# Automated Collocation Suggestion for Japanese Second Language Learners

Lis W. Kanashiro Pereira Erlyn Manguilimotan Yuji Matsumoto

Nara Institute of Science and Technology Department of Information Science {lis-k, erlyn-m, matsu}@is.naist.jp

#### Abstract

This study addresses issues of Japanese language learning concerning word combinations (collocations). Japanese learners may be able to construct grammatically correct sentences, however, these may sound "unnatural". In this work, we analyze correct word combinations using different collocation measures and word similarity methods. While other methods use well-formed text, our approach makes 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 results compared to other methods that use only wellformed text.

## **1** Introduction

Automated grammatical error correction is emerging as an interesting topic of natural language processing (NLP). However, previous research in second language learning are focused on restricted types of learners' errors, such as article and preposition errors (Gamon, 2010; Rozovskaya and Roth, 2010; Tetreault et al., 2010). For example, research for Japanese language mainly focuses on Japanese case particles (Suzuki and Toutanova, 2006; Oyama and Matsumoto, 2010). It is only recently that NLP research has addressed issues of collocation errors.

Collocations are conventional word combinations in a language. In Japanese, *ocha wo ireru* "お茶を入れる<sup>1</sup> [to make tea]" and *yume wo miru*"夢を見る<sup>2</sup> [to have a dream]" 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, the Japanese collocation *yume wo miru* [lit. to see a dream] is unpredictable, at least, for native speakers of English, because its constituents are different from those in the Japanese language. A learner might create the unnatural combination *yume wo suru*, using the verb *suru* (a general light verb meaning "do" in English) instead of *miru* "to see".

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 result compared to other methods that use only wellformed text.

The remainder of the paper is organized as follows. In Section 2, we introduce related work 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. The experimental setup and the results are described in Sections 5 and 6, respectively. Section 7 points out the future directions for our research.

## 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. To constrain the size of the confusion set,

<sup>&</sup>lt;sup>1</sup> lit. to put in tea

 $<sup>^{2}</sup>$  lit. to see a dream

similarity measures are used. To rank the best candidates, the strength of association in the learner's construction and in each of the generated alternative construction are measured.

For example, Futagi et al. (2008) generated synonyms for each candidate string using WordNet and Roget's Thesaurus and used the rank ratio measure to score them by their semantic similarity. Liu et al. (2009) also used WordNet to generate synonyms, but used Pointwise Mutual Information as association measure to rank the candidates. Chang et al. (2008) used bilingual dictionaries to derive collocation candidates and used the loglikelihood 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 (Dahlmeier and Ng, 2011), but similar to Chang et al. (2008), it is essential to identify the learner's first language and to have bilingual dictionaries and parallel corpora for every first language (L1) in order to extend the resulting system. Another problem is that most research does not actually take the learners' tendency of collocation errors into account; instead, their systems are trained only on well-formed text corpora. Our work follows the general approach, that is, uses similarity measures for generating the confusion set and association measures for ranking the best candidates. However, instead of using only wellformed text for generating the confusion set, we use a large learner corpus created by crawling the revision log of a language learning social networking service (SNS), Lang-8<sup>3</sup>. Another work that also uses data from Lang-8 is Mizumoto et al. (2011), which uses Lang-8 in creating a largescale Japanese learner's corpus. The biggest benefit of using such kind of data is that we can obtain in large scale pairs of learners' sentences and their corrections assigned by native speakers.

# 3 Combining Word Similarity and Association Measures to Suggest Collocations

In our work, we focus on suggestions for noun and verb collocation errors in "noun wo verb (noun- $\mathcal{E}$ -verb)" constructions, where noun is the direct object of verb. Our approach consists of

three steps: 1) for each extracted tuple in the second learner's composition, we created a set of candidates by substituting words generated using word similarity algorithms; 2) then, we measured the strength of association in the writer's phrase and in each generated candidate phrase using association measures to compute collocation scores; 3) the highest ranking alternatives are suggested as corrections. In our evaluation, we checked if the correction given in the learner corpus matches one of the suggestions given by the system. Figure 1 illustrates the method used in this study.



Figure 1 Word Similarity and Association Measures combination method for suggesting col-locations.

