自 然 言 語 処 理 153－14
（２００３．１．２１）

「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木構造としてあらかじめ自動生成しておき、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。このようにして作成された抽出規則は、再現率と適合率により評価されるが、各規則が抽出するノイズや、各規則間の包含関係によって、規則の良し悪しを評価する方法を示した。以上に基づき、実験と評価を行い、評価方法の有効性を示した。

The Method to extract Texual “Kansei” Expression in the Customer’s Voice

Masakazu Tateno
Fuji Xerox Co., Ltd.
〒2590157　430 Sakai, Nakai-machi, Ashigarakami-gun
tateno.masakazu@fujixerox.co.jp

Abstract

We propose the method to extract Textual “Kansei” (ability to feel something happens) expression. The method includes tagging to the sentences with the Kansei expression and generating the rules to extract similar sentences to the tagged ones. Each sentence in the corpus is parsed to generate a tree that is not ambiguous as the syntax for the sentence and Kansei expressions are tagged as the dependencies by hand. The extracting rules are generated from the tagged corpus automatically, then they extract Kansei expressions from another corpus. We also showed the method to improve the rules by counting noises produced by the rules and by clustering all the rules to evaluate the rules by recall and precision. The experiment, evaluation and improvement are also shown.

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 「お客様の声」のコーパスの特徴

各企業では、「お客様の声」1件1件がどのように処理されているかを示す履歴がデータベースに記録されている。そこで表現されている日本語は、通常の書き言葉では表現されない、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。

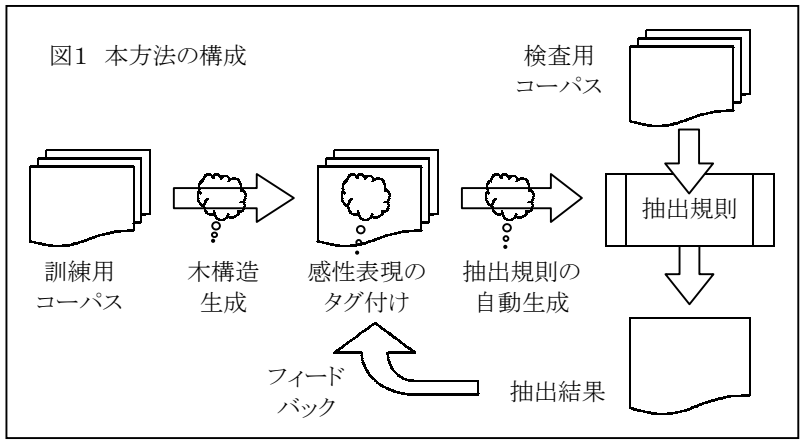
3 タスクの定義

「お客さまの声」のコーパスに含まれる緊急性のある問い合わせに必ず対応する、という目的があるので、再現率を重視する。つまり負の感性表現を漏れなく抽出することが今回のタスクのねらいである。

4 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。今回の報告は、このうち、コーパスへのタグ付け、抽出規則の生成、抽出実験、評価までを順番に述べる。

図1 本方法の構成



4．1 コーパスへのタグ付け

コーパス(訓練用コーパスと検査用コーパス)へのタグ付けを行うために次のような予備実験を行った。初めに2名のタグ付与者に次のような手順を示し、コーパスから感性表現を抽出する作業を実施させた。

手順

(1) 文中で、もっとも重要と思われる文節を特定する。多くの場合、文末の用言節である。

(2) (1)で特定された文節を修飾する節の中で、もっとも重要と思われるものを特定する。多くの場合、「は」「が」「を」「に」「で」などの助詞を伴う。

(3) (1)と(2)で得られた対(または組)に、意見・感想、背景、状況説明、質問、要求、苦情、などの種別を割り当てる。その際、未知のものに関しては人が判断し種別を設定し、それ以降はその種別を使用する。

(4) (1)と(2)で得られた対、あるいはそれぞれを種別で示した対が、感性表現かどうかを判定し、感性表現の場合、肯定・否定の度合い(感性値)を付与する。その際、未知のものに関しては人が判断し、それ以降はその値を使用する。

以上の手順により、文中の1箇所または複数箇所の形態素列を感性表現として特定した。このことにより、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る。この手順は、ある程度なれてくると、感性表現部分だけを抽出することが可能となる。したがって、作業者には、そのような方法でのタグ付けも許した。

