意見文の対象読者を限定する条件の抽出

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

ウェブ上には，ニュース記事に対するコメントやカスタマーレビューなどの意見テキストが膨大に蓄積されている．これらを有効利用するために，意見マイニングの研究が盛んである．意見マイニングとは，意見テキストの集合から知識を発見する処理の総称であり，基盤技術として意見抽出が重要である．以下の例文を用いて具体的に説明する．

(1) ホテル A は，子供連れの家族旅行ならリーズナブルな価格だと思います。

既存の意見抽出手法 [1, 2] は,「対象＝ホテル A，属性＝価格，評価（極性）＝リーズナブル（肯定）」のような要素を抽出することで，意見情報の基本単位を定型化する.「評価」は「リーズナブル」のような評価表現や「肯定／否定」の極性を値とすることが多い．

意見テキストの集合から意見の基本単位が大量に抽出されると，ある対象における評価の概観が容易となる．例えば，ホテル A の品質改善を試みる場合は，ホテル A に関する意見から否定的な評価を受けやすい属性を調べるだろう．

ここで，例文 (1) 中のホテル A に対する肯定的な評価（「リーズナブルな価格」）は,「子供連れの家族旅行なら」という条件の下で成立しており,「出張目的のシングル利用」にとっては「リーズナブル」ではないかもしれない．従来の意見マイニング手法は条件付きの意見を考慮していないため，対象を過大もしくは過小に評価している可能性がある．

旅行サイトのレビュー 7 000 文を人手で調査したところ，意見文のうち 28%が条件付きであった．すなわち，既存の意見マイニングにおける分析結果には最大で 28%の誤差が含まれる．

上記の問題には少なくとも二通りの対処方法がある．一つは，何らかの条件が付いた意見文を検出して分析対象から削除する消極的な方法である．もう一つは，条件付き意見文を検出し，その条件に該当する利用者の属性を特定して役立てる積極的な方法である．例文 (1) では「子供連れ」で「家族旅行」を計画している利用者である．言い換えれば，当該意見文を読むべき読者の属性を特定することである．以降本論文では，上記のような条件を**評価条件**と呼び，例文中の下線は評価条件を表す．

広義の評価条件とは，意見が成り立つための条件である．狭義の評価条件とは，対象読者がある属性に限定される条件である.「女性には ホテル A がお得です。」という意見文は，対象読者が「女性」という属性に限定されており,「男性」がこの意見文を読んでも意味を成さない．よって,「女性には」は狭義の評価条件である．一方,「改装すれば ホテル A に泊まりたい。」という意見文内の「改装すれば」は，対象読者の属性が定義できないため，狭義の評価条件でない．

本研究ではこれまで，広義の条件付き意見文を抽出する手法を提案してきた [5]．対して本論文では，狭義の評価条件に焦点を当て，狭義の評価条件を抽出する手法を提案する．

評価条件を広義と狭義に区別する意義は，応用が異なる点にある．広義は条件付きの意見とそうでない意見の分離に役立つのに対して，狭義に絞ることは対象読者における意見の可視化に役立つ．

本研究では，まず例文 (1) のように意見性のある記述を含む条件文に注目する．加えて,「観光には ホテル A がお得です。」などの条件節を含まない形も評価条件とする．さらに，例文 (2) のような理由も評価条件とする．

(2) 子連れの家族旅行なので，ホテル A はリーズナブルな価格だと思う。

(3) 無料の朝食がもう利用出来ないので，ホテル A は価格面でリーズナブルではない。

例文 (2) は，例文 (1) の意味と厳密に同じではなく，条件節が評価の理由に置き換わっている．この理由もまた対象読者を限定するため，評価条件とする．しかし，例文 (3) の「無料の朝食がもう利用出来ない」は，属性によって対象読者が限定されないので，すべての理由が本論文の抽出対象ではない．

2 関連研究

本研究と混同しやすい事例として，Narayanan ら [3] は例文 (4) や例文 (5) のような文法上の条件文に対する極性分類手法を提案した．

(4) ホテル A は，価格がリーズナブルでなかったら、生き残ることができなかっただろうに。

(5) もしリーズナブルな価格をもつホテルを探しているなら、ホテル A に泊まりなさい。

彼らの焦点は，例文 (4) のように「リーズナブル」といった評価表現を含むにもかかわらずホテル A に対する意見ではない条件文や，例文 (5) のように条件文全体としてホテル A に対する肯定的な意見を識別する点に

