文内の単語共起照合に基づくクレーム抽出方式の性能評価

2X−03

永井 明人　高山 泰博　鈴木 克志

三菱電機株式会社　情報技術総合研究所

1. はじめに

大規模文書から目的の情報を絞り込んで迅速に入手する目的指向テキストマイニング技術を開発している。本稿では、クレームの特徴表現を文内の単語共起で表現したクレーム抽出規則を用いて、インターネット上の Web 文書からクレーム情報を抽出する方式を述べ、クレーム抽出規則の大規模化と、本方式の抽出性能を評価した結果を報告する。

2. クレーム抽出の課題

インターネットを利用した情報発信が盛んになり、一般ユーザからの情報が広く公開されるようになった。また、EC 拡大に伴い、データウエアハウスやコールセンタでは、CRM システムへの顧客メール数が急増している。これらの大量の文書からクレーム情報を抽出して、クレームへの迅速な対応や、顧客の潜在ニーズ発掘などを実現する要求が急速に高まっており、従来から特定の意図や意見を抽出・分類する技術として、意図認識技術[1]、メール自動分類技術[2]、インターネットからの製品評判抽出技術[3]などが提案されている。しかし、[1][2]は分析対象となる業務に依存したテンプレートや辞書などの抽出知識を要し、幅広い内容を含む Web 文書への適用が困難である。また、[3]は、抽出表現を単語として照合するため、複数の単語により意味を成す表現を抽出できなかった。これらに対し我々は、意図(クレーム)を表現する一般的な特徴表現を、複数の単語の共起パタンとして規則化し、意図抽出を行なうアプローチを提案する[4]。以下、本方式の概要を説明する。

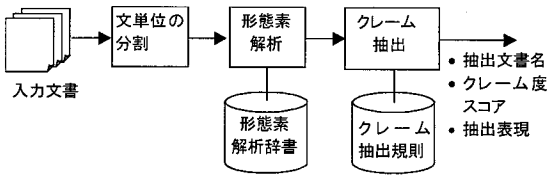

図 1：クレーム抽出方式の全体構成

"Evaluation of a claim extraction method based on verifying word co-occurrence"
NAGAI Akito, TAKAYAMA Yasuhiro, SUZUKI Katsushi
Information Technology R&D Center
Mitsubishi Electric Corporation

3. 提案方式の概要

図 1に本方式の全体構成を示す。まず、文内の単語共起照合を行なうために、入力された文書 D を文単位の解析単位に分割する。次に、形態素解析の後、単語見出しと品詞情報を含む形態素解析結果がクレーム抽出部へ入力される。クレーム抽出部では、クレーム抽出規則を参照して、解析単位中の形態素列と単語共起パタンとの照合を行なう。

クレーム抽出規則は、重み付きの単語共起パタンで表現する。具体的には、表 1に示すように、単語見出しと品詞の複数の組で表現された単語共起パタンに、クレームの度合いを表わす重みを付与して定義される。単語共起パタンの照合では、各単語の順序関係が保持され、また、各単語間は、任意の文字列が許容される。

表 1：クレーム抽出規則の例

番号	単語共起パタン	重み
規則 1	納得(名サ)/でき(活用)/ない(助動詞)	1.0
規則 2	対応(名サ)/腹(名詞)/立(タ五)	1.0
規則 3	良識(名詞)/疑(ワ五)	1.0
…	…	…

これらの単語共起パタンが解析単位の形態素列に存在すれば、文書 D に対するクレーム度スコアにクレーム抽出規則の重みを加算していく。文書 D 全体の照合が終了した際のクレーム度スコアが閾値を越えた場合に、文書 D をクレーム文書と判定し、抽出表現と共に出力する(図 2参照)。

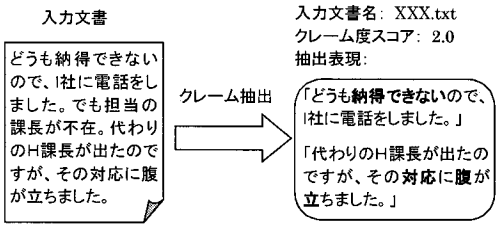

図 2 ：クレーム抽出結果の例

4. 評価

文献[3]のクレーム抽出規則を拡張し、本方式の抽出性能を評価した結果を述べる。

4.1. クレーム抽出規則の拡張

Web 上で公開されているクレーム文書を視察し、クレームの特徴表現を抽出して、規則化を行なった。クレーム抽出規則の重みは主観的に 0.0～1.0 の範囲で付与した。表 2に、視察文書数を増加させた場合に、一文書あたりで得られた新規のクレーム抽出規則の増加率 (規則数の増分/視察文書数)を示す。これより、視察文書数が増加するに伴い、増加率は減少する傾向が見られ、クレーム表現の被覆性が向上していることが分かる。ただし、増加率は飽和していないため、更に規則の拡張が必要である。また、単語共起パタンの構成単語数の分布 (図 3)を見ると、2 単語以上で構成されるクレーム表現が全体の 6 割以上を占めていることが分かる。これらは、従来の単語単位の照合方式では抽出が困難であり、複数の単語からなるクレーム表現を抽出対象とすることの重要性を示している。

