B0TB2045

卒業論文

レビューテキストからの属性-評価値抽出

大江 貴裕

2014年 3月 31日

東北大学
工学部 情報知能システム総合学科

レビューテキストからの属性-評価値抽出*

大江 貴裕

内容梗概

近年、インターネット上でユーザー投稿型の評価サイトが増加している。このようなサイトからユーザーの意見を得ることは重要であり、数多くの評判分析が行われている。しかし、レビューには対象の全体の評価が明確な場合が多いが、より詳細である対象に関する評価を知るためにはレビュー本文のテキスト処理が必要となる。本論文では、照応解析の知識を用いて機械学習をすることにより、評価表現と対応する属性の同定を行う。

キーワード

評判分析、属性抽出、照応解析

*東北大学 工学部 情報知能システム総合学科 卒業論文, B0TB2045, 2014 年 3 月 31 日.

目 次

1 序論 **1**

2 関連研究 **3**

2.1 シーケンスモデル 3

2.2 トピックモデル 4

2.3 頻度と関係性に基づいた手法 4

2.3.1 照応解析と見なす手法 5

3 手法 **9**

3.1 属性候補抽出 9

3.2 素性 10

4 実験 **12**

4.1 実験設定 12

4.2 係り受けパスの素性を追加する実験 12

4.2.1 結果 13

4.2.2 考察 13

4.2.3 エラー分析 14

4.3 二値分類とトーナメントモデルの比較 16

4.3.1 結果 16

4.3.2 考察 17

5 まとめ **18**

謝辞 **19**

1　序論

近年、インターネット上でユーザー投稿型の評価サイトが増加している。このようなサイトには企業が調査した情報ではなく、ユーザーの生の声や意見が集まる場所になっている。そのため、多くの消費者が商品を選択するときに参考にし、また企業も商品の評価を得ようとしている。このように評価サイトのレビューは重要な情報源となっており、ユーザーの意見などを抽出することが注視されている。しかし、大半の評価サイトのレビューの形式は商品の全体の評価を点数または☆の数などで、詳しい商品の評価をテキストという形で表している。商品全体の評価は明確に記されているため簡単に抽出することが出来るが、商品の詳細な評価、すなわち商品の側面や一部の評価を理解するためには、レビューテキストを分析する必要がある。

レビューテキストには評価表現とその評価表現に対応する属性が存在する。評価とは書き手の意見であり、(対象, 属性, 評価極性, 評価者, 時間) の 5 つ組から構成される [1]。対象は意見の対象、属性は対象の一部や側面、評価極性は意見がポジティブかネガティブかの極性、評価者は評価する人物、時間は評価した時間である。図 1 を例にとると、(デジタルカメラ, 画質, ポジティブ, 購入者,2014/2/27) の 5 つ組の評価「充分」が存在する。レビューテキストには、商品全体の評価だけでなく属性についての評価も書かれている。本論文では、評価表現と対応する属性を抽出することに重点を当てる。

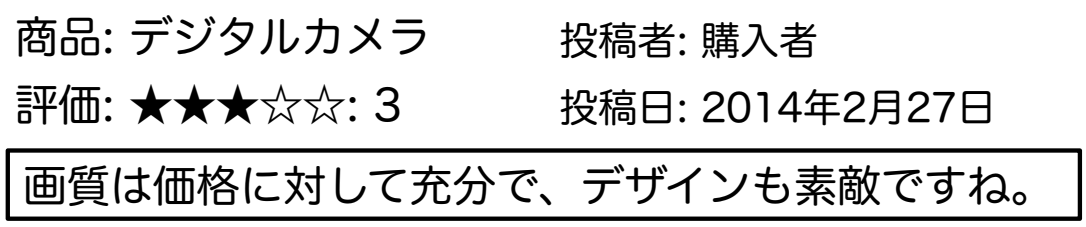

図 1: レビューテキストの例

レビューから評価表現と属性のペアを抽出する手法は多数存在する。それらの手法をまとめたサーベイ論文 [1] があり、論文中にはそれぞれの手法の問題点につ

いて述べられているが、細かいところまでは分からない。そのため本論文では、エラー分析を行うとともに、属性抽出の問題を再整理する。

今回は簡単な手法である照応解析と見なして行う手法に注目する。照応解析と見なす手法では、評価表現を照応詞、属性を先行詞と見なし、照応詞と対応する先行詞を当てる方法と同じようにして評価表現と対応する属性を当てる。これを属性同定と言う。属性同定を既存研究で行われている二値分類で行う手法 [2] とトーナメントモデルを用いる手法 [3] に注目する。これらの手法では、素性として属性候補の表層文字列や品詞などの簡単なものしか使っておらず、構文構造を捉えるような素性は使っていない。そこで Li らの手法 [4] で用いられている構文木のパスの素性を基に係り受けパスの素性を追加する。この係り受けパスと既存の素性をいくつか追加し、二値分類とトーナメントモデルを用いて属性同定を行い、精度が上がることを確認し、エラー分析を行うことでどのような課題があるのかを理解し、問題点を述べる。

