非自立語連鎖における含意を用いた文書分類の方法

小田　弘美
HP 研究所
hiromi.oda@hp.com

1　序

様々な商品や物事についての意見を含んだ文書群が大量に存在している場合に、先ずそれらに含まれる意見がおおまかに肯定的な意見を含むものか、あるいは、否定的な意見を含むものかという情報を抽出することは有益である。その後さらに詳細な処理を行う場合にも、前処理として大別することは非常に有用であると考えられる。従来、対象範囲を絞り効率の良い評価表現を多く収集することによって高い精度を目指す手法が主流であるが、本研究では含意という概念を導入し、内容語に含まれる含意に加え、どのような文書にも必ず現れる非自立的要素の持つ含意に着目することにより、インターネットから収集されるような生のデータや広い分野の文書に適用できる頑健な文書分類方法を構築する事を目指す。

2　問題の所在と背景

いろいろな商品やサービスについてのユーザの意見を含む文書を肯定的な内容のものと否定的内容のものに分類する、という課題を考える。この課題は、「この酒はおいしい」、「このワインの味は嫌いだ」といった分かりやすい表現を含む意見を想定すれば、一見単純なものに思われるが、現実には意見の表明は持って回った表現やニュアンスのようなものに込められている事が多い。例えば、「重たい味」、「まろやかさに欠ける」などと言った表現には単純な肯定・否定的表現はないが、日本語話者には肯定・否定どちらに偏った意見であるかは容易に判定できる。

どのような特徴を捉えて意見の分類を行うかという点が問題であり、多様な方法が提案されているが、最も基本的で広く行われている方法はこのような表現を多く収集して語彙リストを作り、当該の文書に含まれている割合や重み付け等の計算を行った結果によって判定するものである。リストを用いた方法では、当該の文書にリスト中の表現が生起しなければ判定ができないので、リストに収める項目やその重み等を自動的に収集し、拡張して行く方法が重要であり、各種の方法が提案されている [4][5]。また、リストに収める表現に句のレベルのものまで含めるなどの試みも行われているが、特定分野に特有な表現が収集されてしまう可能性もあり、精度を上げつつ広い分野へ応用するにはさらなる工夫が必要である。ここでは、インターネットから収集されるような比較的雑多な文書に含まれる意見の分類に応用できるよう、広い対象範囲への適用を想定した提案を行う。

3　アプローチ

適応分野についての制限の問題を克服する方法 の一案として含意 (connotation) という概念を用いて問題を整理することを提案する。ここでは 内容語の含意と非自立的要素の含意、と呼ぶ２種類の含意を認め、これを用いて分類を行う事を考えてみたい。

内容語の含意：

内容語とは、名詞、動詞、形容詞、副詞など独立して生起する事の出来る自立語、及び表現を意図している。内容語と呼ぶ理由は文法的には自立語であっても「する」、「ある」、「いる」、「なる」など意味内容が乏しいものは除外しているためである。内容語の中には、それ自体は明示的な否定、肯定の表現ではないが、どちらかへの意見的偏りを持つような表現が存在する。例えば、日本語の「いかがなものか、残念ながら」といった表現は否定的な意味合い（＝含意）を持っており、「ふんわり、なめらか、始まって以来の」 といった表現からは肯定的含意が感じ取れる。英語でも、“mixed opinions, unfortunately” といった表現からは否定的

な含意が感じられ、“meticulous, ingenious” 等には肯定的含意が認められる。

非自立的要素の含意：非自立的要素とは、言語の構成要素の中で、独自には出現せず、必ず他の要素について補助的な役割を果たすもののことで、品詞名としては、格助詞、終助詞、助動詞、接頭辞、接尾辞、活用語尾などが含まれる。さらに「する」、「ある」、「いる」、「なる」など意味内容の乏しく助動詞的機能を持つものも含む。

このような非自立的要素の連鎖に含意が生じる事がある。次の最小対を参照：

1a.「それは良い提案 **で** ある。」
1b.「それは良い提案 **では** ある。」

1a の「で」が 1b において「では」に変わっただけで、発言の持つ意味合いが否定的な方向に大きく変わる。提題の「は」単独で否定的含意があるとは考えにくいので、[では] という連鎖を形成することによってそのような含意が発生したと考えられる。この現象は日本語に限定されるものではなく、英語でも同様に次のような最小対が存在する。

2a. “This **is** a killer application.”
2b. “This **could have been** a killer application.”

