D-006

新聞記事からの行動履歴情報の抽出実験と評価
Extraction of One Person's Behavior on Various Matters from Newspaper Articles and its Evaluation

南雲 旭人 †　山田 剛一 †　絹川 博之 †
Akito Nagumo　Koichi Yamada　Hiroshi Kinukawa

1. はじめに

近年，外交問題や政権交代などの大きな政治的問題の発生により，国民の政治への関心が高まっている．国民が政治の情報を得るための手段としては，TV ニュースや新聞，Web 上の各種サイトが挙げられる．TV ニュースや新聞は毎日閲覧しなければ情報を逃してしまうことがあるが，大手の新聞社が公開している Web ニュースサイトは，検索機能により特定の政治家の情報を過去のものを含め得ることができる．しかし，多くの情報を得ようとすると必然的に閲覧の手間が増えてしまう．

本研究は，新聞社が公開しているニュースサイトの政治カテゴリ記事を対象とした情報抽出システム開発を行う．記事テキストを形態素解析，構文解析し，得られた品詞や構文情報を手がかりに政治家の発言や行動情報の記述部分を推定し抽出を行う．抽出した記述はデータベースに格納し，ユーザ側のインタフェースから入力された政治家名・政党名や話題に関して時系列順などに整理して提示する．抽出と提示処理により，政治記事内容をマクロにとらえることをシステムの最終目標とする．

2. 新聞記事からの抽出対象

本研究では，政治記事内容をマクロにとらえるという最終目標を実現するために，以下の情報を抽出対象とする．

2.1 発言の記述

ニュースサイトの政治記事においては，政治家が記者会見などで述べた内容を端的に表す部分が引用されているため，その政治家の考えや方針を表わす記述として抽出する．その際，発言の記述だけでなく，発言に関する日付や発言場所，その政治家の所属政党や役職といった付属情報もあわせて抽出する．

2.2 行動の記述

行動とは，一般に活動や行い全般を示す語であるが，本研究では，政治家の方針などに関わる離党や視察といった発言以外の政治活動としての行動を抽出対象とする．行動の記述も発言と同様の付属情報が存在するため，抽出を行う．

2.3 話題の記述

本研究では，発言の焦点を表す語や句のことを話題とする．政治記事においては，TPP や憲法改正などの特定の語と，「TPP 交渉への参加」のようなそれらの語を含む句のことである．

† 東京電機大学大学院　未来科学研究科，
Graduate School of Science and Technology for Future Life, Tokyo Denki University

3. システム概要

記事収集から抽出データの提示までを行う，図 1 のシステムを提案する．

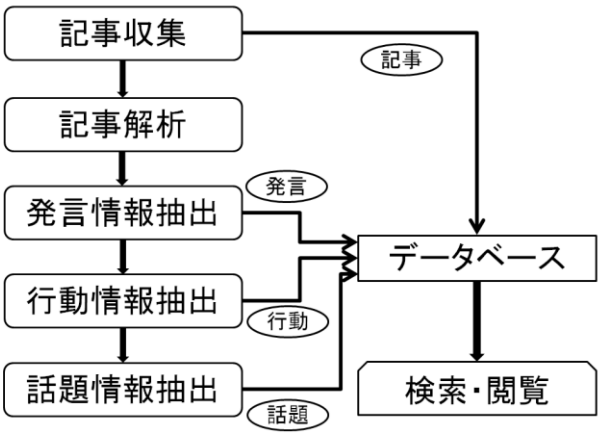

図 1　システム構成

3.1 記事収集

Webstemmer[1]を利用し，対象となる新聞社の Web ニュースサイトの政治カテゴリ記事からタイトルと本文を収集する．収集した記事データはデータベースに格納するとともに，記事テキストの解析を行う処理部へと送る．

3.2 記事解析

発言とその付属情報や行動情報の記述部分を特定するための手がかりとして，記事中の単語の品詞を取得する．形態素解析エンジンには MeCab[2]を利用する．

3.3 発言情報抽出

記事を解析して得られた品詞情報などを手掛かりとして利用し，発言と付属情報の抽出を行う．抽出の際は，発言と付属情報を表す部分の手がかりを登録したテンプレートを用いて，記事のテキストにおける対応する部分を特定する形で行う．発言は鉤括弧付きのものとそうでないものが存在するが，その両方を取得する．付属情報は発言者，発言者の所属，発言者の役職，日付，場所，話題の 6 種類を取得する．付属情報は発見されたもののみを抽出するが，発言者と日付は検索や閲覧の段階で必要な情報であるため，補完処理を行う．発言者の見つからなかった文があった場合，発言者が省略されたと考え，その直前に発見された発言者と同一と判断する．日付が見つからなかった場合は，記事の投稿日時を抽出する．

