修士論文

意見とその根拠を含む有用な商品レビューの分類

CHUANG Po-Min

主指導教員 白井 清昭

北陸先端科学技術大学院大学
先端科学技術研究科
（情報科学）

令和 6 年 3 月

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.

概要

ネット上の商品レビューは，購入意思を決定する消費者と商品の評判を知りたい企業の双方にとって，重要な役割を果たしている．しかし，すべてのレビューが有益な情報を提供するわけではない．大量のレビューから有用なレビューを発見するのは多くの時間を要するため，レビューの有用性を自動的に判定する研究が注目を集めている．一方，レビューの有用性は様々な観点から定義される．レビュアーの意見に対する理由や根拠は重要な特徴の一つである．例えば「デザインがかっこいいので気に入った」というように，意見だけでなくその根拠も述べているレビューは，消費者と企業にとって有益である．しかし，このような観点からレビューの有用性を判定する研究は，これまで行われていなかった．

本研究は，多くの製品レビューから有益なレビューを選別するために，レビューにユーザの意見とその根拠が含まれているかどうかを判定することを目的とする．この目的を達成するために，以下の 2 つのタスクを実行する．1 つ目は「根拠関係分類タスク」である．これは，入力として与えられたレビュー内の文や節の組がユーザの意見とその根拠を含んでいるかどうかを分類するタスクである．2 つ目は「商品言及分類タスク」である．根拠を含むと判定した節の組に対し，それが商品に対する意見の根拠であるかを判定するタスクである．商品に関係のない記述に対する根拠が書かれていても有用なレビューとは言えないため，2 つ目のタスクが必要である．

根拠関係分類タスクの提案手法は以下の通りである．まず，与えられたレビューから依存関係がある節のペアを抽出する．次に，それぞれの節のペアが「根拠関係-節ペア」（意見とその根拠が含まれる）であるか「非根拠関係-節ペア」であるかを分類する．根拠関係分類タスクを解く 4 種類の分類器を学習する．(1) ルールベースの方法．入力とした節の組のいずれかに談話標識「から」もしくは「ので」が出現していれば，その節ペアを根拠関係-節ペアと判定し，それ以外は非根拠関係-節ペアと判定する．(2)Bidirectional Encoder Representations from Transformers(BERT)．訓練データを使用してファインチューニングする．(3)BERT ベースの Intermediate Fine-Tuning(IFT)．まず，比較的大きな関連タスクのデータセットを使用して BERT モデルをファインチューニングし，その後，対象タスクのラベル付きデータセットを使用して再びファインチューニングする．(4) ルールベースと BERT モデルを組み合わせたハイブリッド手法．最初に，ルールベースの手法を用いて入力の節ペアが根拠関係-節ペアか否かを判定する．非根拠関係-節ペアと判定されたとき，BERT による根拠関係分類器を用い，その判定結果を最終の判定結果とする．

根拠関係分類器を学習するために 3 つのデータセットを用いる．1 つ目は既存の談話関係解析データセットである京都大学ウェブ文書リードコーパス (KWDLC) である．KWDLC において,「原因/理由」タグでラベル付けされた文の組を正例 (根拠関係-節ペア) として抽出する．また,「原因/理由」以外の談話関係タグが付与された文の組を負例 (非根拠関係-節ペア) として抽出する．KWDLC は文間の談話関係が人手で付与されたデータセットであるが，本研究が対象とする根拠関係

分類タスクに必ずしも適しているわけではない．また，KWDLC と商品レビューはドメインが異なる．KWDLC では様々なウェブ文書に対して注釈付けされているが，根拠関係分類タスクのドメインは商品レビューである．2 つ目は談話標識によるデータセットである．ラベルなしのレビューから節の組を抽出し，もし節に「から」または「ので」の談話標識が含まれていれば，意見とその根拠が含まれた正例 (根拠関係-節ペア) として抽出する．また，これらの談話標識を含まない節の組を負例 (非根拠関係-節ペア) として抽出する．KWDLC とは異なり，談話標識を手がかりに構築したデータセットはイン・ドメインのデータである．しかし，正例と負例は自動抽出されているため，誤りを含む可能性があることに注意する必要がある．3 つ目は ChatGPT によるデータセットである．意見とその根拠の両方を含む節をシード節として用意する．それぞれのシード節に対し，ChatGPT により，それと類似しかつ意見とその根拠を含む新しい節を 40 個生成する．ChatGPT を用いて生成した節は，同じジャンルのレビューからランダムに選ばれた別の節と結びつけられ，正例の根拠関係-節ペアが作られる．また，談話標識によるデータセットと同じ方法で負例を作成する．なお，IFT では，最初の学習段階でアウト・ドメインのデータセットである KWDLC を用い，第 2 の学習段階でイン・ドメインの 2 つのデータセットを用いる．

商品言及分類タスクに対する提案手法は，キーワードマッチングに基づく単純な手法である．まず，商品カテゴリに関連するキーワード集合をあらかじめ構築する．ここでの商品カテゴリとは,「食品」「本」「家電」など，商品の種類を分類したものである．キーワードとは，そのカテゴリの製品に関するレビューで頻繁に使用される重要な単語を指す．本タスクの分類器は，入力された節ペアに商品カテゴリのキーワード集合のいずれかのキーワードが出現するときには，その節ペアの意見は商品に関連があるとみなす．逆に，節ペアに商品カテゴリのキーワードがひとつも出現していないときは，その節ペアの意見は商品に関連がないとみなす．

商品カテゴリに関するキーワードを抽出するために 5 つの手法を用いる．(1)TF-IDF．TF-IDF スコアの高い単語が抽出される．(2)TF-ENT．TF-ENT スコアの高い単語が抽出される．ここでの ENT とは，単語の商品カテゴリに関する顕現性（単語が 1 つの商品カテゴリのみで使用される傾向がどれだけ強いか）を評価するスコアであり，単語が商品カテゴリのレビューに現れる確率の確率分布のエントロピーによって算出される．(3)YAKE!．既存の教師なしキーワード抽出手法 YAKE! によって抽出された単語．(4)TF-IDF+YAKE!．TF-IDF と YAKE!の組み合わせによって抽出された単語．(5)TF-ENT+YAKE!．TF-ENT と YAKE!の組み合わせ手法である．各手法ごとに候補単語のスコアを計算し，その上位 200 件の単語をキーワード集合として取得する．

提案手法を評価するためにいくつかの実験を行った．根拠関係分類タスクのテストデータは，506 個の節ペアに正解ラベルを人手で付与して作成した．また，商品言及分類タスクのテストデータは，上記のデータセットに含まれる 186 の根拠

関係-節ペアに対して，正解ラベルを人手で付与して作成した．

根拠関係分類タスクの実験結果から，自動的に構築した 2 つのデータセットを用いることで根拠関係分類タスクの性能が大きく改善することを確認した．最高の F1 スコアは IFT が達成した 0.71 であり，既存のデータセット KWDLC で学習した BERT モデルより 0.09 ポイント高かった．これにより，異なる種類のデータセットを組み合わせる我々のアプローチが意見に対する根拠関係の分類に有効であることが示された．さらに，これらの異なる種類のデータセットを利用する際には，データセットの単純な結合よりも Intermediate Fine-Tuning の方が適していることがわかった．一方，ルールベースの手法と BERT モデルを組み合わせたハイブリッド手法は，BERT モデルを上回らなかった．これは，ルールベースの手法は評価データの 23%しか分類できないが，BERT ベースの IFT はこれらのデータに対する分類の精度が高い (0.79) ためである．

商品言及分類タスクについては，TF-IDF と TF-ENT を比較すると，TF-IDF の方が再現率と F1 スコアが高かった．また，YAKE!は TF-IDF と TF-ENT よりも高い性能を示した．特に再現率の差が大きかった．最後に，TF-IDF と YAKE!を組み合わせてキーワードを抽出する手法が最も有効であることがわかった．この手法の再現率は 0.91，F1 スコアは 0.89 であり，5 つの手法の中で最も高く，精度も他の手法と比べて大きな差はなかった．また，2 つのキーワード抽出手法を組み合わせる TF-IDF+YAKE!は，ひとつの抽出手法だけを用いる場合と比べて，より多様なキーワードを獲得していることを確認した．このことが，TF-IDF と YAKE!を組み合わせることによって商品言及分類タスクの性能が向上する主な原因であると考えられる．

目 次

図 目 次

表 目 次

第1章　はじめに

1.1　背景

オンラインショッピングのプラットフォームが増加する中で，ユーザレビューはユーザ (顧客) と企業にとって貴重な情報源となっている.[1]ユーザが商品の購入を検討する際，あるいは企業が自社の製品を改良する際，他のユーザによって書かれた商品のレビューが参考にされる．しかし，ユーザや企業にとって，膨大な数のユーザレビューから必要な情報を取得することは難しい．そのため，レビューから自動的に有用な情報を抽出するための研究や，ユーザや企業が必要な情報を見つけるのをサポートする研究が注目されている．例えば，レビューを肯定的なレビューと否定的なレビューに自動的に分類したり，複数のユーザレビューを要約したりする研究が行われている．

レビューの中には有用なレビューとそうでないレビューがある．ここでの有用ではないレビューとは，製品に関する情報やユーザの意見を全く含まず，ユーザや企業にとって役に立たないことが明らかなレビューとする．レビューの中には有用ではないレビューも数多く存在することが，ユーザレビューからの情報獲得の障害のひとつとなっている．そのため，レビューの有用性を判定する研究，すなわちレビューが有用か否かを判定する研究も近年注目を集めている [9, 12].

レビューの有用性は，様々な視点から定義される．例えば，長いレビューは短いレビューよりも有益な情報を含むことが多いことから,「レビューの長さ」は有用性の観点のひとつである．また,「商品に関する詳細な説明があること」も有用性の観点のひとつと言える．レビューの有用性に関する様々な観点のうち，本研究では,「意見の理由や根拠を含むこと」に着目する．例えば,「デザインがかっこいいので気に入っています.」といったレビューは，ユーザがその商品を好んでいる理由を説明しており，商品の利点を明らかにしているため，有用な情報を含んでいると言える．逆に,「素晴しいです.」「残念です.」といった単に意見を述べているレビューは，その理由や根拠を示していないため，あまり有用であるとは言えない.

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

1.2　目的

本研究は，多くの商品レビューの中から消費者や企業にとって有用なレビューを選別するために，レビューがユーザの意見とそれに対する根拠を含んでいるかを判定する手法を提案する．既に述べたようにレビューの有用性には様々な観点があるが，その中でも意見とその根拠を含むレビューに焦点を当て，これを自動的に判別することを研究の目的とする．この目的を達成するために，以下の 2 つのタスクを実行する．1 つ目のタスクは，レビューに出現する節の組に対し，それが意見に対する根拠を含むかどうかを判定するタスクである．これは，一つの節が意見を表し，もう一つがその根拠を示す，いわゆる 2 つの節間の談話関係解析の一種として定式化される．2 つ目のタスクは，根拠を含むと判定した節の組に対し，それが商品に対する意見の根拠になっているかを判定するタスクである．たとえレビューに意見とそれに対する根拠が書かれていても，それが商品と関係のない意見であれば，有用なレビューとは言えない．例えば,「箱がつぶれていたのでがっかりした.」という文は意見と根拠を含むが，商品の質を評価した意見ではない．本研究では商品に関係のない意見と根拠は有用ではないとし，これを自動的に判別する手法を探究する．

1.3　提案手法の概要

提案手法の概要を図 1.1 に示す．商品に関するユーザレビュー (商品レビュー) を入力とし，それが有用なレビューか否かを判定する．ここでの有用なレビューとは，商品に関する意見とそれに対する根拠を含むレビューと定義する．提案手法は 1.2 節で述べた 2 つのタスクを順に実行する．図 1.1 の「根拠関係判定」では，節の組がユーザの意見とそれに対する根拠を含むかを判定する．レビュー中における全ての節の組に対し，意見と根拠を含まないと判定したときは，その商品レビューは有用ではないと判定する．このモジュールの詳細は 3 章で述べる．次に，図 1.1 の「意見の対象判定」では，レビュー中の意見が商品に対する意見であるか否かを判定する．商品に対する意見と判定したときのみ，入力された商品レビューは有用であると判定する．このモジュールの詳細は 4 章で述べる．

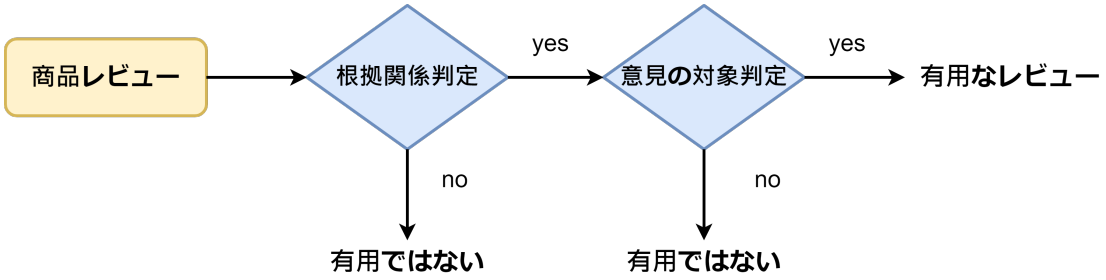

図 1.1: 提案手法の概要

1.4　本論文の構成

本論文の構成は以下の通りである．第 2 章では，本論文の関連研究を紹介する．第 3 章では，レビューに出現する節の組が意見と根拠を含むか否かを分類する手法を詳述する．第 4 章では，意見が商品に関するものであるかを分類する手法の詳細を述べる．第 5 章では，提案手法の評価実験と結果について報告する．最後に，第 6 章では，本論文のまとめと今後の課題について述べる．

第2章　関連研究

本章では，本論文に関連する研究について述べる．2.1 節では，商品レビューの有用性に関する研究を紹介する．2.2 節では，本研究の提案手法は節間の関係を分類する談話関係解析の一種とみなすことができるため，談話関係解析に関する先行研究を紹介する．2.3 節では，本研究ではレビューから重要なキーフレーズを抽出する処理を必要とするため，既存の教師なしキーフレーズ抽出手法を紹介する．2.4 節では，本研究で利用する言語モデルである BERT(Bidirectional Encoder Representations from Transformers) を紹介する．最後に，2.5 節では，本研究と先行研究の違いについて論じる．

2.1　商品レビューの有用性に関する研究

2.1.1　レビューの有用性の自動判定

オンラインの商品レビューは，消費者の購買決定に重要な影響を与える [35] だけでなく，商品を改善するための有用な情報も提供するが，全てのレビューが顧客や企業にとって有用であるわけではない．そのため，レビューが有用であるか否かを判定する多くの研究が行われている．

Diaz らと Ng らは，有用な商品レビューの予測に関する関連研究について調査している [9]．彼らは，有用性予測システムで使用されている特徴を整理し，これらは主にコンテンツ特徴とコンテキスト特徴に分けることができるとしている．コンテンツ特徴として以下の 6 つの特徴を挙げている．(1) レビュー文の長さ．長いレビューはより多くの情報を持っている．(2) 読みやすさ．レビューが読みやすければ，より多くのユーザに役に立つと思われる．(3) 重要単語．情報の重要性を示すキーワードを含むレビューは有用であるという考えに基づいている．(4) 単語カテゴリ．有用なレビューに出現する単語のカテゴリ (グループ) である．(5) コンテンツの多様性．レビューの内容が特定の参照テキストとどの程度異なっているかを測定する．(6) その他．星の数による評価やレビュー文の主観性など．一方，コンテキスト特徴として以下の 2 つの特徴を挙げている．(1) レビュアー特徴．良いレビュアーは繰り返し有用なレビューを書くという考えに基づき，レビュアーの履歴情報からレビューの有用性を予測する．(2)user-reviewer idiosyncrasy 特徴．(レビューを読む) ユーザとレビューアーの類似度を測る．ユーザ自身と嗜好が似ているレビューアーが書いたレビューは有用である．

