

商品比較のための文脈つき評価軸抽出の検討

小橋 賢介[†] 酒井 哲也[†]

[†] 早稲田大学基幹理工学研究科情報理工・情報通信専攻 〒169-8555 東京都新宿区大久保 3-4-1

E-mail: †lelouch-ken@ruri.waseda.jp, ††tetsuyasakai@acm.org

あらかし 通販サイトのレビューは実際に使用した経験に基づいているため、商品を購入するユーザーにも商品を製造している企業にとっても重要なリソースである。類似商品に関する多くのレビューを閲覧することはユーザの負荷となる。この負荷を抑える効果的な方法のひとつが、ユーザに複数の評価軸に基づく類似商品比較結果を提供することである。例えば、ユーザがスマートフォンの比較検討を行う場合、音質、カメラ画質、重量、値段などの評価軸を提供すること考えられる。しかし、従来研究では、これらの各軸についていずれの商品が優れているかというレベルの情報しか提供されていなかった。本研究では、レビューから抽出した文脈を各評価軸に付加することにより、より詳細な商品比較を可能にする手法を検討する。提案手法により自動抽出した文脈つき評価軸をアノテータが作成した正解データとの比較により評価を行ったところ、ジャックカード係数による部分一致では6割程度、完全一致では1割程度にとどまり、大幅な改善の余地があることがわかった。

キーワード 自然言語処理, テキストマイニング, 評判分析, 情報要約

1 はじめに

Amazon¹や楽天市場², Yahoo!ショッピング³などの通販サイトでは、商品のサイズや機能だけでなく、商品を購入したユーザが投稿したレビューを閲覧することができる。通販サイトのレビューは実際に使用した経験に基づいているため、商品を購入するユーザにも商品を製造している企業にとっても重要なリソースである。ユーザは購入する商品を決める際にレビューを参照し、一方で企業は商品の評判を調査するためにレビューを参照する。

しかし、その商品の評価を行うにはその商品のレビューだけでなく、類似商品に関するレビューを閲覧し商品と比較する必要があり、全てのレビューを閲覧することはユーザや企業の負荷となる。この負荷を抑える効果的な方法のひとつが、ユーザに複数の評価軸に基づく類似商品比較結果を提供することである。例えば、ユーザがスマートフォンの比較検討を行う場合、音質、カメラ画質、重量、値段などの評価軸に基づいた商品比較結果を提供することである。

従来研究では、これらの各軸についていずれの商品が優れているかというレベルの情報しか提供されていなかった。そこで本研究では、レビューから抽出した文脈を各評価軸に付加することにより、より詳細な商品比較を可能にする手法を検討する。レビューから文脈を抽出するためにStanford CoreNLP [1]を用いた。レビューの構文解析を行い、評価軸単語と文法的に依存関係にある単語を取得して文脈を抽出する。人手による正解データ文脈と構文木から得られた文脈を比較して、文脈付与の有効性を検証する。

以降、第2章では関連研究を挙げ、第3章では提案手法の詳細を説明する。第4章では提案手法を実データを用いて実験を行い、結果について報告し、第5章で考察したあと、第6章で結論と今後の展望について述べる。

2 関連研究

2.1 比較関係抽出

通販サイトのレビューから商品と比較している文を抽出する研究は数多く行われている。

Jindalら [2] は Class Sequential Rule (CSR) [3] と機械学習を組み合わせて文が比較文か非比較文かどうかの2値分類を行った。データセットはレビューだけでなく、ニュース記事も含まれる。

また、Jindalら [4] は先ほどのCSRによるレビューが比較文かどうかの分類に加え、更に2つのタスクを行った。1つ目は、抽出した比較文を Non-Equal Gradable, Equative, Superlative の3値に分類を行った。2つ目は Label Sequential Rule (LSR) を用いて、比較している商品や比較ワードを抽出し、(relationWord, features, entityS1, entityS2) の形で出力した。以下に例文と出力結果を示す。

例文 1:

Canon's optics is better than those of Sony and Nikon.

例文 1 結果:

(better, optics, Canon, Sony, Nikon)

Xuら [5] は Conditional Random Fields (CRF) [6] を改良した two-level CRF を用いてスマートフォンに関するレビューから比較方向、比較している商品、比較ワードを抽出した。比較方向を入れることで、商品のどちらが優れているかどうかを把握することができる。以下に例文と出力結果を示す。

1 : <https://www.amazon.com/>

2 : <https://www.rakuten.com/>

3 : <https://shopping.yahoo.com/>

例文 2: Nokia N95 has a better camera than iPhone.

例文 2 結果: >(Nokia N95, iPhone, camera, better)

Danone ら [7] は Xu らの手法を用いてレストランのレビューから比較方向, 比較している商品, 比較ワードを抽出した. 出力結果は Xu らと同様の形態である. それに加え, 出力結果を様々なグラフを用いて視覚化を行った.

