国会における議員の発言の自動要約システム

掛谷英紀 **尾崎正宗** **佐藤裕也**

kake@iit.tsukuba.ac.jp s1320805@u.tsukuba.ac.jp s1420791@u.tsukuba.ac.jp

筑波大学大学院システム情報工学研究科

概要 本研究は、国会における議員の発言を要約し、短時間で国会議員の普段の言動の傾向を読み取れるような情報提供を行うシステムの実現を目指す。その具体的手段として、本論文では2つの要約文自動生成手法を提案する。1つ目の手法は、専門用語辞書 TermExtract で重要度を設定してナップサック制約付最適化問題（MCKP）を解くことで要約文を作成する手法である。従来は tf-idf 法で算出した重要度に基づいた要約が提案されているが、国会での議論の特性を踏まえ、専門用語辞書による重要度設定を用いることを試みる。２つ目の手法は、コメント付き動画サイトのコメントが多く付いている部分を抽出して要約文を生成する手法である。この方法で自然な要約文を作成する方法として、コメントが多い部分を段落単位で抽出してつなぐ手法を試みる。以上の方法に基づいて作成した要約文について、従来手法の要約文と混ぜて被験者に提示し、要約文としての評価を問う実験を行ったところ、動画サイトでコメントの多い部分を抽出する要約文が最も高い評価を得た。

1. はじめに

一時期、日本の国政選挙では、各党が掲げるマニフェスト（政権公約）をマスメディアが取り上げ、それに基づいて投票先を選ぶことがマスコミなどから推奨された。その流れで、マニフェストに基づき、自分の考え方と似た政党を選ぶ VoteMatch という仕組みが定着しつつある[1]。

VoteMatchとは, 利用者に１０から３０程度の政策上の争点に関する賛否を回答させ、政策的立場に最も近い政党や候補者などを提示するウェブ上の仕組みの１つである[2]。日本国内でも、２００７年の参議院選挙から静岡大学の佐藤らのグループや毎日新聞社が、独自にVoteMatchを実施している[3,4]。

しかし、マニフェストに基づく投票先の判断には問題もある。その一つとして、各政党が必ずしもマニフェストを守らないことである。また、マニフェストは政党が掲げているものであるため、必ずしも個々の議員の意見と一致しているとはいえない。実際、選挙後政党が分裂することもしばしばである。

本来は、政党の政策だけでなく、各国会議員の普段の言動も参考にして誰に投票するかを決めるのが理想的である。しかし、現実にはそうした情報を短期間で取得するのは難しい。そこで、議員個人の政治思考を手軽に判断できる手段として、橋本らおよび東らの Web 上のレビュー記事による分類 [5, 6]や、東らの国会議員のツイッター分類[7]のように、議員同士の類似度を自己組織化マップ（SOM）によって提示するシステムが提案されている。また、尾崎らは、橋本・東らの研究をもとに投票支援システムの構築を行っている[8]。これは、VoteMatch と同様に、利用者に Web 上でアンケートを行い、そのアンケートのデータを用いて議員の SOM 上に利用者の位置を示すものである。しかし、これらのシステムは議員の意見を読み取れないという欠点がある。

これを解決する手段として、類似度マップ上の興味のある議員に対してマウスオーバーをした際に、その議員の意見を提示するという方法が考えられる。本研究では、有権者が議員の普段の言動を手短に知ることができるようにするために、国会議員の普段の言動である国会内での発言を自動的に要約するシステムを構築することを目的とする。

2. 手法

本研究では、国会会議録を大きく分けて２つの手法によって要約している。１つ目の手法は、要約をナップサック制約付最適化問題（maximum coverage problem with knapsack constraint (MCKP)）に帰着させ、貪欲アルゴリズムを用いる手法[9]である。MKCP は、要約手法として広く知られているものである。本研究では、単語の重要度として、従来用いられている tf-idf 法[10]による重要度の他に、TermExtract[11]を使用する方法を試みる。具体的には、TermExtract により算出された重要度に対数をとったものを用いる。

さらに、本研究では、ニコニコ動画[12]のコメント数に着目して要約を作成する手法を新たに提案し、これを２つ目の手法として１つ目の手法との比較評価を行う。コメント数が多い部分は、その動画の中でも部分である可能性が高い。そこで、本手法はコメント数が多い部分を下記のアルゴリズムで抽出する。

ニコニコ動画では、動画の長さや、動画を１分毎に区切った区間でのコメント数などでコメント保持数が決定されている。しかし、その動画のすべてのコメントは過去ログから取得することができる。今回は、NicomentXenoglossia[13]を用いてニコニコ動画の過去ログを取得する。

