Stepwise Acquisition of Verbal Collocations Considering Nest Structure

Sho Machida^{a,*}, Shiho H.Nobesawa^b, Shuichi Arai^b

^aTokyo City University Graduate Division, 1-28-1 Tamazutsumi, Setagaya-ku, Tokyo 158-8557, Japan

^bTokyo City University, 1-28-1 Tamazutsumi, Setagaya-ku, Tokyo 158-8557, Japan

*Corresponding Author: machida15@ipl.cs.tcu.ac.jp

Abstract

Collocation is a natural sequence of terms used habitually. By collecting collocation comprehensively, it can be used as a resource for automatic elaboration and proofreading of sentence. It can also be used as a resource for the analysis of precise meaning. The acquisition of collocations have been a significant task in natural language processing, and the acquisition of expressions which contain exchangeable terms such as the first noun in the sequence {尊敬(reverence), の念を覚える(feel the sense of)} has not been successful enough. Therefore, to clarify the range of collocations and their related term groups, we focus on verbs as the central unit of collocations and estimate the collocation ranges stepwise. By using our system, we succeeded in estimating the collocation ranges as n -grams ($n > 8$). The extraction accuracy of this system was 80%, which was 20% better than the extraction accuracy by related work.

Keywords: Automatic acquisition, collocation, verbal phrase

1. Introduction

Collocation has various definitions⁽¹⁾, and in this paper, we define a collocation as a natural sequence of terms used habitually⁽²⁾. By collecting collocations comprehensively, it can be used as a resource for automatic elaboration and proofreading of sentences⁽³⁾. It can also be used as a resource for analysis of precise meaning⁽⁴⁾. In collocations, we define the sequence of natural terms with the term to co-occur for certain verb as a verbal collocation. For example, {覚える(remember/feel)} co-occur with many terms, and among which {念(sense), を(postpositional particle to make the prior term an object for a verb), 覚える(feel)} is a verbal

collocation. This is because the sequence of terms {を (particle), 覚える(feel, verb)} is used habitually, and it becomes a natural and meaningful sequence of terms with a preceding noun to be an object for the verb. In addition, {念(sense), を (particle), 覚える(feel)} is used as a part of a phrase {*(wildcard), の(of, particle), 念(sense), を (particle), 覚える(feel)}. Considering this, we find that co-occurrence relationship of terms and phrases should also be acquired as a verbal collocation such as {尊敬(reverence), の(of, particle), 念(sense), を (particle), 覚える(feel)} which is used habitually. Therefore, {尊敬(reverence), の(of, particle), 念(sense), を (particle), 覚える(feel)} can be regarded as two verbal collocation nested, that is, a verbal collocation {尊敬(reverence), の(of, particle), 念(sense), を (particle), 覚える(feel)} contains a smaller verbal collocation {念(sense), を (particle), 覚える(feel)}. Considering the co-occurrence relationship between terms and phrases, our system estimates the range of verbal collocations. In this paper, we focus on the acquisition of co-occurrence of both phrases and terms.

2. Related Work

Suttbs⁽⁵⁾ proposed a classification standard for English collocations. Sonoda et al. used the classification standard for Japanese collocations extraction^(6,7). In Japanese, case particles, not words order, determine the sentence structure; for example, {首相(prime minister), が(is, particle), アメリカ(the United States), に(to, particle), 訪れる(visit)} expresses the same meaning with {アメリカ(the United States), に(to, particle), 首相(prime minister), が(is, particle), 訪れる(visit)}¹. So, when there are two relationships {首相(prime minister), 訪れる(visit)} and {アメリカ(the United States), 訪れる(visit)}, both sentences

¹ Both {首相,が,アメリカ,に,訪れる} and {アメリカ,に,首相,が,訪れる} mean “the prime minister visits the US.”

