リビュー文書を対象とした句単位の日本語評価極性タグ付きコーパス

中澤 真人　池田 可奈子　山田 美知花　吉村 綾馬　鈴木 由衣　小町 守
首都大学東京

{nakazawa-naoto, ikeda-kanako1, yamada-michika, yoshimura-ryoma, suzuki-yui}@ed.tmu.ac.jp, komachi@tmu.ac.jp

1 はじめに

ウェブ上のリビュー文書は，マーケティングに役立つ有益な情報が多く含まれるため，評価極性分析の対象となっている．評価極性分析の中でも，「アメニティがあれば嬉しい」というように，フレーズとして極性のついた単語周辺の文脈を考慮することで，ユーザの嗜好やニーズの獲得ができる [1].

Socher ら [2] が構築した Stanford Sentiment Treebank（以下 SST）は，英語で書かれた映画のリビュー文を対象として句単位の評価極性情報が付与されている．SST は文に対して自動で句構造解析を行い，抽出された構成素に対して 5 段階の評価極性を付与してあるデータセットで，英語の評価極性分析で広く使われている [3].

しかしながら，日本語では文書単位または文単位で評価極性が付与されたコーパスは存在するが，句単位で評価極性情報が付与されたコーパスは存在しない．文単位のアノテーションでは，どの単語がどの単語と結びついて極性を持つかといった局所的な構造や，文中でフレーズの極性が否定表現によって反転したり強意表現によって強まったりするのかが分からず，日本語でこれらの構造を考慮することが評価極性分析に有効かどうかを検証することができない．

そこで我々は，**日本語のリビュー文書に句単位で評価極性情報を付与**した．我々のコーパスは Stanford Sentiment Treebank を参考に，句に対して 5 段階の評価極性をアノテートしたデータセットである．

本研究の主な貢献は，以下の点である．

- 日本語のリビューコーパスに対し，句単位の評価極性情報を付与した
- 本研究で構築したコーパスは，GitHub からダウンロードできる

2 関連研究

日本語に対する評価極性分析では，NTCIR-6 OPINION 意見分析パイロットタスクテストコレクション[1]と NTCIR-7 MOAT 多言語意見分析テストコレクションが広く使われている[2]．このデータは新聞記事を対象として，1 文単位で人手で評価極性情報を付与したものである．新聞記事を対象にしているため，リビュー文書の解析に必要なウェブに特有の表現は含まれず，また，文より小さい単位の評価極性を知ることができない．そこで本研究では，ウェブ文書からフレーズを抽出してアノテーションを行なった．

一方，ウェブ文書に対する評価極性タグ付きコーパスとしては，文単位で評価極性情報が自動付与された ACP Corpus がある[3]．ACP Corpus はウェブ文書集合から評価文を自動収集し，HTML の構造を用いて箇条書きの見出しの「良い点」「悪い点」などの情報を手がかりに評価極性を付与している．しかし，自動付与されたコーパスには誤った情報が含まれる．また，このコーパスは文単位のアノテーションである．本研究では，フレーズを対象に人手によるアノテーションを行ない，アノテーションの信頼性についても調査した．

また，日本語ブログコーパスとして Kyoto University and NTT Blog コーパス[4]（KNB コーパス）がある．KNB コーパスはブログ記事にアノテータの任意の句単位で評判情報が付与されている．本研究ではブログ記事ではなくリビュー記事に対して評価極性情報を付与した．また，記事単位で文脈を見て特定の句に対しアノテートするのではなく，あらゆる句を対象にフレーズ以外の情報を参照せずに評価極性情報を付与した．

[1] http://research.nii.ac.jp/ntcir/permission/ntcir-6/perm-ja-OPINION.html
[2] http://research.nii.ac.jp/ntcir/permission/ntcir-7/perm-ja-MOAT.html
[3] http://www.tkl.iis.u-tokyo.ac.jp/~kaji/acp/
[4] http://nlp.ist.i.kyoto-u.ac.jp/kuntt/

3　句単位の評価極性タグ仕様

3.1　アノテーション対象

評価極性タグのアノテーションは，Socher らの SST [2] にならって句構造解析済みのフレーズ（句）を対象として行う．文に対して句構造解析をして得られた全てのフレーズに極性を付与する．ただし，アノテーションの際にはフレーズのみを見て極性を判断し，フレーズ外の文脈は考慮しない．また，フレーズはトークンではなくタイプとしてアノテーションする．

