Collocational Aid for Learners of Japanese as a Second Language

Lis Pereira and Yuji Matsumoto
Graduate School of Information Science
Nara Institute of Science and Technology
{lis-k, matsu}@is.naist.jp

Abstract

We present Collocation Assistant, a prototype of a collocational aid designed to promote the collocational competence of learners of Japanese as a second language (JSL). Focusing on noun-verb constructions, the tool automatically flags possible collocation errors and suggests better collocations by using corrections extracted from a large annotated Japanese language learner corpus. Each suggestion includes several usage examples to help learners choose the best candidate. In a preliminary user study with JSL learners, Collocation Assistant received positive feedback, and the results indicate that the system is helpful to assist learners in choosing correct word combinations in Japanese.

1 Introduction

Collocational competence is one of the factors which contribute to the differences between native speakers and second language learners (Shei and Pain, 2000). However, studies confirm that the correct use of collocations is challenging, even for advanced second language learners (Liu, 2002; Nesselhauf, 2003; Wible et al., 2003). Since there are no well-defined rules to determine collocation preferences, language learners are prone to produce word combinations that, although they may be grammatically and semantically well formed, may sound "unnatural" to a native speaker. Moreover, the number of tools designed to target language learners' collocation errors is limited or inexistent for many languages, which makes it difficult for learners to detect and correct these errors. Therefore, an application that can detect a learners' collocation errors and suggest the most appropriate "ready-made units" as corrections is an important goal for natural language processing (Leacock et al., 2014).

In this paper, we describe Collocation Assistant, a web-based and corpus-based collocational aid, aiming at helping JSL learners expand their collocational knowledge. Focusing on noun-verb constructions, Collocation Assistant flags possible collocation errors and suggests a ranked list of more conventional expressions. Each suggestion is supported with evidence from authentic texts, showing several usage examples of the expression in context to help learners choose the best candidate. Based on our previous study (Pereira et al., 2013), the system generates corrections to the learners' collocation error tendencies by using noun and verb corrections extracted from a large annotated Japanese learner corpus. For ranking the collocation correction candidates, it uses the Weighted Dice coefficient (Kitamura and Matsumoto, 1997). We add to our previous work by implementing an interface that allows end-users to identify and correct their own collocation errors. In addition, we conducted a preliminary evaluation with JSL learners to gather their feedback on using the tool.

2 The need for collocational aids

Existing linguistic tools are often of limited utility in assisting second language learners with collocations. Most spell checkers and grammar checkers can help correct errors made by native speakers, but offer no assistance for non-native errors. Futagi et al. (2008) note that common aids for second language learners namely, dictionaries and thesauri are often of limited value when the learner does not know the appropriate collocation and must sort through a list of synonyms to find one that is contextually appropriate. Yi et al. (2008) observe that language learners often use search engines to check if a phrase is commonly used by observing the number of results returned. However, search engines are not designed to offer alternative phrases that are more commonly used than the

Proceedings of The 2nd Workshop on Natural Language Processing Techniques for Educational Applications, pages 20–25,
Beijing, China, July 31, 2015.

learner's phrase (Park et al., 2008). Concordancers seem to be an alternative to search engines, but they retrieve too much information because they usually allow only single-word queries. Too much information might distract and confuse the user (Chen et al., 2014). Thus, a computer program that automatically identifies potential collocation errors and suggests corrections would be a more appropriate resource for second language learners.

A few researchers have proposed useful English corpus-based tools for correcting collocation errors (Futagi et al., 2008; Liu et al., 2009; Park et al., 2008; Chang et al., 2008; Wible et al., 2003; Dahlmeier and Ng, 2011). In a user study, Park et al. (2008) observed positive reactions from users when using their system. In another study, Liou et al. (2006) showed that the miscollocation aid proposed by Chang et al. (2008) can help learners improve their knowledge in collocations. One limitation is that these proposed tools rely on resources of limited coverage, such as dictionaries, thesauri, or manually constructed databases to generate the candidates. Another drawback is that most of these systems rely solely on well-formed English resources (except Wible et al., 2003) and do not actually take into account the learners' tendencies toward collocation errors.

