese Collocation Dictionary of Content Words provides collocation patterns for adjectives, nouns and verbs. Collocation patterns for nouns include “noun-verb”, “noun-noun”, “noun-adjective”, “verb-noun”, “adjective-noun” and “quantifier-noun”[24]. Collocation patterns for verbs include “noun-verb”, “verb-noun”, “verb-adjective”, “verb-verb” and “adjective-verb”. Collocation patterns for adjectives include “verb-adjective”, “adjective-verb”, “noun-adjective”, and “adjective-quantifier”. Many collocation extraction researches focus on collocations of certain collocation patterns [8], [14].

3.2 Collocation Frames

Collocations can be classified according to different criteria. When the distance between collocating words is considered, they can be classified into adjacent collocations and distant collocations. When the number of words in a collocation is considered, they can be classified into bigram collocations and multi-gram collocations. Those collocations both distant and multi-gram are referred to as collocation frames in this study. For example, in Example 1, “人际交流 (human communication)” is an adjacent bigram collocation. “增强 * 能力 (to strengthen the ability)” is a distant bigram collocation. “数据充分表明 (data strongly suggested)” is an adjacent multi-gram collocation. “增强 * 在 * 方面的能力 (to strengthen the ability of *)” is a distant multi-gram collocation, thus, collocation frame.

Example 1: 研究人员妮娜克劳斯说这些数据充分表明，音乐训练过程中所产生的神经元连接能增强大脑在人际交流其它方面的能力。（Researcher Nina Kraus said the data strongly suggested that the neural connections made during musical training also primed the brain for other aspects of human communication.)

3.3 Recursive Nature of Collocations

In most studies, collocations are limited to bigrams, but theoretically, collocations should have no length limit [23]. In fact, in different definitions of collocation, a collocation is defined as two or more word combination. Collocations with more than two words are ubiquitous. And according to Heid [25], the components of a collocation can again be collocational themselves, which he refers to as the recursive nature of collocations. Conversely, we can say that many multi-gram collocations are made up of bigram collocations. For example, in Example 1, the collocation “数据充分表明 (data strongly suggested)” is made up of “数据表明 (data suggested)” and “充分表明 (strongly suggested)” and the collocation “增强 * 在 * 方面的能力 (to strengthen the ability of *)” is made up of “在 * 方面 (in the aspect)”, “增强 * 能力 (to strengthen the ability)” and “方面能力 (ability in the aspect)”. The recursive nature of collocations is the theoretical base for our collocation frame extraction work.

3.4 Collocation Rules

Except for syntagmatic relations between collocating words, we collected a group of rules according to the linguistic knowledge and based on statistical methods. These rules are called collocation rules and can be classified into two types. One type is used to filter bigram collocations and referred to as collocation filtering rules and the other is used to extend bigram collocations to multi-gram collocations and referred to as collocation extending rules. These rules are listed in Table 1, in which, the trigger collocations, trigger conditions and the corresponding treatments are described.
in detail. For example, in example 1, the collocation “增强了在*方面的能力” is recursively extended by “*方面 (in the aspects)”, “增强了能力 (to strengthen the ability)” and “方面能力 (ability in the aspect)” with “的 (of)” is extended according to collocation extending rule number 3. Because in bigram collocation extraction, in order to decrease computational complexity, we do not consider function words which are in huge quantity. But in extending bigrams to multi-grams, function words work as necessary cohesions and need to be considered.

4. Algorithm

The extraction of collocation frames in a given sentence is based on the bigram collocation extraction results. Thus, in order to extract collocation frames, we have to first retrieve bigrams in the sentence. The traditional two-phase collocation extraction method is used but with improvement in each phase.