この2名によるタグ付け作業は、お互いがどのようにタグ付けしているかを知らせないように行った。その結果、2名が共通にタグ付けをしている部分は極めて少なく、このようなタグ付けに基づくタグ付きコーパスを準備することは極めて難しいことがわかった。しかし、文を単位として見た場合には、かなりの共通性があることがわかった。そこで、タグ付けしている箇所ではなく、その箇所を含む文が共通なものを正解とするタグ付きコーパスを作成することとした。約 8,600 文からなるコーパスにタグ付け作業をした結果、2名が共通してタグ付けしたものが約 830 文、1名だけがタグ付けしたものが約 80 文、もう1名がタグ付けしたものが約 3,000 文あった。これは人により感性表現と判断する閾値が異なるためで、感性表現のタグ付けは、極めて感覚的であり、正当性の評価は難しいことを示している。

4.2 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。そのためには、文字列の部分一致の抽出規則を記述すればいいのだろうか。あからさまな表現であれば文字列レベルでの抽出が可能であるが、微妙な表現であればその周囲での言葉づかいを木目細かく見る必要がある。この抽出は、結局、再現率・適合率の性能問題に行き着くので、本方法では、あらかじめ日本語の構文構造と素性情報を含めた抽出規則を記述することにより、精度を上げていく方法とし、抽出箇所の指定の木目細かさを上げたり下げたりする。そこで、そのための日本語の構文構造を次項で述べるのように設定した。

4.3 日本語の構文構造

本方法では、解析対象文から、あいまい性のない範囲で構造を作り、次に依存関係として、関係する構成要素を引数とする関数を記述することとした。したがって、多数の構造が候補として得られることはない。また、構造がもつ排他性により、本来得られるべき構造が得られなくなるという現象は、複数の依存関係がもつ矛盾を許すことにより得ることとした。このような前提にたつ場合、どの程度までの構造を生成しておくべきか判断が必要となる。本方法では、被修飾(受け)として、用言節、体言節、体用言節の三通りとし、そのそれぞれの第一の素性を、修飾(係り)として、連用節、連体・終止節、「の」節の三通りを設定した。ここで、助動詞「だ」に継続する体言は、被修飾の立場からは体言節と用言節を兼ねると考え、

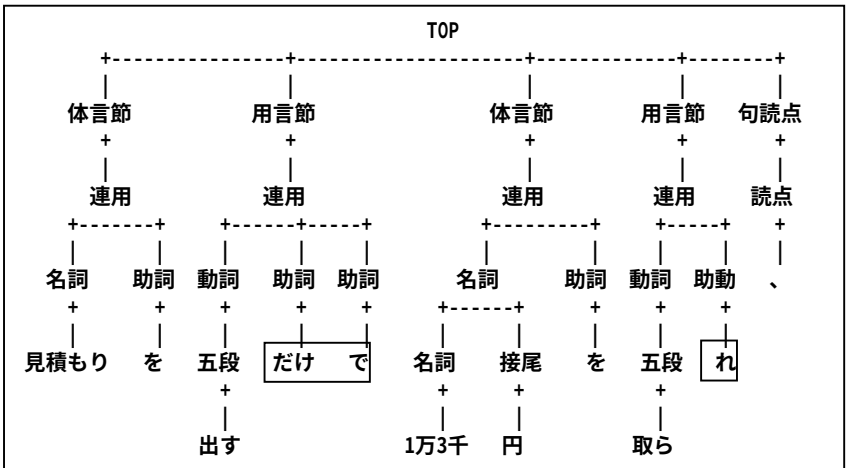


図2　本方法で使用する構文構造例

体用言節とした。(なお、体言とは、名詞および学校文法の形容動詞の語幹とする。また用言とは、動詞とする。その他の品詞(形容詞、連体詞、接続詞など)は、テキスト内での係り受け関係が明確ではないので、そのまま、解析木中の1本の枝とした。また、連体節と終止節を分けず連体・終止節としたのは、ほとんどの場合、形態が同一であるためである。なお、連体か終止かを特定する場合は、直後に体言がないかどうか、あるいは文末かどうかで判断できるので、その時点での処理に委ねる。また「の」節として単独にしたのは、「の」の係り先が直後の体言に限られる訳ではないため、意味の解釈なしには構造化できないことによる。これも、意味が解釈できた時点での処理に委ねる。