3 Collocation Assistant

In the proposed system, we focused on providing Japanese collocation suggestions for potential collocation errors in Japanese noun-verb constructions. Given the noun-verb collocation input by a learner, the system first checks if it exists in the reference corpora. If not, the input is validated as a potential collocation error and a message is displayed to the user. Next, the system suggests more appropriate noun-verb collocations. For instance, if the learner types *夢をする (*yume wo suru* , lit. 'to do a dream'), the system flags a collocation error. When the user clicks on "same noun", the system displays better collocations with the same noun input by the user, such as 夢を見る (*yume wo miru*, 'to dream') and 夢を持つ (*yume wo motsu*, 'to hold a dream'), as shown in Figure 1. Likewise, when the user clicks on "same verb", the system displays better collocations with the same verb input by the user. If the user clicks on "View all suggestions", all possible better collocations with the same noun or the same verb input by the user are displayed. Aside from the collocations, sentence examples for each phrase suggestion are displayed, showing the phrase in context with surrounding text. Showing phrases in context can be crucial in helping users determine which phrase is most appropriate (Park et al., 2008). Even if the learner's input is not flagged as an error, it will undergo the same correction process, since better collocations than the input might exist. In this case, the learner will check the ranked suggestions and sentence examples and choose the most appropriate expression. The current system does not detect which component (noun or verb) is wrong in a noun-verb construction. Therefore, the learner must specify which component would be corrected by the system. This has been the common evaluation setup in collocation error correction, which assumes that the error is already detected and more focus is given on correcting the identified errors. (Dahlmeier and Ng, 2011). In case the learner types only a noun or only a verb, the system will suggest collocations containing words that strongly collocate with this input.

3.1 Approach to collocation suggestion

Based on our previous study (Pereira et al., 2013), our Collocation Assistant generates collocation corrections by using noun and verb corrections extracted from a large annotated Japanese language learner corpus, Lang-8[1]. This approach to generating corrections achieved superior performance compared to methods that generate candidates based solely on the semantic relation of words. Using noun and verb corrections extracted from a learner corpus, our system can explore the learners' tendency to commit collocation errors. For ranking the collocation correction candidates, the tool uses the Weighted Dice coefficient (Kitamura and Matsumoto, 1997). This association measure achieved the best performance in our task among other association measures evaluated (i.e. pointwise mutual information, log-likelihood and dice coefficient).

3.2 Resources used for providing sentence examples

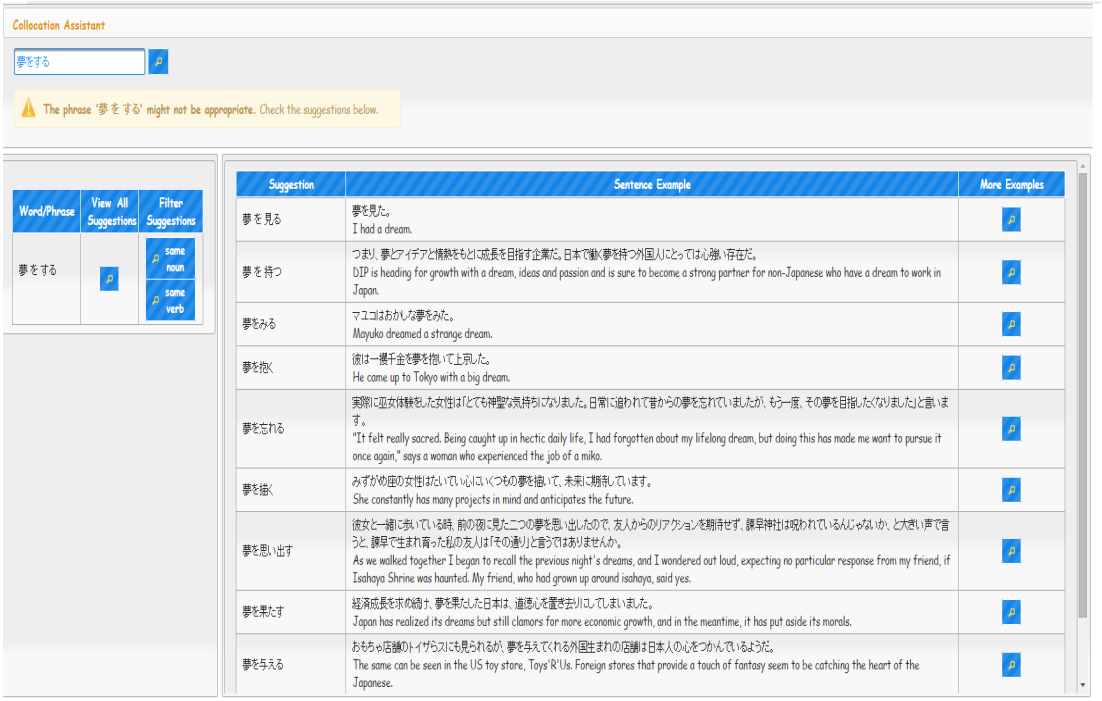
We used several monolingual and bilingual resources for providing useful sentence examples to users. These resources are:

Bilingual resources. 1) Tatoeba Corpus[2], a

[1] http://cl.naist.jp/nldata/lang-8/

[2] https://tatoeba.org/eng/

(a)



(b)

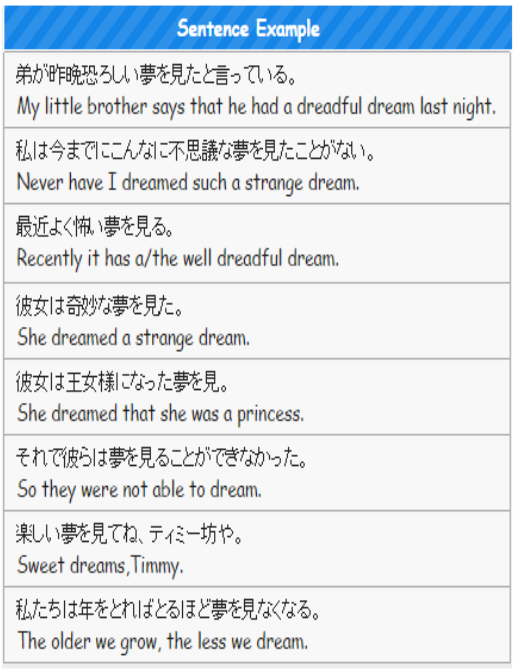


Figure 1: An example of collocation suggestions produced by the system given the erroneous collocation *夢をする (*yume wo suru*, lit. 'to do a dream') as input. (a) Collocation suggestions are shown on the left and an example sentence for each suggestion is shown on the right. In the example, 夢を見る (*yume wo miru*, 'to dream') is the correct collocation. (b) Further examples for each suggestion are shown when the user clicks on "More examples". In the example, further examples for the collocation 夢を見る (*yume wo miru*, 'to dream') are displayed.

free collaborative online database of example sentences geared towards foreign language learners. We used the Japanese-English sentences available in the website. 2) Hiragana Times (HT) Corpus[3], a Japanese-English bilingual corpus of magazine articles of Hiragana Times, a bilingual magazine written in Japanese and English to introduce Japan to non-Japanese, covering a wide range of topics (culture, society, history, politics, etc.). 3) Kyoto Wikipedia (KW) Corpus[4], a corpus created by manually translating Japanese Wikipedia articles (related to Kyoto) into English.

Monolingual resource: the Balanced Corpus of Contemporary Written Japanese (Maekawa, 2008) was used for the noun-verb expressions where no bilingual examples were available.

	# *jp* sentences	# *en* sentences
Tatoeba	203,191	203,191
HT	117,492	117,492
KW	329,169	329,169
BCCWJ	871,184	-

Table 1: Data used as sentence examples.

4 Preliminary User Study of the Collocation Assistant

We conducted a preliminary evaluation with JSL learners to gather their feedback on using the Collocation Assistant. The results gave us insights about the usefulness of the tool, and about the possible interesting evaluations that should be carried out in the future.

4.1 Participants

In this study, 10 JSL learners, all graduate students from the same institution as the authors were invited to participate. Participants' ages ranged from 24 to 33 years, and the average age was 27.5. Among the respondents, 2 were female and 8 were male, and they had different language backgrounds (Chinese, Indonesian, Tagalog, Swahili, Spanish, and Basque). Regarding their proficiency level, three were beginners, three were intermediate, and four were advanced learners, based on the Japanese-Language proficiency test certificate level they previously obtained. All participants were regular computer users.

