大規模言語モデルを用いたその場での要約に基づくレビュー探索インタフェース

藤井真梨乃† 河田 友香†† 山本 岳洋††

† 兵庫県立大学 社会情報科学部 〒 651–2197 兵庫県神戸市西区学園西町 8–2–1
†† 兵庫県立大学 大学院情報科学研究科 〒 651–2197 兵庫県神戸市西区学園西町 8–2–1
E-mail: †{fa20b075,ad22l019}@guh.u-hyogo.ac.jp, ††t.yamamoto@sis.u-hyogo.ac.jp

あらまし 本研究では，類似する商品レビューについて，大規模言語モデルを用いた要約をユーザに提示するシステムを提案する．たとえばドライヤーをインターネットで購入するときに「髪の毛がつやつやになりました！」というレビューを見て，類似する観点のレビューを探したくなるというように文章を読んで調べたい観点が思い浮かぶことがある．しかし，従来の検索手法である文字列マッチングでは，類義語を用いたレビューを検索することができないうえ，検索キーワードとして適切な単語がないという問題がある．そこで，商品レビューから抜粋した文章を入力とし，分散表現を用いて類似すると判定した商品レビューについて要約を行い，提示するシステムを提案する．大規模言語モデルを用いることにより，ドメインに縛られずに要約を行い，同じ観点のレビューを分かりやすく提示できると考えた．提案手法についてユーザ実験を行い，観点や意見の網羅性や結果の見やすさについてアンケートをとった結果，従来の文字列マッチングよりも回答者平均が上回り，結果の見やすさについては有意差が認められた．また，レビューの検索精度についても評価したところ，提案手法ではフレーズで検索を行ったときに文字列マッチングを用いた手法より F_1 値が高くなった．今後はユーザ実験の人数を増やし，プロンプトの改善を行っていく必要がある．

キーワード レビュー，LLM，インタフェース

1 はじめに

インターネットが発達してきた近年，EC サイトで商品を購入する人が増加している．経済産業省の調査によると，物販系分野の BtoC–EC 市場規模は，2013 年の 5 兆 9,931 億円から 2022 年の 13 兆 9,997 億円まで年々増加している [16]．EC サイトで商品を購入する判断材料の 1 つとして商品レビューがある．商品レビューを読むことで，実際の使用感や耐久性などを詳しく知ることができる．

レビューを読む際，読んでいる最中に調べたいことが思い浮かぶことがある．たとえば，新しくドライヤーを買おうとしているユーザが，気になっているドライヤーのレビューを読んでいる途中に,「可愛い色合い」というフレーズが目に留まったとする．このときユーザは「見た目」について調べると考えられ，ユーザは 2 通りの行動をとる．1 つ目は，レビューを最初から見直すという行動である．この場合，レビューの投稿数が非常に多いときには，ユーザが求めている情報を探すことに時間がかかるという問題がある．そのため，求めている情報が書かれたレビューを絞ることや，求めている情報を要約することが必要である．2 つ目は，キーワードで検索をするという行動である．このようなシステムでは文字列マッチングが一般的に用いられている．しかし,「見た目」と「デザイン」といった，同じ意味を表す単語が含まれているレビューが表示されないという問題がある．また，観点を簡潔な単語で表しにくい場合があるという問題もある．たとえば，ドライヤーにおいて「髪の毛がつやつやになりました！」と「(髪が) しっとりと落ち着いたのにビックリしました」という 2 つのレビューは,「髪質」のように同じ観点で書いてあるように見えるが，実際に文字列マッチングで検索するには限度がある．そのため，類似語や類似文でも検索結果に表示されるシステムが必要である．

そこで本研究では，ユーザがレビューを閲覧している途中で気になる観点や文章を抜粋し，分散表現を用いて抜粋した文章と類似する商品レビューを絞り，それらのレビューについて大規模言語モデルを用いた要約を行って提示するインタフェースを提案する．大規模言語モデルを用いることにより，データ量やパラメータ数が多いという特性からドメインを限定せずに要約を行うことが可能になるのではないかと考えた．このインタフェースを用いることで，ユーザは類似文を検索することや，検索結果で内容を一目見て把握することが可能になる．

提案手法において，従来の検索手法と比較して「要約の見やすさ及び精度」,「ユーザの興味に対する網羅性」,「観点の網羅性」,「意見の網羅性」,「結果の見やすさ」,「システムの使用難易度」,「全体的な満足度」の 7 項目にどのような差があるのかを把握するためにユーザ実験を行った．ユーザは 2 種類のドライヤーのレビューを，1 つは比較手法，もう 1 つは提案手法を用いて検索した．ユーザへのアンケートを用いて 7 項目をリッカート尺度で測定し，検索クエリの発行数及び検索にかかった時間なども取得した．

実験を行った結果，7 項目のうち「システムの使用難易度」を除いた 6 項目で，回答者平均が比較手法を上回った．しかし，有意差は認められなかった．また，検索クエリの発行数及び検