ある．しかし，条件節がホテル A に対する評価の対象読者を限定している訳ではない．

それに対して，本研究は文法上の条件文かどうかを問わずに,「ホテル A における評価の対象読者が，ある属性に限定される条件を抽出すること」が目的である．

Kim ら [4] は,「スタッフが無礼だったので、サービスがひどかった。」のような意見中の評価における理由を特定する手法を提案した．しかし，彼らの目的は評価の理由を同定することであり，例文 (3) のように対象読者を限定しない理由も含む．それらの理由は，本論文が対象とする評価条件ではない．

3 提案手法

本論文のタスクは，レビューテキストから狭義の評価条件を抽出することである．提案手法は評価と対応する評価条件が同文に存在するとし，文ごとの抽出を想定する．提案手法は，レビュー文が与えられたとき，まず意見の基本単位を抽出する．そして，ある対象の評価が見つかれば対応する評価条件を探す．なお，意見の基本単位を抽出する処理は既存手法によって既になされているとし，本論文では評価条件の抽出に焦点を当てる．

提案手法は評価条件の抽出を BIO チャンキングとしてモデル化する．しかし，日本語での評価条件は評価条件の開始位置を表す明らかな特徴がないので，B ラベルは使用しない．提案手法は文節をトークンとみなし，I ラベルと出力された文節の連続部分を評価条件として抽出する．本論文では「I」と「O」の代わりに,「cond」と「other」の用語をそれぞれ用いる．

定式化すると，文節の系列 $\boldsymbol{x} = x_1 \dots x_n$ が与えられたときに，ラベルの系列 $\boldsymbol{y} = y_1 \dots y_n$, $y_i \in \{Cond, Other, Target, Aspect, Opinion\ word\}$ を予測するタスクである．なお，$Target, Aspect, Opinion\ word$ はそれぞれ対象，属性，評価表現を含む文節のラベルであり，事前に特定されている．実際のタスクは，それら三つを除いた文節に対する $y_i \in \{Cond, Other\}$ の 2 値分類である．分類器の学習には，条件付き確率場 (CRF) を用いる．提案手法は，unigram モデルと bigram モデルによる linear-chain CRF を用い，条件付き確率は式 (1) によって計算される．

$$p(\boldsymbol{y}|\boldsymbol{x}) = \frac{1}{Z_{\boldsymbol{x}}} \exp\Big(\sum_{i,k} \lambda_k \cdot f_k(y_i, \boldsymbol{x}) + \sum_{i,k} \mu_k \cdot g_k(y_{i-1}, y_i, \boldsymbol{x})\Big) \tag{1}$$

ここで，$Z_{\boldsymbol{x}}$ は，$p(\boldsymbol{y}|\boldsymbol{x})$ を確率分布にするための正規化因子であり，f_k と g_k はそれぞれ unigram と bigram モデルにおける素性関数である．k は素性関数の固有番号である．unigram モデルでは，y_i の出力ラベルは $x_{i-1,v}$ もしくは $x_{i,v}$ に依存し，bigram モデルにおいて y_i の出力ラベルは $x_{i,v}$ と y_{i-1} の組合せ，もしくは $x_{i-1,v}$ と y_{i-1} の組合せに依存する．つまり，素性関数は以下に示すパターンのいずれかから成る．

- P1: unigram $x_{i,v}$
- P2: unigram $x_{i-1,v}$
- P3: bigram $y_{i-1}\ x_{i,v}$
- P4: bigram $y_{i-1}\ x_{i-1,v}$

ここで，$x_{i,v}$ は x_i における素性 v の値である．提案する素性 v は図 1 に基づいて後述する．図 1 は，各文節における素性値の例を示している．図 1 上部分の四角形，その内部の数字，矢印はそれぞれ文節，文節番号，文節間の係り受け関係を表す．また，下部分は各文節における提案素性 1–13 の値を示す．素性関数は，変数 (例えば, f_k における $x_{i,v}$, y_{i-1}, y_i を指す) の値で考えられる全ての組合せパターン数だけ生成され，対応するパターンが真となれば 1，それ以外は 0 を返す．

次に，提案する素性について説明する．素性 1–7 については，紙面の都合上素性値の計算方法のみを説明する．それらの素性は中山ら [5] も用いており，素性の詳細は彼らの論文を参照してほしい．本論文では中山らと異なる素性 8–13 について主に説明する．

素性 1: 評価表現までの係り受け距離　素性 1 として，注目文節と評価表現を含む文節との係り受け距離，つまり評価表現に到達するまでにたどった係り受けの数を用いる．もし，注目文節と評価表現を含む文節との間に係り受けの経路がなければ，素性値は-1 とする．

