Vol.2013-IFAT-109 No.3
Vol.2013-NL-210 No.3
2013/1/11

# Collocation Suggestion for Japanese Second Language Learners

LIS W. K. PEREIRA[†1,a)] ERLYN MANGUILIMOTAN[†1,b)]
YUJI MATSUMOTO[†1,c)]

***Abstract:*** This study addresses issues of Japanese language learning concerning word combinations (collocations), which despite they can form sentences grammatically correct, they can sound unnatural. We analyze correct word combinations using different collocation measures and word similarity methods. Our analysis includes the use of a large Japanese language learner corpus for generating collocation candidates, in order to build a system that is more sensitive to constructions that are difficult for learners. Our results show that we get better precision and recall rates compared to other methods that use only well-formed text.

## 1. Introduction

Automatic grammatical error correction is emerging as an interesting topic of natural language processing (NLP). However, previous research in second language learning focused on restricted types of learners' errors, such as article and preposition errors. Only recently natural language processing research has addressed issues of collocation errors.

Collocations are conventional word combinations in a language. In Japanese, お茶を入れる and 夢を見る are examples of collocations. Even though their accurate use is crucial to make communication precise and to sound like a native speaker, learning them is one of the most difficult tasks for second language learners. For instance, a Japanese language learner may write:

文化を分かるために日本語を勉強している。

(I am studying Japanese to understand the culture.)

However, the correct sentence should be:

文化を理解するために日本語を勉強している。

(I am studying Japanese to understand the culture)

Although both sentences are syntactically correct and have the same meaning, the first one has an unnatural expression, '文化を分かる'. Such combinations can be quite confusing for Japanese second language learners, for which an application to assist in choosing the right collocation would be useful.

So far, most research in collocation error correction has relied on resources of limited coverage, such as dictionaries, thesauri, or manually constructed databases to generate the correction candidates [16], [5], [12], [19]. Better scope was offered by machine translation approaches, which is based on learners' first language (L1), yet unique systems have to be constructed for learners of different L1s [1], [3]. Another problem is that most research does not actually take learners' tendency of collocation errors into account; instead, their systems are trained only on well-formed text corpora. Moreover, most of these research works were done for English language learners, and so far, there is no available system to assist Japanese language learners.

In this work, we analyze various Japanese corpora using a number of collocation and word similarity measures to deduce and suggest the best collocations for Japanese second language learners. In order to build a system that is more sensitive to constructions that are difficult for learners, we use word similarity measures that generate collocation candidates using a large Japanese language learner corpus. By employing this approach, we could obtain a better recall compared to other methods that use only well-formed text.

The remainder of the paper is organized as follows. In Section 2, we introduce work related on collocation error correction. Section 3 explains our method, based on word similarity and association measures, for suggesting collocations. In Section 4, we describe different word similarity and association measures, as well as the corpora used in our experiments. Finally, we show the results of our experiments in Section 5 and in Section 6 and point out the future directions for our research in Section 7.

## 2. Related Work

Collocation correction currently follows a similar approach used in article and preposition correction. The general strategy compares the learner's word choice to a confusion set generated from well-formed text during the training phase. If one or more alternatives are more appropriate to the context, the learner's word is flagged as an error and the alternatives are suggested as corrections [10]. To constrain the size of the confusion set, similarity measures are used, considering that we are dealing with substitution of open class words (nouns, verbs, adjectives and adverbs). To rank the best candidates, we measure the strength of association in the learner's construction and in each of the generated alternative construction.

For example, [5] generated synonyms for each candidate string using Wordnet and Roget's Thesaurus and used rank ratio measure to score them by their semantic similarity. [12] also used Wordnet to generate synonyms, but used Pointwise Mutual Information as an association measure to rank the candidates. [1] used bilingual dictionaries to derive collocation candidates and used the log-likelihood measure to rank them. One drawback of these approaches is that, they rely on resources of limited coverage, such as dictionaries, thesaurus or manually constructed databases to generate the candidates. Other studies have tried to offer better coverage by automatically deriving paraphrases from parallel corpora [3], but similar to [1], it is essential to identify the learner's first language and to have bilingual dictionaries and parallel corpora for L1 in order to extend the resulting system.

