セマンティックアーカイブ高度利用のため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

セマンティックアーカイブ高度利用のために重要な，テキストマイニングについて，２００８年４月以降の主な成果を報告する．セマンティックアーカイブの，より高度な利用のためには，bag of words レベルだけではなく，構文レベルの情報を活用し，さらに，意味的情報にも踏み込んで行く必要がある．その基礎として，コーパスから事象とその原因・根拠表現を抽出する汎用手法を行ってきているが，今回は，第 2 章で特許マイニングへの応用を紹介する。第 3 章では，文中の表現間に因果関係が存在するか否かの判定に関する研究を紹介する。これらの解析を行うための前処理として必要な係り受け解析に関する分析結果を第 4 章で紹介する．意味処理に必要なオントロジーの構築に関して，第 5 章で紹介する．関連した研究として，第 6 章から第 8 章で，企業の公式 Web サイトからの基本情報抽出，Web 掲示板を対象とした質問・回答対応の自動抽出手法の提案，青空文庫を対象とした書き手の識別とその応用について，それぞれ紹介する．

2 特許明細書からの出願目的・技術課題情報の抽出 [7]

特許明細書から出願目的，技術課題情報を自動的に抽出するための手法を訂案した．本手法によって抽出された技術課題情報は，パテントマップ自動生成や，技術トレンドの分析に有用である．本手法では，技術課題情報を抽出するために有効な手がかり表現を使用して抽出するが，この手がかり表現を自動的に獲得することで高い精度 (77.1%) だけでなく，高い (76.3%) をも達成することができた．

3 企業業績要因文における因果関係の有無判定手法の提案 [12]

現在，Web ページや新聞記事を含む大規模な機械可読文書が入手可能となっている．多くの機械可読な文書中には，実アプリケーションに役立つ様々な情報がある。テキストマイニングは，そのような情報を獲得するのが目的である．本研究では，そのような情報のひとつとして，因果関係に関する知識を取り上げ，手がかり表現の前後の表現の間に因果関係があるか否かを判定する手法を提案し評価実験を行った．その結果，精度 92.4%，再現率 84.5%を達成した．

4 日本語係り受け解析における誤りの類型化と文構造の曖昧性について [14]

統計的日本語係り受け解析の誤りの原因を網羅的に検証するために，係り受けの誤りの類型化，ならびに，類型の自動分類のルールを作成した．そして，実際の係り受け解析器で網羅率，ならびに，累計ごとの誤りの分布を調べた．さらに，それぞれの類型について人手での評価を行い，実際に係り先が一意に決定できるか否かについて検証を行った．

5 日本語彙大系と日本語 Wikipedia からの汎用オントロジーの自動構築手法 [10]

日本語語彙大系の意味体系の分類方針に沿うように，日本語 Wikipedia の記事とその所属するカテゴリを選定し，日本語語彙大系の意味分類とそれらの情報を統合する．これにより，日本語語彙大系の意味分類に従って明確に分類された階層構造を持つ，大規模な日本語汎用オントロジーを自動構築する手法を提案した．

6 企業の公式 Web サイトからの基本情報抽出 [9]

企業の公式サイトから会社概要の抽出を行う手法を提案した．これらの情報は，そのデータの意味を機械が理解できるフォーマットで記述されていないことが多い．そのため，データマイニング等の再利用のためには，記述に対して意味のアノテーションを行う必要がある．しかしながら，東京証券取引所に上場する企業だけで 2,372 社という大量に存在する企業に対して人手でアノテーションを行うことは困難である．そこで，このような企業が提供する情報をアルゴリズムを用いて抽出し，再利用可能な形にすることが必要となる．企業の公式サイトから会社概要の情報を属性，属性対の形で抽出手法を考案し，その有効性を実験で確かめた．

7 Web 掲示板を対象とした質問・回答対応の自動抽出手法の提案 [11]

Web 掲示板を対象とした質問回答対応の自動抽出手法を提案する．記事の類似度とスレッドツリーでの距離を考慮した重みを用いた質問と回答の対応付け方法を考案し，実験した所，F 値 0.78 を達成した．

8 青空文庫を対象とした書き手の識別とその応用 [13]

書き手が文章を生成するしくみとして言語学で用いられる Harmonic Grammar に基づき書き手の特徴をとらえる手法を提案した．特徴をとらえるために助詞の出現パターン，読点の打ち方，品詞の出現パターンに着目した．青空文庫を対象に書き手の識別実験を行った結果，比較的高い正解率で書き手の識別を行うことに成功した．

9 まとめ

本稿では，前年度の報告に引き続き，２００８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．なお，これ以外にも企業の業績発表記事からの業績要因抽出と極性付与 [5]，景気動向を示す根拠表現の抽出と分析 [2]，未知のサイトの Web ページからの主要部分抽出手法 [4] があるが，それらの結果は，既に 2007 年度の報告書で，in press として紹介しているので，省略する．また，構文パターンを用いた因果関係の抽出 [3] についても，改良は加えたものの，初期の結果を 2007 年度の報告書で報告済なので省略する．

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, pp.307-320, Cambridge, England.

[2] Hiroki Sakaji, Satoshi Sekine, Shigeru Masuyama, Extracting Causal Knowledge Using Clue Phrases and Syntactic Patterns, Proc. PAKM2008, LNAI5345, pp.111-122, 2008.

[3] 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, 2008.

[4] 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.

[5] 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

[6] 野中尋史，酒井浩之，　増山繁, テキストマイニングを用いた判例文書の分類および情報抽出, 情報ネットワークローレビュー，vol.8，pp.74-85, 2009.

[7] 酒井浩之，　野中尋史，増山繁, 特許明細書からの出願目的・技術課題情報の抽出，人工知能学会第 23 回全国大会論文集，1C3-4, 2009.

[8] 藤村真太郎，酒井浩之，増山繁，企業業績要因文の経常的か否かに基づく分類とイベントスタディ法に基づく分析，人工知能学会第 23 回全国大会論文集，3G3-OS12-2, 2009.

[9] 鶴田雅信，関根聡，増山繁，企業の公式 Web サイトからの基本情報抽出，人工知能学会第 23 回全国大会論文集，3B4-2, 2009.

[10] 小林暁雄，増山繁，関根聡，日本語彙大系と日本語 Wikipedia からの汎用オントロジーの自動構築手法，情報学ワークショップ 2008 論文集, pp.63-66, 2008.

[11] 鈴木佑輔, 酒井浩之，増山 繁，Web 掲示板を対象とした質問・回答対応の自動抽出手法の提案，言語処理学会第 15 回年次大会発表論文集, pp.44-47, 2009.

[12] 坂本大祐, 坂地泰紀, 酒井浩之，増山 繁，企業業績要因文における因果関係の有無判定手法の提案，言語処理学会第 15 回年次大会発表論文集, pp.925-928, 2009.

[13] 太田貴久，増山繁，青空文庫を対象とした書き手の識別とその応用，言語処理学会第 15 回年次大会発表論文集, pp.679-680, 2009.

[14] 山本悠二，増山繁，日本語係り受け解析における誤りの類型化と文構造の曖昧性について，言語処理学会第 15 回年次大会発表論文集, pp.789-792, 2009.

[15] 野中尋史，酒井浩之，　増山繁, テキストマイニングを用いた判例文書の分類および情報抽出, 情報ネットワーク法学会第 8 回研究大会予稿集, pp.39-44, 2008.