4A1-2

表層的言語パタンを用いた階層的評価視点カタログの自動生成

A Method of Sentiment Aspect Tree Construction Using Linguistic Patterns

山下 和輝 *1　乾 孝司 *1　山本 幹雄 *1
Kazuki Yamashita　Takashi Inui　Mikio Yamamoto

*1筑波大学大学院システム情報工学研究科
Graduate school of System and Infomation Engineering, University of Tsukuba

In this paper, we propose a method of sentiment aspect tree construction. While sentiment aspects are very important units in the sentiment analysis, it is hard to understand the whole figure of them because there are large amounts of aspects without any structured formats. To resolve this problem, we propose to manage a set of sentiment aspects by a tree structure, where each node expresses each aspect and each edge expresses each hierarchical relation between nodes(aspects), and propose a tree construction method based on some linguistic patterns and the maximum spanning tree algorithm.

1. はじめに

現在、web 上ではブログや SNS、掲示板、レビューサイトなど、個人が自分の意見を書き込める場が増えてきている。これらのうち、商品やサービスに対しての意見・感想は、ユーザがその商品を買う目安になったり、企業側にとってもマーケティングや商品開発に使えるデータとなる。しかし、それらのデータは膨大であり、また構造化されていないことが多いため、それらから有益な情報を得るためには大きな作業負荷を要するという問題がある。

この問題に対する言語処理的な取り組みに評判分析がある [乾 06][Pang 08]。評判分析は評価文書から意見・感想を抽出し、整理、提示する研究分野である。評判分析を実施する上での基本概念として、評価視点および評価極性と呼ばれる 2 つの概念がある。評価視点とは、評価される対象のポイントとなる項目である。ホテルのレビューを例に挙げると、評価対象は「ホテル」であり、評価視点は「部屋」、「風呂」、「朝食」などが挙げられる。評価極性は、評価対象に対して人々が抱く評価の良し悪しのことであり、通常は肯定 (positive) と否定 (negative) の 2 極の値を想定する。例えば、「部屋が広くて綺麗だった。」という文であれば、「広くて」や「綺麗」などの手がかり表現からその文の評価極性は肯定と判定される。近年では、評価視点を文書中から抽出し、文書単位でなく、評価視点単位での細かい粒度での極性の判定を行っている。そのため、文書中から評価視点を抽出する方法の研究が行われてきた [Hu 04][Liu 05]。しかしながら、一般にある評価対象に対応する評価視点は数多く存在するため、先行研究のように単に評価視点を抽出するだけでは、出力の視認性が悪く、評価視点の全体構造を把握することが困難となってしまう。

そこで、本研究では抽出された評価視点を構造化し、カタログとして整理する方法を検討する。評価視点カタログの構造化の手法として、先行研究ではランキングやグルーピング、ラベリングなどの手法が取られているが、本研究では評価視点が階層性を持つことに注目し、木構造として評価視点を構造化する。提案手法では、2 つの段階を踏まえて評価視点木を作成する。まず第一段階では評価視点の組を抽出する。そして、第二段階はそれを元に木の生成を行う。評価実験を通して提案手法の有効性を検証した結果、最良モデルにおいて、82.6%の適切な評価視点パスを含むカタログを自動生成することができた。

連絡先: 山下和輝, 筑波大学大学院システム情報工学研究科, 茨城県つくば市天王台 1-1-1, yamashita@mibel.cs.tsukuba.ac.jp

2. 関連研究

評価視点の構造化の関連研究として Carenini らの研究がある [Carenini 05]。Carenini らは User-defined-features というユーザが定義した木構造を用意し、評価視点をこの木構造の各ノードに割り振るというクラス分類問題として扱った。この手法では、入力として木構造を与える必要がある。また、どのノードにも当てはまらない評価視点が出現したとしても、いずれかに割り振られてしまうという問題がある。そこで本研究では、評価視点の組から木を自動生成するという手法を提案する。これにより、入力として木構造を与える必要がない。また、木を自動生成するので分類できない評価視点が発生するという問題も解決できる。

