1N-01

お客様情報の視点別分類

加藤 弘之樹, 上田 良寛
富士ゼロックス株式会社

1. はじめに

企業にとってお客様情報の活用は、今後の開発・営業の指針を決定する重要な要因である。従来のお客様情報の分類は、予め用意されたカテゴリのどこに属するかの判断を人手で行うというものであり、コストの他個人差の問題も生じる。また、分析に関しては上記分類の結果から経験則で予測する程度しかなされていない。よって、分析に利用するために、文書を視点別に分類する技術を実現することが課題となる。

2. 関連研究

分類の既存研究で、課題解決に利用できる研究は少ない。[清田 98]では、単語の出現頻度によるベクトルを基本とし、付随する助詞によって重み付けをする。電子ニュース記事を、企業名や製品名に対して重み付けした分類を提供できる。これは視点別分類と見なすことができるが、中心となる述語を固定しなければ意味がない。

[乾 98]では、文末表現を用いて、自由回答アンケートをタイプ別に分類している。文末表現はお客様の声の種類を要望、問題、質問などに分類できるが、それをさらに掘り下げた分類はできない。

[諸橋 98]は、意図表現により顧客からの声を要望や質問などに分類し、さらにキーワードを抽出する。しかし、そのキーワードがどのようなコンテキストで現れるものかはわからないので、視点別分類には利用できない。

3. データ分析

弊社のお客様情報 17 データ 61 文書(以下「分析文書群」とする)を用いて、分類の手法を見出すための分析(分類のハンドシミュレーション)を行った結果、特有の表現パターンがあることがわかった。パターンは係り受けを基本単位とし、係り語と受け語本来の意味、付随の否定表現、関係を要素とする。「問題」に関する視点としては「部位」「症状」がある。

【例 1】「ランプが破裂した」という表現は、

≪[部位] が [異常語] ⇒ 部位:#1, 症状:#2≫

というパターンにマッチし、

部位=「ランプ」、症状=「破裂」

というキーワードが得られる。

例1のようなパターンを用意すれば視点別キーワードが抽出でき、これをシソーラスで分類すれば良い。

4. 分類システムの設計

本システムの基本構成を図 1 に示す。

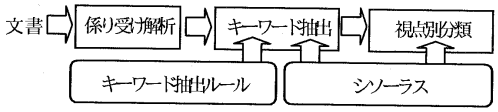

図 1 分類システムの基本構成

4.1. 係り受け解析とキーワード抽出ルール

文書に対して係り受け解析を施す。

一方、分析文書群から例 1 のようなパターンを得て、それらを図 2 のように 16 個の抽出ルールにまとめた。

- [部位]は別に用意するシソーラスにおいて概念「部位」に包含されることを、other は概念「部位」「症状」のいずれにも包含されないことを示す。
- 異常とはその語が本来異常を意図する語であるかどうかの判定であり、この情報を上記シソーラスにおいて保持しておく。
- 否定とは用言の後に続く付属語列が否定の意味を持つかどうかの判定であり、別に用意される否定表現パターンとのマッチングによる。

Classification of Customers' Voices from Multiple Viewpoints
Hiroyuki Kato, Yoshihiro Ueda
Fuji Xerox Co., Ltd.
430 Sakai, Nakai-machi
Ashigarakami-gun, Kanagawa 259-0157, Japan

ルール	係り語			関係	受け語			部位	症状
	表層	異常	否定		表層	異常	否定		
NS1	other	t	f	の,では,た,い	[症状]	t	f		受
NS2	other	f	t	の,では,た,い	[症状]	t	f		受
NS3	other	t	f	の,では,た,い	[症状]	f	t		受
NS4	other	f	t	の,では,た,い	[症状]	f	t		受
NS5	※1	f	f	の	[症状]	f	f		受
PN1	[部位]			では,を,は,の,が	other	f	t	係	
PN2	[部位]			では,を,は,の,が	other	t	f	係	
PN3	[部位]			の	※1		f	係	
PS1	[部位]			では,が,に,の	[症状]	f	t	係	受
PS2	[部位]			では,が,に,の	[症状]	t	f	係	受
SN1	[症状]	f	f	が,は,より,から,も,と,で,に,の	other	f	t		係
SN2	[症状]	t	f	が,は,より,から,も,と,で,に,の	other	f	f		係
SN3	[症状]	f	f	が,は,より,から,も,と,で,に,の	other	t	f		係
SN4	[症状]	f	f	が,は,より,から,も,と,で,に,の	※2		f		係
SP1	[症状]	f	t		[部位]			受	係
SP2	[症状]	t	f		[部位]			受	係