2　関連研究

テキストからの評価表現と属性の抽出は多くの試みが行われている。属性抽出の手法は大まかに以下のような三つのカテゴリに分けることが出来る。

- シーケンスモデルを用いる手法
- トピックモデルを用いる手法
- 単語の頻度と関係性に基づく手法

次からこの 3 つのカテゴリの手法について述べていき、それぞれの手法の問題点を明らかにする。また、本論文で用いる手法として、単語の頻度と関係性に基づく手法に属する照応解析と見なす手法を説明する。

2.1　シーケンスモデル

教師ありの機械学習を行うことによって、属性を抽出しようとする考えがある。その多くは情報抽出のタスクで広く使われているシーケンスモデルを使ったものである。シーケンスモデルとは、文を入力に与え単語毎にラベルを付けていくモデルであり、属性抽出を系列ラベリング問題と捉えて行う。シーケンスモデルとして、隠れマルコフモデル (HMM) や条件付き確率場 (CRF) を使う手法がある。

HMM とは固有表現抽出によく使われるシーケンスモデルであり、隠れ状態の系列を持っており、隠れ状態によって出力の確率分布が異なり、隠れ状態を遷移しながら出力していく。Jin ら [5] は、HMM を使ってレビューからの属性抽出を行っている。

CRF とは、系列ラベリング問題を解くのによく使われるモデルであり、状態を遷移しながら出力を行う。状態は一つ前の出力に依存し、状態と観測の素性によって次の状態と出力を決定する。Jakob ら [6] は素性として語彙、品詞、短い依存構造パス、単語の距離を素性とし CRF を用いて、評価表現を含む文から属性を抽出している。

シーケンスモデルを用いる手法では頻度だけではなく、学習することでモデルのパラメータを決定することが出来るが、人手でラベル付けされた学習データが必要となる。

2.2 トピックモデル

トピックモデルは自然言語の分野で幅広く使われている。トピックモデルはドキュメントに様々なトピックが存在し、単語毎にトピックが割り当てられているという考えに基づいており、その潜在的トピックを推定する統計的モデルがトピックモデルとなる。単語の分類の手順は、各単語の背景トピックの初期値を割り当て、ランダムに単語を選択しトピックを変更することを収束するまで繰り返すことで単語の背景トピックを決定する。トピックモデルとして pLSA(ProbabilisticLatent Semantic Analysis)[7] や LDA(LatentDirichlet Allocation)[8] などのモデルが使われている。

トピックモデルはシーケンスモデルと違い人手でラベル付けされた訓練データを必要とせず、属性抽出と属性のグループ化を同時に行うことが出来る。しかし、膨大のデータが必要になるという欠点もあり、また、詳細な分析は行えず、一般的なものや大まかな属性しか抽出できず、より詳細な属性は抽出できず評価表現と対応する属性までは当てることができない。

2.3 頻度と関係性に基づいた手法

属性抽出を行う手法で、出現する単語の頻度や評価表現と属性の間の関係性を用いて抽出するというものがある。レビューでは、評価者がその商品の属性について多く言及するため、属性の出現頻度が高くなる。そこで、単語の出現頻度を測ることにより属性を抽出することが出来ると考えられる。Hu らの手法 [9] では、文書集合中の出現頻度の高い名詞句を属性と見なし、属性の周辺の周辺の形容詞を評価表現と見なし抽出している。また単に単語の出現頻度を測るのではなく、パターンにマッチした評価表現と属性があるとき、マッチしたパターンとその属

性とのPMIを測ることで属性を抽出する手法[10]がある。頻度に基づく手法は単純で効果的だが、出現頻度の低い属性が抽出することが出来ない問題点がある。

属性と評価表現間にはある特定の関係があると考え、その関係を捉えて属性を抽出する手法がある。その一つとして、属性と評価表現間の関係をルールで表しパターンマッチングを行うことにより属性を抽出する手法がある。Qiuらの手法[11]は、レビューテキストを構文木で表現し、属性と評価表現間に成り立つルールのパターンマッチングを行うことで抽出する。ここでのルールとは属性と評価表現の依存関係や、属性が名詞句かつ評価表現が形容詞であるなどの属性や評価表現についての制約に基づいている。この手法では、属性と評価表現の関係のルールだけではなく、属性同士、評価表現同士の関係のルールも用いている。関係性に基づいた手法では、頻度が低い属性も抽出できる一方、属性でない表現にも同様の関係が成り立つ場合に誤った抽出が増えるという問題点があり、正確に抽出できる関係性を捉えるようなルールなどをを作成する必要がある。

