Sakura: Large-scale Incorrect Example Retrieval System for Learners of Japanese as a Second Language

Mio Arai **Tomonori Kodaira** **Mamoru Komachi**
Tokyo Metropolitan University
Hino City, Tokyo, Japan
arai-mio@ed.tmu.ac.jp, kodaira.tomonori@gmail.com,
komachi@tmu.ac.jp

Abstract

This study develops an incorrect example retrieval system, called Sakura, that uses a large-scale Lang-8 dataset for Japanese language learners. Existing example retrieval systems either exclude grammatically incorrect examples or present only a few examples. If a retrieval system has a wide coverage of incorrect examples along with their correct counterparts, learners can revise their composition themselves. Considering the usability of retrieving incorrect examples, our proposed system uses a large-scale corpus to expand the coverage of incorrect examples and present correct as well as incorrect expressions. Intrinsic and extrinsic evaluations indicate that our system is more useful than existing systems.

1 Introduction

A standard method that supports second language learning effort is the use of examples. Example retrieval systems such as Rakhilina et al. (2016) and Kilgarriff et al. (2004) particularly check for the appropriate use of words in the context in which they are written. However, in such a system, if the query word is incorrect, finding appropriate examples is impossible using ordinary search engines such as Google. Even if learners have access to an incorrect example retrieval system, such as Kamata and Yamauchi (1999) and Nishina et al. (2014), they are often unable to rewrite a composition without correct versions of the incorrect examples. Furthermore, existing example retrieval systems only provide a small number of examples; hence, learners cannot acquire sufficient information when they search. These systems are primarily developed for use by Japanese teachers; therefore, they are not as helpful for learners without a strong background in Japanese.

Another difficulty in learning Japanese as a second language is to learn the use of particles. Particles in Japanese indicate grammatical relations between verbs and nouns. For example, the sentence, "日本語を勉強する。", which means "I study Japanese." includes an accusative case marker "を", which introduces the direct object of the verb. However, in this case, Japanese learners often make mistakes, such as "日本語が勉強する。", which means "Japanese language studies." Thus, the appropriate use of particles is not obvious for non-native speakers of Japanese. Particle errors and word choice are the most common Japanese grammatical errors (Oyama et al., 2013), both of which require a sufficient number of correct and incorrect examples to understand the usage in context. Word n-gram search provides only few or no examples for a phrase because Japanese is a relatively free language in terms of word order, in which a syntactically dependent word may appear in a distant position.

Considering this, our study develops an incorrect example retrieval system, called Sakura[1], that uses the large-scale Lang-8[2] dataset for learners of Japanese as a second language (JSL) by focusing on the usability of incorrect example retrieval. The main contributions of this work are as follows:

- We use a large corpus; hence, the new system has far more examples than previous systems.
- Our system shows the incorrect sentences and the corresponding sentence as corrected by a native speaker. Thus, learners can rectify their mistakes during composition.

Figure 1 illustrates an example of the search result obtained using our system Sakura. Suppose a learner wants to view examples for the usage of "読みたり (*yomitari*, meaning "to read")",

[1] http://cl.sd.tmu.ac.jp/sakura/en

[2] Multi-lingual language learning and language exchange social networking service. http://lang-8.com/

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pages 1–6
Florence, Italy, July 28 - August 2, 2019.

