Identification of Opinion and Ground in Customer Review Using Heterogeneous Datasets

Po-Min Chuang, Kiyoaki Shirai and Natthawut Kertkeidkachorn
Graduate School of Advanced Science and Technology, Japan Advanced Institute of Science and Technology, 1-1 Asahidai Nomi, Ishikawa, Japan

Keywords: Opinion Mining, Ground of Opinion, Customer Review, Discourse Analysis, Weakly-Supervised Learning.

Abstract: Online reviews are a valuable source of information for both potential buyers and enterprises, but not all reviews provide us helpful information. This paper aims at the identification of a user's opinion and its reason or ground in a review, supposing that a review including a ground for an opinion is helpful. A classifier to identify an opinion and a ground, called the opinion-ground classifier, is trained from three heterogeneous datasets. The first is the existing dataset for discourse analysis, KWDLC, which is the manually labeled but out-domain dataset. The second is the in-domain but weakly supervised dataset made by a rule-based method that checks the existence of causality discourse markers. The third is another in-domain dataset augmented by ChatGPT, where a prompt to generate new samples is given to ChatGPT. We train several models as the opinion-ground classifier. Results of our experiments show that the use of automatically constructed datasets significantly improves the classification performance. The F1-score of our best model is 0.71, which is 0.12 points higher than the model trained from the existing dataset only.

1 INTRODUCTION

As online shopping platforms experience growing traffic,[1] customer reviews are valuable information sources for customers and enterprises. When potential customers consider purchasing products, they usually refer to reviews written by other users. Companies also use customer reviews to know what features users like or dislike to enhance their products. However, it is hard for users to find valuable information from a vast amount of reviews. Studies to automatically extract useful information from reviews or to support users in finding helpful information are paid much attention.

Previous research has explored various aspects of analyzing customer reviews such as summarizing reviews (Hu and Liu, 2004). Especially, the research on the helpfulness of reviews has attracted much interest (Diaz and Ng, 2018), since some reviews are useless or meaningless for customers or companies. Techniques to automatically identify helpful reviews are crucial, because it is hard and time-consuming to read a large amount of customer reviews.

The helpfulness of reviews can be defined from various points of view. For example, the helpfulness can be evaluated by the length of a review (a long review may be more helpful than a short one), detailed explanation about a product, and so on. Reasons or grounds for a reviewer's opinion are also one of the important features. On the one hand, a review that just expresses an opinion or sentiment toward a product such as "It is excellent." and "Too bad." provides not so useful information to other people. On the other hand, if a reviewer also writes a reason why she/he thinks a product is good or not, such as "I like it since the design is cool," it can be helpful for both potential customers and companies.

Our goal is to identify an opinion of a reviewer and a ground for it in a customer review written in Japanese in order to select and provide helpful reviews to users. It is formulated as a kind of discourse analysis between two clauses where one clause expresses an opinion and the other represents its ground. A pre-trained language model is used since it has achieved superior results in many Natural Language Processing tasks. However, no labeled data, a collection of Japanese reviews annotated with relations between an opinion and a ground, is available. Therefore, we use three kinds of datasets for training a model: (1) an existing out-domain dataset of dis-

[1] https://www.marketplacepulse.com/stats/amazon-online-stores-sales

Chuang, P., Shirai, K. and Kertkeidkachorn, N.
Identification of Opinion and Ground in Customer Review Using Heterogeneous Datasets.
DOI: 10.5220/0012307000003636

In *Proceedings of the 16th International Conference on Agents and Artificial Intelligence (ICAART 2024) - Volume 2*, pages 68-78
ISBN: 978-989-758-680-4; ISSN: 2184-433X

course analysis, (2) an in-domain dataset automatically constructed by using a discourse marker, and (3) an augmented in-domain dataset made by a generative AI. Our research questions are: what datasets are useful for the identification of an opinion and its ground, and how can those datasets be employed to train an effective model?

The main contributions of this work are summarized as follows:

- We define a new task to identify an opinion and its ground in a customer review toward retrieval of helpful reviews.
- We propose novel methods to automatically construct labeled datasets for the above task.
- We evaluate how the datasets constructed by our methods can contribute to improving the performance of the identification of an opinion and a ground.

The remainder of this paper is structured as follows. Section 2 provides a brief introduction to related work. Section 3 describes details of our proposed method to construct the labeled data for training. Section 4 reports several experiments to evaluate our methods. Finally, we conclude the paper in Section 5.

2 RELATED WORK

2.1 Helpfulness of Review

Online consumer reviews can significantly influence consumers' purchasing decisions (Zhu and Zhang, 2010) and also give useful information to improve products. Since not all reviews are useful for customers and companies, many studies have been carried out to identify whether a review is helpful or not. Diaz and Ng carry out a survey of the relevant work on the prediction of helpful product reviews (Diaz and Ng, 2018).

Several studies try to discover useful reviews based on the helpfulness votes given by other reviewers at an EC website. Kim et al. employ Support Vector Machine (SVM) regression for prediction of helpfulness using 5 classes of features (Kim et al., 2006). Mudambi and Schuff apply Tobit regression for the prediction of customer review helpfulness, indicating that the extremity, depth, and product type of reviews affect the perceived helpfulness of the review (Mudambi and Schuff, 2010). Pan and Zhang utilize a mixed-effects logistic model with a random intercept to analyze various factors that influence the helpfulness. They show that the valence (rating) and length of reviews are positively associated with the perceived helpfulness. In addition, they examine contents of reviews and reveal a positive correlation between the reviewer's innovativeness and the helpfulness (Pan and Zhang, 2011).

