E-010

クラスタリングを利用した評価文のアスペクト推定
Aspect identification of sentiment sentences using a clustering algorithm

波多野 匡[†] 嶋田 和孝[†] 遠藤 勉[†]
Masashi Hadano Kazutaka Shimada Tsutomu Endo

1. はじめに

現在，インターネット上には膨大なテキストデータが存在している．特に，ある製品やサービスへの意見や評価といった情報，いわゆるレビュー情報は増加が著しく，レビューサイトには大量のレビューが蓄積されている．これらのレビューには，企業側，消費者側双方にとって有用な情報が含まれている．そのため，レビューを分析し，有用な情報を抽出して利用する評判分析の研究が活発となっている [1].

評判分析の研究の 1 つとして，レビュー要約の研究がある．Blair-Goldensohn ら [2] は，アスペクト (評価の視点) に基づくレビュー要約を提案している．アスペクトに基づくレビュー要約によって，視覚的にもわかりやすい要約が実現されるとされる．

アスペクトに基づくレビュー要約を実現するために，評価文のアスペクトを推定する必要がある．Blair-Goldensohn らは，教師あり学習によって文やフレーズのアスペクト推定をおこなっている．教師あり学習は高い精度が期待できるが，大量の教師データの整備が必要であり一般に高コストである． 我々は，高い精度と低コストの実現を目指し，クラスタリングを利用した評価文のアスペクト推定手法を提案する．

2. 提案手法

提案手法は，評価文を入力とし，その評価文のアスペクトを出力する．図 1 に提案手法の概要を示す．まず,「類似した文は同じアスペクトを持つ」と仮定する．入力される評価文をクラスタリングし，生成される類似文クラスタの中心にある評価文にアスペクトを付与することで，良質な教師データを生成する．さらに，その教師データを用いて，クラスタ内の評価文から新たな教師データを獲得する．提案手法は，以降で述べる 4 つのフェーズを順に実行するものである．

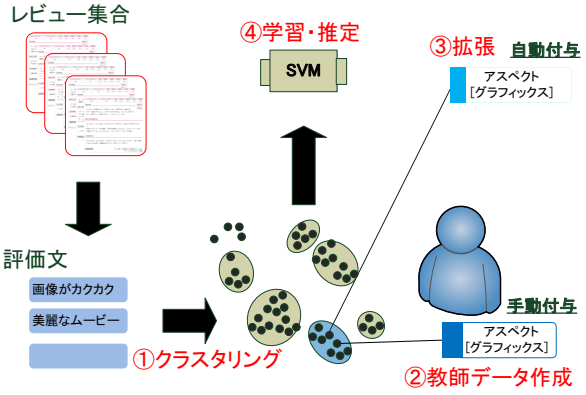

図 1: 提案手法概要

2.1. クラスタリングフェーズ

クラスタリングフェーズでは，評価文をクラスタリングし，類似文クラスタを生成する．クラスタ内の評価文は類似しているため，仮定より同じアスペクトであると考える．

評価文のクラスタリングに用いる素性は，評価文を構成する内容語の頻度である．内容語の出現傾向が類似している評価文をクラスタとして抽出する．

2.2. 教師データ作成フェーズ

教師データ作成フェーズでは，クラスタリングフェーズにおいて作成されたクラスタごとに，教師データの作成をおこなう．クラスタ内の評価文は，それぞれがクラスタに対する所属度を持つ．所属度は，クラスタの中心ベクトルと評価文の素性ベクトルのコサイン類似度である．クラスタへの所属度が高いほど，よりそのクラスタらしい評価文であるといえる．

教師データ作成フェーズでは，各クラスタの所属度最大の評価文を人間に提示し，正解アスペクトを与える．ここで，提示された評価文と割り当てたアスペクトを初期教師データとする．

2.3. 拡張フェーズ

教師データ作成フェーズで，人手でアスペクトを与えた初期教師データは非常に信頼できるものであるが，少量である．拡張フェーズでは，教師データ作成フェーズで作成された初期教師データを基に，クラスタ内の評価文とそのクラスタへの所属度によって拡張する．

所属度の高い評価文ほど，クラスタ内の初期教師データと同一のアスペクトである可能性が高い．そのため所属度の高い評価文に対して，クラスタ内の初期教師データと同一のアスペクト割り当てることで，教師データの拡張をおこなうことができる．しかし，所属度が高いということはクラスタ中心に近いということでもあり，クラスタ内の初期教師データに表層的に類似していることが考えられる．この場合，拡張をおこなうことによるアスペクト推定の精度への効果は小さい．

一方，所属度が低い評価文は所属度が高い評価文に比べて，クラスタ内の初期教師データと同一のアスペクトである可能性は低くなる．しかし，クラスタ内の初期教師データとは類似度が小さいため，アスペクトさえ正しければ，拡張をおこなうことで教師データのバリエーションが広がる．

本稿では，クラスタを中心部，中間部，外縁部に分割し，中心部を拡張に用いる場合 (MTD_{close})，中間部を拡張に用いる場合 (MTD_{mid})，外縁部を拡張に用いる場合 (MTD_{outer}) を提案する．なお，拡張を行わない場合を MTD_{only} とする．

