新聞記事からの政治家の意見抽出

酒井　隆行†　　村上　晴美‡

大阪市立大学大学院創造都市研究科

1. はじめに

近年，多量の情報が発信されている中で，必要な情報を効率良く取得する技術に人々の関心が高まってきている．発信される情報源の中でも，新聞記事は客観的な事実の他に人物の意見情報も多数掲載されており有益な情報源となる．

本研究では，新聞記事からの意見情報の抽出と整理を目的とする．題材として政治家の意見情報の抽出を取り上げる．

政治家は，議員の名簿が公表されているため名前や所属等を特定しやすい．また，評価の際に人手での比較を行いやすい．さらに，新聞記事では，政治家の意見情報が多数記載されている．これらの意見情報を有権者が容易に抽出できるシステムがあれば，選挙時の意思決定の支援ができる．

本稿では，新聞記事から指定したトピック（キーワード）に関する政治家の意見（賛否とその根拠文）を抽出し，表形式で出力する手法について述べる．以下，2 節で提案手法，3 節で「CD-毎日新聞 2005」と「法案」関連の記事を用いた実験，4 節で本稿のまとめについて述べる．

2. 提案手法

2. 1　概要

本研究では，直接引用表現（鉤括弧で括られた表現）と動詞を主体とする評価表現に着目して，政治家の意見抽出を行う．

提案手法は，政治家データ抽出処理と，意見情報抽出・提示処理に大別される．政治家データ抽出処理では，1 年間の新聞記事 DB から，指定したキーワードに関する政治家の意見文を抽出して政治家記事意見文 DB を作成する．

その後，意見情報抽出・提示処理では，政治家記事意見文 DB から，キーワードに対する政治家の賛否に関する意見（賛成，反対，不明）とその根拠文を表形式で提示する．図 1 に提案手法の概要，図 2 に出力イメージを示す．

2. 2　政治家データ抽出処理

政治家データ抽出処理は，キーワードが見出しまたは本文に含まれる記事を新聞記事 DB から抽出する．記事抽出処理，政治家を含む記事のみを抽出する政治家判別処理，記事内の政治家の意見文を抽出する意見文判別処理の順に処理を行う．

2. 2. 1　政治家判別処理

政治家判別処理では，政治家辞書を用いて政治家を含む記事のみを抽出する．政治家辞書として，衆議院，参議院のサイト内にある．2005 年度中，「郵政解散」による衆院選後に在籍した衆議院，参議院の議員全員（722 名）を登録している．

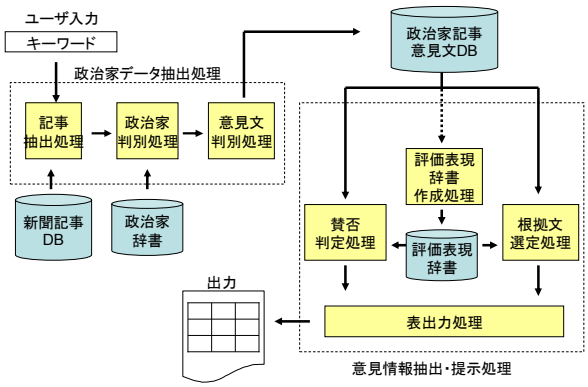

図1　提案手法の概要

議員名	意見文(文)	意見	根拠文
小泉純一郎	134	賛成	小泉純一郎首相は5日、首相官邸で自民党の武部勤幹事長と会い、郵政民営化について「党内の抵抗は強いだろうが、(通常国会の)会期内に上げる」と述べ、今月21日召集予定の通常国会での関連法案成立に向け、党内調整を急ぐよう指示した。
竹中平蔵	34	賛成	竹中平蔵郵政民営化担当相は会見で「(法案骨格は)首相と官房長官が閣内をまとめ上げた。法案にして国会で審議していただき、成立させるのが私の務めだ」との考えを表明した。
岡田克也	20	反対	岡田代表も、首相が「改革の本丸」と位置づける郵政民営化問題を取り上げ、「(首相と自民党の)茶番」と批判、対決姿勢を強く打ち出している。
……	……	……	……

図2　出力イメージ「郵政民営化」の場合

2. 2. 2　意見文判別処理

意見文判別処理は，政治家判別処理後の記事群から政治家の意見文を抽出する．本処理では，以下の考えに基づき，直接引用表現のある文に着目した．新聞記事では，人物名と直接引用表現のある文は，その人物の重要な意見である可能性が高い[1]．さらに，その文に「郵政民営化」等のキーワードが含まれると，キーワードに関する重要な意見文である場合が多くなる．

そこで，意見文判別処理では，「政治家が主語であり，かつ，直接引用表現と入力したキーワードが含まれる文」を政治家の意見文として取得する．政治家が主語であるかどうかの判定は，政治家判別処理後の記事集合に Chasen を用いて形態素解析を行い，名詞と係助詞の連続した関係を用いる．最後に，政治家ごとに記事と意見文のまとまりを政治家記事意見文ＤＢとして出力する．

2. 3　意見情報抽出・提示処理

意見情報抽出・提示処理は，政治家記事意見文 DB

Opinion extraction of politicians from Japanese newspaper articles

†Takayuki Sakai・Graduate School for Creative Cities, Osaka City University

‡Harumi Murakami・Graduate School for Creative Cities, Osaka City University

から，指定したキーワードに対する政治家の賛否とその賛否の意見文である根拠文を表形式で出力する処理群の総称である．政治家の賛否判定を賛否判定処理，根拠文の選定を根拠文選定処理で行う．表出力処理では，賛否判定処理と根拠文選定処理で出力された結果を統合出力する．