We considered only the tuples that contain noun or verb error. A real application, however, should also deal with error detection. For each example of the construction on the writer's text, the system should create the confusion set with alternative phrases, measure the strength of association in the writer's phrase and in each generated alternative phrase and flag as error only if the association score of the writer's phrase is lower than one or more of the alternatives generated and suggest the higher-ranking alternatives as corrections.

## 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. In our work, we used the following three measures to analyze word simila-

<sup>&</sup>lt;sup>3</sup>www.lang-8.com

		Confusion Set									
Word	する	受ける	始める	作る	書く	言う	食べる	やる	持つ		
Meaning	do	accept	begin	make	write	say	eat	do	carry		
Word	ビル	ビール	生ビール	お金	札	金額	景色	料金	建築物		
Meaning	building	beer	beer draft beer		oney bill d		scenery	fee	building		
		of									
		money									

Table 1 Confusion Set example for the words suru ( $\forall 3$ ) and biru ( $\forall \mathcal{V}$ )

	書く	読む	つける		ご飯を	ラーメンを	カレーを
	write	read	put on		rice	ramen noodle soup	curry
日記を diary	15	11	8	 食べる eat	164	53	39

Table 2 Context of a particular noun represented as a co-occurrence vector

rity: 1) thesaurus-based word similarity, 2) distributional similarity and 3) confusion set derived from learner corpus. The first two measures generate the collocation candidates by finding words that are analogous to the writer's choice, a common approach used in the related work on collocation error correction (Liu et al., 2009; Östling and O. Knutsson, 2009; Wu et al., 2010) and the third measure generates the candidates based on the corrections given by native speakers in the learner corpus.

**Thesaurus-based word similarity**: The intuition of this measure is to check if the given words have similar glosses (definitions). Two words are considered similar if they are near each other in the thesaurus hierarchy (have a path within a pre-defined threshold length).

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 similarity models are often used since it gives higher recall. On the other hand, distributional similarity models tend to have lower precision (Jurafsky et al., 2009), because the candidate set is larger. The intuition of this measure is that two words are similar if they have similar word contexts. In our task, context will be defined by some grammatical dependency relation, specifically, 'object-verb' relation. Context is represented as co-occurrence vectors that are based on syntactic dependencies. We are interested in computing similarity of nouns and verbs and hence the context of a particular noun is a vector of verbs that are in an object relation with that noun. The context of a particular verb is a vector

 Table 3 Context of a particular noun represented as a co-occurrence vector

of nouns that are in an object relation with that verb. Table 2 and Table 3 show examples of part of co-occurrence vectors for the noun "日記 [dia-ry]" and the verb "食べる [eat]", respectively. The numbers indicate the co-occurrence frequency in the BCCWJ corpus (Maekawa, 2008). We computed the similarity between co-occurrence vectors using different metrics: Cosine Similarity, Dice coefficient (Curran, 2004), Kullback-Leibler divergence or KL divergence or relative entropy (Kullback and Leibler, 1951) and the Jenson-Shannon divergence (Lee, 1999).

**Confusion Set derived from learner corpus:** In order to build a module that can "guess" common 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 noun and verb corrections for each of the nouns and verbs found in the data. Table 1 shows some examples extracted. For instance, the confusion set of the verb suru "する [to do]" is composed of verbs such as ukeru "受ける [to accept]", which does not necessarily have similar meaning with suru. The confusion set means that in the corpus, suru was corrected to either one of these verbs, i.e., when the learner writes the verb suru, he/she might actually mean to write one of the verbs in the confusion set. For the noun *biru*"ビル [*building*]", the learner may have, for example, misspelled the word  $b\bar{i}ru$  " $\not{\vdash} \neg \nu$ [beer]", or may have got confused with the translation of the English words bill ("お金[money]", "礼 [bill]", "金額 [amount of money]", "料金 [fee]") or view ("景色 [scenery]") to Japanese.

#### 4.2 Word Association Strength

After generating the collocation candidates using word similarity, the next step is to identify the "true collocations" among them. 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 scores indicate strong association, and can be used to select the "true collocations". We adopted the Weighted Dice coefficient (Kitamura and Matsumoto, 1997) as our association measurement. We also tested using other association measures (results are omitted): Pointwise Mutual Information (Church and Hanks, 1990), log-likelihood ratio (Dunning, 1993) and Dice coefficient (Smadja et al., 1996), but Weighted Dice performed best.