Some researchers point out that helpfulness votes are not always good ground-truth indicators due to voting bias (Liu et al., 2007; Tsur and Rappoport, 2009; Yang et al., 2015). For example, Liu et al reveal three kinds of voting biases. The imbalance vote bias is a tendency in which users value positive opinions rather than negative ones. The winner circle bias means that reviews that receive a lot of votes tend to attract more attention, leading them to accumulate votes at a faster rate than others. The early bird bias means that the earlier a review is posted, the more votes it will get.

Due to such biases of the helpfulness votes, datasets of reviews manually annotated with labels of the helpfulness are constructed and used for helpfulness prediction. Almagrabi et al. annotate user reviews, which are written for five electronic products on Amazon.com and CNet.com, with a binary label of helpfulness (Almagrabi et al., 2018). Liu et al. propose a guideline called SPEC that defines four categories of review quality which represent different values of the reviews to users' purchase decision (Liu et al., 2007). They construct the dataset by annotating the reviews in Amazon.com following the SPEC, and then SVM is trained using this dataset as a binary classifier to determine whether a review is helpful or not. Tsur and Rappoport implement an unsupervised algorithm RevRank (Tsur and Rappoport, 2009). The Virtual Core Review (VC) is made by extracting particularly important terms from reviews for a specific book. Then the reviews are ranked by the similarity with the VC as the helpfulness score. Gamzu et al. construct a collection of review sentences labeled with continuous scores of helpfulness and use it for a task to extract one positive and one negative Representative Helpful Sentence (RHS) from a given set of reviews, which is a kind of multiple document summarization (Gamzu et al., 2021). Yang et al. use a regression model to predict the helpfulness score of a product review using only the review text (Yang et al., 2015). They employ two interpretable semantic features and use human scoring of helpfulness as ground truth. The results show the models trained with semantic features are more easily applied to the reviews in the different product categories.

This study defines the helpfulness of a review in a different way than the previous work. That is, a review is helpful if it contains the user's opinion and its ground. Although this definition is not universal, the

Table 1: Example of opinion-ground and non-opinion-ground clause pairs.

	Clause1	Clause2
opinion-ground clause pair	インタビュアーとの対談形式になっているので、(Since this book is written in the form of a conversation with an interviewer.)	内容はわかりやすく説得力があります。(The contents is easy to understand and convincing.)
non-opinion-ground clause pair	マスキングテープは色々と使えて便利です。(Masking tape is useful because it can be used for many purposes.)	メール便で発送してもらえるのも嬉しいサービスです。(It is nice for me that it is shipped by the easy mail service.)
opinion-ground clause pair	とっても可愛いくて大満足だったのですが、(It is very cute, so I am fully satisfied with it,)	近くのショップでもっと安く販売されてました。(but I found that it is on sale at the near shop much cheaper.)

proposed method enables us to evaluate the helpfulness of a review from a new point of view.

2.2 Discourse Relation

Our task is related to the discourse structure analysis where the relation between two sentences or clauses is identified. Prasad et al. construct Penn Discourse TreeBank (PDTB), a corpus of English newspaper articles where discourse relation tags are assigned to pairs of clauses (Prasad et al., 2008; Prasad et al., 2018). Among several relation types in PDTB, "CONTINGENCY.Cause" is the most related to this study. It means the causal relation, i.e., a clause represents a cause of an event in another clause. Discourse structure analysis using PDTB has got lots of attention. However, Kim et al. claim that the lack of consistency in preprocessing and evaluation protocol poses challenges to fair comparison of results in the literature (Kim et al., 2020). They propose the standard label sets and the protocol of the section-based cross-validation for fair comparison of methods for implicit discourse relation classification, where there is no explicit discourse marker between two spans or arguments, and develop two strong baselines using Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019) and XLNet (Yang et al., 2019).

Kishimoto et al. construct Kyoto University Web Document Leads Corpus (KWDLC), a Japanese dataset annotated with discourse relations (Kishimoto et al., 2020). It consists of a small-scale high-quality corpus made by experts and a large-scale corpus made by non-expert crowdworkers. Using KWDLC as the training and test data, they implement and compare four methods of discourse relation analysis: the BERT model, a machine learning classifier called "opal", a classifier based on discourse markers, and an ensemble model of the classifier by discourse markers and the BERT model.

Unlike previous studies on discourse structure analysis where many types of relations are considered, our task concentrates on the relation between an opinion and its ground, which is similar to the causal relation, on the restricted domain, i.e., customer reviews.

3 PROBLEM STATEMENT

Our task aims to identify whether a given pair of clauses contains a reviewer's opinion and a ground or reason for that opinion. The input in this task is not a review or a sentence but a pair of clauses, since an opinion and its ground are often expressed in two related clauses. In the example below, the positive opinion ("fully satisfied") is expressed in the clause c_2, and its ground ("purchase ... at a cheap price") is in the clause c_1.

$$\underbrace{\text{I can purchase this high-end model at a cheap price,}}_{c_1} \underbrace{\text{so I am fully satisfied with it.}}_{c_2}$$

We call this task the "opinion-ground classification task", and a pair of clauses that contain an opinion and its ground "opinion-ground clause pair". It is a new task that can be applied to judge whether a review is helpful or not. It is also a challenging task since there is no labeled data for this task.