Li ら [8] は 2 つ以上の Entity, すなわち比較対象を比較することを目的とする質問 (Comparative question) に対し, Entity の抽出を行った.

比較文書の抽出は英語以外でも研究されている. Wang ら [9] は Chinese Opinion Analysis Evaluation (COAE 2008) から中国語のレビューデータセットを取り出し, 文法から独自の比較パターンから比較レビューを抽出した.

2.2 StanfordNLP と Stanford CoreNLP

StanfordNLP [10] と StanfordCoreNLP [1] はどちらもスタンフォード大学がオープンソースで提供している自然言語処理フレームワークである. 対応言語は StanfordNLP が 53 カ国である一方, StanfordCoreNLP は 6 ケ国語である. どちらもトークン化 (tokenize) や品詞付け (postag), 係り受け解析 (dependency), 構文解析 (parser) を行う.

両者とも品詞付けの機能はあるが, 品詞付けのラベルが異なる. 例文とそれぞれの品詞付け出力結果を図 1 に示す.

Example	Galaxy	S4	is	better	than	iPhone	.
Stanfordnlp	PROPN	PROPN	AUX	ADJ	ADP	PROPN	PUNCT
StanfordCoreNLP	NN	NN	VBZ	JJR	IN	NNP	.

図 1 StanfordNLP と StanfordCoreNLP の品詞付けの違い

Stanford NLP は better を ADJ (形容詞) と品詞付けしたが, StanfordCoreNLP は JJR (形容詞:比較級) と品詞付けを行い, StanfordCoreNLP の方がより詳細に品詞付けを行っている. つまり, StanfordCoreNLP の品詞付けは Stanford NLP より比較文書を取り扱うのに適している. よって, 本研究では StanfordCoreNLP を使用する.

3 提案手法

3.1 用語定義

軸

商品を比較する対象. 例文 3 の場合, 軸は「camera」である. 他にも「screen」, 「sound」, 「cost」などがある.

比較用語

商品の優劣関係を決定する単語. 例文 3 の場合, 比較用語は「better」である. 「smaller」, 「higher」, 「cheaper」などがある.

比較商品

レビュー内で優劣関係を決定する対象. 例文 3 の場合は「Nokia N95」, 「iPhone」である. 本研究では, スマートフォンの名称が比較商品である.

例文 3:

Nokia N95 has a better camera than iPhone.

3.2 問題提起

既存研究では例文 4 から「LG G3」, 「Galaxy Note S5」, 「bigger」, 「screen」が抽出される. この例文の軸は「screen」である. 例文 4 と比較して例文 5 は商品間のスクリーンの差を具体的に述べているが, 例文 5 から抽出される要素は例文 4 と同様の要素が抽出される. それでは, 商品間で具体的にどのくらい差があるのかを把握することができない. 例えば, 「0.4 inch」が先ほどの要素に加わることで具体的な差を把握することができる. 軸「screen」に詳細な説明がついた文「0.4 inch bigger screen than」を文脈付き評価軸と呼ぶ. 本研究では, 商品間の具体的な差を把握するために文脈付き評価軸を抽出する手法を提案する.

文脈付き評価軸を抽出するために Stanford CoreNLP を利用する.

例文 4:

LG G3 is bigger screen than Samsung Galaxy S5.

例文 5:

LG G3 is 0.4 inch bigger screen than Samsung Galaxy S5.

3.3 手法詳細

以下に, 文脈付き評価軸を抽出する手順を示す.

- 1). レビューを StanfordCoreNLP に入力し, 係り受け解析 (dependency parsing) を行う
- 2). 軸, 比較用語と係り受け関係にある単語を取り出す
- 3). 比較商品, 軸, 比較用語, 2) で取り出した単語を順番に並べる

まず 1) でレビューを StanfordCoreNLP に入力することでレビュー内の単語間の係り受け関係を取得する. 図 2 に係り受け関係にある単語を取り出す様子を示す. 係り受け関係抽出の視覚化にはスタンフォード大学が公開しているサービス⁴を用いた. 係り受け解析についての詳細⁵はここでは省略する. 次に, 2) で軸, 比較用語とそれぞれ係り受け関係にある単語を取り出す. 図 2 の場合, 比較用語「bigger」と係り受け関係にある「inch」, その「inch」と係り受け関係にある「0.4」を抽出する. 最後に比較商品, 軸, 比較用語, 2) で取り出した単語を順番に並べて文脈付き評価軸の抽出を行う. 図 2 では, 「LG G3 0.4 inch bigger screen Samsung Galaxy」を文脈付き評価軸とする. 比較商品や軸, 比較用語はアノテータがあらかじめ抽出した単語を用いる.

4 実験と結果

第 3 章の提案手法を基に実験を行った. 実験に用いるデータセット, 評価方法, 得られた結果を以下に示す.