should be a natural terms combinations. In Japanese particles come after nouns and verbs come after particles, and we focus on content words. Also, it is meaningless to acquire collocations that “run” or “fly” follows a noun such as “car” or “airplane,” and “eat” is preceding “food” such as “bread” or “rice”. It is necessary to characterize that a term preceding {食べる (*eat*)} is “food”. Sonoda et al. use n -Xgram to acquire collocations for content words. n -Xgram is an independent term of n terms that appeared next to each other⁽⁸⁾. They acquire verbal collocations from the simultaneous co-occurrence probability between two terms including verbs. However, since co-occurrence between only two terms is captured, it is difficult to acquire co-occurrence relationships between terms and phrases even if co-occurrence relationships between terms can be acquired. Takayama et al. used pointwise mutual information (PMI: equation (1))⁽⁸⁾ as a theory to acquire verbal collocations⁽⁹⁾. PMI is a theory to measure the cohesion between terms⁽⁹⁾. The cohesion of terms and terms is the high probability that terms and terms appear simultaneously.

$$PMI(x, y) = \log_2 \frac{P(x, y)}{P(x)P(y)} \quad (1)$$

$P(x, y)$ is the probability that elements x and y appear at the same time. $P(x)$ is the probability that an element x appears. When the PMI is a positive value ($PMI(x, y) > 0$), x and y should appear together. When the PMI is a negative value ($PMI(x, y) < 0$), it is not likely that x and y appear together. PMI can be judged that it has a stronger tendency when the absolute value of the positive value and the negative value is larger. When a simple simultaneous co-occurrence probability $P(x, y)$ is used for extraction of verbal collocations, combinations of frequently co-occurring terms such as particles can be extracted. On the other hand, PMI is effective to estimate the collocations boundaries.

3. Co-occurrence Relationship between Term and Phrase

Sonoda et al. proposed an extraction method of verbal collocations using n -Xgram and generalization^(6,7). However, we consider that particles include significant information. This is because following verbs are variable depending on particles when focusing on nouns⁽³⁾. A co-occurrence of words should not be always dealt as a collocation. Thus, we define a collocation as follows:

- 1 A sequence of terms such as {尊敬(*reverence*), の念を覚える(*feel the sense of*)} in which the preceding and succeeding phrases for verb are incomplete.
- 2 By preceding and succeeding phrases, a sequence of terms such as {怒る(*get angry*), かも, しれない(かもしれない, *might be*)} exactly including cases where the verb is limited to negative or passive forms.
- 3 A sequence is not incomplete for terms that describe a verb such as {しみ, じみ(しみじみ, *keenly*), 感じる(*feel*)}.

Japanese language contains morphemes in terms as the smallest language unit remaining meaning⁽¹⁰⁾. A term is constructed as a sequence of several morphemes. In collocations extraction, co-occurrence relationships between terms are extracted from a corpus as a sequence of morphemes. However, when considering the co-occurrence relationships between terms and phrases, the ranges of the verbal collocations are not clear enough. Thus, we investigate the relationships between terms and phrases focusing on verbs. For the sequence of terms and phrases such as {尊敬(*reverence*), の念を覚える(*feel the sense of*)}, we use equation (2) to calculate the verb bigram (w_{i+1}, w_i). And we use the equation (3) to calculate PMI of a trigram.

$$PMI(w_{i+1}, w_i) = \log_2 \frac{P(w_{i+1}, w_i)}{P(w_{i+1})P(w_i)} \quad (2)$$

$$PMI(w_{i+2}, w_{i+1}, w_i) = \log_2 \frac{P(w_{i+2}, w_{i+1}, w_i)}{P(w_{i+2})P(w_{i+1})P(w_i)} \quad (3)$$

Table 1 shows examples of PMI values of a verb and its preceding terms. Here we name a verb bigram as a sequence of a verb and a term preceding it, and a verb trigram as a sequence of a verb and two following terms. Table 1 shows that although the relationship becomes lower with {を (particle)} preceding {習い(*acquire*), 覚える(*learn*)}, it raises with considering {念(*sense*)} before {を (particle), 覚える(*feel*)}. We found that the higher PMI values of such verb bigram and verb trigram seems likely to be a sequence of terms. In addition, {念を覚える(*feel the sense*)} is a sequence which is a component of {尊敬の念を覚える(*feel the sense of reverence*)}. We can consider it to have acquired a base of a verbal collocation. For these reasons, we treat the sequence of terms with higher PMI values of verb bigram and verb trigram as the bases of verbal collocations. Also, considering collocation bases, we can estimate the collocation ranges.