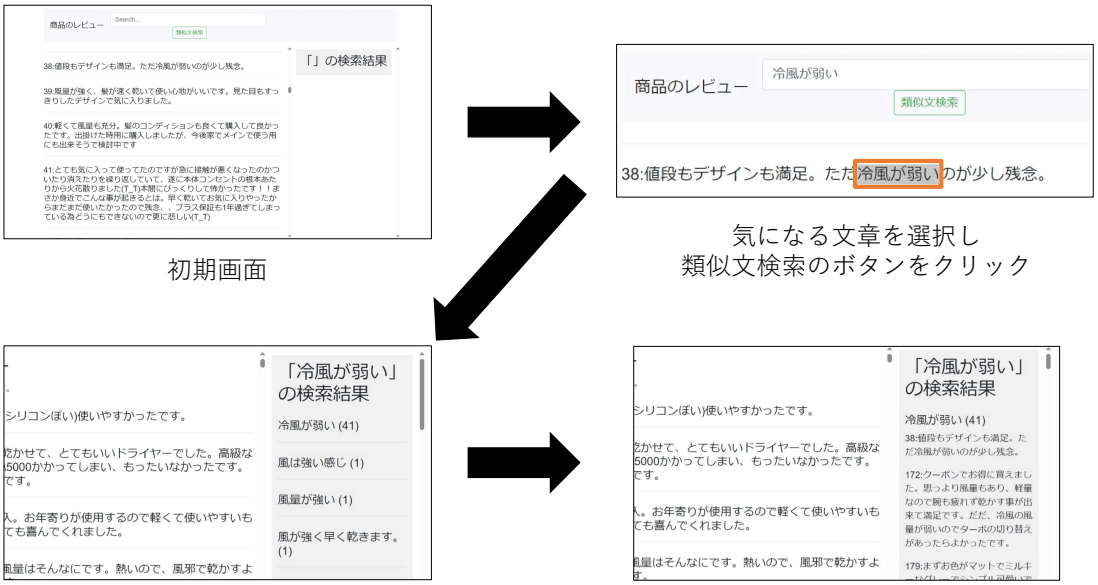

図 1　ユーザ側の利用の流れ.

索にかかった時間についても，有意差は認められなかった.

そして，どのような精度でレビューが検索結果に表示されているかを確かめるために，検索精度の評価も行った．実際の実験で用いられたクエリの中から，名詞 1 つで構成される単語 5 つと，名詞以外の品詞も含まれるフレーズ 5 つを選択して評価を行った．その結果，比較手法では単語での検索の方が高い F_1 値を示し，提案手法ではフレーズでの検索の方が高い F_1 値を示した.

2　関 連 研 究

2.1　商品レビュー分析

商品レビューを分析する研究は数多く存在する．Hu らは，商品の属性を抽出し，その特徴に対して肯定的，否定的な意見を WordNet を用いて収集した [3]．Poria らは，埋め込み表現を用いてレビュー中で述べられている観点を抽出した [10]．Jerripothula らは，商品レビュー及びそれに対する投票と商品の機能レベルへの点数から，商品の機能に対する評価点数を算出した [5]．Deng らは，製品に関する質問に対し，回答生成タスクと意見マイニングタスクをマルチタスク学習を行うことで，高い精度で回答を生成することを可能とした [2]．金兵らは，商品レビューから係り受け解析を用いて評価表現辞書を構築して特徴語を収集し，クラスタリングを行うことで評価軸の構築を行った [14]．渡辺らは，Word2Box を用いて出現単語間の領域の重なりから概念構造における上位語を選択し，教師なし学習で観点の付与を行った [19]．Nguyen らは，200 字以内の短いレビュー文の要約を行う際に，長いレビュー文が投稿されるレビューサイトを用いたアルゴリズムを作成した [9]．ホンらは，商品説明を用いて，レビュー文から教師なしでキーフレーズを抽出した [13]．Angelidis らは，弱い教師あり学習を行った観点抽出器と感情予測器を用いて商品レビューの要約を行った [1]．Liu らは，商品レビューの要約に ChatGPT を用いたとき，人手では他の手法と比較して高い評価を受けたことを報告した [8].

これらの研究では，レビュー中の観点の抽出やレビューの要約を行っているが，本研究では商品レビューを検索して要約を行うインタフェースという部分に焦点を当てる.

2.2　レビューを検索するシステム

商品レビューを検索するシステムについて研究が行われている．Liu らは，商品に対する肯定的，否定的な意見から属性を抽出し，その割合を可視化するシステムを提案した [7]．杉木らは，商品レビュー検索サイトにおいて自然言語表現で入力されたクエリと商品レビューに対し，感性表現シソーラスを用いて（項目，値）の組を抽出して類似度を求める検索方式を提案した [18]．栗原らは，映画レビューサイトに投稿されているレビュー文を，Doc2Vec を用いて映画ごとにベクトル化を行うことで，検索クエリから類似する評価表現を含むレビューを発見することができることを報告した [15].

