A Study of Text Simplification by Paraphrase of Phrases

1910074 Kawarai Tsubasa

Text simplification is a technique to paraphrase a sentence or document into a simpler expression while retaining its meaning. It is useful for children, aged people, and non-native speakers of Japanese. For example, if a text that contains many complex words can be paraphrased into a simpler text, it would be helpful for people whose language skill of Japanese is not high to understand the text. The most common method of text simplification is to replace a complex word, which is hard to understand, in a sentence with a simple word, which is easy to understand. The recent studies of text simplification propose various methodology, such as a method to replace a complex word with another word that is easy and the most similar to it, and a method to paraphrase a complex word with a frequently used word in a definition sentence of the complex word in a dictionary. In addition, several studies aim at constructing a database compiling pairs of a complex and simple word as fundamental knowledge for text simplification. However, the paraphrase of words is insufficient for text simplification, although it is relatively easy to implement and the most of the previous studies focus on it. Even when a complex word is paraphrased into a simple word and it is carefully confirmed that they are similar and the latter is easier than the former by word-by-word comparison, the sentence obtained by paraphrase of the words may not be simple or natural. In such a case, paraphrase of the word is not appropriate when a context of the word is considered. Another problem is that variety of simplified texts obtained by paraphrase of words is rather poor. When people simplify a text, they often paraphrase not only words but also phrases, clauses, or a whole document. It is impossible to automatically generate a wide variety of simplified texts as humans write by applying paraphrase of words only.

This study aims at simplifying a text by paraphrase of a phrase. Note that the phrase to be paraphrased is limited to a sequence of words of "noun-particle-verb". Furthermore, when paraphrasing a complex phrase, an appropriate simple phrase is selected from candidates by considering a context of the complex phrase. Comparing with text simplification based on paraphrase of words, simplification by paraphrase of phrases enables us to generate more various and high quality simplified texts as humans do.

In the proposed method, text simplification is performed in five steps. The first step is preprocessing. Morphological analysis of an input text is performed to obtain lexical information of words such as part-of-speech (POS). Then, phrases in the form of "noun-particle-verb" are extracted by referring

POSs obtained by morphological analysis. The second step is construction of "simply paraphrased word database". The existing language resources are used to obtain pairs of a complex word and its simply paraphrased word, and they are compiled as the database. Two language resources are used: one is "Simple PPDB: Japanese" that is a simple paraphrase dictionary, the other is "SNOW D2" that is a paraphrase dictionary. Only pairs of complex and simple words are extracted from them by checking the level of difficulty of words. In addition, pairs of intransitive and transitive verbs are excluded, since their meanings and levels of the difficulty are almost the same. The third step is generation of simple phrases. Candidates of simple phrases to be replaced with a complex phrase are generated using the constructed simply paraphrased word database. Three patterns are used for generation: paraphrasing both a noun and verb, paraphrasing only a noun, and paraphrasing only a verb. The fourth step is selection of the simple phrase. A score is calculated for each candidate of the simple phrase generated in the previous step, considering both the faithfulness and fluency, then the most appropriate simple phrase is selected based on this score. Here the faithfulness evaluates whether the meaning of the sentence is not changed by paraphrase, while the fluency evaluates how natural the paraphrased simple phrase is. The score of the faithfulness is measured by the cosine similarity of sentence embeddings (vector representation of sentences) before and after the paraphrase. Sentence BERT is used to obtain the sentence embedding. The score of the fluency is calculated by the relative frequency of the simple phrase, since a phrase can be recognized as natural when it frequently occurs in a corpus. The frequency of the phrase is obtained by Kyoto University case frame dictionary. The scores of the faithfulness and fluency are combined in two ways, that is, we define two scores for selection of the simple phrase. In these scores, two parameters are introduced: one is used to adjust the scale of the scores of the faithfulness and fluency, the other is the weight parameter for the faithfulness and fluency. In addition, to generate more simple phrases obtained by paraphrase of both the noun and verb, such a phrase is always chosen when its score is not the maximum but greater than a pre-defined threshold. The threshold is determined using development data. The final step is generation of a sentence. The complex phrase of the original sentence is replaced with the chosen simple phrase to generate a simple sentence.

Two experiments were conducted to evaluate our proposed method. In the first experiment, the performance of the selection of the simple phrase was measured by the accuracy, which was the proportion of cases where simple phrases chosen by the system and by the human subject agree to the total number of cases in the test data. The accuracy was changed for the different scores and parameters used in the system as well as the human sub-

jects. It was around 0.45 and 0.51 at maximum. On the other hand, the inter-annotator agreement of two subjects was 0.61. Therefore, it was rather difficult even for humans to select the best simple phrase from candidates. It was found that the idea to rank simple phrase candidates based on the faithfulness and fluency was effective since the accuracy of the system was close to the inter-annotator agreement. In the second experiment, we evaluated the quality of the simple sentences containing the simple phrases generated by the proposed method. Two subjects compared the simple sentences with the original sentences and gave them a five-point rating with respect to three aspects: simplicity, faithfulness, and fluency. As a result, the average rating of the two subjects was approximately 4 points for each aspect. Since the ratings were high, the effectiveness of the proposed method was confirmed. In addition, the weighted κ coefficient between ratings of two subjects was calculated. They were high, 0.93, 0.89, and 0.95 for the simplicity, faithfulness, and fluency, respectively. It indicated that the evaluation of two subjects was stable. In addition, we performed an error analysis on the results of the first experimental. We investigated the reason why the simple phrases selected by the proposed method and human subjects disagreed, then discussed the current problems of the proposed method and future directions.

The main contribution of this thesis was that it proposed the method of text simplification by paraphrasing phrases considering the faithfulness and fluency, which could generate sufficiently high quality simple sentences for being read by a human.