**Algorithm 1** Bigram Collocation Extraction Algorithm (2GCEA)

<table>
<thead>
<tr>
<th>Input</th>
<th>The sentence S = (W₁, W₂, · · ·, Wₙ). Word embedding file and word frequency file, Collocation Patterns.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Output</td>
<td>Bigram collocations in S.</td>
</tr>
</tbody>
</table>

- Dependency parse S
- Initialize a list to save candidates in S: CandColl = []
- Extract dependency relations from S that are in accordance with collocation patterns and add them in CandColl.
- For every collocation candidate (W_A, W_B) in CandColl, suppose its pattern is (POS_A, POS_B).
  - Count the frequencies of W_A and W_B in S as FreqA and FreqB respectively.
  - Initialize a list to save words in POS_A in S, ListA = []
  - Initialize a list to save words in POS_B in S, ListB = []
  - For each word in POS_A do
    - If the part-of-speech of Word is POS_A do
      - Add it in ListB
    - Elif the part-of-speech of Word is POS_B do
      - Add it in ListA
    - End if
  - End for
  - Compute DS-BAM between W_A and each word in ListA and put it in SimA.
  - End for
  - Compute DS-BAM between W_B and each word in ListA and put it in SimB.
  - End for
  - Rank candidate (W_A, W_B) in SimA and SimB as RankA and RankB
  - If RankA < FreqA or RankB < FreqB then
    - Put candidate (W_A, W_B) in TrueCol
  - End if
  - End for

The first phase is to discover collocation candidates.

We use both dependency relations and collocation patterns to do this task. The theoretical base for this is: on the one hand, although dependency relations can capture the grammatical relations between words, the experimental data show that in all the dependency relations, only one fifth are collocations, the other four fifths are noises; on the other hand, only words in certain type can be collocations, so, it is feasible to use parts-of-speech to generalize collocations. Collocation part-of-speech combinations are called collocation patterns. By using dependency relations and collocation patterns in candidate discovering can decrease the noise and improve the quality of candidates. Collocation patterns in this work are summed up from collocation dictionary [24], including collocation patterns of words in the part-of-speech as prepositions, nouns, verbs and adjectives. Dependency relations are extracted by using Stanford natural language processing toolset CoreNLP [26].

**Algorithm 2** Collocation Frame Extraction Algorithm (CFEA)

<table>
<thead>
<tr>
<th>Input</th>
<th>The sentence S = (W₁, W₂, · · ·, Wₙ). Bigram collocation list in S.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Output</td>
<td>Collocation frames in S.</td>
</tr>
</tbody>
</table>

- Initialize a list to save candidates extracted, CoList = []
- For each bigram collocation in S do
  - If it cannot be extended then
    - Add it in CoList
  - Else
    - For each word in the bigram collocation do
      - Find collocations that include this word and extend the bigram collocation
      - Delete these collocations from bigram collocation list
      - Continue this step until no word can be extended
    - Add extended collocation in CoList
  - End for
  - End if
  - End for

The second phase is to filter noises in the candidates. In our approach, we propose an integrated ranking method, that is, on the basis of distributional semantics based association measure we integrate three best suitable traditional association measures. To use distributional semantic based association measure is because it is proved to be effective in sentence level collocation extraction, and better than state-of-the-art association measures [27]. To integrate traditional association measures is because different association measures play different roles in collocation extraction, and they are complementary to each other. In our approach, we use distributional semantics based similarity computation as the main association measure and combine three other state-of-the-art association measures including t-test, mutual information and log likelihood ratio. The formula for distributional semantics based similarity computation (DSBAM) is:

\[
DSBAM(W_A, W_B) = \frac{\text{Vector}(W_A) \ast \text{Vector}(W_B)}{||\text{Vector}(W_A)|| \ast ||\text{Vector}(W_B)||}
\]

The improvement we make on bigram collocation extraction is as follows: first, we introduce collocation patterns in collocation candidates discovering. Only those dependency relations in accordance with collocation patterns...
are regarded as candidates. Secondly, we also improved distributional semantics based association measure which uses three ranking lists and compares all the word pairs in the sentence but in our proposed method, only two ranking lists are kept and the candidates in the list are chosen according to collocation patterns. Thirdly, we use the frequency of the candidate words as threshold for judging the real collocation which is different for every ranking list and every collocation instead of a fixed empirical value. Finally, we make improvement in bigram noise filtering. There are altogether more than 80 association measures, and those often used in collocation extraction are t-test, mutual information, log likelihood ratio and word-embedding based similarity computing. Different methods have different advantages and can complement each other. We combine the four often used measures in collocation extraction in order to take full advantage of the complementarity, and only when a candidate is decided as true by two measures will it be extracted by our method. In this way, not only can we avoid those errors incurred by certain method and improve the precision, but also we can take advantage of complementarity of different methods and extract some results that certain method cannot and thus improve the recall.