Kim らは，Amazon.com の評価レビューの有用性を判定するために，構造，語彙，統語，意味，メタデータの 5 つのクラスの特徴量を用いて，Support Vector Machine(SVM) を学習する手法を提案している [14]．また，学習に使用した特徴の中で特に重要なものは何かを分析している．その結果，最も有効的な特徴は，レビューの長さ，単語 uni-gram，製品評価であることがわかった．

Mudambi と Schuff は，顧客レビューの有用性の予測に Tobit 回帰手法を適用し，レビューの極端さ，レビューの深さ，製品タイプがレビューの有用性を判断する際の重要な特徴であることを示している [20]．極端な星の評価（非常に高いまたは低い）を受けたレビューは，そうでないレビューと比べて，有用性の度合いが低いと報告している．また，レビューの深さ (レビュアーのコメントの広範さ) がレビューの有用性と正の相関関係があるとしている．

Pan と Zhang は，ランダム切片を持つミックス効果ロジスティックモデルを用いて，有用性に影響を与える様々な要因を分析している [22]．彼らは，他者によるレビューの評価ならびにレビューの長さと有用性との間に正の相関があることを示している．さらに，彼らはレビューの内容を検討した結果，レビューの独創性 (レビューアーが自身の新しい考えを書いていること) と有用性の間にも正の相関関係があることを明らかにしている．

Yang らは，レビューテキストのみを用いて製品レビューの有用性スコアを予測する回帰モデルを学習している [33]．彼らは解釈可能な 2 つの意味素性を採用し，人間の採点による有用性スコアを正解データとして回帰モデルを学習している．その結果，提案された意味素性で訓練されたモデルは，異なる製品カテゴリのレビューの有用性判定にも適用できることを示した．

Tsur と Rappoport は，本に関するレビューの有用性を判定する教師なしアルゴリズム RevRank を実装している [31]．Virtual Core Review (VC) は，レビューの中で特に有用と考えられるレビューの集合である．評価対象とする本のレビューから重要な単語を抽出し，これを元に VC となるレビューを選別する．その後，VC との類似度に基づいてレビューの有用性スコアを算出し，ランキングする．

2.1.2 有用性投票のバイアス

前項で述べた研究の多くは，EC サイト上でレビューに対して他のユーザが与える有用性の投票結果を正解として，有用性を判定するモデルを学習している．一方，有用性の投票はバイアスがあるため，有用性投票の結果が常に良い有用性の指標であるとは言えないことが指摘されている [16, 31, 33]．例えば，Liu らは 3 種類の投票バイアスを明らかにしている [16]．Imbalance vote bias は，ユーザが否定的な意見よりも肯定的な意見を評価する傾向である．Winner circle bias は，多くの投票を受けたレビューがより多くの注目を集め，速いペースでさらに多くの投票がもらえる傾向である．Early bird bias は，レビューが早く投稿されればされるほど，それに対する投票が増加する傾向である．

ユーザによる有用性投票の結果は必ずしも真の有用性を表しているわけではないため，レビューが有用かどうかを人手によってアノテーションしたレビューのデータセットが構築され，有用なレビューを予測するために使されている．Almagrabi らは，Amazon.com と CNet.com の 5 つの電子製品について書かれたユーザレビューに，有用か否かの 2 値のラベルをアノテーションした [2]．一方，Liu らは SPEC と呼ばれるガイドラインを提案している [16]．レビューの品質を表す 4 つのカテゴリを定義し，それぞれの観点からレビューの有用性をアノテーションする．彼らは Amazon.com のレビューを SPEC にしたがってアノテーションし，データセットを構築した．これを訓練データとし，レビューが有用であるか否かを予測する二値分類器を SVM により学習した．

2.1.3 代表的な有用文の抽出

個々のレビューの有用性を判定するのではなく，レビュー集合から，有用かつ代表的な文を抽出することで，ユーザに有用な情報を提供する研究も行われている．本項では Gamzu らによる研究 [10] を紹介する．その概要を図 2.1 に示す．まず，Amazon.com における 123 の製品に関する 22,000 のレビュー文に対し，有用性のスコアをラベル付けしたデータセットを構築する．そして，このデータセットを用いて，文の有用性スコアを予測する BERT[8] をファインチューニングする．次に，Amazon AWS Comprehend[1]によって各文の極性を識別し，レビュー文を肯定的な文の集合と否定的な文の集合に分割する．そして，文の代表性を表すサポートスコアを計算する．肯定的文集合と否定的文集合のそれぞれにおいて，他の多くの文と類似している文に対し，より高いサポートスコアを与える．最後に，有用と判定された文の中から，サポートスコアが最大となる文を肯定的文集合と否定的文集合から 1 つずつ選び，これを代表的な有用文とする．上記の有用文の抽出タスクは一種の複数文書要約であると言える．

2.2 談話関係解析

談話関係解析は，文章における 2 つの文の間の関係を分析するタスクである．自然言語処理の基盤的な解析の一つであり，英語，日本語など様々な言語で研究されている．談話関係の情報がアノテーションされた代表的な英語のコーパスは Penn Discourse TreeBank (PDTB)[25, 26] と RST Discourse Treebank (RST-DT)[5] である．日本語のコーパスとしては京都大学ウェブ文書リードコーパス（KWDLC）がある．本研究は，レビューにおける 2 つの節の間に意見とその根拠という関係が存在するかを判定することを目的としており，2 つの文や節の関係を識別する談話構造解析との関連が深い．以下，談話関係に関連する先行研究を紹介する．

[1]https://aws.amazon.com/comprehend/

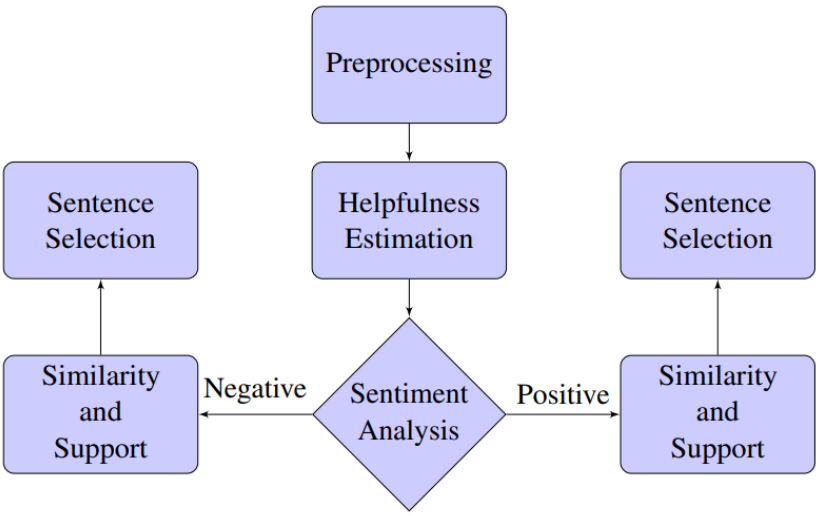

図 2.1: 代表的な有用文を抽出する手法の概要 [10]

2.2.1 英語の談話構造コーパス

Prasad らは，2,159 件の英語の新聞記事を対象に，2 つの談話単位間に談話関係タグを付与したコーパスである Penn Discourse TreeBank(PDTB) を構築した [25, 26]．PDTB における談話関係は，明示的な談話関係と暗黙的な談話関係に分類される．明示的な談話関係とは，談話接続詞（but，however など）によって示される談話関係であり，暗黙的な談話関係とは，談話接続詞がなくても 2 つの引数の間に成り立つ談話関係である．

PDTB 3.0[26] の談話関係タグセットは，図 2.2 に示すとおり，3 階層，30 種類のタグから構成されている．これらの談話関係タグのうち，本研究と最も関連の深い談話タグは，“CONTINGENCY.Cause” である．これは因果関係，すなわち，ある節が別の節が表す出来事が発生する原因となっているといった関係を表す．

PDTB を使用した談話構造解析の研究は数多く行われている．しかし，Kim らは，前処理と評価プロトコルに一貫性がないため，PDTB を用いた実験によって複数の手法を公平に比較できない可能性があると主張している [13]．さらに，2 つのスパンや引数の間に明示的な談話標識が存在しない，非明示的な談話関係分類の手法を公正に比較するために，標準的なラベルセットと，談話構造付きコーパスを元の文書のセクションで分割する交差検証のプロトコルを提案した．また，BERT[8] と XLNet[34] を用いて，非明示的な談話関係を分類する 2 つの強力なベースラインを開発した．

RST-DT は，385 件の英語の新聞記事に対して，隣接する談話単位間に談話関係タグを付与したコーパスである [5]．また，PDTB と RST-DT はともにウォール・ストリート・ジャーナルの記事に対して談話関係を付与したコーパスである

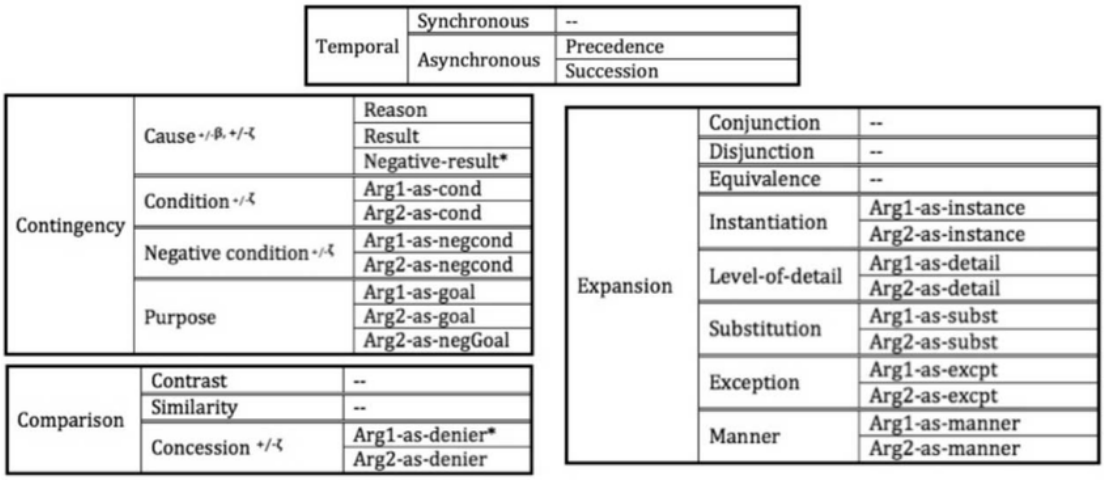

図 2.2: PDTB 3.0 における談話関係タグセット [13]

が，PDTB と RST 両者の違いは，RST-DT では談話単位を階層的に構造化して文書全体の談話関係を 1 つの木構造で表現している点にある．

2.2.2 日本語の談話構造コーパス

京都大学ウェブ文書リードコーパス (Kyoto University Web Document Leads Corpus; KWDLC)[36] は，ウェブページの冒頭 3 文に様々な言語情報を人手で付与したテキストコーパスであり，付与された言語情報の中には談話関係の情報も含まれる．PDTB 2.0[25] を参考に，2 階層 7 種類のタグから構成される談話関係タグセットが用いられている．KWDLC における談話関係タグセットを図 2.3 に示す．KWDLC には 2 つのデータセットが含まれている．1 つは「専門家データセット」であり，専門家がアノーテションした小規模・高品質のデータセットである．3 人の専門家によってアノテーションされた 2,320 の節の組からなる．もう 1 つは「クラウドソーシングデータセット」で，40,467 の節の組からなる．10 人のクラウドワーカーが節の組に対して談話関係のタグを割り当てる．最終的な談話関係は，10 人によって割り当てられたタグの多数決により決定される．岸本らは，KWDLC を訓練データとテストデータに分割し，BERT モデル，機械学習ツール “opal”，談話標識を手がかりとした解析器，これらを組み合わせた解析器を実装し，評価した．

2.2.3 対照学習を利用した談話関係解析

清丸と黒橋は，談話関係解析の精度を向上させるために，複数文からなる入力テキストに対して各文の埋め込み表現を計算するモデルを考え，それを対照学習 (contrastive learning) の枠組みで学習する手法を提案している [15]．彼らの手法の

上位タイプ	下位タイプ	例
順接系	原因・理由	【ボタンを押したので】【お湯が出た.】
	目的	【試験に受かるために】【必死に勉強した.】
	条件	【ボタンを押せば】【お湯が出る.】
	根拠	【ここにカバンがあるから】【まだ社内にいるだろう.】
逆接系	対比	【大阪は雨だが,】【東京は晴れだ.】
	逆接	【あのレストランはおいしいが】【値段は高い.】
関係なしまたは弱い関係		【家に着いてから】【雨が降ってきた.】

図 2.3: KWDLC の談話関係タグセット [36]

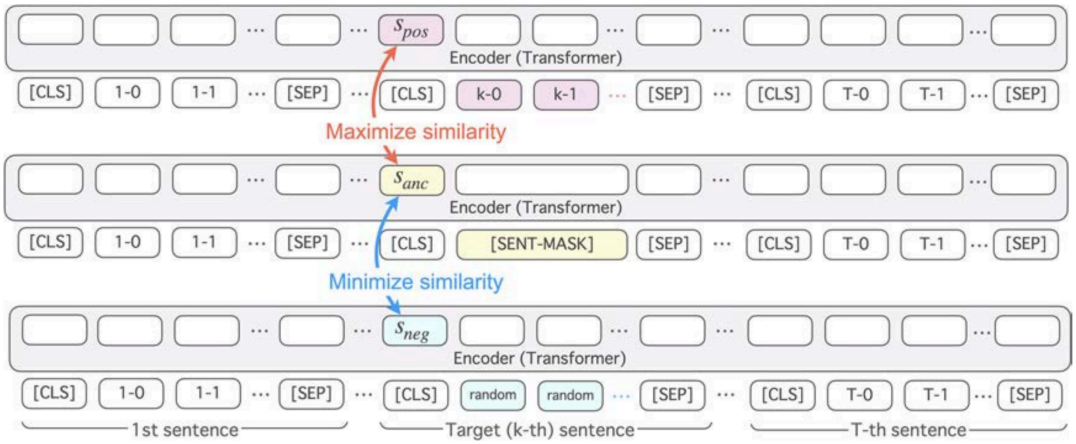

図 2.4: 対照学習の概要 [15]

概要を図 2.4 に示す．最初に，入力テキストからランダムに一文を選び，図 2.4 では $k(1 \leq k \leq T)$ 番目の文をターゲット文として選択する．このターゲット文の埋め込み表現を s_{pos} とする．次に，ターゲット文を特殊トークンである [SENT-MASK] に置き換えた入力テキストを作成する．この [SENT-MASK] の文表現を s_{anc} と記す．最後に，ターゲット文を無作為な文で置き換えた入力テキストを用意し，その無作為な文の表現を s_{neg} と記す．

そして，s_{pos} と s_{anc} の類似度を最大化し，同時に s_{neg} と s_{anc} の類似度を最小化する対照学習が行われる．損失関数は以下の式 (2.1) のように定義する．

$$L = -\log \frac{(\exp \langle \mathbf{s}_{\text{pos}}, \mathbf{s}_{\text{anc}} \rangle)}{\sum_{\mathbf{s} \in \mathcal{S}} \exp (\langle \mathbf{s}, \mathbf{s}_{\text{anc}} \rangle)} \tag{2.1}$$

ここで $\langle \cdot, \cdot \rangle$ はベクトルの内積，$S = \{s_{\text{pos}}, s_1^{\text{neg}}, \ldots, s_N^{\text{neg}}\}$ である．N はターゲット文を無作為に入れ換えて作成する文の総数である．

この手法は，文脈に依存した文の埋め込み表現を対照学習を通じて学習するものである．その有効性を評価するために，PDTB と KWDLC を用いて，本手法を談話

関係を分類するタスクに適用する実験を行った．実験では，BERT[8]，XLNet[34]，RoBERTa[17] が文埋め込みを生成するベースモデルとして使用された．