Table 1: Relationship between a Verb and its Preceding Terms

$\{w_{i+1}, w_i\}$ {term 1, verb}	PMI value	> <	$\{w_{i+2}, w_{i+1}, w_i\}$ {term2, term 1, verb}	PMI value
{習い, 覚える}	8.87	>	{を, 習い, 覚える}	3.34
{を, 覚える}	2.38	>	{顔, を, 覚える}	1.85
{を, 覚える}	2.38	<	{念, を, 覚える}	5.13
{も, 堪る}	3.32	<	{楯, も, 堪る}	6.42

Figure 1: Co-occurrence Relationship between a Verbal Collocation and its Preceding Terms

By comparing the PMI values of the verb bigram and the verb trigram, it becomes a clue to estimate the ranges of verbal collocations. In Figure 1, these should be restriction in part-of-speeches around verbal collocation bases. From these facts, we can expect that it is possible to acquire highly likely verbal collocations by considering the relationship between the base verbal collocations and their surrounding terms.

4 Acquisition of Verbal Collocations

Considering Sonoda’s collocations acquisition method of autonomous terms⁽⁶⁾ and our acquisition method of verb collocations based on PMI, we insist that verbal collocations should be estimated not only by the co-occurrence relationships between terms but also by the co-occurrence relationships between terms and phrases. Thus in this paper we propose a stepwise method to acquire verbal collocations focusing on nest structure. There is no limit on the number of terms in a verbal collocation, so it is necessary to acquire an acquisition method not limited by the number of constituent terms of verbal collocations. Our system to acquire verbal collocations gradually is shown in Figure 2. From the survey, it is possible to gradually acquire the components of verbal collocations based on a sequence with high bond strength, and it is possible to acquire highly probable collocations. From the sequences of a few terms co-

Figure 2: Verbal Collocations Acquisition Method

occurrence with verbs, we estimate the ranges of verbal collocations by using the threshold of the co-occurrence probability. When extracting terms preceding the sequence of terms, the survey found that the limited kinds of part of speeches should appear. So we extract the preceding terms as part of verbal collocations using part of speech conditions.

4.1 Acquisition of the Verbal Collocation Bases

To acquire verbal collocations, our stepwise method first acquire the bases of verbal collocations. We extract verb bigrams and verb trigram, and we measure the bond strength between terms. Our system compares the bond strength of verb bigrams and verb trigrams to find verbal collocation bases. Sequences of terms judged as verb trigrams predicts the components of verbal collocations by preceding the verb trigram and using the stepwise processing of the relationship with terms preceding and following the verb trigram. Sequence of terms judged as verb bigram estimates constituent elements of verbal collocations using stepwise process, with relationship to the terms following verb bigram. Our system acquires verbal collocations using term sequence obtained by stepwise processing. For example, a sequence of terms {の (of, particle), 念 (sense), を (particle), 覚える (feel)} can be considered a part of a verbal collocation {尊敬 (reverence), の (of, particle), 念 (sense), を (particle), 覚える (feel)}. Our system calculates the probability of co-occurrence of preceding and succeeding terms for the verbal collocation bases.

4.2 Stepwise Acquisition of Verbal Collocations

To find the range of a verbal collocation, our system explores the verbal collocations gradually. Suppose that the component of the base s is a sequence of words $w_i \dots w_j$, and let the number of components of s be $length(s)$. Here w_{j+1} co-occur with a certain probability for a base s . $length(s) + 1$ or more that is considered to be the base ($w_i \dots w_{j+1}$), and let the constituent elements of s be $w_i \dots w_{j+1}$, and subsequent s search for the term that follows.

