

複数レビューのアスペクトベースセンチメント分析に基づく アスペクト毎の要約文の生成

Summary Generation for each Aspect Category based on Aspect-based Sentiment Analysis of Multiple Reviews

辻谷 千春¹⁾ 渥美 雅保¹⁾
Chiharu Tsujitani Masayasu Atsumi

1 はじめに

EC サイトにおけるレビューは消費者による投稿からなり、多様な内容が存在する。そのため、消費者が商品選択や供給者が商品分析をする際に、膨大なレビューを効果的に利用することは困難である。本研究では、アスペクトベースセンチメント分析モデル [1] を使用してレビューをアスペクトカテゴリとセンチメントに分類し、それらのレビュー群を、大規模言語モデルを用いて要約する手法を提案する。既存のデータセットには、センチメント分析と要約の両方を満たす教師データがラベル付けされたものが存在せず、既存の手法を適用することができない。要約の自動評価指標として ROUGE[2] や BLEU[3] などがあるが、これらは参照要約を必要とするので、参照要約が存在しないデータセットでは評価が不可能である。そこで、新たな評価手法を設定する必要がある。本研究では、参照要約を必要としない新たな評価手法として、レビューのアスペクトカテゴリ分類に使用したモデルを活用し、要約文と元の文章で同じ分類ができるかどうかを測る手法を提案する。また、レビューは、消費者側の観点から述べられたものが主であるが、消費者のみならず、供給者にとっても重要な情報ソースである。そこで、消費者の観点からの要約に加えて、供給者の観点からの要約を行うことで要約生成の観点適合性も評価する。

2 関連研究

アスペクトベースセンチメント分析 (ABSA) には、アスペクトカテゴリの検出、ターゲットフレーズの同定、極性の推定の 3 つのタスクが存在する。例えば、"This apple is delicious." という文であれば、Food#Quality というアスペクトカテゴリに分類され、そのターゲットは apple であり、Positive な極性文であることが推定される。ここで、アスペクトカテゴリの Food#Quality において、Food をエンティティ、Quality をアトリビュートと呼ぶ。ABSA のアプローチとして、テキストにターゲットとアスペクトカテゴリのセットを補助文として挿入して入力し、極性とアスペクトカテゴリを分類する手法 [4] や、テキストにアスペクトカテゴリとマスクトークンで構成された文を挿入して入力し、そのカテゴリに対する極性の単語を予測する手法 [5] などが挙げられる。

要約には、抽出要約と生成要約の 2 つのアプローチが存在する。抽出要約は、文章から重要とされる文を抽出し、それを要約として扱う手法である。具体的には、文章を句点などの区切り文字で分割し、機械学習を用いて重要文の判定を行う。生成要約は、文章全体の意味を把握し、自然な要約を生成する手法である。代表的な生成要約モデルには、PEGASUS[6] などがある。

1) 創価大学大学院理工学研究科情報システム工学専攻

Encoder-Decoder 型のモデルであり、入力文を Encoder 部でエンコードし、Decoder 部で要約文の生成を行う。抽出要約の特徴としては、要約の内容が正確であるが、文の接続が不自然であり、出力される要約の長さの調整が難しいという点が挙げられる。生成要約の特徴としては、幅広い単語を使用して自然な要約が生成できるメリットがある一方で、破綻した文章や不必要な情報を生成するといったデメリットも存在する。大規模言語モデルとして、最近では ChatGPT(GPT-3.5[7], GPT-4[8]), LLaMA[9] といった数百億から一十億以上のパラメータを持つ大規模テキスト生成モデルが台頭してきており、生成要約でも高い性能を発揮することが確かめられている。本研究では、ABSA と大規模言語モデルを用いた生成要約を組み合わせている点、及び要約の教師データを必要とせずに評価を行う点が新規な点である。