3. 提案手法

3.1 提案手法の概要

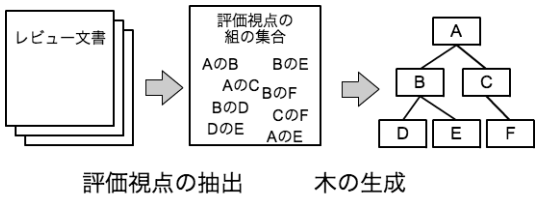

図 1: 提案手法の概要

提案手法では 2 つの段階を踏まえる (図 1)。まず、入力として評価文書が与えられる。第一段階では、入力として与えられた評価文書から、評価視点の組を抽出する。第二段階では第一段階で抽出された評価視点の組から木構造のカタログを生成する。この第二段階において、単純に評価視点の組をつなげると、複雑なグラフとなってしまい木構造が得られない。そこでグラフから木を生成する手法を幾つか検討し、それらの有効性を検証する。

3.2 評価視点の抽出

提案手法の主要モジュールはグラフから木を自動生成する第二段階である。第一段階ではグラフの構成要素となる評価視点を組の形でレビュー文書から抽出する。この組が後述するグラフや木の枝となる。評価視点組の抽出には言語パタン「A の B」を用いた。「A の B」パタンは「名詞の名詞」となるようなパタンである。これに注目した理由として、「A の B」というパタンが全体ー部分関係を表す典型的なパタンであること、また実際のデータ分析から階層性を持つ評価視点対の多くが「A の B」パタンで表現されていることからである。

抽出する際のルールとして、以下のようなものを設定した。

- A,B は名詞が 1 個以上連続したものであること。
- A,B は代名詞、非自立語を含まないこと。
- 「の」の品詞が助詞であること。

また、上記のルールに加え、一部の不適切なノードのフィルタリングを行った。フィルタリングで削除したノードは「人」「他」「存在」「一つ」の４つである。これらのノードは文書中に多く出現しているが、評価視点としては相応しくないため、事前に削除を行っている。これ以外にも評価視点としては相応しくないものは存在するが、木の生成に大きく関わってくるものでは無いため、対処を行っていない。

3.3 評価視点木の生成

前節で抽出した評価視点の組「A の B」から、評価視点の木を生成する。まず、「A の B」のうち、評価視点の A、B をグラフのノード、「A の B」を A から B へのエッジと見なすことで、「A の B」の事例集合から有向グラフを作成する。ここで、「A の B」の出現頻度を枝の重みとする。次に、このグラフとルートノードとなる評価視点を入力として、評価視点木を生成する。例えば、図 2 のようなグラフに対して、木の生成を行う。この図 2 では灰色で塗りつぶされたノードがルートノードである。

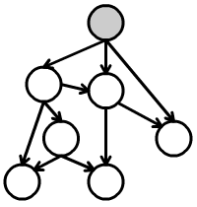

図 2: 「A の B」から作られたグラフの例

3.3.1 幅優先法

幅優先法では、ルートノードから木を幅優先探索で辿る際の順序に従って新規なノードとエッジを採用していき、採用ノード数が上限値となったところで停止する。この際、あるノードを親ノードと見立てた場合の子ノード集合については、エッジの重みに従って降順にソートしておく。幅優先法のアルゴリズムを Algorithm1 に、またその木の生成過程を図 3 に示す。図で灰色に塗りつぶされたノードは追加されたノードを表し、これは各手法の図で同じである。この手法は後述する提案手法との比較のために採用したベースライン手法である。

3.3.2 深さ優先法

深さ優先法は、上記の幅優先法のうち、考慮する探索アルゴリズムを幅優先から深さ優先に変更したものである。深さ優先法のアルゴリズムを Algorithm2 に、またその木の生成過程を図 4 に示す。この手法もベースライン手法としての採用である。