Table 2: Examples of Verbal Collocation Bases

Base	description
{ら, なく, なる} {させ, て, もらう}	Sequence of auxiliary verb.
{の, 念, を, 覚える} {の, そば, に, 住ん, で, いる} {から, 鼻, へ, 抜ける} {の, 前, に, 呼び, ださ, れる}	sequences starting with a particle without a noun preceding them.

Table 3: Part-of-Speeches Preceding the Bases

POSeS	Example	Reason
Noun	{心残り, を, 覚える} { }	Since many of the methods of using verbs are sequence of {noun, particle, verb}
Adverb	{はじめて, 覚える}	Adverbs are terms that modify verbs, because of their deep relationship with verbs
Noun, particle, particle	{いつ, まで, も, 覚える}	Because it is a sequence of natural terms with high probability
Prefix	{両, 腕を広げる}	Although it does not appear more frequently than the above-mentioned part of speech system, there are cases where the part of speech of the term preceding by the verbal collocations is only the prefix
verb	{食べ, させてもらう}	It is a part of speech that is easy to precede verbs and bases, because it is a sequence of natural terms with high probability
Adjective	{重, さを測る}	Parts of speech that are likely to appear before {さ} suffix

When the probability for w_{j+1} to precede s is smaller than a certain probability for s , treat it as a part of the base ($w_i \dots w_{j+1}$) composed of $length(s) + 1$. For example, sequence starting with a particle should take preceding terms to be verbal collocations (Table 2). Continue this

Table 4: Number of Verbal Collocations by Base Size

n -gram	Distinguished Number of Collocations
bigram	1,681
trigram	22,638
4-gram	528
n -gram ($n > 4$)	132
total	26,568

stepwise process until we acquire all bases. With this process, the range of the verbal collocations can be found. From the relationships of the preceding terms to the base and the succeeding terms, the example sequences of terms obtained by stepwise processing is shown in Table 2. As shown in Table 2, preceding and succeeding terms were acquired by stepwise processing with respect to the base. The base of 4-gram is easy to be followed by {れる(passive)}, {られる(passive)}, and the base of trigram and 5-gram has not only passive but also a lot of bases including negative terms such as {ない(negative form)}, {ぬ(negative form)} were acquired. Such as {の念を覚える(*feel the sense of*)} cannot be said to be natural term sequence because they become a verbal collocation in a sequence with the preceding terms. Therefore, our system acquires verbal collocations as bases of n -grams obtained by gradual processing. Our system acquires verbal collocations from the bases of extracted n -grams.

However, it is necessary to extract a sequence of natural terms and phrases, so it is not necessary to acquire all the terms preceding the bases. It is unknown which sequence is natural in terms of the number of co-occurrences of terms preceding a verb, and the range of a verbal collocation. {覚える(*remember/feel*)} tends to be following a particle, even among them it is likely to appear with a sequence of a noun, a particle and a verb, such as {心残り(*regretful*), を(particle), 覚える(*feel*)}. From these facts, it is understood that the part of speech information serves as a clue for extracting verbal collocations. However, since it is necessary to acquire verbal collocations for all bases of n -grams, we need to select versatile part-of-speeches. As a result of investigating a sequence of part-of-speeches of a natural term preceding the base, we acquire a term preceding part-of-speeches. Table 3 is showing the part-of-speech system of the term preceding the base and the reason for acquiring it. We acquire the verbal collocations for the part of speech of Table 3. It is a subjective but a survey result of the part of speech information of terms preceding the base. We obtained verbal collocations using these methods.