Examples of an opinion-ground clause pair and a non-opinion-ground clause pair are shown in Table 1. These clauses are excerpted from product reviews in Japanese. An English translation is shown in the parentheses. As for the opinion-ground clause pair in the second row, Clause 2 expresses the opinion "the book is easy to read", while Clause 1 shows the reason why it is. On the other hand, no such a relation is found between Clause 1 and Clause 2 of the non-opinion-ground clause pair at the third row. Note that a clause can be a whole sentence, which is usually

short, such as the clauses of the non-opinion-ground clause pair.

Although we suppose that an opinion and its ground are in separate clauses, in the preliminary investigation, we found that an opinion and a ground were often appeared in one clause. In fact, one clause expresses both a user's opinion and its ground in nearly 60% of the opinion-ground clause pairs. An example is shown in the last row in Table 1. Clause 1 contains the positive opinion "fully satisfied" and its ground "very cute". In our task, whether an opinion and its ground are in one or two clauses, we aim to classify a pair of clauses whether it is the opinion-ground clause pair or not.

4 PROPOSED METHOD

4.1 Overview

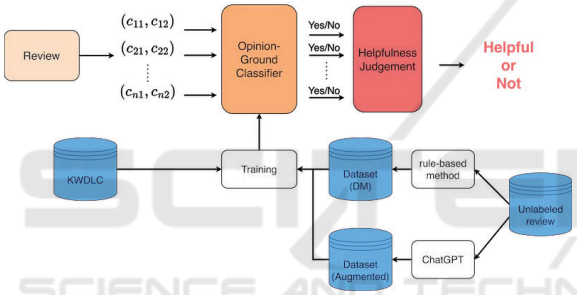

Figure 1: Overview of proposed method.

Figure 1 shows an overview of our proposed method. A review that consists of several sentences is given as an input. It is split into clauses, then pairs of clauses under a dependency relation are extracted, which is denoted as (c_{i1}, c_{i2}) in Figure 1. For each clause pair, the opinion-ground classifier judges whether it contains an opinion and its ground. This procedure is the main problem in this paper as denoted in Section 3. If one of the clauses is classified as "yes", the input review is judged as a helpful review.

The opinion-ground classifier is trained from three datasets. The first is the KWDLC which is the existing dataset of discourse relations. The second is the dataset constructed by using a discourse marker, which is made from unlabeled reviews by a rule-based method. The third is the dataset augmented by ChatGPT.

4.2 Clause Segmentation

Pairs of clauses are extracted from a given review. First, a review is split into sentences by using a period as a sentence boundary. Second, each sentence is split into clauses by a comma. Next, each sentence is analyzed by the Japanese dependency parser KNP (Kurohashi and Nagao, 1994)[2]. KNP outputs the dependency between *bunsetsu*, which is a linguistic unit similar to a base phrase or a chunk. Since *bunsetsu* is a smaller unit than a clause, we convert the dependency between *bunsetsu* to that between clauses.

Then, pairs of clauses under the dependency relation are extracted as intra-sentence clause pairs. Inter-sentence clause pairs are also extracted. Since a head clause of a sentence in Japanese is the last clause, the last clauses of two consecutive sentences are extracted. Figure 2 shows an example of a result of the dependency parsing, where c_i^1 and c_i^2 represent clauses of the first and second sentence, respectively. We extract (c_1^1,c_3^1), (c_2^1,c_3^1), (c_3^1,c_4^1), (c_1^2,c_3^2) and (c_2^2,c_3^2) as the intra-sentence clause pairs, and (c_4^1,c_3^2) as the inter-sentence clause pair.

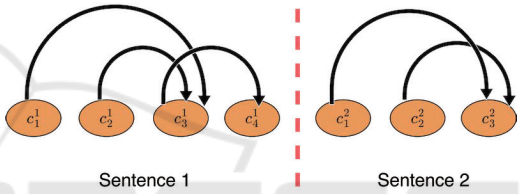

Figure 2: Dependency analysis between clauses.

4.3 Training Data

4.3.1 Kyoto University Web Document Leads Corpus

The Kyoto University Web Document Leads Corpus (Kishimoto et al., 2020) [3] is used as one of our training datasets. This corpus compiles three sentences at the beginning of Web documents in a wide variety of genres. It is annotated with multiple Japanese discourse relations. A set of seven coarse-grained Japanese discourse relation tags is defined referring to that of PDTB 2.0 (Prasad et al., 2008). The tag most related to this study is "cause/reason", which is the same as "CONTINGENCY.Cause" in PDTB 2.0 and PDTB 3.0.

KWDLC includes two datasets. One is the expert dataset which is a relatively small but high-quality dataset. It consists of 2,320 clause pairs annotated by three experts. The other is a large one called the crowdsourcing dataset, which consists of 40,467 clause pairs. Ten crowdworkers assign the discourse

[2]https://github.com/ku-nlp/knp/
[3]https://github.com/ku-nlp/KWDLC

relation tag for each pair of clauses. The final discourse relation is determined by the majority voting of 10 assigned tags. Besides, we define the reliability of the relation in the crowdsourcing dataset as the proportion of the number of majority tags to the total number of tags (i.e., 10).