Our work follows the general approach, that is, use similarity measures for generating the confusion set and association

[1] Nara Institute of Science and Technology
a) lis-k@is.naist.jp
b) erlyn-m@is.naist.jp
c) matsu@is.naist.jp

measures for ranking the best candidates. However, instead of using only well-formed text for generating the confusion set, we use a large learner corpus created from the revision log of a language learning SNS, Lang-8[1]. Another work that also uses data from Lang-8 is [14], which uses it for creating a large-scale Japanese learner's corpus.

## 3. Combining Word Similarity and Association Measures to Suggest Collocations for Japanese Second Language Learners

We combine word similarity measures for generating the confusion set and association measures for ranking the candidates. We also used different corpora in these combinations.

Our work is focused on suggestions for verb collocation errors. Using a dependency parser (Cabocha[2]), we automatically extracted from a large Japanese learner corpus 269 noun-verb collocations(noun and verb paired with the particle を), with incorrect verbs together with their correction (described in sub-section 5.2). For each extracted noun を verb tuple in the second learner's composition, we created a set of candidates using word similarity algorithms. Then, we measured the strength of association in the writer's phrase and in each generated candidate phrase using association measures.When one of the candidates suggested by the system matches the correction given in the corpus, this candidate is added in the result. Figure 1 illustrates the method used in this study.

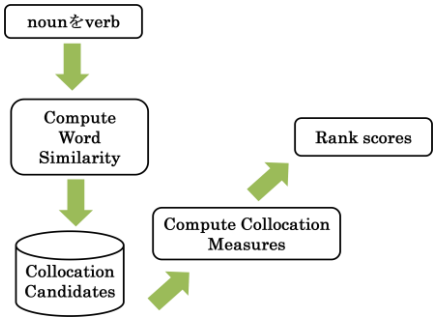


**Figure 1** Word Similarity and Association Measures combination method

## 4. Approaches to Word Similarity and Word Association Strength

### 4.1 Word Similarity

Similarity measures are used to generate the collocation candidates that are later ranked using association measures.

A common approach in similarity measure is to find words that are analogous to the writer's choice [12], [19], [15]. In our work, we analyze two common word similarity measures: thesaurus-based word similarity and distributional similarity.

**4.1.1** Thesaurus-based word similarity

The intuition of this measure is to check if the given words have similar glosses (definitions). Two concepts are considered similar if they are near each other in the thesaurus hierarchy (have a short path between them).

In this work, we used Bunrui Goi Hyo [18], a Japanese thesaurus, which has a vocabulary size of around 100,000 words. We used it to compute word similarity, taking the words that are in the same sub tree as the candidate word.

**4.1.2** Distributional similarity

Thesaurus-based methods produce weak recall since many words, phrases and semantic connections are not covered by hand-built thesauri, especially for verbs and adjectives. As an alternative, distributional models are often used since it gives higher recall. On the other hand, distributional models tend to have lower precision [6], because the candidate set is larger.

The intuition of this algorithm is that two words are similar if they have similar word contexts. In our task, context will be defined by some grammatical relation, specifically, 'noun-verb' relation. Two words that have similar parse contexts can be assumed to have similar meaning.

In our work, context is represented using co-occurrence vectors that are based on syntactic dependencies. The similarity between co-occurrence vectors is computed using cosine, which is a generally accepted metric. Other methods include the KL divergence [9] and the Jenson-Shannon divergence [11]. The cosine similarity formula is represented as follows:

$$\cos(\vec{v},\vec{w}) = \frac{\vec{v}\bullet\vec{w}}{|\vec{v}||\vec{w}|} = \frac{\vec{v}}{|\vec{v}|}\bullet\frac{\vec{w}}{|\vec{w}|} = \frac{\sum_{i=1}^{N} v_i w_i}{\sqrt{\sum_{i=1}^{N} v_i^2}\sqrt{\sum_{i=1}^{N} w_i^2}} \quad (1)$$

Here, the two vectors $\vec{v}$ and $\vec{w}$ indicate the counts of times a particular verb occurs in a noun-verb relation with other nouns.