4: <https://corenlp.run/>

5: https://nlp.stanford.edu/software/dependencies_manual.pdf

Basic Dependencies:

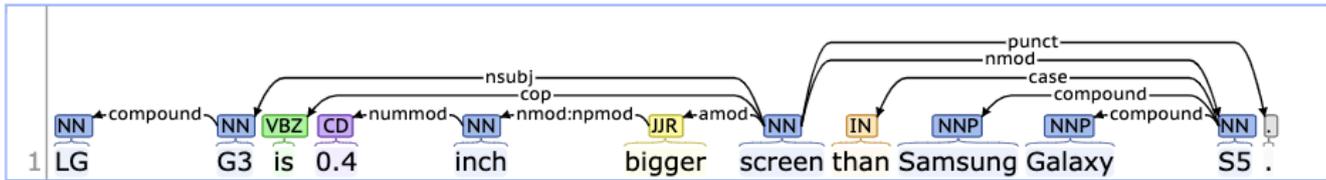


図 2 係り受け関係にある単語抽出の様子

4.1 データセット

Amazon のスマートフォンに関するレビューデータ⁶を文レベルで分割し、比較商品が含まれている比較レビュー 60 件取得した。

アノテーションは 20 代男性と 20 代女性の 2 人が各自 60 件ずつ行い、レビュー 1 件ごとに比較商品、軸、比較用語を抽出し、レビューからサマリーを作成した。サマリーとは、文脈つき評価軸と比較商品を単語順に羅列した文である。

例えば、レビュー文「LG G2 is much better display than my iPhone 4.」に対してアノテーションを行った場合、比較商品 1 に「LG G2」、比較商品 2 に「iPhone 4」、軸に「display」、サマリーに「LG G2 much better display than iPhone 4」と入力する。

4.2 評価方法

提案手法で出力した文脈つき評価軸がレビューからどの程度情報を抽出できたのか精度を評価するために、アノテータが作成したレビュー文のサマリーを正解データとして、提案手法と正解データの比較を行う。サマリーの形式に比較商品が含まれているので、提案手法の出力した文脈つき評価軸に比較商品の単語を組み込んだ。

アノテータのサマリーは統合することができないのでそれぞれ別の正解データとして扱う。それぞれを正解データ 1、正解データ 2 とする。評価指標は完全一致平均とジャックカード係数の平均を用いる。

それぞれの評価指標の説明と式を示す。完全一致平均はレビューごとに提案手法から出力された単語と正解データの単語が全て同じだった場合を 1 と出力し、一つでも異なっていたら 0 と出力して、レビューごとの値の合計を平均した値である。

ジャックカード係数とは $J(A, B)$ とは 2 つの集合 A, B に対して、共通部分を A, B の和集合で割ったものである。本研究では、提案手法から出力される単語の集合を A 、正解データ中の単語の集合を B として算出する。

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \quad (1)$$

4.3 評価

第 4.2 節で述べた方法で、提案手法の精度評価を行う。

評価結果を表 1 に示す。表 1 より完全一致度平均が共に 1

表 1 文脈つき評価軸抽出の精度

	完全一致度平均	ジャックカード係数平均
提案手法と正解データ 1	0.117	0.633
提案手法と正解データ 2	0.100	0.696

割程度でとても低い結果となった。一方、ジャックカード係数平均は共に 6 割を超える結果となった。ここから、ユーザがレビューから抽出したい文脈の大半を取得できるが、文脈と完全一致することは困難であることがわかった。

5 考察

第 4.3 節より提案手法はユーザが取得したい文脈の平均 6 割を取得する結果になった。以下に、ジャックカード係数が全レビューの平均と近い値の例を示す。正解データと抽出結果からお互い共通していない単語に下線を引く。

例文 6:

The battery life is way better than the LG G3 that I was used to.

抽出結果 (アノテータ 1 の単語参照):

The battery life is way better than LG G3

正解データ 1:

battery life better than LG G3

抽出結果 (アノテータ 2 の単語参照):

The battery life is way better than the LG G3

正解データ 2:

The battery better than the LG G3

以上のように、ジャックカード係数で 6 割超という結果は、実用的ではない。そこで、提案手法の精度が低い原因の考察を行った。

5.1 レビュー特有の比較表現

通販サイトのレビューは、記載している商品をもとに作成されているのでその商品名を省略する傾向がある。以下の例文 7 にあるように比較対象の商品名は存在するが、レビューの対象である商品は省略されている。レビューの対象である商品が省略されることで、軸が主語になることがある。この場合、文脈つき評価軸は「the battery lasts longer than」が正しくなり、動詞を入れないと意味が通らなくなる。

例文 8 の文脈つき評価軸は「longer battery than」となり、

6 : <https://nijianmo.github.io/amazon/index.html>

動詞を必要としない。提案手法では、軸が主語になった場合に動詞も文脈つき評価軸に含める処理を行っていないので、精度が下がっている。

例文 7:

the battery lasts longer than the Galaxy S7.

