拡散カーネルを用いた教師なし属性・評価組抽出

筑波大学大学院 システム情報工学研究科 コンピュータサイエンス専攻
本多 波輝　　乾 孝司　　山本 幹雄
honda@mibel.cs.tsukuba.ac.jp,　inui@cs.tsukuba.ac.jp,　myama@cs.tsukuba.ac.jp

1. はじめに

近年、Amazon[1] のレビューや Twitter[2] の投稿を参照することで、商品やサービスに対する個人の意見を比較的容易に収集できるようになった。収集したこれらの文書群から人々の意見を自動抽出し、商品やサービスに対する評価を商品開発にフィードバックすることは、ビジネスにおいて重要である。

本研究では、意見を<対象語、属性語、評価語>の三つ組で定義する。例えば、「ホテル の 夕飯 が おいしかった」の文における<ホテル、夕飯、おいしかった> を意見として扱う。ここで、「対象語」は商品やサービス、「属性語」は対象語の特徴や性質、「評価語」は評価者の主観的な評価を表す表現である。本稿では、対象語を指定して収集した文書から属性・評価組を自動抽出する手法について述べる。

属性・評価組抽出の従来手法としては教師あり学習を用いた手法がある [3]。しかし、教師あり学習に基づく手法では、対象語ごとに学習データを人手で作成するコストが生じる点が問題となる。そこで、本研究では指定した対象語に関連する文書から、属性・評価組を抽出する問題に対して、学習データを必要としない手法を提案する。提案手法では、属性・評価組候補の関連度を用いる。本論文では関連度を、ある属性語とある評価語が文書内において、意見を構成する要素としてどのくらい組になりやすいかを表す度合いと定義する。単純には、関連度の算出に文の係り受け関係や、品詞の情報が利用できる。しかし、意見分析で処理されるソーシャルメディアに含まれる文は文法的に適格でない文も多く、係り受けや品詞情報だけでは適切な関連度を見積もることが困難である。そこで本研究では、この問題に対応するために入力文書を大規模コーパスで補完したグラフを作成し、このグラフを用いて関連度を求める。特に本研究では、関連度の算出にカーネル法を適応することで、グラフ上の情報伝搬に基づく関連度算出法を提案する。楽天トラベルのレビュー文書セットを用いて平均適合率の平均値を求める実験を行った結果、提案手法は 84.34% となり、ベースラインの 72.44% を 11.9 ポイント上回る結果となった。特に、係り受け関係にない属性・評価組を抽出する際は、本手法が有効であることを確認できた。

2. 関連研究

教師なしの属性・評価組抽出の従来研究として、Liu らの手法がある。Liu ら [4] は、指定した対象語に関連する大規模レビュー集合から、Random Walk with Restart[5] を用いて属性・評価組の抽出を行った。Random Walk with Restart を用いることで、グラフにおける任意のノード間の関連度を求めることができる。Liu らは、大規模コーパスを用いて属性語ノードと評価語ノードからなる二部グラフを作成し、Random Walk with Restart で属性語ノードと評価語ノードの関連度を求めることで属性・評価組の抽出を行った。本手法では、大規模コーパスの情報に加えて、入力文書を補完された係り受けのグラフで表現することで文構造を考慮する点が Liu らと異なる。

3. 提案手法

3.1. 属性・評価組抽出手法の概要

提案手法では、入力文書を大規模レビュー集合で補完したグラフ (以後、拡張文書グラフと呼ぶ) を作成し、各ノード間の関連度を求める。関連度はノード間を結ぶエッジの重みとして表現され、入力文書に含まれる属性語と関連度が高い評価語が、属性・評価組として抽出される。

提案手法では、初めに入力文書を係り受け関係で表したグラフ (以後、初期文書グラフと呼ぶ) を作成する。係り受け関係を用いる理由としては、事前調査の結果より、属性・評価組は係り受け関係である場合が多かったことによる。次に、対象語に関連する大規模コーパスから作成したノードとエッジを初期文書グラフと併合することで、拡張文書グラフを作成する。拡張文書グラフを作成することで、係り受け関係にないノード間にエッジを張ることができる。しかし、依然として拡張文書グラフでは抽出すべきノード間にエッジが張られていない場合や、抽出すべきでないノード間に高い重み付けがされている場合は適切な属性・評価組抽出ができない。そこで本研究では、関連度の算出にカーネル法を適応することで、グラフ上の情報伝搬に基づいた関連度を推定する。カーネル法を用いることで、抽出すべきノード間にエッジが張られていない場合も関連度の算出が可能になる。