2.2.4　汎用言語モデルに基づく日本語談話関係解析

植田らは，Kyoto-Waseda Japanese Analyzer(KWJA) と呼ばれる統合的日本語解析器を構築した [32]．自然言語解析には形態素解析，構文解析など様々なタスクがあり，一般には個々のタスク毎にツールが開発されている．しかし，複数のタスクを同時に実行する場合，複数のツールの結果を統合する処理が必要になり，そのコストが高いという問題がある．そのため，植田らは複数のタスクを同時に行うツールを開発している．また，事前学習された汎用言語モデルを積極的に活用し，人手による自然言語解析ツールの開発コストを抑えるため，RoBERTa[17] を利用して統合的日本語解析器を実現している．

KWJA の解析フローを図 2.5 に示す．具体的には，まず「入力誤り訂正」が行われ，次に「分かち書き」と「単語正規化」が行われ，最後に「形態素解析」「固有表現認識」「言語素性付与」「構文解析」「述語項構造解析」「橋渡し照応解析」「共参照解析」「談話関係解析」が行われる．

KWJA で行われる解析の中には談話関係解析も含まれる．7 種類の談話関係クラスを定義し，入力された節の組に対し，その談話関係を分類する．談話関係解析器を実装するため，RoBERTa によって得られる節のベクトル表現を結合したものを入力とし，談話関係を分類する 3 層の feed-forward neural network を学習する．モデルの学習には KWDLC が用いられている．実験の結果，談話関係分類の F 値のマイクロ平均と分散は 55.3 ± 3.6 となり，KWJA の有効性を示した．

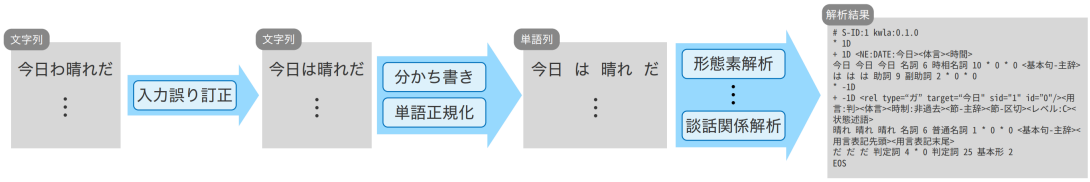

図 2.5: KWJA の解析フロー [32]

2.3　教師なしのキーフレーズ抽出手法

キーフレーズ抽出とは，文章からその主題を良く表現している句を抽出する技術である．これには教師ありと教師なしの手法がある．教師ありキーフレーズ抽出とは，正解のキーフレーズが付与された文書集合を訓練データとしてキーフレーズを抽出するモデルを学習する手法であり，教師なしフレーズ抽出とは，そのよ

うな正解情報が付与されていないテキストからキーフレーズを抽出する手法である．本節では，教師なしキーフレーズ抽出手法をいくつか紹介する．

2.3.1 グラフベースのキーフレーズ抽出手法

Mihalcea と Tarau は，ウェブ検索エンジンにおいてウェブページのランクを決定するアルゴリズム PageRank[21] に基づき，TextRank と呼ばれるキーフレーズ抽出手法を提案した [19]．TextRank では，文章に含まれる単語をノードとし，単語と単語が指定した窓幅内で共起した場合にエッジを張り，無向グラフを形成する．また，グラフのサイズが過度に大きくならないようにするために，品詞による単語 (ノード) の絞り込みを行う．図 2.6 は文書とそれから構築したグラフの例である．次に，無向グラフのノードへの入力エッジ数=ノードからの出力エッジ数として PageRank アルゴリズムを適用し，ランクの高い順に単語を抽出する．最後に，抽出した単語が元のテキスト内で連続して出現するとき，それらを連結して単一のキーフレーズにまとめる．

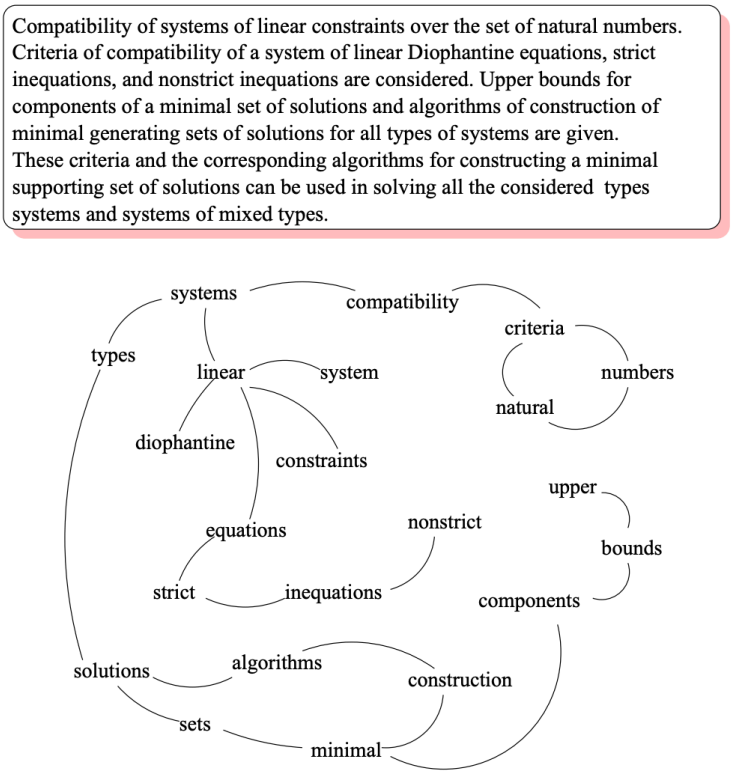

図 2.6: TextRank によって形成された無向グラフの例 [19]

2.3.2 統計ベースのキーフレーズ抽出手法

YAKE!はCamposらによって提案された教師なしのキーフレーズ抽出手法である[4]．以下の6つのステップでキーフレーズを抽出する．

1. テキスト前処理
2. 特徴量抽出
3. 単語の重要度スコアの計算
4. 重要なキーフレーズの抽出
5. 編集距離による重複キーフレーズの除外
6. ランク付けされた重要キーフレーズの出力

「テキスト前処理」では，スペースまたは特殊文字（改行，角かっこ，コンマ，ピリオドなど）を区切り文字として文章を単語に分割する．「特徴量抽出」では，w_{case}，w_{pos}，w_{freq}，w_{rel}，$w_{difSent}$といった5つの単語の特徴量を計算する．

w_{case}は単語の格特徴であり，式(2.2)のように計算される．$TF(U(t))$は大文字で始まる候補語tの出現回数，$TF(A(t))$は候補語tが頭字語（単語のすべての文字が大文字のもの）として出現する回数，$TF(t)$はtの出現頻度である．したがって，候補語が大文字で出現する頻度が高いほど，その候補語は重要であるとみなされる．

$$w_{\text{case}} = \frac{\max(TF(U(t)), TF(A(t)))}{\ln(TF(t))} \tag{2.2}$$

w_{pos}は単語の出現位置を表す特徴量である．キーフレーズは文書の最初によく現れるという仮定に基づき，文書の最初に出現する単語をより重視する．w_{freq}は単語の出現頻度である．より頻繁に出現する単語を高く評価する．w_{rel}は単語の文脈との関連性を表す特徴量である．対象単語の周辺に出現する単語の頻度を算出し，対象単語の周辺に出現する単語の異り数が多いほど，対象単語のw_{rel}を低く算出する．$w_{difSent}$は単語が異なる文に出現する頻度を定量化したものである．単語の出現頻度と似ているが，異なる文で頻繁に出現する単語をより高く評価する．

「単語の重要度スコアの計算」では，式(2.3)によって単語w_iの重要度スコア$S(w_i)$を算出する．このスコアは前述の5つの特徴量から計算され，小さいほど重要な単語であることを表す．

$$S(w_i) = \frac{w_{\text{rel}} \times w_{\text{pos}}}{w_{\text{case}} + \frac{w_{\text{freq}}}{w_{\text{rel}}} + \frac{w_{\text{difSent}}}{w_{\text{rel}}}} \tag{2.3}$$

「重要なキーフレーズの抽出」では，最終的なスコア単語n-gramをキーフレーズの候補kwとし，その重要度のスコア$S(kw)$を式(2.4)によって算出する．

$$S\left(kw\right) = \frac{\prod_{w_i \in \ kw} S\left(w_i\right)}{\mathrm{TF}\left(kw\right) * \left(1 + \sum_{w_i \in kw} S\left(w_i\right)\right)} \tag{2.4}$$

分子は，単語 n-gram 内の各単語のスコア $S(w_i)$ を乗算したものである．一方分母は，単語 n-gram の文中での出現頻度（term-frequency, TF）と $S(w_i)$ のスコアの合計を乗算したものである．この手法により，長い単語 n-gram に無条件に低いスコアを与えることを防ぎつつ，重要な単語を含む単語 n-gram に対して低いスコアを割り当てる．$S(kw)$ が低いほど kw が重要なキーフレーズであることを表す．よって，$S(kw)$ が低いいくつかのキーフレーズを重要なキーフレーズとして抽出する．

「編集距離による重複キーフレーズの除外」では，レーベンシュタイン距離を用いて，抽出された重要キーフレーズ間の距離を計算し，距離が小さくほぼ同じと考えられるキーフレーズを除外する．

最後に,「ランク付けされた重要キーフレーズの出力」では，残されたキーフレーズを重要度スコアによってランキングし，ランク付けされた重要キーフレーズのリストを得る．

2.3.3 文埋め込みに基づくキーフレーズ抽出手法

Peters らは，文埋め込みモデル SIF[3] と事前学習済みモデル ELMo[23] を組み合わせて，短い文書からでもキーフレーズを抽出できる SIFRank と呼ばれる手法を提案している [29]．図 2.7 に SIFRank のフレームワークを示す．SIFRank は以下の 5 つのステップから構成される．

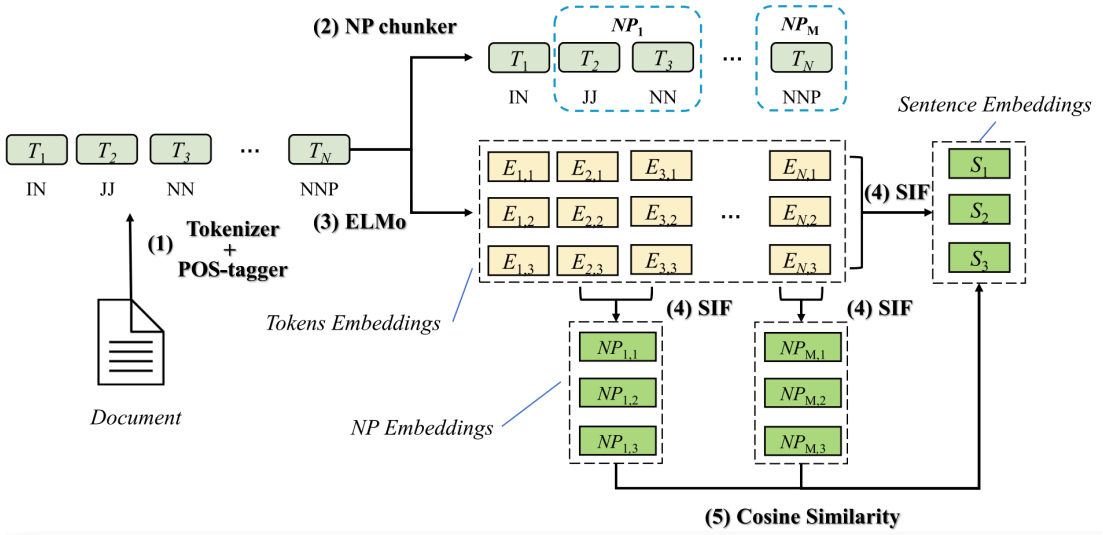

図 2.7: SIFRank のフレームワーク [29]

1. 文書をトークン (単語) に分割し，品詞タグ付けを行い，品詞が付与されたトークン列を得る．

2. 正規表現によるパターンマッチで実装された NP-chunker を用いて，トークン列から名詞句 (Noun Phrases; NPs) を抽出する．抽出された NPs がキーフレーズ候補となる．

3. トークンの列を事前学習済みモデルに入力し，各トークンの埋め込み表現を得る．

4. トークンの埋め込みから文埋め込みを得るモデル SIF を用いて，名詞句の埋め込み (NP 埋め込み) を得る．同様に，元の文書の埋め込み表現も得る．両者は同じ次元数のベクトルである．

5. NP 埋め込みと文書埋め込み間のコサイン類似度を計算する．これをキーフレーズ候補と文書のトピックの類似度とみなす．類似度の大きい上位 N 個のキーフレーズを最終的な重要キーフレーズとする．

SIFRank による NP と文書の類似度は式 (2.5) のように定式化される．

$$\mathrm{SIFRank}(v_{\mathrm{NP}_i}, v_d) = \mathrm{Sim}(v_{\mathrm{NP}_i}, v_d) = \cos(v_{\mathrm{NP}_i}, v_d) = \frac{v_{\mathrm{NP}_i} \cdot v_d}{\|v_{\mathrm{NP}_i}\| \cdot \|v_d\|} \tag{2.5}$$

ここで，d は文書，v_d は文書の埋め込み，NP は名詞句，v_{NP_i} はその埋め込み (NP 埋め込み) である．SIFRank は 0 から 1 の間の値を取り，1 に近いほど候補のキーフレーズは文書のトピックとの関連性が高いことを表す．逆に，値が 0 に近いほど，そのフレーズはトピックと無関係である．

SIFRank は Bag-of-Words 手法の一種であり，文書の先頭に重要なキーフレーズが現われやすいといった位置の情報が考慮されていない．しかし，特に複数の段落がある長い文書の場合には，位置の情報も考慮すべきである．Peters らは，長い文書に対応するために，SIFRank+も提案した．式 (2.6) は名詞句に対する位置のバイアス (重み) である．p_1 は NP_i が文書内で最初に出現した位置である．μ では，ハイパーパラメータであり，位置のバイアスを調整する．

$$p(NP_i) = \frac{1}{p_1 + \mu} \tag{2.6}$$

さらに，softmax 関数を用いて位置バイアスを 0〜1 の間の値に変換する．

$$\tilde{p}(NP_i) = \mathrm{softmax}(p(NP_i)) = \frac{\exp(p(NP_i))}{\sum_{k=1}^{N} \exp(p(NP_k))} \tag{2.7}$$

そして，名詞句 NP_i に対する重要度のランクは，式 (2.8) に示すように，NP_i と d とのコサイン類似度に位置バイアスをかけて計算される．

$$\text{SIFRank+}(NP_i, d) = \tilde{p}(NP_i) \cdot \text{Sim}(v_{NP_i}, v_d) \tag{2.8}$$

長い文書のデータセットである DUC2001 を用いた実験では，SIFRank+は SIFRank よりも優れていることが確認されている．

2.4 BERT

Bidirectional Encoder Representations from Transformers(BERT)[8] は代表的な言語モデルの一つである．Wikipedia や書籍を合わせた大規模なコーパスを使って事前学習されたモデルを，下流タスクのデータセットでファインチューニングすることで，自然言語処理の様々なタスクの性能を大幅に改善した．BERT で入力を作成する際には，入力の開始を表す [CLS] トークンと，入力における文の区切りを表す [SEP] トークンという 2 つの特殊トークンが使われている．BERT の事前学習は Masked Language Model と Next Sentence Prediction の 2 つのタスクで構成される．BERT の概要を図 2.8 に示す．