Table 5: Examples of Extracted Verbal Collocations

Preceding terms				Bases
よく	情を	念を	不安を	覚える
いろいろと	急に	政府に	将来に 対する	不安を 覚える
尊敬	後悔	感謝	憎悪	の念を 覚える

Table 6: Extraction Accuracy by Base Size

n -gram	Dictionary	Dic+manual
bigram	58%	80%
trigram	67%	83%
4-gram	48%	72%
n -gram($n>4$)	70%	85%
Total	58%	80%

5 Results

We used 2,526,860 sentences from Aozora Bunko Corpus whose noise removed, and we acquired verbal collocations using our system. The threshold of appearance probability of preceding and succeeding terms relative to the base ($w_i \dots w_j$) was determined experimentally to 80%. We show the number of the unlike and total number of acquired verbal collocations in Table 4. Verbal collocations were extracted using the part-of-speeches defined in Table 3. The verbal collocations acquired as trigram were the most frequent in our result. A part of the extracted verbal collocations is shown in Table 5. Table 5 shows that co-occurrence relationships of terms and phrases can be acquired as well as co-occurrence relationships between terms. In addition, it is possible to acquire verbal collocations of both {念(*sense*), を覚える(*feel*, particle and verb)} and {尊敬(*reverence*), の念を覚える(*feel the sense of*)}. It can be said that nest verbal collocations has been acquired. Since terms preceding {を覚える(*feel*, particle and verb)} and {の念を覚える(*feel the sense of*)} can also be extracted, it can be said that they have achieved nest verbal collocations, which is the object of this method. For the evaluation method, we use a collocation dictionary created by hand⁽¹¹⁾. The collocations dictionary contains about 600,000 collocations acquired from novels, magazines, and newspapers. It describes the sequence of natural terms and we use it as answer set. As using the collocation dictionary alone is insufficient for evaluation, we also tried manual evaluation according to the restriction of collocations. Since it is difficult to verify manually, we extracted verbal collocations from randomly-selected 1,000 sentences. Table 7 shows the results of evaluating verbal collocations acquired by the proposed

Table 6: Examples of Incorrect Verbal Collocations

Output	Answer	Meaning
{たち, の, 意見, に, よる}	{~, たち, の, 意見, に, よる}	According to ~ opinion
{やって, やる, れる}	{やって, やる, れる, ない}	I cannot take this anymore
{かも, 知れる}	{~, かも, 知れる, ない}	It might be ~
{見当, が, つく}	{皆目, 見当, が, つか, ない}	I have no idea

method using the answer set. The average of accuracy of verbal collocations was 58%. The accuracy of verbal collocations of trigram and n -gram ($n>4$) was high, and the accuracy of verbal collocations of 4-gram was low. As a result of presenting an acquisition method not confined to the number of constituent terms of verbal collocations, it was possible to find 8-gram verbal collocations from verbal collocations of bigram. There was no verbal collocations over 8-gram.

6 Discussion

The dictionary of verbal collocations used a dictionary with a high percentage of accuracy was 58%. By using restriction with dictionary improving accuracy up to 80%. I could show that it was possible to acquire not only the relationship between terms and terms but also the relationship between terms and phrases. The accuracy of n -gram is 85% and natural sequence of terms can be acquired, showing the effectiveness of the acquisition method, which is not captured by the number of constituent terms of verbal collocations. In order to improve the accuracy, we consider the collocations of verbs that was incorrect in the evaluation experiment. Table 7 shows examples of incorrect output verbal collocations. There are two possible reasons why the verbal collocation shown in Table 7 was incorrect. The first is setting of the threshold value used for acquiring verbal collocations. Thresholds were experimented with 100%, 90% and 80%, and the highest threshold value 80% was used when evaluating. In the case of 100%, the total number of acquired verbal collocations was 18,591, and the average of accuracy of verbal collocations was 72%. When compared with the case of 80%, the numerical value became low together with the total number of acquisition of verbal collocations and average of accuracy of verbal collocations. In the case of 90%, the total number of acquired verbal collocations was 22,183, and the average of accuracy of verbal collocations was 75%. Also, as compared with the

