Kashihara, A. et al. (Eds.) (2024). Proceedings of the 32nd International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

Implementation and an Evaluation of a Search Function Allowing Misspelling for a Japanese Learning System

Hidenobu KUNICHIKA[a*] & Miguel Antonio VILLALOBOS ZUNIGA[b]
[a] *Faculty of Computer Science and Systems Engineering, Kyushu Institute of Technology, Japan*
[b] *Graduate School of Computer Science and Systems Engineering, Kyushu Institute of Technology, Japan*
*kunitika@ai.kyutech.ac.jp

Abstract: Search functions are useful in foreign language learning. Learners may sometimes make spelling mistakes. This makes it difficult to find the information they need using a search function. The target of this research is to implement and evaluate a grammar search function allowing misspelling for a Japanese learning system. As the result of an evaluation, we found that our search function is superior to a simple search function in terms of accuracy.

Keywords: search function, misspelling, Japanese learning system, Levenshtein distance

1. Introduction

Search functions are useful in foreign language learning. For example, when a learner encounters unfamiliar words or grammars, she will get necessary information by using a search function on the WWW. The usefulness is the same also in Japanese learning systems, and a retrieval function for example sentences or pages explaining grammar is one of the necessary functions. However, learners may frequently make spelling mistakes, so it is sometimes difficult to get the necessary information. The target of this research is to implement and evaluate a grammar search function allowing misspelling for a Japanese learning system (Villalobos Zuniga, et al., 2022). The search function mitigates the false results generated when a user enters a typo, misspelling or variant spelling into the search tool.

2. Japanese Learning Support

Japanese learning systems should provide a detailed explanation of the grammar, an explanation of the grammar rule, and many example sentences. Many current systems (Kim, 2017; The Japan Foundation, 2022; Vyšný, 2015) have been developed, but most have only one or a few of these features. It also requires a search function that mitigates common mistakes made by non-native learners, as they may produce misspellings. For example, it is common for Spanish speakers to confuse words with long vowels (ありがと(a-ri-ga-to) / ありがとう(a-ri-ga-to-u)), “s” and “z” sounds (さっし(sa-tsu-shi) / ざっし(za-tsu-shi)), or words with small “tsu” (いしょに(i-sho-ni) / いっしょに(i-tsu-sho-ni)).

3. A Japanese Learning System

3.1 Outline

Figure 1 shows the outline of our system (Villalobos Zuniga, et al., 2022). Our system has the tools described above: a detailed explanation of the grammar, grammar rules, and example sentences. In addition, the system has grammar exercises, an explanation of how and when to use the grammar, and a link to the related grammars. The search function will be able to retrieve the correct results even if the user makes typing errors. Finally, it has a custom sorting function that reorders the results by referring to the user's level.

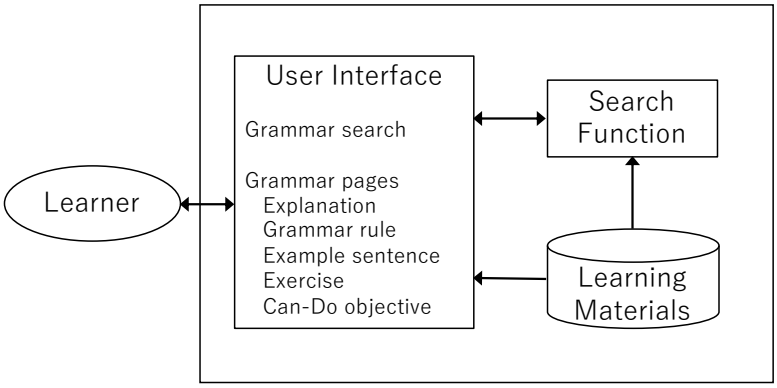

Figure 1. Outline of our system

3.2 The Search Function

Japanese writing systems include kanji, hiragana and katakana. Inconsistent spellings affect retrieval, so our system uses hiragana (hereafter, 'kana') as the internal representation. The search function of the system converts all the text of the grammar pages into its kana version and stores it as the full-text index. When the search queries are executed, the results are retrieved using the full-text index, and then they are filtered using a Levenshtein distance.