In computing the cosine similarity, three main resources were used in this work:

1) Mainichi Shimbun Corpus [13], one of the major newspapers in Japan that provides raw text of newspaper articles used as linguistic resource.
2) Balanced Corpus of Contemporary Written Japanese (BCCWJ Corpus) [7], a balanced corpus of one hundred million words of contemporary written Japanese. Portions of the BCCWJ corpus used in our experiments include magazine, newspaper, textbooks, and blog data. Incorporating a variety of topics and styles in the training data helps to minimize the domain gap problem between the learner's vocabulary and newspaper vocabulary found in the Mainichi Shimbun data.
3) Lang-8 Corpus, a large-scale Japanese learner data set which was created by crawling the revision log of a language learning SNS, Lang-8. It contains pairs of learner's sentence and its correction given by native speakers of Japanese language. For computing cosine similarity, we used only the correction data, the data containing only the corrected sentences.

**4.1.3** Confusion set derived from learner corpus

In order to build a module that can "guess" common

[1] www.lang-8.com
[2] http://chasen.org/ taku/software/cabocha/

construction errors, we created a confusion set using Lang-8 corpus. Instead of generating words that have similar meaning to the learner's written construction, we extracted *all* the possible *verb corrections* for each of the verb found in the data. For example, for the verb 届く, the confusion set is composed of verbs such as 届ける, 送る, もらう, meaning that in the corpus, 届く was always corrected by one of these verbs, i.e., when the learner writes the verb 届く, he/she might actually mean to write verbs 届ける, 送る, or もらう.

### 4.2 Word Association Strength

After generating the collocation candidates using word similarity, the next step is to identify the "true collocations" among generated candidates. Here, the association strength was measured, in such a way that word pairs generated by chance from the sampling process can be excluded. An association measure assigns an association score to each word pair. High association score indicates strong association, and can be used to select the "true collocations". For our work, we adopted Weighted Dice coefficient [8] as our association measurement. We also tested using other association measures (results are omitted for this report): Pointwise Mutual Information [2], log-likelihood ratio [4] and Dice coefficient [17], but Weighted Dice performed best. The resources we used for computing collocation score are: Mainichi Shinbum data, BCCWJ corpus and Lang-8 corpus (2010 year data).

## 5. Experiments and Results

### 5.1 Experiment setup

In computing word similarity and association scores, we used: a) Mainichi Shimbun Corpus; b) Bunrui Goi Hyo Corpus; c) BCCWJ Corpus and d) Lang-8 Corpus.

1) Mainichi Shimbun Corpus: One year of Mainichi Shimbun newspaper data (1991) was used to extract the noun-verb pairs to compute word similarity (using cosine similarity metric) and collocation scores. We extracted around 220,000 pairs composed of 16,000 unique verbs and 37,000 unique nouns.
2) Bunrui Goi Hyo Thesaurus: This thesaurus was used to compute word similarity, taking the words that are in the same sub tree as the candidate word.
3) BCCWJ Corpus: We extracted 194,036 noun-verb pairs composed of 43,243 unique nouns and 18,212 unique verbs. This data is necessary to compute the word similarity (using cosine similarity metric) and collocation scores.
4) Lang-8 Corpus: Consisted of two year data (2010 and 2011).
   A) Year 2010 data, which contains 1,288,934 pairs of learner's sentence and its correction, was used to:
      i) Compute word similarity (using cosine similarity metric) and collocation scores: We took out the learners' sentences and used only the correction data, the data containing only the corrected sentences. We extracted 163,880 noun-verb pairs composed of 38,999 unique nouns and 16,086 unique verbs.
      ii) Construct the confusion set (explained in sub-section 4.1.3 ): We constructed the confusion set for all the 16,086 verbs that appeared in the data.
   B) Year 2011 data was used to construct the test set (described in sub-section 5.2).

