国会議員のツイッター分類とその応用

東　宏一　　掛谷英紀

qq274sw9k@yahoo.co.jp　　kake@iit.tsukuba.ac.jp

筑波大学

概要　本研究では、投票支援システム構築の一貫として、原子力発電所の是非に関する国会議員の立場分類を行った。具体的には、最大エントロピー法（MEM）による議員の分類と、SOM による議員マップ出力の 2 手法を用いた。MEMを用いた手法では、原発推進派寄りと分類された議員と反対派寄りとされた議員のツイートを被験者に読んで判断してもらったところ、反対派寄り議員のツイートの6割以上が「原発反対派」として判断された。SOMを用いた手法では、出力された議員マップ上の各議員クラスタにおいて、“自然エネルギー”や“菅内閣”など原発問題に関連するキーワードの用い方に異なる特徴が見られた。

キーワード：国会議員，ツイッター，投票支援，原子力発電所，MEM，SOM，日本語評価極性辞書

1. はじめに

近年、国政選挙などでは各政党が独自のマニフェストを提示し、有権者にその信を問うというスタイルが一般化している。しかし、有権者の多くは候補者の情報を事前に持っていないことが多い。欧米では、このためにVoteMatchという仕組みにより投票支援を行うことが定着しつつある[1]。日本でも取り入れられつつあるVoteMatchだが、課題も見られる。例えば、選挙期間中に発表されたマニフェストを情報源として投票支援を行っている点である。選挙後にはマニフェストと異なる政策が実施されることも多く、これに頼るだけでは正確な投票支援が可能とは言い難い。

そこで、著者らは各議員が選挙の前後を通し継続的に発信している情報を収集し、統計的に分析するという手法を提案している[2]。本研究では、ある特定の政治テーマに対して、各議員の立場を明らかにすることを目指す。

今回、政治テーマとしては原子力発電所（以下、原発と略記）の是非に関する議論を対象とする。このテーマを選んだのは、電力供給・エネルギー問題という国政における重大なテーマであるにも関わらず、立場を明確にしている議員が少なく、今後の国政選挙において一つの争点となることが予想されるためである。

各議員の原発問題に対する立場を把握するため、本研究では2通りの手法を用いた。1つは、原発問題に関する知識人の発言を教師信号とし、各議員の立場を分類するというものである。この手法では、各議員の特徴を把握することはできるが、議員同士の距離を把握することができない。そこで、SOM を用いた議員マップの出力を行う。これが2つ目の手法である。議員マップの出力に際しては、属性に原発関連のキーワードを用い、同一文中において、これと日本語評価極性辞書中の名詞がどのように共起するかを調べることで属性値を決定した。両手法とも、情報源にはツイッターを利用した。

2. 学習指標の収集と分析

本研究では、情報源として米国Twitter社が提供するマイクロブログサービスである”Twitter”を用いている[3]。2011年7月に知識人19名分のツイートを収集し、2011年10月に国会議員（非現職議員を含む）125 名分のツイートを収集した。これらのツイートはChasen によって形態素解析を行い、品詞は名詞のみを利用した[4]。

3. 最大エントロピー法による議員の立場分類

3.1　教師信号の作成

原発推進・維持派として 9 名、原発反対（脱原発）派として 10 名の知識人を選び、ツイートを収集した。収集したツイートに対し、原発問題に関連するキーワードをまとめた辞書によるフィルタリングと、日本語評価極性辞書によるカテゴリ分けを行い、教師信号を作成した。

3.2.1 「原発問題」関連キーワード辞書

今回、原発問題に関連するツイートのみを抽出するために、原発に関連するキーワードの辞書を作成した（以下、原発用語辞書と呼ぶ）。これは、

知識人のツイートを形態素解析して得られた素性の中から、特に原発問題に関連があると思われる名詞、345 個を抽出したものである。名詞には、「原発」、「東電」、「エネルギー」、「ソーラー」などが含まれる。

3.2.2 日本語評価極性辞書名詞編

日本語評価極性辞書名詞編とは、日本語名詞に対し、人手でポジティブな意味合いをもつもの、ネガティブな意味合いをもつもの、どちらにも属さないもの、の3種類の評価極性を付与したものである[5]。

