対象－属性－評価の３項関係同定による評判情報抽出

土田正明　　水口弘紀　　久寿居大
NEC インターネットシステム研究所
{m-tsuchida@cq,hironori@ab,kusui@ct}.jp.nec.com

1　はじめに

口コミやレビューの情報が商品購入やサービス選択の参考に用いられるようになってきた．掲示板やレビューサイト，ブログなどの情報源から評判を自動的に収集することができれば，商品購入の際の評判検索，マーケティング，企業のリスク管理など，幅広い用途への利用が期待できる．

本研究では,「ある対象物の属性（評価をしている点）についての評価の情報」の対象物，属性，評価の３項を評判情報と定義し，これを抽出することを目的とする.「対象物」は，商品やサービスなど,「属性」は対象物の特徴や性質や構成要素,「評価」は評価者の主観的な評価を表す表現である．

先行研究では，特定の商品やサービスについての掲示板やレビューサイトなどの書き込みから属性と評価の２項同定がなされている [2]．しかしながら，収集できる評判が，掲示板やレビューサイトが存在する対象に限られてしまう．また，収集できる量も限られる．他にもレビューサイトをどのように見つけるか，サイト自体がなくなってしまうリスク等の問題もある．

我々は，幅広く大量の評判情報を収集するため，日々大量の情報が発信され，記事収集も容易であるブログを対象に評判を抽出する研究を進めている．本稿では，まずブログ記事を対象に評判情報の書かれ方を分析する．次に，分析結果を踏まえた３項関係同定法を提案し，その評価について報告する．

2　ブログ記事の分析

ブログ記事における３項関係の特徴を明らかにするため，ブログ記事での評判情報の書かれ方を調査した．具体的には，1) ３項関係にある対象，属性，評価の出現位置の特徴，2) 属性，評価に用いられる表現，を調査した．調査にあたり，ブログ記事から評判タグ付きコーパスを作成した．タグは，記事に対象物，属性，評価がそろって書かれている３項関係に付与した．対象物は TV 番組で，992 記事に 2914 個の評判情報が含まれた．

まず，対象物，属性，評価の出現位置の特徴を，属性と評価の位置関係，対象物と属性の位置関係に分けて調査した．

対応関係にある属性と評価の出現位置を比べると，属性と評価が同じ文で 87 %，属性が評価の１文前で 9 %であり，これらで全体の 96 %を占めていた．また，属性と評価が係り受け関係にあるものは 52 %であった．係り受け関係のみによる抽出では不十分であるが，対応関係は１文前まで探せば十分であると分かった．

次に，対象物と属性の出現位置の特徴を表１に示す．表１の相対距離は，対象物と属性の間の形態素数である．マイナスの値は，属性が対象物の前に書かれていること示す．評判数は，相対距離を 100 毎に区切り集計した．表１に示す範囲外にも評判は存在するが少量なので省略する．表１の −100〜0 と 1〜100 には評判数に顕著な差がある．このことから,「評判は先に対象物を明記して書かれることが多い」と言える．また，距離が離れるに従い評判数が単調に減っていることから,「評判は対象物から近い範囲に書かれやすい」と分かる．対象が属性や評価と同じ文に存在する例はほとんど無く，係り受けなど文構造の手がかりを使うことはできない．

表 1: 対象物と属性の相対距離

相対距離	-100〜0	1〜100	101〜200	201〜300	301〜400
評判数	94	1200	529	294	153

次に，評判を構成する属性と評価の表現について調査した．2914 個の評判情報のうち，属性表現の異なり数は 1602 個，評価表現の異なり数は 1793 個であった．また，1 回しか用いられなかった表現が全体の約 80 %を占めた．これは，対象が多様な表現で評価されているため，事前に表現を網羅的に用意するのは困難であることを意味する．

以上から分析結果を整理すると，1) 属性と評価は比較的近くに現れるが係り受け関係のみでは不十分，2) 対象物と属性は離れている場合が多く同じ文に存在することはほとんどない，3) 属性と評価は多様な表現で評判がかかれ，事前にそれぞれの表現を網羅的に用意するのは困難，の 3 点である．次節より，分析結果を踏まえた 3 項関係同法の方針とその詳細を説明する．