2.3.1 照応解析と見なす手法

関係性に基づく手法の一つとして評価表現を照応詞に属性を先行詞と見なし、照応解析の手法を用いることで評価表現と対応する属性を抽出する手法がある。この手法は以下の二つのステップで属性抽出を行う。

- 評価表現、属性候補抽出
- 属性同定

この手法の概要は図2のようになり、赤枠で囲っている単語が評価表現、青枠で囲っている単語が属性候補である。テキスト「映りは価格に対して充分で、デザインも素敵ですね。」が与えられると、評価表現と対応する属性候補集合のペア (評価表現:「充分」属性候補:「映り」「価格」「デザイン」) と (評価表現:「素敵」属性候補:「映り」「価格」「デザイン」) を抽出する。次に、評価表現「充分」に対応するペアを属性候補集合 (「映り」「価格」「デザイン」) から選択し、その結果、属性候補「映り」が選択され (評価表現:「充分」属性:「映り」) の評価表現

と属性のペアが抽出される。評価表現「素敵」に関しても同様の処理を行い (評価表現:「素敵」属性:「デザイン」) のペアが抽出される。

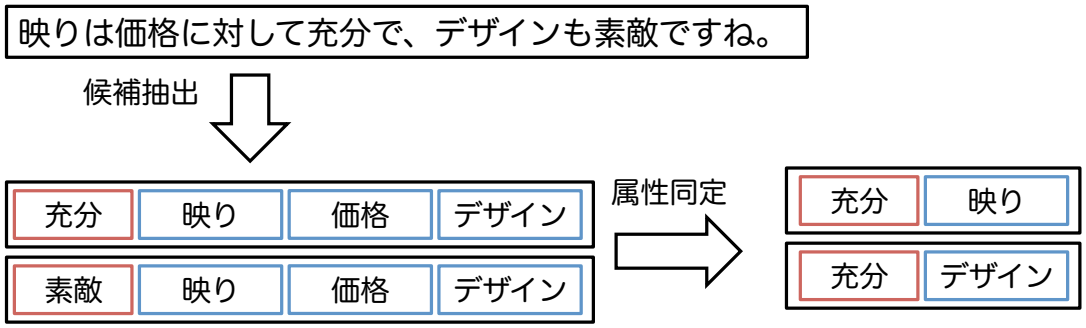

図 2: 照応解析と見なす手法の概要

図 2 を例にして、評価表現、属性候補抽出を説明する。まず、評価表現の辞書を用いることで評価表現「充分」「素敵」を抽出する。次に、評価表現「充分」「素敵」の一定の範囲に対して属性の辞書を用いて属性候補「映り」「価格」「デザイン」を抽出する。この処理によって、(評価表現:「充分」属性候補:「映り」「価格」「デザイン」) と (評価表現:「素敵」属性候補:「映り」「価格」「デザイン」) の評価表現と属性候補集合のペアが完成し、次のステップで評価表現に対応する属性を候補から決定し、評価表現と属性のペアを抽出する。

評価表現と属性候補集合のペアが完成すると、属性候補集合から対応する属性を決定する。ここでは二値分類で行う手法 [2] とトーナメントモデル [3] で行う手法を紹介する。最後に機械学習に用いられている素性について説明する。

二値分類で属性同定を行う手法では、図 3 のように評価表現と属性候補集合が与えられると、各属性候補に対して、評価表現の属性であるか属性でないかの二値分類を行う。図 3 では、評価表現「充分」と対応する属性候補「映り」「価格」「デザイン」が与えられており、評価表現と各属性候補に対して二値分類を行い、属性であると判定された属性候補と評価表現のペアを抽出する。その結果、「充分」と「映り」、「充分」と「価格」の 2 つのペアに対して属性であると判定され、抽出する。この手法では一つの評価表現に対して、複数の候補が属性と判定され抽出することが出来る。属性判定には各候補ごとに行うため、属性になりやすい

と思われる候補だけを抽出する。

図 3: 二値分類

飯田らの手法 [3] では属性候補から属性の決定にトーナメントモデルを用いている。トーナメントモデルとは最尤の属性を複数の中から決定するために、候補間で比較を行い勝ち抜き方式で属性を決定する。図 4 にトーナメントモデルの例を示す。まず「充分」の評価表現に対して (「映り」,「価格」,「デザイン」) の属性候補集合が与えられている。この候補集合から属性を決定する為に、勝ち抜き方式のトーナメントを行い、勝ち残った候補を属性として抽出する。最初に属性候補である「価格」と「デザイン」間で比較を行い、その結果「デザイン」が勝ち上がり、次に勝ち上がった「デザイン」と「映り」間で比較を行い、「映り」が勝ち上がり、最終的にトーナメントを勝ち抜いた「映り」と「充分」が評価表現と属性のペアとして抽出される。このようにして、候補間での比較を繰り返し勝ち抜き方式で評価表現に対応する属性を決定する。この手法では、候補間の比較を行うことで属性を決定することで、より属性らしい候補が選択される。このモデルでは一つの評価表現に対して、ただ一つの属性を出すことで誤った抽出を少なくしているが、対応する属性が複数となる場合も少なからず存在し、その場合に全ての属性を抽出できない。