2.4. 学習・推定フェーズ

拡張フェーズにおいて拡張された教師データを用いて学習をおこない，分類器を構成する．分類器としては SVM を用いる．

[†]九州工業大学 Kyushu Institute of Technology

3. 評価実験

提案手法の有効性を検証するために評価文のアスペクト推定実験をおこなった．

3.1. 実験内容

本研究では現在，ゲームレビューを研究対象としている．Tadano ら [3] の研究によって，ゲームレビューの評価文にアスペクトタグが付与されている．このアスペクトタグが付与されている記述 4607 件を実験データとし，その付与されているアスペクトタグを推定する実験をおこなう．

アスペクトとしては，「オリジナリティ (o)」，「グラフィックス (g)」，「音楽 (m)」，「熱中度 (a)」，「満足感 (s)」，「快適さ (c)」，「難易度 (d)」の 7 種があり，それらの複合も認められている．アスペクトの分布は一様でなく，「o」と「s」の頻度が突出している．

実験データに対するアスペクト推定を 10 分割交差検定で実施し，適合率を評価する．提案手法の初期教師データ数と同数の教師データをランダム抽出して教師あり学習をおこなう場合 (同数教師あり) をベースラインとして，提案手法の有効性を検討する．また，提案手法同士の比較によって，教師データの拡張に有効な所属度の範囲の調査もおこなう．

クラスタリングには軽量データクラスタリングツール bayon[*1] を用い，クラスタリング手法には Repeated Bisection 法を用いた．提案手法におけるクラスタリングフェーズのクラスタ数は指定せず，クラスタが一定のまとまりとなった場合にクラスタリングを終了するようにした．結果としてクラスタリングフェーズでは，10 分割交差検定の平均で 219 クラスタが生成された．すなわち，初期教師データは 200 件強である．さらに，拡張フェーズを適用した場合，例えば，MTD_{close} では，200〜300 件程度の教師データが新たに獲得された．

3.2. 結果と考察

各手法の適合率の比較を表 1 に示す．ランダム抽出した教師データで学習する場合より，クラスタリングに基づいて抽出した教師データを用いて学習をおこなう場合の適合率が高いことがわかり，提案手法の有効性が確認できた．同数教師ありでは，教師データをランダム抽出しているため，内容語が類似した評価文が教師データとなることがある．提案手法ではそのような評価文がクラスタとしてまとめられ，その中の 1 文が教師データとなる．そのため，類似した評価文が教師データとなることが少なくなり，教師データのバリエーションが広がり，適合率が向上したと考えられる．また，提案手法に関しては，概ねクラスタ中心部での拡張が有効であるといえる．

表 1: 評価実験結果

手法	適合率 [%]
同数教師あり	67.28
MTD_{only}	73.80
MTD_{close}	**73.97**
MTD_{mid}	71.30
MTD_{outer}	67.30

提案手法では，高い精度と低コストの実現を目指した．我々は，先行研究において，アスペクト推定について，いくつかの機械学習の精度を比較した [4]．同様に，4607 件のデータを単純に 10 分割交差検定により評価した場合の適合率を表 2 に示す．表 2 より，400 件強のデータで学習した MTD_{close} が約 4000 件のデータで学習した C4.5 と同程度の適合率を実現できていることが確認でき，高い精度と低いコストを実現しているといえる．一方で，SVM の適合率は 80%強であり，提案手法において教師データの拡張が理想的におこなわれた場合には，適合率は向上する可能性があるといえる．

表 2: 教師あり学習における適合率

機械学習器	適合率 [%]
SVM	**80.93**
C4.5	73.86

前述のように，提案手法は 10 倍程度の教師データで学習した場合と同等の精度が得られることがある．一方で，教師データが少ないことで，テストデータを分類する際にそもそも素性が 1 つも存在しない問題 (ゼロベクトル) が生じる．しかし，MTD_{only} と比較して，拡張フェーズの導入によって教師データが増加し，結果としてこのゼロベクトルの問題は改善されていることが確認された．ただし，精度の面では十分な貢献を見ることができず，教師データの有効な拡張方法については，今後さらなる考察が必要である．

4. おわりに

本研究では，評価文のアスペクト推定というタスクに対して，クラスタリングを用いた手法を提案した．評価実験により，提案手法の有効性を確認した．

Titov ら [5] は，文書集合の生成モデルである LDA を拡張した MG-LDA を用いて，教師データなしで単語とアスペクトの対応付けをおこなっている．今後は，この手法を参考に，教師なし学習の検討も必要であると考えている．

謝辞

この研究の一部は栢森情報科学振興財団の助成を受けて遂行された．

参考

[1] 乾，奥村，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] S. Blair-Goldensohn et al., “Building a sentiment summarizer for local service reviews”, WWW Workshop on NLPIX, 2008.

[3] R. Tadano, K. Shimada and T. Endo, “Effective construction and expansion of a sentiment corpus using an existing corpus and evaluative criteria estimation”, PACLING2009.

[4] 波多野，嶋田，遠藤，レビュー文のアスペクト分類における機械学習器の精度比較，第 17 回電子情報通信学会九州支部学生会, 2009.

[5] I. Titov and R. McDonald, “A Joint Model of Text and Aspect Ratings for Sentiment Summarization”, ACL, 2008.

[*1] http://code.google.com/p/bayon/