3.4 行動情報抽出

行動の記述には，離党の場合は「離党届」や「除籍」など，視察の場合は「視察先」や「訪問」といったそれぞれの行動の種類とあわせて表れやすい表現が存在する．本システムで扱う行動は政治活動に限定されているため，離党や視察といった政治活動に関する手がかりをあらかじめテンプレートに登録し，記事テキストから一致するものを行動情報として抽出する．また，あらかじめ登録されていないものについては，動詞や出現位置といった行動の種類に関わらない表現を手がかりとして利用し抽出を行う．行動情報抽出の際も発言と同様に付属情報を抽出する．処理は図２の流れで行う．

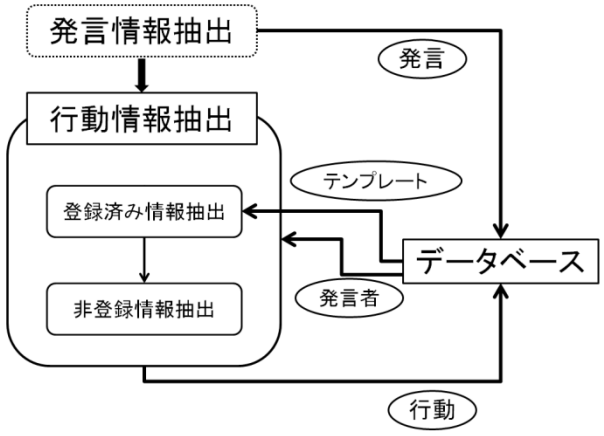

図２　行動情報抽出

3.5 話題情報抽出

話題の抽出は，記事テキストの単語それぞれに重要度をつけ，ある一定以上の重要度である単語を話題語であると判断し，抽出する．重要度は，その単語と発言の記述部分の間にいくつの単語が含まれるかという距離や，記事のタイトルに含まれる単語である等の条件から決定する．抽出の際はその単語のみと，その単語を含んだ名詞句それぞれを取得する．単語のみの話題語データを利用することで，その話題に一致する発言のみを抽出して閲覧することができる．また，単語を含んだ名詞句を利用することで，特定の単語以外の検索ワードに対応することができる．図３の例では，語だけの抽出では２０１２年度補正予算案という限定された話題になってしまうが，句を含むことで参院や審議といった範囲に話題を拡大することができる．

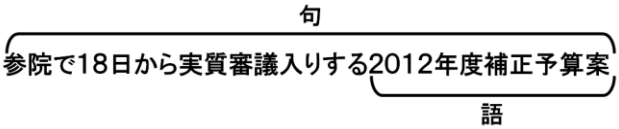

図３　話題の定義例（語と句）

発言と付属情報，行動情報，話題情報の抽出例を図４と表１に示す．

公明党の山口代表は２日、東京都内で街頭演説し、今夏の参院選について、「与党で過半数を得ることが安定した政治への第一歩だ。（政治の）停滞を招く対立が、参院を中心に行われることがあってはならない」と述べ、自民党と公明党を合わせて過半数（非改選議員含む）を目指す考えを強調した。
（2013 年 1 月 3 日 21 時 45 分　読売新聞）

図４　記事テキスト例

表１　発言と付属情報の抽出例

発言者	山口
所属	公明党
役職	代表
日付	2013/1/3
場所	東京都内
発言	与党で過半数を得ることが安定した政治への第一歩だ。（政治の）停滞を招く対立が、参院を中心に行われることがあってはならない
行動	街頭演説
話題	参院選（語），今夏の参院選（句）

3.6 検索・閲覧

ユーザ側のインタフェースとして，政治家名や政党名，話題といった単語から検索できる機能を用意する．入力された検索ワードに関するデータをデータベースから取得し，整理して出力する．政治家名であればその人物の発言を時系列順に並べて出力し，話題であれば特定の人物に限らずその話題に関して発言されているものをまとめて出力する．

4. おわりに

本稿では，Web ニュースサイトの政治カテゴリ記事を対象とした，政治記事内容をマクロにとらえるための情報抽出システムを提案した．今後は各情報抽出部の評価実験と精度向上，検索と閲覧部分の実装を進めていく．

謝辞

本研究に使用させていただいた Webstemmer と MeCab の開発者様に感謝致します．

参考文献

[1] Webstemmer:http://www.unixuser.org/~euske/python/webstemmer/index-j.html
[2] MeCab:http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html