The detailed bigram collocation extraction algorithm is shown in Algorithm 1. Based on bigram collocation extraction results, we make full use of the recursion nature and linguistic rules of the collocations in extending collocation frames. By its recursive nature, we extend the extracted bigrams to multi-grams, and at the same time consider the linguistic rules. In our study, linguistic rules are got by linguistic experts according to statistics and they are grouped into two types. One is collocation filtering rules, the other is collocation extending rules. At the end of the approach, only the longest collocation frames in the sentence are output. Algorithm 2 shows the detail.

5. Experiments and Analysis

5.1 Experimental Preparation

The aim of this study is to extract collocations from a given sentence, including bigrams, multi-grams and collocation frames. The extraction of collocation frames is based on bigram extraction results. Thus, the whole procedure includes two parts: bigram discovering and collocation frame extending. Bigram discovering needs two tools, one is dependency parser provided by Stanford CoreNLP toolkit[26] and the other is word embedding training model by using gensim word2vec model, the python version of word2vec [28].

In training word embedding using word2vec, several important parameters need to be considered. We choose an optimal parameter value set which is got empirically by [27]. They include skip-gram model, window size being 10, vector size being 600, and iteration time being 30. The resources necessary for this method are a training corpus to train word embedding and get word frequencies and a testing corpus to evaluate the performance of the method. The training corpus is provided by [15], the Chinese part of a self-compiled parallel corpus of Chinese and English, which includes 457,899 Chinese sentences and 8.4 million words and almost covers most commonly used words in modern Chinese. The testing corpus is also provided by [15], which is a set of 600 sentences including “能力 (ability)” in each sentence and in which all “能力 (ability)” related collocations are tagged manually.

5.2 Evaluation Metrics

The proposed method is used to extract not only bigram collocations but also multi-grams and collocation frames. So, we evaluate the method in two perspectives. For bigrams, there is a lot of complete work in literature that is comparable with our work. We choose Varro as baseline1 [29], since it is dependency based, the same as our approach, and the extracted results record the source of the collocations thus they can be easily located into specific sentences which are comparable to our sentence level collocation extraction, besides, it can extract both bigrams and multi-grams. Another baseline (baseline2) we choose is the latest collocation extraction work since our method is improved based on it[15]. The evaluation metrics we choose are the commonly used ones: Precision (P), Recall (R), and F1-value (F1).

$$P = \frac{\text{accurately extracted bigrams}}{\text{all extracted bigrams}}$$  \hspace{1cm} (2)

$$R = \frac{\text{accurately extracted bigrams}}{\text{all bigrams in the test set}}$$  \hspace{1cm} (3)

$$F1 = \frac{2PR}{P + R}$$  \hspace{1cm} (4)

For multi-grams, since different studies define them differently, it is difficult to get a fair competition between different methods. So on the one hand, we invite linguistic experts to tag the extracted results as true or false collocation in order to compute the Accuracy (A), on the other hand, since in the test set, the multi-grams have been labeled, so we can also evaluate our result by Precision (P), Recall (R) and F1-value (F1).

$$A = \frac{\text{True multi-grams tagged by experts}}{\text{all extracted multi-grams}}$$  \hspace{1cm} (5)

$$P = \frac{\text{accurately extracted multi-grams}}{\text{all extracted multi-grams}}$$  \hspace{1cm} (6)

$$R = \frac{\text{accurately extracted multi-grams}}{\text{all multi-grams in the test set}}$$  \hspace{1cm} (7)

$$F1 = \frac{2PR}{P + R}$$  \hspace{1cm} (8)