例文 8:

product A has a longer battery than the Galaxy S7.

5.2 レビュー内における軸、比較用語の曖昧性

第 4.1 節で述べたように、比較商品や軸、比較用語の抽出はアノテータが行った。つまり、文脈つき評価軸はアノテータが抽出した比較商品や軸、比較用語を利用して文脈つき評価軸を出力する。レビューによって軸や比較用語が曖昧であることから抽出が困難になり、正解データとはかけ離れた結果になる。以下に軸と比較用語が曖昧なレビューの例を示す。また、レビュー本文の他に、各アノテータが抽出した比較商品、軸、比較用語、正解データを以下に示す。正解データと抽出結果からお互い共通していない単語に下線を引く。

例文 9:

I found this phone to be more user-friendly than the iPhone 3 it replaced.

・アノテータ 1 が抽出した軸と比較用語

軸: user-friendly 比較用語: user-friendly

・抽出結果 (アノテータ 1 の単語参照):

this phone to be more user-friendly user-friendly than the iPhone 3 replaced

・正解データ 1:

this phone more user-friendly than iPhone 3

・アノテータ 1 が抽出した軸と比較用語

軸: more user-friendly 比較用語:

・抽出結果 (アノテータ 2 の単語参照):

phone to be more user-friendly than iPhone 3

・正解データ 2:

this phone more user-friendly than the iPhone 3

以上のように「user-friendly」という普通の文法に無い単語が出現することで、その単語が軸か比較用語かの判断が困難になり、正解データとはかけ離れた結果になった。実際に、アノテータ 1 とアノテータ 2 で異なる抽出を行った。この一因としてアノテータへの指示が不明確であったことが挙げられる。

6 結論と今後の展望

本研究では、商品比較の具体的な差を把握するために、レビューから StanfordCoreNLP を用いて文脈付き評価軸抽出方法を提案し、評価を行った。結果はユーザがレビューから取得したい情報の 6 割以上を抽出することができたが、正解データと完全一致する確率は 1 割と低い精度で実用的でない。文脈つき評価軸の抽出精度の向上が今後の課題である。

本研究は、あらかじめ比較文と軸、比較商品、比較用語が与えられている状態で文脈つき評価軸抽出を行った。実用的にするには、レビューから軸、比較商品、比較用語の自動抽出と文脈つき評価軸抽出の精度の両方を考慮に入れる必要がある。

7 謝 辞

本論文の執筆にあたり、貴重なご意見を頂いた酒井研究室の河東宗祐先輩やアノテーションを引き受けていただいた 2 名に感謝の意を表します。

文 献

- [1] The Stanford Natural Language Proceeding Groups, Stanford Parser, (<https://stanfordnlp.github.io/CoreNLP/>), 2018.
- [2] Jindal, N. and Liu, B. “Identifying comparative sentences in text documents.” SIGIR-06, 2006.
- [3] Bing Liu. “Web Data Mining Exploring Hyperlinks, Contents, and Usage Data Second Edition” Springer, 55–56, 2011.
- [4] N. Jindal, B. Liu, “Mining comparative sentences and relations”, in: K.D. Forbus (Ed.), Proceedings of 21st National Conference on Artificial Intelligence, Boston, Massachusetts, USA, pp. 1 - 6, 2006.
- [5] Xu, K., Liao, S. S., Li, J., Song, Y. “Mining comparative opinions from customer reviews for Competitive Intelligence.” Decision support systems, 50(4), pp. 743-754, 2011.
- [6] John D. Lafferty, Andrew McCallum, Fernando C Pereira. “Conditional random fields: Probabilistic Models for Segmenting and Labeling Sequence Data” Proceedings of the Eighteenth International Conference on Machine Learning, 2001.
- [7] Yaakov Danone, Tsvi Kuflik, Osnat Mokryn, “Visualizing Reviews Summaries as a Tool for Restaurants Recommendation,” In 23rd International Conference on Intelligent User Interfaces (IUI '18). ACM, New York, NY, USA, 607–616, 2018.
- [8] Shasha Li, Chin-Yew Lin, Young-In Song, Zhoujun Li “Comparable Entity Mining from Comparative Questions,” IEEE Transactions on Knowledge and Data Engineering, 25, 1498–1509, 2010.
- [9] Wang, W., Zhao, T., Xin, G., Xu, Y. “Recognizing Comparative Sentences from Chinese Review Texts,” International Journal of Database Theory and Application, 7(5), 29–38, 2014.
- [10] The Stanford Natural Language Proceeding Groups, Stanford Parser, (<https://stanfordnlp.github.io/stanfordnlp/>), 2018.