レストランドメインにおける意見情報抽出

廣瀬峰史　　小林のぞみ　　乾健太郎　　松本裕治
奈良先端科学技術大学院大学 情報科学研究科
{takafu-h,nozomi-k,inui,matsu}@is.naist.jp

1 はじめに

近年，Weblog を公開する人が急増し，個人が発する有用な情報源の一つとして注目を浴びている．その中でも特に重要視されているのが意見情報である．Web 上に存在する商品やサービスに関する意見情報には，書き手がその商品やサービスを評価しているものなど，主観的な要素が含まれており，抽出を行う対象が明確ではない．そのため意見を抽出するにあたり，意見を何らかの形で定義する必要がある．

意見を形式化して抽出する研究には [3][9] などがある．峠ら [9] は意見文は対象表現，属性表現，評価表現などの組合せから成ると定義し，意見を判別するために手がかりとなる評価表現や強調表現を収集し，その表現を用いて主観的な文章に文単位で意見らしさのスコアを付与する．そしてそのスコアをもとに意見文であるか否かを判定し，抽出している．また，Kobayashi ら [3] は，意見を [対象, 属性, 評価] の三つ組みの形式で定義し，評価の対象となる語の同定を照応解析の先行詞同定の問題に置き換え，先行詞候補をトーナメント形式で戦わせて最尤の候補を同定し，意見性判定を行い，意見の抽出を試みている．

このように三つ組で意見を抽出する問題に焦点があてられてきたが，三つ組で記述できない組も多く存在することが分かってきた．そこで，本研究では，小林ら [4] が提案する意見情報の形式に基づいて，ドメインをレストランに限定し，意見抽出を行う．小林らは，記事の対象を評価している記述のみに着目し，対象，対象の側面・構成要素，構成要素の属性，評価という対象から評価までの関係の連鎖を一つの意見の形式として定義している．本稿ではレストランドメインに限定して抽出を行うため，評価の対象を記事で取り上げられている店名とする．また，対象 (店) を構成する構成要素やその構成要素の部位，属性を総称して**アスペクト**と定義し，対象またはアスペクトに対する評価を示す表現を**評価値**とし，評価値と対となる対象やアスペクトを**評価対象**とする．

図 1 の抽出例から分かるように， 上述の意見抽出には「薄い」という評価値とその評価対象となる「味」を同定する問題と，評価対象である「味」から対象である店名までの階層的な関係の連鎖を同定する問題が存在する． そこで，意見抽出問題を (a) 評価値と評価対象の対を抽出する問題 (**値関係抽出問題**) と (b) 対象から評価対象までの階層的な関係の対を抽出する問題 (**階層関係抽出問題**) の二つの部分問題に分けて考える．本稿では，ドメインに特化した辞書や共起情報などを用いて形式化された意見がどの程度抽出できるのか，逆にどれだけ資源を用意しても抽出できない意見はどのようなものかを分析し，明らかにする．

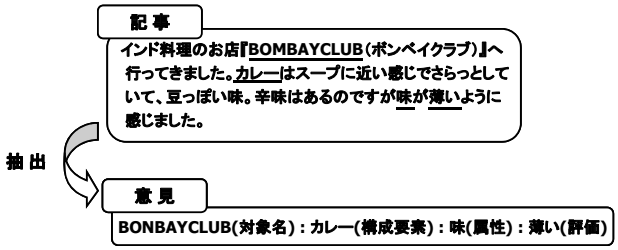

図 1: 意見の抽出例

2 抽出手法

2.1 問題設定

今回は評価値は与えられていると仮定し，以下の手順で意見を抽出する．

1. 候補となる評価対象を辞書を用いて同定する
2. 評価値と対となる評価対象を抽出する
3. 評価対象が店名でなければ，店名にたどりつくまで評価対象間の階層関係を同定する

前処理として，店名や料理名の識別，同一指示対象の同定が必要であるが，今回はそれらは解けていると仮定する．店名，料理名の識別に関しては，以下で述べるドメイン辞書に，コーパス中に存在する料理名と店名を追加し，疑似的に店名と料理名が識別できたと仮定する．

同一指示対象の同定については，以下の例にあるように,「麺」と「手作り麺」は同一指示の関係にあり,「弾力」は両方のアスペクトになるので，あらかじめ評価用の記事内で同一指示対象となる表現を同定しておき，学習と評価の際には，同定された同一指示対象の情報を使用する．

$\langle$麺$\rangle_i$ はやや太目の押し出しの $\langle$手作り麺$\rangle_i$，かなりの 弾力 で...