原発推進派・反対派の各ツイートを、評価極性ポジティブの名詞が含まれているもの、評価極性ネガティブの名詞が含まれているもの、の二種類に分類した。以後行うクロスバリデーションによる実験では、このそれぞれのカテゴリについて実験を行っている。

3.3 クロスバリデーションによる実験

原発用語辞書、日本語評価極性辞書を利用して収集した知識人のツイートをフィルタリングし、さらにこれらを学習データとして 10 分割クロスバリデーションによる実験を行った。学習アルゴリズムには最大エントロピー法を用いた[6]。実験のパターンとそれぞれの結果について表 1 に示す。ここで、両方の辞書を利用する場合とは、原発用語辞書でフィルタリングした後に、残ったツイートを評価極性辞書によりポジティブ・ネガティブに分類してそれぞれ実験したものである。

表 1　各分類器と正解率

No.	学習パターン	正解率(%)
1	辞書を利用しない	57.87
2	原発用語辞書のみを利用	92.67
	日本語評価極性辞書のみ利用	
3	ポジティブ	90.39
4	ネガティブ	90.34
	両方の辞書を利用	
5	ポジティブ	89.84
6	ネガティブ	90.44

最も正解率が高かったのは、No. 2 の分類器であった。以後行う議員の分類では、これを利用して実験を行った。

3.4 政治家のツイートによる実験

No.2 の分類器に対し、国会議員のツイートをテストデータとして実験を行った。今回、実験に際して収集した議員 125 名の中で、比較的著名な議員 20 名を選んだ。

表 3 に示した 6 つの分類器の中から、特に正解率が高かった、原発用語辞書のみを利用した場合を利用して実験を行った。実験にあたっては、テストデータである議員のツイートについても教師信号と同様のフィルタリングを行った。実験結果を図 1 に示す。

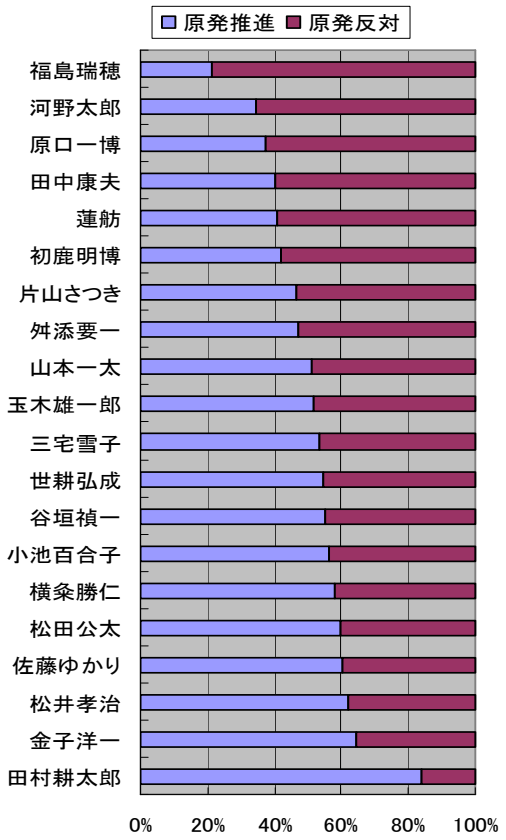

図 2　原発用語辞書のみを利用

4. 実験結果の検証

前節での実験結果について検証を行った。検証実験は、被験者に抽出したツイートを読んでもらい、原発推進、原発反対、どちらとも言えない、の 3 種類に分類してもらうという方法で行った。検証実験は原発推進派・反対派に特に近いと分類された各カテゴリ上位 4 名の議員ツイートを利用して行った。テスト用ツイートは、原発用語辞書とのマッチ度が高い発言を各議員につき 10 件抽出し、合計 80 件を用いた。

実験は、政治的な話題にはあまり詳しくない２０代の男性 1 名を被験者として行った。実験に際して、ツイート内に含まれる他の web サイトへのリンクを参照せず、本文のみを読んで判断してもらった。図 2 に結果を示す。