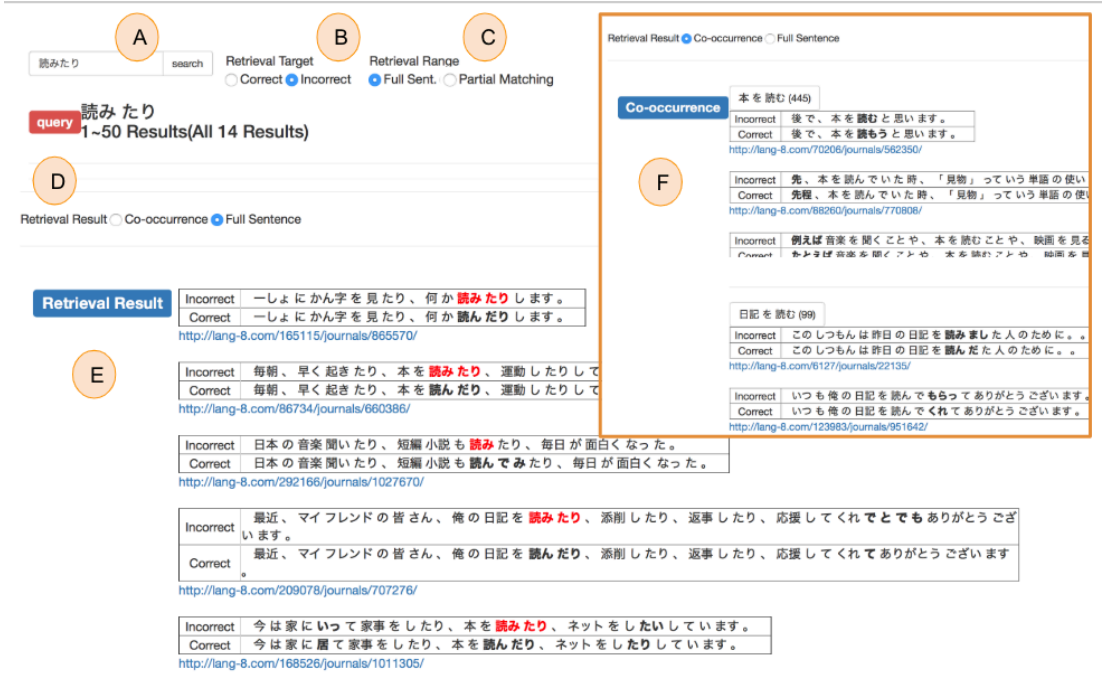

Figure 1: User interface of Sakura.

which is an incorrect or archaic expression. As shown in E of Figure 1, Sakura displays the incorrect examples with "読みたり" written in red and presents the correct examples using "読んだり (*yondari*, which is the correct euphonic form of "to read")". The learner can then identify that "読みたり" is the incorrect expression, and "読んだり" is the correct phrase. F in Figure 1 shows the collocation for "読む (*yomu*, which is the basic form of "読んだり")". The learner can grasp the common ways of using "読む", such as "本を読む (*hon wo yomu*, which means "I read a book.")".

2 Sakura: Large-scale Incorrect Example Retrieval System for JSL

This section describes the dataset and user interface of our proposed system, Sakura. Our system uses the data explained in Section 2.1 as the database for example retrieval. The user interface illustrated in Section 2.2 allows learners to search for incorrect examples.

2.1 Lang-8 Dataset

In this study, we used the Lang-8 Learner Corpora created by Mizumoto et al. (2011). The developers of the dataset used it for Japanese grammatical error correction, whereas we used it as an example retrieval database for JSL.

Each learner's sentence has at least one revised sentence. A learner's sentence is combined with a revised sentence to make a sentence pair. If a learner's sentence has more than one revised sentence, each of the revised sentences is paired with the learner's sentence as a separate sentence pair. Sentences with a length exceeding 100 words or with a Levenshtein distance of more than 7 are eliminated to remove the noise in the corpus.

We extracted 1.4 million pairs of the learner's sentences written by Japanese language learners and the revised sentences corrected by Japanese native speakers. The total number of included Japanese essays was 185,991.

The learner's sentences and the revised sentences are tokenized and POS-tagged by the morphological analyzer, MeCab (ver. 0.996)[3] with UniDic (ver. 2.2.0). Furthermore, we used the dependency parser CaboCha (ver. 0.69)[4] for the revised sentences to extract triples of a noun phrase, particle, and verb to be used as a co-occurrence.

2.2 User Interface

Figure 1 shows the user interface of Sakura. Its components are explained below.

A. Query Input the word to be searched for. The input query is assumed as a word or a phrase (se-

[3]https://github.com/taku910/mecab
[4]https://github.com/taku910/cabocha

quence of words).

B. Retrieval target Choose the target of retrieval as correct or incorrect usage. The default option is correct usage.

C. Retrieval range Choose the retrieval range from full sentence or partial matching with revised string. The system searches the entire sentence when a full sentence is selected. When partial matching with the revised string is selected, it searches the sentences where the query overlaps the revised string. The default option is full sentence. Learners can verify if their expressions are correct by selecting this option.

D. Retrieval result Choose the priority of displaying the retrieval result from the co-occurrence and the full sentence. The default option is co-occurrence.

E. Full Sentence The retrieval results that match the query are displayed when the user selects Full Sentence. The incorrect sentences written by learners are shown in the upper part, paired with the correct examples revised by native speakers. The query is represented in red, and the revised part is represented in bold.