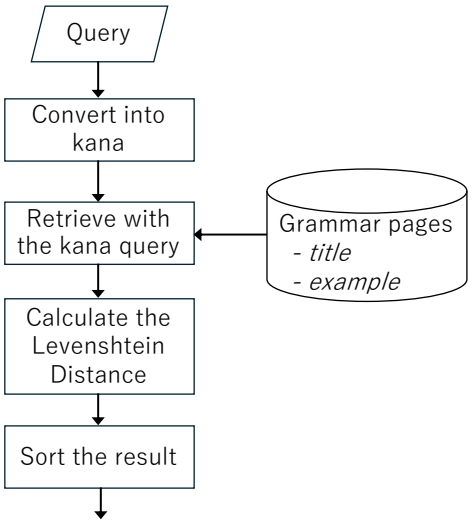

Figure 2. The processes of retrieving a grammar page

The search function consists of the following five sub-processes. First, a morphological analyzer MeCab (Kudo, et al., 2004) is used for word segmentation. Next, the system creates an inverted index from the words in the example field. Figure 2 shows the processes of retrieving a grammar page. After a user enters a query, the system converts the query into its kana form. The system then searches both the title field and the examples field for the query by using Elasticsearch (Elastic, 2024). Finally, filtering is performed by using the Levenshtein distance. The system calculates the Levenshtein distance for each match on a grammar obtained from the results. The summatory of all the Levenshtein distances for all the matches

in a particular grammar will be the total Levenshtein distance for that grammar. After that, the system calculates a score that ranks the grammar using the Inverse Document Frequency formula. The more matches in a query, the more likely it is that the grammar is correct.

Optionally, the user can apply the custom order based on the user level. Each grammar page in the system contains the level information of the grammar. The system estimates the user's level based on the level of the page the user has visited, and customizes the order of search results so that results corresponding to that level appear at the top of the results.

4. An Evaluation of the Search Function

In order to investigate the accuracy of the search function, we prepared a file containing sixty queries. Each of the queries has one misspelling and the corresponding grammar IDs. Those misspellings are based on common mistakes made by Spanish speakers: long vowels, small tsu, and, “s” and “z” sounds, as mentioned above. For each of the queries, grammar pages are retrieved from the search function and ordered by Levenshtein distance.

As a result of the experiment, the percentage of queries where the correct grammar was found within the top ten search results was 98.24% (56 out of 57), while the percentage for a simple search using Elasticsearch was 31.25% (15 out of 48). Our function was able to include the correct grammar in the top results for most queries. On the other hand, in the simple search, the number of cases where the correct grammar was placed at the top of the results was limited. Thus, in the case of using simple search, it will be difficult for users to find the correct grammar.

5. Conclusion

This paper has described a grammar search function allowing misspellings. As an evaluation of the search function, we experimented with queries containing one misspelling. In the experiment, the accuracy rate was 98.24%.

Spelling/grammar checking tools are capable of correcting misspellings, and are easily available, e.g., from Google (Google, 2024). While such tools may be able to make appropriate corrections, they may not be able to correct a word which is misspelled but correct (i.e., misspelled as the word the user intended to write, but correct as a different word). Also, even if the tools present the correct word candidates, the user may not be able to select the appropriate one. Furthermore, since our system retrieves words that are close to the input word, it is possible to compare or study several variations of the word, including its variants.

Although our system allows for misspellings, there is a limitation that users need to select the appropriate result by themselves. In the future, we plan to think about how to support the user in selecting the appropriate one from among the search results.

References

Elastic. (2024). Elasticsearch: The Official Distributed Search & Analytics Engine. https://www.elastic.co/elasticsearch/

Google. (2024). Google Docs. https://www.google.com/docs/about/

Kim, T. (2017). Learn Japanese. https://guidetojapanese.org/learn/grammar

Kudo, T., Yamamoto, K., & Matsumoto, Y. (2004) Applying Conditional Random Fields to Japanese Morphological Analysis, *Proceedings of EMNLP 2004* (pp.230-237).

The Japan Foundation. (2022). MARUGOTO Plus. https://marugotoweb.jp/en/

Villalobos Zuniga, M. A., & Kunichika, H. (2022). Implementation of a Japanese Learning System Equipped with a Grammar Search Function Allowing Misspelling, *Proceedings of ICFULL 2022* (p. 28).

Vyšný, M. (2015). Aedict Online. https://aedict-online.eu