The dataset for our task is constructed as follows. Pairs of clauses labeled with the "cause/reason" tag are extracted as positive samples. We use all pairs in the expert dataset and only pairs whose reliability is higher than or equal to 0.7 in the crowdsourcing dataset. The same number of negative samples are chosen from clause pairs annotated with the discourse relation tag other than "cause/reason".

It is worth noting that the aforementioned dataset is not exactly fit for the opinion-ground classification task. First, KWDLC is the out-domain. Web documents in various domains are annotated in KWDLC, while a customer review is the target domain of the opinion-ground classification task. Second, the definitions of the target relation are similar but not exactly the same. The relation "cause/reason" in KWDLC represents a general causality, while the opinion-ground concept in our task focuses on a more specific relation. Table 2 shows an example of a clause pair annotated with the "cause/reason" tag in KWDLC. Looking for suppliers (in Clause 1) is the reason to ask people to contact (in Clause 2), but it is not a relation between an opinion and its ground.

Table 2: Example of cause/relation clause pairs in KWDLC.

Clause 1	Clause 2
当社では、新たな発想で商品提案をして頂けるお取引先様を募集しています。(We are looking for suppliers who can propose products with new ideas.)	ご希望の方は下記メールアドレスからご連絡をお願い致します。(If you are interested in working with us, please contact us at the e-mail address below.)

4.3.2 Dataset Constructed by Using Discourse Marker

We construct an in-domain dataset, i.e., customer reviews labeled with the opinion-ground relation tag, by a rule-based method. First, unlabeled reviews are prepared, then pairs of clauses are extracted as described in Subsection 4.2. If the clause contains "から''(*kara*) or "ので''(*node*) that means "because", this clause pair is extracted as a positive sample in which an opinion and its ground are included. Both "から''(*kara*) and "ので''(*node*) are the Japanese discourse markers that represent the causal relation. Table 3 shows examples of the extracted positive samples where the discourse marker is underlined. In addition, pairs of clauses not including these discourse markers are extracted as negative samples. To form the dataset, an almost equal number of positive and negative samples are excerpted.

Table 3: Example of clause pairs including discourse markers.

Clause 1	Clause 2
送料無料でレビューも良かったので、(Because the shipping charge is free and I found positive reviews,)	評判が良ければまた買いたいと思います。(I'd buy it again if its reputation is good.)
今回はきちんとショップを選んで購入したつもりですから、(This time, because I intended to carefully choose the store and buy from it,)	お店の言うとおりの品質のお米が送られてくることと期待しています。(I expect to receive rice of the same quality as the store says.)

The disadvantage of this approach is that the relation indicated by "から''(*kara*) and "ので''(*node*) is not exactly the same as the opinion-ground relation. In addition, the positive samples always contain these discourse markers, although an opinion and its ground can often appear without accompanying them. However, the substantial advantage is that the dataset can be constructed fully automatically without heavy human labor. Another merit is that it is the in-domain dataset, unlike KWDLC.

4.3.3 Dataset Augmented by ChatGPT

Dai et al. and Gilardi et al. show that ChatGPT can provide high-quality augmentation data (Dai et al., 2023; Gilardi et al., 2023). This study also utilizes ChatGPT-3.5 for data augmentation toward the opinion-ground classification task. Especially, we aim to obtain positive samples that do not contain the discourse markers, which are not included in the dataset constructed in 4.3.2.

Figure 3 shows how the dataset is constructed by ChatGPT. First, 17 single clauses that contain both an opinion and its ground are excerpted from unlabeled customer reviews for products of 9 different product types (genres). For each clause, we give a prompt to ChatGPT so that ChatGPT generates 40 similar clauses that include an opinion and its ground. Specifically, the following prompt is used to create new clauses:

君はデータ拡張生成器、これと似た意見に対する根拠を含む商品レビュー 1 節だけ 40 個を生成してください。節：⟨original clause⟩
(You are a data augmentation generator, generate 40 similar one-clause product reviews that include a ground of an opinion. Clause: ⟨original clause⟩)

Since our task is a clause pair classification and what we generate is a single clause, the generated clauses as well as the original clause are coupled with another clause that is randomly chosen from the review of the same product genre to form a pair of clauses. A half of the constructed clause pair contains the clause generated by ChatGPT as the first clause and another half as the second clause.

The above method only generates positive samples. Almost an equal number of negative samples are produced by the same method used in the development of the dataset described in 4.3.2.

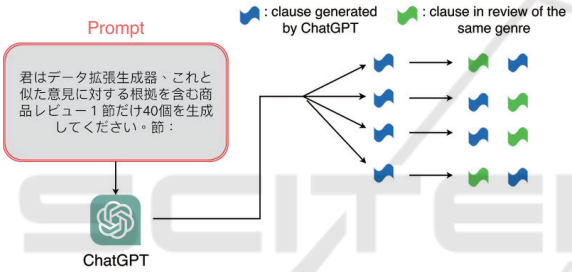

Figure 3: Data augmentation by ChatGPT.

Table 4 shows an example of augmentation. c_o is the original clause, while c_{g1} and c_{g2} are the clauses generated by ChatGPT. These clauses have almost the same meaning with the original one, preserving the opinion and its ground.