Algorithm 1 幅優先法

入力

入力としてグラフとルートノードが与えられる

Step.1

ルートノードから幅優先探索の順序に従ってノードとエッジを採用し、木を生成する。ノードが上限値に達した時点で停止する。

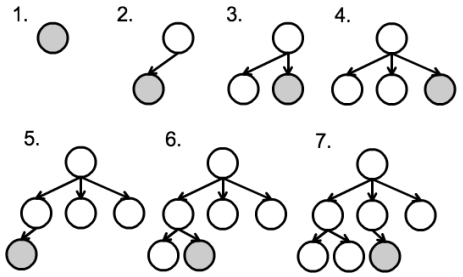

図 3: 幅優先法での木の生成過程

Algorithm 2 深さ優先法

入力

入力としてグラフとルートノードが与えられる

Step.1

ルートノードから深さ優先探索の順序に従ってノードとエッジを採用し、木を生成する。ノードが上限値に達した時点で停止する。

3.3.3 貪欲法

貪欲法では、ルートノードからエッジの重みが大きいノードから順に新規なノードとエッジを採用していき、採用ノード数が上限値となったところで停止する。なお、当然ながら採用することで木の制約を違反してしまうノードとエッジは採用されない。貪欲法のアルゴリズムを Algorithm3 に、またその木の生成過程を図 5 に示す。この手法は先の 2 つのベースライン手法よりもバランスの良い木が生成できると期待できる。一方で生成過程が貪欲的に進むため、大域的な木の良さは生成時には考慮されていない。

3.3.4 MST 法

MST 法は、基本的に貪欲法と同じであるが、前処理として、入力グラフから最大全域木 (Maximum Spanning Tree;MST) を生成するステップが追加されている。MST 生成の際はエッジの向きは無視し、プリム法 [Prim 57] を適用する。

3.3.5 各手法の比較

幅優先法と深さ優先法は単純な手法である。グラフからそれぞれ幅と深さを優先して木の生成を行っている。

貪欲法では、ルートノードからエッジの重みが大きい順にノードを選択する。このため、木全体としては、良くない枝を選択する可能性がある。また、間違った枝を選択した後、その直下のエッジの重みが大きいと連鎖的に誤りエッジを伸ばしてしまう。　また、枝の重み順にノードを選択していくため、相

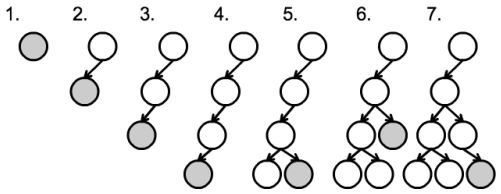

図 4: 深さ優先法での木の生成過程

Algorithm 3 貪欲法

入力

入力としてグラフとルートノードが与えられる

Step.1

ルートノードからエッジの重みに従って貪欲的にノードとエッジを採用し、木を生成する。ノードが上限値に達した時点で停止する。

対的な頻度の違いによって、一部だけ深くなりすぎるという問題があると考えられる。

それに対し MST 法では全域木を作成したのちに最終的な木を決定するため、貪欲法で起こるような間違いは少なくなると考えられる。

4. 評価実験

4.1 実験設定

4.1.1 データセット

楽天トラベル公開レビューデータ [*1] から 50 万件のレビュー文を無作為に抽出して実験に利用した。このデータに対して評価視点抽出を行った結果、評価視点の数が 127,058 個、組の数が 329,843 個となった。また、エッジの重みの平均値は 2.68 となった。抽出した評価視点の組から、4 つの手法で木を生成する実験を行った。比較を簡単にするため、いずれも上限ノード数を 300 に設定した。

4.1.2 評価手法

提案手法では、作成されるカタログが木の形をとっている。そのため、評価の際はルートから各ノードまでのパスが適正かどうかで判断を行う。その際、それ以上枝が伸びない終点のノードまでのパスを”terminal”、その下に子ノードを持つノードを”non-terminal”として扱う。”terminal”、”non-terminal”