また，検索を行うインタフェースについての研究も行われている．平山らは，商品レビューに形態素解析を用いて語の共起関係を抽出し，評価属性ごとの記述および関係性を可視化したインタフェースを作成することで，利用したユーザが商品購入に対して有益な文章を従来より多く発見できることを明らかにした [20]．市村は，係り受け解析を用いて飲食店のレビューから料理に関する文の要約を 3 種類行い，インタフェースを作

成して提示したところ，料理に関する抽出を行った 1 文をそのまま載せたシステムが最も満足度が高いことを報告した [17]. Jasim らは，肯定的，中立，否定的なレビューの訪問率を可視化してバイアスを減らすインタフェースを提案し，ユーザが自信を持った意思決定をより行いやすくなったことを報告した [4].

本研究は，市村 [17] のように要約を提示するインタフェースを作成するが，要約に大規模言語モデルを利用する部分が異なる．なお本研究では，先行研究 [15] [17] と同様に商品レビューに対してベクトル化を行い，類似度を求める検索方式を採用している．

2.3 大規模言語モデルを用いた検索システム

最近では，大規模言語モデルを用いた検索システムについての研究が行われている．Spatharioti らは，商品の特徴量の検索に対し，大規模言語モデルである ChatGPT を用いた対話型の検索インタフェースと従来のキーワード検索インタフェースを用いて比較した [12]．Jianqi は，文献検索において，自然言語で書かれた質問から大規模言語モデルを用いて検索クエリとなり得る単語を出力し，キーワード検索を行うことが有効であると報告した [6]．Pride らはそれに加え，質問に対する回答も大規模言語モデルで生成した [11].

要約結果を提示する検索インタフェースについての研究は存在するものの，商品レビューに対し，大規模言語モデルを用いてリアルタイムで要約を行っている検索インタフェースはまだない．そのため，本研究では大規模言語モデルを用いた要約を提示する部分に着目する．

3 その場での要約に基づくレビュー探索インタフェース

本節では，まず提案インタフェースにおけるユーザ側の利用手順を述べる．次に，提案インタフェースの概要を述べ，その後レビューの要約という段階について詳細に述べる．最後に，実際のシステムの動作例について述べる．以下では，1 ユーザが書いたレビュー文章全体を「レビュー」，その中の 1 文を「レビュー文」と定義する．

3.1 ユーザ側の利用手順

図 1 にユーザが実際にシステムを使用する時の流れを示す. まず，レビューが表示されたウェブサイト上で，ユーザが調べたい文章をドラッグして選択する．たとえば，ドライヤーのレビューを見ながら,「冷風が弱い」という文を見て他のレビューが気になったとき，ユーザはその文章を選択する．このとき，ユーザが選択した「冷風が弱い」という文章はウェブサイト上部の検索窓に自動的に入力される．次に，類似文検索ボタンを押すと，クエリである「冷風が弱い」と類似するレビューの要約が表示される．そして，要約された結果をクリックすると要約前のレビューが表示されるようになっている．これは，要約文のみだとレビューが異なる意味で解釈される可能性があるためである．

3.2 提案インタフェースの概要

提案インタフェースの概要を図 2 に示す．まず，あらかじめレビューを 1 文ずつ分散表現に変換する．1 文ごとに類似度を求めることにより，検索クエリにより近いレビュー文が結果に表示できると考えられる.「髪がつやつやになりました．デザインも可愛くて良かったです.」というレビューの場合,「髪がつやつやになりました.」と「デザインも可愛くて良かったです.」のように句点ごとに分割し，OpenAI の embedding 用モデルである text–embedding–ada–002 [1]を用いて各文のベクトル化を行う．このモデルを用いると，レビュー文 1 文を 1,536 次元のベクトルで表現することができる．次に，検索窓にクエリを入力する．このとき，入力するクエリはユーザがレビュー文から引用した文章とする．そして，クエリとレビュー文のコサイン類似度を求め，0.85 以上となる文が含まれているレビュー文を要約対象にする．そこから大規模言語モデルを用いて，対象となるレビュー文を要約する．最後に要約結果をユーザに提示する．

3.3 商品レビューの要約

類似度を求めることで得たレビュー文を，OpenAI の gpt–35–turbo–16k [2]を用いて要約した．このモデルは，プロンプトと呼ばれる自然言語による指示を入力すると，その回答を自然言語で行う大規模言語モデルである．

要約を行う理由は，最初に表示される画面で，レビュー文に書かれている評価を一括で見られるようにするためである．

要約には，few–shot プロンプトを用いる．few–shot プロンプトとは，プロンプトの中で入力と出力の例をいくつか示すことで，回答の精度を上げるプロンプトのことである．プロンプトは以下の通りである．

```
#説明文
以下で入力された文章を，出力形式に従って 10 文字程度で要約してください.
#入力
(類似度が 0.85 以上のレビュー文)
#入力例
(検索クエリが含まれているレビュー文)
#出力例
(検索クエリ)
#出力形式
10 文字程度の要約結果
```

本研究では，入力例として検索クエリの元となったレビュー文，出力例として検索クエリを与える．これにより，ファインチューニングなどで大量のデータを要することがない上，商品のドメインに縛られずに生成を行うことができる．

同一のレビュー文および検索クエリを入力したときに同じ結

1：https://openai.com/blog/new-and-improved-embedding-model
2：https://platform.openai.com/docs/models/gpt-3-5

