係り受け解析を用いたテキストマイニング支援システム

倉田 早織・加納 敏行・牧野 恭子・早川 ルミ
東芝ソリューション株式会社　ＩＴ技術研究所
{kurata.saori, kano.toshiyuki, makino.kyoko, hayakawa.rumi}@toshiba-sol.co.jp

1　はじめに

テキストマイニングとは、単なる検索や分類整理とは異なり、大量の文書データの内容を総合的にとらえることで初めて得られる知見を抽出するための内容分析の技術である[1]。例えば、アンケートの自由記述回答の意見分析[2]、営業担当者の日報分析[3]、コールセンターの問い合わせ分析[1]等に用いられる。

一般にテキストマイニングでは、見つけたい表現をエントリに持つ辞書を用いる。しかし辞書の構築には時間がかかり、コスト低減が課題となっている。本論文ではこの課題を解決するためのテキストマイニング支援システムに関して報告する。

テキストマイニングに用いられる辞書の構築方法としては、語彙ネットワークを利用した手法、共起情報を利用した手法、周辺文脈の情報を利用した手法がある[4]。また、共起パタンを用いて、文書から評価表現と評価対象表現を抽出することにより、評価表現辞書を作成する手法[5]や、表現頻度分析機能が搭載されている辞書構築支援ツール[6]がある。

表現分析の精度向上を目指した上記研究と異なり、本研究では、テキストマイニングの経験が浅いユーザを想定し、業務全体の効率化を目指している。

2　テキストマイニングのプロセス

テキストマイニングでは、分析対象文書を収集することも重要な作業であるが、ここでは収集された文書の分析に焦点を当ててプロセスを紹介する。また、評価表現と評価対象表現の分析を例に挙げて、テキストマイニングのプロセスを紹介する。評価表現は「良い」、「使いやすい」等、ある物事に対する評価を述べる際に使われる表現である。評価対象表現は評価の対象となった物事を記述するのに使われる表現である。例えば、「サービスが良い」という文では、評価表現「良い」に対する評価対象表現は「サービス」である。

テキストマイニングの作業プロセスを、エキスパートの経験に基づき分析した。この結果、評価カテゴリの設定、辞書の構築、評価対象表現の分析、の３つのプロセスからなることが明らかとなった。それぞれのプロセスの詳細を以下に述べる。

2.1 評価カテゴリの設定

分析作業者は、評価を整理するためのカテゴリを決める。例えば、製品に対するユーザの評価をアンケートの自由記述回答から分析する場合、カテゴリは「好評」、「不評」、「要望」等となる。

2.2 辞書の構築

テキストマイニングでは、所望の評価表現を文書から見つける際に評価表現辞書を用いる。評価表現辞書の例を表 1 に示す。評価表現辞書には、「良い」、「いい」、「美味しい」などの評価を示す表現とそれに対応するカテゴリが登録されている。辞書の表現とマッチする文字列を文書から見つけ出し、表現に対応するカテゴリが「好評」であれば、この文字列を好評カテゴリの評価表現と解釈する。

あらかじめ用意されている一般的な評価表現辞書を用いる方法も考えられるが、分析対象のドメインに特化した評価表現辞書を使用した方が

表 1 評価表現辞書の例

カテゴリ	表現
好評	良い
好評	いい
好評	美味しい
不評	使いづらい
要望	欲しい

より正確で詳細な分析ができる。このような場合には分析対象の文書に基づいた評価表現辞書を構築することが必要となる。

評価表現辞書の構築は 2 つのサブプロセスからなる。各サブプロセスの内容を説明する。

2.2.1 評価表現の抽出

大量にある分析対象文書の中から人手で解析できる程度の量のサンプル文書を選択する。サンプル文書を読んで、評価表現を抽出する。さらに、その評価表現が属するカテゴリを決める。例えば、「サービスが良い」という文の場合、評価表現は「良い」、対応するカテゴリは「好評」とする。

2.2.2 表現の登録

評価表現とそれに対応するカテゴリの組を辞書に登録する。しかし、単純にサンプル文書から抽出された評価表現をそのままの文字列で辞書の表現として登録しただけでは、評価表現の取りこぼしが多い。例えば、辞書に登録された表現が「良い」の場合、「フロントのサービスが良かった」の「良かった」を見つけ出すことができない。そこで「良い」の活用形「良かった」という表現の登録も必要となる。このように、活用形も辞書の表現として登録する必要がある。ただし、例えば「良くない」という否定形の場合は、カテゴリは「不評」となるので、活用形ごとに対応するカテゴリを決める必要がある。

