セマンティックアーカイブ活用のための基礎技術としてのテキストマイニングとその周辺

増山繁，酒井浩之 (第７工学系, セマンティックアーカイブコア)

1　はじめに

情報洪水の時代と言われるようにWeb上に溢れている情報に溺れることなく積極的に活用するための情報技術として、テキスト自動要約、テキストマイニング，特に，因果関係などの情報抽出を重点的に研究している。これらはいずれも，セマンティックアーカイブ活用のための基礎技術として重要なものである．本稿では，前年度の報告に引き続き，２００７年４月以降のこれに関連した主な成果を報告する．

2　企業の業績発表記事からの業績要因抽出 [5, 4] と極性付与 [1]

投資などの参考にするためには，企業業績の情報だけでなく，とりわけ，その業績要因が重要であるが，人手によって全ての企業の業績要因を取得するのは大変な労力を要する．そこで，我々は，経済新聞記事から企業の業績記事を抽出し，その中から，業績要因（たとえば,「ハイブリッド車の売り上げが好調だった」）を抽出する手法を提案した [5, 4]．しかし，より有効な情報として利用するためには，抽出した業績要因に対して業績に対する極性（業績が向上する要因なら「ポジティブ」，業績が悪化する要因なら「ネガティブ」）を付与する必要がある．例えば，業績要因「ソフト販売の利益が寄与する」に対しては「ポジティブ」、「繊維部門の不審が響く」に対しては,「ネガティブ」のラベルを付与する手法を開発した [1]．

3　景気動向を示す根拠表現の抽出と分析 [2]

本研究では, 新聞記事から景気動向を示す「根拠となる表現」を統計的手法を用いて自動的に抽出する手法を提案する．また, 抽出された景気動向を示す「根拠となる表現」を景気が回復することを示すPositive表現と悪化することを示すNegative表現に分類する手法も併せて提案する．企業や投資家にとって, 株価や商品の売行きを予測するために, 景気動向を知ることは重要なことである．そこで, 我々は景気動向に関する記事から景気動向を示す「根拠となる表現」を抽出し, それを用いることにより, 景気動向の予測が可能ではないかと考えた．そこで, 景気動向を予測するための材料として, 景気動向の根拠となる表現を抽出し, それが景気が回復することを示すのか, 悪化することを示すのかの判定手法を開発した．

4　仕事文推敲支援に向けた連体修飾部不足に対する受容性判定法 [6]

文章推敲に関する従来研究では，主に，タイプミス，構文構造の複雑さ，表記の揺れを指摘する手法など，表記レベルと統語レベルの手法に重点がおかれていた．それに対して，本研究では，読みやすさを向上させるために，説明が不足していて論理展開が読み取りにくいと感じられる箇所を検出する技術を扱った．情報を正確に伝達するための仕事文（仕事用の文）を対象としていて，文単位での情報不足を推敲対象とする．

5　統計的手法に基づく講義音声書き起こし文書の文境界推定 [7]

本研究では，教師あり学習を使用した形態素解析などを用いない話し言葉を対象とした文境界推定法を提案する．提案手法は，日本語ではモダリティが文末に集中することに着目して文境界の推定を行う．本研究では，多くの統計情報を得ることが出来る講義音声書き起こした文書を対象とした．提案手法を実際の大学における講義音声書き起こし文書に適用した結果を報告する．

6　未知のサイトのWebページからの主要部分抽出手法 [3]

Webページのテキスト情報を有効活用するためには，その主要部分を取り出す必要がある．そこで、平均的なWebページにおいて, 主要なDOMノードがウィンドウのどの位置を被っているかという情報を用いて，それを抽出する手法を提案した．この手法は，サイト全体のスクロールを必要とせず，セグメンテーションに依存しないのが特徴である．

7　係り先候補の相対的な距離を反映した統計的日本語係り受け解析 [6]

係り先候補の相対的な距離を反映した統計的日本語係り受け解析手法を提案した．統計的係り受け解析手法は，文節間の係りやすさを訓練データから推定する．その際，従来手法では，文節間の距離はいくつかのカテゴリに分けられ，推定に用いられる素性として明示的に与えられる．しかし，複数の文節間候