素性 2: 評価表現までの文節距離　素性 2 の値として注目文節と評価表現を含む文節間における文節番号の差を用いる．注目文節と評価表現を含む文節における文節番号の差が負であれば，素性値を-1 とする．

素性 3: 属性への到達可能性　注目文節と属性を含む文節間に係り受け経路があれば 0，それ以外は 1 とする．

素性 4: 属性までの文節距離　注目文節と属性を含む文節間における文節番号の差を用いる．属性までの文節距離が負であれば，素性値を-1 とする．

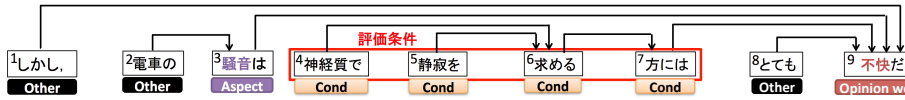
素性 5: 素性 2 と素性 1 の素性値の差　素性 5 の値として，素性 2 と素性 1 における素性値の差を用いる．

素性 6: 文頭の文節　素性 6 の値は文頭の文節であれば 1，そうでなければ 0 とする．ただし，主辞の品詞が接続詞である文節は例外として素性値を 0 にする．

素性 7: 主辞の品詞

素性 8: 手掛り表現の有無　評価条件は，しばしば特定の助詞や助動詞などの語で終わる．そこで素性 8 の値として，手掛り表現の有無 (1/0) を用いる．本手法は，網羅的な手掛り表現を収集するために，日本語機能表現辞書 [6] で収録されている表現を用いる．日本語機能表現辞書は「に」のような機能語と「にたいして」といった機能語のように働く複合辞を見出し語として持ち，各見出し語に一つ以上の意味カテゴリが付与されている．日本語機能表現辞書の中で，本論文の目的にあう意味カテゴリ 19 種類を選択し，それらのカテゴリに属する用語 388 語を手掛り表現とする．選択した意味カテゴリの代表例と手掛り表現の例を表 1 に示す．

素性 9: 手掛り表現の意味カテゴリ　素性 8 は，各表現の頻度がスパースになる危険性があるため，各表現を意味カテゴリに抽象化した素性を用いる．また，手掛り表現の意味カテゴリを素性に加えることは，狭義の評価



素性1	1	2	Aspect	4	3	2	1	1	Opinion word	-1
素性2	8	7	Aspect	5	4	1	2	1	Opinion word	-1
素性3	1	0	Aspect	1	1	1	1	1	Opinion word	1
素性4	2	1	Aspect	-1	-1	-1	-1	-1	Opinion word	-1
素性5	7	5	Aspect	1	1	1	1	0	Opinion word	0
素性6	0	0	Aspect	0	0	0	0	0	Opinion word	0
素性7	接続詞	名詞	Aspect	名詞	名詞	動詞	名詞	副詞	Opinion word	動詞
素性8	手掛り表現なし	手掛り表現なし	Aspect	手掛り表現なし	手掛り表現なし	手掛り表現なし	には	手掛り表現なし	Opinion word	手掛り表現なし
素性9	手掛り表現なし	手掛り表現なし	Aspect	手掛り表現なし	手掛り表現なし	手掛り表現なし	話題	手掛り表現なし	Opinion word	手掛り表現なし
素性10	手掛り表現なし	手掛り表現なし	Aspect	には	には	には	には	手掛り表現なし	Opinion word	手掛り表現なし
素性11	手掛り表現なし	手掛り表現なし	Aspect	話題	話題	話題	話題	手掛り表現なし	Opinion word	手掛り表現なし
素性12	対象読者表現なし	対象読者表現なし	Aspect	神経質	対象読者表現なし	対象読者表現なし	対象読者表現なし	対象読者表現なし	Opinion word	対象読者表現なし
素性13	対象読者表現なし	対象読者表現なし	Aspect	外見・能力・性格	対象読者表現なし	対象読者表現なし	対象読者表現なし	対象読者表現なし	Opinion word	対象読者表現なし

図 1: 各文節における素性値の例

条件と広義の評価条件を区別する手掛りとして期待できる．例えば，表 1 の「に」と「ために」はそれぞれ状況，目的の意味をもつため，狭義の評価条件に偏って出現することが予想される．

素性 10: 手掛り表現を含む文節との係り受け経路の有無　素性 8 で説明したように，評価条件は末尾の文節に一つ以上の手掛り表現を含む．加えて，評価条件は一つ以上の文節で構成される．これらにより，手掛り表現を含む文節を修飾する文節は，cond 文節の可能性がある．そこで，注目文節と手掛り表現を含む文節との係り受け経路の有無 (1/0) を素性 10 とする．なお，注目文節が手掛り表現を含む場合は，手掛り表現を含む文節に到達していると考え，係り受け経路があるとする．例えば，図 1 の文節 4-7 は,「には」の手掛り表現を含む文節 7 との間に係り受け経路をもつ．