4．4　抽出処理

以上の解析は、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。ここで、入力は、形態素解析結果であり、具体的には、表層表現、レンマ、素性列の繰り返し(つまり{表層表現、レンマ、素性列}+)である。処理は、最初に、入力された素性が、必要に応じて変換規則により、XIP の素性表現に置き換えられる。次に、これらの表層表現、レンマ、素性が、レキシコン規則により試され、必要に応じて新たな素性が追加される。このようにして加除修正された形態素解析結果から、各形態素が一つのノードで表現され、その下に表層表現、レンマ、素性列が吊り下がる形式に変換される。このようなノードからなる列を対象に、ある条件を満たすかどうかを試し、満足する場合には、ノードを一つ生成し、その下にそれらの条件を満足するノード列を吊り下げる。図2に例を示す。このようにノードを一まとめにするための規則を塊化規則と呼ぶ。塊化規則は、複数の層に分けて記述し、同じ層の塊化規則は1回の解析処理で同時に適用されるが、層の異なる規則は順番が来るまで用いられない。したがって、ノード列は、複数回の解析を経過していろいろなところで枝を作り、それらを合わせて木となる。このようにして生成された木構造を対象に、枝同士の特定の関係を導き出す。そのために記述される規則を依存規則と呼ぶ。

4．5　木構造の生成と感性表現のタグ付け

正解コーパス中のタグ付けされた文を対象に、前述の方法により解析を行い、木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。その手順は次の通りである。

手順

(1)　できるだけ助詞・助動詞などの機能語をタグ付けすること。(例：～とは～だ)

(2) 係り受け関係で表現されているものは、その対をタグ付けすること。(例:せっかく～のに。腹が～立つ。頭に～きた。)

(3) 必要があれば、動詞で表現されているものをタグ付けする。(例:訴えてやる。)

(4) 上記以外でも、必要であれば、タグ付けしていい。

4.6 抽出規則の自動生成

タグ付けされた木構造上から、抽出規則を自動生成する。抽出規則は XIP の依存規則である。上記の例からは、次のような抽出規則が自動生成される。

| 用言節#1{連用{?*, 助詞[lemma:だけ], 助詞[lemma:で]}, ?*, 用言節#2{連用{?*, 助動[lemma:れる]}| 不満表現(#1, #2) //(1)

この抽出規則により、

不満表現(出すだけで, 取られ) //(2)

が抽出されることになる。なお、係り受け関係も依存規則により記述できる。

4.7 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。そこで、抽出規則の性能を見る指標として、まず抽出規則ごとの抽出正文数と抽出負文数をあげることができる。抽出正文数が多いものは抽出規則としての一般性がある。また少ないものは、個別的であり、もっと一般性のあるものに書き換えるかどうか検討すべきものである。一方、抽出負文数が少ないものは、適合率が高くて良い抽出規則である。抽出負文数が多いものは、適合率が低く、おそらく一般化しすぎているのであろう。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。複数の抽出規則があるノードでは、1個を残して他の抽出規則を消去する。その際、できるだけ抽出する負文の少ないものを残す。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。

4.8 抽出規則の改良

クラスター上で抽出文数の多い抽出規則は、この木の深いところに配置されるが、その抽出規則が正文のみを抽出するのであれば、それより浅いノードにある抽出規則は不要である。しかし、実際には、正文のみを抽出する抽出規則は少なく、クラスターを見ながら、抽出規則の改良を行うこととなる。

5 評価

5.1 評価尺度

本タスクの評価尺度は、再現率(抽出された正文数/真の正文数)と適合率(抽出された正文数/抽出された文数)であるが、本タスクの性質上、

再現率を重視する。したがって、F スコアを求めるとすれば、再現率を重視するよう、重み付けを変える必要がある。

5．2 評価対象のコーパス

「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。まず約4,300件ずつをそれぞれ文集合1、文集合2として二つに分割し、感性表現にタグ付けを行った。タグ付けされたものは、それぞれ約400文あった。ここではそれらを規則集合1、規則集合2と呼ぶ。

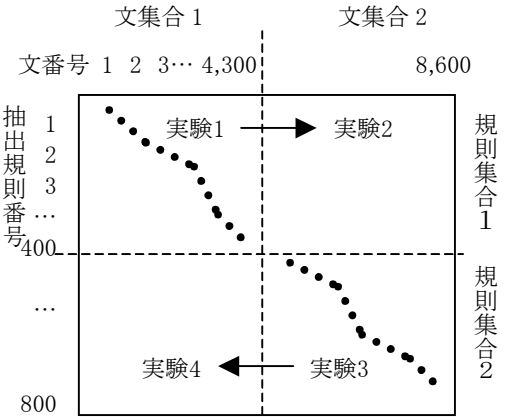


図3 実験計画

5．3 実験計画

二つの文集合と二つの規則集合を用いて、次のような4回の評価実験を行った。実験1では、文集合1を対象に規則集合1を適用した。実験2では、文集合2を対象に規則集合1を適用した。文集合と規則集合を入れ替えて、実験3と実験4を同様に行った。なお、各実験とも、文集合1と文集合2をそれぞれ分野ごとの23個のサブコーパスに分けて、累積値を測定した。ここで、実験1は、訓練用コーパスで抽出規則が機能しているかを見る実験である。このようすを図3に示した。