飯田らの手法 [3] では、素性に属性候補の情報として、表層文字列、品詞を用いており、属性と評価表現間の関係として、属性候補と評価表現が直接係り受けの関係にあるか、属性候補と評価表現の文節間距離を用いている。この手法では単純な素性しか用いていないのに対して、Li らの手法 [4] では述語構造解析で使われる素性を追加し、性能を上げている。Li らの手法 [4] では、Soon のモデルと同様に

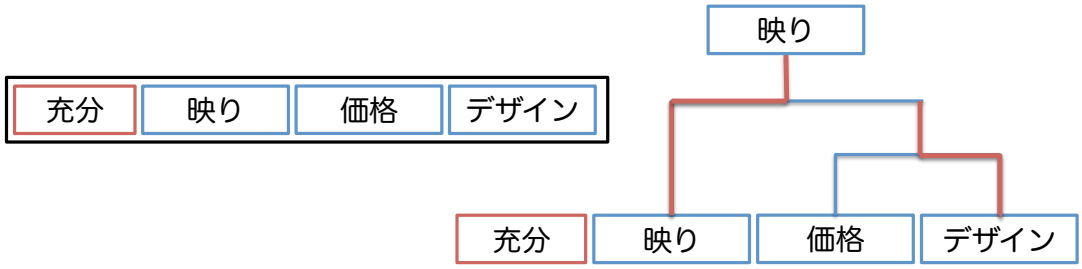

図 4: トーナメントモデル

評価表現の各属性候補に対して二値分類を行い評価表現と属性のペアを抽出している。述語構造解析で使われている構文木のパスに関する素性を追加した。構文木のパスとは、テキストを構文木で表したときの、属性候補から評価表現への道筋のことである。図 5 に構文木のパスの例を示す、ここでは属性候補「design」と評価表現「good」のパスが実線で表されており、このパスを考えると、「design」から順に「NP」「NP」「S」「VP」「ADJP」と辿ることで「good」にたどり着けるので、パスは「NP → NP → S ← VP ← ADJP」と表せる。テキストを構文木で表すことで構文構造を表せて、その構文木のパスを使うことで属性候補と評価表現間の構文関係を捉えられるので、より正確に候補から属性を選択できる。

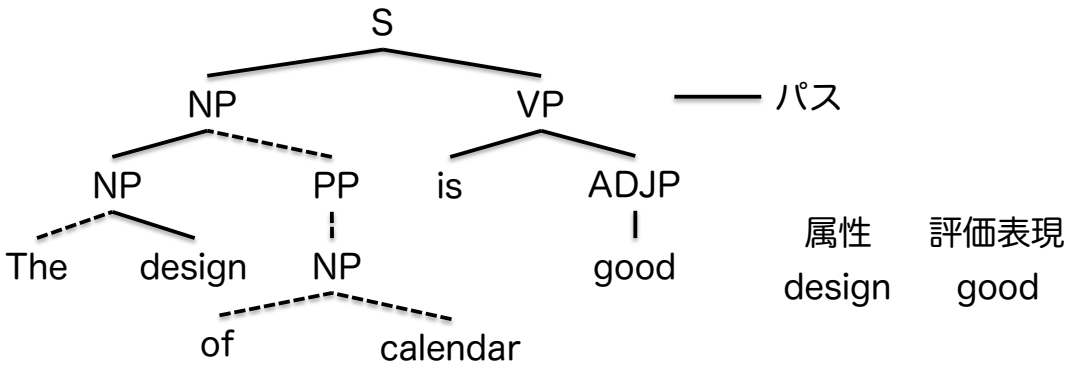

図 5: 構文木のパス

3　手法

属性抽出の問題点を調べるために、今回は簡単な手法である属性同定を照応解析と見なして行う手法を用いる。この手法では、評価表現、属性抽出と属性同定の二つのステップで属性抽出を行う。飯田らの手法 [3] では評価表現と属性候補の抽出に辞書を用いていたが、今回は評価表現は辞書を用いて抽出し、属性候補は辞書を用いずに抽出を行う。属性同定では、先行研究で用いられてる Soon の二値分類のモデルと飯田らのトーナメントモデルを用いる。また、Li らは構文木のパスを素性にし精度を上昇させていたため、構文木のパスを基に係り受けパスに関する素性を追加する。係り受けパスとは、テキストを係り受け木で表したときの属性と評価表現間のパスのことである。係り受け木によりテキストの構文構造が表現され、係り受けパスを使うことで属性と評価表現間の構文関係を捉えるような素性となる。