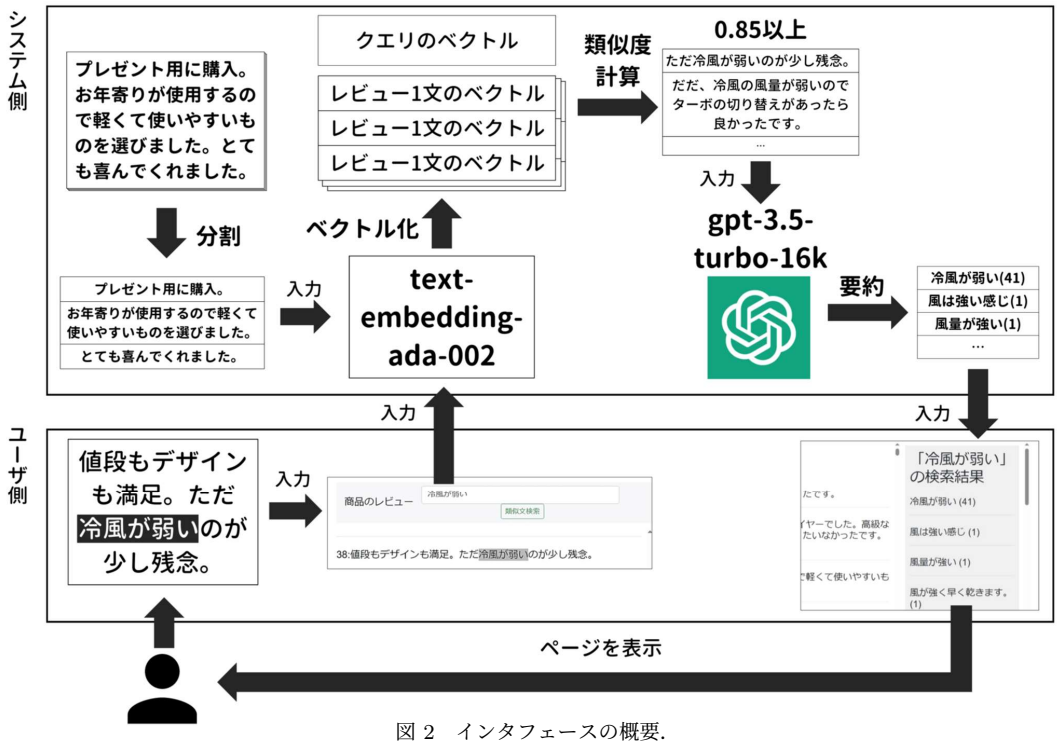

図 2　インタフェースの概要.

果を出力するために，値が高いほど回答が変化しやすくなるパラメータである temperature を 0 とする．

3.4　システムの動作例

本節では，実際のシステムの動作例を述べる．まず,「髪の毛の仕上がりも結構パサつく感じがあって子供のきれいな髪にはいいけど，傷んだ髪質には向かないなって思いました.」というレビュー文から「傷んだ髪質には向かない」というクエリが入力されたとする．このとき,「傷んだ髪質には向かない」との類似度が 0.85 以上になるレビュー文は 5 件存在する．このうち「乾かした時の髪の質ですが，微妙.」というレビュー文の要約を行う場合，以下のプロンプトがシステムに入力される．

> #説明文
> 以下で入力された文章を，出力形式に従って 10 文字程度で要約してください.
> #入力
> 乾かした時の髪の質ですが，微妙.
> #入力例
> 髪の毛の仕上がりも結構パサつく感じがあって子供のきれいな髪にはいいけど，傷んだ髪質には向かないなって思いました.
> #出力例
> 傷んだ髪質には向かない
> #出力形式
> 10 文字程度の要約結果

このプロンプトが実行され,「髪の質は微妙」という要約が出力される．今回は「傷んだ髪質には向かない」と類似するレビュー文が 5 個存在するために同様の処理が 5 回行われ，出力された要約とその頻度をまとめてユーザに提示する．

4　ユーザ実験

4.1　使用したデータ

本研究では，楽天市場 [3]で収集したドライヤー 2 商品のレビューデータを用いた．ドライヤーを用いた理由は，複数の類義語を持つ観点や，明確に言語化できない観点のレビューが多いと考えたからである．価格は 2023 年 12 月 11 日現在，それぞれ 5,918 円と 6,490 円であり，ほとんど同じ価格帯の商品である．今回は商品レビューの文章のみを使用しており，レビューのタイトルおよび評価点数は利用していない．また，訓練タスク用にオーブントースター 1 商品のレビューデータも使用した．いずれも 2023 年 7 月 26 日時点での最新 200 件を収集した．

4.2　実験参加者

実験参加者は，兵庫県立大学に所属する学生 4 名 (男性 2 名，女性 2 名) である．2024 年 1 月 9 日から 2024 年 1 月 16 日にかけて実験を行い，参加者には約 1,500 円を支払った．

4.3　比較手法

ベースラインとして，文字列マッチングで検索を行い，結果に全文を表示するインタフェースを用いた．検索の流れとしては，検索窓に単語を入力し，文字列マッチングで検索した結果

3：https://www.rakuten.co.jp/

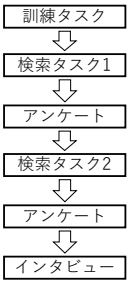

図 3　実験の手順.