2b. においては、話し手はその製品が *“killer application”* ではない、と考えているというニュアンスが伝わるが、a. と b. の違いは *“is”* と *“could have been”* の違いだけである。ここで ‘*“could”* は単に *“can”* の過去形であり、この助動詞一語に否定的含意を担わせる事には無理がある。他の 2 語についても同様であり、ここでも非自立的要素の連鎖の中に否定的含意があると考えられる。このように、個々の非自立要素では発生しない含意が、他の要素とつながることによって発生する。このような肯定・否定を示唆する含意を持つ連鎖を大量に検知することによって文書全体の肯定・否定への偏りを判定することを考える。具体的には、内容語の含意と非自立的要素の持つ含意の両方を反映したベクトルを構築し、サポートベクタマシンを用いて分類を行うという手法を提案する。

4 手法の提案

4.1 ベクトルの構築

4.1.1 内容語の含意の反映方法

個々の文書は一個のベクターで表現される。ベクターの最初の 2 個の要素は、肯定表現比率、否定表現比率であり、内容語の含意を反映する。この比率は、次のように計算される。

$$\text{肯定表現比率} = \frac{\text{当該文書における肯定表現の総数}}{\text{当該文書における内容語の数}}$$

$$\text{否定表現比率} = \frac{\text{当該文書における否定表現の総数}}{\text{当該文書における内容語の数}}$$

4.1.2 非自立的要素の含意の反映方法

ベクターの残りの部分は、非自立的要素の含意を用いた素性を表す。非自立的要素の n-gram 連鎖を素性とし、そこから素性選択と次元圧縮を行った結果によって非自立的要素の含意を反映する。まず、当該文書から非自立的要素のみを取り出し、 bi-gram, tri-gram、及び間隔を許す bi-gram 連鎖を抽出する。肯定・否定のサンプル文書（＝実験における訓練データ）を準備し、この連鎖の中からどちらか一方に有意に多く生じる連鎖のみを取り出し、その連鎖を素性とする。

素性選択の方法は、比率の差の検定を用いる [3]。今、ある連鎖 W が 2 つの文書集合 d_1, d_2 に共に表れたと考え、その頻度が w_1, w_2 であったとする。また、文書集合 d_1 に表れた連鎖の総数を n_1, 文書集合 d_2 のそれを n_2 とする。すると W がそれぞれの文書集合に表れた割合は、$p_1 = \frac{w_1}{n_1}$, $p_2 = \frac{w_2}{n_2}$ となる。今 $p_1 > p_2$ である場合に、これが有意であるかどうかを検定する。検定を行うために、まず実際には知られていない母比率 $\hat{\pi}$ を標本比率から推定する。

$$\hat{\pi} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$$

ここから Z 値を次の式で計算する。

$$z = \frac{p_1 - p_2}{\sqrt{\hat{\pi}(1-\hat{\pi})(\frac{1}{n_1} + \frac{1}{n_2})}}$$

5 %の危険率を想定し、$z > 1.65$ の値のものを選択する。このようにして肯定・否定のサンプル文書どちらかにより多く有意に出現する連鎖を素性として採用し、