3　３項関係同定による評判抽出法

テキストから（対象物－属性－評価）の 3 項関係を同定する方法を述べる．2 節から，対象物と属性の出

現位置の特徴は，属性と評価の特徴と大きく異なることがわかった．そこで，対象物と属性と評価の 3 項関係を，1) 対象物と属性の関係，2) 属性と評価の関係，に分けて同定する．また，多様な表現で評判が書かれているため，属性と評価の表現辞書を自動増殖させる．

概要を図 1 に示す．**対象物辞書**，**属性辞書**，**評価辞書**はあらかじめ用意する．入力は，処理対象の**テキスト文書**である．

我々は，3 項関係同定を，1) 対象物の出現位置を特定し，対象物について書かれた部分を特定する (**対象物トピック限定**)，2) 属性らしい表現と評価らしい表現を抽出し，属性と評価の辞書に未登録の表現を追加 (**属性・評価辞書自動増殖**)，3) 対象物トピックテキスト内の属性と評価の関係を同定し，トピックの対象物との 3 項関係として同定 (**属性－評価の関係同定**)，の 3 ステップで行なう．

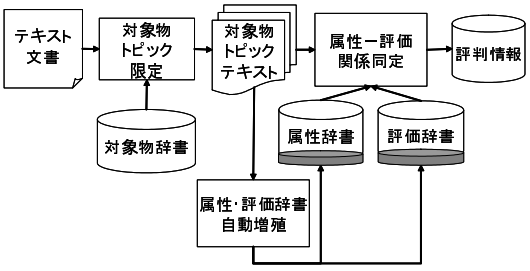

図 1: 処理フロー

対象物トピック限定　対象物トピック限定では，属性－評価の対と対象物を対応付けるため，記事中で対象物がトピックとなっている部分を特定する．2 節の分析から，評判は「対象物の出現箇所の後ろに書かれる」，「対象物から離れるほどかかれなくなる」，という知見が得られた．

そこで，1) 記事の先頭から対象物辞書の表現を探索し，対象物がマッチしたら別の対象物が出現するまでその対象物のトピックとする,2)1) で，出現箇所から後 n 形態素を超えても別の対象物が出現しない場合には n 個目の形態素を含む文までをトピックとする，の 2 つの戦略で特定する．対象物より前にも少量の評判が存在していることもあるので，対象物の前 m 形態素を含む文も対象物トピックとして含める．

属性・評価表現の辞書自動増殖部　事前に属性や評価の辞書を網羅的に登録することは困難であるため，入力記事から属性や評価らしい表現を抽出することで，属性や評価表現の自動増殖を行う．問題設定が，固有表現抽出に類似しているため，固有表現抽出法を応用する．

具体的には，単語を 1）属性の始まり (B-ATTR)，2）途中 (I-ATTR)，3）評価の始まり (B-EVAL)，4）途中 (I-EVAL)，5）1〜4 以外 (NONE)，に分類する．素性は，分類対象の単語と前後 m 単語の「表層文字列，単語の原形，品詞，カタカナを含むか，アルファベットを含むか，数字を含むか」と，前 n 単語の分類ラベルを用いる．

属性－評価の関係同定　対象物トピック限定により，掲示板やレビューサイトからの属性－評価の関係同定と同じ問題設定と見なせるようになる．また，2 節の分析により，属性と評価が係り受け関係にないものが半数以上あることが確認されている．そこで，照応解析の手法を応用し，係り受けにない属性と評価の関係同定を実現した飯田ら [2] の手法をベースに拡張する．

飯田らは，属性－評価の関係同定を，1）属性－評価の対同定，2) 同定された対の主観性判定，の 2 ステップで行なっている．以下概要を説明して，問題点を明らかにし，拡張方法を説明する．

