階層非循環有向グラフを用いた文章の類似度に基づく評価文抽出

橋本 大吾　嶋田 和孝　遠藤 勉
九州工業大学 情報工学部 知能情報工学科
{d_hashimoto, shimada, endo}@pluto.ai.kyutech.ac.jp

1　はじめに

近年，Weblog や Web 掲示板等の普及により，個人が製品やサービスなどに対するレビューを投稿する機会が増えている．これに伴い，自然言語処理において，評価情報を対象とした処理（評価情報分析）の重要性が増している [1]．評価情報分析を行うためには，評価情報を含む文（評価文）と評価情報を含まない文（非評価文）が混在している文書の中から，評価文のみを収集する技術が必要となる．

本研究では，Web 掲示板等のレビュー記事から自動的に評価文を抽出することを目的とする．評価文抽出に関する研究はこれまでに数多く行われている．峠ら [2] は，人手により作成した評価表現辞書と，複数の規則及びパターンを用いて評価文抽出を行った．峠らの手法は，人手による評価表現辞書の作成が必要であり，多大なコストがかかる．鍛冶ら [3, 4] は，大量の HTML 文書から評価文コーパスと評価表現辞書を自動構築した．この辞書は，極めて大規模な評価表現辞書であり，さらにドメイン非依存であるため，非常に有用である．しかし，これをある特定のドメインに適用した場合，必ずしも十分な結果が得られない場合がある．そのため，ある特定のドメインに対して何らかのシステムを作る場合には，ドメインに依存した評価文抽出を行うことが重要になる．

そこで，本稿では評価表現辞書を使用せずに，ドメイン依存／非依存に対応可能な評価文抽出手法を提案する．提案手法では，評価表現辞書を使用しない代わりに，人手により少量の評価文を与え，文章の類似度に基づいて評価文を抽出する．文章の類似度の算出には，鈴木ら [5] が提案した階層非循環有向グラフと呼ばれるテキストの表現形式を用いる．

2　階層非循環有向グラフ

本研究では，文章の類似度に基づいて評価文の抽出を行うため，正確な類似度の算出が不可欠である．従来のテキスト処理では，単語の集合（Bag-of-Words）をテキストの特徴として処理する方法が一般的であったが，情報が不十分であり，必ずしも良い結果が得られるとは限らなかった．これに対して，単語の系列 [6, 7] や，単語の依存構造に基づく木 [8] を用いてテキストを表現する手法が提案されている．しかし，これらの単純な構造では，複雑なテキスト内の文法・意味的な情報を統一的に表現することができない．そこで鈴木ら [5] は，テキスト内の文法・意味的な情報を統一的に扱える「階層非循環有向グラフ」と呼ばれるテキストの表現形式を提案した．階層非循環有向グラフとは，グラフ内のノードがサブグラフにより表現されるような，階層的な構造を持つ有向グラフである．鈴木らは，階層非循環有向グラフを用いることで，従来のテキスト表現形式を用いる場合よりも，テキスト処理タスクの高精度化が可能になるとしている．

そこで本研究では，より正確な類似度を算出するために，鈴木らが提案した階層非循環有向グラフを用いて文章の類似度を算出する．

3　類似度の計算方法

本節では，階層非循環有向グラフを用いて文章の類似度を算出する方法について述べる．本稿では，階層非循環有向グラフに関する用語のうち，文章の類似度を算出する際に最低限必要となる用語についてのみ述べる．階層非循環有向グラフの詳細については，鈴木らの論文 [5] を参照していただきたい．

3.1　文章からグラフへの変換

まず，文章の形態素解析及び係り受け解析を行い，文章を階層非循環有向グラフに変換する．本研究では形態素解析ツール JUMAN[1] と係り受け解析ツール KNP[2] による解析結果を使用する．これらの解析結果によって得られた「文節」や「普通名詞」

[1] http://nlp.kuee.kyoto-u.ac.jp/nl-resource/juman.html
[2] http://nlp.kuee.kyoto-u.ac.jp/nl-resource/knp.html