## 5 Experiment setup

We divided our experiments into two parts: verb suggestion and noun suggestion. For verb suggestion, given the learners' "noun wo verb" construction, our focus is to suggest "noun wo verb" collocations with alternative verbs other than the learner's written verb. For noun suggestion, given the learners' "noun wo verb" construction, our focus is to suggest "noun wo verb" collocations with alternative nouns other than the learner's written noun.

## 5.1 Data Set

For computing word similarity and association scores for verb suggestion, the following resources were used:

1) Bunrui Goi Hyo Thesaurus (The National Institute for Japanese Language, 1964): a Japanese thesaurus, which has a vocabulary size of around 100,000 words, organized into 32,600 unique semantic classes. This thesaurus was used to compute word similarity, taking the words that are in the same subtree as the candidate word. By subtree, we mean the tree with distance 2 from the leaf node (learner's written word) doing the pre-order tree traversal.

2) Mainichi Shimbun Corpus (Mainichi Newspaper Co., 1991): one of the major newspapers in Japan that provides raw text of newspaper articles used as linguistic resource. One year data (1991) were used to extract the "*noun wo verb*" tuples to compute word similarity (using cosine similarity metric) and collocation scores. We extracted 224,185 tuples composed of 16,781 unique verbs and 37,300 unique

nouns.

3) Balanced Corpus of Contemporary Written Japanese, BCCWJ Corpus (Maekawa, 2008): composed of one hundred million words, portions of this corpus used in our experiments include magazine, newspaper, textbooks, and blog data<sup>4</sup>. Incorporating a variety of topics and styles in the training data helps minimize the domain gap problem between the learner's vocabulary and newspaper vocabulary found in the Mainichi Shimbun data. We extracted 194,036 "noun wo verb" tuples composed of 43,243 unique nouns and 18,212 unique verbs. These data are 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 contain 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 corrected sentences. We extracted 163,880 "*noun wo verb*" tuples composed of 38,999 unique nouns and 16,086 unique verbs. ii) Construct the confusion set (explained in Section 4.1): We constructed the confusion set for all the 16,086 verbs and 38,999 nouns that appeared in the data.

B) Year 2011 data were used to construct the test set (described in Section 5.2).

# 5.2 Test selection

We used Lang-8 (2011 data) for selecting our test set. For the verb suggestion task, we extracted all the "noun wo verb" tuples with incorrect verbs and their correction. From the tuples extracted, we selected the ones where the verbs were corrected to the same verb 5 or more times by the native speakers. Similarly, for the noun suggestion task, we extracted all the "noun wo verb" tuples with incorrect nouns and their correction. There are cases where the learner's construction sounds more acceptable than its correction, cases where in the corpus, they were corrected due to some contextual information. For our application, since we are only considering

<sup>&</sup>lt;sup>4</sup> Although the language used in blog data is usually more informal than the one used in newspaper, maganizes, etc., and might contain errors like spelling and grammar, collocation errors are much less frequent compared to spelling and grammar errors, since combining words appropriately is one the vital competencies of a native speaker of a language.

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. After applying those conditions, we obtained 185 tuples for the verb suggestion test set and 85 tuples for the noun suggestion test set.

#### 5.3 Evaluation Metrics

We compared the verbs in the confusion set ranked by collocation score suggested by the system with the human correction verb and noun in the Lang-8 data. A match would be counted as a true positive (tp). A false negative (fn) occurs when the system cannot offer any suggestion.

The metrics we used for the evaluation are: precision, recall and the mean reciprocal rank (MRR). We report precision at rank k, k=1, 5, computing the rank of the correction when a true positive occurs. The MRR was used to assess whether the suggestion list contains the correction and how far up it is in the list. It is calculated as follows:

$$MRR = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{rank(i)}$$
(1)

where *N* is the size of the test set. If the system did not return the correction for a test instance, we set  $\frac{1}{rank(i)}$  to zero. Recall rate is calculated

with the formula below:

$$\frac{tp}{tp+fn} \tag{2}$$

## 6 Results

Table 4 shows the ten models derived from combining different word similarity measures and the Weighted Dice measure as association measure, using different corpora. In this table, for instance, we named *M1* the model that uses thesaurus for computing word similarity and uses Mainichi Shimbun corpus when computing collocation scores using the association measure adopted, Weighted Dice. *M2* uses Mainichi Shimbun corpus for computing both word similarity and collocation scores. *M10* computes word similarity using the confusing set from Lang-8 corpus and uses BCCWJ and Lang-8 corpus when computing collocation scores.