3.1　属性候補抽出

評価表現が属性についてのある事柄を述べていると見なし、属性が名詞句という仮定をおく。属性抽出の範囲は評価表現が出現した文内を対象とし評価表現の後ろを 3 文節までを属性候補抽出の範囲とする。これは、属性候補が評価表現の前に出現する場合は文節が遠く離れていても属性となりうる場合があるが、評価表現の後ろに出現する場合は遠い文節にはほとんど出現しないためである。今回の実験で使用したデータでは後ろに出現する属性 108 個中の 107 個は 3 文節までに出現したため、後ろ 3 文節までを抽出範囲とした。

前述の範囲で名詞句を属性候補と抽出するが、今回はいくつかの決まりを設けた。名詞の連接を名詞句と見なし、名詞の主辞、すなわち最右の名詞を候補として抽出を行う。「起動速度」という名詞句が抽出範囲に出現すると「起動」と「速度」の名詞の連接であるため、以下のように最右の名詞の「速度」を候補として抽出する。名詞句の主辞を候補として抽出すると、「携帯性」からは「性」、「機能的」からは「的」が候補として抽出されてしまう。これだけでは属性としての情報が少ないという問題が起こってしまうため、最右の名詞の品詞細分類が接尾

の場合は直前の名詞も一緒に候補と見なし抽出する。また、「大きさ」などといった形容詞と名詞の接尾が組み合わさる属性も存在するため、名詞の品詞細分類が接尾で直前の単語が形容詞の場合、直前の形容詞も一緒に候補として抽出する。

3.2 素性

使用する素性は飯田らの手法 [3] で使われている素性に加えていくつかの新たな素性を用いる。元の属性候補に関する素性では属性候補の原型や品詞などの単純な素性しか使っていないため、属性候補の品詞細分類などのより詳細な属性の情報の素性を追加する。Li らの手法 [4] では構文木パスを素性に加えることで精度が上がったため、構文木パスを基に係り受けパスに関する素性を追加する。係り受け木は図 6 のように表される。係り受けパスは属性と評価表現間の構文構造を捉えられることができ、精度が上がるものと思われる。

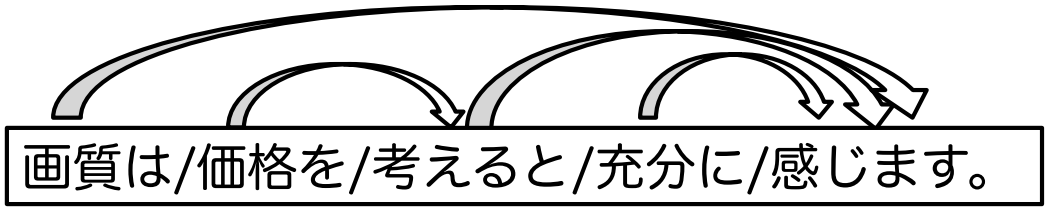

図 6: 係り受け構造

係り受けパスに関する素性の説明をする。図 6 の評価表現「充分」の属性候補である「画質」を例に考えていく。係り受けパスは属性を始点にして評価表現までの係り受け関係となり、係り受けの方向も考慮したものとする。係っている場合は「→」、係られている場合は「←」で表す。例では、始点である属性の候補の文節「画質は」が「感じます」に係っており、ここまでの係り受けパスは「→」と表される。次に「感じます」の係り先、係り元を調べると評価表現の文節「充分に」が係り元になっており、「感じます」が「充分に」に係られているので、ここでの係り受けパスは「←」で表される。これで属性の候補から評価表現までた

どり着けたので、係り受けパスは「→」と「←」を組み合わせて「→←」となる。係り受け距離は属性候補から評価表現までの係り受け関係の数、すなわち係り受けパスの矢印の数とする。例の係り受け距離は「2」となる。主辞付き係り受けパスは属性から評価表現までの係り受け関係で途中の文節の主辞の原型も係り受けパスに記録する。例の主辞付き係り受けパスは途中に「感じます」の文節を通るので、この文節の主辞である「感じる」を記録して、「→感じる←」となる。品詞付き係り受けパスは、主辞付き係り受けパスでは途中の文節の主辞を記録したが、この素性では主辞の品詞を記録する。例の品詞付き係り受けパスは、途中に「感じます」の文節を通るので、この文節の主辞である「感じ」の品詞である「動詞」を記録して、「→動詞←」となる。