F. Co-occurrence When the user searches for the query, including a noun or verb, Sakura displays up to 10 triplets of the noun, particle, and verb that co-occur with the query under the Co-occurrence tab. These triplets are shown with some example sentences, and the user can view up to 10 examples by scrolling. If the user searches for a POS other than a noun or a verb, Sakura shows the message "Not found" under the Co-occurrence tab, and retrieval results can be found under the Full Sentence tab.

3 Experiment

We compared Sakura with the search system for the "Tagged KY corpus" (hereafter, KYC) [5] in our evaluation experiment to confirm the effectiveness of presenting pairs of correct and incorrect examples. We evaluated our system in two parts, intrinsic and extrinsic evaluation.

3.1 Intrinsic Evaluation

We compared the accuracies of two systems, Sakura and KYC. We searched for the phrases in each system (KYC and Sakura) and counted the number of matches for each system that led to correct expressions to ensure accuracy.

We randomly extracted 55 incorrect phrases from the learner's erroneous sentences with correct phrases from the Lang-8 dataset, which were not included in the corpus we used for our system. We classified the incorrect examples into seven types: alternating form (A), lexical choice (L), omission (O), misformation (M), redundant (R), pronunciation (P), and others (X). Table 1 lists examples of the test phrases.

Table 2 shows the frequency and accuracy of each type of error. Although KYC searches for incorrect and correct examples, it cannot find correct answers because it has very few examples. Even if it finds some examples that match the query, it cannot find the correct examples because it does not contain revised sentences corresponding to an incorrect sentence.

The accuracy was high for superficial errors, such as omission and redundant errors, because learners tend to make similar errors. For example, an incorrect word "ニュージランド" requires "ー" to make the correct phrase "ニュージーランド (New Zealand)." In contrast, the incorrect word "みんなさん" has an additional character "ん" when compared with the correct phrase "みなさん (everybody)." Such error patterns are common among learners of Japanese; therefore, our system can provide correct answers to JSL learners.

However, it is difficult for our system to find the correct answer for Types A (alternating form) and L (lexical choice) because they have too many error forms, which makes identifying the appropriate answer challenging. For instance, an incorrect phrase "本がもらえる (I can get a book)" is corrected to "本しかもらえない (I can only get a book)" in the test corpus, but "本がもらえる" can be paraphrased in many ways, such as "本をもらう (I get a book)." Thus, it is difficult for learners to determine the most appropriate phrase.

3.2 Extrinsic Evaluation

We recruited 10 Japanese non-native speakers majoring in computer science in a graduate school in Japan to solve 10 Japanese composition exercises. Participation was voluntary and unpaid. These prompts are shown in Table 3. We assigned

[5] http://jhlee.sakura.ne.jp/kyc/corpus/

incorrect phrase	pronunciation	correct phrase	pronunciation	Sakura	Error Type
おねさん	onesan	おねえさん (sister)	oneesan	×	O
ニュージランド	nyu-jirando	ニュージーランド (New Zealand)	nyu-ji-rando	✓	O
みんなさん	min'nasan	みなさん (everybody)	minasan	✓	R
大体に	daitaini	大体 (roughly)	daitai	×	R
疑問をして	gimonwoshite	疑問に思って (in doubt)	gimon'niomotte	×	M
驚い	odoroi	驚き (surprise)	odoroki	✓	M
がもらえる	gamoraeru	しかもらえない (only get this)	shikamoraenai	×	A
稼ぐ	kasegu	稼いだ (earned)	kaseida	✓	A
ちさい	chisai	少ない (few)	sukunai	✓	L
助けられる	tasukerareru	できる (can)	dekiru	×	L
しましだ	shimashida	いました (there was)	imashita	×	P
死んちゃう	shinchau	死んじゃう (will die)	shinjau	✓	P
ハウス	hausu	家 (house)	ie	✓	X

Table 1: Examples of test results. The column "Incorrect phrase" contains the phrases written by the learner. These are extracted from the Lang-8 test set. The column "Sakura" shows whether or not Sakura could identify the correct answer for that phrase.

system	type	frequency	accuracy
Sakura	ALL	55	0.44
	Alternating Form	19	0.37
	Lexical Choice	16	0.38
	Omission	8	0.75
	Misformation	6	0.40
	Redundant	3	0.67
	Pronunciation	2	0.50
	Others	1	1.00

Table 2: Frequency and accuracy of each type.

No.	Prompt
1	The event of your country.
2	The most impressive adventure in your life.
3	Your favorite feature about Japanese.
4	The most favorite movie or book.
5	The food of your country.
6	The pros and cons of English as a universal language.
7	Japanese supermarket.
8	Major incident in your country's history.
9	The place you'd like to visit.
10	Your favorite season and the reason.

