レビューテキストを用いた条件付き意見文の抽出

中山 祐輝　　藤井 敦
東京工業大学大学院情報理工学研究科計算工学専攻

1 はじめに

情報化技術の発展により，個人が様々な情報を手軽に発信できるようになった．その中には，利用者の意思決定支援に有益であるような主観的な評判情報が膨大に含まれている．評判情報とは「商品 A の○○が良い」という「評価」や「商品 B の××がうれしい」という「感情」が記述されている情報である．

しかし，情報過多により有益な評判情報を利用者自身で発見することは，多大な時間コストを要する．よって，評判情報の抽出や整理を自動的に行い，評判検索や商品推薦に応用できる技術が求められている．本論文では，これらの技術を総称して評判分析と呼ぶ．近年，レビュー投稿サイト，ブログサイト，Twitter などのメディアから，評判分析が盛んに行われている．そのタスクの一つに，テキストから対象，属性，評価表現を抽出し，極性を付与するタスクがある [1]．例えば,「ホテル A のアメニティが良い」において，対象は評判の対象を示す「ホテル A」である．属性は「アメニティ」のような対象の特徴や構成要素を表す．評価表現は「良い」のような主観的な評価を表す語である．極性は，評判が肯定的か否定的かを示し，上記の例文では肯定が付与される．

既存の研究では，評判に関する記述が全ての利用者や状況に当てはまることを想定している．しかし，評判はある条件をもって成立する場合がある．例えば,「ホテル B は、ミナミエリアで遊ぶには 良いロケーション。」という意見文において,「良い」という評価表現は「ミナミエリアで遊ぶには」という条件をもって成立する評価表現であり，ビジネス目的の宿泊客に当てはまるとは限らない．つまり，このような条件を考慮することで，ある利用者における評判情報の過大もしくは過小評価を軽減し，個々の利用者に当てはまる評判をより獲得できる．本論文では，意見文中に評価表現が成り立つための条件を含む文を**条件付き意見文**，その条件を**評価条件**と呼ぶ．

我々は評価条件を 5 種類に分類する．評価条件を分類することで，特定の利用者に向けた評判のみを知りたいなど，評判検索時の条件の絞り込みや極性を持たない評価表現を排除することに役立つ．以下で，評価条件を含む例文を種類ごとに示す．なお，本論文中で示される全ての例文は，実際に楽天トラベルのレビューサイトに投稿された例文である．本論文中の全ての下線は，評価条件を表す．

- **利用者**：特定の利用者で評価表現が成立する
  （例）身長が 185cm あるので ベッドが小さい．
- **場面**：特定の場面や時点で評価表現が成立する
  （例）朝食は、クロワッサンやパンで、一泊だけならいいですが、連泊するときは 物足りない。
- **目的**： 特定の利用目的で評価表現が成立する
  （例）部屋は綺麗でこじんまりと ビジネス利用には十分です。
- **反実仮想**：現在は条件を満たしていない．しかし，いつか満たせば評価表現が成立する
  （例）料理は 金額が安ければ 満足です。
- **視点**：ある視点から見ると評価表現が成立する
  （例）ホテルの規模を考えれば, この値段は満足。
  （例）ビジネスホテルで東京では 一番です。

本研究は評価条件を考慮することで，評判分析の高精度化を目的としている．そのために，本研究では評価条件の抽出と分類,「出張には」と「ビジネスには」のような評価条件の同義表現をまとめるためのクラスタリングを行っていく．その第一段階として，本研究ではレビューテキストから評価条件を抽出する手法を提案した [2]．本論文では条件付き意見文および評価条件を抽出する新たな手法を提案し，評価実験によって本手法の有効性を示す．

2 関連研究

評価条件は，評判の妥当性に影響を及ぼすという点で原因や理由と関連している．本研究と類似するタスクに因果関係抽出 [3, 4, 5] がある．因果関係抽出とは入力文から二つのイベントを抽出し，イベント間の関係を特定するタスクである．例えば,「津波」と「地震」というイベントを抽出し,「地震」が「津波」の「原因」という関係を同定する．しかし，Inui ら [5] は＜節 1，手がかり語（ため），節 2 ＞，Chang ら [3] と Girju[4] は＜名詞句 1，動詞，名詞句 2 ＞の構文パターンに絞っており，1 節で挙げた例文のすべてがこれらの構文パターンに当てはまらないため，既存手法を評価条件の抽出に応用することは困難である．Kim ら [6] は，“The service was terrible because the staff was rude” のような，理由を含む意見文を同定する手法を提案している．しかし，彼らの目的は評判を正当化する理由を同定することであり，我々の目的とは異なる．我々は，理由の中で利用者に依存する理由を評価条件として扱う．利用者に依存す