使用する素性を表 1 に示す。図 6 の評価表現「充分」の属性候補である「画質」の場合の素性を例として載せる。

表 1: 使用した素性

素性名	内容	例
属性原型	属性候補の原型	画質
属性品詞	属性候補の品詞	名詞
属性品詞細分類	属性候補の品詞細分類	一般
文節間距離	属性候補と評価表現の文節間距離	3
前後関係	属性候補が評価表現より前、後ろどちらに出現するか	前
属性機能語	属性候補の文節の機能語	は
係り受け	属性候補と評価表現が直接係り受けにあるかどうか	False
係り受け距離	属性候補と評価表現の係り受け距離	2
係り受けパス	属性候補と評価表現の係り受けパス	→←
主辞付き係り受けパス	係り受けパスに中間ノードの主辞を記録したもの	→感じる←
品詞付き係り受けパス	係り受けパスに中間ノードの主辞の品詞を記録したもの	→動詞←

4　実験

レビューテキストからの評価表現と属性抽出に関して、次の 2 つの実験を行った。

- 係り受けパスの素性を追加する実験
- 二値分類とトーナメントモデルの比較

照応解析と見なす手法の二値分類とトーナメントモデルに係り受けパスに関する素性を追加して、精度が上がることを確かめ、エラー分析を行うことで問題点を理解する。また、二値分類とトーナメントモデルとの比較実験を行うことで、それぞれのモデルの特徴を調査する。

4.1　実験設定

使用するデータとして楽天商品レビューのデジタルカメラのレビュー 500 文書を用いる。500 文書のうち、レビューの全体の評価毎に均等になるようにしたかったのだが、評価が 1 と 2 のレビューが少ないために、評価 1 が 56 文書、評価 2 が 109 文書、評価 3 が 115 文書、評価 4 が 110 文書、評価 5 が 110 文書の配分となっている。この 500 レビューに人手で評価表現と対応する属性に対してアノテートを行ったところ属性と評価表現のペアが 934 個になった。実験設定として、評価表現は予めアノテートしたものをわかっているとし、評価表現に対応する属性を抽出する。今回は属性が文外に出現したり、属性が文内に出現しない評価表現については無視し、同一文内に対応する属性が出現する評価表現のみを用いて実験を行う。L2 ロジスティック回帰モデルで学習を行い、学習器として classias[12] を用いる。

4.2　係り受けパスの素性を追加する実験

二値分類とトーナメントモデルを用いて、それぞれに係り受けパスの素性を追加した場合と追加しない場合との比較を行う。ベースラインの素性として表 1 の素性から係り受け距離、係り受けパス、主辞付き係り受けパス、品詞付き係り受

けパスを除いた素性で機械学習させる。係り受けパスの素性を追加した場合は表1の全ての素性を用いて機械学習させる。

4.2.1 結果

実験の結果、表2のようになった。係り受けパスの素性を加えることで二値分類ではprecisionが下がったがrecallは上昇し、F値もわずかではあるが上昇した。トーナメントモデルではprecision、recallどちらとも上昇した。

表 2: 素性の比較

手法	Precision	Recall	F 値
二値分類	0.829	0.658	0.733
+係り受けパス素性	0.811	0.671	0.734
トーナメントモデル	0.810	0.765	0.787
+係り受けパス素性	0.832	0.786	0.808

4.2.2 考察

二値分類では係り受けパスの素性を追加することで抽出できなかった属性が抽出できるようになったが、誤った抽出が増えてしまった。これは、属性と評価表現によく成り立つ係り受け関係が属性でない候補に成り立つ場合に誤って抽出したためだと思われる。一方、トーナメントモデルでは係り受けのパスの素性を追加すると性能が上昇した。他の候補間と比較することで評価表現と同じ係り受け関係にあっても、係り受け以外の要素によって正しく選択できるためだと思われる。係り受けパスを入れることにより抽出できる属性が増え、属性抽出に対して有用な一面もあるが、属性ではない候補にも同じ係り受け関係が成り立つ場合があり、その候補だけをみている手法では誤った抽出が起きてしまう。しかし、トーナメントモデルのように他の候補と比べることにより属性になりやすい係り受け関係にあった場合でも誤りをなくすことが出来る。

トーナメントモデルでは一つの評価表現に対して複数の属性が抽出できないという問題点がある。抽出できなかった属性208個のうち、約27%の57個が複数

の属性のためであった。ここで複数の属性の文を見てみると、「対応、梱包には満足」、「機能や画質には文句無し」といったように他の候補と並立関係にある場合が多かった。そのため、選択した属性候補の周囲の候補に対して簡単なルールを適用することで並立関係にある候補も抽出することを考える。次の実験ではこの処理を加えたトーナメントモデルも比較対象に用いる。

4.2.3 エラー分析

この実験に対してエラー分析を行ったところ、主に次のようなタイプの誤った検出が見られた。() には全体のエラーでの割合を、「正」には抽出されてほしい正しいペア、「誤」には実際に抽出された誤ったペアを示している。次からこれらの誤りのタイプについて詳しく述べる。