3 アスペクトベースセンチメント要約モデル

本研究で提案するモデルの構成を図 1 に示す。本モデルは、アスペクトベースセンチメント分析モデル (ABSA モデル)、文選択アルゴリズム、要約モデル、及び要約評価モデルの 4 つのサブモデルに分かれている。大まかな流れとして、まず学習済みの ABSA モデルを用いてレビュー文集合をアスペクトカテゴリとセンチメントに分類する。次に、アスペクトカテゴリのセンチメント毎にそれらに分類されたレビュー文を既定の数だけ抽出し、要約モデルに渡す。要約モデルに渡されたレビュー文集合は、1 つの文章に要約され、要約の評価モデルへと渡される。評価モデルでは、学習済み ABSA モデルを用いて、要約文のアスペクトカテゴリとセンチメント極性を推定する。この推定値が、入力に用いた要約文集合のアスペクトカテゴリとセンチメント極性と同じものであるかを判定することで、モデルの評価を行う。

3.1 アスペクトベースセンチメント分析モデル

ABSA モデルとして、三浦ら [1] が提案したモデルを本研究に向けて改良して使用する。このモデルは、テキストの入力のみで 3 つの ABSA タスクを解決できる点がメリットである。モデルは、アスペクトカテゴリ分類ネットワークと、アスペクトセンチメント分類ネットワークの 2 つからなる。アスペクトカテゴリ分類ネットワークでは、入力文をトークンに変換し、BERT[10] によるエンコードを行い、エンコードされた特徴ベクトルを用いて、アスペクトカテゴリの分類を行う。もう一方のアスペクトセンチメント分類ネットワークでは、アスペクトカテゴリ分類ネットワークから得たアスペクトカテゴリと単語特徴ベクトル列を入力に用いる。Transformer エンコーダ層のアテンション機構に基づき推論を行い、ターゲットフレーズの位置と極性を推定する。本研究では、アスペクトカテゴリ分類と極性推定のみを用いるので、ターゲットフレーズの位置推定は利用しない。

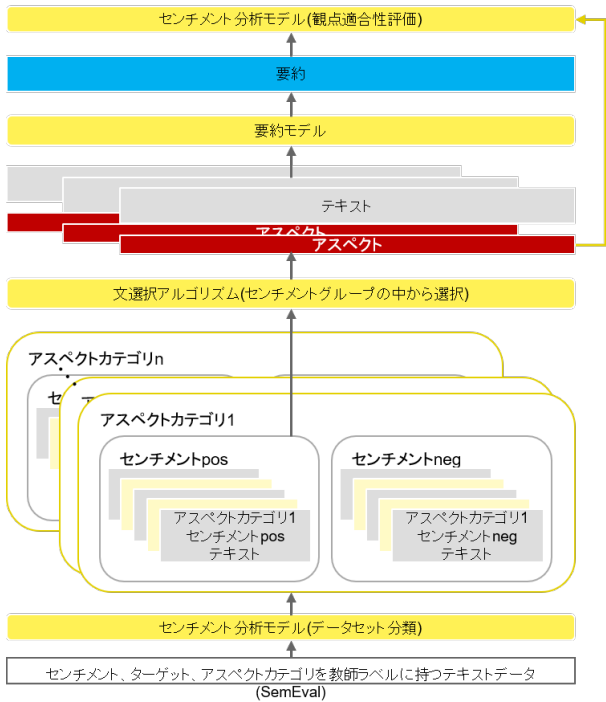


図 1 モデル構成

3.2 文選択アルゴリズム

ABSA モデルを用いて分類されたレビュー文集合の中から、アスペクトカテゴリのセンチメント毎に既定の数のレビュー文を選択し、要約モデルに適合する形式に変換する部分である。本研究では、ランダムにレビュー文を選択し、要約モデルへと受け渡す。

3.3 要約モデル

要約モデルには ChatGPT の GPT-3.5 モデルを用いる。文選択アルゴリズムでランダムに選択されたレビュー文集合を一つのテキストとして入力し、要約を出力する。このモデルは、プロンプトによって要約以外にも様々なタスクを解決することが可能なため、本研究では要約を出力するプロンプトを最初に入力している。

