E-001

政治情報サイトからの政策表現文の抽出

Extraction of Manifest Sentences from the Political Information Websites

永井 茅希† 山田 剛一† 絹川 博之†

Kayaki Nagai Koichi Yamada Hiroshi Kinukawa

1 はじめに

今日の日本の課題の一つとして,投票率の低さが挙げられる.特に20代の投票率の低さは顕著であり,第47回衆議院議員総選挙における20代の投票率は60代の投票率の半分以下である[1].また,平成27年6月の公職選挙法の改正により,選挙権年齢が20歳から18歳に引き下げられた[2]ため,今後はよりいっそう若者の政治参加が求められる.

本研究では,政治的な情報を取り扱うサイトから,政策を表す文を抽出するシステムを作成する.このシステムにより,各政党の政策についての文章を容易に閲覧できるようにし,若者の投票の意思決定を支援できるようにすることを目的とする.

その目的を達成するためには,文を抽出した後,政治的な課題ごとに比較できるような,見やすい形式に整形することが必要である.その第一歩として,本研究においては,政治的な情報を扱うサイトから政策を表現する文を抽出する機能の開発を行う.

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

政治情報の中でも,公約に関連した情報が重要である.そのため,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[3] 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などから,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窺える.

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,以下のような文を政策表現文と定義した.

(1) 政党の政策
(2) 政策に付随する意見
(3) 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 政策表現文を含むページの特徴

政策表現文を含むページの特徴として,政党ごとに同一の課題に関連するページの見出しが異なっていても,本文には政党を問わずに共通する単語が使用されているという特徴がある.この特徴に着目して,各政党から同一の課題に対する政策表現文を抽出すれば,その政策の比較が容易にできるようになると考えられる.加えて,ある課題に関するページ内に,その課題と関係のない文が含まれることがある.その際にも政党間に共通する単語という特徴に着目することにより,課題と関係のない文を除去できると考えられる.

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作成した.作成したシステムの流れを**図1** に示す.

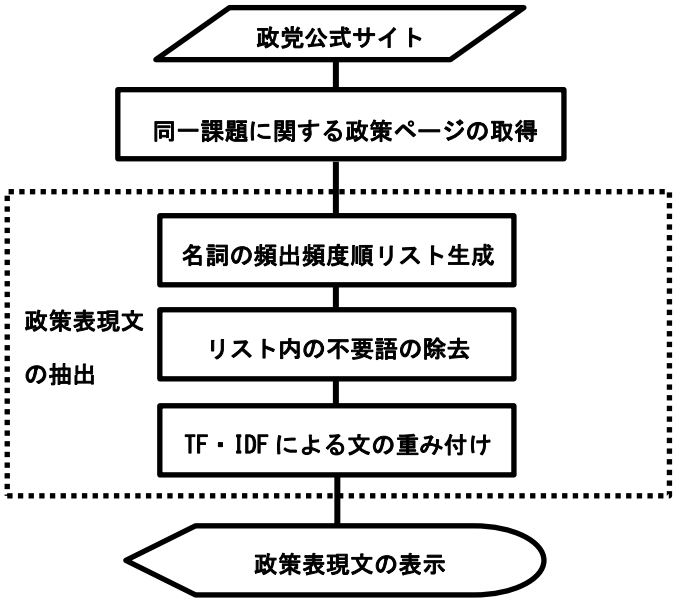

図1 政策表現文抽出システムの流れ

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 [4] 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 [5] 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.課題ごとの名詞出現頻度順リストを比較し,一定数の課題で共通して出現する語を不要語として除去する.

†東京電機大学大学院 未来科学研究科
Graduate School of Science and Technology for Future Life, Tokyo Denki University

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

文の重みを算出した後, 長文であるほど値が大きくなることを防ぐために係数kを掛けて正規化を行う. 係数kは文長をもとにした正規化係数である. 正規化した値をnwとして定義式を以下に示す.

$$\mathrm{k} = \frac{2\sum_{i=1}^{m}\frac{\text{文}\mathrm{S_i}\text{の文字数}}{\mathrm{m}}}{\text{文}\mathrm{S_i}\text{の文字数}} \quad \text{ただし,} \mathrm{k} > 1 \text{のとき } \mathrm{k} = 1$$

m = システムが取得したテキスト中の文の総数

$$\mathrm{nw} = \mathrm{w} \times \mathrm{k}$$

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8つの課題を扱う. (1)TPP, (2)アベノミクス, (3)マイナンバー, (4)憲法改正, (5)原発, (6)子育て, (7)集団的自衛権, (8)消費税, これらの課題に関して,自由民主党,民進党,共産党の公式サイトからページを取得する.なお民進党の公式サイトは設置されて間も無く, 十分なデータが取得できないため, 民主党のサイトを利用する. また**3.3**の不要語の除去の際には, 文書頻度が5以上のものを不要語とし, 文を出力する際には, 重み付けした値の上位5%を出力する. これらの値は予備実験によって設定した値である. 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り, 文数は, 自由民主党12,581文, 民進党15,285文, 日本共産党8,086文となっている. これらを対象に政策表現文抽出システムを用いて, 政策表現文を抽出する実験を行った.

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

評価実験の結果を**表1** に示す.

表1 政策表現文抽出の評価結果

	自由民主党	民進党	日本共産党
精度	52.5%	46.2%	50.5%
再現率	61.0%	70.6%	63.7%
F値	53.2%	54.8%	54.4%

4.4 考察

政策表現文抽出システムが出力した政策表現文ではない文は, 事実だけを述べている文や, 誰かに対する問いかけのような文が多かった. これらの特徴として, 文末が助動詞「た」であることや, 終助詞「か」であることが挙げられる. そのため,これらのような文から不要な文の特徴を見つけ, 処理する方式を検討することが, 精度の向上に繋がると考えられる.

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, 得られた結果の文字列から政策表現文を抽出して表示するシステムを提案した. 自由民主党, 民進党, 日本共産党の公式サイトを用いて実験をした結果, 3党平均のF値が54.1%という値を得た. 今後の課題として, 抽出の精度の向上, 抽出した文の整形手法の検討が挙げられる.

謝辞

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

参考文献

[1] 総務省, 国政選挙における年代別投票率について, http://www.soumu.go.jp/senkyo/senkyo_s/news/sonota/nendaibetu/
[2] 総務省, 選挙権年齢の引き下げについて, http://www.soumu.go.jp/senkyo/senkyo_s/news/senkyo/senkyo_nenrei/
[3] 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
[4] boilerpipe, https://code.google.com/p/boilerpipe/
[5] kuromoji, http://www.atilika.org/