※1　薄, 濃, 悪, わる, 不可, な(形容詞), 丸ま, 荒れ, 故障
※2　生じ, 発生, でき, 出, 出やす

図 2 表現パターン(キーワード抽出ルール)

ルール PS2 は、係り語概念「部位」に包含され、関係が「では」「が」「に」「の」のいずれかであり、受け語が概念「症状」に包含され異常であり否定でなければ、係り語を視点「部位」の、受け語を視点「症状」のキーワードとして各々抽出する。

【例 2】ルール PS2 は例 1 の表現にマッチし、「部位」のキーワード「ランプ」と「症状」のキーワード「破裂」を抽出する。これは例 1 にて得られるパターンをルール化したものを包含する。

4.2. 視点別分類

抽出したキーワードを用いて文書を分類する。抽出に用いたものと同じシソーラス上に文書を配置する。その後、ユーザの指定するクラスタ数に近くなるようにクラスタのマージ/細分割を行う。

5. 分類システムの実装

この手法に基づく分類システムを実装した。ここでは「問題」に関する 2 視点による分類を行う。ユーザが視点と生成クラスタ数を指定するとその分類結果を表示するものである。抽出ルールは図 2 の 16 個であり、シソーラスは分析文書群から構成した。

6. 結果と評価

テスト結果を図 3 に示す。

6.1. システムの正当性と手法の有効性

分析文書群のハンドシミュレーション結果とシステムのキーワード抽出結果を比較した。システムの再現率は 100%であり、これはシステムが設計通りに実装されていることを示す。

適合率も 98.1%と高く、シソーラスが十分に詳細であったといえる。

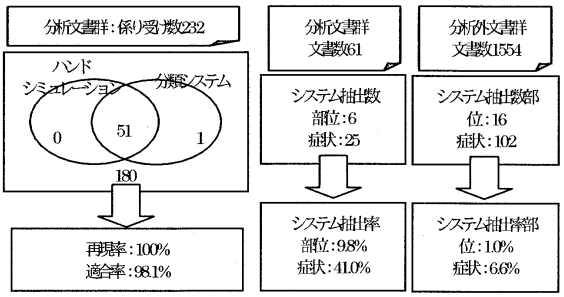

図 3 テスト結果

6.2. シソーラスとキーワード抽出ルール

分析文書群と分析外文書群(409 データ 1554 文書)に対するシステムのキーワード抽出率[ii]を比較したところ、図 3 のように部位と症状のいずれにおいても後者が著しく低い。これは、分析対象外文書に適したシソーラスや抽出ルールの不足のためであり、これらを整備することで改善されると考えられる。

7. まとめ

文書中の係り受けを用いた視点別分類の技術を開発した。評価の結果、手法は効果的であるがシソーラスと抽出ルールの整備が必要であることがわかった。今後の課題はその整備に加え、視点の種類や対象文書を拡充することである。

参考文献

[乾 98] 乾裕子, 内元清貴, 村田真樹, 井佐原均: 文末表現に着目した自由回答アンケートの分類, 情報処理学会研究報告, pp. 181-188, 1998.

[清田 98] 清田陽司, 黒橋禎夫, 中村順一, 長尾真: 構文情報を利用した電子ニュース記事のクラスタリングシステムの作成と評価, 情報処理学会研究報告, pp. 77-84, 1998.

[諸橋 98] 諸橋正幸, 那須川哲哉, 長野徹: 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 回全国大会, No. 3, pp. 75-76, 1998.

---

[i] このデータは 1 データ中にお客様の声、補足、対処など 5 個のテキストフィールドを含み、一部のテキストフィールドが空の場合もある。この各々を単独の文書として扱っている。分析上異なる扱いをする必要が生じる可能性を考えた処置であるが、現在のところ同じパターンで対処できている。

[ii] 全文書に対する、その視点のキーワードを 1 個以上抽出した文書の割合を示す。