Table 4: Example of the augmented clauses by ChatGPT.

c_o	良いものが安く買えて良かったです。(It was good to buy a good one cheap.)
c_{g1}	予想以上に良い商品が手頃な価格で手に入って嬉しいです。(I'm delight to find that this product not only exceeded my expectations but also comes at an affordable price.)
c_{g2}	手ごろな価格で、品物も良いのでとても満足しています。(I am very satisfied because the product is good and its price is reasonable too.)

4.4 Opinion-Ground Classifier

We implement four classifiers for the opinion ground classification task.

4.4.1 Rule-Based Method

A simple rule-based method is implemented. The Japanese morphological analysis engine Janome[4] is used for word segmentation of input clauses. Then, this classifier checks whether an input pair of clauses contains the causal discourse marker "から"(*kara*) or "ので"(*node*). If an input contains the discourse marker, it is classified as the opinion-ground clause pair.

4.4.2 BERT

BERT is fine-tuned for the opinion-ground classification task. The pre-trained BERT model[5] that is obtained from the Japanese version of Wikipedia is used.

Two kinds of BERT models called BERT-2C and BERT-1C are fine-tuned. BERT-2C uses BERT as a sentence pair classification model where a pair of clauses are given as the input, while BERT-1C is a single sentence classification model where the concatenation of two clauses is given. The input of each model is shown as follows:

BERT-2C: ⟨CLS⟩ Clause1 ⟨SEP⟩ Clause2 ⟨SEP⟩
BERT-1C: ⟨CLS⟩ Clause1 Clause2 ⟨SEP⟩

4.4.3 Intermediate Fine-Tuning

Intermediate fine-tuning (IFT) is the process of fine-tuning a model using an intermediate dataset before fine-tuning using the target dataset in order to enhance the model performance (Cengiz et al., 2019; Poth et al., 2021). Cengiz et al. showed that using IFT improved the performance on the medical NLI task (Cengiz et al., 2019). IFT on a dataset of the same or related task transfers more task-specific knowledge to the model for the target domain.

IFT is applied to train a better classification model. Two datasets are prepared: one is the out-domain dataset which is KWDLC, the other is the in-domain dataset which is the union of the dataset constructed by using the discourse marker and the dataset augmented by ChatGPT. The BERT model is fine-tuned using the out-domain dataset first. Using the trained model to provide initial parameters, the BERT model is fine-tuned again using the in-domain dataset.

4.4.4 Hybrid Model

Rule-based methods are often more interpretable and transparent than deep learning models like BERT. In addition, the rule-based method is expected to achieve

[4]https://github.com/mocobeta/janome
[5]https://huggingface.co/cl-tohoku/bert-base-japanese-v2

high precision in compensation for low recall. By leveraging the strengths of both approaches, we first use the rule-based method to classify an input clause pair. If the rule-based method classifies it as a non-opinion-ground clause pair, we use a fine-tuned BERT model to classify it again.

5 EVALUATION

5.1 Experimental Setup

5.1.1 Dataset

Following five datasets are used to train the models for the opinion-ground classification task.

D_{kwdlc} **:** KWDLC (Subsubsection 4.3.1). It is the out-domain dataset.

D_{dm} **:** Dataset constructed by using the discourse marker (Subsubsection 4.3.2). It is the in-domain dataset.

D_{gpt} **:** Dataset augmented by ChatGPT (Subsubsection 4.3.3). It is also the in-domain dataset.

D_{dm+gpt} **:** $D_{dm} + D_{gpt}$, the union of two in-domain datasets.

D_{all} **:** $D_{kwdlc} + D_{dm} + D_{gpt}$.

D_{dm} and D_{gpt} are automatically constructed from unlabeled reviews. We use reviews in Rakuten Ichiba (the EC website in Japan) in the Rakuten Data Collection (Rakuten Institute of Technology, 2023). Table 5 reveals the number of clause pairs in the training and validation data of five datasets, where "og" and "non-og" indicate the opinion-ground and non-opinion-ground clause pairs, respectively. D_{dm} and D_{gpt} are split into the training and validation data in the ratio of 80:20. D_{kwdlc} are split into the training, validation, and test data in the ratio of 80:10:10. We call the test data "KWDLC test data", whose statistics are also shown in Table 6.

Table 5: Number of clause pairs in datasets.

Dataset	Training		Validation	
	og	non-og	og	non-og
D_{kwdlc}	1,044	1,043	131	130
D_{dm}	582	560	146	140
D_{gpt}	558	560	140	140
D_{dm+gpt}	1,140	1,120	286	280
D_{all}	2,184	2,163	417	410

Another test data is manually constructed to evaluate our proposed methods. From unlabeled reviews in Rakuten Ichiba, 162 reviews are randomly chosen.

Table 6: Test data.

	og	non-og
KWDLC test data	131	131
Rakuten test data	186	329

These reviews are mutually exclusive with ones used for the construction of D_{dm} and D_{gpt}. Then, pairs of clauses are extracted from the reviews by the process described in Subsection 4.2. If the preceding clause ends with a comma, we convert it to a period and also convert a conjugated form of the last verb into its base form so that the clause becomes a complete sentence. For each pair of clauses, two human annotators discuss and judge whether it includes a reviewer's opinion and ground. We call it "Rakuten test data" whose statistics are shown in Table 6.