5.3 Experimental Results

The evaluation on bigram collocation extraction results is shown in Table 2. It is clearly that by integrating both collocation patterns and collocation rules, all the three evaluating indicators show significant improvement.
Table 2 Evaluation of bigram collocation extraction

<table>
<thead>
<tr>
<th>Methods</th>
<th>P</th>
<th>R</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline1</td>
<td>0.310</td>
<td>0.451</td>
<td>0.367</td>
</tr>
<tr>
<td>Baseline2</td>
<td>0.321</td>
<td>0.480</td>
<td>0.385</td>
</tr>
<tr>
<td>Ours</td>
<td>0.368</td>
<td>0.527</td>
<td>0.433</td>
</tr>
</tbody>
</table>

Table 3 Evaluation of multi-gram collocation extraction

<table>
<thead>
<tr>
<th>Methods</th>
<th>A</th>
<th>P</th>
<th>R</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline1</td>
<td>0.688</td>
<td>0.343</td>
<td>0.281</td>
<td>0.309</td>
</tr>
<tr>
<td>Ours</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The evaluation on multi-grams is shown in Table 3. The multi-gram collocation extraction is based on the extraction results of bigram collocations. As in extracting bigrams, some sentences may have no bigram extracting results, thus no multi-gram results. If we include all multi-grams in all test sentences, the precision is 34.3%, and recall is 28.1%. But if we only consider the test sentences that have bigram extracting results, the precision of multi-gram collocation extraction is 58.4% and the recall is 48.2%.

Because of its recursive characteristics, tagging the collocation frames in test set is influenced. Thus in evaluating, to only compare the extracting result with the test set is not enough. So we resort to manual evaluation method too. Linguistic experts are asked to decide if the extracted multi-grams are correct or not and we calculate the accuracy for our method. The accuracy is 68.8%. In extracting multi-grams, our method aims to extract the longest collocation frames in the sentences, but the baseline system only extracts multi-grams. So in comparing, both the precision and recall of the baseline system are quite low. In order to prove that our method is really good, we compare the extracting results of the two methods in detail. That is, we count the adjacent trigrams, distant trigrams, adjacent four-grams, distant four-grams, adjacent five- and more-grams and distant five- and more-grams extracted by each method. The statistics in Table 4 show that baseline system mostly extracted adjacent tri- and four-grams and it can extract a few distant tri- and four-grams, but it cannot extract more-than-four-grams, while our method can extract more distant multi-grams. This elucidates the superiority of our method over the baseline.

5.4 Error Analysis

Experimental results show that both the precision and recall of dependency-based collocation extraction methods are quite low. By comparing the dependency relations and the collocations in the test set, we find that in all the collocations in the test set, only 72.3% of them are in dependency relations, the other 27.7% have no dependency relations. That is to say, in all the collocations in the test set, more than one fourth of them cannot be extracted by dependency-based method. This problem is caused on the one hand by the performance of the dependency parser and on the other hand by the inequality between dependency relations and collocations. This problem greatly affects the recall of dependency-based collocation extraction method.

One factor that affects the performance of the proposed method, especially on the extraction of multi-grams, is the recursive characteristic of collocations. Collocations are recursive, that is to say, long collocations include shorter collocations. Thus a problem exists in defining our longest collocation frame. Where should be the delimiter? This problem exists in labeling the testing set. Different linguistic experts have different opinions on the longest collocation frame in the sentence. And this problem stands out in evaluating the extracted results. By comparing, we can see that some extracted results are either longer or shorter than the collocations in the test set and evaluated as wrong results but in themselves are complete in meaning and can be regarded as correct collocations. For example, in the extracted collocation frames, there are “影响*对*的探测能力(（infuence the inspection ability of))” and “强的抗*能力((strong anti-* ability))”, but the tagging results in the sentences are “严重影响对*的探测能力(badly influence the inspection ability of)” and “具备*强的抗冲击能力(be well able to take the hit)”.