関連度を求めた後は、入力文書内の各属性語候補ごとに、文書内の評価語候補を関連度の降順にソートする。

その後、関連度を用いて Top-k により属性・評価語を抽

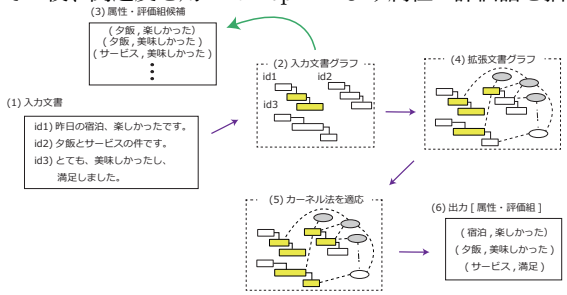

図 1　提案手法のフローチャート

出する。ここで、Top-k とは、属性語候補に対する評価語候補の関連度の上位 1 番目から k 番目までを抽出する操作を表している。提案手法のフローチャートを図 1 に示す。図 1 の黄色のノードは、各手順で属性・評価語抽出の対象となるノードを表している。図 1 の (1) は、入力文書を示す。図 1 の (2) では、入力文書の文の係り受け構造に従って、入力文書グラフを作成する。図 1 の (3) では、入力文書グラフから、属性・評価組を抽出する際に用いる属性・評価組候補を作成する。図 1 の (4) では、対象語に関連する大規模コーパスを用いて拡張入力文書グラフを作成する。図 1 の (5) では、拡散カーネルにより、到達可能な全てのノード間の関連度を算出する。図 1 の (6) では、(3) で作成した属性・評価組候補の関連度を用いて、入力文書に含まれる属性・評価組を抽出する。

3.2. 初期文書グラフの作成

初期文書グラフ作成の前処理として、入力文書を文節の主辞集合 C に変換する。C は、解析するレビュー文書を入力とした場合の、係り受け関係のフレーズの集合とする。また、C 間の係り受け関係の集合を R で表す。

3.2.1. 初期文書グラフの作成方法及び属性語候補と評価語候補の生成

作成するグラフは重み付き無向グラフで、$G_{\mathrm{IN}} = (V_{\mathrm{IN}}, E_{\mathrm{IN}})$ と表す。

- $V_{\mathrm{IN}} = C, v^i_{\mathrm{in}} \in V_{\mathrm{IN}}$ は初期文書グラフのノード集合を表す。
- E_{IN} は初期文書グラフの無向エッジの集合である。$(e_{\mathrm{in}} : v^i_{\mathrm{in}} \rightarrow v^j_{\mathrm{in}}) \in E_{\mathrm{IN}}$ となるようにエッジが張られる。ここで、$r_{i,j} \in R$ である。

同時に、属性語辞書と一致するノードを属性語候補とし、評価語辞書と一致するノードを評価語候補とする。ここで、属性語辞書および評価語辞書とは、対象語に関する大規模コーパスから抽出した属性語と評価語が登録されている辞書である。

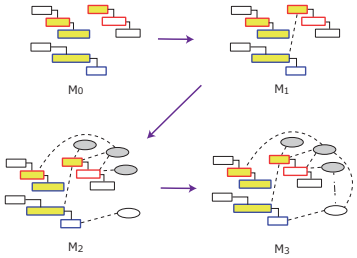

図 2　拡張文書グラフの各手法

3.3. 拡張文書グラフの作成

初期文書グラフの拡張手法は M_0 から M_3 まであり、M の添え字が大きくなるに従い、拡張文書グラフのノード数及びエッジ数が増加する。M_1 は M_0 の属性・評価語辞書とマッチするノード間に重み付きエッジを張る手法である。M_2 は M_1 の手法の後に、属性・評価組辞書と片方がマッチする場合にノードと重み付きエッジを追加する手法である。ここで、属性・評価組辞書とは、対象語に関する大規模コーパスから抽出した属性・評価組が登録されている辞書である。拡張文書グラフの各手法により、グラフが拡張されていく過程を図 2 に示す。図 2 の黄色のノードは、各手順において属性・評価語抽出の対象となるノードを表している。赤枠の実線のノードは入力文書グラフの属性語候補のノードを、青枠の実線のノードは入力文書グラフの評価語候補のノードを表している。白色の楕円のノードは属性語辞書から作成されるノードを、灰色の楕円のノードは評価語辞書から作成されるノードを表している。点線のエッジは拡張文書グラフの各手法で貼られるエッジを表している。