表 1　各被験者が検索を行う順番.

被験者	検索タスク 1	検索タスク 2
1, 5	比較手法/商品 1	提案手法/商品 2
2, 6	比較手法/商品 2	提案手法/商品 1
3, 7	提案手法/商品 1	比較手法/商品 2
4, 8	提案手法/商品 2	比較手法/商品 1

を要約せずに表示するというものである．提案手法との違いは以下の 3 点である．

- **検索クエリとして単語を入力すること．**
- **クエリと全く同じ単語が含まれているレビューを検索結果に表示すること．**
- **検索結果に最初に表示されるものがレビュー全文であること．**

4.4　実 験 手 順

実験手順を図 3 に示す．まず，実験参加者に対して検索インタフェースの概要と収集するデータについて説明を行った．収集を行うデータは，検索クエリと検索時間のログ，及びインタビューの音声データである．次に，参加者が検索インタフェースに慣れてもらうために，訓練タスクを行った．訓練タスクでは，オーブントースター 1 商品のレビューに対し，提案手法であるインタフェースを用いたタスクを行ってもらった．具体的には，3 つの検索クエリを提示して検索を行ってもらい，結果に表示されるレビュー全文を Google Form で送信してもらうというものである．また，このタスクが終わった後に比較手法についても説明を行い，2 つの手法それぞれで自由に検索する時間を設けた．訓練タスクが完了した後，検索タスクを行い，ドライヤー 2 商品のレビューを見てどのような評価がされているのかを検索してもらった．検索タスクでは，2 つの商品のうち 1 つを比較手法，もう 1 つを提案手法のインタフェースで検索してもらった．実験参加者に対して検索してもらった商品とインタフェースの順番を表 1 に示す．本実験は，被験者内実験として行った．検索タスクの終了後，アンケートに回答してもらった．アンケートでは，インタフェースに対する満足度や，気になった観点についてくまなく検索を行うことができたかなどを回答してもらった．最後に，アンケートの回答理由を詳細に知るために口頭でインタビューを行い，ユーザ実験を終了した．

表 2　アンケート項目.

評価項目	質問
要約の見やすさ及び精度	レビューに含まれる意見を簡単に確認できたと思う.
ユーザの興味に対する網羅性	自分の調べたいことを調べることができたと思う.
観点の網羅性	様々な観点で調べることができたと思う.
意見の網羅性	同じ観点の中で漏れなく意見を調べることができたと思う.
結果の見やすさ	結果の表示方法が分かりやすかったと思う.
システムの使用難易度	システムの使用は簡単だったと思う.
全体的な満足度	システムの機能に満足している.

4.5　検索タスク

実験参加者には次の文言が書かれた紙を渡し，想定する状況を把握したのちにレビュー検索を行ってもらった．

> あなたは母親に誕生日プレゼントとしてドライヤーをプレゼントすることにしました．値段などを考慮して，2 つの商品に絞りました．
> それぞれの商品についてレビューを読み，どのような観点でどのような評価がされているのか調べて下さい．なるべく多くの観点で調査を行ってください．

制限時間は 1 商品につき 10 分とした．理由としては，1 つの商品あたり 200 件のレビューを用意しており，それを読むためにかかる時間として適切であると判断したためである．そして，充分に検索し終えたと感じたら検索を切り上げても良いこととした．また，自由にメモをとってもらい，実験終了後に回収した．

4.6　アンケート

検索タスクを終了するごとにアンケートを行った．アンケートはそれぞれの商品につき 7 問であり，その項目は表 2 の通りである．全ての問いで 5 段階リッカート尺度を用いた (1：そう思わない，2：あまりそう思わない，3：どちらともいえない，4：ややそう思う，5：そう思う).

4.7　アンケートの分析

アンケートおよびインタビューを対象として，アンケートの分析を行った．まず，アンケートの結果として，手法と設問ごとに回答者の平均と標準偏差を算出した表を以下表 3 に示す．この結果から，システムの使用難易度以外の 6 項目において提案手法が比較手法を上回ったことが分かった．特に 5%有意水準でウェルチの t 検定を行ったところ,「結果の見やすさ」については有意差が認められた．インタビューの中で，提案手法が比較手法を上回っている部分としてあげられることが多かった項目は「観点の網羅性」,「意見の網羅性」,「結果の見やすさ」の 3 つであった.「観点の網羅性」については，比較手法について言及するときに以下のような発話がみられた．

“自分のボキャブラリーにないと検索ができないので，そこがちょっと難しいかなと思ったので，そこは一番やりづらさを感じた部分だったかなと思います．”（参加者 1）

表 3　アンケート結果.