素性 11: 係り受け経路によって到達する手掛り表現の意味カテゴリ　素性 9 と同様の目的で，素性 10 を意味カテゴリに抽象化した素性を用いる．例えば「には」の手掛り語は，表 1 より「話題」の意味を持つため，図 1 の文節 4–7 における素性値は「話題」となる．

表 1: 提案手法で用いる機能表現と意味カテゴリの例

意味カテゴリ	手掛り表現
話題	には，となると，にいたると，とくると
目的	ために，のに，に
状況	に，にあたって，になると，さいに
対象	にかんして，にたいして，にかけると

素性 12: 対象読者の限定に関する表現　文節中に含まれる対象読者に関する表現を素性とする．対象読者に関する表現の収集方法として,「ちょんまげ英語塾[1]」に収録されているすべての用語 8414 語からなる辞書を構築する．ちょんまげ英語塾は，対象読者の手掛りとなる「身分・世代」などのカテゴリ別に用語が収録されているため，本論文の目的に合った様々な用語を収集できる．表 2 にその代表的なカテゴリと表現を示す．例えば，図 1 の文節 4 は「神経質」という素性値になる．

[1]http://mage8.com/tango/(閲覧日: 2015 年 1 月 20 日)

素性 13: 対象読者の限定に関する表現の意味カテゴリ　素性 12 は表層形を用いており，各表現の頻度がスパースになる危険性があるため，各表現を意味カテゴリに抽象化した素性を用いる．図 1 の文節 4 は「外見・性格・能力」という素性値になる．

表 2: 対象読者の限定に関する表現と意味カテゴリ例

意味カテゴリ	対象読者の限定に関する表現
外見・性格・能力	身長，太った，社交的な，素人の
人間関係・身分・世代	子供，父，男性，結婚記念日
体の各部位	上半身，背中，腰，足
病気・ケガ	アレルギー，腰痛，認知症

4 評価実験

提案手法の有効性を評価するために，楽天トラベルのレビューテキスト[2]34 万 8564 件の一部を用いた．このデータセットから，580 件のレビューテキストを選択し，人手で意見の基本単位を同定した．本論文の焦点は評価条件の抽出なので，意見抽出の出力として人手で特定された対象，属性，評価表現を用いた．なお，すべての文節が対象，属性，評価表現のいずれかを含む文は，定義により評価条件を含まないため除外してコーパスを作った．表 3 にコーパスの詳細を示す．このコーパスを用いて 10 分割交差検定により学習とテストを行った．

表 3: 実験で用いるコーパスの詳細

(a) 文単位

意見文	条件付き	504
	条件なし	1 553
意見文でない		1 099
総文数		3 156

(b) 文節単位

cond 文節数		1 266
other 文節数		17 575
意見単位	評価表現	3 764
	属性	3 400
	対象	132
総文節数		26 137

正解基準として「部分正解」と「正解」を用いる.「部分正解」は，テスト文が評価条件を含むか否かを検出できるかの評価基準であり，比較的緩い評価である．対して「正解」は，評価条件の範囲を完全に特定することが

[2]http://www.nii.ac.jp/cscenter/idr/rakuten/rakuten.html

要求され，厳しい評価である．評価尺度は精度，再現率，F 値，正解率を用いた．

比較手法として二つの手法を用いた．一つ目の手法は，二つの規則に基づいた手法で， 本論文では「Rule」と呼ぶ．一つ目の規則は，評価条件が評価表現を修飾しやすいという特徴と末尾の文節に手掛り表現を含みやすいという特徴に基づいている．二つ目の規則は，cond 文節が手掛り表現を含む文節と係り受け経路をもつという特徴に基づいている．具体的にはまず，評価表現までの係り受け距離が 1 であり，かつ手掛り表現を含む文節を cond 文節と出力する．次に，その文節と係り受け経路が存在する全ての文節を cond 文節と出力する．二つ目の手法は，3 節で説明した素性 1–13 を用いて SVM を学習した．提案手法は，13 種類の素性と素性関数における 4 種類のパターンを用いて CRF を学習した．

表 4 に実験結果を示す．表 4 より，提案手法は「部分正解」と「正解」の両基準におけるすべての評価尺度で，いずれの比較手法よりも優れた結果を達成した．また，F 値と正解率について両側 t 検定を行い，両基準で Rule と提案手法との間に有意水準 1%で統計的に有意な差があった．また，両基準の F 値について SVM と提案手法との間に有意水準 1%で統計的に有意な差があった．

