「お客様の声」に含まれる テキスト感性表現の抽出方法

舘野昌一
富士ゼロックス株式会社
〒2590157 神奈川県足柄上郡中井町境430
tateno.masakazu@fujixerox.co.jp

要約
テキストに含まれる感性表現を抽出する方法を提案する。具体的には、コーパスの中で感性表現を含む文をタグ付けし、これと同類の文を抽出する規則を自動生成する。そのために、文は、構文としてあいまい性がない範囲までを構文構造として自動生成し、その中に含まれる感性表現を、要素間の依存関係として人手によりタグ付けする。このようにして表現されたタグ組から、自動的に抽出規則を生成し、その規則に基づいて、コーパス内の感性表現を抽出する。各抽出規則が抽出するノイズや各抽出規則間の包含関係によって、規則の良し悪しを評価する。

1 背景

企業が負う社会的責任は日増しに高まってきている。何かしたことによる責任だけでなく、何もしないことによる責任も追求されることが当たり前になってきている。このことは国や地方自治体においても同様である。つまり組織がもつ社会的責任は重大でありかつ増大している。ここで、企業であれば、サービスや商品の提供を受ける人、国や地方自治体であれば、国民がお客様であるが、そのお客様からの電話や email による問い合わせに潜んでいる、肯定・否定または満足・不満足の表明には、組織の経営トップが見落とすことのできない重要な情報が含まれている。本稿ではこれらの問い合わせがテキスト化されたものを「お客様の声」と呼び、そこに含まれる肯定・否定または満足・不満足の表現をテキスト感性表現と呼ぶこととする。組織が提供する商品やサービス、あるいは組織そのものに向けられたお客様からの負のテキスト感性表現には、組織経営の視点から見て緊急性の高いものが多い。したがって、他の情報を差し置いてでも、そのための対処のフローを組織内に作り、即座に対応していくことが、双方の利益となる。本稿では、そのようなテキスト感性表現を抽出するための方法を提案する。

2 タスクの定義

「お客さまの声」のコーパスに含まれる負の感性表現を抽出することが今回のタスクである。そこで表現されている日本語は、通常の書き言葉では使用されない、いわゆるくだけた表現が多く、また誤字・脱字・変換ミスなどの表記の誤りも多く含まれる。

3 本方法の概要

これを行うために、図1に示したように大きく、(1)負のテキスト感性表現(感性表現と略す)を抽出するための規則を生成する過程(図の横方向)と、(2)生成された抽出規則により、コーパスから感性表現を抽出する過程(図の縦方向)の二つに処理を分けた。本稿では、このうち、訓練用コーパスへのタグ付け、感性表現のタグ付け、抽出規則の自動生成、改良のための評価方法、評価までを順番に述べる。

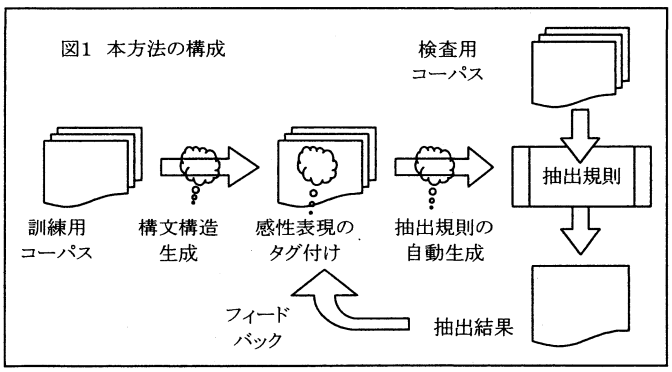

図1 本方法の構成

3.1 訓練用コーパスへのタグ付け

2名のタグ付与者に、前述の背景を丁寧に説明した。そして、抽出手順を示し、コーパスから感性表現を抽出する作業を実施させた。その結果、各文内の特定の形態素列またはそれらのn項の共起関係が指定された。これに基づき、2名が共通にタグ付けしている文を正解とするタグ付きコーパスを作成した。約 8,600 文からなるコーパスにタグ付け作業をした結果、約 830 文が得られた。

3.2 抽出規則の記述

コーパス中のある正解文と同等の文を抽出するには、正解文にタグ付けされている表現と同じ表現を含む文を抽出することが必要である。そのため、本方法では、

構文構造と素性情報を含めた抽出規則を記述することとした。構文解析と表現抽出は、Xerox Incremental Parser (XIP) [1]により行った。XIPは、あらかじめ記述された複数の規則の順序付き集合に基づいて、テキストを解析する。