Considering that the size of the candidate set generated by different word similarity measures vary considerably, we limit the size of the confusion set to 270 for verbs and 160 for nouns, which correspond to the maximum values of the confusion set size for nouns and verbs when using Lang-8 for generating the candidate set. Setting up a threshold was necessary since the size of the candidate set generated when using Distributional Similarity methods may be quite large, affecting the system performance. When computing Distributional Similarity, scores are also assigned to each candidate, thus, when we set up a threshold value n, we consider the list of n candidates with highest scores. Table 4 reports the precision of the k-best suggestions, the recall rate and the MRR for verb and noun suggestion.

## 6.1 Verb Suggestion

Table 4 shows that the model using thesaurus (M1) achieved the highest precision rate among the other models; however, it had the lowest recall. The model could suggest for cases where the wrong verb written by the learner and the correction suggested in Lang-8 data have similar meaning, as they are near to each other in the thesaurus hierarchy. However, for cases where the wrong verb written by the learner and the correction suggested in Lang-8 data do not have similar meaning, M1 could not suggest the correction.

In order to improve the recall rate, we generated models M2-M6, which use distributional similarity (cosine similarity) and also use corpora other than Mainichi Shimbun corpus to minimize the domain gap problem between the learner's vocabulary and the newspaper vocabulary found in the Mainichi Shimbun data. The recall rate improved significantly but the precision rate decreased. In order to compare it with other distributional similarity metrics (Dice, KL-Divergence and Jenson-Shannon Divergence) and with the method that uses Lang-8 for generating the confusion set, we chose the model with the highest recall value as baseline, which is the one that uses BCCWJ and Lang-8 (M6) and generated other models (M7-M10). The best MRR value obtained among all the Distributional Similarity methods was obtained by Jenson-Shannon divergence. The highest recall and MRR values are achieved when Lang-8 data were used to generate the confusion set (M10).

	Similarity used for Confusion Sets	Thesaurus	Cosine Similarity					Dice Coefficient	KL Divergence	Jenson- Shannon Divergence	Confusion Set from Lang-8
	K-Best	Mainichi Shimbun	Mainichi Shimbun	BCCWJ	Lang-8	Mainichi Shimbun + BCCWJ	BCCWJ + Lang-8	BCCWJ + Lang-8	BCCWJ + Lang-8	BCCWJ + Lang-8	BCCWJ + Lang-8
		M1	M2	<i>M3</i>	M4	M5	M6	M7	M8	M9	M10
Verb Suggestion	1	0.94 1	0.48	0.42	0.60	0.62	0.56	0.59	0.63	0.60	0.64
	Recall	0.20	0.91	0.30	0.50	0.90	0.00	0.80	0.35	0.00	0.97
	MRR	0.19	0.26	0.19	0.49	0.36	0.50	0.58	0.26	0.53	0.75
Noun Suggestion	1	0.16	0.20	0.42	0.58	0.50	0.55	0.30	0.63	0.57	0.73
	5	1	0.66	0.94	0.89	1	0.91	0.83	1	0.84	0.98
	Recall	0.07	0.17	0.22	0.45	0.04	0.42	0.35	0.12	0.38	0.98
	MRR	0.03	0.06	0.13	0.33	0.02	0.29	0.18	0.10	0.26	0.83

Table 4 The precision and recall rate and MRR of the Models of Word Similarity and Association Strength method combination.

#### 6.2 Noun Suggestion

Similar to the verb suggestion experiments, the best recall and MRR values are achieved when Lang-8 data were used to generate the confusion set (M10).

For noun suggestion, our automatically constructed test set includes a number of spelling correction cases, such as cases for the combination *eat ice cream*, where the learner wrote *aisukurimu wo taberu* "アイスクリムを食べ る" and the correction is *aisukurīmu wo taberu* " アイスクリームを食べる". Such phenomena did not occur with the test set for verb suggestion. For those cases, the fact that only spelling correction is necessary in order to have the right collocation may also indicate that the learner is more confident regarding the choice of the noun than the verb. This also justifies the even lower recall rate obtained (0.07) when using a thesaurus for generating the candidates

## 7 Conclusion and Future Work

We analyzed 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, instead of only using wellformed text. By employing this approach, we could obtain better recall and MRR values compared to thesaurus based method and distributional similarity methods.