Another factor that influences the tagging and evaluating of collocations is to what degree it is generalized. In tagging the multi-gram collocation frames, some parts should be generalized. Then the problem appeared: to what degree should this collocation be generalized? For instance, in Example 2, two multi-grams “拥有*优良的*能力(（have good ability of))” and “具有*理想的*能力(（have ideal ability of))” are labeled as the collocation frames. But the extracting results of our method is “拥有*优良的*能力(（have good *hardening ability of))” and “具有*理想的*能力(（have ideal hardening ability of))”. It seems both the labeled collocations and the extracted ones are correct and the only difference lies in the degree of generalization. But in evaluating, these two results are judged as wrong. More often, some extracted results are estimated as wrong because of both the factors mentioned above. For example, we extract a collocation frame “得益*出*的*协调能力(（owe *to extraordinary *coordination))” from a sentence in which the tagged collocation frame is “出*手眼协调能力(（extraordinary hand-eye coordination))”. So, in computing, we put all the wrong extracted collocations because of these two factors together and find that in all the extracted results, almost one fourth (that is 24.1%) belong to this situation, and most of them are distant four-and-more-gram collocations. This great number has significant influence on...
the precision and recall of our method.

Example 2: DP 钢拥有更加优良的加工硬化能力, TRIP 钢则具有较为理想的烘烤硬化能力。(DP steel has better work hardening ability while TRIP steel has better bake hardening ability)

6. Conclusion

Accurate recognition and extraction of collocation is of great significance to many natural language processing tasks. This study aims to recognize and extract collocations in a given sentence. The major contributions of this study are: First, we improve the distributional semantics based collocation extraction method by introducing collocation patterns both in candidate discovering phase and in candidate filtering phase. Secondly, we make improvement in bigram noise filtering. Four different methods are combined in order to take full advantage of the complementarity between them. Thirdly, we enhance the multi-gram collocation extraction performance of the system. We introduce collocation frame into our system and extract them by extending the bigram collocation extraction results recursively according to certain collocation rules. Experimental results show that the proposed method does a very good job in multi-gram collocation extraction and compared with baseline it has significant improvement in all the three evaluation metrics, precision, recall and f-value.

But there are still some problems which affect the performance seriously. First, it’s unequal between dependency relations and collocations and because of this more than one fourth collocations cannot be extracted by dependency-based method. Then, there are inconsistencies in testing set collocation tagging and multi-gram extraction results because of recursive characteristic of collocations and generalization degree of collocation frames. In the future, we will improve the proposed method by solving the above problems. For the first problem, by using collocation patterns, select those word pairs that are not in dependency relations but in accordance to certain collocation patterns as candidates too. For the problem of recursion and generalization, we will try to introduce generalization into the method in preprocessing phase and extract collocations in generalized sentences.

Acknowledgments

This work has been supported in part by the National Natural Science Foundation of China under Grant 61672127.

References


Xiaoxia Liu received her M.A. in foreign linguistics and applied linguistics from Dalian University of Technology in 2011 and she is now a Ph.D. candidate in computer application technology of Dalian University of Technology. Her research interests include natural language processing and computational linguistics.

Degen Huang received the B.S. degree in computer science from Fuzhou University, China, in 1986, the M.S. and Ph.D. degrees in computer science from the Dalian University of Technology, China, in 1988 and 2004 respectively. He is currently a Professor with the School of Computer Science, Dalian University of Technology. His research interests include natural language processing, machine learning, and machine translation. He is now a senior member of CCF, CIPS, ACM, CAAI and an associate editor of Int. J. Advanced Intelligence.

Zhangzhi Yin received his Master of Computer Science in Dalian University of Technology, the main research areas are named entity recognition and information extraction.

Fuji Ren was born in 1959, Ph.D. He is currently a Professor with Department of Information Science & Intelligent Systems, Tokushima University, 2–1 Minamijosanjima, Tokushima, 770–8506 Japan. His current research interests include natural language processing, machine translation, artificial intelligence, language understanding and communication, multi-lingual multi-function multi-media intelligent systems, super-function methodology.