属するカテゴリの評価表現取得の再現率を向上させ、適合率を保ちながら、表現を登録しなければならない。

2.3 評価対象表現の分析

評価表現辞書と分析対象文書を用いて、評価表現と、その評価表現に対応している評価対象表現を見つけ出す。例えば、「フロントのサービスが非常に良い」の場合、評価表現は「良い」、評価対象表現は「フロント」や「サービス」である。この後、評価表現や評価対象表現の出現頻度の集計等を行い、自由記述回答を分析する。

3 テキストマイニングにおける課題

前章で述べた辞書の構築は、人がサンプル文書を読み評価表現を抽出しなければならない。この作業は文書の内容を読んで意味を理解し、評価表現を正しく認識して、辞書に登録する表現を抽出する必要がある。また、評価表現の取りこぼしを減らすために、辞書に登録する表現の活用形などのバリエーションを考える必要がある。

このように人手による辞書の構築は時間がかかり、そのコスト削減が課題となっている。

4 テキストマイニング支援システム

辞書の構築プロセスのコスト削減という課題を解決するために、このプロセスを支援するシステムを開発した。機能構成を図 1 に示す。このシステムは、表現頻度分析、辞書登録・編集、辞書検証、評価対象表現取得の 4 つの機能から構成される。

4.1 表現頻度分析

文書に出現する共起表現や単語の頻度を算出する。ここでは同一文中において係り受け関係にある 2 つの単語を共起表現としている。算出は、品詞別だけでなく、共起表現の場合は係り受け関係別にも行う。

共起表現の頻度の算出結果から、そこに含まれる評価表現と評価対象表現を見て、評価表現辞書に登録する表現(辞書登録表現)の候補を選別でき

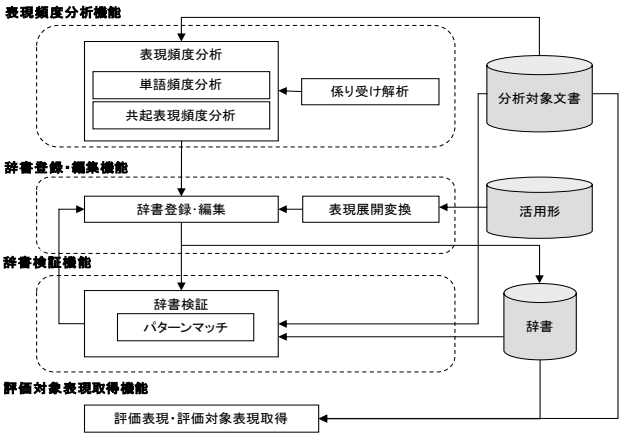

図 1 システムの機能構成

る。さらに、単語頻度の算出結果から、辞書登録表現の候補を選別することもできる。この 2 つの分析結果を利用することにより、辞書登録表現の候補を、漏れなく見出せる効果が期待される。

なお、辞書登録表現の候補を漏れなく出力することを重視しているため、係り受け解析においては、係り受けグラフの結果を全部使用している[7]。

4.2 辞書登録・編集

表現頻度分析で得られた辞書登録表現の候補を辞書に登録する。辞書に登録すべき活用形を生成する機能(図 1 における表現展開変換)等により、2.2.2 表現の登録で行うべき作業が自動でできる。更に、手動で任意の文字列を表現として登録することもできる。

4.3 辞書検証

評価表現辞書とサンプル文書を用いてパターンマッチを行い、辞書の表現とサンプル文書のマッチした箇所の対応を出力する。主に、手動で辞書に表現を登録した際に、意図通りに辞書登録がされたかどうかを確認するために用いる。

4.4 評価対象表現取得

評価表現辞書を用いて、分析対象文書から、評価表現と、その評価表現に対応している評価対象表現を取得する。この機能においても、係り受け解析を用いている。

5 実験による評価

表現頻度分析機能によって、人がサンプル文書を読まなくても辞書登録表現の候補が自動で得られるため、人が読んで表現を抽出する作業コストの削減が期待される。

実験において想定した作業は、アンケート自由記述回答を分析するための、「好評」、「不評」、「要望」の 3 つのカテゴリに関する評価表現辞書の構築、および、構築した評価表現辞書を用いた評価対象表現の分析である。