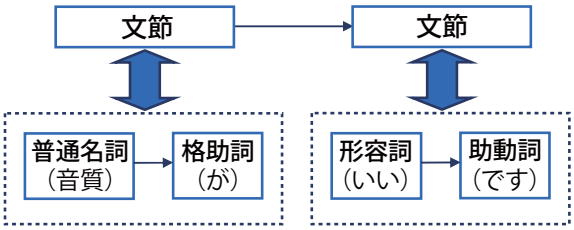

図 1: 文章を変換して得たグラフ

表 1: 階層部分属性列とその重みの例

階層部分属性列	重み
＜文節＞	$\sqrt{\beta}$
＜文節，文節＞	β
＜文節＜普通名詞＞＞	β
＜文節＜形容詞，助動詞＞＞	$\sqrt{\beta}^3$
＜文節＜普通名詞＞，文節＜助動詞＞＞	β^2

といった情報を属性とし，グラフの各ノードに付与する．図 1 に「音質がいいです.」という文章を変換したグラフを示す.

3.2　階層部分属性列の抽出

階層部分属性列とは，グラフのノードに付与された属性を，グラフの係り受け構造と階層構造に基づいて抽出したリストのことである．係り受け構造は「,」で，階層構造はリストの入れ子で表現する．

階層部分属性列はそのサイズに基づいて $\sqrt{\beta}^m$ の重みが付与される．ここで，β は 1 属性あたりの一致度 $(\beta > 0)$，m は階層部分属性列中に出現する属性の数を表す．表 1 に図 1 のグラフから抽出される階層部分属性列の一部とその重みの例を示す.

また，階層非循環有向グラフでは，ノードスキップを許して階層部分属性列の抽出を行う枠組みが導入されている．これにより，完全一致する構造だけでなく，類似した構造も評価することが可能となる．ノードスキップを許した階層部分属性列の抽出を行う場合，ノードのスキップ数に基づく減衰関数 $\Lambda(v) = \lambda^{n+1}$ を階層部分属性列の重みにかける．ここで，λ はノードスキップに対する基本減衰率 $(0 \leq \lambda \leq 1)$，n はサブグラフ中のノード数を表す．図 2 にノードスキップを許して抽出された階層部分属性列の例を示す.

以上の手順により，比較を行う文章のグラフからすべての階層部分属性列を抽出する.

3.3　2 つの文章の類似度の算出

文章の類似度算出には，比較を行う 2 つの文章から抽出される階層部分属性列のうち，両方の文章から共通に抽出されたもののみを用いる．まず，これらの階層部分属性列それぞれについて一致度を算出する．階層部分属性列の一致度はそれぞれが持つ重みを掛けることで算出することができる．図 2 において，2 つのグラフから抽出された「＜文節＜普通名詞＞, 文節＜形容詞＞＞」という階層部分属性列の一致度は $\lambda^2\beta^4$ となる．次に，算出した一致度の総和を求める．最後に，一致度の総和を 2 つの文章の文節数の積で除算したものを，2 つの文章の類似度とする.

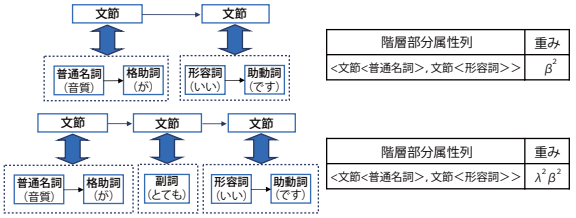

図 2: ノードスキップを許した場合の例

4　グラフの階層

階層非循環有向グラフに新たな階層を追加することは，テキストの意味情報及び文法情報を付加することに相当する．本節では，階層非循環有向グラフを用いてより正確な類似度計算を行うために新しく追加した階層について述べる.

品詞情報階層

品詞情報階層は，文節内の単語の品詞情報を用いた階層である．文節内の各単語の品詞情報をまとめて 1 つのノードに付与し，文節のノードと単語のサブグラフの間に配置する．これにより，階層部分属性列を抽出する際に文節に対して意味情報を付与することができる.

単語階層

単語階層は，単語を属性として付与したノードを含む階層である．これにより，表層の一致も考慮に入れて類似度計算を行うことが可能となる.

複合名詞階層

「ファイル転送ソフト」のような複合名詞については,「転送ソフト」,「ソフト」などのように表記揺れが発生する可能性がある．このような表記揺れに対応するために複合名詞階層を追加した．複合名詞階層は，複合名詞を構成する各単語をまとめて配置する階層であり，複合名詞の品詞を属性として持つノードの階層下に配置する.