個々の文書についてその出現回数を計算し、当該文書の非自立要素素性ベクターとする。

さらに、素性間の間接的共起関係を捉え、より頑健なシステムとする目的で特異値分解を用いる Latent Semantic Indexing (LSI) の手法によって、次元圧縮を行う。特異値分解は、$m \times n$ の行列 $\mathbf{A}$ について $\mathbf{A} = \mathbf{D} \cdot \mathbf{S} \cdot \mathbf{T'}$ のように３つの行列に分解する手法である。ここで $\mathbf{D}$ は、 $m \times n$ の行列、$\mathbf{S}$ は、左上から右下の対角要素に特異値が大きい順に表れる $n \times n$ の行列をなし、また、$\mathbf{T'}$ は、$n \times n$ の行列である。$\mathbf{D}$ と $\mathbf{T}$ はそれぞれの列が直交関係にある直交行列となる。ここで、$\mathbf{S}$ の特異値を大きい方から r 個取り ($r \leq n$)、$r \times r$ の行列 $\mathbf{S}_r$ とし、$\mathbf{D}$ から、$m \times r$ の部分行列を取り出し $\mathbf{D}_r$ とし、$\mathbf{T'}$ から、$r \times n$ の部分行列を取り出し、$\mathbf{T'}_r$ とすると $\hat{\mathbf{A}} = \mathbf{D}_r \cdot \mathbf{S}_r \cdot \mathbf{T'}_r$ とAのランク r における近似 $\hat{\mathbf{A}}$ が得られる。Deerwester et al.[2] 等では、元の $\mathbf{A}$ の行列が、m 個の文書、n 個の用語に対応する情報を持った行列である場合に、$\mathbf{D}_r$ は、r 次元における文書の新しい配置を示し、また $\mathbf{T}_r$ は r 次元における新しい用語の配置を示し、その重要な特徴を抽出した表示となっており、用語間の間接的共起関係も反映されていると主張する。結果として、この手法を用いれば、n 次元の素性を用いた文書の表現が、$\mathbf{D}_r$ の行ベクターにおいて、r 次元の表現に圧縮されたことになる。これを非自立的要素の含意を反映する素性とする。実験では r=15 を用いている。

4.2 機械学習による分類

結果として、それぞれの文書は、[肯定表現比率, 否定表現比率, 次元圧縮された非自立要素素性] というベクターで表現されている。訓練用文書、テスト用文書を全てこのようなベクターで表現した後、機械学習を用いて、訓練データを基に与えられた文書を２つのカテゴリーに分けるための判断基準を学習し、その基準を用いて新しいデータ（テスト文書）を分類する。ここでは、現在最もパターン識別能力の高い方式の一つとされているサポートベクタマシンを用いる [1]。サポートベクタマシンは、訓練文書に基づいて判断基準となる超平面から双方のクラスの最も近いデータポイント（＝サポートベクタ）までの距離（マージン）を最大にする、という方法によって高い汎化能力を持つモデルを学習する。学習後のモデルにより、分類器としてテスト文書の分類を行わせる。サポートベクタマシンは基本的に２クラスへの分類を行うものであり、本課題に好適な学習方式である。実験では工藤拓氏によって公開されている TinySVM[*1] を使わせて頂いた。

5 実験

提案手法の有効性を検証するために小規模な実験を行った。

5.1 実験データ

データ収集：データ領域として「日本酒」を例として選び、インターネットからデータを収集した。「出羽桜」「土佐鶴」などの銘柄のリストを準備し、このリストの用語をキーワードとしてインターネット検索を行った。検索結果から、該当のキーワードを含む html ページをダウンロードし、蓄積した。ここから html タグ等の除去を行った後、意見を含む文書を抽出するために以下のような文書タイプを識別しながら排除し、フィルタリングを行った：ガーベージ型文書（リンク切れ等）、リスト型文書（商品名の列挙等が主なページ）、日記型文書（多くの話題の扱われている文書のごく一部に目的のキーワードが現れるもの）。次に、主観的判断や直感的、感覚的表現を含む評価表現リスト等を利用し、素人の文書と専門家の文書を区別するという文書分類を行うことによって商品を売る側の文書と区別し、一般ユーザの評価を含む文書を抽出し、最終的に人手による検証を行った。

収集された意見を含む文書から 1000 個の文書を選び、内容の肯定・否定の観点から五段階に分類してみたところ、次のような結果となった。

$$[Bad\ 31, Worse\ 64, Neutral\ 16, Better\ 470, Good\ 419]$$

結果として、大きく肯定の方に偏っており、非常にバランスの悪い結果となった。これは、blog や電子掲示板などに投稿する際には、多くの人が読む事を想定するために丁寧な表現となるという事情があるように思われる

内容語リスト：訓練文書から、肯定的含意を含む内容語リストと否定的含意を含む内容語リストを人手によって抽出する。実験に用いた肯定的含意を含む内容語リストは、「おいしい、良い、大好き、好ましい、甘

[*1] http://chasen.org/ taku/software/TinySVM/

い、風味、最高、素晴らしい」等 66 個、否定的内容語リストは、「まずい、臭い、もの足りない、ざらざら、べたべた」といった表現 120 個が記載されている。

5.2 実験方法

比較のためのベースラインとして、肯定表現比率と否定表現比率の 2 つの素性のみを用いた分類装置を準備した。これと以上で説明した全ての素性を用いた分類装置を比較することによって非自立的要素の含意を加えた分類装置の効果を測定する。