属性－評価の対同定で用いられるトーナメントモデルは，１つの評価に複数存在する属性候補間で勝ち抜き戦を行い，最後まで勝ち残った候補と評価を対応関係とする．図 2 に具体例を示す．図 2 の上側は対象物トピックテキストであり，属 n は属性の候補、評 m は評価の候補である．図 2 の下側は「かっこいい」に対するトーナメントモデルの解析例である．

今回の月 9 は (対象物：ドラマ A) だ。(属 1:予告) を見る限り相変わらず (属 2:俳優 A) は (評 1:かっこいい) けど、(属 3:ストーリー）は（評 2：面白い）のかな？

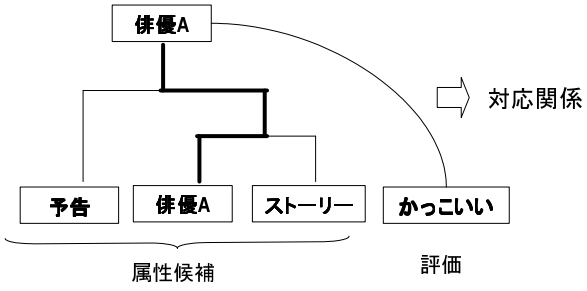

図 2: トーナメントモデルの解析例

しかしながら，例からも分かるとおり，トーナメントモデルは評価に複数の属性が対応する場合，原理的に同定できない．例えば,「俳優 A や俳優 B はかっこいい。」の両方を同定することは出来ない．

我々は，複数の属性の対応を同定するため，トーナメントモデルと属性の言語的出現パターンを併用する．具体的には，トーナメントモデルにより属性候補が同定された後に，パターンを適用し，当てはまる属性を対応関係として同定する．典型的なパターンには,「(属 1) と (属 2)」や,「(属 1) も (属 2) も」，など並立関係が挙げられる．パターンは再帰的に適用する．例えば,「映像もストーリーも音楽も最高！」という文から,（音

楽－最高）が同定されたとして,「(属 1) も (属 2) も」というパターンを再帰的に適用すると,(映像－最高),(ストーリー－最高）が抽出される．

2) の主観性判定では，対応と同定された属性と評価の対が主観的に書かれたか否かを判定する．例えば，図 2 の（ストーリー－面白い）は，対応関係にはあるが,「ストーリが面白い」と評価しているわけではない．主観的でない属性－評価対の代表的な種類には，仮定，伝聞，疑問がある．

属性－評価対の同定，主観性判定は 2 値分類問題であり，任意の教師つき学習法を用いることが出来るが，我々はサポートベクターマシンを用いる．学習方法や使用する素性などの詳細は，文献 [2] に従う．

最終的には，対応関係にあり，かつ主観性がある対と，トピックの対象物を 3 項関係として抽出する．図 2 の例では，(ドラマ A －俳優 A －かっこいい) となる．

4 評価実験

本節では，対象物トピック限定，属性・評価辞書自動増殖，属性－評価の対応同定の拡張の有効性を評価する．最後に，提案法全体と，飯田ら [2] の方法を適用した場合との比較で総合的な評価を行なう．

実験データには 2 節で述べた TV 番組の評判タグ付きコーパスを用いる．記事の前処理は，形態素解析に juman と構文解析に knp を用いた．また，全実験を通し，学習にはサポートベクターマシンの多項 2 次カーネル，5 分割交差検定で評価を行なう．また，対象物トピック限定のパラメータである前 m 形態素の m は経験的に 10 とした．

実験１：対象物トピック限定の有効性評価

対象物トピック限定による精度と再現率へ影響を評価する．各種辞書は，タグ付きコーパスのタグから対象物，属性，評価の表現を全て抽出して生成した．つまり，評判を構成する全ての表現が登録されている状態である．対象物トピック限定のパラメータである後 n 形態素の n を 200,400, ∞ (限定なし) で精度と再現率を測定した．

結果を表 2 に示す．トピックの範囲を広げると精度は低下し，再現率が上昇することが確認された．限定する場合は，しない場合に比べて精度が約 +0.20，再現率が −0.10 であるため，再現率への悪影響よりも精度への効果が大きい．

