認知言語学的アプローチによる新しいテキストマイニング手法の提案
The method of Text Mining using Cognitive Linguistics approach

秋山 優†1 深谷 昌弘†2
†1 慶應義塾大学大学院 政策メディア研究科 †2 慶應義塾大学 総合政策学部

†1 Yu Akiyama †2 Masahiro Fukaya
†1 Graduate School of Media and Governance, Keio University
†2 Faculty of Policy Management, Keio University

E-mail:maooon@sfc.keio.ac.jp

概要
近年，インターネットの発展により，電子化された大量のテキストデータが蓄積されつつある．その結果，大量のテキストデータから有益な情報を抽出するテキストマイニング技術が注目されている．近年の傾向としては，形態素情報のみではなく，係り受け関係を用いたテキストマイニング手法が発展しつつある．しかし，係り受け関係をどのように利用すれば分析者による妥当な意味解釈を支援できるかについては，現在研究途上である．本稿では，田中，深谷が提唱する助詞の操作子機能と用言句の図式構成機能に着目し，両機能によって取り纏められる最小の係り受け構造を基礎単位としたテキストマイニング手法を提案する．

１．はじめに

近年，World Wide Web 等インターネット上の各種サービスを用いた情報発信，情報収集が容易となり，電子化された大量のテキストデータが蓄積されつつある．その結果，大量のテキストデータから有益な情報を抽出するテキストマイニング技術が注目されており，様々な商品([TRUETELLER]など)が開発されている．

テキストマイニングにはデータマイニングや自然言語処理等の手法が用いられてきた．データマイニング手法に基づくアプローチでは，形態素解析されたテキストデータを単語の集合として扱い，キーワードやパターンを抽出する手法が研究されてきた([MDA93]など)．しかし，本来テキストデータ中の個々の単語が担う意味は不確定性を持つ．そのため，例えば単語の共起関係が発見されたとしても，係り受け関係等の情報がないため，分析者が元の文の意味を解釈することは困難である．

最近では自然言語処理技術の発展を受け，形態素解析のみならず係り受け解析を利用したテキストマイニング手法が発展しつつある．例えば，係り受けを含むチャンクを対象にパターン抽出を行うことにより，特定の対象に対する印象や評価を抽出しようとする研究が行われている([舘野 02]など)．

しかしながら，係り受け関係等の情報をどのように利用するかは経験によるところも多く，分析者による妥当な意味解釈を指向した手法が確立されているとは言い難い．本稿では，分析者の意味解釈支援を重視し，認知言語学の知見を援用したテキストマイニング手法を提案する．

２．提案する手法

２－１ 手法

現在テキストマイニングは企業におけるコールセンタ業務からブランドイメージ調査等の社会調査に至るまで，様々な分野で利用されつつある．テキストマイニングでは，コンピュータによってテキストデータから情報やパターンを抽出するが，最終的な意味解釈は分析者が行う．そのため，分析者の意味解釈が曖昧にならないよう，適切な情報抽出が重要となる．

本稿では，田中，深谷が提唱する助詞の操作子機能と用言句の図式構成機能[田中・深谷 98]に着目し，両機能によって取り纏められる最小の係り受け構造(意味図式)を基礎単位としたテキストマイニング手法を提案する．意味図式とは，端的には名詞，助詞，用言句からなる表現を指す(例「携帯を使用する」「携帯が普及する」「親が子供に携帯電話を持たせる」など)．

意味図式を分析の基礎単位とする考え方は，人間の言葉の意味解釈がチャンキングによって行われているという認知言語学における知見に

基づいている．言葉が意味図式までチャンキングされれば，意味解釈の揺れが大幅に縮減され，分析者の解釈の恣意性を抑止することが可能であると考えられる．テキストデータから特徴的な意味図式や頻出する意味図式，または意味図式間の関係を抽出し，整理することによって，分析者はより妥当な意味解釈を行うことができるのではないか．

２－２　関連研究

本手法と関連した研究として，述語(動詞)に注目し述語とその引数(述語と係り受けの関係をもつ単語)からなるグループを基礎単位として扱う研究[松澤 99]等がある．しかし，語順の自由度が高い日本語を対象とした場合，意味解釈の際助詞が果たす機能は極めて重要であると考えられる．そのため，本研究では助詞の操作子機能に着目した意味図式を分析の基礎単位とした．

３．本手法を用いた分析の紹介

３－１　分析システム概要

本手法に基づいた分析を行うにあたり，XRCE(Xerox Research Centre Europe)において開発されている XIP(Xerox Incremental Parser)[AMCR02]を分析システムのプラットフォームとして採用した．その上で，富士ゼロックス株式会社において XIP 上で日本語文法を規則集合として柔軟に記述可能なよう実装された WebGee をカスタマイズしていくことにより，本手法に基づく分析を行っている．

３－２　分析結果

筆者らの研究グループでは，上記分析システムを利用して各種社会調査を行っている[山澤 02]．ここでは，携帯電話に関して述べられたテキストデータ 2,441 件(株式会社国連社による携帯電話に関するアンケート自由解答欄．調査期間：2002 年 4 月 11 日～2002 年 4 月 24 日)を対象として，本手法を用いて話題抽出を行った結果を簡略に述べる．分析は以下の手順で行われた．

1. ターゲット語(携帯電話)を指定し，これを含む意味図式を抽出する
2. 抽出された意味図式に含まれる用言句の類似性に基づきカテゴライズ，話題の分類を行う
3. 各話題のもとでどのような主張がなされているかについて，意味図式間の関係から分析する
4. サブターゲット語(ターゲット語と共起頻度の高い語)についても同様の分析を繰り返し，話題の詳細化を行う

分析の結果，各話題のもとなされた主張について，主張の理由にまで遡って明らかにすることができた．例えば，「子供に携帯を持たせるべきか否か」という話題のもと，賛成反対の主張がどのような理由に基づいているのかについて分析することができた．

また，本分析を行った３人の分析者の間で，分析結果に対する意味解釈に大きな矛盾や違い等は見られなかった．

４．まとめ・展望

本稿では，認知言語学的アプローチから，助詞と用言句の機能に着目し，両者によって取り纏められる最小の係り受け構造(意味図式)を基礎単位とした手法を提案した．今後本手法の有効性を実証するためには，分析システムの機能拡充及び分析事例の蓄積が必要である．

また，展望として本手法を応用した検索システムの開発に取り組む予定である．

・参考文献

[AMCR02] Salah A.t-Mokhtar, Jean-Pierre Chanod, and Claude Roux. Robustness beyond shallowness: incremental deep parsing. In Natural Language Engineering, No. 8(2), pp. 121・44, 2002.

[MDA93]K.Muraki, S.Doi, and S.Ando. Description of the Veniex system as used for MUC-R. In Proceedings of MUC5, pp.147-159, 1993

[TRUETELLER] 株式会社野村総合研究所「TRUETELLER」http://www.trueteller.net/

[舘野 02] 舘野昌一．テキスト感性表現の抽出におけるお客様の声の活用方法．第 4 回日本感性工学会大会，p. 242, 2002.

[田中・深谷 98] 田中茂範，深谷昌弘．＜意味づけ論＞の展開　情況編成・コトバ・会話．紀伊国屋書店，1998.

[松澤 98]松澤裕史．テキストデータからの頻出パターンのマイニング．知識発見のための自然言語処理シンポジウム，1999.

[山澤 02]山澤美由起，テキストマイニングを用いた商品/モノに関するイメージ分析，認知言語学会,2003.