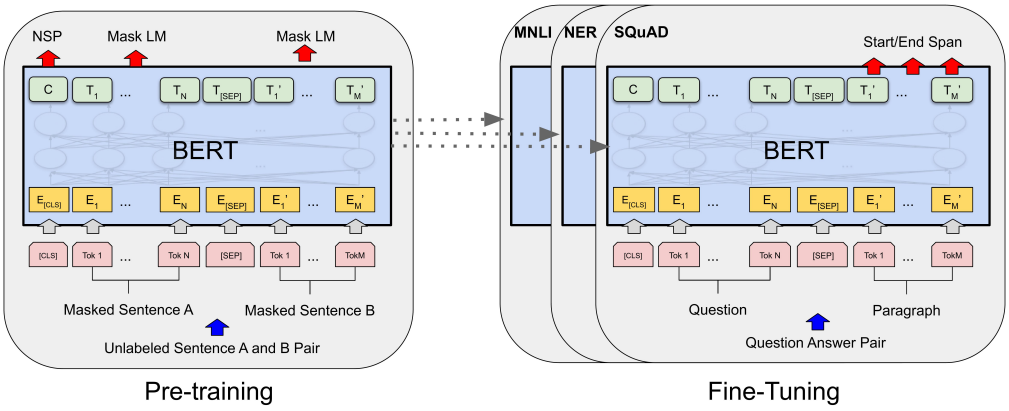

図 2.8: BERT の概要 [8]

Masked Language Model タスクは，トークンの穴埋めを行うタスクである．トークン列中のいくつかのトークンをランダムに選び，それを隠して，先行するトークン列と後続するトークン列の双方の文脈情報を使用して隠したトークンを予測することで，双方向から文脈を考慮したモデルの学習を実現している．Next Sentence Prediction タスクでは，BERT の事前学習の際に，用意された文書から 2 つの文が取り出され，これらを [SEP] で連結し，モデルに入力する．そして，入力を構成する 2 つの文が元の文書で連続して出現するか否かを予測する．BERT の事前学習は，訓練データのテキストに対して，Masked Language Model タスクと Next Sentence Prediction の両方が同時に適用され，損失関数はこの 2 つのタスクの損失関数を加算したものが使われる．

BERT のファインチューニングは，事前学習済みのモデルに対して下流タスクに合わせた全結合層を追加し，下流タスクのデータセットを使ってモデル全体のパラメータを更新する処理である．例えば，テキスト分類のタスクに BERT を適用するとき，最初の [CLS] トークンに対する最終の隠れ層の埋め込みを文全体の埋め込み表現とする．BERT の上部には単純な全結合層とソフトマックス分類器が追加され，正解ラベル c の確率を予測する．W を全結合層の重み行列としたとき，クラス c の確率は式 (2.9) のように求められる [28].

$$p(c|h) = \mathrm{softmax}(Wh) \tag{2.9}$$

BERT は，自然言語処理における様々なタスクに応用でき，高い正解率が得られることが報告されている．談話関係解析のタスクに BERT を適用することも可能である．本研究では，談話関係解析のタスクの一種として，節の組が意見とその根拠を含むかを BERT を用いて分類することを試みる．

2.5 本研究の特徴

2.1 節で述べたように，レビューの有用性を判定する様々な先行研究があるが，本研究はこれらの研究とは異なる基準でレビューの有用性を定義する．すなわち，レビューにユーザの意見とその根拠が含まれている場合，そのレビューは有用であると定義する．これは，レビューの有用性を包括的に定義するものではなく，特定の視点から有用性を定義したものであるが，意見と根拠を含むか否かを自動的に分類する技術が確立されれば，新しい視点からレビューの有用性を評価することが可能になる．

2.2 節で紹介した談話関係解析に関する先行研究は，数ある談話関係の一つとして根拠関係を扱っており，根拠関係の解析に焦点を当てた研究ではない．本研究は，根拠関係の解析のみに注力し，根拠関係を分類するモデルを学習するための訓練データを自動的に構築する手法を提案する．

第3章　意見と根拠を含む節の組の分類

3.1　概要

本章では，商品レビューに意見とその根拠が含まれるかどうかを判定する手法について述べる．意見とそれに対する根拠は 2 つの節によって記述されていることが多い．以下にその例を示す．ポジティブな意見 (「十分に満足しています.」) は節 c_2 に現われており，その根拠 (「安い価格で購入した.」) は節 c_1 に現われている．

$$\underline{\text{この高級モデルを安い価格で購入したので,}}_{c_1}\ \underline{\text{十分に満足しています.}}_{c_2}$$

提案手法の概要を図 3.1 に示す．複数の文から構成される商品レビューを入力とし，これが意見とそれに対する根拠を含むか否かを出力とする．まず，レビューから係り受け関係にある節の組を抽出する．図 3.1 では (c_{i1}, c_{i2}) と示されている．次に，それぞれの節の組に対し，それが意見と根拠を含む (Yes) か否か (No) を判定する．以下，この分類器を「根拠関係分類器」と呼ぶ．そして,「最終判定」のモジュールでは，ひとつでも Yes と判定された節の組がある場合，その商品レビューは意見と根拠を含むと判定し，それ以外は含まないと判定する．

図 3.1 において最も重要な処理は根拠関係分類器の学習である．本研究では，意見とその根拠を含む節の組，すなわち意見とその根拠という関係を含む節の組を「根拠関係-節ペア」と呼ぶ．また，意見と根拠を含まない節の組を「非根拠関係-節ペア」と呼ぶ．そして，2 つの節が与えらたとき，それが根拠関係-節ペアか非根拠関係-節ペアのいずれであるかを判定する新しいタスクを定義する．以下，このタスクを「根拠関係分類タスク」と呼ぶ．根拠関係分類器はこのタスクを解くモデルである．

表 3.1 に根拠関係-節ペアと非根拠関係-節ペアの例を示す．(1) の根拠関係-節ペアにおいて，レビューの対象となっている商品は書籍であるが，節 2 は「その本は読みやすい」という意見を表しており，節 1 はその理由を説明している．一方，(2) の非根拠関係-節ペアについては，節 1 と節 2 の間にそのような関係は見られない．なお，節は通常短い文である可能性があり，(2) の例のように節ペアは実際には文の組になることがある．

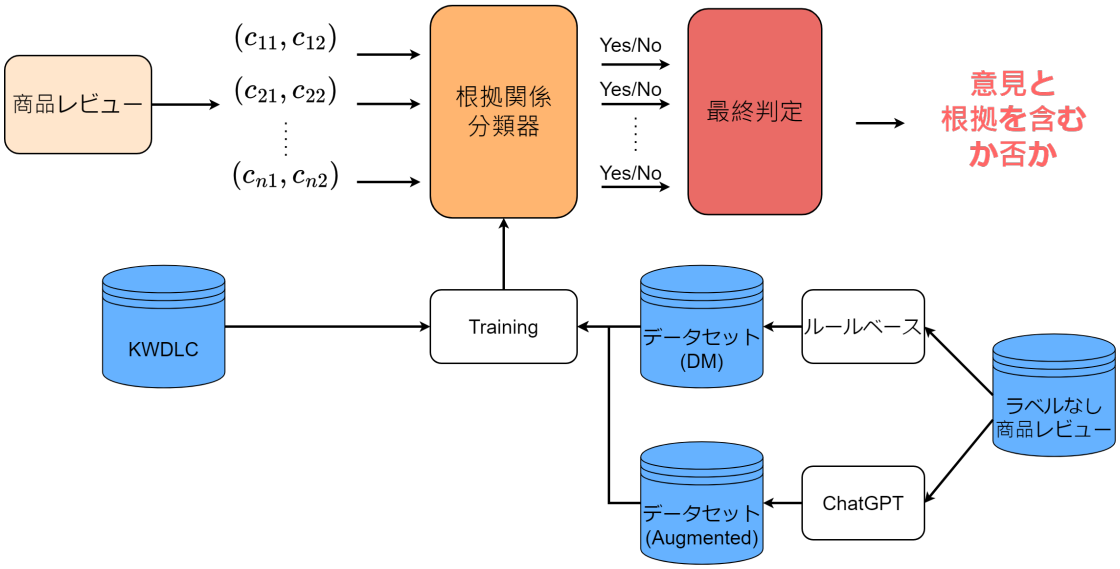

図 3.1: レビューが意見とその根拠を含むかを判定する手法の概要

表 3.1: 根拠関係-節ペアと非根拠関係-節ペアの例

	節 1	節 2
(1) 根拠関係-節ペア	インタビュアーとの対談形式になっているので,	内容はわかりやすく説得力があります.
(2) 非根拠関係-節ペア	マスキングテープは色々と使えて便利です.	メール便で発送してもらえるのも嬉しいサービスです.
(3) 根拠関係-節ペア	とっても可愛いくて大満足だったのですが,	近くのショップでもっと安く販売されてました.

研究の初期の段階では，意見とその根拠は別々の節にあると仮定していた．しかし，レビューにおいて意見とその根拠がどのように現われるかを調査したところ，意見と根拠が 1 つの節に現われることも多いことがわかった．予備調査では，根拠関係-節ペアの 60%において，意見とその根拠が 1 つの節に出現していた．その例を表 3.1 の (3) に示す．節 1 には肯定的な意見「完全に満足」とその根拠「とても可愛い」が含まれている．本研究が提案する根拠関係分類タスクでは，意見とその根拠が 1 つの節にあるか 2 つの節に分かれているかどうかに関係なく，節ペアが意見とその根拠を含むか否かを分類することを目的とする．

本章の以降の節では提案手法の詳細について述べる．3.2 節では，レビューから節の組を抽出する手法について述べる．3.3 節では，根拠関係分類器を学習するための訓練データを構築する手法について述べる．本研究では，既存の談話関係のデータセットである「KWDLC」，談話標識 (Discourse Marker; DM) を手がかりに構築した「データセット (DM)」，ChatGPT を使って構築した「データセット

(Augmentation)」の 3 つのデータセットを訓練データとして用いる．3.4 節では，これらのデータセットから根拠関係分類器を学習する手法について述べる．

3.2 節の分割

本節では，与えられたレビューから節の組を抽出する方法について述べる．まず，句点を文の区切りとして，レビューを文に分割する．次に，読点を節の区切りとして，文を節に分割する．その後，日本語構文・格・照応解析システム KNP[37] により文節の係り受け解析を行う．KNP は，文節の間の係り受け関係を出力するが，文節は基本的なフレーズやチャンクに類似した言語的単位であり，節よりも小さな単位である．文節間の係り受け関係から，(先ほど読点で分割した) 節間の係り受け関係を決定する．

次に，同じ文内で係り受け関係にある節の組を抽出する．係り受け関係にある節は互いに関連性が高く，意見とそれに対する根拠を含む可能性が高いと考えられる．また，意見と根拠は同じ文ではなく，異なる文に出現する可能性もある．この場合，日本語における主節は最後の節であることから，意見や根拠も文の最後の節に出現する可能性が高いと考えられる．そこで，連続する 2 つの文におけるそれぞれの最後の節を節ペアとして抽出する．これらの節ペアが根拠関係分類タスクの対象となる．

節ペアの抽出の例を示す．図 3.2 は係り受け解析の結果の例である．ここで c_i^1 と c_i^2 はそれぞれ最初の文 (Sentence 1) と二番目の文 (Sentence 2) における節を表す．また，矢印は 2 つの節の間に係り受け関係があることを示す．同一文内の節ペアとして，(c_1^1,c_3^1)，(c_2^1,c_3^1)，(c_3^1,c_4^1)，(c_1^2,c_3^2)，(c_2^2,c_3^2) が抽出される．文間の節ペアとして，(c_4^1,c_3^2) が抽出される．

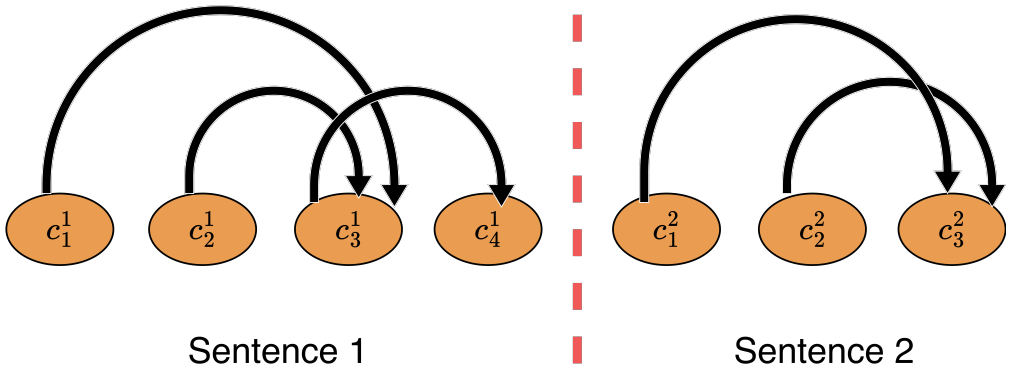

図 3.2: 節間の係り受け解析の例

3.3 訓練データ

3.1 節で述べたように，本研究では根拠関係分類器を学習するために 3 つのデータセットを用いる．以下，それぞれのデータセットの詳細について述べる．

3.3.1 京都大学ウェブ文書リードコーパス

2.2.2 項で紹介した京都大学ウェブ文書リードコーパス [36][1]は，ウェブ上の文書に対して文間の談話関係が人手で付与されたデータセットである．同コーパスに付与されている談話関係タグのうち，本研究と最も関連のあるタグは「原因/理由」である．これは PDTB 2.0 および PDTB 3.0 における談話関係タグ「CONTINGENCY.Cause」に相当する．

KWDLC は専門家とクラウドソーシング 2 つのデータセットから構成されている．前者は専門家が談話関係をアノテーションしたデータであり，データ数は少ないが，その談話関係タグは信頼できる．後者はクラウドワーカーが談話関係をアノテーションしたデータであり，データ数は多いが，クラウドワーカーは専門的な知識を持っていないため，談話関係タグの信頼性は低い．

以上を踏まえ，KWDLC から根拠関係分類器の学習データを以下のように構築する．まず，KWDLC において,「原因/理由」タグでラベル付けされた文の組を正例 (根拠関係-節ペア) として抽出する．ただし，専門家データセットにおける全てのペアと，クラウドソーシングデータセットの中で信頼度が 0.7 以上のペアのみを使用する．クラウドソーシングデータセットでは，談話関係タグは複数のクラウドワーカーによって選択されたタグの多数決によって決定され，またその談話関係タグの信頼度はそのタグを選択したクラウドワーカーの人数の割合で算出される．一方,「原因/理由」以外の談話関係タグが付与された文の組を負例 (非根拠関係-節ペア) として抽出する．ただし，正例と負例の数は同じとし，負例はランダムにサンプリングする．

KWDLC から構築されたデータセットは，本研究が対象とする根拠関係分類タスクに必ずしも適しているわけではないことに注意が必要である．まず，KWDLC と商品レビューはドメインが異なる．KWDLC では様々なウェブ文書に対して注釈付けされているが，根拠関係分類タスクのドメインは商品レビューである．すなわち KWDLC はアウト・ドメインのデータセットである．テストデータと異なるドメインの訓練データから学習された分類器は，同じドメインの訓練データから学習した分類器と比べて，分類の性能が低くなることが知られている．また，KWDLC の「原因・理由」関係と我々が定義した根拠関係は似ているが，完全に同じではない．KWDLC における「原因/理由」の関係は一般的な因果関係を表しているが，本研究における根拠関係分類タスクでは意見とそれに対する根拠といったより具体的な関係を対象としている．表 3.2 に，KWDLC において「原因/理由」のタグ

[1] https://github.com/ku-nlp/KWDLC

でアノテーションされた文の組の例を示す．(文 1) において「取引先を探していること」は，(文 2) における「連絡を求める」ことの理由になってはいるが，意見とその根拠の関係ではない．

表 3.2: KWDLC における「原因/理由」関係の例

文 1	文 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．	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．

3.3.2 談話標識によるデータセット

本項では，イン・ドメインのデータセットをルールベースの手法で自動構築する手法について述べる．イン・ドメインのデータセットとは，テストデータと同じドメインのデータ，すなわち商品レビューに対して根拠関係のラベルが付与されたデータセットを指す．ここでは理由を表す談話標識である「から」と「ので」に着目する.「から」と「ので」は接続詞であるが，これらの接続詞の前にはある事柄についての理由が説明されていることが多い．したがって，これらの理由を表す談話標識を含む節の組を根拠関係-節ペアとして抽出することでデータセットを構築する．

まず，ラベルなしのレビューを用意する．次に，3.2 節で説明した手法によって節の組を抽出する．もし節に「から」または「ので」が含まれていれば，意見とその根拠が含まれた正例 (根拠関係-節ペア) として抽出する．また，これらの談話標識を含まない節の組を負例 (非根拠関係-節ペア) として抽出する．ただし，正例と負例の比は 1:1 とする．負例とする節の組は正例と同じ数だけランダムにサンプリングする．

談話標識を手がかりに抽出された根拠関係-節ペアの例を表 3.3 に示す．下線の単語は理由を表す談話標識である．いずれの例も，節 1 は節 2 が表す事実の理由を説明している．

KWDLC とは異なり，談話標識を手がかりに構築したデータセットはイン・ドメインのデータであり，根拠関係分類器の学習に適していると言える．しかし，正例と負例は自動抽出されているため，誤りを含む可能性があることに注意する必要がある．すなわち，談話標識「から」「ので」は必ずしも理由を表すわけではない．また，理由を表すときでも，ユーザの意見に対する理由でないこともある．表 3.3 の 2 番目の例では，節 2 には商品 (お米) に関する意見は含まれていない．

表 3.3: 談話標識が含まれる節ペアの例

節 1	節 2
送料無料でレビューも良かったので,	評判が良ければまた買いたいと思います.
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,	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.

