新聞記事からの法案の賛否に関する意見抽出
Extracting Politician Opinions about Legislation from Newspaper Articles

西村 昌和[*1*2] 酒井 隆行[*3] 村上 晴美[*1]
Masakazu NISHIMURA Takayuki SAKAI Harumi MURAKAMI

[*1] 大阪市立大学大学院創造都市研究科 [*2] 株式会社 西村屋 [*3] 大和総研ビジネス・イノベーション
Graduate School for Creative Cities, Osaka City University Nishimura-ya Inc. Daiwa Institute of Research Business Innovation

We propose a method that extracts and organizes politician opinions about legislation from newspaper articles. Our method is based on direct quotations and a p/n dictionary. We present a prototype system that displays a list of politician names and opinions (positive or negative) about selected legislation.

1. はじめに

情報爆発時代において，必要な情報だけを収集したいニーズは高まり続けている．あらゆる情報の中で，新聞記事は信憑性が高く，情報量も多く，良質な情報源である．

本研究は，新聞記事から人物の意見を抽出して整理することを目的とする．題材として政治家の法案に対する意見に着目する．政治家は議員の名前や所属が公開されているため特定しやすく，また法案に対する意見は賛成か反対に分類できる．さらに新聞記事には政治家の意見が多く掲載されている．

本稿では，新聞記事から指定したいくつかの法案(トピック)に関する政治家の意見(賛否)を抽出する手法を提案する．これらの情報は選挙時に有権者の意思決定に役立てられる．以下，2 節で提案手法，3 節で実験について述べ，4 節では試作したプロトタイプシステムを紹介する．

2. 提案手法

2.1 概要

本研究では，新聞記事から法案に関する政治家の意見(賛否)を抽出する手法を提案する．図 1 に提案手法の概要を示す．

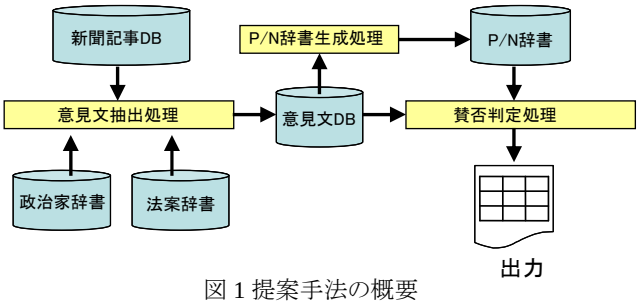

図 1 提案手法の概要

本研究の主要なアイディアは，直接引用表現と P/N 辞書の利用である．政治家名と法案が出現する文に，政治家が主語となり直接引用表現(鈎括弧で括られる部分)が含まれる場合，その直接引用表現内のテキストに，政治家の意見を判定するために必要な情報が含まれると考える．本研究では，このテキストを意見文と呼ぶ．次に意見文に含まれる動詞に着目し，P/N 辞書を用いて意見(賛成・判定・不明)判定を行う．政治家が法案に対して賛成であればポジティブな動詞，反対であればネガティブな動詞が多く含まれると考える．

連絡先：西村昌和,大阪市立大学大学院創造都市研究科,
m08uc514@ex.media.osaka-cu.ac.jp

提案手法は，意見文抽出処理と賛否判定処理に大別される．意見文抽出処理では，政治家辞書と法案辞書を用いて新聞記事 DB から政治家の意見文を抽出する．賛否判定処理では，P/N 辞書を用いて政治家の法案に対する意見を判定する．

2.2 意見文抽出処理

政治家辞書と法案辞書を用いて，新聞記事 DB から政治家の意見文を抽出する．

(1) 新聞記事 DB

CD-ROM で提供される新聞記事などから，見出し・本文・キーワードをデータベースに格納して新聞記事 DB とする．

(2) 政治家辞書の作成

2008 年 10 月 31 日時点での衆議院・参議院の政治家 618 名の姓名と所属政党を，構想日本が運営する WEB サイト「政治家・政策データベース」から抽出した．この際，新聞記事中の政治家名はフルネームとは限らないため，姓と名に分ける．

(3) 法案辞書の作成

新聞記事データにはキーワードが与えられている．政治家名を含む新聞記事に与えられたキーワードで，語尾が法案・法・案のいずれかになるものを抽出し頻度順に並べる．頻度の高いものから目視により法案を選定する．新聞記事中の法案名の表記の揺れを吸収するため，語彙を与えて，法案辞書とする．たとえば郵政民営化法案の場合は「郵政民営化」が含まれる．新聞記事から意見文を抽出する際，「郵政民営化」が出現すれば，「郵政民営化法案」として扱う．

(4) 意見文の抽出方法

井上ら[井上 08]は文章の段落単位で意見抽出を行っているが，本研究では政治家の意見が直接書かれていると考えられる直接引用表現を意見抽出の対象とする．

まず政治家名がフルネームで含まれかつ法案名が含まれる記事文を抽出する．

その中で，政治家の姓名もしくは姓を含み，それに助詞「は」または「が」が掛かっている文に絞り込む．フルネームが出現している文中には，続けて「小泉氏」のように姓のみで表現される場合があるが，この場合は本人であると判断をする．

さらに同一文中に直接引用表現が含まれていれば，鈎括弧内の文を意見文として抽出する．同一政治家，同一法案に対し複数の意見文が存在するが，全ての意見文を抽出し，それら全てを計算の対象とする．

