E-16 単語共起照合に基づくクレーム抽出方式の改良

Improvement of a claim extraction method based on verifying word co-occurrence

永井 明人† 高山 泰博† 鈴木 克志†
Nagai Akito Takayama Yasuhiro Suzuki Katsushi

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パタン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る。方式改良として、クレーム抽出規則の大規模化と、単語共起パタンの照合処理の改良を行ない、本方式の抽出性能を評価した結果を報告する。

2．クレーム抽出への要求と課題

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した[4]。さらに、クレーム抽出規則の拡張を行なって抽出精度の評価を行なった[5]。以下、本方式の概要を説明し、実施した改良検討と実験結果について報告する。

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。

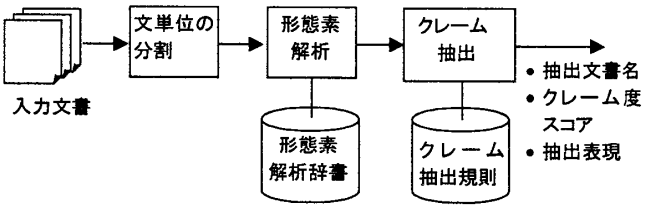

図 1：クレーム抽出方式の全体構成

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)。

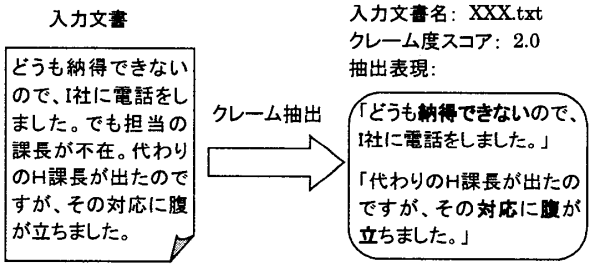

図 2：クレーム抽出結果の例

4. 提案方式の改良課題

抽出精度の向上のためには、再現率と適合率の両者の観点から改善する必要がある。文献[5]における本方式の実験結果を分析し、以下の改良課題を抽出した。

(1) 再現率の向上

- 文献[5]では、クレーム抽出規則を 1024 ルールへ拡張したが、多様なクレーム表現をカバーするには不十分な規模であるため、更に増強する必要がある。
- 規則の増強にあたっては、単語共起パタンで定義された単語に関する類義語も包含する必要がある。

(2) 適合率の向上

- 単語共起パタン内の単語間では、任意数の形態素の存在を許容しているため、長い文内で遠く離れた位置の単語同士がミスマッチする原因となっていた。このため、単語間の距離に制約をかける必要がある。
- 一単語のみからなる単語共起パタンでは、クレームか否かの判定が曖昧になるため、規則の作成では、なるべく 2 単語以上で定義する必要がある。

† 三菱電機株式会社, Mitsubishi Electric Corp.

これらの課題に対し、下記の改良を行なった。

4.1. クレーム抽出規則の大規模化

既存のクレーム抽出規則 1024 ルールに対して、各ルールの単語共起パタンを人手により類義語展開させて追加し、クレーム抽出規則の大規模化を行なった。類義語展開の作業では、一単語のみからなる単語共起パタンは抑制し、なるべく複数の単語からなる単語共起パタンで定義するように留意した。この結果、総ルール数は 10268 となり、類義語展開により約 10 倍のルール規模になった。

4.2. 照合処理の改良

単語共起パタンの照合処理において、単語共起パタンで定義した各単語間の距離に対して制約を行なった。各単語間では、指定された N 個の形態素数まで許容するものとし、距離が N を越えた場合、該当する単語共起パタンの規則適用仮説を破棄するように照合処理を改良した。

5. クレーム抽出実験

上記の改良効果を検証するために、Web 文書に対するクレーム抽出実験を行なった。

5.1. 実験条件

全文検索エンジンにより Web から収集した 3215 文書を評価文書セットとした。これらに対し、表 2に示す判定基準に従ってクレーム文書と非クレーム文書とを人手で判定して、正解文書リストを作成した。クレームと判定された文書数は 97 である。なお、クレームか、非クレームかの判断に迷った場合は、非クレーム文書とした。