3.3.3 ChatGPT によるデータセット

Dai らと Gilardi らは，ChatGPT がデータ拡張に適しており，高品質な拡張データを提供できることを示している [7, 11]．ここでのデータ拡張とは，自動的にラベルを付与したデータを生成し，訓練データの量を増やす処理を指す．本項では，根拠関係分類タスクについて，ChatGPT-3.5 を利用してデータ拡張を行う．特に，談話標識「から」「ので」を含まない根拠関係-節ペアを獲得することを狙う．このような正例は 3.3.2 項で述べた談話標識を手がかりに構築したデータセットに含まれていない．

図 3.3 は，ChatGPT によりデータセットがどのように構築されるかを示している．まず，9 つの異なる商品ジャンルの商品レビューから，意見とその根拠の両方を含む節を抽出する．それぞれの節に対し，ChatGPT により，それと類似しかつ意見とその根拠を含む新しい節を 40 個生成する．具体的には，ChatGPT に以下のプロンプトを与える．

> 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：[元の節]

本研究における根拠関係分類タスクでは節の組を分類するが，この手法で生成されるのは単一の節である．節の組のデータセットを構築するため，ChatGPT を用いて生成した節は，同じジャンルのレビューからランダムに選ばれた別の節と結びつけられ，正例の根拠関係-節ペアが作られる．作成された正例の半分は，ChatGPT によって生成された節を最初の節とし，残りの半分は 2 番目の節とする．

上記の方法は正例のみ生成する．そこで，3.3.2 項で説明した手法と同じ手法で負例を作成する．負例の数は正例の数と同じとする．

表 3.4 は ChatGPT によるデータ拡張の例を示す．c_o は元の節であり，c_{g1} および c_{g2} は ChatGPT によって生成された節である．これらの節はほぼ元の節と同じ意味を持ち，意見とその根拠を含んでいる．

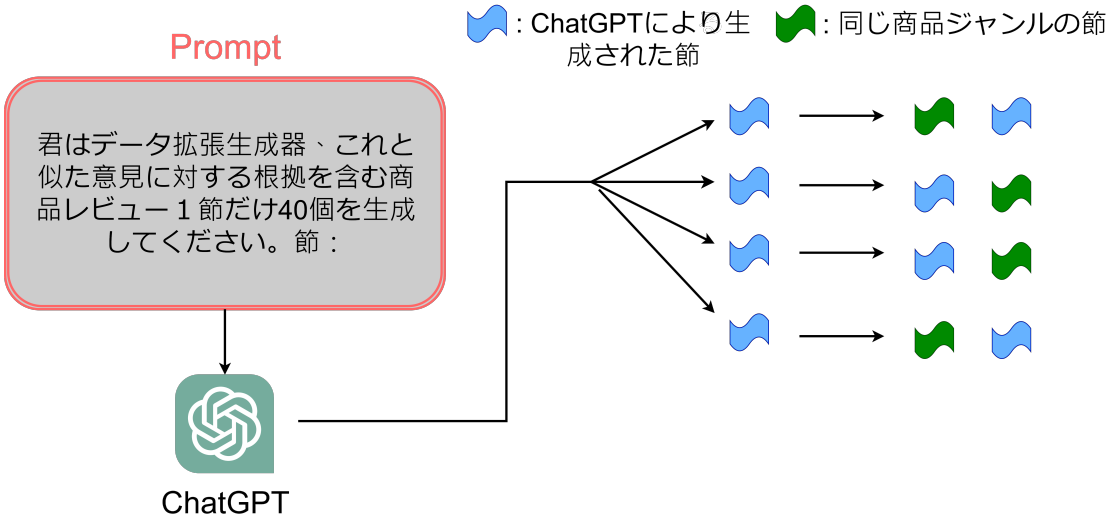

図 3.3: ChatGPT によるデータ拡張

表 3.4: ChatGPT によって生成された節の例

c_o	良いものが安く買えて良かったです.
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.
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.

3.4 分類器の学習

根拠関係分類タスクを解くための分類器，すなわち根拠関係分類器を学習あるいは実装するための 4 つの手法を提案する．本節では，その 4 つの手法の詳細を順に述べる．

3.4.1 ルールベースの手法

この手法では，談話標識を手がかりとしたルールによって，節ペアが根拠関係-節ペアか否かを判定する．具体的には，入力とした節の組のいずれかに談話標識「から」もしくは「ので」が出現していれば，その節ペアを根拠関係-節ペアと判定し，それ以外は非根拠関係-節ペアと判定する．

まず，入力の節ペアにおけるそれぞれの節を日本語形態素解析器 Janome[30] で単語分割する．次に，分割された単語をひとつずつ走査し，それが「から」または「ので」と一致するかを判定する．談話標識のいずれかと一致した単語が見つかった場合のみ，その節ペアは意見とその根拠を含むと判定する．

3.4.2 BERT

根拠関係分類タスクを下流タスクとし，事前学習済み BERT をファインチューニングすることで，根拠関係分類器を学習する．BERT のファインチューニングには 3.3 節で構築したデータセットを用いる．

BERT は，(1 つの) 文の分類タスクにも適用できるし，文間の関係を分類するタスクにも適用できる．これを踏まえ，BERT に与える入力として以下の 2 つを提案する．ひとつは，入力とする節の組が与えられたとき，2 つの節を特殊トークン [SEP] で区切り，これを入力とする方式である．これは節間の関係 (本研究では意見と根拠の関係) を分類することに相当する．以下，これを入力とする BERT モデルを「BERT-2C」と呼ぶ．もうひとつは，入力とする節の組が与えられたとき，2 つの節を連結として 1 つの文とし，これを入力とする方式である．これは 1 つの文が意見と根拠を含むかを分類することに相当する．以下，これを入力とする BERT モデルを「BERT-1C」と呼ぶ．BERT-2C と BERT-1C における入力の書式を以下に示す．[CLS] は BERT で用いられる特殊トークンで，これに対する埋め込みが節または節の組の意味表現として利用される．

BERT-2C: [CLS] 節 1 [SEP] 節 2 [SEP]
BERT-1C: [CLS] 節 1 節 2 [SEP]

事前学習済みモデルとして，日本語 Wikipedia から事前学習され，東北大学によって公開されている BERT base Japanese モデル [38] を用いる．

3.4.3 Intermediate Fine-Tuning

Intermediate Fine-Tuning（IFT）は，モデルの性能を向上させるために，対象タスクのデータセットを用いてファインチューニングを行う前に，中間データセットを使用してモデルをファインチューニングする手法である [6, 24]．ここでの中間データセットとは，対象タスクと関連が深い別のタスクのデータセットである．対象タスクのデータセットが少なく，かつ比較的大規模な別の関連タスクのデータセットが中間データセットとして用意できる場合，IFT が適用される．まず，中間データセットでモデルをファインチューニングすることで，事前学習モデルに対象タスクと関連する知識を転移させ，その後対象タスクのデータセットを用いてファインチューニングすることで，対象タスクの知識を転移させる．Cengiz らは，IFT を使用することで医療 NLI(Natural Language Inference) タスクのパフォーマンスが向上することを示した [6]．本研究では，IFT を BERT をベースとした根拠関係分類器の学習に応用する．中間データとして，アウト・ドメインのデータセットである KWDLC を用いる．また，対象タスクのデータとして，イン・ドメインの 2 つのデータセット，すなわち談話標識を手がかりに構築したデータセットと ChatGPT によるデータ拡張によって構築したデータセットを用いる．まず，アウ

ト・ドメインのデータセットを用いてBERTをファインチューニングし，その後，イン・ドメインのデータセットを用いてBERTをファインチューニングする．本研究が対象とする根拠関係分類タスクにおいて，アウト・ドメインとイン・ドメインのデータセットを結合し，これを用いてBERTを一度だけファインチューニングする手法と比べて，IFTが優れているかを検証する．

3.4.4 ハイブリッド手法

一般に，ルールベースの手法は，BERTのようなディープラーニングモデルよりも解釈可能で透明性が高い．さらに，ルールベースの手法は，特に少ないルールしか用いない場合，再現率は低いが精度が高い性質を持つことが多い．ルールベースの手法と機械学習に基づく手法の両方のアプローチの強みを活かすために，両者を組み合わせて用いる方法を提案する．

最初に，精度が比較的高いと予想されるルールベースの手法を用いて，入力の節ペアが根拠関係-節ペアか否かを判定する．根拠関係-節ペアと判定されたとき，それをそのまま最終の判定結果とする．非根拠関係-節ペアと判定されたとき，BERTによる根拠関係分類器を用い，その判定結果を最終の判定結果とする．

第4章　商品への言及の有無の分類

本章では，1.3 節の図 1.1 に示した「意見の対象判定」のモジュールの詳細について述べる．

4.1　問題設定

意見に対する根拠関係を含む節ペアに対し，その意見が商品に関連があるか否かを判定する．言い換えれば，レビューが商品に言及しているか否かを判定する．1.2 節で述べたように，本研究では単に意見とその根拠を含むレビューではなく，評価対象の商品に対する意見と根拠を含むレビューを有用なレビューとして検出する．例えば,「別の店より安かったので，得をした.」という節ペアは根拠関係を含んでいるが，商品とは直接関係がないと考えられるため，有用なレビューではない判定とする．

以下，上記の問題を「商品言及分類タスク」と呼ぶ．

4.2　提案手法

レビューにおける節ペアで書かれている意見が商品に対する意見であるときには，商品に関連する言葉が使われていると予想される．例えば，評価の対象となる製品が「パスタソース」であるとき,「味」「おいしい」「調理」といった商品に関連する単語が用いられると考えられる．そこで，本研究では，商品に関連するキーワードの集合をあらかじめ用意し，それが節ペアに出現するかによって，その意見が商品に関連するか否かを判定する．

提案手法の処理の流れを図 4.1 に示す．まず，商品に関連するキーワード集合をあらかじめ構築する．ただし，商品毎にキーワード集合を構築するのは，評価対象となる商品の数が非常に多いことから，現実的には難しい．そこで，本研究では，商品カテゴリ毎に関連するキーワード集合を構築する．ここでの商品カテゴリとは,「食品」「本」「家電」など，商品の種類を分類したものである．

本研究では，楽天市場のデータセット [27] で定義されている商品カテゴリを利用する．同データセットにおける商品カテゴリは階層構造になっているが，本研究では最上位のカテゴリを商品カテゴリと定義する．それぞれの最上位カテゴリについて，それに属する商品のレビューを全て取得し,「レビュー集合」とする．次

に，レビュー集合から重要なキーワードを抽出し,「キーワード集合」を得る．評価対象の節ペア (ここでは意見とその根拠を含む節の組) が与えられたとき，その節ペアを含むレビューの対象商品ならびにその商品カテゴリも与えられると仮定する．レビューの商品カテゴリから，カテゴリの階層構造を上に辿り，最上位の商品カテゴリを得る．図 4.1 の例では，入力されたレビューの商品カテゴリは「その他」であり，階層構造を「パスタ」→「麺類」→「食品」と辿り，最終的に最上位のカテゴリ「食品」を得る．

最後に，その最上位のカテゴリに対応する「キーワード集合」と「レビュー」に対して「照合」の処理を行う．具体的には，節ペアに商品カテゴリのキーワード集合のいずれかのキーワードが出現するときには，その節ペアの意見は商品に関連があるとみなす．逆に，節ペアに商品カテゴリのキーワードがひとつも出現していないときは，その節ペアの意見は商品に関連がないとみなす．

次節では，商品レビューの集合から重要なキーワードを抽出する手法について述べる．

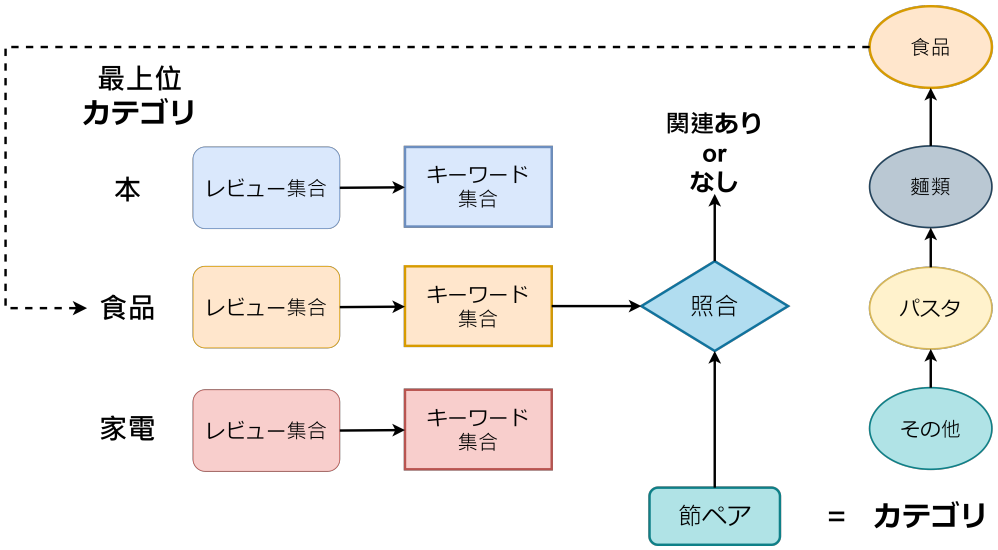

図 4.1: 節ペアが商品に言及しているかを判定する提案手法の概要

4.3 キーワードの抽出

本節では，ある商品カテゴリのレビューの集合から，その商品に関連するキーワードを抽出する手法について述べる．まず，レビュー集合から自立語 (名詞，動詞，形容詞など) を抽出し，これをキーワードの候補とする．次に，教師なしキーワード抽出手法を利用して，それぞれの単語の重要度を計算する．最後に，重要度の大きい上位 200 件の単語をカテゴリのキーワードとして抽出する．

キーワードを抽出する手法として，TF-IDF，TF-ENT，YAKE![4]，TF-IDF+YAKE!，TF-ENT+YAKE!の 5 つの手法を用いる．以下，それぞれの詳細を説明する．なお，以降の説明では，カテゴリを c，キーワードの候補となる単語を w_i と記す．

TF-IDF TF-IDF は,「ある文書に出現するある単語の重要度」を表す統計的な尺度であり，情報検索の分野で広く使われている．ここでは，あるカテゴリ c に属する全てのレビュー集合を 1 つの文書とみなし，カテゴリ c(文書) における単語 w_i の重要度を TF-IDF を用いて計算する．TF-IDF の定義を式 (4.1) に示す．

$$\text{TF-IDF}(w_i, c) = \text{TF}(w_i, c) \times \text{IDF}(w_d, D) \tag{4.1}$$