2.3 賛否判定処理

賛否判定処理は，政治家の意見文に含まれる動詞をもとにP/N 辞書を参照し，賛否に関する意見（賛成，反対，不明）を判定する．

藤村ら[藤村 05]は，文単位で評判の肯定もしくは否定の分類を行うことにより，評判抽出を肯定・否定の評判，ノイズといった 3 値分類問題への置き換えを検討している．本研究では上記手法を参考にした処理を提案する．

(1) P/N 辞書の作成

P/N 辞書生成処理により，あらかじめ P/N 辞書を作成しておく．本研究では機械学習により辞書作成を行う．被験者に意見文を含む新聞記事を提示し，政治家の意見（賛成・反対・不明）を判定させたデータを訓練データとする．ここでは「CD-毎日新聞 2005（後述）」に出現する法案の中で意見文数の最も多い「郵政民営化法案」を学習に用いた．辞書作成の詳細は以下に示す．

1. 郵政民営化の意見文中に出現する動詞を抜き出す．
2. 動詞群を以下の優先順位でソートする．(a)～(d)をそれぞれ第 1～第 4 優先キーとする．(d)で漢字コードを用いている理由は実験の再現性を保つためである．
   (a) 動詞が出現する意見文の評価が常に賛成か反対のいずれかである
   (b) 意見文の中での出現頻度の高さ
   (c) 動詞が出現する意見文の評価のうち，賛成と反対のそれぞれの出現頻度の差の大きさ
   (d) 単語の漢字コード（日本語 EUC）順
3. 上位の n 個を P/N 辞書とする

n は経験上 170 とした．なお，「する」という動詞は直前の名詞と組み合わせることで政治家の意見を反映した語になるため，名詞+「する」は一つの動詞として扱う．

(2) 賛否判定方法

賛否判定は意見文中の動詞に対して P/N 辞書によるマッチングを行い，スコアを算出する．政治家と法案の組に与えられた意見文の中に現れる全ての動詞に対してスコアリングを行う．このとき動詞直後の「ない」や「ず」などの否定語はスコアを正負逆転させて処理をする．スコアの式は

$$score(o) = E_P(o) - E_N(o) \qquad (1)$$

とする．score(o)が正であれば賛成，負であれば反対，0 であれば不明と判定する．E_p は肯定評価表現の総和，E_N は否定評価表現の総和，o は意見文とする．

ここで 2.2 節の(4)において抽出した文を 2 件例としてあげる．

1. 小泉純一郎首相は冒頭，「抵抗，反対を恐れず，既得権にひるまず，過去の慣例にとらわれず，郵政民営化（関連法案）を今国会で成立させたい」と述べ
2. 小泉純一郎首相は27日の衆院予算委員会で，郵政民営化法案について「今国会で成立させるよう努力するのが私の立場であり，成立しないとは全く考えていない．必ず成立させる」と明言した．

直接引用表現内の動詞（下線部）を原型に変換する．P/N 辞書において「成立する」という動詞は肯定表現であるため，この単語が現れた場合は E_p に 1 ポイント加算する．「恐れる」は否定表現であるが，「恐れず」のように否定助動詞が伴う場合は，肯定表現として扱う．よって E_p に 1 ポイント加算する．

「成立しないとは全く考えていない」のような二重否定文は現段階で対応しておらず，今後の課題である．

3. 実験

「CD-毎日新聞 2005 データ集本社版」を用いた．18 件の法案を選定した．3 名の被験者が政治家名と法案名とそれらを含む新聞記事を与えられ，政治家の法案に対する賛否を判定した．被験者の中で判定が異なる場合は，被験者以外の人物が統一を行った．

651 件の意見文が抽出された．郵政民営化法案を除く 17 件の法案のうち 1 名以上の政治家の意見が抽出できた法案は 12 件であった．

「郵政民営化法案」から作成した P/N 辞書を用いて上記 17 件で実験した判定結果の精度は 73%，再現率は 43%であった．精度と再現率の計算式は以下の通りである．

$$\text{精度} = \frac{\text{出力された正解の数}}{\text{本手法による出力結果の数}} \qquad (2)$$

$$\text{再現率} = \frac{\text{出力された正解の数}}{\text{正解組の数}} \qquad (3)$$

ただし正解組とは政治家名と法案の組み合わせのうち被験者による正解が与えられている組の数である．

4. プロトタイプシステム

Web ベースのプロトタイプシステムを構築した（図 2）．ユーザが法案名を選択すると，選択された法案に対して意見文を持つ政治家の一覧が表示され，賛成・反対・不明のいずれかが出力される．選択できる法案名はあらかじめ登録されているものであるが，システム管理者は任意に追加変更することができる．

選挙時の有権者の意思決定支援に役立てられると考える．

図 2 プロトタイプシステム

5. おわりに

本研究では新聞記事から政治家の意見情報を抽出・整理する手法を提案し，実験を行った．主に P/N 辞書の構築方法について述べた．精度と再現率を高めることを今後の課題とする．

参考文献

[藤村 05] 藤村滋，豊田正史，喜連川優: 文の構造を考慮した評判抽出手法，電子情報通信学会第 16 回データ工学ワークショップ(DEWS2005), 6C-i8 (2005).

[井上 08] 井上結衣，藤井敦: Web 世論からの意見抽出と賛否に基づく分類，言語処理学会第 14 回年次大会 ，C2-7 (2008).