表 2：クレーム/非クレームの判定基準

クレーム	・一般的なクレーム表現。 ・製品の故障や不具合情報に関する表現。 ・口論内容での相手への明確なクレーム表現。
非クレーム	・一般的なクレーム表現を含まない文書。 ・製品の故障や不具合情報を含まない文書。 ・クレームの相談窓口情報を紹介する文書。

5.2. 実験結果

まず、規則を大規模化した場合の再現率に与える効果を図 3に示す。図中、横軸はクレーム判定のスコア閾値を表わし、縦軸は各閾値における再現率と適合率を表わす。また、(a)はクレーム抽出規則(1024 ルール)を適用した場合、(b)は大規模化したクレーム抽出規則(10268 ルール)を適用した場合を示す。これより、再現率が高い領域であるスコア閾値 10 において、82%から 87%への向上を確認した。ただし、顕著な効果ではなく、1 万ルール規模ではクレーム表現の被覆性がまだ十分でないと考える。

次に、単語間の距離制約が適合率に与える効果を図 4に示す。実験では、単語間の距離制約を N = 3、5、7、10 の 4 通りについて評価し、これらの中で適合率が最良であった N = 7 の場合を図 4の(c)で示している。これより、再現率の低下が若干見られたものの、(b)と比較して(c)の場合は抽出誤り数が減少し、スコア閾値が 20 から 50 までの範囲で適合率が改善する効果が見られた。

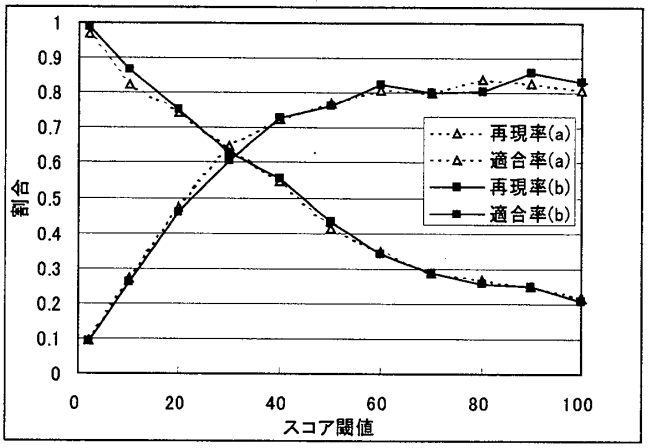

図 3：抽出精度（規則の大規模化の効果）

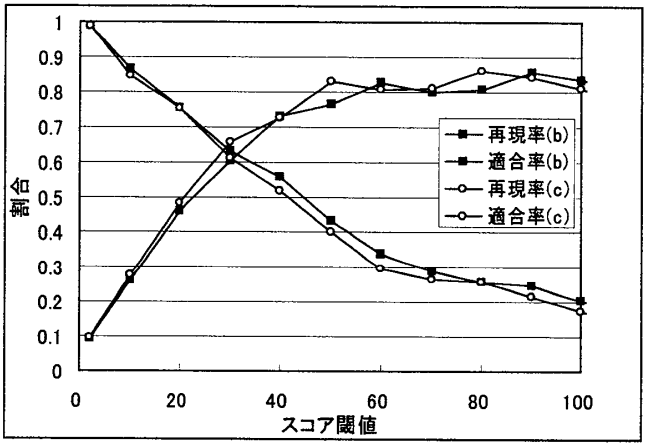

図 4：抽出精度（単語間の距離制約の効果）

6. おわりに

複数単語のクレーム表現の抽出方式改良を検討し、実験により評価した。その結果、改良の効果を確認した一方で、ルール規模、照合方式ともに改善の余地があることが分かった。今後は、継続して改良と詳細評価を進めるとともに、抽出対象をクレーム以外にも拡張していく予定である。

[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.
[5] 永井, 他 “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,” 情報処理学会 第 64 回(平成 14 年後期) 全国大会 3-17, 2002.