D は一般に全文書集合を表す．本研究の場合，カテゴリ毎にレビューをまとめた仮想文書の集合である．

TF と IDF の計算式を，それぞれ式 (4.2) と式 (4.3) に示す．

$$\text{TF}(w_i, c) = \frac{\text{カテゴリ } c \text{ のレビューにおける } w_i \text{の出現回数}}{\text{カテゴリ } c \text{ のレビューの全文字数}} \tag{4.2}$$

$$\text{IDF}(w_i, D) = \log\left(\frac{|D| + 1}{w_i \text{が出現するカテゴリの数}}\right) \tag{4.3}$$

TF は，キーワード w_i がカテゴリ c のレビューによく出現するとき，高い値を取り，その単語が重要であることを表す．一方，IDF は，キーワード w_i が多くのカテゴリのレビューに出現するほど低く，逆に少数のカテゴリのレビューにしか出現しないときは高くなる．IDF によるスコア付けは，キーワードが限られたカテゴリのレビューに出現するときは，キーワードはそのカテゴリにおいて重要な役割を果たし，重要であるという考えに基づく．

全ての候補単語について TF-IDF を計算し，その上位 200 件の単語をキーワード集合として取得する．

TF-ENT TF-IDF は，文書内での単語の重要性を評価する際に広く使用されているが，本研究ではカテゴリ c に当てはまるレビュー全体を仮想的な文書として使用しているため，IDF が効果的に働かない可能性がある．例えば,「食品」のカテゴリの重要なキーワードとして「味」があるとする．もし「味」が「食品」カテゴリのみに出現するなら，その IDF は高く見積られる．しかし，実際には「味」という単語は，少ない回数ではあるが，他のカテゴリのレビューにも出現すると考えられる.「味」が「食品」カテゴリのレビューで非常によく出現する場合でも，他のカテゴリでも例外的に 1 回でも出現すれば，式 (4.3) における分母が大きくなり，IDF も低く見積られる．

この問題に対処するために，ある単語が特定のカテゴリに出現する傾向をIDF の代わりにエントロピー (ENT) を用いて評価する手法を提案する．ここでのエントロピーは，単語がカテゴリに出現する確率分布のエントロピーであり，式 (4.4) のように定義される．

$$ENT(w_i) = -\sum_{c} P(c|w_i) \cdot \log_2 P(c|w_i) \tag{4.4}$$

確率 $P(c|w_i)$ は，単語 w_i がレビューに出現したとき，そのレビューのカテゴリが c である確率である．その定義を式 (4.5) に示す．

$$P(c|w_i) = \frac{\text{カテゴリ } c \text{ における } w_i \text{の出現頻度}}{\text{全カテゴリにおける } w_i \text{の出現頻度の和}} \tag{4.5}$$

$ENT(w_i)$ は，キーワード w_i が全てのカテゴリに均等に出現するときに高くなり，逆に特定のカテゴリのみに偏って出現するときに低くなる．本研究は，エントロピーはIDF よりも，キーワードが特定のカテゴリのみに現われるかを評価するのに適していると考える．そのため，IDF を ENT に置き換えた式 (4.6) によってキーワードの重要度を算出する．

$$\text{TF-ENT}(w_i, c) = \text{TF}(w_i, c) \times \frac{1}{ENT(w_i) + \alpha} \tag{4.6}$$

α は分母が 0 になることを避けるための定数である．本研究では α=0.1 と設定する．

全ての候補単語について TF-ENT を計算し，その上位 200 件の単語をキーワード集合として取得する．

YAKE! 2.3.2 項で紹介した統計ベースの抽出手法 YAKE!を利用する．YAKE!はもともと英語を対象に文書からキーワードを抽出する手法であり，単語がスペースで区切られていることを想定している．日本語ではスペースによって単語が区切られていないため，前処理として，日本語形態素解析器 janome[30] を用いて事前に分かち書きを行う．さらに，YAKE!のパラメータを設定する際には，言語を日本語に指定し，n-gram の最大のサイズは 3 に設定する．また，重複削除の閾値は 0.9 と設定する．YAKE!によって候補単語の重要度のスコアを算出し，その上位 200 件の単語をキーワード集合として取得する．

TF-IDF+YAKE! TF-IDF によるキーワード抽出と YAKE!によるキーワード抽出を組み合わせる手法である．まず，TF-IDF を計算し，その上位 100 件のキーワードを抽出する．次に，YAKE!を用いて 100 件のキーワードを抽出する．これらの和集合を最終的なキーワード集合とする．合計のキーワードの

数は 200 であるが，2 つの手法で同じ単語が選ばれたときには重複を除くため，実際のキーワード数は 200 よりも小さくなる．

TF-ENT+YAKE! TF-ENT によるキーワード抽出と YAKE!によるキーワード抽出を組み合わせる手法である．TF-ENT によって 100 件，YAKE!によって 100 件，キーワードを抽出し，その和集合 (最大で 200 件) を最終的なキーワード集合とする．

上記の 5 つの手法のうち，TF-ENT と TF-ENT+YAKE!は本研究における提案手法，他の手法は既存の手法である．

第5章　評価

本章では，提案手法の評価実験について述べる．まず，5.1 節では，本実験のテストデータについて述べる．5.2 節では，3 章で述べた根拠関係分類タスクを解く提案手法の評価実験の実験設定と結果を報告し，提案手法の有効性について考察する．5.3 節では，4 章で述べた商品言及分類タスクを解く提案手法の評価実験の設定を説明し，実験結果と考察について報告する．また，抽出されたキーワードの例を示す．

5.1　テストデータ

本章の冒頭で述べた通り，本実験では根拠関係分類タスクと商品言及分類タスクについて評価する．ここではそれぞれのタスクのテストデータについて述べる．

根拠関係分類タスクについては2つのテストデータを用いる．ひとつは「KWDLC テストデータ」である．3.3.1 項で述べた KWDLC から構築したデータセットを 8:1:1 の割合で分割し，それぞれ訓練データ，検証データ，テストデータとする．「KWDLC テストデータ」は上記の分割によって作成されたテストデータである．KWDLC のラベルは人手で付与されたものであるが，既に述べたように，KWDLC における「原因/理由」の関係は必ずしも本研究が目指す意見とその根拠の関係ではないこともある．

もうひとつは「楽天テストデータ」である．これは，実際のレビューから抽出された節の組に対し，それが根拠関係-節ペアか非根拠関係-節ペアであるかを人手でラベル付けしたデータである．まず，楽天市場のレビューから，162 件のレビューをランダムに選択する．次に，3.2 節で説明した方法に従い，レビューから節の組を抽出する．「句点」もしくは「述語＋読点」で節に分割し，分割した節の末尾が終止形ではないときには終止形に直す．そして，係り受け関係が成立する全ての節の組を抽出する．抽出された節ペアについて，2 人の注釈者が，これらの節がユーザの意見とそれに対する根拠を含むかを判定する．

2 つの評価データの統計を表 5.1 に示す．

商品言及分類タスクの評価データは以下のように構築する．まず，楽天テストデータで「根拠関係-節ペア」とラベル付けされた節ペアを抽出する．次に，被験者 1 名が，これらの節ペアに書かれている意見と根拠が商品に言及しているか否かを判定する．以上の手続きで商品への言及があるか否かをラベル付けしたデー

表 5.1: 根拠関係分類タスクのテストデータ

	根拠関係-節ペア	非根拠関係-節ペア
KWDLC テストデータ	131	131
楽天 テストデータ	186	329

表 5.2: 商品言及分類タスクのテストデータ

商品への言及あり	商品への言及なし
160	26

タセットを作成する．その統計を表 5.2 に示す．商品に言及がある節ペアはない節ペアと比べて約 6 倍多い．

楽天テストデータを作成する際に使用したレビューは，様々な商品カテゴリからランダムに選択している．その最上位の商品カテゴリを表 5.3 に示す．商品カテゴリの数は 28 である．

表 5.3: 楽天テストデータにおけるレビューの最上位の商品カテゴリー

家電	水，ソフトドリンク
車用品・バイク用品	食品
本・雑誌・コミック	インテリア・寝具・収納
スイーツ・お菓子	パソコン・周辺機器
ビール・洋酒	TV・オーディオ・カメラ
美容・コスメ・香水	レディースファッション
キッズ・ベビー・マタニティ	バッグ・小物・ブランド雑貨
花・ガーデン・DIY	ペット・ペットグッズ
日用品雑貨・文房具・手芸	医薬品・コンタクト・介護
メンズファッション	靴
CD・DVD・楽器	腕時計
おもちゃ・ホビー・ゲーム	キッチン用品・食器・調理器具
インナー・下着・ナイトウエア	ジュエリー・アクセサリー
スポーツ・アウトドア	ダイエット・健康

5.2 意見・根拠関係の分類の評価

5.2.1 訓練データ

本実験では，以下の 5 つのデータセットを訓練データとして用いる．これらのデータセットから学習した根拠関係分類器を比較する．

D_{kwdlc}：KWDLC から構築したデータセット (3.3.1 項)．テキストはウェブ文書であり，商品レビューではないため，アウト・ドメインのデータセットである．

D_{dm}：談話標識を手がかりに構築したデータセット (3.3.2 項)．テキストは商品レビューであり，イン・ドメインのデータセットである．

D_{gpt}：ChatGPT によって構築したデータセット (3.3.3 項)． イン・ドメインのデータセットである．

D_{dm+gpt}：$D_{dm} + D_{gpt}$． 2 つのイン・ドメインのデータセットを合わせたデータセット．

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

D_{dm} と D_{gpt} はラベルのないレビューから自動的に構築される．テストデータと同様に，ラベルなしレビューとして楽天データにおける楽天市場のレビュー [27] を使用する．表 5.4 は，5 つのデータセットの訓練データと検証データに含まれる節ペアの数を示している．“og” と “non-og” は「根拠関係-節ペア」ならびに「非根拠関係-節ペア」を表す．D_{dm} と D_{gpt} は，自動構築した後，8:2 の割合で訓練データと検証データに分割している．

表 5.4: データセットの統計

データセット	訓練データ		検証データ	
	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

5.2.2 実験設定

実験では，3.4 節で述べた以下の分類器を比較する：

Rule 3.4.1 項で述べたように，談話標識「から」「ので」の有無によって根拠関係の有無を判定するルールベースの手法．

BERT-2C 3.4.2 項で述べたように，BERT をファインチューニングすることで根拠関係分類器を学習する手法．2 つの節の組を入力として与える．

BERT-1C 同様に，BERT をファインチューニングすることで根拠関係分類器を学習する手法．節の組を連結した一文を入力として与える．

IFT-2C 3.4.3 項で述べたように，Intermediate Fine-Tuning によって BERT を学習する手法．訓練データが D_{all} のときのみ学習する．まずアウト・ドメインの D_{kwdlc} を用いて BERT をファインチューニングし，次にイン・ドメインの D_{dm+gpt} を用いて BERT をファインチューニングする．BERT への入力は BERT-2C と同じく 2 つの節の組とする．

IFT-1C 上記の IFT-2C とほぼ同じだが，BERT への入力を BERT-1C と同じく節の組を連結した一文とする手法．

Rule+BERT-2C 3.4.4 項で述べたハイブリット手法．まず Rule によって判定を行い，非根拠関係-節ペアと判定したときには BERT-2C を用いて最終判定を行う．

Rule+IFT-2C Rule と IFT-2C を組み合わせたハイブリット手法．

BERT のファインチューニングする際のハイパーパラメータの決定には，Preferred Networks 社が開発した optuna[1] という自動最適化フレームワークを使用する．Optuna は，ベイズ最適化によってパラメータの探索を効率的に行うツールである．学習回数の上限を 50 回と設定し，バッチサイズの探索範囲を $\{8, 16, 32\}$，学習率の探索範囲を $\{1e^{-5}, 2e^{-5}, 3e^{-5}\}$，検証データでのエポック数の探索範囲を $\{1, 2, 3, 4, 5\}$ とし，訓練データによるモデルの訓練と検証データによるモデルの評価を繰り返し，最適なハイパーパラメータの組み合わせを探索する．BERT のパラメータ学習の最適化アルゴリズムとして AdamW[18] を使用し，荷重減衰 (weight decay) は 0.01 に設定する．50 回の試行の中で検証データに対する F 値が最も良かったハイパーパラメータの組み合せを選択する．

上記の手法に加え，ChatGPT-3.5 もベースラインとして比較する．プロンプトとして，人手で作成した 3 つの正例サンプルと 3 つの負例サンプルを ChatGPT に与えて，テストデータの節ペアに根拠関係が含まれるかを尋ねる．

5.2.3 結果と考察

異なるデータセットと手法の組み合わせについて，根拠関係分類タスクの性能を評価した結果を表 5.5 に示す．評価指標は，根拠関係-節ペア検出の精度 (precision)，再現率 (recall)，F1 スコア (F1-score)，および分類の正解率 (accuracy) であり，それぞれ (a)-(d) の表にまとめている．Rule と ChatGPT の性能は，ラベル付きデータセットを使用しないため，全ての訓練データに対して同じ結果を示している．一方，IFT と Rule+IFT の性能は D_{kwdlc} と D_{dm+gpt} の両方のデータセットを必要とするため，D_{all} のみの結果を示す．各モデルにおいて一番結果が良かったデータ

セットの評価値を太字で示す．また，D_{all} で学習したモデルのうち最も結果が良かった評価値を太字かつイタリック体で示す．

イン・ドメインのデータセット（D_{dm} と D_{gpt}）で学習されたモデルは，アウト・ドメインのデータセット（D_{kwdlc}）で学習されたモデルよりも精度が優れているが，再現率に関しては性能が低い．したがって，これらのデータセットを組み合わせて用いることは，お互いの長所を補い合う効果が期待される．

F1 スコアと正解率に関しては，全体的に，イン・ドメインのデータセットの方がアウト・ドメインのデータセットよりも優れている．これは，イン・ドメインのデータセットを自動的に構築する我々の提案手法が有効であることを示している．D_{dm} と D_{gpt} を比較すると，BERT-2C モデルでは D_{gpt} の方が，BERT-1C では D_{dm} の方が結果が良い．しかし，BERT-2C は BERT-1C よりも F1 スコアや正解率が高いことから，総合的に見れば D_{gpt} の方が優れていると言える．これは ChatGPT によるデータ拡張の有効性を示している．D_{dm+gpt} は，個々のデータセット D_{dm} または D_{gpt} と比較して，再現率は向上しているが，F1 スコアはほぼ同等である．

最後に，D_{all} で学習したモデルの F1 スコアは，他のデータセットよりも優れている．3 つのデータセットを結合した訓練データを用いる場合と比べて，Intermediate Fine-Tuning は F1 スコアと正解率をさらに向上させる．最高の F1 スコアは D_{all} で学習した IFT-2C が達成した 0.71 であり，D_{kwdlc} で学習した BERT-2C より 0.09 ポイント高い．

これらの結果から，異なる種類のデータセット，すなわち，人手によってラベル付けされたアウト・ドメインのデータセットと，弱教師あり学習手法によってラベル付けされたイン・ドメインのデータセットを組み合わせる我々のアプローチが，意見に対する根拠関係の分類タスクに有効であることが示された．さらに，これらの異なる種類のデータセットを利用する際には，データセットの単純な結合よりも Intermediate Fine-Tuning の方が適していることがわかった．

これまではデータセットの違いについて考察していたが，手法の違いについて考察すると，以下のようにまとめられる．

- 全体的に，入力として節ペアを受け入れるモデル (*-2C) は，単一の節を受け入れるモデル (*-1C) より優れていることが分かった．意見と根拠はしばしば 1 つの節に現れるが，BERT モデルに入力するときは，2 つの節を分離するほうがよい．

- ハイブリッド手法は，BERT モデルを上回らない．特に，Rule+IFT-2C の精度は，Rule や IFT-2C よりも悪い．これは，Rule は評価データの 23%しか分類できないが，IFT-2C はこれらのデータに対する分類の精度が高い (0.79) ためである．したがって，Rule と IFT の組み合わせは改善をもたらさなかった．

