2Q-4

新聞記事からの発言・行動履歴情報の抽出

南雲旭人† 山田剛一†† 絹川博之†††
東京電機大学大学院 未来科学研究科‡

1 はじめに

近年，社会情勢への不安や政権交代などをうけ，国民の政治への関心が高まっている．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 上の各種サイトなどが挙げられる．TV ニュースや新聞に対し，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得ることなどができる．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．

新聞記事からの情報抽出に関する研究には，パターンマッチング処理をベースとした人物・企業情報の抽出[1]や，定型表現を利用した下位概念単語の自動出力[2]などが行われている．本研究では，新聞社が公開しているニュースサイトを対象に，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の開発を目指す．記事中における重要な政治家・政党関係者の発言や意見表明，その焦点となる話題などを抽出する．ユーザは政治家名・政党名や話題を入力し，システムは入力された政治家の発言や行動履歴，または話題に関しての政治家の様々な発言や意見を時系列順に整理して提示する．

2 新聞記事における発言と行動の記述

新聞記事では，政治家の発言や行動，政党の方針等が簡潔に記述される．本章では，記事における発言と行動の記述について述べる．

2.1 発言の記述

ニュースサイトの記事においては，政治家が記者会見や会談，演説等で述べた内容のうち特に考えや方針を端的に表した部分が発言として抜粋され，鉤括弧付きで記述されることが多い．発言が記述される一文には主語となる発言者の人名が記述される他に，日付や政治家の所属政党・役職，発言をした場所，発言の対象となる話題といった付属情報が記述される場合がある．発言単体ではその発言に至った背景や対象がわからないため，抽出した発言を整理して閲覧するためにはこれらの付属情報を抽出することが重要である．

2.2 行動の記述

記事中に記述される行動情報には，一般的な動詞で表される行動のほか，離党や視察，などの政治活動の一部としての行動が挙げられる．本研究では，ある政治家の今後の活動方針や事象への態度も行動の記述として扱う．

3 発言の抽出手法

今回は，ニュースサイトから記事を収集し，発言とその付属情報を抽出するシステムを構築した．構成を図１に示す．

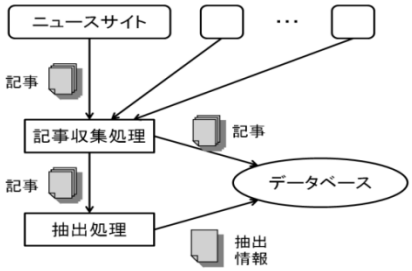

図１．システム構成

記事収集部には，Webstemmer[3]を利用した．記事収集の対象となるニュースサイトとして YOMIURI ONLINE と朝日新聞デジタルを登録し，収集した記事をタイトル・本文・投稿日時に分けデータベースに格納するとともに抽出部に送る．抽出部は記事に次項の処理を施し，抽出データを記事と対応した形でデータベースに格納する．

3.1 発言・付属情報抽出

発言と付属情報の抽出は，記事のテキストを形態素解析し，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．形態素解析エンジンには MeCab[4]を利用した．

はじめに，発言部分の抽出を行い，そこで特定した発言部分を中心に付属情報を抽出する．記事中の発言は鉤括弧付きで記述される他，「述べる」「訴える」といった動詞や「指摘する」等の，発言の後に記述されることの多い手がかり表現が直後に出現するため，これらの表現に当てはまった場合に発言部分と特定し，鉤括弧内を発言として抽出する．

鉤括弧が存在しない発言の記述の後には，「とする」などの発言部分以外でも多く出現する汎用的な語句が記述されることがある．そのため，鉤括弧が存在しない発言を抽出する際は，鉤括弧が存在するという条件のかわりにこれらの語句も手がかり表現に追加している．これにより発言と関係のない部分も抽出してしまう可能性が高まるので，後述する付属条件のうち半分以上が検出された場合のみ，鉤括弧が存在しない場合も発言であると判定し，付属情報と手がかり表現を除いた部分を抽出する．

発言部分を特定後，その発言に関する付属情報を抽出する．付属情報の分類は発言者，発言者の所属，発言者の役職，日付，場所，話題の６つである．このうち，発言者と日付は後の閲覧に当たり必要不可欠な情報であるので，抽出されなかった場合は補完処理を行う．話題に関しては抽出の手法が異なるため，次節で述べる．

発言者は発言が記述された一文中の冒頭に主語として出現するため，その条件に一致した上で人名であるものを抽出する．発言部分の周辺に人名が発見されなかった場合は，発言者が一度記述されたために省略されたと考