評価項目	比較手法 平均	比較手法 標準偏差	提案手法 平均	提案手法 標準偏差
1. 要約の見やすさ及び精度	4.00	1.31	4.80	0.48
2. ユーザの興味に対する網羅性	3.63	1.19	4.20	0.52
3. 観点の網羅性	3.88	1.55	4.40	0.52
4. 意見の網羅性	2.88	1.55	3.60	1.14
5. 結果の見やすさ	3.25	1.28	4.20	0.84
6. システムの使用難易度	4.63	0.74	4.40	0.71
7. 全体的な満足度	3.38	1.51	4.40	0.67

“どこをドラッグして検索しようか悩んでしまった自分がいました．”（参加者 4）

「意見の網羅性」については，以下のような発話がみられた．

“（提案手法について）結構広い範囲で（結果が）出てくれるので，そこは割といいなと思いました．”（参加者 1）
“1（比較手法）は「風量が大きい」で調べると,「大きい」と書いてあるレビューしか出てこないけど，2（提案手法）はそういうことがなかった．”（参加者 2）
“比較手法だと,「満足度」で調べると「満足度」と書いてあるやつは出てくるけど，大満足なのかどうかというのは，ワードが違うと出てこないから，同じ意見を出すという部分ではいい感じかも．”（参加者 3）

「結果の見やすさ」については，以下のような発話がみられた．

“後半（提案手法）の方は．何件それについて書いてあるかが分かった．”（参加者 2）
“（提案手法について）大量に検索（結果）が出たときに，こっちはめちゃくちゃ見やすいなと思った．”（参加者 3）

また，6 の「システムの使用難易度」のみ比較手法が提案手法を上回った．使用難易度について，以下のような発話がみられた．

“シンプルさっていうのは変わらないかな．”（参加者 1）

表 3 より平均が 4.4 点であることを踏まえると，提案手法でも使用は難しくないといえる．

4.8　クエリの分析

それぞれの参加者のクエリ発行数は表 4 の通りである．クエリ発行数とは，クエリの重複なしで検索ボタンが押された回数を指す．

参加者の中では比較手法のクエリ発行数が多い者が多かったが，手法ごとに有意差はみられなかった．

クエリごとの滞在時間についても分析を行った．クエリごとの滞在時間を，本研究では「クエリの検索ボタンが押されたとき」から「次のクエリの検索ボタンが押されたとき」までとする．そのため，最後から 2 番目までに検索されたクエリを対象として滞在時間の計算を行った．また，検索ボタンのダブルクリックを行っているものや，レビューに関係のない記号について検索しているものは除外した．比較手法と提案手法におけるそれぞれの参加者の平均滞在時間を表 5 に示す．提案手法の方が滞在時間が長い者が多かったが，5%有意水準でウェルチの t 検定を行ったところ有意差はみられなかった．

表 4　クエリ発行数.

参加者	比較手法	提案手法
1	9	8
2	18	22
3	26	14
4	15	6
5	13	10
6	9	13
7	15	9
8	10	8

表 5　クエリごとの平均滞在時間（秒）.

参加者	比較手法	提案手法
1	56.62	79.86
2	33.29	27.95
3	21.20	40.08
4	40.43	105.80
5	48.25	44.78
6	63.38	47.45
7	28.36	76.12
8	48.33	72.71

表 6　選択したクエリ.

単語	使いやすかった
	音があまり気にならない
	こげたような匂い
	軽くていい
	冷風にすると風量が弱くなってしまう
フレーズ	風量
	カラー
	プレゼント
	故障
	温度

5　レビュー検索の精度評価

5.1　概　　要

どの程度正しく検索結果にレビューが表示されるのかを調べるために，システムで用いられる検索の精度を評価した．使用するクエリは，予備実験と本実験で商品 1 のレビュー検索において使用されたクエリ 10 個である．そのうち 5 個を名詞 1 つで構成される単語のクエリ，残りの 5 個を名詞以外の品詞も含まれるフレーズとした．選択したクエリは表 6 の通りである．それぞれのクエリについて，レビューが検索結果に表示されて欲しいと著者が判断したものを真の値とし，適合率，再現率，F_1 値を算出した．

5.2　結　　果

それぞれの手法で単語とフレーズの適合率，再現率，F_1 値の平均を以下の表 7 で示す．なお，提案手法の単語における適合率および F_1 値については,「故障」「温度」クエリの検索結果が 0 件であったため，適合率の値を 0 として平均を算出した．

表 7　レビュー検索の精度評価の結果.

	比較手法		提案手法	
評価項目	単語	フレーズ	単語	フレーズ
適合率	0.94	1.00	0.60	0.72
再現率	0.42	0.06	0.20	0.60
F_1 値	0.56	0.11	0.27	0.61

比較手法では単語で検索を行うと F_1 値が高く，提案手法ではフレーズで検索を行うと F_1 値が高くなった.

6 考　　察

6.1 アンケートの分析

「結果の見やすさ」という部分において高い評価を受けたのは，提案手法において要約した結果を一目で可視化できたからであると考えられる．比較手法だと検索結果に大量の文字が表示されるため，全てのレビューを読むより軽減されるものの，レビューからユーザが知りたいことを探すには大変である．一方で提案手法では結果を短く要約し，件数を表示することで，ユーザにとってレビューを読むことに対する負担を減らしている．このため，アンケートにおいて有意差が認められ，多くのインタビューで言及された.