1. 他の候補と同じ係り受け関係にある (20%)

   例 表現力はソニーに劣る。

   正 「表現力」-「劣る」

   誤 「ソニー」-「劣る」

2. 属性が評価表現の後ろに出現 (18%)

   例 高齢者の方でも使いやすい仕様です。

   正 「仕様」-「使いやすい」

   誤 「方」-「使いやすい」

3. 意味が合わない組み合わせが検出される (17%)

   例 [画質] きれい [機能性] 高い。

   正 「画質」-「きれい」

   誤 「機能性」-「きれい」

属性が他の候補と同じ係り受け関係にある場合が誤りの約 20%を占めていた。上記の例では属性「表現力」と他の候補の「ソニー」のどちらとも評価表現「劣る」に係っている。この場合、係り受けに関する素性は同じになってしまう。このように係り受け関係が同じになってしまうと誤った抽出が増えてしまう。二値分類で係り受けパスに関する素性を加えて precision が下がったのは、属性とならない候補が属性と評価表現間でよく成り立つ係り受け関係にあり、誤った抽出が増えたものだと思われる。この問題を解決するためには格構造を考慮する必要があると思われる。上記の例では、「表現力」が「劣る」のガ格になり、「ソニー」が「劣る」のニ格になっおり、ガ格の「表現力」が対応する属性となる。このように格構造を考慮することで同じ係り受け関係でも違いが生じて、より正しい抽出が行えるようになると考えられる。

全体の誤りのうち約 18%が評価表現の属性の後ろに出現していた。属性は評価表現の前に出現するものが大半なため、前後情報の素性が強く働き、後ろに出現する候補が抽出されにくくなってしまい、後ろに出現する属性の場合誤った抽出が多くなってしまう。今回の実験のデータでは全属性 934 個のうち約 88%の 826 個が評価表現の前に出現していた。この問題を解決するためには後ろに出現する属性の特徴をつかむ必要がある。例えば、評価表現が属性に係っているということが考えられ、直接係り受けに関する素性に前後情報の素性を組み合わせることによって、後ろに出現した場合の誤りを減らせるのではないかと考える。

全体の誤りのうち約 17%が評価表現と属性が意味の合わないであった。上記の例では属性「機能性」と評価表現「きれい」と全く意味の合わない組み合わせが抽出されてしまった。これは素性として評価表現に関するものを一切使っていないため、属性と評価表現の組み合わせを考慮されていないためである。この問題は評価表現の情報を素性に加えると解決できると考える。例えば、tf-idf を基にして、文書集合中での属性候補と評価表現が共起する頻度に属性候補が出現する頻度の逆数をかけた積を考える。意味の合う組み合わせの共起頻度は高くなり、積は高くなると考えられる。意味の合わない組み合わせは共起頻度が低くなるか、もしくは属性候補がよく使われる単語であり、その属性候補単体の出現頻度が高くなり、その逆数は低くなり、共起頻度との積は低くなる。このため、共起頻度

と属性候補単体の頻度の逆数の積は、意味の合う組み合わせは高くなり、意味の合わない組み合わせは低くなると考えられ、誤った抽出が減ると思われる。

4.3 二値分類とトーナメントモデルの比較

二値分類、トーナメントモデルの比較を行う。また、実験1でトーナメントモデルでは複数の属性に対応できるように、抽出した属性の周囲の候補に対して簡単なルールを適用することで複数の属性を抽出する手法も考え、性能が上がるかどうかを調べる。よって、二値分類、トーナメントモデル、最後にルールを適用する処理を加えたトーナメントモデルの3つで比較実験を行う。実験には表1の全ての素性を用いる。

今回の実験設定では、評価表現に対応する属性が必ず存在する。二値分類では、一つの評価表現に対して全ての属性候補に属性でないと判定してしまうと、その評価表現に対応する属性を一つも抽出できないことになり不利になってしまう。そのため、比較実験では二値分類で評価表現の属性候補が全て属性でないと判定された場合に、一番スコアの高い属性候補を属性として抽出を行う。

トーナメントモデルで並立関係の属性候補を抽出するために用いるルールをレビュー中でよく見られたパターンを参考に、次のようにした。選択された属性候補に対してこのルールを適用し、マッチした周囲の属性候補も一緒に抽出する。

- 隣接する文節が、候補+並立助詞、候補 +「、」、候補 + 格助詞「と」

  例: 機能性が抽出された属性　携帯性や機能性　→　携帯性

- 同一文節内で、抽出された候補 +「・」+ 候補

  例: 防塵性が抽出された属性　防塵性・防水性　→　防水性

4.3.1 結果

実験の結果表3のようになった。二値分類とトーナメントモデルを比較するとrecallでは二値分類が上回り、precisionではトーナメントモデルが上回った。F値

で見るとトーナメントモデルが高い値を示した。トーナメントモデルにルールを適用し、並立関係の候補を抽出できるようにした場合、precision は少し下がってしまったが、recall を上げることができ、F 値も上昇した。