句構造は自動解析により付与するが，形態素・句構造解析誤りと思われる事例においても，フレーズ区切りは正しいものとして極性を付与する．例えば「この部屋は木が邪魔できれいな海がいまいちでした。」という文から「この部屋は木が邪魔できれい」というフレーズが自動抽出されたとする．句構造解析が正しければこのフレーズは抽出されないが，このような場合にも正しい句構造を付与するのではなく，抽出されたフレーズを対象に極性を付与する．

3.2　アノテーション基準

リビューからの情報抽出のために，Socher らの SST [2] にならい読み手として判断した極性を付与する．

SST では人手によるアノテーションの際には 7 段階で分けて付与していたが，最終的なデータは 5 段階に変換されていたため，本研究では表 1 に示すように最初から 1 から 5 の 5 値で極性を付与する．

1 (very negative) や **5 (very positive)** はそれぞれ，ネガティブまたはポジティブなフレーズが強意語などの強める表現で修飾されている場合，または 2 つ以上の同じ極性を持ったサブフレーズがフレーズの中に含まれている場合に付与する．本コーパスでは 1 フレーズだけで 1 または 5 の極性が付与されることはない．

3 (neutral) はポジティブでもネガティブでもない場合，多義語で文脈不明の場合などに付与する．ポジティブでもネガティブでもない場合には，「レストランに入る」などの一般的なものや，「いまだかつて」などのような強意表現であるフレーズが含まれる．多義語で文脈不明の例は，「食後はトレイを下げてくれ」があり，述語の活用によって事実を述べる場合（連用形）と，要望を述べる場合（命令形）があるが，3 と判定する．また，書き手の変換・タイプミスは元の表現が推測できる場合は推測して極性を付与する（「不陰気」→「雰囲気」と判断して 3）が，元の表現が推測できない場合も 3 を付与する．

表 1: ラベルとフレーズの例.

ラベル	フレーズ
1 (very negative)	二度と宿泊しない
2 (negative)	不便ですが、
3 (neutral)	貸し出し用
4 (positive)	豊富なメニュー
5 (very positive)	価格的にも非常に満足

2 (negative) と **4 (positive)** はフレーズの中に 1 つ以上の極性を示す表現があった場合に付与する．2 つ以上極性を示す表現があった場合，アノテータの主観に委ねる．同じ極性を持つサブフレーズがある場合は 1 や 5 になる場合もあれば，相反する極性を持つサブフレーズが含まれる場合は 3 になる場合もある．

リビューには条件や限定付きで評価する文も頻出するが，条件節の有無によって極性を変化させるかどうかはアノテータの主観に委ねた．例えば，「マラソン後の宿泊先としては最高だと思います」というフレーズの場合，「最高だと思います」の部分の極性は 5 となるが，「マラソン後の宿泊先としては」という条件節によってフレーズ全体の極性を 4 とすることも認める．

また，リビュー文の中にはニュートラルともネガティブとも取れる要求を述べる表現も含まれるが，今回のデータのうちの 5%程度の文にしか含まれないため，特別なラベルは付与せず極性の判定を行なった．

4　句単位の評価極性タグ付け

前節で述べた仕様に基づき，日本語のリビューコーパスに対して句単位の評価極性アノテーションを行った．単語分割には KyTea 0.4.7 [5]，句構造解析には Ckylark [6]を使用した．句構造解析の結果得られた構成素のうち，末尾が P で終わる非終端記号（例: “動詞 P”）のカバーする単語列をフレーズとして抽出した．フレーズの前後の文脈を考慮せずにアノテーションするため，SST と同様に抽出したフレーズをランダムに並べ替えてからアノテーションを行った．

本研究では日本語リビューコーパスとして筑波大学文単位評価極性タグ付きコーパス[7]（以下, TSUKUBA コーパス）を使用した．楽天トラベルのレビューデータの 1,000 件（4,309 文）に対して，文単位の評価情

[5] http://www.phontron.com/kytea/index-ja.html
[6] https://github.com/odashi/ckylark
[7] http://www.nlp.mibel.cs.tsukuba.ac.jp/~inui/SA/corpus/readme-euc.txt

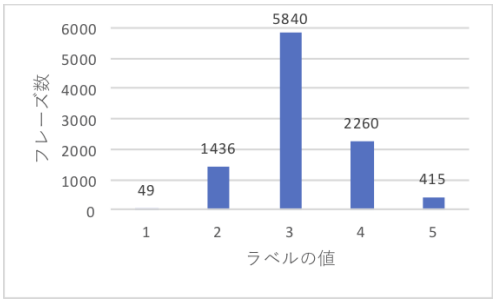

図 1: TSUKUBA コーパスから抽出してアノテートした 10,000 フレーズの極性ラベルのヒストグラム.