表 2：クレーム抽出規則の増加率 (一文書平均)

視察文書数	特徴表現数	総規則数	増加率
10	103	350	11.0
40	215	600	6.3
50	255	859	5.2
37	163	1024	4.5
合計 137	合計 852	1024	－

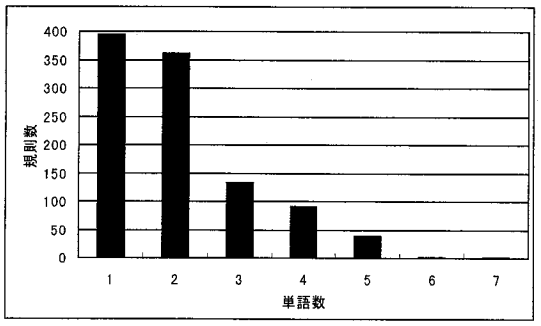

図 3：単語共起パタンの構成単語数の分布

4.2. 抽出性能

視察文書とは別途収集した評価文書 (携帯電話に関する掲示板の発言 1000 件)に対してクレーム抽出を行ない、再現率と適合率を評価した。図 4に、クレーム抽出性能を示す。図中、横軸はクレーム判定の閾値として設定したクレーム度スコアを示し、縦軸は各閾値における再現率と適合率を示す。これより、クレーム度スコアの閾値が3.5以上の領域で適合率80%以上の精度が得られたことが確認できた。なお、クレームと判定された文書を調査した結果、非クレーム表現にも適合している曖昧な単語共起パタンが若干存在しており、クレーム/非クレームの分離性を高めるための規則の洗練化を今後実施することで、更に適合率の向上が見込まれる。また、再現率に関しては、クレーム度スコアの閾値が 1 の付近で 80%程度となっているものの、全体的に高くない値となっている。これは、

(1) 評価対象とした掲示板の発言は一般に短いため、適合したクレーム抽出規則数が少なく、全体的なクレーム度スコアが低い。
(2) クレーム抽出規則の被覆性が十分でない。

などの原因があげられる。今後、(1)に対しては、抽出対象の文書を十分なテキスト量を持つような単位として収集する、(2)に対しては、クレーム抽出規則数を更に増強させて被覆性を高め、抽出漏れを減少させる等の改良課題が考えられる。

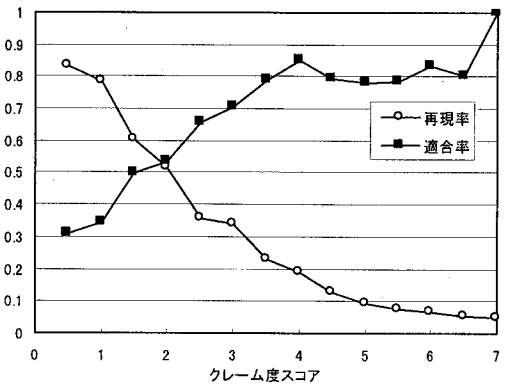

図 4：クレーム抽出性能

5. おわりに

複数の単語からなるクレーム表現の抽出方式を評価し、良好な抽出精度が得られる見通しを得た。今後は更にクレーム抽出規則数の増強を継続して行ない、方式改良と詳細評価を進める予定である。

[参考文献]

[1] 諸橋, 他 “テキストマイニング：膨大な文書データからの知識獲得 － 意図の認識 －,” 情報処理学会 第 57 回(平成 10 年後期) 全国大会 3-75, 1998.
[2] “日本語完全対応 e メール自動分類・配信ソリューション － MatchMail-CallCenter －,” ビジネスコミュニケーション Vol. 37, No. 5, 2000.
[3] 立石, 他 “インターネットからの評判情報検索,” 情報処理学会 研究会資料 (NL 144-11), pp. 75, 2001.
[4] 永井, 他 “CRM における顧客メール分析手法の検討,” 情報処理学会 第 62 回(平成 12 年後期) 全国大会 3-81, 2000.