「意見の網羅性」という部分において高い評価を受けたのは，「見た目」と「デザイン」といった同じ意味を表す単語が含まれているレビューが表示されないという問題や，実際に文字列マッチングで検索するには適切なキーワードが浮かばないという問題が解消したからであると考えられる．実際に文字列マッチングで検索するには適切なキーワードが浮かばないという問題を解消できた例を挙げる,「購入後、1 年 2ヶ月で断線しました.」というレビューが気になったとき，ユーザは耐久性に関連するレビューを検索しようとする．しかし，故障の中でも「火花が出た」や「動かなくなった」など表現が様々であり，キーワード検索において一括で検索するのは難しい．提案手法では,「1 年 2ヶ月で断線」で検索を行うと「1 年ちょっとしか使ってないのに、コード根元から火花が散り、使えなくなりました.」「2 月に買って 7 月の頭にいつも通りターボで髪を乾かしていたらいきなり電源が切れスイッチがガチガチに固まってそれ以降動かなくなりました.」といった，様々な壊れ方をしたレビューが表示される．このため，意見の網羅性という項目において高い評価となり，多くのインタビューで言及された.

6.2 クエリの分析

提案手法の方が 1 つのクエリに対する滞在時間が長くなった要因は 2 つあると考えられる．1 つ目は，提案手法の方が処理に費やす時間が長くなるためである．提案手法では，gpt–35–turbo–16k を用いるために OpenAI の API を呼び出している．一方で，比較手法でクエリに一致するレビュー文を見つけるためには，API を呼び出す必要はない．クエリの検索ボタンを押してから次のクエリの検索ボタンが押されるまでを測定しているため，滞在時間には検索結果が表示されるまでの時間が含まれる．そのため，処理時間が滞在時間に影響を及ぼしている可能性があると考えられる．今後，結果が表示された時刻を記録することで，処理時間に起因せず滞在時間を測定する必要がある．2 つ目は，提案手法の方が検索結果に出力されるレビューの件数が多いためである．レビューの件数が多くなると，関係のないレビューが出力される可能性もある．ただ表 7 では，フレーズで検索したときの再現率が高くなっており，関係のあるレビューが出力される確率は高い．そのため，比較手法と比べて関係のあるレビューが多く出力され，それらを読むために滞在時間が長くなると考えられる.

6.3 検索精度の評価

比較手法でフレーズより単語の方が F_1 値が高くなったのは，クエリの文字数や情報量が多くなるにつれて完全にクエリと一致するレビューが少なくなるからである．たとえば「風量」の検索結果には,「風量が強い」というレビューも「風量が弱い」というレビューも含まれる．しかし,「風量が強い」の検索結果には「風量が弱い」というレビューが含まれない．したがって，単語で検索する方がクエリと一致するレビューの数が多くなったと考えられる.

また提案手法において，単語で検索したときに F_1 値が低くなった理由として，1 つの単語と文では言葉の単位が異なり，類似度が低下した組合せが多いことが考えられる．たとえば「音があまり気にならない」と「音量」の場合，類似度は 0.80 となり，要約の対象にはならない．ただ,「音」と「音量」は類似する単語であり，類似度も 0.87 と要約対象となる基準を上回っている.「音」の後に単語が続くことによって,「音」以外の単語も多く考慮した分散表現に変換されてしまい，類似度が 0.85 を満たさなくなったと考えられる.

7 まとめと今後の課題

本研究では，レビューの中でユーザが気になる部分を選択し，大規模言語モデルを用いて生成された類似するレビューの要約を提示するインタフェースを提案する．このインタフェースを用いることで，ユーザは類似文を検索することや，検索結果で内容を一目見て把握することが可能になる．実際にユーザに使用してもらい,「要約の見やすさ及び精度」,「ユーザの興味に対する網羅性」,「観点の網羅性」,「意見の網羅性」,「結果の見やすさ」,「システムの使用難易度」,「全体的な満足度」の 7 つの項目でアンケートを行った．その結果「システムの使用難易度」以外の 6 項目において，提案手法の回答者平均が比較手法を上回ったが，統計的な有意差はみられなかった．また，クエリについて分析を行った結果，提案手法において 1 つのクエリごとの滞在時間が長いことが多かったが，有意差はみられなかった.

そして，レビューの検索精度の評価も行った．文字列マッチングを用いた手法ではクエリに単語を用いた検索において F_1 値が高く，提案手法ではクエリにフレーズを用いた検索において F_1 値が高いという結果になった.