case of 80%, the numerical value became low together with the total number of acquisition of verbal collocations and average of accuracy of verbal collocations. In the case of 70%, the total number of acquired verbal collocations was 31,562. Also, as compared with the case of 80%, the numerical value became low together with the total number of acquisition of verbal collocations and average of accuracy of verbal collocations. In the case of 70%, the total number of acquired verbal collocations was 31,562. And the average of accuracy of verbal collocations was 68%. Compared to the case of 80%, the numerical value of the total acquisition number of the verbal collocation naturally rises, but the average of accuracy of the verbal collocations was low. Therefore, the experiment was carried out using the threshold value 80% at which the average of accuracy was the highest. However, because we only refer to the appearance probability of preceding terms, we ignored co-occurrence with terms with low probability of occurrence. For that reason, for example, although {皆目(*all/totally*)} precedes by 16% against {見当(*aim/guess*), が³(*particle*), つく(*have, verb*)}, our system cannot guess that {皆目(*all/totally*), 見当(*aim/guess*), が³(*particle*), つく(*have*)} is always followed by {ない¹(*no, negative form*)}. Secondly, since the collocation dictionary was created manually, it lacks in comprehensiveness and consistency. Because the collocation dictionary is constructed for research purposes, it is not a dictionary for computers, and it is published as a clue for people to proofread. Therefore, there are cases where incompatibilities become incorrect because natural term sequence is not described.

7 Conclusions

We proposed a stepwise method to acquire verbal collocations considering co-occurrence relationships of terms and phrases. We proposed an acquisition method that is not confined to the number of constituent terms of verbal collocations, the experimental results include that the accuracy of verbal collocations is 80% as a whole, 22% higher than Sonoda's result. In Table 5, there are {よく(*well.*)}, {念を(*sense, none and particle*)}, and so on. As natural term sequence that co-occur with {覚える(*remember/feel*)}. Also, we can acknowledge that a series of words such as {将来に対する(*about the future*)} precedes by {不安を覚える(*feel uneasiness*)}, and in the evaluation of human hands, unnatural words do not appear in the acquired verbal collocations. Thus we conclude that our method succeeded in the acquisition of verbal collocations effectively enough.

References

- (1) Kazundo Fukushima: "Definitions of Collocations and Their Usefulness : From the Viewpoint of Writing Collocations Dictionaries," written in Japanese, Information and Communication Studies, Vol. 36, pp. 253-282, 2007
- (2) Takumi Sonoda, and Takao Miura : "Data Mining for Japanese Collocation," Vol. 1, No. 2, pp. 20-35, 2010
- (3) Tanomura Tadaharu: "Retrieving Collocational Information from Japanese Corpora: An Attempt Towards the Creation of a Dictionary of Collocations," written in Japanese, Osaka Univ. Nihongo-kenkyu Vol. 21, written in Japanese, pp. 21-41, 2009
- (4) Hiroki Oda, and Kenji Kita : "Automatically Extracting Collocations Using Words Position Information in Corpora," IPSJ SIG Technical Reports, Vol. 1997-NL-85, pp. 75-82, 1997
- (5) Michael Stubbs : "Collocations and Semantic Profiles : On The Cause of The Trouble with Quantitative Studies," written in Japanese, Functions of Language, Vol. 2, pp. 23-55, 1995
- (6) Takumi Sonoda, and Miura Takao : "Data Mining for Japanese Collocation," DEIM Forum, Vol. E10, No. 2, 2012
- (7) Takumi Sonoda, Satoshi Shimada, and Takao Miura : "Extracting Collocational Information from Japanese Corpus Using Statistical Indicator," Forum on Information Technology, Vol. E-021, pp. 201-206, 2012
- (8) Hiroya Takamura : "Introduction to Machine Learning for Natural Language Processing," written in Japanese, Corona Publishing, 2015
- (9) Hiroki Takayama, Yoshihide Kato, Tomohiro Ohno, Shigeki Matsubara, and Yoshiharu Ishikawa : "Collocation Extraction Using a PMI-based Association Measure for Dependency Tree Pattern," The 10th International Symposium on Natural Language Processing (SNLP 2013), pp. 136-141, 2013
- (10) Manabu Okumura: "Introduction to Natural Language Processing," written in Japanese, Corona Publishing, 2010
- (11) Hajime Onai: "Te ni o ha jiten," written in Japanese, Sanseido, 2015