表 2: 対象物トピック限定による精度と再現率

	精度	再現率
$n = 200$	0.63(750/1193)	0.26(750/2914)
$n = 400$	0.59(877/1483)	0.30(877/2914)
限定なし	0.41(1101/2658)	0.38(1101/2914)

実験２：属性・評価の辞書自動増殖の有効性評価

属性と評価を自動抽出すると当然間違いも含まれるため，再現率は向上するが精度が下がってしまうと考えられる．そこで，辞書自動増殖の有効性を評価する．

人が事前に属性，評価の辞書を作成した場合を擬似的に実現させるため，実験 1 の属性，評価の辞書から，ランダムに 8 割を抽出して新たな辞書を作成した．つまり，2 割の登録漏れがある状態である．対象物トピック限定のパラメータは 400 とし，精度と再現率を測定する．辞書増殖は，交差検定の各テスト前に，全テスト記事から対象物トピック限定により対象物トピックテキスト生成して，それら全てに対して行なった．

結果を表 3 に示す．精度は低下し，再現率が上昇している．

表 3: 辞書自動増殖による精度と再現率

	精度	再現率
増殖なし	0.56(633/1135)	0.21(633/2914)
増殖あり	0.51(759/1476)	0.26(759/2914)

実験３：属性－評価の対応同定の拡張方法の有効性評価

属性の言語的出現パターンを適用して，複数の属性の対応を同定する方法の有効性を評価する．言語的出現パターンは,「(属性 1) と (属性 2)」など，並立関係の 10 パターンを定義した．

属性と評価の関係同定は，対応同定と主観性判定からなる．対応同定にトーナメントモデルを用いた場合と提案方法を用いた場合で，精度と再現率を測定する．各種辞書は，実験 1 と同様のものを用いた．対象物トピック限定のパラメータは 400 で行なった．

結果を表 4 に示す．提案方式により，新たに 28 の属性－評価の対が抽出された．一方で，精度は低下する結果となった．

表 4: トーナメントモデルと拡張方式の比較

	精度	再現率
トーナメントモデル	0.59(877/1483)	0.30(877/2914)
提案方式	0.56(905/1609)	0.31(905/2914)

実験４：従来法と提案法の総合評価

従来法は，飯田ら [2] の方法で同定された属性と評価の対に最も近い対象物を対応付ける（対象物トピック限定で $n = \infty$ と同等）方法とし，提案法との精度と再現率，F 値を比較する．

提案法では，対象物トピック限定のパラメータは 400 として，1) トピック限定のみを行なう場合と，2) さらに辞書自動増殖，複数属性の同定を用いる場合で比較する．辞書は，実験 2 と同じものを用いる．これは，現実的には，全ての表現を網羅的に登録することが不可能であり，実環境での精度や再現率を測定することを目的としたためである．

結果を表 5 に示す．精度と再現率を総合した尺度である F 値では，提案法を全て用いた提案法 2 が最も良い結果となった．

表 5: トーナメントモデルと拡張方式の対応判定

	精度	再現率	F 値
提案法 1	0.56(633/1135)	0.22(633/2914)	0.31
提案法 2	0.50(786/1582)	0.27(786/2914)	0.35
従来法	0.39(786/2025)	0.27(786/2914)	0.31

5 考察

対象物トピック限定 今回のデータでは少数であったが，対象物の出現のみを手がかりにすると,「TV 番組 B でも」や「TV 番組 B にも出演している」など，対象物が主題となっていない場合にミスをする．このように主題となっていない対象物の出現を判定することでさらに改善できる．

また，対象物のトピックが長い記事もあれば短い記事もあるので，距離のみを手がかりにすると，トピックの範囲を的確に特定できない．距離のみではなく，内容をみてトピックの範囲を的確に判断できれば，精度も再現率も向上できる．