また、拡張文書グラフのエッジの重みは、PMI 辞書に問い合わせることにより割り当てる。ここで、PMI 辞書とは、対象語に関する大規模コーパスから作成したノード間の重みを参照できる辞書である。

3.3.1. 拡張文書グラフの作成方法 (M_0)

M_0 で作成するグラフは重み付き無向グラフで、$G_{\mathrm{M}_0} = (V_{\mathrm{M}_0}, E_{\mathrm{M}_0}, W_{\mathrm{M}_0})$ と表す。

- $V_{\mathrm{M}_0} = V_{\mathrm{IN}}, v^i_{\mathrm{m0}} \in V_{\mathrm{M}_0}$ は拡張入力文書グラフのノード集合を表す。
- E_{M_0} は拡張入力文書グラフの無向エッジの集合である。$E_{\mathrm{M}_0} = E_{\mathrm{IN}}$ となるようにエッジが張られる。

3.3.2. 拡張文書グラフの作成方法 (M_1)

M_1 で作成するグラフは重み付き無向グラフで、$G_{\mathrm{M}_1} = (V_{\mathrm{M}_1}, E_{\mathrm{M}_1}, W_{\mathrm{M}_1})$ と表す。

- $V_{\mathrm{M}_1} = V_{\mathrm{M}_0}, v^i_{\mathrm{m1}} \in V_{\mathrm{M1}}$ は拡張入力文書グラフのノード集合を表す。

- E_{M_1} は拡張入力文書グラフの無向エッジの集合である。E_{M_1} は E_{M_0} に属性・評価語辞書に登録されているノード間にエッジを追加することで作成する。

3.3.3. 拡張文書グラフの作成方法 (M_2)

M_2 で作成するグラフは重み付き無向グラフで、$G_{\mathrm{M}_2} = (V_{\mathrm{M}_2}, E_{\mathrm{M}_2}, W_{\mathrm{M}_2})$ と表す。

- V_{M_2},$v^i_{\mathrm{m}_2} \in V_{\mathrm{M}2}$ は拡張入力文書グラフのノード集合を表す。V_{M_2} は属性・評価組辞書に登録されているキーに、(v_{m_2}, v_*) または (v_*, v_{m_2}) が存在していれば v_* を V_{M_1} に追加する、という操作を行うことにより作成する。

- E_{M_2} は拡張入力文書グラフの無向エッジの集合である。E_{M_2} は属性・評価組辞書に登録されているキーに、$(v^i_{\mathrm{m}_2}, v_*)$ が存在していれば $(e_{\mathrm{m}_2} : v^i_{\mathrm{m}_2} \rightarrow v_*) \in E_{\mathrm{M}_2}$ を、$(v_*, v^i_{\mathrm{m}_2})$ が存在していれば $(e_{\mathrm{m}_2} : v_* \rightarrow v^i_{\mathrm{m}_2}) \in E_{\mathrm{M}_2}$ を E_{M_1} に追加する、という操作を行うことにより作成する。

3.3.4. 拡張文書グラフの作成方法 (M_3)

M_3 で作成するグラフは重み付き無向グラフで、$G_{\mathrm{M}_3} = (V_{\mathrm{M}_3}, E_{\mathrm{M}_3}, W_{\mathrm{M}_3})$ と表す。

- $V_{\mathrm{M}_3} = V_{\mathrm{M}_2}$,$v^i_{\mathrm{m}_3} \in V_{\mathrm{M}3}$ は拡張入力文書グラフのノード集合を表す。

- E_{M_3} は拡張入力文書グラフの無向エッジの集合である。E_{M_3} は E_{M_2} に属性・評価語辞書に登録されているノード間にエッジを追加することで作成する。

3.4.exponential カーネルの適応及び属性・評価組抽出方法

拡散カーネルの種類は多々あるが、少ないホップ数で遷移できるノード間に比重を重く持たせるという理由で、本研究では exponential カーネル [6] を選択する。exponential カーネルの式は式 1 の通りである。

$$K(\lambda) = M\exp(\lambda M) = M\lim_{n\to\infty}\left(1+\tfrac{\lambda M}{n}\right)^n \quad (1)$$