以下では，それぞれを抽出する手法について詳しく述べる．

2.2 使用する辞書

候補となる評価対象は辞書を用いて同定するが，その他にも評価値表現辞書，共起辞書を使用する．

2.2.1 評価値表現辞書 [5]

評価値表現辞書は，小林ら [5] が作成した評価を表す表現 (形容詞句，副詞句，動詞句，名詞句) 合計 5,402 表現を含む辞書に，さらに「塩辛い」や「水っぽい」などレストランに特化した評価値表現を，レストランのレビュー記事 3160 件から収集し，それを人手でフィルタリングして，281 表現を追加し，合計 5684 表現を含む辞書を作成した．

2.2.2 ドメイン辞書

ドメイン辞書は，レストランドメインに関する名詞，名詞句を，階層的なクラスに分類した辞書である．分類語彙表 [6] の材料，料理のカテゴリから人手で表現を収集し，表 1 に示すような 16 のクラスに分類した．さらに LivedoorBlog と Cookpad.com よりレストランと料理に関する記事，合計 57,702 件 (831,858 文) を収集し，専門用語自動抽出システム [10] を用いて，スコアが平均以上のものレストランに特徴的な名詞，名詞句として抽出した．そして，それらの表現を人手でフィルタリングし，上記のクラスに分類した．

また今回は店名と料理名を識別する処理ができていると仮定するため，意見タグつきコーパス内にある店名と料理名を追加した．辞書は名詞，名詞句など合計 9,993 表現の表現から成る．

2.2.3 共起辞書

評価値とアスペクト，アスペクトとアスペクトがどの程度共起したかを対数尤度比で計算した．その際，以下の式で同時確率をスムージングした．ここで，A はアスペクト，Ca はアスペクトのクラス，X はアスペクトもしくは評価値とする．$C()$ は頻度を表す．

$C(A,X) = 0.5 \times C(A,X) + 0.5 \times C(Ca,X) \times P(A|Ca)$

共起の頻度は，レストランと料理に関する記事 57,702 件 (831,858 文) から，評価値表現辞書とドメイン辞書の中に存在する表現の組み合わせで計算した．

2.3 値関係抽出問題

値関係を抽出する処理の流れを以下に示す．

ステップ 1: 評価値に対し，同一文内から評価対象の候補を抽出し，最も評価値と対になりそうな評価対象を同定する．

ステップ 2: 同定した評価対象が評価値と対となるか否かを判定する．

ステップ 3: 対とならないと判定された場合，前の文もしくは後ろの文から候補を抽出し，最も評価値と対になりそうな評価対象を抽出する．

2.3.1 同一文内における値関係抽出問題

同一文内の値関係同定モデルは Bact[7] を用いて学習する．半構造化情報を考慮した分類モデルである Bact を用いることで，値関係は「A が評価値」のような形で出現しやすいというパタンを学習できると期待できる．訓練の際には，評価値から評価対象までの係り受け木 (正例) と，評価値から対とならないアスペクトまでの係り受け木 (負例) を訓練事例とし，評価値と対となりやすいアスペクトを学習する．ステップ 2 では，ステップ 1 で学習したモデルを用いて評価値に対する評価対象を同定し，その同定した評価対象が真の評価対象となるか否かを分類するモデルを学習する．1 と同様，モデルの学習には Bact を用いる．

表 1: ドメイン辞書の構成

クラス	表現数	例
店名	973	風月，天下一品
店の種類	242	酒屋，カフェ
料理名	3423	うな重，ハヤシライス
材料名	2442	白ごま，豚バラ肉
味	147	味付け，甘味
香り，食感，外観	471	コシ，臭み
食器，調理機具	180	茶碗，伊万里焼
料理の価格	61	料金，時価
店の造り	196	サロン，バルコニー
店の装備	455	客席，レンガ
店の景観	77	眺め，夜景
店の雰囲気	188	クラシカル，BGM
スタッフ	312	店長，女将
サービス	383	食べ放題，オーダーストップ
店の場所	311	裏通り，海辺
店全体	78	自営業，無休

2.3.2 文間における値関係抽出問題

文内に対となる評価対象がないと判断された場合，以下の手順で対となる評価対象を同定する．文間における値関係の同定には，トーナメントモデル [1] を使用する．このモデルは，候補同士を比較しながら，最尤の候補を同定するモデルである．