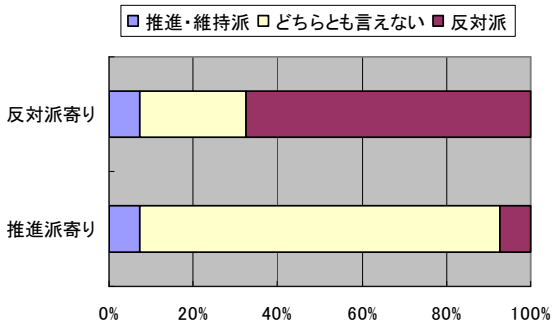

図 2　被験者実験の結果

原発反対派寄り議員は「反対派」と判断された割合が最も高く、推進派寄り議員は「どちらとも言えない」と判断された割合が最も高かった。原発反対派の知識人の発言を読むと、明確に原発への反対姿勢とを示しており、これと同様の傾向が反対派寄りと判定された議員にも見られた。

5. 名詞との共起を用いた議員マップ出力

ここまで用いてきた手法では、各議員の発言傾向は把握できたが、議員同士の距離は分からない。そこで、原発関連のキーワードと前出の日本語評価極性辞書との共起を用いて議員マップの出力を行う。議員マップの出力には SOM を用いる。SOM のアルゴリズムには Torus 型を用いる。

5. 1　入力ベクトルの生成

今回、入力ベクトルを生成するに当たり、原発関連のキーワードを属性とし、属性値には共起によるキーワードの出現頻度を用いた。

原発関連のキーワードは以下の手法により抽出した。まず、東京大学の中川らが開発した専門用語抽出エンジンである TermExtract[7]を用いて知識人のツイートから専門用語を抽出した。次に、その中で特に原発問題に関連が深いと考えられるキーワードを抜粋した。この結果得られたキーワードは、原子力発電、自然エネルギー、太陽光発電、などである。次に、各文においてこれらのキーワードと日本語評価極性辞書中の名詞との共起を調べた。あるキーワードがネガティブな名詞と共起していれば出現数をマイナス値とし、ポジティブなものと共起していればプラス値とした。全ての文についてこの共起を調べた上で、出現数を合算し、発言数で割って出現頻度とした。以上の手法により入力ベクトルを生成した。

5. 2　実験

生成した入力ベクトルを用いて議員マップの出力を行った。学習には、市販本に付属のプログラムを用いた[8]。学習結果を図 3 に示す。このマップでは色が白に近いほど隣接ノード同士の距離が近く、黒に近いほど隔たりが大きい。また、今回 Torus 型アルゴリズムを利用しているため、対称な辺同士は距離が近い。

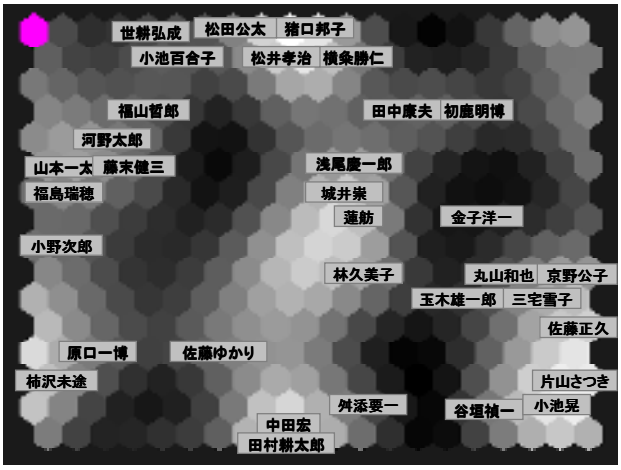

図 3　出力された議員マップ

5. 3　実験結果の検証

出力された議員マップを検証するため、マップ上の議員集団ごとに特徴的なキーワードの抽出を行った。まず、マップ上で見られる議員集団ごとに 4 つのクラスタを設定した。これを図 4 に示す。

