1ZA-8

文間の類似性を用いた国会会議録のトピック別要約の検討

金丸 浩司 † 西崎 博光 ‡ 関口 芳廣 ‡

† 山梨大学大学院医学工学総合教育部 ‡ 山梨大学大学院医学工学総合研究部

1. はじめに

膨大な情報が溢れる現代において，大量の文書から必要な情報（知りたい情報）を検索，抽出，再構成できる要約器の登場が望まれている．我々は，複数のトピックが存在し，かつ一般の人間の関心が高いと考えられる国会会議録[1] を対象とした要約器の作製を目指している．そこで本稿では，会議録から発話内容を検出し，トピック毎のクラスタリングを行なって要約文を生成する処理について報告する．

2. 国会会議録の特徴

国会会議録には以下のような特徴がある．

1. 国会会議録の内容は質疑応答型が多いが，質問と回答は近くにあるとは限らず，その対応付けは非常に難しい，
2. 国会会議録では，複数の質問者，回答者が現れるために，類似の質問・回答が繰り返し出現する，
3. 国会会議録ではキーワード（特に名詞）の羅列が多く，文構造が曖昧な場合が多い．

よって，1) それぞれの質問部分と回答部分を対応付けする．2) 質問・回答の組であるサブトピックをクラスタリングし，トピックを抽出する．3) キーワードとなる名詞句をトピック中から抽出することで，国会会議録を要約するシステムを作製する．本稿では，質問と回答の組をサブトピックと定義し，その類似性を考慮したサブトピックの塊りをトピックと定義する．

3. 要約文生成のためのトピックの特定

トピック別要約文生成処理の流れを図 1 に示す．現在は，文章中に含まれる特徴的な表現から質問と回答の対応付けを行ないサブトピックを抽出し，サブトピック中の文からそれに類似するサブトピックを抽出し，トピックを見つける処理まで行なっている．

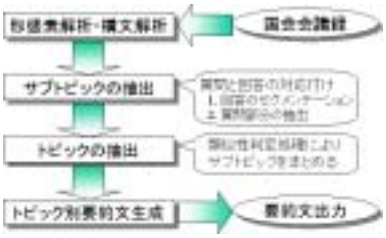

図 1: トピック別要約文生成の流れ

Topic Summarization of the Minutes of the National Diet using Similarity between Sentences , by Koji KANEMARU† , Hiromitsu NISHIZAKI‡ and Yoshihiro SEKIGUCHI‡

†Department of Education Interdisciplinary Graduate School of Medicine and Engineering , University of Yamanashi

‡Department of Research Interdisciplinary Graduate School of Medicine and Engineering , University of Yamanashi

[1] *http://kokkai.ndl.go.jp*

3.1 質問と回答の対応付け

一般に国会会議録では，一人の発話に複数の内容（年金問題，環境問題など）が含まれる．トピック毎の要約文を生成するためには質問発話・回答発話それぞれに含まれている複数の内容を認識して対応付けをすることが必要であり，これにより会議録の閲覧が非常に容易となる．また，国会会議録では，質問，回答に特徴的な表現が現れる．本稿ではこれを「手がかり表現」と記述する．そして，発話中に現れる手がかり表現を利用して，上記の問題の解決を試みる．

3.1.1 回答のセグメンテーション

「手がかり表現」は，質問発話中の手がかり表現と回答発話中の手がかり表現に分類できる．質問発話における手がかり表現の例を以下に示す．

- 重ねて 所見を伺いたい。
- 今年度補正について措置すべきだが、国土交通大臣の 見解をお伺いいたします。

次に，回答発話における手がかり表現の例を示す．

- 特殊法人改革 についてのお尋ねであります。
- 社会保障の基盤が損なわれるためリストラをやめさせるべきだ という御指摘がありました。

このような手がかり表現は，発話内容が変わるタイミングで多く出現する．実際に 3 年分の国会会議録を分析し，これらの手がかり表現を抜き出した．その結果，回答発話中においては，内容が変わる時ほぼ決まった手がかり表現が出現することが分かった（質問発話中の手がかり表現は内容が変わった部分の 7 割程度で用いられている）．よって，回答発話では手がかり表現が出現するタイミングでセグメンテーションを行なうことが出来る．また，この処理により回答に対応する質問がどの程度存在するかも分かる．

3.1.2 回答に対応する質問部分の抽出

回答のセグメンテーションを行なった後，各回答に対応する質問部分の抽出を行なう．処理としては，各回答の手がかり表現を含む文中に出現する名詞（キーワードと呼ぶ）を抽出し，質問発話内にこれらのキーワードが出現するかを調べる．これは，手がかり表現を含む文中で用いられる名詞はサブトピックの内容をよく示す単語と考えられ，それらキーワードが固まって出現する場所が，回答部分に対応する質問部分の可能性が高いためである．ただし,「社会保障」など，複合名詞は，名詞を分離させると別の意味を持つ単語になるため，このまま扱う．また，多くのサブトピックに出現する可能性が高い高頻度で出現する名詞を用いないように，出現回数 5 回未満の名詞をキーワードとしている．しかし，キーワー

ドは対象の質問部分以外でも出現する可能性があるため，上記の処理だけでは関係ない質問まで抽出するおそれがある．よって，各質問部分の候補を比較することで，質問部分の絞り込みを行なう．

