2Q-09

# 政治情報サイトからの政策表現文の抽出方法とその評価

永井 茅希　　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

政治情報サイトの中で,特に日本の政党の公式サイトでは，政策が政策課題(以下，課題とする)ごとに整理されておらず分かりにくいという問題がある.そこで本研究は,各政党の政策についての文章を整理し,政策が理解しやすくなることを目的としたシステムを開発する.その目的を達成するためには政策を表す文のみを抽出した上で,課題ごとに比較しやすい形式に整形することが必要である.本研究はその第一歩として，政党の公式サイトから，機械学習を用いて政策を表す文を抽出するシステムを開発し,その評価を行う.

## 2　政党公式サイト上の政策表現文

### 2.1　政党公式サイトと政策表現文

政党の政策を知るには,公約に関連した情報が重要であると考え,政党公式サイトに着目した.政党公式サイトでは,日々の活動や,政党の方針などを有権者に向けて発信している.しかしながら，総選挙・無投票者調査[1] の「今後の選挙に希望する改善点」において,政策の伝え方が上位になっていることから,有権者は現状の政党の政策の伝え方に不満を抱いていることがうかがえる.

そこで本研究においては,政党の公式サイトから,まず課題ごとに政策表現文を抽出することを目的とする.

### 2.2　政策表現文の定義

本研究では，以下の文を政策表現文と定義する.

(a)　政党の政策
(b)　政策に付随する意見
(c)　政策に付随する意見の引用

意見の引用に関しては,政策を理解するための参考になると考え，定義に加えた.

### 2.3　政策表現文の抽出方針

複数の政党の公式サイトから,特定の課題の政策表現文を含むページを得る際,他の課題の政策表現文も含んでしまうことがある.しかしながら,元の課題の政策表現文には政党を問わずに共通する単語があるという特徴がある.この特徴に着目すれば,当該課題を表す政策表現文だけをそれぞれ抽出することが可能になる.その結果,各政党から同一の課題に対する政策表現文を抽出し,当該課題に対応する政策の比較が容易にできるようになると考えられる.

Extraction of Manifest Sentences from the Political Information Websites and Its Evaluation
Kayaki Nagai, Koichi Yamada, Hiroshi Kinukawa
Graduate School of Science and Technology for Future Life, Tokyo Denki University

### 2.4　政策表現文の特徴

政策表現文の特徴を調査したところ,以下の特徴が見つかった.

(a)政党名，党首名，政策課題名が主語になりやすい，(b)「批判」や「主張」などのサ変名詞の出現頻度が高い，(c)長文は政策表現文になりにくい，(d)意見の引用も政策表現文であるため，文中にカギ括弧が出現しやすい，(e)過去の出来ごとや疑問は政策表現文でないため,文末に助動詞の「た」や「か」が出現する文は政策表現文になりにくい．これらの特徴に着目して政策表現文の抽出を行う.

## 3　政策表現文抽出システム

本研究では，政党公式サイトから課題ごとの政策表現文を抽出し，出力するシステムを提案する．(**図1**)

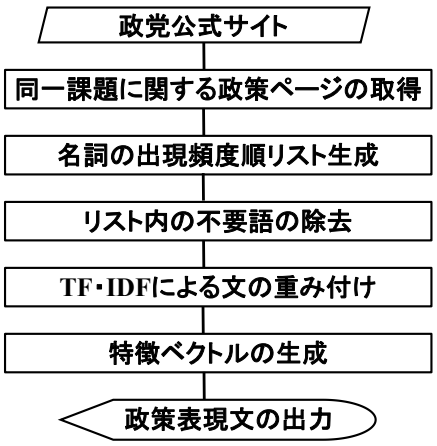


**図1　政策表現文抽出システムの流れ**

### 3.1　政党公式サイトからの政策ページの取得

政党公式サイトから課題を検索語にしてページを検索する．その結果から boilerpipe [2] を使用してテキスト部分を取得する.

### 3.2　名詞の出現頻度順リストの生成

公式サイトから取得したテキストを kuromoji [3] を使用し形態素解析する.政策表現文における重要な概念はサイト内の名詞であるため,形態素解析の結果から名詞のみを取得する.1つの課題に関するテキスト全体を1文書と見なし，名詞の出現頻度順リストを生成する.

### 3.3　出現頻度順リスト内の不要語の除去

作成したリストには不要語が含まれていることがある.課題ごとの名詞の出現頻度順リストを比較し,一定数以上の課題で共通して出現する語を不要語として除去する.