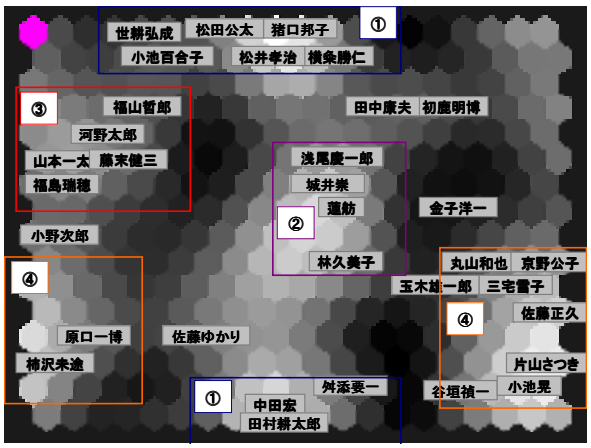

図 4　議員マップの各クラスタ

この各クラスタにおいて中心となるノードのベクトル要素と、各ベクトル要素の平均値を比較し、絶対値の差分が大きいものからキーワードを抽出した。ここで、差分が負の値のものはネガティブキーワード、正の値のものをポジティブキー

ワードとした。まず、特徴的なネガティブキーワードを抜粋したものを表 2 に示す。数字は各クラスタ番号を表している。

表 2　各クラスタにおけるネガティブキーワードの例

①	②	③	④
エネルギー	エネルギー	日本	日本
委員会	委員会	金	金
エネルギー政策	停電	子ども	原発
自然エネルギー	エネルギー政策	原発	民主党
リスク	自然エネルギー	原発事故	水
交付金	再生可能エネルギー	避難	避難
供給	計画停電	民主党	震災
菅内閣	大震災	光	汚染
電力	東日本大震災	プロジェクト	反対
仏	厚労省	放射能	事故
技術	東北地方	汚染	被災者
危機管理	原子力	行政	賠償
電力会社	復興	放射能汚染	子ども
研究者	電力会社	原発事故対策	福島
安全委	損害賠償	東電	津波

次に、各クラスタにおける特徴的なポジティブキーワードを抜粋したものを表 3 に示す。

表 3　各クラスタにおけるネガティブキーワードの例

①	②	③	④
日本	金	エネルギー	財政
被災	子ども	エネルギー政策	交付金
金	原発	委員会	安全性
原発	技術	自然エネルギー	電力会社
事故	節電	原子力	火力
震災	民主党	大震災	保安院
避難	避難	水	電力自由化
被災地	反対	再生可能エネルギー	発送電分離
反対	行政	被災地	安全神話
アメリカ	汚染	安全委員会	権力闘争
汚染	危機	電力会社	健康管理
水	東電	東日本大震災	空間線量
放射能	放射能	原子力安全委員	震災対応
賠償	被災者	原子力安全委	事故調査委
被災者	水	リスク	被災地復興

この 2 つの表を見ると、①，②のクラスタと③，④のクラスタで自然エネルギーに対する姿勢の違いがあることが読み取れる。同様に、民主党、菅内閣に対する評価もクラスタによって異なることが分かる。

6 おわりに

本研究では、議員のツイッター上での発言を利用して投票支援システムを構築することを目指し、そのための 2 つの手法を提案した。実際に各手法により議員の発言を分析したところ、最大エントロピー法による分類で特に原発反対派、賛成派に近いと判断された議員に関しては、議員マップにおける特徴的なキーワードの傾向も一致していることが分かった。

今後はこの手法を TPP 問題など他の政治テーマにも応用し、結果を総合的に判断することで、精度の高い投票支援システム実現を目指したい。

参考文献

[1] Instituut voor Publiek en Politiek (IPP), http://www.stemwijzer.nl

[2] 東宏一,橋本悠,掛谷英紀(2011), Web 上の言語資源に基づく自己組織化マップの作成,言語処理学会第 17 回年次大会

[3] twitter.jp, http://twitter.com/

[4] 奈良先端科学技術大学院大学 松本研究室 ChaSen http://cl.aist-nara.ac.jp/

[5] 乾孝司，奥村学. テキストを対象とした評価情報の分析に関する研究動向. 自然言語処理 Vol.13, No.3, pp.201-241.2006

[6] 内元清貴，村田真樹，関根聡，居佐原均(1999): 日本語係り受け解析に用いる ME モデルと解析精度，言語処理学会第 5 回年次大会併設ワークショップ論文集.

[7] 中川裕志、森辰則、湯本紘彰: ″出現頻度と連接頻度に基づく専門用語抽出″, 自然言語処理、Vol.10 No.1，pp. 27 － 45，2003 年 1 月

[8] 大北正昭，徳高平蔵，藤村喜久郎，権田英功編(2008)『自己組織化マップとそのツール』シュプリンガージャパン