3.4 評価モデル

評価モデルには、システムに一貫性を持たせるために、分類で用いた ABSA モデルを使用する。そして、要約文を入力してそのアスペクトカテゴリと極性を求める。評価モデルが出力した要約文のアスペクトとセンチメントが、要約モデルの入力レビュー文集合のアスペクトとセンチメントと同じものかを評価するモデルである。

4 実験

4.1 実験構成

実験では、ABSA モデルによるレビュー文のアスペクトカテゴリとセンチメント極性の分類性能の評価と、文選択アルゴリズムにより得られたアスペクトと極性毎のレビュー文集合をプロンプトと共に要約モデルに入力して生成した要約文の評価を行う。また、供給者の観点からの要約の観点適合性を評価するために、要約モデルに与えるプロンプトを変更して要約を生成し、得られた要約文の評価も行う。

4.2 データセット

本研究では、Semeval 2016 task 5:Aspect Based Sentiment Analysis[11] の Restaurant ドメインのデータセッ

トを用いて実験を行った。このデータセットは、訓練データ 2000 文、テストデータ 676 文で構成されており、それぞれの文にアスペクトカテゴリ、極性、ターゲットの位置がアノテーションされている。アスペクトカテゴリとしては、12 種類のうちのいずれか 1 つがラベル付けされている。12 種類のアスペクトカテゴリを表 1 に示す。また、極性には、Positive, Negative, Neutral の 3 つのうちいずれかがアノテーションされている。データセットの中には、ラベル付けされていないテキストのみのデータも存在し、それらを除くと、訓練データ 1708 文、テストデータ 587 文となる。また、本研究では、要約の良し悪しをはっきりさせやすくするために、Positive, Negative がアノテーションされている文章のみを使用した。

表 1 アスペクトカテゴリー一覧

FOOD#PRICES
FOOD#QUALITY
FOOD#STYLE_OPTIONS
DRINKS#PRICES
DRINKS#QUALITY
DRINKS#STYLE_OPTIONS
SERVICE#GENERAL
AMBIENCE#GENERAL
LOCATION#GENERAL
RESTAURANT#GENERAL
RESTAURANT#PRICES
RESTAURANT#MISCELLANEOUS

4.3 ABSA モデルによる分類実験

表 2 に、Semeval データセットを用いて学習した際のハイパーパラメータを示す。また、表 3 にアスペクトカテゴリと、センチメント極性の分類結果を示す。評価には、precision, recall, f 値を使用している。また、support は文章の数を表している。

表 2 ABSA モデルのハイパーパラメータ

ハイパーパラメータ	値
センチメント分類 Transformer 層数	1
センチメント分類 TransformerFFN 隠れ層次元数	2048
センチメント分類 Transformer 内ドロップアウト率	0.1
Attention ヘッド数	8
特徴量次元数	768
活性化関数	ReLU
最適化関数	Adam
損失関数	CrossEntropyLoss
バッチサイズ	8
エポック数	200
学習率	0.001

表 3 ABSA モデル分類結果

	precision	recall	f1	support
アスペクトカテゴリ分類性能	0.87	0.81	0.84	1708
センチメント分類性能	0.93	0.93	0.93	1708

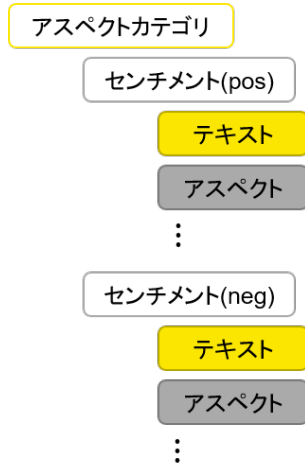


図 2 データセットの構成

4.4 要約文評価実験

ABSA モデルによりアスペクトカテゴリとセンチメントごとに分類したレビュー文を、文選択アルゴリズムによって選択し要約に使用するために、図 2 のような構成のデータセットに配置する。