- Rule の正解率は比較的高い．これは，テストデータのクラスの不均衡が原因である．表 5.1 に示したように，非根拠関係-節ペアの数は根拠関係-節ペ

表 5.5: 楽天テストデータにおける根拠関係分類の結果

(a) 精度 (precision)

モデル	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	0.47	0.49	**0.61**	0.53	0.63
BERT-1C	0.49	0.53	0.49	0.49	**0.54**
IFT-2C	–	–	–	–	0.67
IFT-1C	–	–	–	–	0.60
Rule+BERT-2C	0.48	0.50	**0.57**	0.51	**0.57**
Rule+IFT-2C	–	–	–	–	0.63

(b) 再現率 (recall)

モデル	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	**0.90**	0.55	0.70	0.84	0.76
BERT-1C	0.83	0.82	0.89	**0.92**	0.83
IFT-2C	–	–	–	–	0.75
IFT-1C	–	–	–	–	0.84
Rule+BERT-2C	0.82	0.63	0.81	0.82	**0.84**
Rule+IFT-2C	–	–	–	–	0.76

(c) F1 スコア (F1-score)

モデル	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	0.62	0.52	0.65	0.65	**0.69**
BERT-1C	0.62	0.64	0.63	0.64	**0.65**
IFT-2C	–	–	–	–	***0.71***
IFT-1C	–	–	–	–	0.70
Rule+BERT-2C	0.60	0.56	0.67	0.63	**0.68**
Rule+IFT-2C	–	–	–	–	0.69

(d) 正解率 (accuracy)

モデル	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	0.60	0.63	0.73	0.67	**0.75**
BERT-1C	0.63	0.67	0.62	0.63	**0.68**
IFT-2C	–	–	–	–	***0.77***
IFT-1C	–	–	–	–	0.74
Rule+BERT-2C	0.61	0.64	**0.71**	0.65	**0.71**
Rule+IFT-2C	–	–	–	–	0.72

アのほぼ2倍である．Ruleは談話標識が見つかった場合のみ根拠関係ありと判定するため，全体的には節ペアを非根拠関係-節ペアと分類する傾向がある．以上の理由からRuleの正解率は高くなっていると思われるが，これは必ずしもRuleが有効であることを示すものではない．RuleのF1スコアはBERTベースのモデルよりも低くなっている．

- ChatGPTは節ペアを根拠関係-節ペアと分類する傾向があり，再現率が高いが，精度は低い．F1スコアはファインチューニングしたBERTより低い．

5.2.4 ドメインの違いに関する考察

根拠関係分類タスクを解くための単純な手法は，既存のラベル付きデータセットKWDLCを使用することである．しかし，3.3.1項で議論したように，KWDLCのドメインはウェブ文書，本研究で想定する根拠関係分類タスクのドメインは商品レビューであり，両者は一致していないため，KWDLCは本研究のタスクに完全に適しているわけではない．

そこで，訓練データとテストデータの違いの影響を調査するために，追加実験を行った．D_{kwdlc}を訓練データとして使用し，楽天テストデータとKWDLCテストデータでの性能を比較する．Rule，BERT-2C，およびRule+BERT-2Cの精度（P），再現率（R），F1スコア（F），および正解率（A）を表5.6に示す．

表5.6: ドメインの異なるテストデータによる評価結果

手法	テストデータ	P	R	F	A
Rule	KWDLC	0.74	0.40	0.52	0.63
	楽天	0.72	0.47	0.57	0.74
BERT-2C	KWDLC	0.78	0.85	0.82	0.81
	楽天	0.47	0.90	0.62	0.60
Rule+BERT-2C	KWDLC	0.70	0.82	0.76	0.74
	楽天	0.48	0.82	0.60	0.61

アウト・ドメインのテストデータ(楽天)に対するBERTモデルの性能は，イン・ドメインのテストデータ(KWDLC)に対する性能よりも明らかに劣っている．ルールベース手法の性能は，この2つのテストデータではほぼ同等であるが，ハイブリッド手法ではアウト・ドメインの方が評価値が低い．

この実験結果は，レビューにおける意見と根拠を正確に識別するためには，イン・ドメインのデータセットを使用する必要があることを示唆する．

5.2.5　分類例とエラー分析

提案手法による根拠関係分類の例

表 5.7 に提案手法による根拠関係分類の例を示す．Gold は正解のクラスを，Pred. は F1 スコアが一番高かった IFT-2C(訓練データは D_{all}) によって予測されたクラスを表す．また，yes は根拠関係-節ペア，no は非根拠関係-節ペアを表す．

表 5.7: D_{all} から学習した IFT-2C による根拠関係分類の例

	Gold	Pred.	節 1	節 2
#1	yes	yes	品質，梱包，全て問題無く大変満足しています.	メールでの問い合わせにも丁寧に対応して頂きました.
#2	yes	no	良い所はやはりデザインと，	それからキャスターつきの所です.
#3	no	yes	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，	美味しかったです.

例 #1 は真陽性の例，すなわち提案手法によって根拠関係-節ペアであると正しく分類した例である．モデルは，意見「大変満足しています」と根拠「問題は見つかりません」が 1 つの節に含まれ，原因や理由を表す明示的な談話標識がないにも関わらず，節ペアに意見と根拠があると正しく分類している．

例 #2 は偽陰性の例，すなわち正解は根拠関係-節ペアであるが提案手法によって非根拠関係-節ペアと誤って判定した例である．提案手法のモデルは意見「良いところ」とその根拠「良いデザイン」,「キャスターつきの所」を捉えることができていない．これは，根拠を示す具体的な単語やフレーズが明示的に表現されていないためであると考えられる．

例 #3 は偽陽性の例，すなわち正解は非根拠関係-節ペアであるが提案手法によって根拠関係-節ペアと誤って判定した例である.「美味しかった」という意見に対する根拠がないにも関わらず，モデルは根拠関係があると判断している．

誤り分析

提案手法の誤り分析を行った．F1 スコアが最も高かったモデル IFT-2C について，偽陰性と偽陽性の例を分析し，誤りの原因を考察した．

偽陰性の誤りは 48 件あった．このうち，39 件 (81%) に肯定的な意見が含まれていた．表 5.7 の例#2 はそのような例であり，ユーザは製品のデザインが良いと述べている．データセットには，ユーザの肯定的な意見とそれに対する根拠を含む根拠関係-節ペアが多く含まれている．モデルは，肯定的な意見があるときにそれに対する根拠も存在すると誤認識する傾向が見られた．

偽陽性の誤りは 70 件あった．分析の結果，談話標識「ので」が誤りの主要な原因であることがわかった．70 の偽陽性のうち，17(24%) 件の節ペアにはこの談話標識が含まれていた.「ので」は因果関係を示すが，必ずしも意見と根拠の関係とは限らない．例えば，表 5.7 の例 #3 では,「ので」は利用者が不安になった理由 (この店で買うのは初めてだった) を表すが，ユーザの意見 (美味しかった) の理由にはなっていない.

5.3 商品への言及の有無の分類の評価

本節では，商品言及分類タスク，すなわち意見とその根拠を含む節ペアが与えられたとき，それが商品に関して言及しているか (商品に対する意見と根拠であるか) を判定する二値分類問題を解く提案手法を評価する.

5.3.1 実験設定

実験では，4.3 節で述べた教師なし 5 つの手法，TF-IDF，TF-ENT，YAKE!，TF-IDF+YAKE!，TF-ENT+YAKE!を比較する．これらの手法によって商品カテゴリ毎に重要なキーワード集合を抽出し，入力とする節ペアがそれらのキーワードを含むか否かで商品への言及の有無を判定する．テストデータとして，表 5.2 に示した 186 件の節ペアからなる商品言及分類タスクのデータセットを用いる．評価基準は,「言及あり」のクラスに対する精度 (precision)，再現率 (recall)，F1 スコア (F1-score) とする.

5.3.2 結果と考察

表 5.8: 商品言及分類タスクの評価結果

手法	精度	再現率	F1 スコア
TF-IDF	0.87	0.74	0.80
TF-ENT	0.87	0.68	0.76
YAKE!	**0.88**	0.84	0.86
TF-IDF+YAKE!	0.86	**0.91**	**0.89**
TF-ENT+YAKE!	0.86	0.89	0.87

実験結果を表 5.8 に示す．まず，TF-IDF と TF-ENT について比較すると，精度は同じであったが，再現率には明確な差が見られた．TF-IDF の再現率が 0.74 であったのに対し，TF-ENT の再現率は 0.68 であった．したがって，F 1 スコアも

TF-IDF の方が高くなった．一方，YAKE!は TF-IDF と TF-ENT よりも高い性能を示した．特に再現率は 0.84 と高く，TF-IDF と比べて 0.10 ポイント，TF-ENT と比べて 0.16 ポイント上回った．

次に，TF-IDF または TF-ENT と YAKE!との組み合せ手法について考察する．TF-ENT+YAKE!は，TF-ENT と比較して，F1 スコアが 0.11 ポイントが向上した (0.76 対 0.87)． また，YAKE!と比較したとき，F1 スコアは 0.01 ポイント高かった (0.86 対 0.87)．TF-IDF+YAKE!についても，TF-IDF や YAKE!と比べて，再現率や F1 スコアが高いことが確認された．以上から，複数のキーワード抽出手法を組み合わせることにより，商品言及分類の性能が向上することが確認された．

最後に，TF-IDF+YAKE!は，再現率 (0.91) と F1 スコア (0.89) が最も高く，精度も他の手法と比べて大きな差はなかった．今回の実験では，TF-IDF と YAKE! を組み合わせてキーワードを抽出する手法が最も有効であることがわかった．

5.3.3 キーワードの例

教師なしキーワード抽出手法によって実際に抽出されたキーワードの例を示す．合計 28 個の商品カテゴリのそれぞれについて重要なキーワードを抽出し，これらの中から頻度が最も高い 20 個のキーワードを可視化した．図 5.1 および図 5.2 は，TF-IDF および TF-ENT によって抽出された最頻出のキーワードとその頻度を示している．上位 3 個のキーワードは同じであり，それらは「値段」,「商品」,「気」である．また，商品に関連する名詞が多く見られる．例えば,「値段」,「商品」,「色」,「感じ」,「価格」などが TF-IDF，TF-ENT の両方によって抽出されている．全体的に，TF-IDF と TF-ENT では，抽出されるキーワードに大きな違いはなかった．

一方 YAKE!によって抽出された出現頻度上位の 20 個のキーワードとその出現頻度を図 5.3 に示す．YAKE!によって抽出されたキーワードには，商品に関連する評価語が多く見られる．例えば,「とても」「やすい」「気に入り」「可愛い」「良かった」などの単語がこれに該当する．また，YAKE!によって抽出されるキーワードは，TF-IDF や TF-ENT によって抽出されるキーワードとかなり異なることがわかる．したがって，2 つのキーワード抽出手法を組み合わせる TF-IDF+YAKE! と TF-ENT+YAKE!では，ひとつの抽出手法だけを用いる場合と比べて，商品に関する名詞と評価語の両方を抽出することができる．すなわち，より多様なキーワードを獲得することができる．このことが，表 5.8 において，2 つの手法を組み合わせる手法が単一の手法と比べて再現率や F1 スコアが高いことの主な原因であると考えられる．

5.3.4 TF-IDF と TF-ENT の比較

ここでは，TF-IDF と TF-ENT の 2 つの手法について，個々の商品カテゴリに対して抽出されたキーワードを比較する．表 5.8 の実験結果では，TF-IDF は TF-

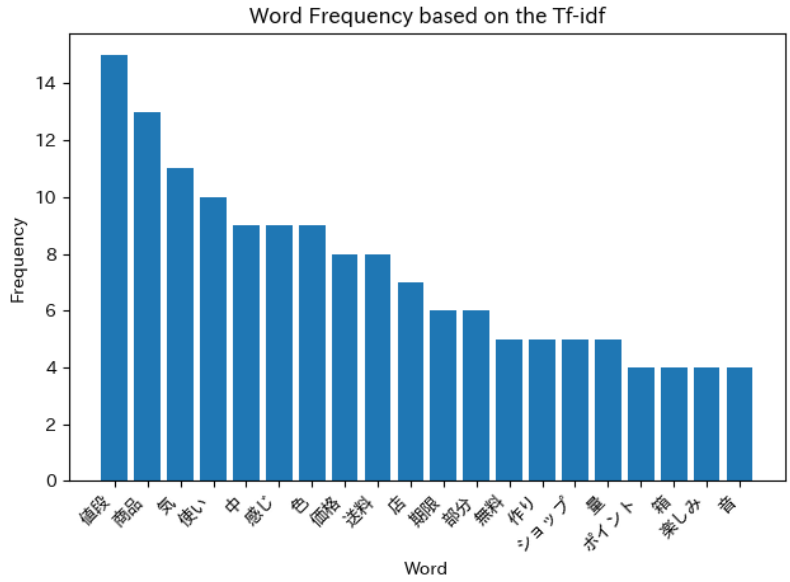

図 5.1: TF-IDF によって抽出されたキーワードとその出現頻度

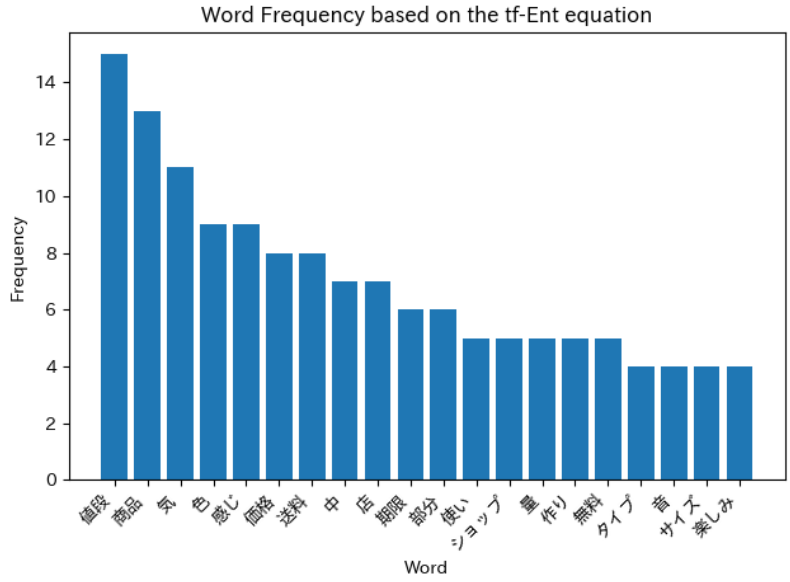

図 5.2: TF-ENT によって抽出されたキーワードとその出現頻度

ENT よりも高い F1 スコアを示したが，実際に抽出されたキーワードが本当に商品に関連しているかどうかという観点から，両者の違いを検証する．比較の対象とする商品カテゴリは,「本・雑誌・コミック」「パソコン・周辺機器」「スイーツ・お菓子」「車用品・バイク用品」「食品」の 5 つとする．

表 5.9 は，5 つの商品カテゴリに対して TF-IDF および TF-ENT によって抽出された出現頻度上位 20 個のキーワードを表示している．商品カテゴリーとあまり関係のないキーワードを太字で示す．TF-ENT によって抽出された単語は，TF-IDF