[3]http://www.hiraganatimes.com/

[4]https://alaginrc.nict.go.jp/WikiCorpus/

4.2 Procedure

A collocation test was designed to examine whether or not the tool could help JSL learners find proper Japanese collocations. This included 12 Japanese sentences from the Lang-8 learner corpus and from another small annotated Japanese learner corpus, NAIST Goyo Corpus (Oyama, Komachi and Matsumoto, 2013). The sentences and their corrections were further validated by a professional Japanese teacher. Each sentence contained one noun-verb collocation error made by JSL learners. The participants were asked to use the Collocation Assistant to identify and correct the errors. Next, they were asked to write a small paragraph in Japanese and to use the tool if they needed. After performing the task, a survey questionnaire was also administered to better understand the learners' impressions of the tool. The questionnaire contained 43 questions answerable by a 7-point Likert-scale (with 7 labeled "strongly agree" and 1 labeled "strongly disagree"). The second part of the questionnaire contained 7 open-ended questions. Our survey questionnaire inquired on the difficulty of Japanese collocations, the usefulness of Collocation Assistant, the design of Collocation Assistant and the quality of the retrieved data.

4.3 Results on the Collocation Test and Survey Questionnaire

The participants successfully found corrections for an average of 8.9 (SD=1.6) out of 12 cases. The average time participants took to complete the task was 29 (SD=16) minutes. The average score of beginner and intermediate learners was 9.6 (SD=0.5). They scored higher than advanced learners, who obtained an average score of 8.2 (SD=2.0). Analyzing the log files of their interactions with the system, we observed that intermediate and beginner learners used the system 40% more times (on average) than the advanced learners. We noticed that two advanced learners tried to answer the questions without using the system when they felt confident about the answer, whereas the beginners and intermediate learners used the system for all sentences and obtained higher scores. The participants had difficulty in correcting two particular long sentences in the test. The noun-verb collocations in the sentences alone were not incorrect, but they were not appropriate in the context they appeared. The

participants had difficulty in finding sentence examples close to the meaning of the sentences in the test. Although we need to evaluate this tool with a larger number of users, we observed that the system was effective in helping the learners choose the proper collocations. In the questionnaire administered, all participants acknowledged their difficulty in using Japanese collocations appropriately and stated that the software aids they used did not provide enough information about the meaning of Japanese phrases nor help in correcting errors in Japanese expressions. Their attitude toward the usefulness of Collocation Assistant was mostly positive and they thought the tool was useful to help choose the proper way to use Japanese expressions. Most participants considered the interface easy to use (M=6.3, SD=0.8). Regarding the quality of the retrieved data, the participants expressed satisfaction with the retrieved collocations, with an average score of 6.5 (SD=0.7). They also expressed satisfaction with the ranking of the collocations presented, with an average score of 5.8 (SD=0.6). Additionally, they reported that the sentence examples further helped them understand in which context an expression should be used. However, some participants expressed dissatisfaction with the complexity of some example sentences: some of the sentences were too long and difficult to understand. In the second part of the questionnaire, some participants stated that the Collocation Assistant could be helpful when learning new words and when one does not know which word combinations to use. They also suggested that the tool could be useful for teachers too when giving feedback to their students about the common errors they make and when providing alternative ways of expressing the same idea. Lastly, they suggested several improvements regarding the sentence examples and the interface: show shorter and simpler example sentences, highlithgt the user's input in the sentence examples and allow English search.

5 Conclusions

In this paper, we presented a collocational aid system for JSL learners. The tool flags possible collocation errors and suggests corrections by using corrections extracted from a large annotated Japanese language learner corpus. Our Collocation Assistant received positive feedback from JSL learners in a preliminary user study. The system can be used independently as a phrase dictionary, or it can be integrated into the writing component of some bigger CALL systems. For example, Collocation Assistant can be used by teachers as a way to obtain better understanding about learners' errors and help them provide better feedback to the students. One limitation of our experiments is the limited contextual information (only the noun, particle, and verb written by the learner). In the future, to verify our approach and to improve on our current results, we plan to consider a wider context size and other types of constructions (e.g., adjective-noun, adverb-verb, etc.). We also intend to investigate how to adjust the difficulty level of the sentences according to the user's proficienSolution cy level. Finally, we plan to conduct a more extensive evaluation with JSL learners to verify the usefulness of the tool in practical learning scenarios.