る理由とは，理由を命題として与えたときに，全ての利用者で命題が真とならないことが明らかな理由であると定義する．例えば，利用者における評価条件の例で挙げた「身長が 185cm ある」という命題は「ベッドが小さい」の理由を表しており，特定の利用者にしか真とならないことは明らかである．一方,「駅直結なので立地は良いです。」では,「駅直結である」という命題は利用者によらず真になるので，利用者に依存しない理由とする．

本研究と同様に，評判分析における条件に焦点を当てた研究として Narayanan ら [7] がある．彼らは，条件文の極性が肯定，否定，中立かを判定する手法を提案している．Narayanan らは，入力を条件文としているのに対し，本研究はレビュー記事全文を入力としている点で異なる．また，本研究は条件付き意見文を検出し，その中で評価条件となる表現を抽出するので，タスクが全く異なる．さらに，本研究では “If” のような明示的な接続詞を含む条件に加えて,「〜には」など接続詞を含まない条件も対象としている点でも異なる．

3 条件付き意見文の抽出手法

本研究はレビュー文を文節毎に区切り，各文節が評価条件を構成する文節か否かを同定する問題として扱う．対象，属性，評価表現の抽出は本研究の対象外であるので，既存の手法によって予め抽出する．Nakayama ら [2] は，対象，属性，評価表現のいずれかを含む文節を除く各文節が I 文節か O 文節かを分類する 2 値分類問題として扱い，分類器の学習に SVM を用い，五つの素性を提案した．本論文では，評価条件を構成する文節を I 文節と呼ぶ．O 文節とは，対象，属性，評価表現のいずれかを含む文節と I 文節以外の文節を指す．また，対象，属性，評価表現のいずれかを含む文節は，I 文節でないという仮定をおき，分類の対象とはしない．

しかし，Nakayama らは以前までの出力結果を素性として利用しておらず，文全体でみたときに尤もらしい出力結果かどうかは明らかでない．それに対して本手法は，図 1 のように入力列 $\boldsymbol{x} = x_1 ... x_n$ に対して尤らしいラベル列 $\boldsymbol{y} = y_1 ... y_n$（$y_t \in \{I, O, \text{対象}, \text{属性}, \text{評価表現}\}$）を出力する系列ラベリング問題として扱うことで精度向上を試みる．学習アルゴリズムとして CRF を用いる．本手法は Nakayama らと同様に，一つ以上の I 文節を含む文を条件付き意見文として検出し，I 文節が連続した文節のかたまりを評価条件として抽出する．

次に，図 1 に基づき，CRF における素性関数の設計の際に用いる 8 個の素性について説明する．素性 A〜C は，Nakayama らが提案した素性である．それらに加えて，我々は新たに素性 D〜H の素性を提案する．素性 A〜E は構文に基づく素性，素性 F〜H は語彙に基づく素性である．図 1 の例文は，文節ごとに区切られており，文節番号および各文節の係り受け結果が付加されている．文節番号は，後の文節になるにつれて大きくなるように各文節に付加されているとする．

- **素性 A：評価表現との到達距離**

  評価条件は評価表現を修飾し，O 文節と比べて評価表現の近傍に出現しやすい．よって，I 文節は評価表現との到達距離が小さくなる傾向にある．到達距離とは，ある文節から係り受けをたどって，目標の文節にたどりつくまでに，通過した文節数を指す．我々は到達距離を素性 A として用いる．図 1 で，文節番号 1 と 4 の素性値はそれぞれ 2 と 3 になる．なお，評価表現の後に続く文節は「だと思う」などのモダリティ表現が付加されやすく，評価条件になりにくい．よって，文節番号 9 の文節のように評価表現へ到達しない文節の素性値は-1 とする．