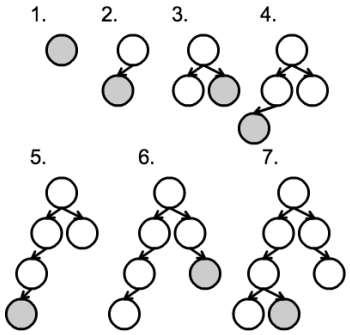

図 5: 貪欲法での木の生成過程

∗1 http://rit.rakuten.co.jp/rdr/

Algorithm 4 MST 法

入力

入力としてグラフとルートノードが与えられる

Step.1

グラフから最大全域木を作成する

Step.2

最大全域木から貪欲法により、エッジの向きを考慮した枝刈りをする。

のそれぞれのパスが正解かどうかを人手で判断し、正解・不正解の判定を行う。判定を行った結果から、各手法で生成した木それぞれのパス正解率を求める。パス正解率は以下のように定義される。

- パス正解率 = (木に含まれる正解のパスの数)/(木に含まれる全てのパスの数)

パス正解率は [0.0,1.0] の間の値になる。1.0 に近ければ近いほど良い結果と言える。

ある手法によって得られた評価視点パスが適切であるか否かの判断は以下で述べる言語テストに基いて判断した。言語テストとは、虫食いスロットを持つテンプレート文を用意し、虫食いの部分にテキストを差し込んだ時に適切な文になるかを判断することによって、虫食い部分に埋め込んだテキストの良否を判断する手法である。例えば「ホテル-部屋-風呂」などのパスに対し、各ノードを「の」でつなぎ、「(X) が良かった」などの文章にはめ込む。これにより、「ホテルの部屋の風呂が良かった。」という文ができあがるが、これは意味が通る正しい文だと言えるため、「ホテル-部屋-風呂」は適切であると判断できる。

言語テストのテンプレート文として、「(X) が良かった。」、「(X) が汚い。」、「(X) が便利である。」の 3 つを用意し、いずれかに当てはまれば適切なパスであると判断するようにした。

4.2 実験結果

各手法の結果は表 1 のようになった。最も良い結果となったのは MST 法である。それにつづいて、貪欲法が続き、その次に幅優先法となった。もっとも悪かったのは深さ優先法で、ほとんど良いパスが得られなかったことが分かる。

また、各手法で得られた木を木の深さという観点から見たところ、表 2 のようになった。これはルートを 0 とした時に、もっとも深い”terminal”ノードまでの深さを最大、もっとも浅い”terminal”ノードまでのパスを最小とし、”terminal”ノードまでのパスの長さの平均を取ったものである。この表から、幅優先法では浅い木しか得られていないことが分かる。ルートノードの直下に他のノードが全て付いている状態である。逆に深さ優先法では、かなり深い木ができていることが分かる。これに対して、MST 法と貪欲法では平均 2.1 と平均 2.3 と、比較的バランスがとれた木が作成できていることが分かる。

表 2: 木の深さ

手法	平均	最大	最小
幅優先法	1	1	1
深さ優先法	153.5	245	26
貪欲法	2.32	5	1
MST 法	2.13	4	1

表 1: パス正解率

手法	パス正解率 (all)	terminal	non-terminal
幅優先法	0.72(216/300)	0.72(215/299)	1.0(1/1)
深さ優先法	0.006(2/300)	0.0(0/43)	0.007(2/257)
貪欲法	0.773(232/300)	0.76(186/246)	0.85(46/54)
MST 法	0.826(248/300)	0.81(212/262)	0.95(36/38)

実際に生成された木構造カタログを調べたところ、貪欲法では以下のようなパスが間違いとなっていることがわかった。

- ホテル-部屋-バス-便
- ホテル-部屋-窓-外-車-駐車
- ホテル-部屋-冷蔵庫-飲み物-自販機-ビール

一つ目の間違いは、お風呂のバスと、車のバスを区別していないことによって起きた間違いである。これは抽出する際の問題と言える。二つ目や三つ目は、外-車や、飲み物-自販機という枝が発生したため、そこから下の枝が間違ったものになるというものである。MST 法ではこのような間違いはほとんど起こらない。車や飲み物というノードをつなぐエッジとしてもっと他にふさわしい部分があるため、貪欲法で起こる局所的な間違いが発生しないためである。