5.1 実験作業

被験者として、テキストマイニング未経験の日本語母語話者 3 名を対象とした。その内 1 名は、対照データ取得のため、システムを使わずに手作業で辞書構築作業を行った。分析対象データは、駅の自動券売機に関するアンケートの自由記述回答 400 件である。この中から無作為に取り出した 200 件のデータをサンプル文書、残り 200 件を評価用データとした。

5.2 実験結果

5.2.1 辞書の構築時間

評価表現辞書の構築にかかった時間を表 2 に示す。システム不使用の場合の時間は、被験者がサンプル文書を読み、「好評」、「不評」、「要望」のいずれか 1 つに対応する評価表現の抽出を行い、辞書への登録までにかかった時間である。

システムを使用した場合の辞書の構築時間は、使用しなかった場合の半分以下となった。

5.2.2 構築された評価表現辞書の有効性

評価表現とその評価表現に対応している評価

表 2 評価表現辞書の構築時間

	構築時間(分)
システム不使用	61
システム使用	23
	22

対象表現の組(評価組)が分析の目的に合致していること(有効性)が、その後の評価対象表現の分析において重要である。

今回の実験で構築した評価表現辞書を用いたときに得られる評価組の有効性を確認するため、正解数と精度を算出した。結果を表 3 に示す。正解数は、評価組の要素が両方とも正しかった数、精度は正解数を取得された評価組の数で割ったものである。

表 3 評価表現辞書の有効性評価結果

辞書構築条件	辞書のエントリ数	正解数		精度(%)	
		サンプル文書	評価用データ	サンプル文書	評価用データ
システム不使用	335	139	78	50.1	34.8
システム使用	159	114	74	62.2	58.3
	147	105	48	50.5	40.4

正解数と精度の両方とも、システム不使用と使用で大きな差はないといえる。その一方で、辞書のエントリ数は、システムを使用した場合、およそ半分以下となっている。

5.3 考察

上述の実験結果より、本システムを使用した辞書構築の作業間が半分となるが、構築された辞書の有効性は、システムを使用しない場合と同等のものであるといえる。また、システムを使用して作成された辞書のエントリ数が小さいので、評価対象表現の取得にかかる時間が短くなる効果も予想される。

今回の実験における正解は、分析対象データから人手で抽出した。その結果、複合語や句単位の表現を正解とするものが多く含まれていた。一方、今回の実験は、システム使用時と不使用時の比較を目的としたため、複合語等の処理を考慮していない状態で行い、出力と正解との比較を厳密に行っているので、精度等の値がやや低くなったと考えられる。

6 まとめと今後の検討課題

テキストマイニング支援システムを開発し、辞書の構築コスト削減の効果を評価し、辞書の品質を保ったまま、構築時間が削減できることを確認した。システムを使うことにより、構築された辞書の品質が均一になり、作業の標準化、作業効率の向上が期待できる。また、実験を通じて、ユーザインターフェース等に関して様々な課題が得られた。得られた課題を解決し、更なる作業時間の削減を目指す。

参考文献

[1] 那須川哲哉､『テキストマイニングを使う技術/作る技術』、東京電機大学出版局、2006

[2] 磯島昭代､「テキストマイニングを用いた米に関する消費者アンケートの解析」､農業情報研究、Vol.15 No.1 pp.49-60、2006

[3] 市村由美、鈴木優､「テキストマイニング技術と応用」､東芝レビュー､Vol.56 No.5 pp.19-22、2001

[4] 乾孝司、奥村学､「テキストを対象とした評価情報の分析に関する研究動向」、自然言語処理、Vol.13 No.3 pp.201-241、2006

[5] 小林のぞみ、乾健太郎、松本裕治、立石健二、福島俊一､「テキストマイニングによる評価表現の収集」、情報処理学会自然言語処理研究会、NL154、2003

[6] 市村由美、鈴木優、酢山明弘、折原良平、中山康子､「日報分析システムと分析用知識記述支援ツールの開発」、電子情報通信学会論文誌、Vol.J86-D-II No.2 pp.310-323、2003

[7] 平川秀樹、天野真家､「日本語解析における最適解探索」、情報処理学会自然言語処理研究会、NL074、1989