5.1.2 Models

The models described in Subsection 4.4 are compared: the rule-based method (Rule), the BERT model obtained by the ordinary fine-tuning (BERT), the BERT model obtained by intermediate fine-tuning (IFT), and the hybrid method of Rule and BERT or IFT (Rule+BERT or Rule+IFT). As for the hyperparameters for fine-tuning the BERT, we simply set the batch size to 16 and the learning rate to $2e^{-5}$, but optimize the number of epochs among $\{1,2,3\}$ on the validation data. The AdamW optimizer (Loshchilov and Hutter, 2019) is used for the optimization, where the weight decay is set to 0.01. Training and evaluation of the models are performed only once to shorten the time for the experiment, although several runs should be done considering influence of random initialization of parameters.

The system using ChatGPT-3.5 is also compared. As a prompt, we give ChatGPT three positive and three negative samples, which are artificially made, and ask ChatGPT to judge whether clause pairs in the test data are the opinion-ground clause pairs or not.

5.2 Result and Discussion

5.2.1 Results of Opinion-Ground Classification

Table 7 shows the results of the several methods trained on five different datasets. The precision, recall, and F1-score of the detection of the opinion-ground clause pairs and the accuracy of the classification on the Rakuten test are measured. Each of them is shown in the individual tables. The suffix "-2C" or "-1C" in the model names indicates that a pair of clauses or a concatenated single clause is given

Table 7: Results of opinion-ground classification on Rakuten test data.

(a) Precision

Model	D_{kwdlc}	D_{dm}	D_{gpt}	D_{dm+gpt}	D_{all}
Rule	0.72	0.72	0.72	0.72	***0.72***
ChatGPT	0.40	0.40	0.40	0.40	0.40
BERT-2C	0.43	0.47	**0.59**	0.52	0.56
BERT-1C	0.40	**0.47**	0.46	0.45	0.45
IFT-2C	–	–	–	–	0.67
IFT-1C	–	–	–	–	0.51
Rule+BERT-2C	0.44	0.49	**0.59**	0.50	0.58
Rule+IFT-2C	–	–	–	–	0.63

(b) Recall

Model	D_{kwdlc}	D_{dm}	D_{gpt}	D_{dm+gpt}	D_{all}
Rule	0.47	0.47	0.47	0.47	0.47
ChatGPT	0.92	0.92	0.92	0.92	0.92
BERT-2C	**0.94**	0.52	0.69	0.82	0.85
BERT-1C	**0.95**	0.87	0.91	0.91	***0.95***
IFT-2C	–	–	–	–	0.75
IFT-1C	–	–	–	–	0.94
Rule+BERT-2C	**0.91**	0.65	0.78	0.80	0.82
Rule+IFT-2C	–	–	–	–	0.76

(c) F1-score

Model	D_{kwdlc}	D_{dm}	D_{gpt}	D_{dm+gpt}	D_{all}
Rule	0.57	0.57	0.57	0.57	0.57
ChatGPT	0.56	0.56	0.56	0.56	0.56
BERT-2C	0.59	0.49	0.64	0.64	**0.68**
BERT-1C	0.57	**0.61**	**0.61**	0.60	**0.61**
IFT-2C	–	–	–	–	***0.71***
IFT-1C	–	–	–	–	0.66
Rule+BERT-2C	0.59	0.56	0.67	0.62	**0.68**
Rule+IFT-2C	–	–	–	–	0.69

(d) Accuracy

Model	D_{kwdlc}	D_{dm}	D_{gpt}	D_{dm+gpt}	D_{all}
Rule	0.74	0.74	0.74	0.74	0.74
ChatGPT	0.42	0.42	0.42	0.42	0.42
BERT-2C	0.52	0.61	**0.71**	0.66	**0.71**
BERT-1C	0.48	**0.59**	**0.59**	0.56	0.57
IFT-2C	–	–	–	–	***0.77***
IFT-1C	–	–	–	–	0.65
Rule+BERT-2C	0.55	0.63	**0.73**	0.64	0.72
Rule+IFT-2C	–	–	–	–	0.72

as the input of BERT.[6] The performance of "Rule" and "ChatGPT" is shown as the same for all datasets since they do not use the labeled dataset, while the performance of "IFT" and "Rule+IFT" is shown only for D_{all} since it uses both D_{kwdlc} and D_{dm+gpt}. The best dataset for each model is indicated in bold, while the best model trained on D_{all} is indicated in bold and italics.

[6]The results of Rule+BERT-1C and Rule+IFT-1C are omitted since their performance was poorer than Rule+BERT-2C and Rule+IFT-2C.

Table 8: Results of the models trained from D_{kwdlc} for the in-domain (KWDLC) and out-domain (Rakuten) test data.