Acknowledgments

We would like to thank anonymous reviewers for their insightul comments and suggestions.

References

Chang, Y. C., Chang, J. S., Chen, H. J. and Liou, H. C. 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. doi: 10.1080/09588220802090337.

Chen, M.-H., Huang, C.-C., Huang, S.-T., Chang, J.S. and Liou, H.-C. 2014. An Automatic Reference Aid for ImprovingEFL Learners' Formulaic Expressionsin Productive Language Use. *IEEE Transactions on Learning Technologies*, 01/2014; 7(1):57-68. doi:10.1109/TLT.2013.34.

Dahlmeier, D. and Ng, H. T. 2011. Correcting semantic collocation errors with L1-induced paraphrases. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*- (pp. 107-117). Edinburgh, Scotland, UK, July 27–31, 2011.

Futagi, Y., Deane, P., Chodorow, M. and Tetreault, J. 2008. A computational approach to detecting collocation errors in the writing of non-native speakers of English. *Computer Assisted Language Learning* 21(4), 353–367. doi: 10.1080/09588220802343561

Kitamura, M. and Matsumoto, Y. 1997. Automatic extraction of translation patterns in parallel corpora. Automatic extraction of translation patterns in parallel corpora. *Information Processing Society of Japan Journal*, 38 (4), 727–735.

Leacock, C., Chodorow, M., Gamon, M., and Tetreault, J. 2014. *Automated grammatical error detection for language learners*. Synthesis lectures on human language technologies, 7(1), 1-170.

Liou, H., Chang, J., Chen, H., Lin, C., Liaw, M., Gao, Z., Jang, J., Yeh, Y, Chuang, T. and You, G. (2006) 2006. Corpora processing and computational scaffolding for a Web-based English learning environment: The CANDLE Project. *CALICO Journal*, 24 (1), 77-95.

Liu, A. L. 2002. *A Corpus-based Lexical Semantic Investigation of VN Miscollocations in Taiwan Learners' English*. Master Thesis, Tamkang University, Taiwan.

Maekawa, K. 2008. Balanced Corpus of Contemporary Written Japanese. In *Proceedings of The 6th Workshop on Asian Language Resources*- (pp. 101–102). Association for Computational Linguistics, Stroudsburg, PA, USA.

Nesselhauf, N. 2003. The use of collocations by advanced learners of English and some implications for teaching. *Applied Linguistics*, 24, 223–242.

Oyama,H., Komachi, M. and Matsumoto, Y. 2013. Towards Automatic Error Type Classification of Japanese Language Learners' Writings. In *Proceedings of the 27th Pacific Asia Conference on Language, Information, and Computation*, pp.163-172, Taipei, Taiwan.

Park, T., Lank, E., Poupart, P. and Terry, M. 2008. "Is the Sky Pure Today?" AwkChecker: An Assistive Tool for Detecting and Correcting Collocation Errors. In *Proceedings of the 21th Annual Assotiation for Computing Machinery Symposium on User Interface Software and Technology*- pages 121-130. Monterey, CA, USA.

Pereira, L., Manguilimotan, E. and Matsumoto, Y. 2013. Automated Collocation Suggestion for Japanese Second Language Learners. In *Proceedings of the Student Research Workshop 51st Annual Meeting of the ACL*, pages 52-58, Sofia, Bulgaria.

Shei, C.-C. and Pain, H. 2000. An esl writer's collocational aid. *Computer Assisted Language Learning*, 13(2):167–182.

Wible, D., Kuo, C., Tsao, N., Liu, A. and Lin, H. 2003. Bootstrapping in a Language Learning Environment, *Journal of Computer-Assisted Learning*, 19(1), pp. 90-102. SSCI, LLBA.

Yi, X., Gao, J. and Dolan, W. A web-based English proofing system for English as a Second Language users. 2008. In *Proceedings of the 3rd International Joint Conference on Natural Language Processing*-pages 619–624. Association for Computational Linguistics, Stroudsburg, PA, USA.