この要約実験では、データセットからアスペクトカテゴリとセンチメントごとに 10 文ずつ選択して、ChatGPT により要約を行う。10 文に満たない文章群については、その数だけで要約を生成する。要約生成の際に、要約の命令の旨を示すプロンプトを先頭に付加して要約を行う。また、プロンプトの違いによる影響を調べるために、アスペクトカテゴリとセンチメントの情報を与えるプロンプトと与えなかったプロンプトで比較を行った。それらプロンプトを表 4 に示す。表 4(2) のプロンプトの 1 つ目の波括弧にはセンチメント、2 つ目の波括弧にはアトリビュート、3 つ目の波括弧にはエンティティがそれぞれ記される。

表 4 各実験のプロンプト

タイプ	プロンプト文
(1) 要約命令文	Generate a summary for this review.
(2) 要約命令文と入力文の説明	These are reviews with {POSITIVE} opinions about the {QUALITY} of the {FOOD} the restaurant offers. Generate a summary for this review.
(3) 供給者視点での要約命令文	Generate a summary for this review based on the perspective of the store's management's.

これらの条件の下、1708 文の入力文から 177 文の要約文が生成された。要約文が元の入力文と同じアスペクトカテゴリとセンチメントを表現できているかを調べるために、分類実験で使用した ABSA モデルを用いて、入力

文と同じアスペクトカテゴリ、センチメントに分類できるかどうかを評価した。その結果を表 5、表 6 に示す。実験結果から、アスペクトカテゴリ分類において、表 7、表 8 に示すように、プロンプトに入力文の説明を与えない要約の方が、入力文の説明を与えた要約に比べて高い分類性能が達成された。一方、センチメント分類性能については、プロンプトに入力文の説明を与えても与えなくても同じ性能を示した。ABSA により分類されるレビュー文には、正しくないアスペクトカテゴリとセンチメントに分類されるものが存在し、それらが要約文生成の性能を下げていると考えられる。そこで、アスペクトカテゴリとセンチメント極性分類の両方で正解した文章のみで要約を生成し、評価をする追加実験を行った。その結果、表 7、表 8 の (3) に示すように、アスペクトカテゴリとセンチメントの分類性能が、全文を使う場合の分類性能と比べて僅かではあるが上回った。

表 5 要約実験結果 (アスペクトカテゴリ分類)

プロンプトタイプ	入力文タイプ	precision	recall	f1	support
(1) 要約命令文	全文	0.84	0.80	0.80	177
(2) 文情報と要約命令文		0.75	0.72	0.72	177
(3) 要約命令文	分類成功文	0.86	0.79	0.80	136

表 6 要約実験結果 (センチメント分類)

プロンプトタイプ	入力文タイプ	precision	recall	f1	support
(1) 要約命令文	全文	0.97	0.97	0.97	177
(2) 文情報と要約命令文		0.97	0.97	0.97	177
(3) 要約命令文	分類成功文	0.98	0.98	0.98	136

4.5 観点変更実験

要約文生成の観点を供給者の視点に変えた際の要約の観点適合性を調べるために、供給者の観点をプロンプトに設定する実験を行った。4.4 の要約文評価実験の結果から、入力文に説明を与えずに要約を生成した方が適切な要約が生成されることがわかったので、供給者の観点から要約する旨の命令の文章を、プロンプトの先頭に付加して要約を行った。具体的なプロンプトを、表 4 に示す。その結果を表 7、表 8 に示す。結果として、センチメント分類性能は、供給者の観点とすることで少し下がる結果となり、アスペクトカテゴリの分類性能についても低下した。

5 考察

要約文評価実験では、アスペクトカテゴリ分類性能において、プロンプトに要約の命令と入力文の説明を与えて生成した方が、アスペクトカテゴリとセンチメントを正しく反映した要約が生成されにくいことがわかったが、これについては、文章にバイアスがかかってしまい、かえって要約の自由度を狭めてしまう結果になってし

表7 アスペクトカテゴリ分類結果

プロンプトタイプ	入力文タイプ	precision	recall	f1	support
要約命令文	全文	0.84	0.80	0.80	177
供給者視点での要約の命令		0.76	0.69	0.68	178