Method	Test data	P	R	F	A
Rule	KWDLC	0.74	0.40	0.52	0.63
	Rakuten	0.72	0.47	0.57	0.74
BERT-2C	KWDLC	0.78	0.85	0.82	0.81
	Rakuten	0.43	0.94	0.59	0.52
Rule+BERT-2C	KWDLC	0.70	0.82	0.76	0.74
	Rakuten	0.44	0.91	0.59	0.55

The models trained on the in-domain datasets (D_{dm} and D_{gpt}) achieve better precision than the models trained on the out-domain dataset (D_{kwdlc}), but less perform with respect to the recall. Thus the ensemble of these datasets is expected to compensate each other. As for the F1-score and accuracy, in general, the in-domain datasets are better than the out-domain dataset. It indicates that the effectiveness of our proposed approach to construct the in-domain datasets automatically. Comparing D_{dm} and D_{gpt}, the latter performs better, indicating the adequate ability of ChatGPT for data augmentation. D_{dm+gpt} achieves better recall but comparable F1-score compared with the individual dataset D_{dm} or D_{gpt}. Finally, the F1-score of the models trained on D_{all} is better than or comparable to the other datasets. The intermediate fine-tuning (IFT) further boosts the F1-score and the accuracy. The best F1-score is 0.71 achieved by IFT-2C trained on D_{all}, which is 0.12 points higher than BERT-2C trained on D_{kwdlc}. It proves that our approach to combine the heterogeneous datasets, i.e., the manually annotated but out-domain dataset and weakly supervised but in-domain dataset, is effective for the opinion-ground classification task. In addition, the IFT is more appropriate than simple combination to utilize those heterogeneous datasets.

The findings on the comparison of the different models are enumerated as follows.

- The models that accept a pair of clauses as the input (*-2C) usually outperform the models that accept a single clause (*-1C). Although an opinion and ground often appear in one clause, it is better to separate two clauses when they are entered to the BERT model.
- The hybrid methods do not usually outperform the BERT models. Especially, the precision of Rule+IFT-2C is worse than Rule or IFT-2C. This is because Rule can classify only 23% of the test data, while IFT-2C performs well on this portion of the test data (its precision is 0.79). Thus the ensemble of Rule and IFT does not yield any improvement.

Table 9: Example of opinion-ground classification by IFT-2C trained from D_{all}.

	Gold	Pred.	Clause 1	Clause 2
#1	og	og	品質、梱包、全て問題無く大変満足しています。(I find no problem on everything including the quality and the packing, so I am very satisfied.)	メールでの問い合わせにも丁寧に対応して頂きました。(They also responded politely to my email inquiry.)
#2	og	non-og	良い所はやはりデザインと、(The strong points are good design as everyone thought,)	それからキャスターつきの所です。(and the wheels are attached.)
#3	non-og	og	初めて利用したお店だったので少し不安がありましたが、(Since it was my first time to buy from this shop, I was in anxiety, but)	美味しかったです。(it was delicious.)

- The accuracy of the rule-based method is relatively high. This may be caused by the imbalance of the classes in the test data. As shown in Table 6, the number of the non-opinion-ground clause pairs is almost twice than that of the opinion-ground clause pairs. Rule tends to classify a clause pair as non-opinion-ground since it judges the input as an opinion-ground clause pair only when the discourse marker is found. Thus the accuracy of Rule is estimated high. However, the F1-score of Rule is inferior to the BERT-based models.
- ChatGPT shows the tendency to classify a clause pair as positive and achieves the high recall but low precision. The F1-score is much less than the BERT-based models.

5.2.2 Domain Difference

A naive approach to solve the opinion-ground classification task is to use the existing manually labeled dataset, KWDLC. As discussed in 4.3.1, however, KWDLC is not fully appropriate for this task due to the disagreement of the task definition and the domain (Web document vs. customer review). To investigate the influence of such differences in the training and test data, an additional experiment is carried out. D_{kwdlc} is used as the training data, then the performance on Rakuten test data and KWDLC test data are compared. The precision (P), recall (R), F1-score (F), and accuracy (A) of Rule, BERT-2C, and Rule+BERT-2C are shown in Table 8. The performance of the BERT model on the out-domain test data (Rakuten) is obviously poorer than that on the in-domain test data (KWDLC). The performance of the rule-based method is rather comparable on the two test data, but the hybrid model also suffers from the gap of the domains. It indicates the necessity to use the in-domain dataset to precisely identify an opinion and ground in a review.

5.3 Case Study and Error Analysis

Table 9 shows examples of the opinion-ground classification. "Gold" stands for the gold label, while "Pred." means the class predicted by our best model IFT-2C. The example #1 is the true-positive. The model successfully classifies it as the opinion-ground clause pair although the opinion ("I am satisfied") and the ground ("I find no problem") are in one clause and there is no explicit discourse marker. The example #2 is the false-negative. The model fails to capture the opinion ("strong point") and the ground ("good design", "attached wheels"), since there is no explicit word or phrase to indicate the ground. The example #3 is the false-positive. Although there is no ground for the opinion, the model judges it as the opinion-ground clause pair.

We also carried out error analysis. Among 48 false-negative samples, 39(81%) samples show positive opinions. The example #2 in Table 9 is such an example; the user says that the design of the product is good. In the dataset, many opinion-ground clause pairs express user's positive opinion. The model might learn irrational relation between a positive opinion and existence of an opinion and ground.

As for false-positive errors, it was found that the discourse marker "ので''(*node*) was a major causes of errors. Among all 70 false-positive samples, 17(24%) samples include this discourse marker. Although "ので''(*node*) indicates causal relation, it is not always relation between an opinion and ground. For example, in the example #3 in Table 9, "ので''(*node*) shows a reason (it was my first time to buy from this shop) why a user was in anxiety, but it is not reason for the user's opinion (it was delicious).