2．3．1　賛否判定処理

賛否判定処理は，政治家記事意見文ＤＢ内の意見文をもとに評価表現辞書を参照し，賛否に関する意見(賛成,反対,不明)を判定する処理である．

藤村ら[2]は，文単位で評判の肯定・否定分類を行うことによって評判抽出を肯定・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．本研究における評価表現辞書は,「動詞（基本形）と賛否ラベル（賛成，反対，不明)」の組から構成される．動詞は「(法案成立を）目指す」「(法案内容を）問う」等，政治家の意見を反映している語が多い．また,「する」という動詞は，名詞との組み合わせによって政治家の意見を直接反映する語となる．例えば，法案Ａに対して，政治家Ｂが法案に賛成である場合,「賛成する」という表現で記述される．そこで，本研究では，名詞と動詞「する」の組み合わせ表現も「動詞」とみなして評価表現とする．

賛否判定処理は，入力キーワードで抽出できた政治家記事意見文ＤＢ内の全ての意見文を入力データとする．入力データに形態素解析をかけ，後述する評価表現辞書作成処理と同様の処理方法で動詞を政治家ごとに取得する．評価表現辞書を参照し，取得した全ての動詞でスコアリングを行って政治家の意見の賛否を判定する．スコアリング $score(o)$ の式は，

$$score(o)=\frac{E_P(o)-E_N(o)}{E_P(o)+E_N(o)+k} \qquad (1)$$
$$(-1\le score(o)\le 1)$$

とする．$score(o)$ が 0 より上なら賛成，0 未満なら反対，0 なら不明とする．なお，E_P は辞書内の肯定評価表現の総和，E_N は辞書内の否定評価表現の総和，k は経験的に 0．01，o は意見文とする．

2．3．2　評価表現辞書作成処理

政治家の賛否を判定するための評価表現辞書は，事前に評価表現辞書作成処理を用いて作成する．あるキーワードを入力して政治家データ抽出処理を用いて作成した政治家記事意見文ＤＢ内の全ての意見文を入力データとして形態素解析を行い，動詞に着目して抽出する．2．3．1 で述べたように,「動詞」と，名詞＋「する」の組み合わせを抽出し，基本形に統一する．次に，抽出された動詞に人手で賛成，反対，不明のラベル付けを行う．

例えば,「目指す」には賛成,「拒む」には反対を付与する．また,「出演する」や「述べる」等，政治家の意見を反映しないと思われる動詞は削除する．

この方法では，特定のキーワードごとに，出力された動詞への人手による賛成，反対，不明のラベル付けが必要となる．ラベル付け作業は非常に手間が掛かるので，評価表現辞書作成処理の省力化が望まれる．

本研究では，政治家の意見の中でも「法案」に着目する.「法案」というキーワードを用いて，評価表現辞書作成処理を行い，評価表現辞書を作成する．これにより,「法案」関連のトピックであれば，キーワード「法案」で作成された評価表現辞書を用いることで，キーワード入力から表出力までの処理を自動化できると考える．

2．3．3　根拠文選定処理

根拠文選定処理は，政治家記事意見文ＤＢ内の意見文の中から，政治家の意見を最も表している文を根拠文として選定する処理である．以下の過程により，意見文の中から根拠文を１つ選定する．

まず，評価表現の出現頻度(賛否判定処理で計算)が最も多い意見文を根拠文として選定する．次に，評価表現の出現頻度が等しい文が複数ある場合，意見文を含む記事の重要度を tf-idf 法によって計算し，スコアの最も高い記事に含まれる意見文を選定する．なお，計算はキーワードの記事中への出現回数を tf，政治家データ抽出処理後の全記事数 N に対するキーワードの出現回数を df で表す．tf-idf 法によるキーワードの重み w は，

$$w=tf\cdot idf=tf\cdot\left(1+\log\left(\frac{N}{df}\right)\right) \qquad (2)$$

とする．まだ意見文が複数ある場合，日付が最も新しい記事に含まれる意見文を根拠文として選定する．

3．実験

現在，新聞記事として「CD-毎日新聞 2005」,「法案」に関連する例題として「郵政民営化」,「国民投票法案」,「憲法改正案」,「人権擁護法案」で実験中である．抽出の際に用いるキーワードは，それぞれ「郵政民営化」,「国民投票」,「憲法改正」,「人権擁護」，である．

提案手法によって抽出できた政治家記事意見文ＤＢ内の政治家数，記事数及び意見文数を表１に示す．

表1　法案関連記事の抽出結果

例題	政治家数(人)	記事数(記事)	意見文(文)
人権擁護法案	8	51	12
国民投票法案	10	107	20
憲法改正法案	26	236	77
郵政民営化法案	56	1657	373

4．おわりに

本研究では新聞記事からの政治家の意見情報の抽出・整理手法を提案した．現在は，法案関連のトピックに関して，提案手法の評価実験中である．

参考文献

[1] 村上晴美, 平田高志: Talking-永田町-政治家エージェントの擬似会話システム-, 川村洋次, 浜田秀, 村上晴美編, 文学と認知・コンピュー16-社会がつくる物語-,日本認知科学会テクニカルレポート, pp61-64 (2003).

[2] 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).