Although only noun-wo-verb construction is examined, the model is designed to be applicable to other types of constructions, such as adjectivenoun and adverb-noun. Another straightforward extension is to pursue constructions with other particles, such as "noun ga verb (subject-verb)", "noun ni verb (dative-verb)", etc. In our experiments, only a small context information is considered (only the noun, the particle wo (E) and the verb written by the learner). In order to verify our approach and to improve our current results, considering a wider context size and other types of constructions will be the next steps of this research.

# Acknowledgments

Special thanks to Yangyang Xi for maintaining Lang-8.

#### References

- Y. C. Chang, J. S. Chang, H. J. Chen, and H. C. Liou. 2008. An automatic collocation writing assistant for Taiwanese EFL learners: A case of corpus-based NLP technology. *Computer Assisted Language Learning*, 21(3):283–299.
- K. Church, and P. Hanks. 1990. Word Association Norms, Mutual Information and Lexicography, *Computational Linguistics*, Vol. 16:1, pp. 22-29.
- J. R. Curran. 2004. From Distributional to Semantic Similarity. Ph.D. thesis, University of Edinburgh.
- D. Dahlmeier, H. T. Ng. 2011. Correcting Semantic Collocation Errors with L1-induced Paraphrases. In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pages 107–117, Edinburgh, Scotland, UK, July. Association for Computational Linguistics
- T. Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics* 19.1 (Mar. 1993), 61-74.
- 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. *Computer Assisted Language Learning*, 21, 4 (October 2008), 353-367.
- M. Gamon. 2010. Using mostly native data to correct errors in learners' writing: A meta-classifier approach. In Proceedings of Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL, pages 163– 171, Los Angeles, California, June. Association for Computational Linguistics.
- D. Jurafsky and J. H. Martin. 2009. Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics. 2nd edition. Prentice-Hall.
- K. Maekawa. 2008. Balanced corpus of contemporary written japanese. In *Proceedings of the 6th Workshop on Asian Language Resources (ALR)*, pages 101–102.
- M. Kitamura, Y. Matsumoto. 1997. Automatic extraction of translation patterns in parallel corpora. In *IPSJ*, Vol. 38(4), pp.108-117, April. In Japanese.
- S. Kullback, R.A. Leibler. 1951. On Information and Sufficiency. *Annals of Mathematical Statistics* **22** (1): 79–86.
- L. Lee. 1999. Measures of Distributional Similarity. In *Proc of the 37th annual meeting of the ACL*, Stroudsburg, PA, USA, 25.

- 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, pages 47–50, Boulder, Colorado, June. Association for Computational Linguistics.
- Mainichi Newspaper Co. 1991. Mainichi Shimbun CD-ROM 1991.
- 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. AFNLP.
- R. Östling and O. Knutsson. 2009. A corpus-based tool for helping writers with Swedish collocations. In Proceedings of the Workshop on Extracting and Using Constructions in NLP, Nodalida, Odense, Denmark. 70, 77.
- H. Oyama and Y. Matsumoto. 2010. Automatic Error Detection Method for Japanese Case Particles in Japanese Language Learners. *In Corpus, ICT, and Language Education*, pages 235–245.
- A. Rozovskaya and D. Roth. 2010. Generating confusion sets for context-sensitive error correction. In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pages 961–970, MIT, Massachusetts, USA, October. Association for Computational Linguistics.
- F. Smadja, K. R. Mckeown, V. Hatzivassiloglou. 1996. Translation collocations for bilingual lexicons: a statistical approach. *Computational Linguistics*, 22:1-38.
- H. Suzuki and K. Toutanova. 2006. Learning to Predict Case Markers in Japanese. In Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL, pages 1049–1056, Sydney, July. Association for Computational Linguistics.J. Tetreault, J. Foster, and M. Chodorow. 2010. Using parse features for preposition selection and error detection. In Proceedings of ACL 2010 Conference Short Papers, pages 353-358, Uppsala, Sweden, July. Association for Computational Linguistics.
- The National Institute for Japanese Language, editor. 1964. Bunrui-Goi-Hyo. Shuei shuppan. In Japanese.
- 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*, pages 115-119, Uppsala, Sweden, July. Association for Computational Linguistics.