- **素性 B：評価表現との文節距離**

  素性 A は係り受け解析の誤りに頑健ではない．我々は素性 A の欠点を補完するために，到達距離を評価表現との文節距離によって近似し，評価表現との文節距離を素性 B として用いる．文節距離とは文節番号の差を表す．もし，文中に評価表現が複数含まれていれば，文節距離の最小値を素性値とする．図 1 で，文節番号 1 の素性値は 7 である．文節番号 9 のように，評価表現との文節距離が負となる文節の素性値は-1 とする．

- **素性 C：属性への到達可能性**

  評価条件は評価表現を修飾する傾向にあるので，属性を修飾することは少ない．よって，素性 C は属性に到達可能かを素性値とする．図 1 において，文節番号 1，4 そして 7 の素性値はそれぞれ 0，1，1 となる．

- **素性 D：属性との文節距離**

  素性 C もまた，係り受け解析の誤りに頑健ではない．我々は，素性 C を属性との文節距離によって近似する．図 1 で，文節番号 0 の素性値は 2 である．文節番号 3〜9 のように，属性との文節距離が負となる文節の素性値は-1 とする．

- **素性 E：素性 B と素性 A の差**

  評価条件は一つ以上の連続した文節からなる．また，I 文節は末尾の I 文節を除いて他の I 文節に係る．すなわち，I 文節における素性 A と素性 B による素性値の差が比較的小さくなると考えられる．図 1 で，文節番号 1 と文節番号 3 の素性値は，それぞれ 5 と 1 である．

- **素性 F：末尾表現の有無**

  評価条件を構成する文節のうち，末尾の I 文節には特定の助詞もしくは助動詞が含まれやすい．これは助詞や助動詞が，動作の行われる相手，時，場所，目的を限定したり，条件節をなす用法を持っているからである．本論文では特定の助詞と助動詞を末尾表現と呼び，16 種類の末尾表現を辞書に登録する．辞書には,「に」,「は」,「なら」などが登録されている．素性 F は，文節中にどの末尾表現が含まれるかを素性にする．

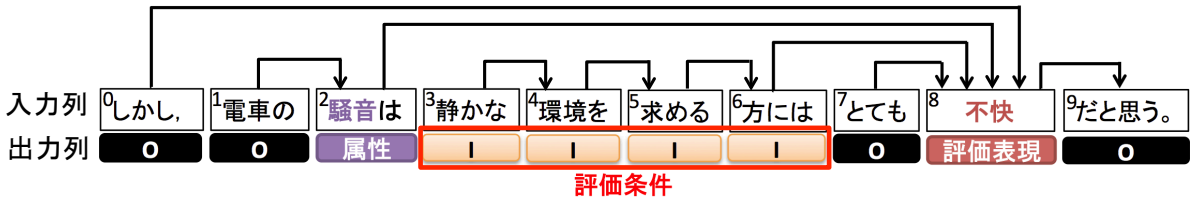


図 1: 各素性の説明に用いる例文

- **素性 G：末尾表現への到達可能性**

  I 文節は，末尾表現を含む文節に到達する傾向がある．例えば図 1 のように，文節番号 3〜5 の I 文節は，「に」と「は」の末尾表現を含む文節に到達可能である．そこで，どの末尾表現を含む文節に到達可能であるかを素性値とする．ただし文節番号 6 の文節のように，文節内に末尾表現を含む場合は，その文節内に含まれる末尾表現を素性値とする．なお，係り受けは評価表現に到達するまでたどる．

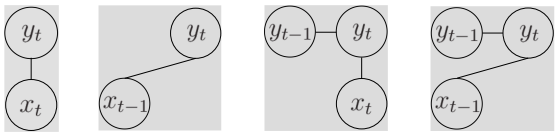
- **素性 H：主辞の品詞**

  主辞の品詞が副詞や接続詞である文節は，I 文節となりにくい．図 1 のように「とても」という文節は，評価表現の極性を強調する働きを持つため，I 文節ではない．一方，主辞の品詞が名詞や動詞である文節は I 文節になりそうである．