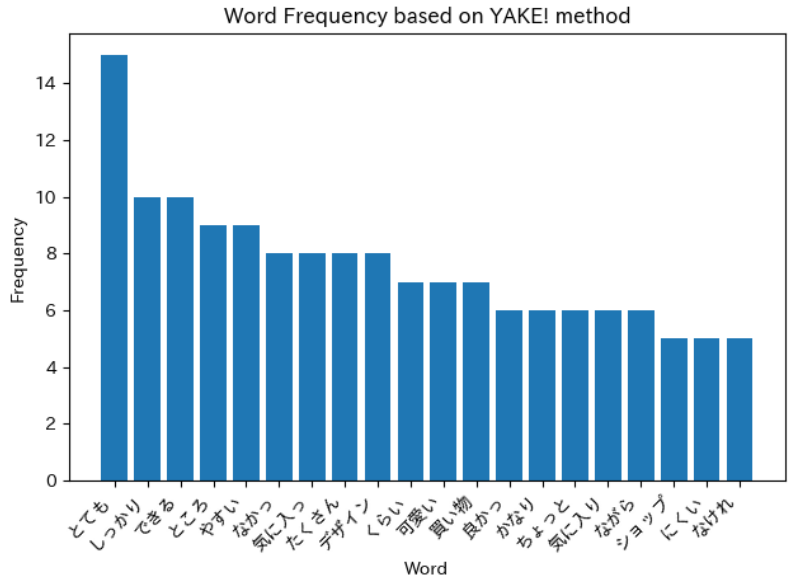

図 5.3: YAKE!によって抽出されたキーワードとその出現頻度

よりも商品カテゴリに特有のものが多い．一方，TF-IDF によって抽出されたキーワードには「他」「商品」「価格」「送料」「感じ」など，他のカテゴリでも用いられるようなキーワードも多く含まれる．このような単語は，IDF もしくは ENT の値が低くなり，重要なキーワードとして抽出されないことが期待されるが，TF-IDF よりも TF-ENT の方がそのような単語が少ないことから，ENT は IDF よりも優れた指標であると言える．FI スコアによる評価では TF-ENT は TF-IDF に劣るが，カテゴリに固有の単語を抽出するという観点では，本研究で提案した TF-ENT は TF-IDF より優れていると言える．

表 5.9: TF-IDF と TF-ENT によって抽出されたキーワードの例

商品カテゴリ	TF-IDF	TF-ENT
本・雑誌・コミック	マンガ，**息子**，**楽しみ**，表紙，**娘**，シリーズ，全巻，商品，書店，**感じ**，作者，本屋，**自分**，漫画，**人**，写真，作品，**子供**，内容，本	ストーリー，主人公，家計，数学，コミック，続き，本書，新刊，ブックス，著者，内容，古本，作品，マンガ，全巻，書店，漫画，作者，本屋，本
パソコン・周辺機器	HDD，本体，製品，**店**，メーカー，LAN，サイズ，**他**，**気**，ケーブル，無線，USB，**感じ**，音，PC，パソコン，**送料**，**価格**，**値段**，**商品**	Edy，**感じ**，MB，インストール，音，**送料**，SATA，edy，HDD，価格，ケーブル，キーボード，LAN，SSD，パソコン，**値段**，PC，USB，無線，**商品**
スイーツ・お菓子	お菓子，ケーキ，**家族**，**得**，抹茶，レンジ，量，食，キャラメル，**子供**，安納，**値段**，リピ，**送料**，**感じ**，チーズ，**商品**，焼き芋，味，芋	**送料**，**ベチャ**，食，**感じ**，リピ，スイートポテト，さつま芋，いも，**商品**，カチョカバロ，キャラメル，キャラメルト，**花畑**，チーズ，安納，味，カチョカヴァロ，アートピアミルク，焼き芋，芋
車用品・バイク用品	耐久，バイク，シート，ワッペン，**他**，LED，**雨**，**価格**，色，HID，取り付け，ランプ，ポジション，**感じ**，サイズ，LED，車，自転車，**値段**，**商品**	サイズ，車，取り付け，**横殴り**，KINGWOOD，自転車，LED，**値段**，Gear，ダッツ，ポジション，ヴォクシー，SMD，矢崎，ひったくり，HID，ダッシュウ，アストンマーティン，ユアーズ，**商品**
食品	**価格**，リピート，スーパー，**自分**，香り，**家族**，甘酒，酒粕，**得**，**無料**，**量**，**子供**，**他**，リピ，**感じ**，**店**，**値段**，**送料**，**商品**，味	高島屋，歳暮，ラー油，**店**，こんにゃく，粕汁，イセエビ，味噌汁，メンマ，**値段**，羊羹，桃屋，酒粕，豚まん，**送料**，松阪，**商品**，ごま油，とらや，味

第6章　おわりに

6.1　まとめ

本論文では，商品に対する意見とその根拠を含むという観点からレビューの有用性を判定する新たなアプローチを提案した．まず，レビュー中の節の組に意見とその根拠が含まれているかどうかを分類する「根拠関係分類タスク」を定義した．また，意見と根拠を含む節ペアに対し，その意見が商品に対するものであるかを分類する「商品言及分類タスク」を定義した．そして，これらのタスクを解決する手法を探究した．

「根拠関係分類タスク」については，このタスクの分類器を学習するために，3つの異なるデータセットを構築した．(1)「原因/理由」の関係にある節の組を正例として利用する KWDLC データセット，(2) 談話標識を手がかりに構築されたデータセット，(3)ChatGPT によるデータ拡張によって構築されたデータセット，の 3 つであった．これらのデータセットから根拠関係分類器を学習する手法として，ルールベースの手法，BERT をファインチューニングする手法，Intermediate Fine-Tuning によって BERT を学習する手法，ハイブリット手法，の 4 つを提案した．

実験の結果，既存の人手で注釈されたアウト・ドメインデータセット（KWDLC）と，自動的に構築されたイン・ドメインデータセット (談話標識によるデータセット，ChatGPT によるデータセット) を組み合わせることで，根拠関係分類タスクの F1 スコアが向上した．最も優れた F1 スコアは，3 つのデータセットを用いて Intermediate Fine-Tuning より学習された BERT モデルによって得られ，その値は 0.71 であった．これは，KWDLC だけを用いて学習されたモデルよりも 0.09 ポイント高かった．

「商品言及分類タスク」については，商品カテゴリ毎に固有のキーワードを抽出し，そのキーワードが節の組に出現するか否かによって商品への言及の有無を判定する手法を提案した．キーワードを抽出する手法として，単語が複数のカテゴリに出現する確率分布のエントロピーによってキーワードの重要度を評価する手法 (TF-ENT) を提案した．さらに，TF-IDF，TF-ENT，YAKE!，TF-IDF と YAKE!の組み合わせ (TF-IDF+YAKE!)，TF-ENT と YAKE!の組み合わせ (TF-ENT+YAKE!)，の 5 つの手法を実装し，これらを比較した．

実験の結果，既存の統計ベースのキーワード抽出手法である TF-IDF と YAKE!よりも，2 つの手法を組み合わせる TF-IDF+YAKE!と TF-ENT+YAKE! の方が F1

スコアが高いことが示された．F1 スコアが最も高かったのは TF-IDF+YAKE!であり，その値は 0.89 であった．これは，単一のキーワード抽出方法である TF-IDF ならびに YAKE!と比べて，それぞれ 0.09 ならびに 0.03 ポイント高かった．

また，5 つの商品カテゴリについて，TF-IDF と TF-ENT により抽出されたキーワードが商品カテゴリに関連しているかどうかを検証した．TF-ENT によって抽出されたキーワードは，TF-IDF よりも特定の商品カテゴリに固有のものが多いことがわかった．また，TF-IDF によって抽出されたキーワードには他のカテゴリでよく使われている単語も多かった．これらの結果から，TF-ENT は TF-IDF と比べて F1 スコアは低いものの，カテゴリに固有のキーワードを抽出するのに適していることがわかった．

6.2 今後の課題

評価実験では提案手法の有効性が確認できたものの，改善が必要な点もいくつか見つかった．これを踏まえ，今後の課題を以下にまとめる．

訓練データの拡張 本研究では，談話標識と ChatGPT を使用してデータセットを構築し，根拠関係分類器の性能を向上させた．しかし，より多様な表現を捉えるために，訓練データを増やし，意見と根拠の関係を表す言語表現が豊富に含まれたコーパスを構築する必要がある．これらの手法では，負例は節の組をランダムに組み合わせることで作成しているが，明らかに無関係な節の組を訓練データに加えることは必ずしも妥当ではないため，より良い負例の作成方法も検討するべきである．

モデルの改善 今回の実験では，Intermediate Fine-Tuning が最も高い F1 スコアを達成したが，モデルの更なる改善が必要と考えられる．例えば，対照学習の手法を導入することで，関連する複数のタスク間の類似性を学習したり，メタ学習の一種である Prototypical Networks を適用することで，未知の節ペアに対するモデルの分類性能を向上させたりすることが，検討課題として挙げられる．

文埋め込みベースの抽出手法の利用 本論文では，商品言及分類タスクについて，統計ベースの手法により抽出されたキーワードを利用して商品への言及の有無を判定した．しかし，商品カテゴリに特有のキーワードや表現が必ずしも抽出できていないという課題も明らかになった．この問題に対処するために，今後は SIFRank などの文埋め込みベースの手法を利用してキーワードを抽出することを検討する．これにより，文埋め込みが持つ多様な特徴量を利用し，商品カテゴリにより深く関連する多様なキーワードやキーフレーズを抽出することが期待できる．

Publication

Po-Min Chuang, Kiyoaki Shirai, and Natthawut Kertkeidkachorn. "Identification of Opinion and Ground in Customer Review Using Heterogeneous Datasets". To appear in The 16th International Conference on Agents and Artificial Intelligence (ICAART 2024), 2024, Feb.

謝辞

本研究を行うにあたり，主指導教員である白井清昭准教授には，丁寧かつ熱心なご指導を頂くなど大変お世話になりました．この場を借りて深謝の意を表します．

同研究室指導教員である Natthawut Kertkeidkachorn 講師には，本論文の作成にあたり，副査として適切なご助言を賜りました．この場を借りて感謝申し上げます．

また，本研究を進めるにおいて多くの意見をくださった白井研究室のメンバーには，本研究の遂行にあたり多大なご助言，ご協力頂きました．この場を借りて誠意の意を表します．

参考文献

[1] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 2623–2631, 2019.

[2] Hana Almagrabi, Areej Malibari, and John McNaught. Corpus analysis and annotation for helpful sentences in product reviews. *Computer and Information Science*, Vol. 11, No. 2, pp. 76–87, May 2018.

[3] Sanjeev Arora, Yingyu Liang, and Tengyu Ma. A simple but tough-to-beat baseline for sentence embeddings. In *International conference on learning representations*, 2017.

[4] Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Mário Jorge, Célia Nunes, and Adam Jatowt. Yake! collection-independent automatic keyword extractor. In *Advances in Information Retrieval: 40th European Conference on IR Research, ECIR 2018, Grenoble, France, March 26-29, 2018, Proceedings 40*, pp. 806–810. Springer, 2018.

[5] Lynn Carlson, Daniel Marcu, and Mary Ellen Okurovsky. Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory. In *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue*, 2001.

[6] Cemil Cengiz, Ulaş Sert, and Deniz Yuret. 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*, pp. 427–436, Florence, Italy, August 2019. Association for Computational Linguistics.

[7] Haixing Dai, Zhengliang Liu, Wenxiong Liao, Xiaoke Huang, Yihan Cao, Zihao Wu, Lin Zhao, Shaochen Xu, Wei Liu, Ninghao Liu, et al. AugGPT: Leveraging ChatGPT for text data augmentation. *arXiv preprint arXiv:2302.13007*, 2023.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[9] Gerardo Ocampo Diaz and Vincent Ng. 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)*, pp. 698–708, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[10] Iftah Gamzu, Hila Gonen, Gilad Kutiel, Ran Levy, and Eugene Agichtein. 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*, pp. 678–691, Online, June 2021. Association for Computational Linguistics.

[11] Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. Chatgpt outperforms crowd-workers for text-annotation tasks. *arXiv preprint arXiv:2303.15056*, 2023.

[12] Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 168–177, 2004.

[13] Najoung Kim, Song Feng, Chulaka Gunasekara, and Luis Lastras. Implicit discourse relation classification: We need to talk about evaluation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5404–5414, Online, July 2020. Association for Computational Linguistics.

[14] Soo-Min Kim, Patrick Pantel, Tim Chklovski, and Marco Pennacchiotti. Automatically assessing review helpfulness. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pp. 423–430, Sydney, Australia, July 2006. Association for Computational Linguistics.

[15] Hirokazu Kiyomaru and Sadao Kurohashi. Contextualized and generalized sentence representations by contrastive self-supervised learning: A case study on discourse relation analysis. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell,

Tanmoy Chakraborty, and Yichao Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 5578–5584, Online, June 2021. Association for Computational Linguistics.

[16] Jingjing Liu, Yunbo Cao, Chin-Yew Lin, Yalou Huang, and Ming Zhou. 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)*, pp. 334–342, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[17] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[18] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

[19] Rada Mihalcea and Paul Tarau. TextRank: Bringing order into text. In Dekang Lin and Dekai Wu, editors, *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pp. 404–411, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[20] Susan M Mudambi and David Schuff. Research note: What makes a helpful online review? a study of customer reviews on amazon.com. *MIS quarterly*, pp. 185–200, 2010.

[21] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking : Bringing order to the web. In *The Web Conference*, 1999.

[22] Yue Pan and Jason Q Zhang. Born unequal: a study of the helpfulness of user-generated product reviews. *Journal of retailing*, Vol. 87, No. 4, pp. 598–612, 2011.

[23] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In Marilyn Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,*

Volume 1 (Long Papers), pp. 2227–2237, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[24] Clifton Poth, Jonas Pfeiffer, Andreas Rücklé, and Iryna Gurevych. What to pre-train on? Efficient intermediate task selection. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 10585–10605, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[25] Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. The Penn Discourse TreeBank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May 2008. European Language Resources Association (ELRA).

[26] Rashmi Prasad, Bonnie Webber, and Alan Lee. Discourse annotation in the PDTB: The next generation. In *Proceedings 14th Joint ACL - ISO Workshop on Interoperable Semantic Annotation*, pp. 87–97, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics.

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

[28] Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. How to fine-tune BERT for text classification? In *Chinese Computational Linguistics: 18th China National Conference, CCL 2019, Kunming, China, October 18–20, 2019, Proceedings 18*, pp. 194–206. Springer, 2019.

[29] Yi Sun, Hangping Qiu, Yu Zheng, Zhongwei Wang, and Chaoran Zhang. Sifrank: a new baseline for unsupervised keyphrase extraction based on pre-trained language model. *IEEE Access*, Vol. 8, pp. 10896–10906, 2020.

[30] Tomoko Uchida. 日本語形態素解析器 janome. `https://mocobeta.github.io/janome/`, 2015. (2024 年 1 月閲覧).

[31] Oren Tsur and Ari Rappoport. Revrank: A fully unsupervised algorithm for selecting the most helpful book reviews. *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 3, No. 1, pp. 154–161, Mar. 2009.

[32] Nobuhiro Ueda, Kazumasa Omura, Takashi Kodama, Hirokazu Kiyomaru, Yugo Murawaki, Daisuke Kawahara, and Sadao Kurohashi. KWJA: A unified Japanese analyzer based on foundation models. In Danushka Bollegala,

Ruihong Huang, and Alan Ritter, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pp. 538–548, Toronto, Canada, July 2023. Association for Computational Linguistics.

[33] Yinfei Yang, Yaowei Yan, Minghui Qiu, and Forrest Bao. 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)*, pp. 38–44, 2015.

[34] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. XLNet: Generalized autoregressive pre-training for language understanding. In *NIPS'19: Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pp. 5753–5763, 2019.

[35] Feng Zhu and Xiaoquan Zhang. Impact of online consumer reviews on sales: The moderating role of product and consumer characteristics. *Journal of marketing*, Vol. 74, No. 2, pp. 133–148, 2010.

[36] 岸本裕大, 村脇有吾, 河原大輔, 黒橋禎夫. 日本語談話関係解析: タスク設計・談話標識の自動認識・コーパスアノテーション. 自然言語処理, Vol. 27, No. 4, pp. 889–931, 2020.

[37] 京都大学. 日本語構文・格・照応解析システム KNP. `https://nlp.ist.i.kyoto-u.ac.jp/?KNP`, 2012. (2024 年 1 月閲覧).

[38] 東北大学. BERT-base-japanese-model-v2. `https://huggingface.co/cl-tohoku/bert-base-japanese-v2`, 2021. (2023 年 10 月閲覧).