補が同一の距離カテゴリに属する場合，距離による弁別ができないため，最尤の係り先を決定することが困難である場合が多い．そこで提案モデルでは，文節候補集合中の二つの文節候補を逐次的に取り出し，どちらが係り元に近いかを明示させて係りやすさの推定を行う．

8 学生レポート剽窃検出のための著者解析の試み [9]

学生レポートの剽窃自動検出において剽窃元の文章が機械処理可能な状態でない場合に剽窃箇所を検出するためには，文書に表れる書き手の特徴を捉え，レポート提出者が記述していない部分（以下，非著作部分）を推定する方法が考えられる．本研究では，まず，このようなレポート中の非著作部分推定のための予備実験とその考察を行う．その後，ベースラインとなるような単純な非著作部分自動推定法を提案した．

9 構文パターンを用いた因果関係の抽出 [10]

因果関係に関する知識は，自動要約や質問応答システムなど，幅広い自然言語アプリケーションにとって重要な知識のひとつである．しかしながら，大規模な文章集合から因果関係に関する知識を人手で獲得するのはコストと時間がかかり，かつ、世の中にの因果関係をすべて書き出すのは不可能である．そこで，我々は４つの構文パターンと手がかりとなる表現（手がかり表現）を用いて因果関係に関する知識を自動的に抽出する手法を提案する．データには因果関係に関する知識が多く含まれている景気動向に関する記述がある記事（景気動向記事）を用いた．我々は因果関係を抽出する上で重要となる手がかり表現の調査をおこなった後，因果関係の抽出を試みた．

10 まとめ

本稿では，前年度の報告に引き続き，２００７年４月以降のセマンティックアーカイブ活用のための基礎技術としてのテキストマイニングとその周辺に関するおもな成果を報告した．

発表論文

[1] Hiroyuki Sakai, Shigeru Masuyama, Polarity Assignment to Causal Information Extracted from Finantial Articles Concerning Business Performance of Companies, to appear in Proc. AI-2008, 28th SGAI International Conference on Artificial Intelligence, Cambridge, England.

[2] Hiroki Sakaji, Hiroyuki Sakai, Shigeru Masuyama, Automatic Extraction of Basis Expressions That Indicate Economic Trends, Proc.12th Pacific-Asia Conference, PAKDD, LNAI5012, Springer, pp.987-984,

[3] Masanobu Tsuruta, Hiroyuki Sakai, Shigeru Masuyama, An Informative DOM Subtree Identification Method from Web Pages in Unfamiliar Web Sites, IEICE Trans. Information and Systems, Vol.ED.No.4, pp.986-989, 2008.

[4] Hiroyuki Sakai, Shigeru Masuyama, Cause Information Extraction from Financial Articles Concerning Business Performance, IEICE Trans. Information and Systems, Vol.ED.,No.4, pp.959-968, 2008.

[5] Hiroyuki Sakai, Shigeru Masuyama, Extraction of Cause Information from Newspaper Articles Concerning Business Performance, Proc. of the 4th IFIP Conference on Artificial Intelligence Applications & Innovations (AIAI2007), pp.205-212, Athens, Greece, 2007.

[6] 梅村 祥之, 増山 繁, 仕事文推敲支援に向けた連体修飾部不足に対する受容性判定法. 自然言語処理, Vol.14, No.14, pp.43-65, July, 2007.

[7] 太田　貴久，増山繁，統計的手法に基づく講義音声書き起こし文書の文境界推定，第２回音声ドキュメント処理ワークショップ講演論文集、pp.137-142, 2008.

[8] 山本　悠二，増山　繁，係り先候補の相対的な距離を反映した統計的日本語係り受け解析，情報処理学会研究報告,Vol.2007, No.113，pp.15-22, 2007 年 11 月.

[9] 太田貴久，増山繁，学生レポート剽窃検出のための著者解析の試み，言語処理学会第 14 回ワークショップ「教育・学習を支援する言語処理」論文集，pp.43-46, 2008 年 3 月.

[10] 坂地泰紀，竹内康介，関根聡，増山繁，構文パターンを用いた因果関係の抽出，言語処理学会第 14 回年次大会発表論文集，pp.1144-1147, 2008.