3.2 トピックのクラスタリング

これまでに得られた各サブトピックから，サブトピック間の類似性を判定し，最終的なトピックを抽出する．トピッククラスタリングのために，サブトピック内に含まれる各文を対象とし，コサイン尺度と格構造を組み合わせた手法をとる [3]．コサイン尺度が閾値以上の場合，お互いの文の述語とそれに付属する格要素を照応し，全てが一致した場合に二つの文は類似文であると判定する．ここで，コサイン尺度と格構造の処理の精度をあげるために，文献 [1] を参考にした文中の冗長部分の削除・表記統一，EDR 日本語単語辞書 [2] の語釈文を利用した類義語辞書の作成を前処理として行なっておく．類似文であると判定された二つの文を含むサブトピックは類似トピックであるとし，以後同一トピックとして扱っていく．これにより，同じ内容を持つトピックの要約文が複数存在することを防げる．ただし，現時点では「構造改革」など大きな分野でのトピッククラスタリングではなく,「構造改革の有効性」「構造改革特区」などのように細かい内容についてのクラスタリングレベルにとどまっている．

4. トピック判定実験

以下に示す実験を行なう．ここで，実験データは平成 13 年から平成 15 年までに開かれた計 6 会議録（364 セグメント）である．

4.1 回答のセグメンテーション実験

回答部分に出現する手がかり表現を利用したセグメンテーション実験では，364 個の内容中，9 つの内容に関してはさらに細かくセグメンテーションが行なわれてしまったため，セグメンテーションの再現率は 1.0，適合率は 0.98 となった．

4.2 質問と回答の対応付け実験

回答に対応する質問部分の抽出実験では，364 個の回答に対して 289 の質問部分を得ることができた．対象となる質問部分が抽出できなかった原因として，手がかり表現中のキーワードが適当でなかった（質問発話中でキーワードを含む表現が用いられていなかった）ことなどがあげられる．

4.3 トピッククラスタリング実験

実験で用いている会議録中には類似文は 93 セットあり，コサイン尺度の閾値を 0.26 に設定し，格構造で類似文を判定した．結果として，抽出組 42，抽出組中の正解 39 であり，再現率 0.42，適合率 0.93，F 尺度 0.58 という結果を得た．ただし，抽出組中で誤った 3 組はいずれも質疑が終った後の報告発話であり，会議の内容とは全く関係がないため実際には適合率は 1.0 といえる．

この結果を元にトピッククラスタリングを行なうと，364 個のサブトピックが 337 個にまとめられる（再現率・適合率ともに 1.0 だとサブトピックは 323 個までまとめられる）．この 337 トピックについて各要約文を作成する．

5. トピック別要約の試み

質問と回答を再構成した要約を生成するための特徴を調べるために，9 名の被験者に国会会議録から要約文を作成してもらった．例を図 2 に示す．

原文
「質問」:
国際テロ組織との闘いは、周辺国家の協力や国連の強化など、これまでにない新たな知恵が求められます。テロは犯罪であり、犯罪を見逃さず闘うのは政治に課せられた責任であることは言うまでもございません。二十一世紀の新たな戦争とも言うべきこうした行為に、総理はこれまでにないどのような新たな戦略を必要とお考えでしょうか。
「回答」:
まず、国際テロに対する戦略についてのお尋ねであります。今回のアメリカで発生しましたテロに対しましては、これは米国のみならず、全世界に対する自由、平和、民主主義に対する攻撃である、そういう認識から、我が国もみずからの問題として主体的に取り組むべき課題であると思っております。そういう観点から、今後、テロ組織の侵入を許さない出入国管理の徹底、テロ組織に関する情報収集の強化のほか、テロ組織の資金活動面を含めて取り締まりに関する法制度や装備、資機材の整備に努めて、我が国において国際テロの活動を許さない毅然たる対応が重要であると思っております。さらには、御指摘のように、世界の国々や国際機関と一致団結して対応するとともに、テロを起こさせないような国際環境の形成を目的としてさまざまな外交努力等の適切な対応を行いまして、テロリズム根絶のためのあらゆる努力を尽くす必要があると考えます。

要約例
テロ組織との闘いに求められる新たな戦略としては、入出国管理の徹底や情報収集の強化、テロ組織に対する法制度などの対応が重要であり、周辺国家の協力など、テロ根絶のための国際環境の形成に努力を尽くす必要がある。

図 2: 国会会議録と人手による要約例

人手による要約の特徴は下記の通りである．

- 質問部分は殆んど要約に使われないが，回答の補足として挿入されることがある
- 要約では文構造はあまり残らない．
- 具体的なキーワードは省略せずに要約文に入れる

主に文や文節のつなぎ変えによって要約文を作成できること [4] と上記の特徴を参考にし，現在要約文を作成するアルゴリズムとその検討を行なっている．

6. おわりに

本稿では，国会会議録における質問-回答の対応づけとトピッククラスタリングについて報告した．今後，トピック別要約文生成とその評価を行なう予定である．

参考文献

[1] 足達康昭, 山本和英. 特徴的冗長表現に着目した国会会議録要約. 情報処理学会研究報告, NL157-15/FI72-15, pp. 107–114. 情報処理学会, 9 2003.

[2] 日本電子化辞書研究所. EDR 日本語単語辞書 version2.0. 1998.

[3] 金丸浩司, 関口芳廣, 西崎博光. 国会会議録要約文生成のための文間の類似度計算. 第 3 回情報科学技術フォーラム講演論文集, E-029, pp. 175–176, 9 2004.

[4] 竹内和広, 松本裕治. 自動文節対応付けを用いた要約中の文再構成操作の調査. 言語処理学会論文誌「自然言語処理」, Vol. 9, No. 3, pp. 87–108, 7 2002.