### 5.2 Test set selection

We used Lang-8 (2011 data) for selecting our test set. First we extracted all the **noun を verb** constructions with incorrect verbs and their correction. From the pairs extracted, we selected the ones where the verbs were corrected to the same verb 5 or more times by the native speakers. Table 1 shows some examples of the extracted **noun を verb** pairs.

| Noun を Verb (Learner) | Noun を Verb (Correction) | Frequency |
|---|---|---|
| 質問を**聞く** | 質問を**する** | 208 |
| 写真を**取る** | 写真を**撮る** | 108 |
| 日記を**書く** | 日記を**書ける** | 96 |
| 試験を**取る** | 試験を**受ける** | 91 |
| 勉強を**続く** | 勉強を**続ける** | 89 |
| 試験を**する** | 試験を**受ける** | 84 |

**Table 1** Examples of noun-verb pairs where the verbs were corrected to the same verb alternative 5 or more times by the native speakers.

One problem of the above selection criterion is that there are cases where the learner's construction sounds more acceptable than its correction. For example, cases such as 日記を書く and its correction 日記を書ける. 日記を書く sounds more correct than 日記を書ける. However in the corpus, it was corrected due to some contextual information. One example for that case is shown below:

Learner's sentence: 最近ちょっと忙しいから、***日記を書きません***.
(I have been a bit busy lately, so I don't write my diary)
Sentence correction:最近ちょっと忙しいから、***日記を書けません***.
(I have been a bit busy lately, so I can't write my diary)

For our application, since we are only considering the noun, particle and verb that the learner wrote, there was a need to filter out such contextually induced corrections. To solve this problem, we used the Weighted Dice coefficient to compute the association strength between the noun and all the verbs, filtering out the pairs where the learner's construction has a higher score than the correction. For the example above, 日記を書く got higher score than 日記を書ける, hence that pair was excluded from our test set. After applying those conditions, we selected 269 pairs for our test set.

### 5.3 Evaluation

We compared the verb suggested by the system with the human suggested verb in the Lang-8 data. A match would be counted as a true positive.

### 5.4 Experiment Results

Table 2 shows the ten models derived from combining word similarity measures and association measures and using different corpus. In the table, the brackets [] indicate what corpus was used. Table 3 shows the precision of k-best suggestions, Table 4 shows the recall rate and Table 5 shows the F-score values for each model.

| Model | Word Similarity+Association Strength method considered |
|---|---|
| M1 | Thesaurus Similarity+ WD(Weighted Dice) [Mainichi Shimbun] |
| M2 | (Cos)Cosine Similarity + WD [Mainichi Shimbun] |
| M3 | Cos + WD [BCCWJ] |
| M4 | Cos + WD [Lang-8] |
| M5 | Cos + WD [Mainichi Shimbun+BCCWJ] |
| M6 | Cos + WD [BCCWJ+Lang-8] |
| M7 | CS(Confusion Set from Lang-8)+WD[BCCWJ] |
| M8 | CS+WD[Lang-8] |
| M9 | CS+WD[Mainichi Shimbun+BCCWJ] |
| M10 | CS+WD[BCCWJ+Lang-8] |

**Table 2** Models of Word Similarity and Association Strength method combination.

| K-Best | M1 | M2 | M3 | M4 | M5 | M6 | M7 | M8 | M9 | M10 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | **0.9473** | 0.5040 | 0.5792 | 0.5945 | 0.4545 | 0.5945 | 0.6973 | 0.6791 | 0.6766 | 0.6716 |
| 2 | **1** | 0.7154 | 0.7500 | 0.7104 | 0.6060 | 0.7297 | 0.8544 | 0.8358 | 0.8120 | 0.8358 |
| 3 | **1** | 0.8048 | 0.8170 | 0.7876 | 0.6926 | 0.7876 | 0.9118 | 0.8992 | 0.8721 | 0.8955 |
| 5 | **1** | 0.9186 | 0.9024 | 0.8648 | 0.7922 | 0.8725 | 0.9770 | 0.9514 | 0.9436 | 0.9552 |
| 10 | **1** | 0.9918 | 0.9634 | 0.9498 | 0.8874 | 0.9459 | 0.9961 | **1** | 0.9924 | **1** |