トーナメントでは，新たに抽出された候補にステップ 2 で負例だと判定された候補を追加する．これは，前段階でシステムが誤って負例だと判定した場合に，正しい対が選ばれる可能性を残すためである．

訓練事例を作成する際には，文内に評価値と対となるアスペクトが存在しないと判断された評価値に対し，文内で同定された評価対象の候補と，新たに候補を前の文もしくは後ろの文から抽出し，対となるアスペクトを正例，他を負例として学習する．

2.4 階層関係抽出問題

ドメイン辞書を用いて同定したアスペクトに対して，2.3 と同様の流れで，文内と文間に出現する場合に分けて，階層関係にあるアスペクトを学習器を用いて決定する．

3 実験

対象とする意見のどのくらい抽出できるのか，逆にどのような意見が抽出できないのかを明らかにするために，評価実験を行った．評価/訓練用のデータとして，意見タグ付きコーパスのレストランドメインの記事 1,389 件を使用し，前節で述べた抽出手法で値関係抽出問題と，階層関係抽出問題を解く実験を行う．比較のため，文内では係り受け関係になっている対を正解として抽出し，文間では一番近くに存在するアスペクトを正解として抽出したものをベースラインモデルとした．評価方法は 5 分割交差検定を用いる．

3.1 意見タグ付きコーパス

意見タグ付きコーパスには，評価値から評価値と対となる評価対象，評価対象からアスペクトを介して店名までリンクが張られている．本稿では，同一指示対象を仮定するため，この意見タグ付きコーパスにさらに，アスペクト同士が同じものを示している場合，照応というタグを新たに付与したものを使用する．

3.2 値関係抽出

Bact に用いた入力素性は共起辞書，日本語語彙大系 [2] の動詞の選択制限，EDR コーパス [8] における人または組織の分類と係り受け関係，ドメイン辞書のクラスの ID，形態素，表層文字列，文頭の文字列，文末の文字列などの統語的情報である．トーナメントモデルに用いた入力素性は共起辞書，ドメイン辞書のクラスの ID，形態素，評価値と評価対象の距離である．

どの素性が有効であるかを検証するために (a) 評価対象のクラス ID を除き，スムージングを行わない共起辞書を入力素性とした場合と (b) 評価対象のクラス ID を除き，スムージングを行った共起辞書を入力素性とした場合，(c) 評価対象のクラス ID を入れ，スムージングを行った共起辞書を入力素性とした場合で比較を行った．それぞれの場合の精度と再現率と，同一文内と文間における精度と再現率の結果を表 2 に示す．

$$\text{精度} = \frac{\text{正しく抽出できた値関係の数}}{\text{システムが対と判断した値関係の数}}.$$

$$\text{再現率} = \frac{\text{正しく抽出できた値関係の数}}{\text{記事の中に存在する値関係の総数}},$$

3.3 階層関係抽出

Bact とトーナメントモデルに用いた入力素性は 3.2 で示したものに，さらに候補の探索元となるアスペクトのクラス ID を追加したものとする．そして，どの素性が有効であるかを検証するため値関係と同様に入力素性を変えて比較を行った．それぞれの場合の精度と再現率と，同一文内と文間における精度と再現率の結果を表 2 に示す．

$$\text{精度} = \frac{\text{正しく抽出できた階層関係の数}}{\text{システムが対と判断した階層関係の数}}.$$

$$\text{再現率} = \frac{\text{正しく抽出できた階層関係の数}}{\text{記事の中に存在する階層関係の総数}},$$

4 解析結果の分析と問題点の考察

4.1 値関係抽出問題

表 2 の結果より，提案手法はベースラインと比較して，精度を下げることなく再現率を向上させることができたことがわかった．特に文間の抽出では，精度と再現率ともにベースラインよりも高い値となった．

文内において，提案手法は文の構造を考慮した学習を行ったのにもかかわらず，係り受け関係のみで抽出を行ったベースラインとほぼ同精度であった．これに関しては，抽出に成功した事例と失敗した事例を比較したところ，失敗した事例では構造の深さが低い特徴量が複数効いて，それがノイズとなって抽出に失敗するという例が多くみられた．このことから，学習の際にある一定の深さをもつ構造木以外は特徴量としないなどの制約を入れるか，または分類を行う際に，特徴量の重みだけでなく，素性の構造の大きさを考慮した分類を行うことを考えている．