5．4 評価結果

5．4．1 再現率と適合率

実験結果1から4までを図4と図5に示した。まず、実験結果1を見てみると、いくつかの文に関して抽出規則を記述しなかったため、再現率が 100% ではないが、抽出規則を記述した文はすべて抽出されていた。適合率はだら下がりである。これは、ある文に対する抽出規則が他の感性表現で

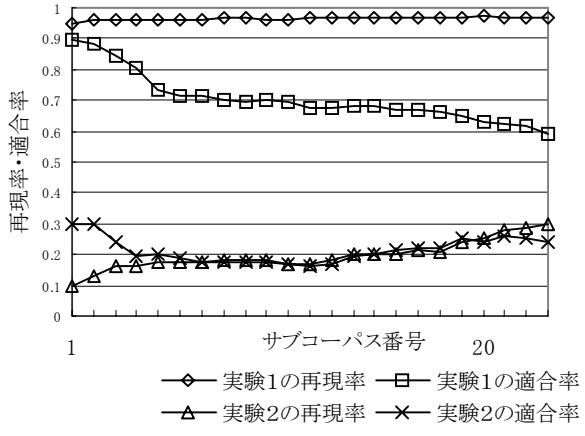


図4 実験1と実験2における累積の再現率・適合率

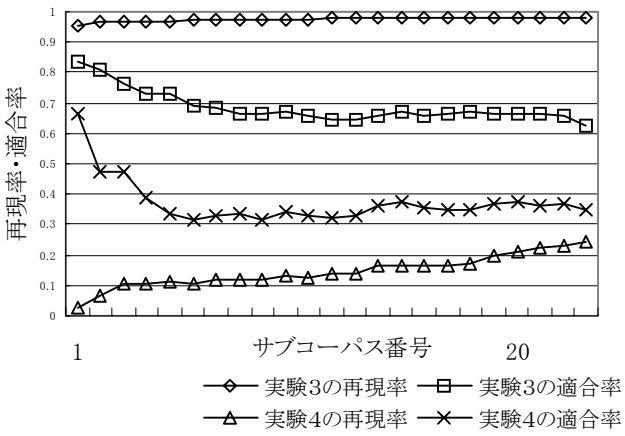


図5 実験3と実験4における累積の再現率・適合率

ない文を抽出しているからである。その場合、負文とされたものが、本当に感性表現でないと言い切れるかは疑問が残る。また、抽出規則の記述が不十分な場合もある。次に、実験2を見てみると、再現率が徐々に上がってきているのがわかる。これは、実験1で得られた抽出規則が、他の文にも適用されていることを示している。適合率に関しては、上がり下がりはあるものの、特に傾向を見てとることはできない。したがって、さらに多くの文にタグ付けをしていくことにより、適合率を下げずに、再現率を上げることが可能であろう。また、再現率曲線の傾きを大きくするには、抽出規則の内容を一般化するなど、定性的な改良が必要である。以上のことは実験3と実験4についてもあてはまる。

5．4．2．　抽出正文数と抽出負文数

次に、改良のための評価尺度である抽出正文数と抽出負文数に関する検討例を示す。次の抽出規則

|用言節{動詞{連用{?*,助詞#1[lemma:ても]}}}| 不満表現(#1)　　//(3)

は、抽出正文数が24文で抽出規則中最多の正文を抽出しているが、同時に、抽出負文数が168文もある最悪の抽出規則でもある。この規則は助詞「ても」を含む文を必ず抽出する。したがって抽出規則を記述する際に「ても」だけではなく、たとえば受けを記述することが必要となることがわかる。抽出正文が多い規則は、必ずしもそれ以上に抽出負文が多いというわけではない。次の規則、

|体言節{名詞{連用{名詞#1[lemma:腹],助詞#2[lemma:が]}}},用言節{動詞{終止連体{動詞#3[lemma:立つ]}}}| 不満表現(#1,#2,#3)