表8 センチメント分類結果

プロンプトタイプ	入力文タイプ	precision	recall	f1	support
要約命令文	全文	0.97	0.97	0.97	177
供給者視点での要約の命令		0.94	0.94	0.94	178

まったのではないかと考えられる。プロンプトの設定については、大規模言語モデル全体の課題でもあるので、より適切な要約を生成できるようなプロンプトを模索したい。センチメント分類結果では、いずれの実験結果においても元文の分類結果(表3)を上回る結果となった。これは、同じセンチメントであると分類された文章を集めて要約しているため、そのセンチメントに分類されやすい表現が選択されているのではないかと考える。また、正誤に関わらず分類されたレビュー文群から要約した結果と、分類に成功したレビュー文のみから要約した結果の比較において、後者の結果がわずかに上回る結果となった。分類に成功している文のみで要約を行なっているため、よりアスペクトカテゴリやセンチメントを正しく反映した要約が生成されていると考えられる。観点変更実験では、供給者の視点から要約を生成したが、この要約の性能は、要約文評価実験の結果に比べて、観点到適合するとは言いえない結果となった。本データセットは、消費者のレビューデータを元としているので、供給者の視点の教師となるようなデータが少ないことが、観点到に沿わない原因の一つとして挙げられる。

6 結論

本研究では、アスペクトベースセンチメント分類モデルと要約モデルを組み合わせて、アスペクトセンチメント毎の要約生成の評価を行った。具体的には、ABSAモデルを用いて Semeval 2016 task 5 restaurant ドメインデータセットの分類性能を評価し、その分類結果を元に、文選択アルゴリズムを使用して要約を生成して、要約の評価、及び要約が観点到に適合するかの評価を、同じ分類モデルを用いて行った。その結果、要約の命令のみをプロンプトに与えることで、高い精度でアスペクトカテゴリやセンチメントに合致した要約が生成可能なことが確かめられた。一方、供給者の観点をプロンプトに加えた実験で

は、要約にそれがうまく反映されなかった。今後、要約の質と観点到依存性を高めるためのプロンプトの工夫などが課題である。また、正しい分類のレビュー文群からの要約が良い要約を生成するということがわかった。そのため、分類モデルの汎化性能向上のためのデータセットの拡充が求められる。今後の展望として、この成果を生かして、センチメント分類と要約を組み合わせた情報抽出や意思決定支援システムの構築が期待される。

参考文献

- [1] 三浦 義栄, 他. 自己注意ニューラルネットワークによる文に含まれる複数のアスペクトのセンチメント分析. 第82回全国大会講演論文集. 2020, vol.2020, no.1, p.511-512.
- [2] Chin-Yew Lin. ROUGE: A Package for Automatic Evaluation of Summaries. In Text Summarization Branches Out, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- [3] Kishore Papineni, et al. BLEU: a Method for Automatic Evaluation of Machine Translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- [4] Chi Sun, et al. Utilizing BERT for Aspect-Based Sentiment Analysis via Constructing Auxiliary Sentence 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 380–385, Minneapolis, Minnesota. Association for Computational Linguistics.
- [5] Jian Liu, et al. Solving Aspect Category Sentiment Analysis as a Text Generation Task. arXiv preprint arXiv:2110.07310, 2021.
- [6] Jingqing Zhang, et al. PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization arXiv preprint arXiv:1912.08777, 2019.
- [7] Wayne Xin Zhao, et al. A Survey of Large Language Models. arXiv preprint arXiv:2303.18223, 2023.
- [8] OpenAI. GPT-4 Technical Report. arXiv preprint arXiv:2303.08774, 2023.
- [9] Hugo Touvron, et al. LLaMA: Open and Efficient Foundation Language Models. arXiv preprint arXiv:2302.13971, 2023.
- [10] Jacob Devlin, et al. 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.
- [11] Maria Pontiki, et al. SemEval-2016 Task 5: Aspect Based Sentiment Analysis. In Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), pages 19–30, San Diego, California. Association for Computational Linguistics.