Table 3: Prompts for extrinsic evaluation.

five learners to solve the odd-numbered exercises using KYC and the even-numbered exercises using Sakura. The other five learners solved the even-numbered exercises using KYC and the odd-numbered exercises using Sakura. The number of sentences in each exercise was three to ensure a fair comparison.

The results were evaluated using the following method. The composition exercise was scored by deducting points from an initial 30 points. One point was deducted per error. The total score of each system was summed up over five exercises.

Learner	KYC	Sakura
A	22	**25**
B	25	**28**
C	26	**27**
D	21	**24**
E	27	27
F	21	**26**
G	20	20
H	**26**	24
I	21	**27**
J	7	**22**
ave.	21.6	**25.0**

Table 4: Extrinsic evaluation. The points assigned to the Japanese compositions of the participants. A higher point indicates a better score.

Table 4 shows the score for each composition. The average writing score of the learners using Sakura was **25.0** and that with KYC was 21.6. In addition, 7 out of 10 learners received a higher score when using Sakura than when using KYC. These results indicate that Sakura is useful as a learner support system for writing a Japanese composition.

KYC had no revised sentences corresponding to the incorrect sentences; hence, the composition using KYC tended to include mistakes related to verb conjugation and lexical choice errors. In contrast, Sakura did not display the POS; thus, the composition using Sakura tended to contain particle errors.

4 Related Works

Web-based search engines are the most common search systems that can be used to search for example sentences. However, these search engines

Name	Correct Sent.	Incorrect Sent.	Revised Sent.	Co-occurrence
Learner's Error Corpora of Japanese Searching Platform	✓	✓	✓	×
Tagged KY corpus	✓	✓	×	×
Natsume	×	×	×	✓
Sakura	✓	✓	✓	✓

Table 5: Features of example retrieval systems for Japanese language learners. "Correct Sent." indicates whether the system can display the correct sentences or not; "Incorrect Sent." indicates whether the system can display the incorrect sentences or not; "Revised Sent." indicates whether the system can display the revised sentence corresponding to the incorrect sentence; and "Co-occurrence" denotes whether the system can provide co-occurrence examples.

are not intended to retrieve examples for language learners; therefore, the search engines show neither example sentences nor the correct version of a given incorrect sentence to aid learners.

Language learners can use several example retrieval systems. The following subsections present information on some of those systems for learners of English and Japanese as a second language.

4.1 Example Retrieval System for English as a Second Language

FLOW (Chen et al., 2012) is a system that shows some candidates for English words when learners of English as a Second Language (ESL) write a sentence in their native language by using paraphrase candidates with bilingual pivoting. In contrast, our system suggests incorrect examples and their counterparts based on corrections from the learner corpus.

Another system, called StringNet (Wible and Tsao, 2010), displays the patterns in which a query is used, together with their frequency. The noun and the preposition are substituted by their parts of speech, instead of the words themselves, to eliminate data sparseness. Our system shows collocation patterns for each query by using parts of speech information and dependency parsing; however, our system does not explicitly present the parts of speech because our dataset is sufficiently large and need not replace and display the part-of-speech tag.

The ESCORT (Matsubara et al., 2008) system shows example sentences to learners based on the grammatical relations of queries. The syntactic structure of the English sentences is stored in the database of a raw corpus. ESCORT analyzes the dependency relations of the input queries and only displays appropriate examples that match the relations. Our system does not parse the query; instead, it parses the learner corpus to present collocations and overcome data sparseness.

Furthermore, ESL learners can check examples while writing an English sentence by using WriteAhead (Yen et al., 2015). This system provides pattern suggestions based on collocation and syntax. For example, when the user writes "We discussed," the system displays the patterns for the use of the word "discussed." In our system, we also employ collocation patterns; however, we use a large-scale learner corpus to search for dependency structures.

Sketch Engine (Kilgarriff et al., 2004) displays grammar constructs associated with words along with thesaurus information. As previously mentioned, our system presents incorrect examples by using a learner corpus apart from the correct examples extracted from a raw corpus.

4.2 Example Retrieval System for Japanese as a Second Language

Recently, various Japanese example retrieval systems have been proposed. However, in practice, learners find them difficult to use. Herein, we explain why these systems are ineffective when used by JSL learners.

Table 5 lists the features of each system. Our proposed system, Sakura, employs a large-scale Japanese JSL corpus for correct and incorrect example sentences along with revisions for the incorrect example.