6 CONCLUSION

This paper proposed the novel method to identify an opinion of a reviewer and its ground in a customer review. First, we newly defined the opinion-ground classification task that aimed to classify a pair of clauses whether it contained an opinion and its ground. To train the classifiers of this task, three heterogeneous datasets were constructed: (1) the part of KWDLC that consisted of pairs of clauses under the "cause/reason" relation as the positive samples, (2) the dataset constructed by checking the existence of the discourse markers, and (3) the augmented dataset including the clauses generated by ChatGPT. In addition, the rule-based method, BERT, and the hybrid of these two methods were emprically compared as the classification models. Results of the experiments showed that the use of not only the existing manually annotated out-domain dataset (KWDLC) but also the automatically constructed in-domain dataset could improve the F1-score of the opinion-ground classification task. The best F1-score, 0.71, was obtained when the BERT model was trained by the intermediate fine-tuning using three datatsets. It was 0.12 points higher than the model using only KWDLC.

In the future, the in-domain datasets should be enlarged so that the classifier can trained from a corpus including a wide variety of linguistic expressions that represent the opinion-ground relation. In addition, instead of just checking two causality discourse markers, we will investigate a more sophisticated rule-based method to extract the opinion-ground clause pairs from unlabeled reviews more precisely.

REFERENCES

Almagrabi, H., Malibari, A., and McNaught, J. (2018). Corpus analysis and annotation for helpful sentences in product reviews. *Computer and Information Science*, 11(2):76–87.

Cengiz, C., Sert, U., and Yuret, D. (2019). KU_ai at MEDIQA 2019: Domain-specific pre-training and transfer learning for medical NLI. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 427–436, Florence, Italy. Association for Computational Linguistics.

Dai, H., Liu, Z., Liao, W., Huang, X., Cao, Y., Wu, Z., Zhao, L., Xu, S., Liu, W., Liu, N., et al. (2023). AugGPT: Leveraging ChatGPT for text data augmentation. *arXiv preprint arXiv:2302.13007*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Diaz, G. O. and Ng, V. (2018). Modeling and prediction of online product review helpfulness: A survey. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 698–708, Melbourne, Australia. Association for Computational Linguistics.

Gamzu, I., Gonen, H., Kutiel, G., Levy, R., and Agichtein, E. (2021). Identifying helpful sentences in product reviews. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 678–691, Online. Association for Computational Linguistics.

Gilardi, F., Alizadeh, M., and Kubli, M. (2023). Chatgpt outperforms crowd-workers for text-annotation tasks. *arXiv preprint arXiv:2303.15056*.

Hu, M. and Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177.

Kim, N., Feng, S., Gunasekara, C., and Lastras, L. (2020). Implicit discourse relation classification: We need to talk about evaluation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5404–5414, Online. Association for Computational Linguistics.

Kim, S.-M., Pantel, P., Chklovski, T., and Pennacchiotti, M. (2006). Automatically assessing review helpfulness. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 423–430, Sydney, Australia. Association for Computational Linguistics.

Kishimoto, Y., Murawaki, Y., Kawahara, D., and Kurohashi, S. (2020). Japanese discourse relation analysis: Task definition, connective detection, and corpus annotation (in Japanese). *Journal of Natural Language Processing*, 27(4):899–931.

Liu, J., Cao, Y., Lin, C.-Y., Huang, Y., and Zhou, M. (2007). Low-quality product review detection in opinion summarization. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 334–342, Prague, Czech Republic. Association for Computational Linguistics.

Loshchilov, I. and Hutter, F. (2019). Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Mudambi, S. M. and Schuff, D. (2010). Research note: What makes a helpful online review? a study of customer reviews on amazon.com. *MIS quarterly*, pages 185–200.

Pan, Y. and Zhang, J. Q. (2011). Born unequal: a study of the helpfulness of user-generated product reviews. *Journal of retailing*, 87(4):598–612.

Poth, C., Pfeiffer, J., Rücklé, A., and Gurevych, I. (2021). What to pre-train on? Efficient intermediate task selection. In *Proceedings of the 2021 Conference on*

Empirical Methods in Natural Language Processing, pages 10585–10605, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A., and Webber, B. (2008). The Penn Discourse TreeBank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Prasad, R., Webber, B., and Lee, A. (2018). Discourse annotation in the PDTB: The next generation. In *Proceedings 14th Joint ACL - ISO Workshop on Interoperable Semantic Annotation*, pages 87–97, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Rakuten Institute of Technology (2023). Rakuten data release. https://rit.rakuten.com/data_release/. (last accessed in Sep. 2023).

Tsur, O. and Rappoport, A. (2009). Revrank: A fully unsupervised algorithm for selecting the most helpful book reviews. *Proceedings of the International AAAI Conference on Web and Social Media*, 3(1):154–161.

Yang, Y., Yan, Y., Qiu, M., and Bao, F. (2015). Semantic analysis and helpfulness prediction of text for online product reviews. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 38–44.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., and Le, Q. V. (2019). XLNet: Generalized autoregressive pre-training for language understanding. In *NIPS'19: Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 5753–5763.

Zhu, F. and Zhang, X. (2010). Impact of online consumer reviews on sales: The moderating role of product and consumer characteristics. *Journal of marketing*, 74(2):133–148.