データが非常に偏っているので、次のような方針で分類実験を行う事とした。条件 1 として、$bad+worse \Leftrightarrow good$ の訓練データとテストデータを準備し、SVM により学習、分類実験を行う。データを訓練データとテストデータに 2 分し、訓練データは、210 個の肯定意見と 48 個の否定意見、テストデータは 209 個対 47 個という構成とした。この訓練データで学習した SVM により、$bad + worse \Leftrightarrow better + good$ のデータを分類させる実験も試みる。

条件 2 として、変則的ではあるが、肯定・否定のバランスを揃えたデータを、$bad+worse \Leftrightarrow good$ のデータから作って実験を行う。条件 1 で good について、訓練データ、テストデータ共約 200 個ある肯定意見を約 50 個づつに 4 分割し、否定意見は再利用して、約 50 個づつの 4 対のデータセットをそれぞれ作成した。結果として 50 個づつのデータセットによる 4×4 の訓練、テストの組み合わせを実験する事が可能となる。

5.3 実験結果

条件 1：線形カーネルを用いた SVM による訓練学習の結果、全ての素性を用いた分類器では、訓練データについて正解率 (accuracy) が 86.8%, 判定精度 (precision) 86.6%, 再現率 (recall) 99%、テストデータについて正解率 83.2%, 精度 85.8%, 再現率 95.2% であった。肯定・否定表現比率のみを用いた分類器は収束しなかった。また、同一の訓練データによって学習した SVM によって、better と分類されたものを含むテストデータ、すなわち、$better+good \Leftrightarrow bad+worse$ 468 対 47 の組み合わせを分類させたところ、正解率 89.7%, 精度 91.8%、再現率 97.1% であった。

条件 2：肯定・否定でデータ数を揃えた分類実験の結果を表 1 に示す。これは、4×4 の組み合わせの実験の平均値である。バランスの取れたデータセット条件ではベースラインの分類器も平均 71.8% の正解率を出しており、全ての素性を用いた分類器 (BOUND-FORM) は 76.5% の正解率であった。50 個対のみの訓練データを用いた結果としてはベースラインの精度もかなり良いと思われるが、標準偏差 (STD) を見ると、どの指標においても非自立的要素の素性を用いた分類器よりも大きなばらつきがあることがわかる。以上の結果から、非自立的要素を用いた分類器は、バランスの崩れた訓練データについても頑健な学習を可能にし、また、50 個対という小さな訓練データであっても安定した分類性能を示すということが示唆される。

6 結論

いろいろな商品やサービスについてのユーザの意見を収集し、肯定的意見のものと否定的意見のものに分類するという課題を取り上げ、肯定・否定的含意を含む表現のリストに加え、非自立的要素の連鎖の持つ含意を用いて分類を行う方法を提案した。インターネットから収集したデータによって小規模な実験を行った結果、バランスの崩れたデータについて頑健な性能を持ち、50 個対の小さな訓練データについても安定した分類能力を持つ事が明らかとなった。

参考文献

[1] Critianini, N., & Shawe-Taylor, J. An Introduction to Support Vector Machines. Cambridge: Cambridge U. P.(2000).

[2] Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., & Harshman, R. Indexing by latent semantic analysis. Journal of the American Society for Information Science, 41(6),(1990).

[3] 池田央編、「統計ガイドブック」新曜社 (1989)

[4] 乾孝司、乾健太郎、松本裕治「出来事の望ましさ判定を目的とした語彙知識獲得」言語処理学会第 10 回年次大会発表論文集 pp. 91-94,(2004).

[5] Taboada, M. & Grieve, J. Analyzing Appraisal Automatically. AAAI Spring Symposium on Exploring Attitude and Affect in Text. Stanford, March 2004.

表 1　条件 2 による実験結果

BOUND-FORM	Accuracy	Precision	Recall
Average (%)	76.51	75.17	74.98
STD	2.01	3.168	2.776
BASELINE	Accuracy	Precision	Recall
Average (%)	71.79	73.24	65.93
STD	2.909	6.328	6.913