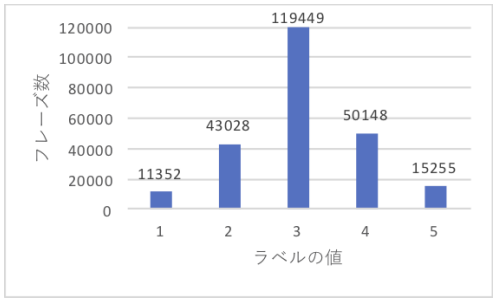

図 2: SST の 239,232 フレーズの極性値を 5 値に置換した際のヒストグラム.

報が付与されたコーパスである．TSUKUBA コーパスには，文単位で 6 値の評価極性情報が付与されているが，本コーパスでは，このコーパスに対し，改めて句単位で 5 値の評価極性情報を付与した．

TSUKUBA コーパスから抽出した異なり 59,758 フレーズに，情報系の大学生 3 人で評価極性を付与した．全体のアノテーションをする前にトレーニング期間を設け，アノテータ間で揺れる部分について議論した．10,000 件については 3 人全員が，残りの 49,758 件は 2 人ずつでアノテーションを行った．

全員が極性を付与した 10,000 件については，最大と最小の極性の差が 2 以上あるフレーズを抽出し，極性を見直した．すり合わせをした後のアノテータ間の一致率 (Fleiss' κ) は 0.70 であり，高い一致率と判断できる．3 人のアノテータが付与した極性の平均値をとり四捨五入したヒストグラムを，図 1 に示す．

全 59,758 フレーズそれぞれに付与された極性の平均値を正式なスコアとし，それぞれのアノテータが付与した極性とともに公開する．

5 考察

日本語のリビューコーパスに句単位で評価極性のアノテーションを行なった結果について以下で考察する．

5.1 SST との比較

本研究では多くの点を SST にならってアノテーションを行った．英語と日本語で言語は異なるが，同様の仕様で行ったアノテーションにどのような特徴が見られるか比較する．SST で公開されている各アノテータの極性値を仕様どおりに 5 値に置換し，その平均値を四捨五入した際の分布を図 2 に示す．我々が行ったアノテーションの一部の結果（図 1）と分布を比較すると，どちらもニュートラルの割合が約半分を占めていること，ネガティブよりポジティブのほうが多いことなど傾向が似通っていることがわかる．相違点として，TSUKUBA コーパスは約半分の文がポジティブと判定される一方，ネガティブと判定される文は 2 割程度なので，文単位ではポジティブとネガティブの割合がほぼ半々である SST と比較すると，ネガティブなフレーズの数が少ないと考えられる．また，1 と 5 が SST と比較して少ないのは，本コーパスでは 1 フレーズのみで 1 と 5 をつけるのを許さなかったせいだと考えられる．

5.2 アノテーションの不一致

推定する文脈が一致しない場合　推定する文脈が複数考えられる場合に，どう推定するかで判定が一致しない場合があった．例えば，「宿探し中に見つけたこのプランは 2 名利用で目を疑う価格でした」のように，評価を表す語が含まれているにもかかわらずフレーズ外の推定する文脈の違いで，いい意味（価格が安い）で捉えるか悪い意味（価格が高い）で捉えるか判定が揺れる事例があった．今回のアノテーションではフレーズ外の文脈を参照しないため，これらのケースの曖昧性を解消することは困難である．

複数の極性のフレーズがある場合　フレーズの中に極性に影響を与える複数のサブフレーズが存在した場合，極性の判断が揺れる場合があった．例えば，「設備の古さが多少気になりますが、清掃は行き届いていました」は，「設備の古さが多少気になる」がネガティブ，「清掃は行き届いていた」がポジティブであるが，これらのサブフレーズの極性の度合いがアノテータによって異なるため，フレーズ全体としての極性が変わった．これらのケースではそれぞれの極性をどの程度重視するかは主観によるため，本質的に曖昧性がある．

