Abstract

Online product reviews play a crucial role for both consumers who make purchase decisions and companies who look for the reputation of their products. However, not all reviews do not provide useful information, i.e., are not helpful. Therefore, the research on the helpfulness of reviews has attracted much interest. Techniques to automatically identify helpful reviews are crucial because it is hard and time-consuming to read a large amount of product reviews. Besides, the helpfulness of reviews can be defined from various points of view. Reasons or grounds for a reviewer's opinion are one of the important features; a review that expresses not only an opinion but also its grounds, such as "I like it since the design is cool," can be helpful for both consumers and companies. However, the previous studies have not paid much attention to the identification of the helpfulness of reviews from this point of view.

The goal of this research is to classify whether a review contains the user's opinion and ground for it to select helpful reviews from many product reviews. Our proposal to achieve this goal consists of two subtasks. The first one is "opinion-ground classification task". It is a task to classify whether a given pair of sentences or clauses in a review include a user's opinion and its ground. The second one is "opinion-target classification task". It is a task to classify whether a target of an opinion is a product or not (whether a user expresses his/her opinion about a product). We believe that the second task is necessary, since a review may not be helpful when it includes a user's opinion but the opinion is not related to a product.

Our proposed method for the opinion-ground classification task is as follows. First, pairs of clauses under the dependency relation are extracted from a given review. Then each pair of clauses is classified as whether it is an "opinion-ground clause pair" (an opinion and its ground appear in it) or a "non-opinion-ground clause pair". Four kinds of classifiers are trained for the opinion-ground classification. (1) Rule-based method. It judges an input as an opinion-ground clause pair when it contains the discourse markers "*kara*" and "*node*", which are Japanese discourse markers indicating causal relation. (2) Bidirectional Encoder Representations from Transformers(BERT) fine-tuned using the training data. (3) Intermediate Fine-Tuning(IFT) of BERT. The BERT model is fine-tuned using another task-related and relatively large dataset first, then it is fine-tuned again using the labeled dataset of the target task. (4) Hybrid method that combined the rule-based and BERT model. The rule-based method is applied first. Then, when it classifies an input as a non-opinion-ground clause pair, the BERT model is applied to make the final decision.

To train the models, we use or construct three labeled datasets. The first is

the existing dataset for discourse analysis, Kyoto University Web Document Leads Corpus (KWDLC). The clause pairs under the "cause/reason" relation in KWDLC are extracted as positive samples (opinion-ground clause pair) and ones under other relations as negative samples. Although KWDLC is a manually annotated corpus, the "cause/reason" relation in KWDLC is not the same as the opinion-ground relation in this study. In addition, KWDLC is an out-domain dataset; the domain of KWDLC is Web documents, while the domain of the opinion-ground classification task is product reviews. The second is the dataset constructed by discourse markers. From unlabeled reviews, clause pairs that include the discourse marker "*kara*" or "*node*" are excerpted as positive samples, while negative samples are made by concatenating randomly chosen two clauses. It is an in-domain dataset (dataset of product reviews), although it may contain incorrect samples. The third is another in-domain dataset augmented by ChatGPT. A small number of sentences/clauses including an opinion and its ground are prepared as seed clauses. Then, we give a seed clause and a prompt to ChatGPT so that ChatGPT generates 40 clauses similar to the seed. Positive samples are made by concatenating the automatically generated clause or the seed clause with another randomly chosen clause, while negative samples are made in the same way in the dataset obtained by the discourse marker. For IFT, the out-domain dataset (KWDLC) is used in the first training phase, and two in-domain datasets are used in the second phase.

Our proposed method for the opinion-target classification task is a simple method based on keyword matching. First, for each product category such as "food" and "books", a set of keywords is constructed. Here the keyword refers to an important word that is frequently used in reviews about products of that category. Our classifier judges an input clause pair as positive or negative by checking whether it contains one of the keywords in the keyword set of the product category or not.

Five methods are used for keyword extraction. (1)TF-IDF. Content words with high TF-IDF scores are extracted. (2)TF-ENT. Words with high TF-ENT scores are extracted. Here ENT is the score that evaluates the salience of the product category (i.e., how likely a word is used only in one product category) by measuring the entropy of the distribution of probabilities that a word appears in product categories. (3)YAKE!. Words are extracted by YAKE!, the existing unsupervised keyword extraction method. (4)TF-IDF+YAKE!. Words are extracted by TF-IDF method and YAKE!. (5)TF-ENT+YAKE!. It is an ensemble of TF-ENT and YAKE!. For each method, the 200 top-ranked words are extracted as a keyword set.

Several experiments were carried out to evaluate our proposed method. The test data for the opinion-ground classification task was made by manually annotating 506 clause pairs with gold labels. For 186 opinion-ground clause pairs in this

dataset, another gold label is manually added to make test data of the opinion-target classification task.

As for the opinion-ground classification task, the results showed that the use of automatically constructed datasets significantly improves the classification performance. The F1-score of our best model is 0.71 by using IFT, which is 0.09 points higher than the BERT model trained from the existing dataset KWDLC only. It proves that our approach to combining the heterogeneous datasets is effective. In addition, the IFT is more appropriate than a simple combination to utilize those heterogeneous datasets. Besides, the hybrid method that combines the rule-based method and the BERT model did not usually outperform the single BERT model since the rule-based method could classify only 23% of the test data, while BERT obtained by IFT performed well on this portion of the test data (its precision was 0.79).

As for the opinion-target classification task, comparing TF-IDF and TF-ENT, TF-IDF achieved better recall and F1-score. YAKE! outperformed both TF-IDF and TF-ENT methods, especially in recall. Finally, TF-IDF+YAKE! achieved the best performance. Its recall and F1-score were 0.91 and 0.89, respectively, which were the highest among five methods, while its precision was comparable to other methods. It was found that the different keywords were extracted by TF-IDF and YAKE!. We guessed that it is the reason why the combination of TF-IDF and YAKE! can improve the performance of the opinion-target classification task.