次に，8 個の素性に基づく素性関数の設計について説明する．我々は，図 2 で示す四つのパターンが条件付き意見文の抽出に寄与すると仮定する．U は unigram の素性パターン，B は bigram の素性パターンを表す．例えば，U00 は x_t と y_t の観測素性であり，位置 t を図 1 の文節番号 4 に対応させると，次のような素性関数 f_k が生成される．

$$f_k = \begin{cases} 1 & \text{if } x_{t,A} = 3 \ \& \ y_t = I \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

$x_{t,A}$ は位置 t の文節における素性 A の素性値である．素性関数は，残りの素性 B〜H についても同様に計算される．すなわち，U00 と U01 からは（$\sum_i$ ラベル数 × 素性 i における素性値の異なり数）個の素性関数がそれぞれ生成される．同様に B00 と B01 からは，（$\sum_i$ ラベル数 × ラベル数 × 素性 i における素性値の異なり数）個の素性関数が生成される．



U00　U01　B00　B01

図 2: 素性関数に用いる四つの素性パターン

4 評価実験

4.1 実験方法

本手法の有効性を検証するために，楽天と国立情報学研究所が提供している楽天トラベルのレビューデータ (2010 年 7 月 13 日時点で 34 万 8564 件) の一部を用いた．このレビューデータから 580 件に対して対象，属性，評価表現タグを付与する．その後，作業者 2 人に評価条件のタグ付けを行ってもらった．作業者間の一致率は κ 値で 0.73 を達成した．形態素解析と係り受け解析は MeCab[1] と CaboCha[2] をそれぞれ用いた．

実験 1：条件付き意見文検出の評価

実験 1 では，本手法がコーパスから条件付き意見文をどれだけ検出できるかを評価する．我々の構築したコーパスの総文数は 3,157 文で，評価条件を含む文 792 文，含まない文 2,365 文で構成されている．

実験 2：評価条件抽出の評価

実験 2 は，実験 1 において条件付き意見文が完全に検出できたことを想定し，条件付き意見文 792 文を入力として，本手法が評価条件を正しく抽出することができるかを評価する．792 文の総文節数は 7741 文節であり，I 文節 2,210 文節，O 文節 2,984 文節，対象，属性，評価表現のいずれかを含む文節 2,547 文節から成る．

実験 1，2 ともに文を単位とした 10 分割の交差検定により学習を行う．構築されたモデルの入力をテスト事例として，テスト事例が正しく予測されるかを検証する．本論文は，ベースラインと Nakayama らによる手法との比較を行う．ベースラインはルールベースの手法であり，以下の手順にしたがう．まず，末尾表現を含み，かつ評価表現との到達距離が 1 である文節を I 文節として抽出する．そして，そのI 文節から係り受けを逆にたどることで到達した文節を I 文節として抽出する．

4.2 実験結果と考察

表 1 に実験結果を示す．Rule と SVM はそれぞれベースラインと Nakayama らの手法を表している．本手法は，8 個の素性と四つの素性パターンに基づく素性関数を生成し，CRF を学習した手法である．表 1 から，本手法はベースライン手法と Nakayama らの手法に比べ，優れた F 値を達成したことがわかった．また，実験 1 の正解率と両方の実験における F 値について，両側 t 検定

[1] http://code.google.com/p/mecab/
[2] http://code.google.com/p/cabocha/

表 1: 実験結果

	実験 1				実験 2		
	精度	再現率	F 値	正解率	精度	再現率	F 値
Rule	.667	.686	.674	.836	.477	.503	.482
SVM	.657	.799	.720	.847	.603	.646	.615
本手法	.868	.729	.791	.903	.715	.736	.721

を行ったところ，Rule と本手法ならびに SVM と本手法の間に有意水準 1%で有意な差があった．

我々は，8 個の素性が有効に働いているかを検証するために，素性のいずれかを除いたときの手法との比較を行った．その結果，本手法が最も高い F 値を示した．よって，8 個の素性は有効に働いていることがわかった．CRF において，実験 1 では U01 と B00 のパターンを除いて学習した手法が最も F 値が高かった．実験 2 では本手法が最も高い F 値を達成した．

我々は実験 2 における本手法の誤り分析を行い，誤りの原因を類型化した．誤りは，構文解析の誤りに起因する誤りが 98 事例，本手法に起因する誤りが 163 事例，原因不明が 39 事例の計 300 事例である．構文解析の誤りに起因する誤りは，係り受け解析誤り，品詞タグ付けの誤り，そして不正な文末区切りが原因である．これらの誤りは，本手法に起因する誤りではないので詳細は割愛する．以下，本手法に起因する誤りの一部を説明する．