**Table 3** The precision rate of Model 1-10

| M1 | M2 | M3 | M4 | M5 | M6 | M7 | M8 | M9 | M10 |
|---|---|---|---|---|---|---|---|---|---|
| 0.1412 | 0.4572 | 0.6096 | 0.9628 | 0.8587 | 0.9628 | 0.9702 | **0.9962** | 0.9888 | **0.9962** |

**Table 4** The recall rate of Model 1-10

| K-Best | M1 | M2 | M3 | M4 | M5 | M6 | M7 | M8 | M9 | M10 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.2458 | 0.4795 | 0.5940 | 0.7351 | 0.5944 | 0.7351 | **0.8114** | 0.8076 | 0.8035 | 0.8023 |
| 2 | 0.2475 | 0.5579 | 0.6725 | 0.8175 | 0.7106 | 0.8302 | 0.9086 | **0.9090** | 0.8917 | **0.9090** |
| 3 | 0.2475 | 0.5831 | 0.6982 | 0.8664 | 0.7667 | 0.8664 | 0.9401 | **0.9452** | 0.9268 | 0.9432 |
| 5 | 0.2475 | 0.6105 | 0.7277 | 0.9112 | 0.8241 | 0.9154 | 0.9736 | 0.9733 | 0.9656 | **0.9753** |
| 10 | 0.2475 | 0.6259 | 0.7467 | 0.9562 | 0.8728 | 0.9543 | 0.9830 | **0.9981** | 0.9906 | **0.9981** |

**Table 5** The F-score of Model 1-10

The highest values are shown in bold type. Table 3 shows that M1 achieved the highest precision rate among the other models; however, it had the lowest recall, as seen in Table 4. The recall was low because the confusion set generated using the thesaurus did not include the correction suggested in Lang-8 data for most cases. In order to improve the recall rate, we generated models M2-M6 using distributional similarity and also using other corpora than Mainichi Shimbun corpus to minimize the domain gap problem between the learner's vocabulary and newspaper vocabulary found in the Mainichi Shimbun data. The recall rate improved significantly, but the precision rate decreased. The best results are achieved when using Lang-8 data for generating the confusion set (M7-M10). The best F-score value for k=1was achieved by M7, which uses Lang-8 data for generating the confusion set and BCCWJ for computing collocation scores. Regarding the effects of corpus size, M10, which uses BCCWJ and Lang-8 for computing collocation scores, provides the highest recall rate, together with M8, which uses Lang-8 data for generating the confusion set and for computing collocation scores.

## 6. Discussion

Model M1 could suggest cases such as the ones shown below (Table 6):

| Noun を Verb (Learner) | Noun を Verb (Correction) | K-Best |
|---|---|---|
| 仕事を**変わる** | 仕事を**変える** | 1 |
| 計画を**作る** | 計画を**立つ** | 1 |
| 体を**動く** | 体を**動かす** | 1 |
| 日本語を**独学する** | 日本語を**勉強する** | 1 |
| 薬を**食べる** | 薬を**飲む** | 1 |

**Table 6** Suggestions given by M1

M1 can suggest such cases because the wrong verb written by the learner and the correction suggested in Lang-8 data have similar meaning, being also near each other in the thesaurus hierarchy. In other words, the confusion set generated using the thesaurus includes the correction suggested in Lang-8 data.

However, for cases such as the ones shown in Table 7, M1 could not suggest any correction, since the wrong verb written by the learner and the correction suggested in Lang-8 data do not have similar meaning. The models M6 and M7, for example, suggested the correction among the 10 best ranked candidates.

| Noun を Verb (Learner) | Noun を Verb (Correction) | K-Best (M6) | K-Best (M7) |
|---|---|---|---|
| ご飯を**作る** | ご飯を**炊く** | 2 | 1 |
| スープを**食べる** | スープを**飲む** | 1 | 1 |
| 仕事を**働く** | 仕事を**する** | 9 | 3 |
| 試験を**参加する** | 試験を**受ける** | 1 | 1 |
| 大学を**出る** | 大学を**卒業する** | 1 | 1 |
| 夢を**する** | 夢を**見る** | 4 | 3 |

**Table 7** Suggestions given by M6 and M7

## 7. Concluding Remarks

In this report, we analyzed correct word combinations using different collocation measures and word similarity methods. The best results were achieved when using a large learner corpus, Lang-8, for generating the confusion set, fine-tuning the system before training only on well formed text, to become more sensitive to constructions that are difficult for learners. In this work, we only examined **noun**を**verb** constructions; in order to verify our approach and improve our current results, many other construction types are considered for future work.