コメントを入力する際、ほとんどの視聴者は１秒単位の誤差は意識しないことが考えられる。その為、コメントを収集するのは、ある程度まとまった時間で行う必要がある。また、コメントの対象となる部分がそのコメントの時間より前の時間になっていることがしばしばある。そのことも考慮して、今回は１秒毎にその時間から５秒後までの平均コメント数の値を計算することとした。ニコニコ動画では、コメントが流れる速度はコメントの長さによって異なるため、一概には言えないが、左端から右端までコメントが流れるまで約３～５秒程度となっている。コメント書き込み密度が多いシーンを抽出するため、今回は５秒毎の平均値を算出する。

また、ニコニコ動画に基づく要約では、段落単位の抽出を行うことで要約文を作成する。文単位の選択による抽出を行うことも考えられるが、特に指示語が多く含まれる場合、文単独で意味が分からない場合が多くある。そこで、内容の一まとまりである段落単位で抽出を行うこととした。

さらに、例外処理として、必ずしも１つの段落で内容が完結していない場合もあるため、その段落が２００字以内かつ先頭に接続詞または指示語がある場合は前の段落も抜き出すようにする。さらに、動画の最初や最後の部分において、拍手コメントによってコメントのピークが来ている場合がしばしばあるため、最初や最後の拍手コメントはコメント数に計上しないことにした。

3. 評価実験と結果

評価実験では、平成 23 年 1 月 28 日の有村治子議員、平成 25 年 1 月 30 日の平沼赳夫議員の代表質問を使用した。コメント数は動画[14],[15]のものを使用している。

この実験では、Word による要約、ニコニコ動画のコメント数の推移に着目して要約を行う手法、参照要約に加え、MCKP の手法を４つ用意した。MCKP の手法は、TermExtract に基づき重要度 W_l' を決定し、文字数 c_l をで割った W_l'/c_l を使用してナップサック問題を解く手法と、そのまま W_l' を用いる手法の２つと、tf-idf に基づき重要度を決定し、文字数で割った W_l'/c_l を使用してナップサック問題を解く手法と、そのまま W_l' を用いる手法の２つの合計４つである。

Word 要約、ニコニコ動画のコメント数による要約、参照要約と MCKP 要約の４つのモデル、計７つの手法に対して評価実験を行った。なお、いずれのモデルも文や段落の順序は出現順になっているものを提示した。また、文抽出を行う MCKP の各手法に対して、元の発言を参照し、同じ段落であるならば、同じ段落として段落分けを行い被験者に提示した。評価基準は３つで、政治的主張が読み取れるかを基準α、興味深い表現が抽出されているかを基準β、自然な日本語になっているかを基準γとして、５段階評価（５が最高）で評価させた。被験者は日本語の母国語話者の成人の男女１４人である。

表１、表２、表３は、それぞれ評価基準α、β、γのそれぞれについて、２つの文書の要約に対する評価の度数を合算したものを示している。MCKP は括弧の中に詳細な手法の違い（文字数で割ったタイプ⇒*Wl'/cl*、そのままのタイプ⇒*Wl'*）と用いた重要度（TE⇒TermExtract, tfidf ⇒tf-idf 法）を記述している。

表１　基準αの評価結果（評価別人数）

手法＼評価値	1	2	3	4	5
Word	6	10	5	6	1
MCKP（*Wl'*,tf-idf）	1	10	11	5	1
MCKP（*Wl'*,TE）	3	7	14	3	1
MCKP（*Wl'/cl*,tf-idf）	4	12	7	3	2
MCKP（*Wl'/cl*,TE）	3	14	5	6	0
Niconico	2	4	14	8	0
Reference	0	2	11	10	5

表２　基準βの評価結果（評価別人数）

手法＼評価値	1	2	3	4	5
Word	6	12	6	4	0
MCKP（*Wl'*,tf-idf）	3	8	13	4	0
MCKP（*Wl'*,TE）	2	9	10	3	4
MCKP（*Wl'/cl*,tf-idf）	2	8	11	6	1
MCKP（*Wl'/cl*,TE）	5	14	4	5	0
Niconico	1	6	10	7	4
Reference	2	1	8	15	2

表３　基準γの評価結果（評価別人数）

手法＼評価値	1	2	3	4	5
Word	10	10	5	2	1
MCKP（*Wl',tfidf*）	5	10	7	4	2
MCKP（*Wl'*,TE）	6	9	9	2	2
MCKP（*Wl'/cl,tfidf*）	3	7	8	6	4
MCKP（*Wl'/cl*,TE）	12	9	6	1	0
Niconico	1	5	6	10	6
Reference	2	2	8	8	8