貪欲法と MST 法それぞれでよく見られた傾向として、一部のノードに多くのノードがつながってしまうという問題がある。幅優先法とくらべて、本研究の目的である階層性を得た評価視点カタログは得られているが、一部にノードが集まってしまう問題については検討する必要があるだろう。

5. おわりに

本研究では、評価視点の階層性に焦点を当て、木構造を持ったカタログの生成を行う手法を提案した。提案手法では、文書中から表層的な言語パタンを用いた、評価視点の組を抽出し、木構造のカタログを生成した。

評価実験の結果、MST 法はベースラインより高いパス正解率である 82.6%という結果を得た。この結果より、MST 法は階層的な木構造カタログを生成する手法として有効であると言える。

今回の結果では、得られた評価視点カタログの中に類似の評価視点が多く見られた。例えば、「風呂」「浴槽」「バス」や、「空調」「エアコン」、「従業員」「スタッフ」などである。意味が似通った評価視点をカタログの自動生成の前もしくは後にマージすることにより、より良い評価視点カタログを得られると考えられる。シソーラスを利用し、評価視点をマージする場合、木の生成の前だと数が膨大なため時間がかかってしまい、後だと必ずしも同じ階層にない場合などが考えられ、類似な評価視点のマージは難しいと考えられる。今後、類似な評価視点の取り扱いに取り組むことが重要な課題と言える。

また、今回の手法では結果を単純にするために木の形を取った。しかし、同じ評価視点でも複数の場所に出てくる場合がある。例えば「コーヒー」という評価視点に着目すると、「レストラン」のコーヒーや、「モーニングサービス」のコーヒー、など様々な場合が考えられる。このため、木という形を取らず、他のグラフの形を検討する必要があると言える。

さいごに、今回の研究の結果として、MST 法で得られたカタログの一部を図 6 に表す。

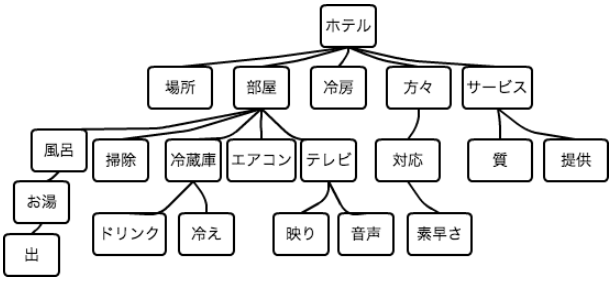

図 6: 評価視点カタログの一部

謝辞

本研究を実施するにあたり、楽天トラベル株式会社から施設レビューデータを提供して頂きました。ここに記して感謝の意を表します。

参考文献

[Carenini 05] Carenini, Giuseppe and Ng, Raymond T and Zwart, Ed: Extracting knowledge from evaluative text, pp.11–18(2005)

[Hu 04] Hu, Minqing and Liu, Bing: EMining opinion features in customer reviews, in AAAI, Vol.4, pp.755–760(2005)

[Liu 05] Liu, Bing and Hu, Minqing and Cheng, Junsheng: Opinion observer:analyzing and comparing opinions on the web, pp.342–351(2005)

[Pang 08] Pang, Bo and Lee, Lillian: Opinion mining and sentiment analysis,Foundations and trends in information retrieval, Vol.2, No.1-2, pp.1–135(2008)

[Prim 57] Prim, Robert Clay: Shortest connection networks and some generalizations,Bell system technical journal, Vol.36, No.6, pp.1389–1401(1957)

[乾 06] 乾 孝司, 奥村 学: テキストを対象とした評価情報の分析に関する研究動向, 自然言語処理, Vol.13, No.3, pp.201–241(2006)