**Acknowledgments** Special thanks to Yangyang Xi for maintaining Lang-8.

## References

[1] Y. C. Chang, J. S. Chang, H. J. Chen, and H. C. Liou. An automatic collocation writing assistant for Taiwanese EFL learners: A case of corpus-based NLP technology. *Computer Assisted Language Learning*, 21(3):283–299, 2008.

[2] K.Church, and P. Hanks. 1990. Word Association Norms, Mutual Information and Lexicography, *Computational Linguistics*, Vol. 16:1, pp. 22-29.

[3] D. Dahlmeier, H. T. Ng. 2011. Correcting Semantic Collocation Errors with L1-induced Paraphrases. *EMNLP 2011*: 107-117.

[4] T. Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics* 19.1 (Mar. 1993), 61-74.

[5] Y. Futagi, P. Deane, M. Chodorow, and J. Tetreault. 2008. A computational approach to detecting collocation errors in the writing of non-native speakers of English. *Journal of Computer-Assisted Learning*, 21.

[6] D. Jurafsky, Daniel, and James H. Martin. 2009. *Speech and Language Processing: An Introduction to Natural Language Processing*, Speech Recognition, and Computational Linguistics. 2nd edition. Prentice-Hall.

[7] Kikuo Maekawa. 2008. Balanced corpus of contemporary written japanese. *In Proceedings of the 6th Workshop on Asian Language Resources (ALR)*, pages 101–102.

[8] M. Kitamura, Y. Matsumoto.1997. Automatic extraction of translation patterns in parallel corpora. In *IPSJ*, Vol. 38(4), pp.108-117, April 1997. In Japanese.

[9] S. Kullback, R.A. Leibler. 1951. On Information and Sufficiency. *Annals of Mathematical Statistics* **22** (1): 79–86.

[10] C. Leacock, M. Chodorow, M. Gamon and J. Tetreault. 2009. *Automated Grammatical Error Detection for Language Learners*. Morgan & Claypool Synthesis Lectures on Human Language Technologies, volume 9.

[11] L. Lee. 1999. Measures of Distributional Similarity, *Proc of the 37th annual meeting of the ACL*, Stroudsburg, PA, USA, 25

[12] A. L. Liu, D. Wible, and N. L. Tsao. 2009. Automated suggestions for miscollocations. In *Proceedings of the NAACL HLT Workshop on Innovative Use of NLP for Building Educational Applications*.

[13] Mainichi Newspaper Co. 1991. Mainichi Shimbun CD-ROM 1991.

[14] T. Mizumoto, K. Mamoru, M. Nagata, Y. Matsumoto. 2011. Mining Revision Log of Language Learning SNS for Automated Japanese Error Correction of Second Language Learners. In *Proceedings of the 5th International Joint Conference on Natural Language Processing*, pp.147-155. Chiang Mai, Thailand, November 2011.

[15] R. Östling and O. Knutsson. 2009. A corpus-based tool for helping writers with Swedish collocations**,** *Proceedings of the Workshop on Extracting and Using Constructions in NLP*, Nodalida, Odense, Denmark.

[16] C. C. Shei and H. Pain. 2000. An ESL writer's collocational aid. *Computer Assisted Language Learning*, 13.

[17] F. Smadja, K. R. Mckeown, V. Hatzivassiloglou. 1996. Translation collocations for bilingual lexicons: a statistical approach. *Computational Linguistics*, 22:1-38.

[18] The National Institute for Japanese Language, editor. 1964. Bunrui-Goi-Hyo. Shuei shuppan. In Japanese.

[19] J. C. Wu, Y. C. Chang, T. Mitamura, and J. S. Chang. 2010. Automatic collocation suggestion in academic writing. *In Proceedings of the ACL 2010 Conference Short Papers*.