表 3: 二値分類とトーナメントモデルの比較

手法	Precision	Recall	F 値
二値分類	0.758	0.807	0.781
トーナメントモデル	0.832	0.786	0.808
+ルール適用	0.818	0.819	0.819

4.3.2 考察

トーナメントモデルは二値分類に比べて、候補間の比較を行うことによってより属性らしい候補を選択することによって高い precision が得られ、複数の属性に対応していないため、recall は二値分類よりも低くなったと考えられる。最後にルールを適用して並立関係の候補を抽出できるようにすると並立関係にある属性も抽出できるようになったが、最初に選択された属性候補が誤りであるとその候補に並立関係にある属性も誤って抽出され、結果的に誤った抽出が増えてしまうという問題も発生した。

5　まとめ

本論文ではレビューテキストを対象に評価表現と属性のペアの抽出を二値分類とトーナメントモデルを用いて、係り受けパスの素性を追加することにより性能が上がることを確認し、エラー分析を行い問題点を述べた。また、二値分類とトーナメントモデルとの比較実験を行い、各手法についての特徴を述べ、トーナメントモデルに抽出した属性の周囲の候補にルールを適用することで並立関係にある候補も抽出できるようにし、precision が少し下がったが、recall と F 値を上げることに成功した。

今後の課題として、エラー分析で述べた問題点の解決が望まれ、以下のようなことが必要となってくる。京大格フレームなどの外部知識を用いて、候補の格構造を考慮した素性を追加する。属性が評価表現の後ろに出現した際の特徴を捉えるために、係り受け関係の素性と前後情報の素性などの組み合わせを試していく。意味の合わない属性と評価表現の組み合わせの抽出を防ぐために、属性と評価表現の共起情報と属性候補の頻度の逆数の積などを使って削減する。また、今回は属性が文外や出現しない評価表現を除いての実験を行ったが、このような評価表現も実際のレビューテキストには少なからず出現する。属性が出現しない評価表現については、対応する属性を推定する必要があり、難しい問題となっており、これらに対しての対策が必要となってくる。

謝辞

本研究を進めるにあたり、ご指導をいただいた乾健太郎教授、岡崎直観准教授に感謝致します。研究全般に渡り、直接のご指導と適切な助言を頂いた高瀬翔氏に感謝致します。日常の議論を通じて多くの知識や示唆を頂いた乾・岡崎研究室の皆様に感謝致します。

参考文献

[1] Lei Zhang and Bing Liu. Aspect and entity extraction for opinion mining. In *Data Mining and Knowledge Discovery for Big Data*, pp. 1–40. Springer, 2014.

[2] Wee Meng Soon, Hwee Tou Ng, and Daniel Chung Yong Lim. A machine learning approach to coreference resolution of noun phrases. *Computational linguistics*, Vol. 27, No. 4, pp. 521–544, 2001.

[3] 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. 情報処理学会自然言語処理研究会予稿集, NL-165-4, pp. 21–28, 2005.

[4] Shoushan Li, Rongyang Wang, and Guodong Zhou. Opinion target extraction using a shallow semantic parsing framework. In *Twenty-Sixth AAAI Conference on Artificial Intelligence*, 2012.

[5] Wei Jin, Hung Hay Ho, and Rohini K Srihari. Opinionminer: a novel machine learning system for web opinion mining and extraction. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 1195–1204. ACM, 2009.

[6] Niklas Jakob and Iryna Gurevych. Extracting opinion targets in a single-and cross-domain setting with conditional random fields. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pp. 1035–1045. Association for Computational Linguistics, 2010.

[7] Thomas Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine learning*, Vol. 42, No. 1-2, pp. 177–196, 2001.

[8] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, Vol. 3, pp. 993–1022, 2003.

[9] Minqing Hu and Bing Liu. Mining opinion features in customer reviews. In *AAAI*, Vol. 4, pp. 755–760, 2004.

[10] Ana-Maria Popescu, Bao Nguyen, and Oren Etzioni. Opine: Extracting product features and opinions from reviews. In *Proceedings of HLT/EMNLP on interactive demonstrations*, pp. 32–33. Association for Computational Linguistics, 2005.

[11] Guang Qiu, Bing Liu, Jiajun Bu, and Chun Chen. Opinion word expansion and target extraction through double propagation. *Computational linguistics*, Vol. 37, No. 1, pp. 9–27, 2011.

[12] Naoki Okazaki. Classias: a collection of machine-learning algorithms for classification.