表記揺れと語感が不一致の場合　読みは同じでも，表記が違うことによって評価が分かれる場合もあった．例として，「面白い臭いが」は，「匂い」ではなく「臭

い」という表記を用いることでネガティブに判定された場合と,「臭い」自体はニュートラルな表記だと判定された場合があった．さらに，語感の違いによりアノテータ間で判定が揺れる場合があった．例えば,「リンスインシャンプー」については，使いやすいと思うかどうかが読み手によって異なるため，アノテータ間で判定が一致しなかった．今回は読み手の主観でアノテーションしているため，文脈の少ない短いフレーズに対し揺れを解決することは難しい．

5.3 アノテーションの限界

ラベルの確信度 今回，ポジティブとネガティブの曖昧性があるフレーズはニュートラルを付与することとしたが，極性に曖昧性がない事例との区別ができないという問題があった．アノテートの際，複数のラベルの付与を許すような仕様にしたり，あるいはそれぞれのラベルの確信度を付与できるようにすることで，曖昧性のないニュートラルとそれ以外を区別できるようになると考えられる．

強意表現 強意表現は，普通は「非常に」や「すごい」のような短いフレーズとして後続する表現を連用修飾するが，本コーパスでは「ビジネスホテルでは見たことの無い」など節単位で強意表現になるものがあり，先行する文脈に強意対象がある場合もあった．今回のアノテーションは極性を持つフレーズを対象としてアノテーションを行なったが，修飾先のフレーズの極性を変化させるフレーズ（e.g. 強意表現，否定表現）については分けてアノテートし，どのような振る舞いをしているか検討することが今後の課題である．

モダリティ モダリティにかかわる表現を含むフレーズに関しては，今回は特別に考慮せず極性を付与したためアノテータ間で判定に揺れが生じる事例があった．例えば,「絶景の渓谷、もう少しライトあった方が綺麗」のように**条件**（もう少しライトあった方が）と**極性をもつ語**（綺麗）が含まれているフレーズでは，条件が付いているため実際にはそうではないとの判断でニュートラルとするか,「綺麗」の極性を採用しポジティブとするかで揺れた．また,「ぜひ男性スタッフも笑顔で同様に対応して下さい。」という**要望**を含むフレーズでは，男性ができていないという事実に着目してネガティブなのか，女性はできているという含意部分に着目してポジティブなのかアノテータ間で判断の揺れが生じた．このような事例について，より詳細な情報を抽出するためには，Matsuyoshi ら [4] で用いられている拡張モダリティのようなラベルを別途用意するとよいと思われる．

6 おわりに

本研究では，日本語のレビューコーパスを対象とし，句単位の評価極性付与に取り組んだ．今回構築したコーパスは，Stanford Sentiment Treebank（SST）のアノテーション基準に則り，TSUKUBA コーパス（楽天トラベル）全体にアノテーションしている．アノテーションの分析結果から，今回の仕様に基づくアノテーションでフレーズ単位の評価極性の一致率が高いコーパスを構築可能であることが分かった．また，アノテーションの不一致が起きやすい箇所についても議論した．

本データセットのアノテーションは SST に従ってフレーズ単位でランダムにアノテータに提示し，読み手の主観的な評価をアノテートしたデータセットであるが，文単位で書き手の主観的な評価をアノテートした KNB コーパスに同様の基準で付与することで，読み手にとっての評価と書き手にとっての評価の違いについて比較する必要がある．フレーズより広い文脈を見ることによるフレーズ単位のアノテーションの揺れの確認を含めて検討していきたい．

参考文献

[1] Hiroshi Kanayama and Tetsuya Nasukawa. Textual demand analysis: Detection of users ’ wants and needs from opinions. In *Proc. of COLING 2008*, pp. 409–416, 2008.

[2] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proc. of EMNLP 2013*, pp. 1631–1642, 2013.

[3] Jiwei Li, Thang Luon, Dan Jurafsky, and Eduard Hovy. When are tree structures necessary for deep learning of representations? In *Proc. of EMNLP 2015*, pp. 2304–2314, 2015.

[4] Suguru Matsuyoshi, Megumi Eguchi, Chitose Sao, Koji Murakami, Kentaro Inui, and Yuji Matsumoto. Annotating event mentions in text with modality, focus, and source information. In *Proc. of LREC 2010*, pp. 1456–1463, 2010.