は、抽出正文が14文、抽出負文が1文、したがって適合率が 93.3%といういい規則である。

5．4．3　クラスタリング

クラスタリングを行うと、次の規則、

//(39)

|用言節{動詞{連用{?*,助詞#1[lemma:だけ],助詞#2[lemma:で]}}},?*,用言節{動詞{連用{?*,助動#3[lemma:れる]}}},?*,体言節{名詞{連用{名詞#4[lemma:納得]}}},用言節{動詞{終止連体{?*,助動#5[lemma:ない]}}}| 不満表現(#1,#2,#3,#4,#5)

の下に、次の規則、

////(182)

|体言節{名詞{連用{名詞#1[lemma:納得]}}},用言節{動詞{終止連体{?*,動詞#2[lemma:いく],助動#3[lemma:ない]}}}| 不満表現(#1,#2,#3)

が配置され、この規則と一緒に、

//(399)

|体言節{名詞{連用{名詞#1[lemma:納得]}}},用言節{動詞{終止連体{動詞#2[lemma:いく],助動#3[lemma:ない]}}}| 不満表現(#1,#2,#3)

が配置される。ここで、規則(39)は正文を1文抽

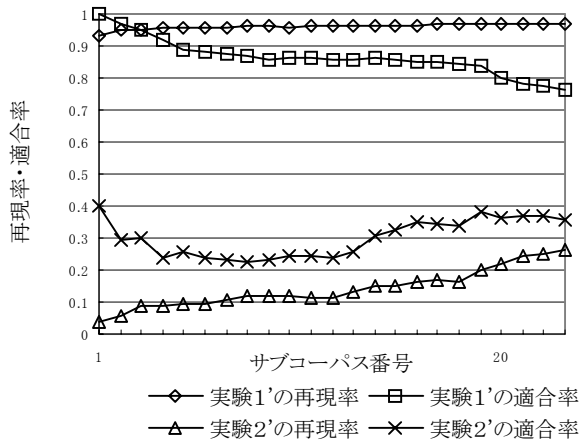


図6 実験1'と実験2'における累積の再現率・適合率

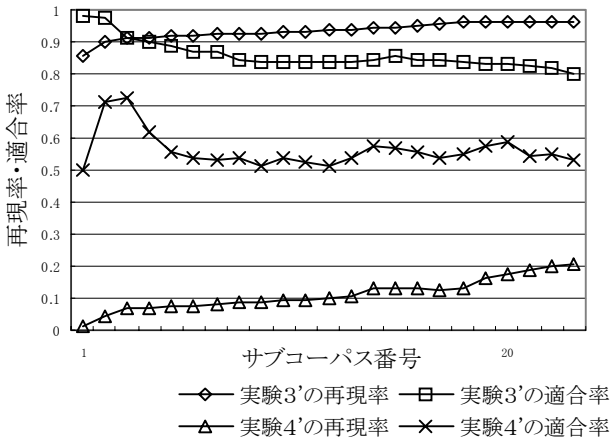


図7　実験3'と実験4'における累積の再現率と適合率

出するが、規則(182)と(399)は、他に四つの正文も抽出する。また、これら三つの抽出規則は一つも負文を抽出しないことがクラスター解析からわかる。したがって、(182)か(399)のみを残して、他の二つを消去できることがわかる。ただし、(182)の方が一般化されており、より多くの文を抽出できる規則である。

5．4．4　改良の試行

ここでは、定量的な検討に基づき、改良を試行してみて、次のように適合率を向上させた。適合率が極端に悪い抽出規則を、規則集合1から一つ、規則集合2から七つ取り除いて、さきほどと同様の実験を実験1’から実験4’まで行い評価した。その結果を図 6 と図7に示した。また、実験1から4までと、実験1’から4’までの、再現率・適合率と、それぞれの差を表1に示した。その結果、8つの規則を省いただけであるが、適合率はすべての場合で、10%以上向上していることがわかった。

（単位:%）

	再現率	差	適合率	差
実験 1	97.1	−0.3	59.1	17.2
実験 1’	96.8		76.3	
実験 2	30.0	−3.9	24.3	11.4
実験 2’	26.1		35.7	
実験 3	98.2	−1.8	62.2	17.6
実験 3’	96.4		79.8	
実験 4	24.2	−3.4	34.9	18.4
実験 4’	20.8		53.3	

表1　実験計画間での再現率・適合率と差

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。すでに述べたように、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。それでも、再現率を向上させることが期待できるので、コーパスの量を大きくしていくことは重要である。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。また、本報告で述べた感性表現と、そのような表現をするにいたった理由や原因に関する依存関係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