図 3 に「転送速度が遅いです.」という文章を，全ての階層を使用したグラフに変換した例を示す.

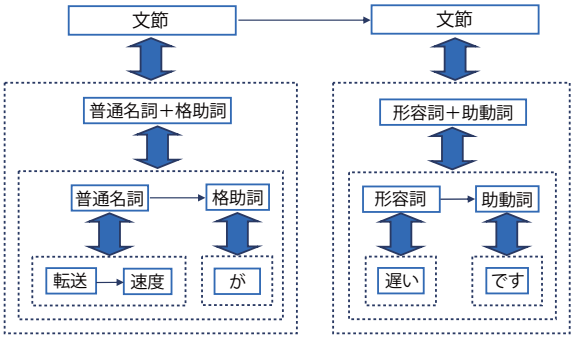

図 3: 全ての階層を使用したグラフの例

5 評価文の抽出

本節では，階層非循環有向グラフ，新たに追加した階層，及び文章の類似度の計算方法を用いて，レビューから評価文を抽出する方法について述べる．

5.1 抽出文集合の作成

以下の手順に沿って，抽出文集合を作成する．

1. 人手により少量の評価文を用意する．
2. 用意した評価文とレビュー内の文との類似度を総当たりで計算する．
3. 用意した評価文 1 文毎に類似度上位 N 文を抽出文集合の要素とする．

このようにして得られた抽出文集合が，文章の類似度に基づいて抽出された評価文ということになる．

5.2 β の調整と類似度の関係

類似度に基づいて評価文抽出を行う際，ドメイン依存の場合は文の表層を重視し，ドメイン非依存の場合は文の構造を重視する必要がある．鈴木らの論文 [5] では，過学習を抑制するために，β の値の範囲は $0 < \beta \leq 1$ と定めている．しかし，この定義のままでは，表層を重視する場合と構造を重視する場合とを明確に区別することができない．そこで本研究では，4 節で述べた 3 つの階層を，上になるほど構造に，下になるほど表層に近づくように設定する．ここで，β の値の範囲を $\beta > 1$ とすれば，3.2 節の階層部分属性列の抽出の定義に基づくと，下の階層，すなわち，表層に近い部分で一致するほど階層部分属性列の重みが大きくなる．一方で，鈴木らの定義に従い，$0 < \beta \leq 1$ とすると，品詞レベル，すなわち，構造的に一致するものについて階層部分属性列の重みが大きくなるということになる．言い換えれば，$\beta > 1$ を許すことで，構造を重視する場合（構造重視）と表層を重視する場合（表層重視）の二通りの類似度を算出することが出来る．これは，ドメイン依存の評価文が求められる場合と非依存の評価文が求められる場合の両方に提案手法が適用できることを意味する．

「音質はもちろんいいです.」という文を用いて，構造重視と表層重視のそれぞれで評価文を抽出した結果の例を表 2 に示す．

5.3 構造重視と表層重視の組み合わせ

構造重視と表層重視のそれぞれで抽出文集合が作成されるが，これらを組み合わせることで，新たな抽出文集合を得ることができる．実験では，以下の組み合わせを使用した．

組み合わせ (AND)
構造重視と表層重視の両方で抽出した文の集合を新たな抽出文集合とする．

組み合わせ (OR)
構造重視と表層重視のいずれかで抽出した文の集合を新たな抽出文集合とする．

6 実験

提案手法の有効性を検証するために，Web 上のレビュー記事を用いて実験を行った．

6.1 実験環境と評価基準

今回の実験では，人手で作成した評価文 20 文と，価格.com[3] に掲載されているポータブルオーディオプレーヤーのレビュー（評価文：610 文，非評価文：442 文）を使用した．また，β については，構造重視の場合 $\beta = 0.5$，表層重視の場合 $\beta = 1.5$ とし，λ 及び N は $\lambda = 0.5$，$N = 5$ と固定して実験を行った．

評価には，以下の尺度を用いた．

抽出評価文数
実際に抽出できた評価文の数．

抽出非評価文数
誤って抽出された非評価文の数．

抽出精度
抽出した文における評価文の割合で，以下の式で表される．

$$\text{抽出精度} = \frac{\text{抽出評価文数}}{\text{抽出評価文数} + \text{抽出非評価文数}}$$

[3] http://www.kakaku.com/

表 2: 構造重視と表層重視による抽出結果の例

順位	構造重視 ($\beta < 1$)	表層重視 ($\beta > 1$)
1	音質はそこそこ良いです.	音質はそこそこ良いです.
2	画面はとても見やすいです	音質はすばらしくいいです.
3	本体は少し重いです.	音質がすごくいいです.