最後に，今後の課題について 3 点述べる．1 点目は類似レ

ビューを判定するシステムの調整である．6.3 節で述べたように，提案手法では文に対してベクトル化を行っているため，単語との類似度が低くなってしまった．そのため，類似レビューと判定する閾値の変更や，文に対して助詞といった単語の除外処理を行ってから分散表現に変換する必要があると考えられる．2 点目はプロンプトの改善である．出力例と入力例としてクエリとその元となったレビュー文を ChatGPT に入力すると，出力がクエリと同様である事例が多数あった．また，10 文字以内で要約できていない事例も存在していたため，現状とは異なるプロンプトの記述方法でも要約を行う必要がある．3 点目は要約の精度評価である．今回，システムの機能のうち，検索精度についての評価は行ったが，要約の精度評価はまだ行っていない．そのため，2 点目で述べた課題も踏まえ，プロンプトの改善前と改善後における要約の精度評価が必要である．

謝辞　本研究は JSPS 科学研究費助成事業 JP21H03774, JP21H03775，JP22H03905，による助成を受けたものです．ここに記して謝意を表します．

文　献

[1] Stefanos Angelidis and Mirella Lapata. Summarizing opinions: Aspect extraction meets sentiment prediction and they are both weakly supervised. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, p. 3675. Association for Computational Linguistics, 2018.

[2] Yang Deng, Wenxuan Zhang, and Wai Lam. Opinion-aware answer generation for review-driven question answering in e-commerce. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pp. 255–264, 2020.

[3] 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.

[4] Mahmood Jasim, Christopher Collins, Ali Sarvghad, and Narges Mahyar. Supporting serendipitous discovery and balanced analysis of online product reviews with interaction-driven metrics and bias-mitigating suggestions. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pp. 1–24, 2022.

[5] Koteswar Rao Jerripothula, Ankit Rai, Kanu Garg, and Yashvardhan Singh Rautela. Feature-level rating system using customer reviews and review votes. *IEEE Transactions on Computational Social Systems*, Vol. 7, No. 5, pp. 1210–1219, 2020.

[6] Shou Jianqi. Towards known unknowns: GPT large language models empower human-centered information retrieval. *Journal of Library and Information Sciences in Agriculture*, Vol. 35, No. 5, p. 16, 2023 (in Chinese).

[7] Bing Liu, Minqing Hu, and Junsheng Cheng. Opinion observer: analyzing and comparing opinions on the web. In *Proceedings of the 14th international conference on World Wide Web*, pp. 342–351, 2005.

[8] Junling Liu, Chao Liu, Renjie Lv, Kang Zhou, and Yan Zhang. Is ChatGPT a good recommender? a preliminary study. *arXiv preprint arXiv:2304.10149*, 2023.

[9] Thanh-Son Nguyen, Hady W Lauw, and Panayiotis Tsaparas. Review synthesis for micro-review summarization. In *Proceedings of the eighth ACM international conference on web search and data mining*, pp. 169–178, 2015.

[10] Soujanya Poria, Erik Cambria, and Alexander Gelbukh. Aspect extraction for opinion mining with a deep convolutional neural network. *Knowledge-Based Systems*, Vol. 108, pp. 42–49, 2016.

[11] David Pride, Matteo Cancellieri, and Petr Knoth. CORE–GPT: Combining open access research and large language models for credible, trustworthy question answering. In *International Conference on Theory and Practice of Digital Libraries*, pp. 146–159. Springer, 2023.

[12] Sofia Eleni Spatharioti, David M Rothschild, Daniel G Goldstein, and Jake M Hofman. Comparing traditional and llm-based search for consumer choice: A randomized experiment. *arXiv preprint arXiv:2307.03744*, 2023.

[13] ホンヘジン, 木村優介, 神田悠斗, 吉丸直希, 波多野賢治. 商品説明に着目したレビュー文からの教師なしキーフレーズ抽出. 第 15 回データ工学と情報マネジメントに関するフォーラム, 1b–5–5, 2023.

[14] 金兵裕太, 沼尾雅之. ネットショッピングサイトの商品レビューを利用したジャンル毎の評価軸の自動構築とその応用. 第 8 回データ工学と情報マネジメントに関するフォーラム, C2–3, 2016.

[15] 栗原光祐, 荘司慶行, 藤田澄男, MartinJ.Dürst. Doc2vec 手法による映画レビューサイトからのクエリと意味的に類似した評価表現の発見. 第 11 回データ工学と情報マネジメントに関するフォーラム, C4–2, 2019.

[16] 経済産業省. 令和 4 年度 電子商取引に関する市場調査 報告書. `https://www.meti.go.jp/policy/it_policy/statistics/outlook/230831_new_hokokusho.pdf`. 2023 年 11 月 27 日閲覧.

[17] 市村哲. 口コミから美味しい料理店を手早く探すシステム. 情報処理学会論文誌, Vol. 61, No. 11, pp. 1748–1756, 2020.

[18] 杉木健二, 松原茂樹. カスタマーレビューに基づく商品検索のための感性表現シソーラスの構築. 言語処理学会第 15 回年次大会発表論文集, pp. 781–784, 2009.

[19] 渡辺一生, 楠和馬, 吉丸直希, 波多野賢治. レビュー要約を目的とした単語の意味領域に基づく上位語選択. 第 15 回データ工学と情報マネジメントに関するフォーラム, 1b–5–5, 2023.

[20] 平山拓央, 湯本高行, 新居学, 佐藤邦弘. 語の共起と極性に基づく商品レビュー閲覧支援システム. 情報処理学会研究報告, Vol. 2012–DBS–155, No. 3, pp. 1–9, 2012.