### 3.4 TF･IDF による文の重み付け

**3.2**で生成した同一課題のテキストにおける名詞をもとに TF･IDF を算出する．その際，IDF値は同一課題のテキスト全体を1文書とみなし算出する．これらの定義式を以下に示す．

$\mathrm{TF_i}$ = 取得したテキストにおける単語$\mathrm{t_i}$の出現数

$$\mathrm{IDF_i} = \log\left(\frac{\text{課題の総数}}{\text{単語}\mathrm{t_i}\text{が出現する課題の数}}\right) + 1$$

$$\mathrm{TF{\cdot}IDF_i} = \mathrm{TF_i} \times \mathrm{IDF_i}$$

次に**3.1**で取得したテキストを句点で文単位に区切る．各文に対して重み付けを行う．重み付けの値を表す係数をwとして以下に定義式を示す．

$$\mathrm{w} = \sum_{i=1}^{n} \text{文構成名詞の}\mathrm{TF{\cdot}IDF}_i$$

n = 当該文を構成する名詞の種類数

### 3.5 特徴ベクトルの生成

重み付けを行った後， **2.4** の特徴に着目してベクトルを生成する．ベクトルの要素である特徴量を以下に示す．(a)文の重み， (b)文末の助動詞の出現の有無， (c)文中にカギ括弧が出現するか， (d)出現頻度が高いサ変名詞の出現の有無， (e)文字数， (f)政党名，党首名，政策課題が主語になっているか， (g)出現したすべての単語に関する出現の有無 (政策ページを取得した際に出現したすべての単語)

最後の特徴量は，予備実験を行った結果，抽出性能を安定させるために必要と判断し，追加した．

### 3.6 機械学習を用いた文の出力

**3.5** で生成した特徴ベクトルに基づき SVM と Random Forest[4]により分類モデルを作成，これにより政策表現文か否かの判定を行い，出力する．その際 SVM には liblinear[5]を用いる．

## 4 評価実験と考察

### 4.1 評価データ

実験に用いる課題として，現在話題になっている以下の7つの課題を扱う． (a)TPP， (b)アベノミクス， (c)消費税， (d)憲法改正， (e)原発， (f)子育て， (g)集団的自衛権．これらの課題に関して，自民党，民進党，共産党の公式サイトからページを取得する．なお民進党に関しては公式サイトのデータ数が少ないため，民主党の公式サイトを利用する．

また**3.3**の不要語は，当該語が出現する課題の数が5以上の語としている．実験に使用するデータは2016年5月4日に取得したものである．文数は，自民党12,581文，民進党15,285文，共産党8,086文の計35,952文であり，このうち政策表現文は1,535文であった．

政策表現文に同数の非政策表現文を加えた計3,070文を対象に政策表現文抽出システムを用いて，5分割交差検定を行った．

### 4.2 評価指標

精度，再現率，F値により評価した．定義式は以下の通りである．

$$\text{精度} = \frac{\text{システムが出力した政策表現文数}}{\text{システムが出力した文数}}$$

$$\text{再現率} = \frac{\text{システムが出力した政策表現文数}}{\text{政策表現文数}}$$

ただし，文全体が政策表現でなくても，政策表現が含まれていれば政策表現文とする．

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

### 4.3 評価実験結果

5分割交差検定の結果の平均を**表1** に示す．

**表1 政策表現文抽出の評価結果**

| | SVM | Random Forest |
|---|---|---|
| 精度 | 79.7% | 89.3% |
| 再現率 | 83.4% | 92.7% |
| F値 | 80.9% | 91.0% |

### 4.4 考察

我々の以前の研究[6]において54.1%であったF値を，本研究では30%以上向上させることができた．なお特徴量として，取得文の出現位置，文末の助動詞の種別なども検討したが，予備実験によって有効でないことが分かった．

## 5 おわりに

本研究では，政党の公式サイトから課題ごとの政策表現文を検索し，得られた結果のテキストから機械学習を用いて政策表現文を抽出するシステムを提案した．自民党，民進党，共産党の公式サイトを用いて実験をした結果F値は，SVMで 80.9%，Random Forestで 91.0%を得た．

今後の課題として，抽出文の整形手法の検討がある．

## 謝辞

使用させていただいたboilerpipe，kuromoji， liblinearの開発者の方々に深く感謝致します．

## 参考文献

[1] 株式会社 ジャパン・マーケティング・エージェンシー，“総選挙・無投票者調査”，https://www.jma-net.com/reports/総選挙・無投票者調査データの公開， 2012

[2] boilerpipe， https://code.google.com/p/boilerpipe/

[3] kuromoji， http://www.atilika.org/

[4] L. Breiman “Random Forests，” Machine Learning， vol. 45， no. 1 pp. 5-32， 2001.

[5] liblinear， https://www.csie.ntu.edu.tw/~cjlin/liblinear/

[6] 永井茅希，山田剛一，絹川博之， “政治情報サイトからの政策表現文の抽出”， 第15回情報科学技術フォーラム講演論文集， 第2分冊， pp. 109-110， 2016