一つ目は，素性 F における末尾表現を含むことによる誤りである．「野菜たっぷりで 親は 満足、子供は 不満という感じです。」の文では，「野菜たっぷりで」の文節に「で」という末尾表現が含まれていたことにより I 文節として誤抽出された．この誤りは事例数が一番多く，優先的に対策を講じるべきである．

二つ目は，素性 F，G において「ので」の末尾表現が有効に働かなかったことによる誤りである．「ので」の接続助詞は，利用者に依存する理由に現れる．しかし，「駅直結なので」など評価条件ではない理由が学習データに多く含まれていたため，I 文節として誤抽出された．

三つ目は，素性 F と遷移素性を用いたことによる誤りである．「ビジネスホテルで東京では 一番です。」の「ビジネスホテルで」という文節に末尾表現を含むため，「東京では」とう文節が誤って O 文節とみなされた．つまり，末尾表現を含みラベルが I から O に遷移する素性におけるパラメータのコストが高くなったため，誤ったと考えられる．

四つ目は，素性 G において到達する末尾表現の種類に起因する誤りである．「京都や奈良の文化などに興味のある方でしたら」という評価条件中には，「に」と「たら」の末尾表現を含む．しかし「京都や」の文節は，この場合末尾表現としてふさわしくない「に」に到達してしまう．すなわち，望ましい末尾表現に到達することができかなったことにより，すべての文節が O 文節であると出力された．

五つ目は，素性 H に起因する誤りである．「朝、ゆっくりしたい方にも おすすめです。」という文において，「ゆっくり」の文節が O 文節であると誤って出力された．3 節で説明したように，ある文節における主辞の品詞が副詞である場合，O 文節である可能性が高い．したがって，素性 H の重みが強くなり誤った．この誤りに対する対策として，評価条件内の副詞は評価表現に係ることはないので，評価表現に係るかを素性として追加することが考えられる．

5 おわりに

評判は全ての利用者に当てはまらず，ある条件をもって成立する場合があり，従来のタスクでは考慮されていなかった．本論文ではこのような条件を含む文を検出し，それらの条件を抽出する手法を提案した．実験において，提案手法はいくつかの誤りのパターンが見受けられた．

今後は，条件付き意見文検出において有効な素性の設計を行っていく．また，一つ目に挙げた誤りの軽減，利用者に依存する理由の抽出に有効な手法を検討する．

謝辞

本研究の一部は，科学研究費補助金基盤研究（B）（課題番号 22300050）によって実施された．

参考文献

[1] B. Liu and L. Zhang, A Survey of Opinion Mining and Sentiment Analysis. In C.C. Aggarwal and C.X.Zhai, editors, *Mining Text Data*, pp.415–463, Springer, 2012.

[2] Y. Nakayama and A. Fujii : Extracting Evaluative Conditions from Online Reviews: Toward Enhancing Opinion Mining, In *Proc. of the 6th International Joint Conference on Natural Language Processing*, pp.878–882, 2013.

[3] D-S. Chang and K-S. Choi : Causal Relation Extraction Using Cue Phrase and Lexical Pair Probabilities, In *Proc. of 1st International Joint Conference on Natural Language Processing*, pp.61–70, 2004.

[4] R. Girju : Automatic Detection of Causal Relations for Question Answering, In *Proc. of the ACL 2003 workshop on Multilingual summarization and question answering*, pp.76–83, 2003.

[5] T. Inui, K. Inui, and Y. Matsumoto : Acquiring Causal Knowledge from Text Using the Connective Marker tame, *ACM Transactions on Asian Language Information Processing*, No.4, Vol.4, pp.435–474, 2005.

[6] S-M. Kim and E. Hovy : Automatic Identification of Pro and Con Reasons in Online Reviews, In *Proc. of the COLING/ACL 2006 Main Conference Poster Sessions*, pp.483–490, 2006.

[7] R. Narayanan, B. Liu, and A. Choudhary : Sentiment Analysis of Conditional Sentences, In *Proc. of the 2009 Conference on Empirical Methods in Natural Language Processing*, pp.180–189, 2009.