また，抽出結果から，文内に存在する値関係には,「味が最高」や「最低の店」など，評価値と評価対象が係り受け関係になっている対が多く見られた．学習時に，評価値と評価対象が係り受け関係にある事例は多く存在するため，係り受け関係にあるという特徴量には高い値の重みが付けられている．その特徴量が効いて，文内の値関係同定では，係り受け関係になっている値関係はかなり高い割合で同定できている．しかし評価値と評価対象の候補が係り受け関係になっていても，間違って抽出された例もある．例えば,「カウンターの狭い店」や「料理が最低の店」など，評価値の係り元と係り先に候補が存在している場合に，誤って判定されていた．このように，文が「[アスペクト $_1$] が [評価値] の [アスペクト $_2$]」という形をしているときは，評価値と対となる候補としてアスペクト $_1$ を選択する必要がある．しかし今回は上記のような構造をうまく学習できていなかった．このような問題は「料理が最低の店」の例では,「最低」のガ格がアスペクトとなるという格の制約，つまり格フレームの情報を素性に組み込むことで，係り受けの関係にあるという情報だけではなく，評価値がどの候補について評価しているのかを明示的に示すことができるため，より評価値と対となる候補を同定できるようになると考えられる．

文間における抽出では，共起情報やトーナメントモデルが効いて，ベースラインよりも高い精度と再現率で抽出することができた．実験結果から評価対象のクラスの ID を素性に入れた場合，精度と再現率が向上した．しかしそれとは逆に文内における精度と再現率が下がった．これはクラスの ID が評価対象の候補を同定する上で有効な素性であるが，逆に文内においてはノイズとなる素性となることが分かった．また，正解となる評価対象は文内に存在するのだが，文内を解くモデルで文間にあると判断されて，それで不正解となる事例が多かった．

4.2 階層関係抽出問題

表 2 の結果から，階層関係に関して，ベースラインよりも精度，再現率ともに向上したことがわかる．文内において，値関係ではあまり精度と再現率の向上が見られなかったが，階層関係ではどちらも向上している．これは，文内の階層関係では係り受けの関係になっている対や「A の B」の関係の構造が学習できているためと考えられる．

また文間においても精度と再現率の向上がみられた．文内と文間それぞれスムージングをした共起情報を素性に入れたことで，少しではあるが精度と再現率が上がった．しかし，クラスの ID を階層関係の両方のアスペクトに素性として入れた場合，文内と文間ともに精度と再現率が下がるという結果になった．これは，今回

表 2: 値関係と階層関係の抽出結果

		ベースライン	Smoothing なし	Smoothing あり	Smoothing+Class
値関係	精度 (全体)	0.5 (1224/2446)	0.506 (1894/3743)	0.456 (1706/3743)	0.496 (1856/3743)
	再現率 (全体)	0.271 (1224/4521)	0.42 (1894/4521)	0.377 (1706/4521)	0.411 (1856/4521)
	精度 (文内)	0.612 (1211/1977)	0.595 (1651/2774)	0.528 (1442/2732)	0.592 (1230/2079)
	再現率 (文内)	0.392 (1211/3086)	0.535 (1651/3086)	0.467 (1442/3086)	0.399 (1230/3086)
	精度 (文間)	0.028 (13/469)	0.251 (243/969)	0.261 (264/1011)	0.376 (626/1664)
	再現率 (文間)	0.009 (13/1435)	0.169 (243/1435)	0.184 (264/1435)	0.436 (626/1435)
階層関係	精度 (全体)	0.074 (276/3755)	0.291 (1423/4882)	0.299 (1461/4882)	0.198 (969/4882)
	再現率 (全体)	0.072 (276/3855)	0.369 (1423/3858)	0.379 (1461/3858)	0.251 (969/3858)
	精度 (文内)	0.168 (242/1442)	0.286 (261/914)	0.293 (276/943)	0.188 (234/1243)
	再現率 (文内)	0.293 (242/826)	0.316 (261/826)	0.334 (276/826)	0.283 (234/826)
	精度 (文間)	0.047 (34/726)	0.293 (1162/3968)	0.301 (1185/3939)	0.202 (735/3639)
	再現率 (文間)	0.011 (34/3022)	0.383 (1162/3022)	0.391 (1185/3032)	0.243 (735/3022)

の実験ではクラスの情報を使用しただけで，クラス間の階層関係の情報を使用しなかったため，クラスの情報が逆にノイズとなってしまったと考えられる．これに関しては，クラスの分類を再考するのと同時に，アスペクト同士の階層性を考慮した素性を追加し，学習を行う必要がある．また，学習の際に対となる候補は上位の階層からしか取らないなどの制約を入れて訓練事例を作成することで，より対象から属性までの関係を考慮した学習モデルができると考える．