3.3 感性表現のタグ付けと抽出規則の自動生成

正解コーパス中のタグ付けされた文を対象に、構文構造を生成しておく。そして人手によりその中の抽出したい箇所にタグ付けをする。そして、タグ付けされた構文構造から、抽出規則を自動生成する。

4 改良のための評価方法

このようにして記述された抽出規則は、一つの抽出規則が一つの文を抽出するが、さらに副作用として類似の文を抽出する。そこで、抽出規則を評価する必要があるが、それは、検査用コーパスを用いて抽出結果の文単位での再現率と適合率により行う。副作用の大きさは、抽出規則の記述が大掴みであれば大きいし、詳細であれば小さい。ここで、抽出すべき文を正文と呼び、抽出すべきでない文を負文と呼ぶこととすると、各抽出規則が正文をいくつ抽出し、負文をいくつ抽出しているかは、個々の抽出規則の性能を示す。さらに、抽出規則間には、次のような上下関係がある。つまり、共通する正文を複数の抽出規則が抽出する場合、より少ない正文を抽出する抽出規則に対応付けたノードを配置し、少なくともそのノードで抽出される文を抽出する抽出規則を、そのノードの下に集める。このようにしてクラスターを生成することにより、同一の正文集合を抽出する抽出規則は、一つのノードに集まる。以上、3つの指標である、抽出正文数、抽出負文数、クラスターに基づいて、抽出規則の詳細化、一般化、選別を行う。

5 評価

「お客様の声」は、各企業が保有しているが、本稿ではこのような実際の情報とかなり近い表現が収集されているウェブサイトである不満リサーチ.com (http://www.fuman-r.com/) から、インターネット(ウェブサイト、PC、プロバイダー)、製品(自動車、家電など)、娯楽(コンサート、ゲーム、カラオケなど)、仕事(企業、業務)、お金(税、預貯金、ローン、クレジットカードなど)、マスコミ(広告・キャンペーン、新聞雑誌、テレビ・ラジオ)、コミュニティ(政府、公共施設)などの7ジャンル 23 分野合計約8,600件の文を使用した。まず各分野の文を二つに分けてそれぞれ合わせて約4,300件ずつとし、訓練用コーパスと検査用コーパスとした。これを用いて、訓練と検査を行った。その結果、訓練用コーパスでの再現率が 100%、適合率が 59.1%であった。また、検査用コーパスでの再現率が 29.6%、適合率が 24.3%であった。検査用コーパスでの再現率は、抽出規則の個数が多くなると増加する傾向があったが、適合率にはそのような傾向は認められなかった。次に、適合率が極端に悪かった抽出規則を一つ取り除いて、さきほどと同様の実験を行った結果、検査用コーパスでの再現率が 26.1%、適合率が 35.7%となった。検査用コーパスでの再現率は若干減少したものの、適合率が大幅に増加した。(表1)

実験番号	訓練用コーパス		検査用コーパス	
	再現率	適合率	再現率	適合率
1	100.0	59.1	29.6	24.3
2	100.0	76.3	26.1	35.7

表1 再現率と適合率

6 関連研究と課題

テキスト中の感性表現に関しては、形容詞など内容語に注目した研究[2]があるが、本稿で示したようなコーパスに基づく方法はないようである。感性表現を含む文は、人によるばらつきがあり、その文のどこで感性を表現しているかとなると、さらに大きくばらつく点で、固有名などに対するタグ付けに比べ難しさがある。コーパスの量を大きくしていくことにより、再現率を向上させることが期待できる。このことが、抽出性能を向上させる原動力となる。コミュニティの中で共有できるコーパスは現状では存在しないが、評価結果を比較するためにも、公開のコーパスが必要である。

7 将来の活動

本稿で述べた方法により、抽出性能の高い感性表現抽出規則を獲得し、それらにどのような語が含まれているかを見ることにより、日本語の特性を把握していきたい。このことは、組織の経営トップが必要とする将来の言語アプリケーションのコア技術として重要である。また、本稿で述べた感性表現と、そのような表現をするにいたった原因や理由とを結びつけることにより、抽出される情報の付加価値が増す。また、そのような項目を追跡することも重要である。副詞や形容詞が不満や満足、肯定・否定の表現に使われており、そのような語にタグ付けをすることも有効な手段と思われる。このような方向への活動を進めていく予定である。

参考文献

[1] Salah Aït-Mokhtar, Jean-Pierre Chanod, and Claude Roux. "Robustness beyond shallowness: incremental deep parsing". In Natural Language Engineering, 8(2): 121–144, 2002.

[2] 自然言語処理のための形容詞の意味表現, 内海, 堀, 大須賀, 人工知能学会誌 Vol. 8 No. 2, pp192-200, 1993