ここで、λ は指数拡散カーネルのパラメータである。式 1 の計算の際に、推移行列 M を半正定値が保証されているラプラシアン行列 L に変換する。本研究で用いるラプラシアン行列は重み付きラプラシアン行列である。推移行列 M をラプラシアン行列 L に変換する際は、式 2 を用いる。

$$l_{i,j} = \begin{cases} -w^{i,j}_{\mathrm{any}} & (i \neq j \text{ のとき}) \\ d_i & (i = j \text{ のとき}) \end{cases} \quad (2)$$

ここで、$l_{i,j}$ はラプラシアン行列 L の要素を、$m^{i,j}_{\mathrm{any}}$ は任意のグラフ G_{any} のノード i とノード j の重みを表している。また、d_i は $w^{i,j}_{\mathrm{any}}$ の i 要素の和を表している。式 1 を L を用いて表した式を、式 3 に示す。

$$K(\lambda) = L\exp(-\lambda L) = V^{-1}exp(\lambda A)V \quad (3)$$

本研究では式 3 を計算するにあたり、L を対角化して、$L = V^{-1}AV$ とした。このようにして求めた $K(\lambda)$ の要素 $k_{i,j}(\lambda)$ は、v_i と v_j の関連度を表している。なお、$K(\lambda)$ を求めた後は、L=$K(\lambda)$ とし後続の手続きへ進む。次に、ラプラシアン行列を用いて、属性語候補ごとに、属性語候補に対する全ての評価語候補の関連度を求める。最後に、関連度の降順に Top-k により属性・評価組を抽出する。

4. 評価実験

提案手法とベースラインを比較した結果を示す。ベースラインは 2 種類あり、入力文書の全ての属性・評価組候補の PMI スコアを用いて属性・評価組を抽出する実験 (以後ベースラインと記す) と、ベースラインに、係り受け関係のみ抽出する制限を加えて Top-k により属性・評価組を抽出する実験 (以後、ベースライン + 係り受け関係と記す) である。比較する提案手法は、M_0、M_1、M_2、M_3 に対してカーネルを適応した際の結果とする。また、カーネルを適応せずに、グラフの重みを関連度とした場合と、入力を文とした場合の Top-k による抽出結果も比較対象とする。

各実験で評価する際に用いる平均適合率の平均値を式 4 に示す。

$$MAP = \frac{\sum_{i\in Attr} MAP_i \cdot x_i}{\sum_{i\in Attr} x_i} \quad (4)$$

$$AP_i = \frac{1}{R_i}\sum_{j=1}^{N}\frac{y_j}{j}\left(1+\sum_{k=1}^{j-1} y_k\right) \quad (5)$$

MAP は平均値適合率の平均値を、AP_i は i 番目の属性語セットの平均適合率を表している。属性語セットとは、抽出した属性・評価語のうち、属性語ごとに作成する属性語に対する評価語からなる集合である。属性語セットは、$Attr$ と表す。x_i は i 番目の属性語セットに抽出すべき属性・評価組があるときは 1 を、そうでないときは 0 を取る 2 値変数である。N_i は i 番目の属性語セットの要素数であり、j の添え字が小さいほど高い関連度になるようにソートされているものとする。y_j は j 番目の属性・評価組の文字列が、正解の文字列と部分一致しているときは 1 を、そうでないときは 0 を取る 2 値変数である。

属性語と評価語の出現傾向調査に使用したデータセットは、楽天株式会社が提供している TSUKUBA コーパス [7] を使用する。TSUKUBA コーパスの 1,000 件のレビュー (総文数：4,309) に対して、抽出すべき属性・評価組の正解データを作成した。

はじめに、実験結果より、係り受け関係にない属性・評価組を抽出する際は、提案手法が有効であるということを示す。入力を文とした場合について、正解データのうち係り受け関係にある組のみを正解とした場合と、係り受け関係にない組のみを正解とした場合の平均適合率を求める。この結果を表 1 に示す。表 1(a) の結果より、提案手法がベースラインと同等の平均適合率の平均値であることがわかる。また、M_3-文_NonKernel と M_3-文_Kernel を比べると、平均適合率の平均値の差が少ないことから、係り受け関係にある場合には拡散カーネルの効果が少なくなると考える。これは、係り受け関係にある組の関連度の初期値を高く設定しているため、拡散カーネルによる関連度の増減が少ないからである。表 1 の結果より、本手法は係り受け関係にない属性・評価組を抽出することを目的としているが、係り受け関係にある属性・評価組の抽出もベースラインと同等の精度で行えていることがわかる。次に表 1(b) の結果より、提案手法がベースラインよりも、高い平均適合率の平均値となっていることがわかる。カーネルなしの結果がベースラインの平均適合率の平均値を上回った理由としては、拡張文書グラフにより係り受け関係にない属性・評価組を抽出できたためであると考える。また、カーネルありの結果がカーネルなしの平均適合率の平均値を上回っている理由としては、拡散カーネルにより適切な関連度を求めることができたためであると考える。表 1 の結果より、係り受け関係にない属性・評価組を抽出する際は、本手法が有効であることを確認できた。