他に，階層関係を抽出する上で問題となっているのが，ランチセットなどのセットメニューとセット名の間の関係である．現在は「前菜のコンフィ」と書かれれば，前菜とコンフィは同一指示との関係にあるとしている．しかしながら,「デザートのフォンダンショコラとフレンチトースト」とあると，デザート=フォンダンショコラではなく，これらの間には階層関係があるとしている．このように，集合に関しては，タグの仕様も固まっておらず，今の抽出モデルではうまく扱えていないのが現状である．例えば，アスペクト同士が「ランチ (ご飯，味噌汁，フライ盛り合わせ)」などのパターンを取る場合はある程度抽出可能である．しかし,「これら」などの指示詞は集合を示しており，それらを全て同定するのは人手でも難しい問題である．

今後は，このような集合的な要素も考慮して，どのようにタグを付与すべきか議論していく必要がある．

4.3 訓練/評価データの信頼性

実験結果を解析した結果，照応タグの付与洩れと，タグ付けミスが少なくなかった．タグ付けのミスには「たこ焼 正解 の名店として有名 評価値 な店 A システムの解 である」など，システムの解が正解とみなせる事例が見られた．そのため，タグ付けミスと照応タグの付与洩れに関しては，実験結果の不正解例を元に，間違いだと判断されたタグを人手で修正し，コーパスを洗練していく．

5 おわりに

本稿では意見情報を意見の対象から評価までの階層化された関係の連鎖でとらえ，抽出するドメインをレストランに限定し，ドメインに特化した資源を用いて抽出を行った．値関係においては精度を下げることなく再現率をあげることができた．また，階層関係においてもベースラインよりもよい結果が得られた．しかし，ドメイン辞書を人手で分類したものを用いて抽出を行ったが，それでも抽出は難しく，精度と再現率ともに高いとは言えない結果となった．このことから，このタスクは現在の技術で解くには難しい問題であり，今後はもう少しタスクの簡略化を考えていくことと，効果的な辞書を構成する方法，さらに文の構造を学習する際に，ノイズとなる構造を減らす手法などを考えていく必要がある．

参考文献

[1] 飯田龍, 乾健太郎, 松本裕治. 文脈的手がかりを考慮した機械学習による日本語ゼロ代名詞の先行詞同定. 情報処理学会論文誌, Vol. 45, No. 3, pp. 906–918, 2004.

[2] 池原悟, 宮崎正弘, 白井諭, 横尾昭男, 中岩浩巳, 小倉健太郎, 大山芳史, 林良彦. 日本語語彙大系. 岩波書店, 1999.

[3] Nozomi Kobayashi, Ryu Iida, Kentaro Inui, and Yuji Matsumoto. Opinion extraction using a learning-based anaphora resolution technique. In *The Second International Joint Conference on Natural Language Processing (IJCNLP), Companion Volume to the Proceeding of Conference including Posters/Demos and Tutorial Abstracts*, pp. 175–180, 2005.

[4] 小林のぞみ, 乾健太郎, 松本裕治. 意見情報の抽出/構造化のタスク仕様に関する考察. 情報処理学会研究報告 NL-171, pp. 111–118, 2006.

[5] Nozomi Kobayashi, Kentaro Inui, Yuji Matsumoto, Kenji Tateishi, and Toshikazu Fukushima. Collecting evaluative expressions for opinion extraction. In *Proceedings of the 1st International Joint Conference on Natural Language Processing (IJCNLP)* , pp. 584–589, 2004.

[6] 国立国語研究所. 分類語彙表. 国立国語研究所資料集 6. 秀英出版, 1993.

[7] Taku Kudo and Yuji Matsumoto. A boosting algorithm for classification of semi-structured text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural language Processing, a Meeting of SIGDAT, a Special Interest Group of the ACL Held in Conjunction with ACL 2004*, pp. 301–308, 2004.

[8] 日本電子化辞書研究所. EDR 電子化辞書. EDR, 1997.

[9] 峠泰成, 大橋一輝, 山本和英. ドメイン特徴語の自動取得による web 掲示板からの意見文抽出. 言語処理学会 第 11 回年次大会, pp. 672–675, 2005.

[10] 湯本紘彰, 森辰則, 中川裕志. 出現頻度と連接頻度に基づく専門用語抽出. 情報処理学会研究報告, pp. 111–118, 2001.