これらの結果からわかる通り、ニコニコ動画のコメント数による要約は自動評価手法の中で最も高く評価された。αの項目では、参照要約に大きく劣り、U 検定での *p* 値 1.8e-03 と有意差がはっきりと出ている。しかし、β、γの項目で参照要約との有意差はない。一方、MCKP（*Wl'/cl*,TE）による要約が Word 要約と同程度悪い結果となった。原因としては、短めで重要度が高い文が抽出されやすいため、他の手法と比較して文のつながりが悪いことが考えられる。U 検定を行ったところ、すべての項目について MCKP（*Wl'/cl*,TE）要約と Word 要約との有意差はなかった。MCKP（*Wl',tfidf*）要約、MCKP（*Wl'*,TE）要約、MCKP（*Wl'/cl,tfidf*）要約については、U検定を行ったところ、３つの手法間ですべての項目について有意差はなかった。

表４に参照要約と比較して求めた各自動要約手法で得られた要約の ROUGE 値[16]を示す。今回の評価実験では、Word 要約以外はある程度 ROUGE 値と人手の評価値の間に相関関係が見られている。

表4 各手法の ROUGE 値の比較

	ROUGE-1	ROUGE-2
Word	0.369105	0.186655
MCKP（*Wl',tfidf*）	0.365015	0.112135
MCKP（*Wl'*,TE）	0.40952	0.20196
MCKP（*Wl'/cl,tfidf*）	0.351015	0.18064
MCKP（*Wl'/cl*,TE）	0.33597	0.16336
Niconico	0.413235	0.22174

4. まとめ

本研究では、コメント付き動画サイトのコメント数による自動要約手法を提案し、国会会議録の議員の発言を対象に MCKP による要約手法の比較を行った。

評価実験の結果、ニコニコ動画のコメント数による要約手法は、政治的主張に関しては参照要約と比較すると十分取り出せていないが、興味深い部分を参照要約と同程度抽出できているとの評価が得られた。また、日本語としての読みやすさについても参照要約と同程度の高い評価が得られた。

コメント付き動画サイトのコメント数による要約手法は、コメント、ユーザのフィルタリング処理などをすることで、より興味深い要約を生成できる可能性がある。現在は、段落区間切り分けを人手の作業で行っているが、音声認識で自動的に切り分けることもできる。それらの作業の自動化が今後の課題である。

参考文献

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

[2] 上神貴佳, 堤 英敬 (2008): 投票支援のためのインターネット・ツール ―日本版ボートマッチの作成プロセスについて― ,『選挙学会紀要』10 号, pp. 27‐48.

[3] 静岡大学情報学部 佐藤哲也研究室 http://tai.ia.inf.shizuoka.ac.jp/

[4] 毎日新聞 えらぼーと http://mainichi.jp/select/seiji/eravote/

[5] 橋本悠, 掛谷英紀（2010): 自然言語処理を用いた Web 上のレビュー記事の分析とその応用, 言語処理学会第 16 回年次大会発表論文集.

[6]東宏一, 橋本悠, 掛谷英紀（2011): Ｗｅｂ上の言語資源に基づく国会議員の分類, 言語処理学会第 17 回年次大会発表論文集.

[7] 東宏一, 掛谷英紀 (2012): 国会議員のツイッター分類とその応用, 言語処理学会第 18 回年次大会発表論文集.

[8] 尾崎正宗, 掛谷英紀 (2013): Web 上のレビュー記事に基づく投票支援サービスの構築,第 9 回メディア情報検証学術研究会講演論文集.

[9] Filatova, E. and Hatzivassiloglou, V. (2004): A Formal Model for Information Selection in Multi-Sentence Text Extraction, in Proceedings of the 20th International Conference on Computational Linguistics (COLING), pp. 397-403.

[10] 天野正家, 石崎俊, 宇津呂武仁, 成田真澄, 福本淳一 (2007):『ITText 自然言語処理』P. 138

[11] 東京大学情報基盤センター・中川裕志, 東京大学経済学部・前田朗, 専門用語（キーワード）自動抽出用 Perl モジュール"TermExtract" http://gensen.dl.itc.u-tokyo.ac.jp/termextract.html

[12] ニコニコ動画, http://www.nicovideo.jp/

[13] NicomentXenoglossia, http://xenog.web.fc2.com/

[14] 平成 23 年 1 月 28 日 参議院本会議 有村治子（自） 代表質問【圧倒的な人材】 http://www.nicovideo.jp/watch/sm13428422

[15] 平成 25 年 1 月 30 日 衆院本会議代表質問・平沼赳夫(日本維新の会) http://www.nicovideo.jp/watch/sm19954845

[16] Lin, C.-Y. and Och, F. (2004): Automatic Evaluation of Machine Translation Quality Using Longest Common Subsequence and Skip-Bigram Statistics, in Proc. of the 42ndAnnual Meeting of the Association for Computational Linguistics, pp. 606–613.