First, the "Learner's Error Corpora of Japanese Searching Platform"[6] was constructed by the Corpus-based Linguistics and Language Education at Tokyo University of Foreign Studies. This system displays sentences that includes incorrect sentences in the keyword in context (KWIC) format based on the learner's information, such as native language, age, and gender. Japanese language teachers can identify the features of the learner's mistakes by using this system. However, this sys-

[6] http://ngc2068.tufs.ac.jp/corpus_ja/

tem is primarily intended for educators rather than learners. As such, learners might find it confusing to use. In addition, this system has few examples; hence, learners cannot acquire sufficient information when they search.

Second, the "KY corpus" is a transcribed speech corpus for JSL learners. "Tagged KY corpus" (Kamata and Yamauchi, 1999) supersedes the "KY corpus" with a search engine using POS. It displays correct and incorrect examples for text written by learners. However, it oftentimes fails to provide results even for high-frequency words, because the number of incorrect examples is small; therefore, it is difficult for language learners to use the limited set of examples as a reference.

Third, a Japanese co-occurrence retrieval system, called "Natsume" (Nishina et al., 2014) [7], presents the words and particles that tend to co-occur with the searched word (e.g., verb and adjective for noun and noun for verb and adjective). However, this system only shows words, and it does not indicate concrete examples; therefore, using this system to write an actual composition is difficult. In addition, it does not include incorrect examples.

5 Conclusion

This study constructed an incorrect example retrieval system using the Lang-8 Learner Corpora. Our proposed system, Sakura, displays incorrect examples along with the revised sentences and example sentences. The results of our experiment indicated that Sakura was useful for JSL learners when writing Japanese compositions. Each example includes incorrect sentences; hence, language teachers can identify the difficulty faced by learners and use this information for language education.

Although this system was constructed for JSL learners, it can easily be customized for other languages. We plan to extend our system to support ESL learners (Tajiri et al., 2012).

Acknowledgements

We would like to thank the Lang-8 web organizer for providing the text data for our system.

[7] https://hinoki-project.org/natsume/

References

Mei-Hua Chen, Shih-Ting Huang, Hung-Ting Hsieh, Ting-Hui Kao, and Jason S. Chang. 2012. FLOW: A first-language-oriented writing assistant system. In *Proceedings of the ACL 2012 System Demonstrations*, pages 157–162.

Osamu Kamata and Hiroyuki Yamauchi. 1999. KY corpus version 1.1. Report, Vocaburary Acquisition Study Group.

Adam Kilgarriff, Pavel Rychly, Pavel Smrž, and David Tugwell. 2004. The sketch engine. In *Proceedings of EURALEX*, pages 105–116.

Shigeki Matsubara, Yoshihide Kato, and Seiji Egawa. 2008. ESCORT: example sentence retrieval system as support tool for English writing. In *Journal of Information Processing and Management*, pages 251–259.

Tomoya Mizumoto, Mamoru Komachi, Masaaki Nagata, and Yuji Matsumoto. 2011. Mining revision log of language learning SNS for automated Japanese error correction of second language learners. In *Proceedings of IJCNLP*, pages 147–155.

Kikuko Nishina, Bor Hodošček, Yutaka Yagi, and Takeshi Abekawa. 2014. Construction of a learner corpus for Japanese language learners: Natane and Nutmeg. *Acta Linguistica Asiatica*, 4(2):37–51.

Hiromi Oyama, Mamoru Komachi, and Yuji Mastumoto. 2013. Towards automatic error type classification of Japanese language learners' writing. In *Proceedings of PACLIC*, pages 163–172.

Ekaterina Rakhilina, Anastasia Vyrenkova, and Elmira Mustakimova. 2016. Building a learner corpus for Russian. In *Proceedings of the Joint Workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition*, pages 66–75.

Toshikazu Tajiri, Mamoru Komachi, and Yuji Matsumoto. 2012. Tense and aspect error correction for ESL learners using global context. In *Proceedings of ACL*, pages 198–202.

David Wible and Nai-Lung Tsao. 2010. StringNet as a computational resource for discovering and investigating linguistic constructions. In *Proceedings of the NAACL HLT Workshop on Extracting and Using Constructions in Computational Linguistics*, pages 25–31.

Tzu-Hsi Yen, Jian-Cheng Wu, Jim Chang, Joanne Boisson, and Jason Chang. 2015. WriteAhead: Mining grammar patterns in corpora for assisted writing. In *Proceedings of ACL-IJCNLP 2015 System Demonstrations*, pages 139–144.