Extraction of One Person's Speaking and Behavior on Various Matters from Newspaper Articles
†Akito Nagumo , ††Koichi Yamada , †††Hiroshi Kinukawa
‡Graduate School of Science and Technology for Future Life, Tokyo Denki University

え，直前に抽出した発言者と同一人物の発言であると推定する．発言者の記述にはそれに接続する形で所属や役職が出現する場合があり，所属している政党や幹事長などの役職を表す部分をそれぞれ抽出する．

記事中において日付は年と月が省略され，発言者の記述の後に「～日」とだけ記述されることがほとんどである．省略された年と月は記事の投稿年月日と同じであると考え，収集処理で取得した投稿日時データと合わせることで日付データとする．場所に関しては，発言者もしくは日付の後に「～で」といった形で現れるため，これを抽出する．

これらの条件を基に図２のテキスト例を処理すると，表１の情報を抽出する．

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２．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

3.2 話題抽出

本研究での話題とは，発言の焦点を表す語や句のことをいう．例えば，政治記事においては，TPP や憲法改正などと，「TPP 交渉への参加」のようなそれらの協議や方針を表す語や句のことである．

話題の抽出は，発言と付属情報抽出とは別の手法を用いる．初めに，記事中の単語に対して，発言の記述との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重要度をつける．その後，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，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，TPP などの特定の単語以外の検索ワードに対応することができる．

４　評価

発言と付属情報の抽出手法に関しての評価を行った．今回の評価では話題を除いた５つの付属情報で行っている．人手で記事から発言と付属情報部分を特定したデータと，システムの出力を比較し，結果を以下の５つに分類した．

(1) 正しく抽出
(2) 不要な修飾語句などを含んで抽出
(3) 必要な情報が欠けて抽出
(4) 異なる情報を抽出
(5) 未抽出

５つのうち，正解に数えるものは，(1)と(2)の２つとした．(2)のデータから修飾語句などを除去する手法は検討中である．はじめに発言を抽出し，次に抽出した発言の付属情報を探索する手法であるので，付属情報は抽出した発言に関するもののみを評価した．評価には，YOMIURI ONLINE と朝日新聞デジタルの計 400 件の政治カテゴリ記事を利用した．

表２に評価結果を示す．表２における精度と再現率，F 値の定義は以下の通りである．

$$\text{精度} = \frac{(1)+(2)}{(1)+(2)+(3)+(4)}$$

$$\text{再現率} = \frac{(1)+(2)}{(1)+(2)+(5)}$$

$$\text{F 値} = \frac{2\times \textit{精度}\times \textit{再現率}}{\textit{精度}+\textit{再現率}}$$

表２．評価

	精度	再現率	F 値
発言	0.907	0.737	0.813
発言者	0.933	0.750	0.832
所属	0.841	0.720	0.776
役職	0.802	0.866	0.833
日付	0.888	0.814	0.850
場所	0.866	0.949	0.906

全体的に再現率に比べ精度が高い結果となり，特に発言と発言者の精度が高くなっている．発言と発言者を抽出するための厳密な条件を設定したためと考えられる．役職と場所は修飾語句なども抽出してしまうことが多く，修飾語句を含む抽出を不正解とすると精度が低くなってしまうため，今後はこの２つに関して不要な情報を除去する方法を検討する．また，今回行わなかった話題抽出に関する手法の研究と評価，行動情報の抽出も進める．

５　おわりに

政治記事内容をマクロにとらえるシステムを開発するにあたっての基幹部分として，発言と付属情報の抽出プログラムを構築し，抽出性能の評価を行った．全体として精度は高めの数値を示したが，再現率が低くなったことと，抽出の際に不要な情報を含めて取得してしまうことがあった．今後は未抽出や誤抽出の改良と話題・行動情報の抽出手法の研究を進める．

謝辞

本研究に使用させていただいた Webstemmer と MeCab の開発者様，読売新聞社様と朝日新聞社様に感謝致します．

参考文献

[1]　西野，落谷，″新聞記事からの人物・企業情報の抽出，″ 情報処理学会研究報告．自然言語処理研究会報告，vol．98，pp．125-132 ,1998.
[2]　安藤，関根，石崎，″定型表現を利用した新聞記事からの下位概念単語の自動抽出(オントロジ・抽出(2))(セマンティックウェブと自然言語処理その他一般)，″ 情報処理学会研究報告．情報学基礎研究会報告，vol．2003，pp．77-82 ，2003.
[3]Webstemmer:http://www.unixuser.org/~euske/python/webstemmer/index-j.html
[4]MeCab:http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html