最後に、平均適合率の平均値を表 2 に示す。表 2 より、実験結果の平均適合率の平均値は、M_3_Kernel が一番高いことがわかる。M_3_Kernel の平均適合率の平均値は 84.34% となり、ベースラインの 72.44% を 11.9 ポイント上回る結果となった。M_3_Kernel の平均適合率の平均値一番高くなった理由としては、提案手法により係り受け関係にない属性・評価組を抽出することができたためであると考える。このことより、拡張文書グラフを最大まで拡張した M_3_Kernel の提案手法が最も良いと考える。

5. まとめ

本稿では指定した対象語に関連する文書から、属性・評価組を教師なしで抽出する手法を提案した。楽天トラベルのレビュー文書セットを用いて平均適合率の平均値を求める実験を行った結果、提案手法は 84.34% となり、ベースラインの 72.44% を上回る結果となった。特に、係り受け関係にない属性・評価組を抽出する際は、本手法が有効であることを確認できた。

表 1　入力を文とした場合の平均適合率の平均値

手法	MAP(%) [(a):係り受けあり]	MAP(%) [(b):係り受けなし]
ベースライン+係り受け関係	87.12	0.00
M_0-文_NonKernel	87.12	0.00
M_1-文_NonKernel	84.79	48.73
M_2-文_NonKernel	90.22	48.52
M_3-文_NonKernel	89.14	48.20
M_0-文_Kernel(λ=0.01)	89.74	53.29
M_1-文_Kernel(λ=0.01)	80.36	53.12
M_2-文_Kernel(λ=0.01)	90.05	52.90
M_3-文_Kernel(λ=0.01)	89.85	54.62

表 2　各手法の平均適合率の平均値

手法	MAP(%)	手法	MAP(%)
ベースライン	72.44	ベースライン+係り受け関係	38.74
M_0-文_NonKernel	38.74	M_0_NonKernel	41.54
M_1-文_NonKernel	79.90	M_1_NonKernel	77.17
M_2-文_NonKernel	79.83	M_2_NonKernel	78.28
M_3-文_NonKernel	72.34	M_3_NonKernel	77.83
M_0-文_Kernel(λ=0.01)	79.76	M_0_Kernel(λ=0.01)	83.47
M_1-文_Kernel(λ=0.01)	78.00	M_1_Kernel(λ=0.01)	59.97
M_2-文_Kernel(λ=0.01)	80.45	M_2_Kernel(λ=0.01)	63.77
M_3-文_Kernel(λ=0.01)	80.66	M_3_Kernel(λ=0.01)	84.34

検討事項としては、提案手法に exponential カーネル以外のカーネルを適応することが可能であるため、他のカーネルとの比較実験を行うことを今後の課題としたい。また、提案手法では PMI スコアを用いることにより組のなりやすさを求めたが、評価語の評価極性を用いて組のなりにくさを求めることも可能である。平均適合率の平均値の改善のために、評価語の評価極性を用いた属性・評価組候補に対する選択制限を検討することも今後の課題としたい。

参考文献

[1] Amazon, http://www.amazon.co.jp/

[2] Twitter, http://twitter.com/

[3] Nozomi Kobayashi, Kentaro Inui, and Yuji Matsumoto,2007. Extracting aspect-evaluation and aspect-of relations in opinion mining. In Proc of EMNLP, pp.1065-1074.

[4] Kang. L, Liheng. Xu, and Jun. Zhao,2014. “Extracting Opinion Targets and Opinion Words from Online Reviews with Graph Co-ranking,” In Proceedings of ACL 2014, Baltimore, pp.22-27.

[5] Pan. J. Y, Yang. H. J, Faloutsos. C, and Duygulu, P. 2004. “Automatic Multimedia Cross- modal Correlation Discovery,” In Proc. of ACM Intl. Conference on Know ledge Discovery and Data Mining, pp.653-658.

[6] Shawe-Taylor. J, and Cristianini. N, 2004. “Kernel Methods for Pattern Analysis,” Cambridge University Press.

[7] 楽天データ公開, http://rit.rakuten.co.jp/opendataj.html/