今後は，主題でない対象物の出現の判定，内容に応じたトピック範囲の特定方法，を検討する必要がある．

属性・評価の辞書自動増殖 表 3 から分かるとおり，精度が低下している．原因は，ノイズとなる表現が抽出されたため考えられる．しかしながら，属性や評価を「候補になりうる全ての表現」と考えると，ノイズであるか否かを判定することは容易でない．

今後，辞書自動増殖は，1) ノイズの基準を定量化して属性や評価の表現にスコアをつけて判定，2) 辞書構築支援として出力を人がチェックして手動追加，といった方向がよいと考えられる．基準の例としては，対象ドメインとの関連が浅く高頻度な語，などが考えられる．

属性－評価の対応同定の拡張方法 タグ付きコーパスから複数属性が対応する例を分析したところ,「属性 A や属性 B」などの並立関係,「属性 A(属性 B)」の表現がほとんどであり，パターンにより同定できる場合がほとんどであった．

パターンによる対応同定ミスには，1) トーナメントモデルのミスに対してパターンを適用することでミスを増幅させてしまう場合，2) パターンそのものによるミス，の 2 種である．以下では，パターンにより改善できる，2 について考察する．

パターンの精度と再現率を測定するために，前段で全て正しく同定された場合を想定し，タグのある属性にパターンを適用し，精度は「新たに同定された属性が同じ評価を有する割合」，再現率は「複数属性を持つ評価の対応関係を同定できる割合」として測定した．結果，精度 83 %，再現率 76 %であった．パターンにより新たに同定された属性の 17 %はミスという結果であった．

パターンがミスしていた主原因を述べる．タグ付けコーパスでは，属性に階層性がある場合に最下層を正解としてタグをつけている．理由は,「俳優 A の演技が最高！」では,「俳優 A」そのものではなく「演技」を「最高」と評価しているように，属性に階層性がある場合は，最下層の属性を評価しているためである．つまり,「ストーリーと俳優 A の演技は最高！」とある場合,「ストーリー」と「演技」が正解であり,「属性 A と属性 B」にマッチするのは,「ストーリー」と「俳優 A」で,「俳優 A」は不正解となる．

上記の解決には，属性の階層性を考慮することが求められる．小林ら [1] は，属性の階層性の問題を機械学習によりアプローチしている．我々は，属性の階層性を同定すること自体が目的ではないので，パターンの増強により解決できるか，機械学習による方法がよいか検討する必要がある．

総合評価 辞書自動増殖と複数属性の対応同定は，再現率に効果がある．一方，対象物トピック限定の方法を改善すると，精度と再現率の両方を向上できる．さらに，現在の対象物トピック限定の方法は，対象物からの距離のみを手がかりにしているが，出現単語の分布の推移や接続詞など，記事内容に基づきトピックの範囲を特定することで改善できると考えられる．

6 おわりに

本稿では，記事中の対象物，属性，評価の 3 項関係同定による評判情報抽出方法について述べた．3 項関係同定を，1) 入力記事を対象物トピック部分に限定，2) 属性，評価表現抽出による辞書自動増殖，3) 属性－評価関係同定，の 3 つに分けてアプローチした．

ブログ記事を対象とした提案法の評価実験では，対象物トピック限定により精度，辞書自動増殖により再現率，複数属性の関係同定の提案方式による再現率，へ有効であることが分かった．総合では，従来法と同等の再現率で精度を 10 %向上でき，有効性を確認した．

今後，記事内容に基づいたトピック分割の先行研究の成果などを取り入れて対象物のトピック範囲の特定法を改善し，精度と再現率を向上させると共に，属性と評価の関係同定の改善にも取り組んで行きたい．

参考文献

[1] 小林のぞみ, 乾健太郎, 松本裕治. 意見情報抽出のための評価対象・評価視点間の関係同定. 言語処理学会第 12 回年次大会, 2006.

[2] 飯田龍, 小林のぞみ, 乾健太郎, 松本裕治, 立石健二, 福島俊一. 意見抽出を目的とした機械学習による属性-評価値対同定. 情報処理学会研究報告 NL165-4, pp. 21–28, 2005.