Estimation of Difficulty of Japanese words for Chinese Native Speakers

1910438 LIN Miaoyu

Identification of difficulty of words plays an essential role in teaching Japanese. There are many situations where the difficulty level of a word is considered, such as prioritizing the words to be taught to learners or not using difficult words for beginners. In general, the difficulty of words depends on the learner's background. For example, since learners whose mother tongue is Chinese know Chinese characters, they can learn Japanese words written in Chinese characters more easily than learners whose mother tongue is a language that does not use Chinese characters. However, previous studies on the difficulty of Japanese words did not consider the difference in learners' native languages.

This study aims at estimating the difficulty of Japanese words for native Chinese speakers. We focus on estimation of the difficulty level when learners study a word for the first time, and also focus on only Japanese words written in Chinese characters (Kanji). In order to measure the difficulty level of Japanese words for native Chinese speakers, for a given Japanese word, a Chinese word written in the same Chinese characters is searched first. When it is found, we measure how similar the senses of Japanese and Chinese words are. The more similar the senses are, the lower the difficulty level is regarded to be. The similarity of word senses is measured by the alignment of the senses in a Japanese dictionary and a Chinese dictionary. In addition, more fine-grained difficulty levels are identified by checking whether the glyphs of the Japanese and Chinese words are completely identical or not.

In this study, we define the difficulty level of Japanese words as follows. Based on the classification of Sino-Japanese words in the past document, four difficulty classes are defined in ascending order of difficulty: S (all senses in Chinese and Japanese words are the same), O (some senses in Chinese and Japanese words are overlapped), D (senses in Chinese and Japanese are totally different), and N (the same word does not exist in Chinese). In addition, even when a Japanese Kanji character and Chinese one are the same, they are sometimes represented by different glyphs. Therefore, the classes S, O, and D are subdivided into X-1 when the glyphs of Japanese and Chinese words are the same, and X-2 when they are different, where X stands for S, O or D.

The classification of S, O, D, and N for Japanese Kanji words is carried out as follows. First of all, in a Chinese dictionary, we search the same Chinese word as a given Japanese Kanji word. If the word is not found, we convert the Japanese Kanji character to Simplified Chinese using the

Chinese-Japanese Kanji mapping table, then search again. When the word is not found even after Kanji conversion, it is classified as N. Otherwise, we extract definition sentences of the senses of the words from the Japanese dictionary and two Chinese dictionaries. The Iwanami Japanese Dictionary is used as the Japanese dictionary, while the Hakusuisha Chinese Dictionary and the Contemporary Chinese Dictionary are used as Chinese dictionaries. The senses are written in Chinese in the Contemporary Chinese Dictionary, so they are translated into Japanese using the Baidu Translation API. Next, we calculate the similarity between the word senses in the Japanese and the Chinese dictionary, and align similar word senses. A word sense is represented as a vector, which is the average vector of the distributional representations of words in a definition sentence. The similarity between the word senses in the two dictionaries is defined as the cosine similarity of the two word sense vectors. Among all the combinations of Japanese and Chinese senses, the pair of senses with the highest similarity is aligned first. After removing the aligned word senses from the two dictionaries, the same process is repeated for the rest of the word senses. However, when the similarity between the senses is less than a threshold T_m, that is, when the sense similarity is not sufficiently large, we do not consider that the two senses have the same meaning and terminate the alignment of the senses. After the alignment of word senses is completed, when none of the pairs of word senses can be aligned, it is judged as D. When not all but some of word senses are aligned, it is judged as O. When all of the word senses are aligned, it is judged as S. We perform the sense alignment for each of Chinese dictionaries and Japanese dictionary. If the results of the two dictionaries are different, we calculate the score of the word sense alignment, which is the average of the similarity between the aligned word senses, and choose the result with the highest score. When the Japanese word has two or more parts-of-speech (POSs), the sense sets are subdivided according to its POS, then the sense alignment is carried out for each subset of the senses. Finally, the difficulty class is determined as S-1, O-1 or D-1 when the glyphs of the Japanese and Chinese word are the same, while S-2, O-2, or D-2 when they are not. In addition, we propose another method to use Word Mover's Distance to measure the similarity between the senses. Also, we propose another algorithm that considers many-to-many alignment between the senses.

The proposed method was evaluated from two points of view. First, we evaluated the performance of the proposed method in identifying the difficulty level. In the experiment, only S, O and D were considered as the difficulty level, since the discrimination between N and others, and between X-1 and X-2 was obvious. A test data consisting of 279 Japanese words with its gold difficulty level was manually constructed. Then the accuracy

of the classification of the difficulty level was measured on this test data. As a result, the accuracy was 0.763 when the threshold T_m was set to 0.196 by heuristics. Although there was still room for improvement, it was confirmed that the proposed method could identify the difficulty level of Japanese words with the reasonably high accuracy.

Second, the validity of the proposed word difficulty classes was verified by a questionnaire survey of native speakers of Chinese. We extracted 5 words for each class S, O, D and N from the test data and asked 22 native Chinese speakers to rate the difficulty of these 20 Japanese words on a 5-point scale by a questionnaire. The rank correlation coefficient between the difficulty of the proposed method and the average of the rating evaluated by the Chinese native speakers in the questionnaire survey were calculated. The results showed that for beginners, the correlation coefficient was 0.788 with a p-value of 0.00004. In other words, a strong correlation between them was found, which was statistically significant. Therefore, it is found that the proposed framework is appropriate to measure the difficulty of Japanese words for beginners.

The main contribution of this thesis was to establish the method for automatically estimating the difficulty of Japanese words with the high accuracy, taking into account the characteristics of native Chinese speakers who know Kanji characters.