表 3: 実験結果

	抽出評価文数	抽出非評価文数	抽出精度
構造重視	52	9	85.2
表層重視	63	7	90.0
組み合わせ (AND)	31	1	96.9
組み合わせ (OR)	84	15	84.8

6.2 実験結果

実験結果を表 3 に示す．構造重視，表層重視ともに高い精度で評価文を抽出できていることが分かる．また，全ての指標について表層重視の方が良いという結果を得た．

組み合わせについては，組み合わせ (OR) は全ての抽出方法の中で抽出精度が最も低いものの，抽出評価文数は最も多かった．一方，組み合わせ (AND) は全ての抽出方法の中で最も良い抽出精度を得た．抽出評価文数は大幅に減少したが，組み合わせ (AND) によって抽出された文は非常に高い確率で評価文であると言える．

7 考察

本節では，構造重視と表層重視の組み合わせについて考察を行う．組み合わせ (AND) は，非常に高い精度で評価文を抽出できることが分かった．このことから，組み合わせ (AND) による抽出結果を新たに評価文として与えて，もう一度抽出を行う，ブートストラップによる抽出が可能になるのではないかと考えられる．

次に，組み合わせ (OR) は比較的高い精度でより多くの評価文を抽出できることから，人手により与える評価文が少ない場合でも，多くの評価文を抽出できると考えられる．また，前述のブートストラップによる抽出は，回数を重ねるごとに抽出できる評価文の数が減少すると推測される．しかし，組み合わせ (OR) を用いることで，その後でも多くの評価文を抽出できるのではないかと考えられる．

これらのことから，構造重視と表層重視を組み合わせることで，目的および状況に応じた評価文抽出が可能になると考えられる．

8 おわりに

本稿では，評価表現辞書を作成せずに，文章の類似度を利用して Web 掲示板などのレビューの中から自動的に評価文を抽出する手法について述べた．実験により，提案手法が高い精度で評価文を抽出できることを実証し，その有効性を確認した．

今後は，グラフの階層構造の変更，類似度計算に使用するパラメータの調整を行うことにより，さらなる精度の向上を目指す．さらに，他の対象のレビュー記事を用いて実験を行い，提案手法の汎用性を検証する予定である．また，ブートストラップによる抽出の有効性の検証を行う必要があると考えている．

参考文献

[1] 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. 201–242, 2006.

[2] 峠泰成, 山本和英. 手がかり語自動取得による web 掲示板からの評価文抽出. 言語処理学会第 10 回年次大会, pp. 107–110.

[3] N. Kaji and M. Kitsuregawa. Automatic construction of polarity-tagged corpus from html documents. In *Proceedings of the 21st International Conference on Computational Linguistics, Poster Sessions (COLING/ACL2006)*, pp. 452–459.

[4] N. Kaji and M. Kitsuregawa. Building lexicon for sentiment analysis from massive html documents. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-CoNLL2007)*, pp. 1075–1083.

[5] 鈴木潤, 佐々木裕, 前田英作. 階層非循環有向グラフカーネル. 電子情報通信学会論文誌, D-II, Vol. J88-D-II, No. 2, pp. 230–240, 2005.

[6] H. Lodhi, C. Saunders, J. Shawe-Taylor, N. Cristianini, and C. Watkins. Text classification using string kernel. *J. Machine Learning Research*, Vol. 2, pp. 419–444, 2002.

[7] N. Cancedda, E. Gaussier, C. Goutte, and J.-M. Renders. Word-sequence kernels. *J. Machine Learning Research*, Vol. 3, pp. 1059–1082, 2003.

[8] M. Collins and N. Duffy. Parsing with a single neuron: Convolution kernels for natural language problems. Technical report, Technical Report, UCS-CRL-01-10UC, Santa Cruz, 2001.