表 4: 実験結果

	部分正解				正解			
	精度	再現率	F 値	正解率	精度	再現率	F 値	正解率
Rule	.345	.754	.473	.734	.180	.203	.186	.671
SVM	.397	.879	.546	.768	.230	.248	.235	.849
CRF	.672	.638	.651	.891	.391	.385	.385	.861

図 2 に，「正解」における 13 種類の素性と素性関数における 4 種類のパターンの有効性を示す．横軸は，素性 X もしくは素性関数のパターン X を除いた場合の手法である．縦軸は，各評価値における提案手法と各手法との比率である．すなわち，もし素性 X を除いた手法の値が 1 より小さければ，素性 X は評価条件の抽出において有効であることを意味する．図 2 より，提案手法はいずれかを除いた手法と比較しても最も高い F 値を達成した．したがって，13 種類の素性と素性関数における 4 種類のパターンは，評価条件の抽出において独立して有効であり，併用することがさらに有効であった．「部分正解」においても同様の傾向が得られた．

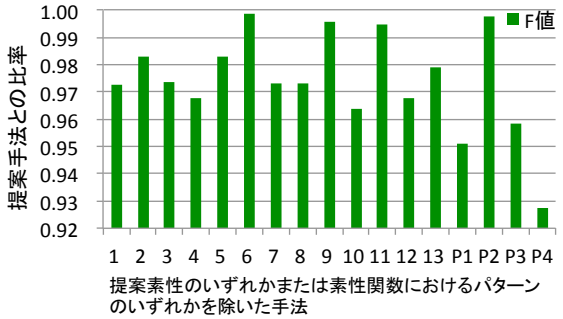


図 2: 提案した「素性と素性関数のパターン」の有効性

提案手法が正しく評価条件を抽出できた例を示す．図 2 より，対象読者の限定に関する表現を用いた素性 12 と素性 13 が特に有効である．そこで，二つの素性を加えたことによって新たに抽出できた事例を示す．「怖がりの娘が一番喜んでいました。」と「学生にとっては ありがたいホテルです。」は，辞書にそれぞれ「娘」や「学生」という世代に関する表現が登録されていたため，抽出できた．

一方，二つの素性を加えても抽出できなかった例として「潔癖症の方でもない限り 許容できる範囲の設備。」と「老齢のため 大浴場、レストランの近くで助かりました。」は，それぞれ「潔癖症」と「老齢」という対象読者に関する表現が辞書に登録されておらず，抽出できなかった．辞書に登録されていない表現への対処方法は今後の課題である．

素性 12 と素性 13 を加えたことで誤った例もある．「食事の量は 酒飲みには じゅうぶんな量です。」は，「酒」という「料理のメニュー・食材」のカテゴリを持つ用語として認識され誤った．現状では，単純な文字列マッチングによる認識方法をとっており，「酒飲み」という形態素単位でマッチングを取る必要がある．また，辞書にはない「〜飲み」や「〜好き」などの嗜好に関する表現を収集することでこの誤りに対応できる．

5 おわりに

意見マイニングの基盤技術として意見抽出を行う様々な手法が提案されてきた．しかし，既存の意見マイニングでは意見の妥当性が条件によって限定される場合を考慮しておらず，意見マイニングの質の低下が懸念される．この問題を解決するために，本論文では対象読者を限定する評価条件の抽出手法を提案した．本論文の貢献は，評価条件の概念を導入した点，対象読者を限定する評価条件の抽出手法を提案した点にある．

参考文献

[1] Pang, B., Lee, L.: Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, Vol.2, No.1-2, pp.1–135 (2008)

[2] Liu, K., Xu, L., Zhao, J.: Syntactic patterns versus word alignment: Extracting opinion targets from online reviews. In *Proc. of ACL*, pp.1754–1763 (2013)

[3] Narayanan, R., Liu, B., Choudhary, A.: Sentiment analysis of conditional sentences. In *Proc. of EMNLP*, pp.180–189 (2009)

[4] Kim, S.-M., Hovy, E.: Automatic identification of pro and con reasons in online reviews. In *Proc. of COLING/ACL on Main Conference Poster Sessions*, pp.483–490 (2006)

[5] 中山祐輝, 藤井敦: 意見マイニングにおける条件付き意見文の抽出. 第 7 回 Web とデータベースに関するフォーラム (2014)

[6] 松吉俊, 佐藤理史, 宇津呂武仁: 日本語機能表現辞書の編纂. 自然言語処理, Vol.14, No.5, pp.123–146 (2007)