E-029

国会会議録要約文生成のための文間の類似度計算
Similarity between Sentences for Summarization of the Minutes of the National Diet

金丸 浩司 † **関口 芳廣** ‡ **西崎 博光** ‡
Koji Kanemaru Yoshihiro Sekiguchi Hiromitsu Nishizaki

1. はじめに

近年，インターネットの普及により，様々な情報を享受することが出来るようになった．しかしその反面，爆発的に情報量が増加しており，効率よく情報を入手することが困難になりつつある．このような問題に対して，大量の文書から必要な情報（重要な情報）を検索，抽出，再構成できる要約器の登場が望まれている．性能のよい要約器を作成するには，例えば同じような内容を持つ文の組が，解析対象の文章に出現した場合，計算機上でそれらを特定できれば，冗長性が少ない要約文を作成することが可能になる．本稿では，複数文書，特に冗長な言い方が多い国会会議録 [1] を対象とし，コサイン尺度と格文法を組合せた簡便な手法を用いて文間の類似性を判定する方法を提案し，類似文判定実験を行なった結果，その有効性を確かめられたので報告する．

2. 類似文判定処理

要約文生成と類似文判定の処理の流れを図 1 に示す．本稿では，意味解析・文脈解析の一処理として行なっている類似文の判定について説明する．

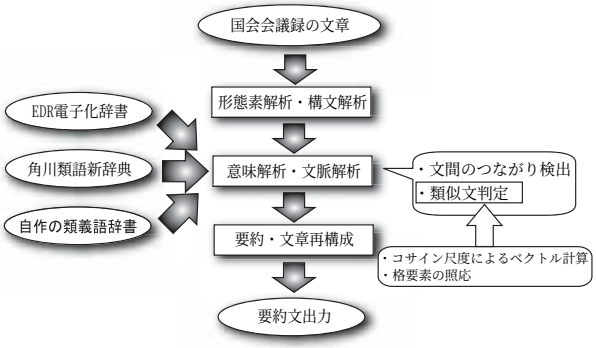

図 1: 要約文生成と類似文判定処理の流れ

2.1 前処理

文間の類似度計算を行なう前に，以下に述べる前処理を行なう．

2.1.1 冗長箇所の削除

一般的な話し言葉と同様，国会会議録にも多くの冗長表現が見られる．冗長表現の例として，

- ～というふうに考えております。
- ～こととしております。

などが挙げられる．こうした冗長部分は，作成した規則により削除する．規則は基本的には文献 [2] を参考にして，本研究では特に文末表現の冗長部削除を重視した．これは，本研究で述語部分の情報を用いるため，前述のような冗長表現を述語部分と誤認識させないためである．また,「～につきまして」「～に関しまして」などは ChaSen[3] では格助詞として扱われるが，前述の表現はほぼ同じ格の意味を持っていると考えられるので，換言処理を行ない，表記を「～について」に統一する．

2.1.2 類義語辞書の作成

名詞の類義語辞書として，角川類義語辞典 [4] を用いるが，さらに類義語辞書の精度を挙げるために，EDR 日本語単語辞書 [5] の概念説明を用いて，類義語辞書を作成した．例えば,「犠牲」と「損害」の場合,「犠牲」の概念説明「災害, 事故などにおける損失」と「損害」の概念説明「事故や災害などで受ける損失」に出てくる名詞を使って，名詞が出現したら 1，しなかったら 0 とおき，コサイン尺度（後述）で計算を行なう．類似判定は閾値を用いるが，現在はヒューリスティックに 0.8 を採用している．

2.2 類似文判定手法

文間の類似度計算は，以下のコサイン尺度で定義する．

$$sim(\boldsymbol{x},\boldsymbol{y}) = \frac{\boldsymbol{v}_x \cdot \boldsymbol{v}_y^T}{\sqrt{(\boldsymbol{v}_x \cdot \boldsymbol{v}_x^T)(\boldsymbol{v}_y \cdot \boldsymbol{v}_y^T)}} \qquad (1)$$

ここで，$\boldsymbol{x}$，$\boldsymbol{y}$は文，また $\boldsymbol{v}_x$，$\boldsymbol{v}_y$ は文$\boldsymbol{x}$，$\boldsymbol{y}$に含まれる各名詞の出現頻度を要素とする単語ベクトルである．

式 (1) の値が大きいと類似文としているが，この手法は出現単語の頻度のみを用いているので，類似文の抽出はよくできる．ただし，類似文以外も類似文であると多く判定されてしまう．そこで本研究では，以下の順序で類似文の判定を行なう．

1. コサイン尺度による類似度計算で類似文の候補を出す
2. 候補文に対して，格文法を用いて格要素同士を照応させ，格要素の照応度合で類似文と判定する

以下に具体的な手法を説明する．

2.2.1 類義語辞典を用いたベクトル計算

文間のベクトル計算は，基本的には式 (1) を用いる．ただし，表記は違うが意味が同じ単語が文中に出現する可能性がある．例えば,「テロの犠牲になられた方々」と「テロの被害に遭われた方々」という文中で,「被害」と「犠牲」はほぼ同じ意味を表している．このような同義

†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

語の問題に対処するため，類義語辞典を参照し，類義語と認められた場合には，同じベクトル要素として扱う．計算結果が適当な閾値以上の場合，類似文の候補として，次の処理へ渡す．

2.2.2 格文法による類似文判定

ここでは，活用語（動詞，形容詞，断定の助動詞「だ」）の持つ格情報から，類似かどうかを判断する．ただし，話し言葉という性質上，深層格を調べることは非常に難しいため，本研究では格助詞と名詞の組合せが一致するかどうかを調べるという簡単な手法をとっている．ただし，活用語の照応は，述語か，もしくは重文とみなされた部分の活用語を対象とし，IPAL[6] の情報から，上位語，類義語などの単語も含めて格要素の照応を行なう．

格要素が完璧に一致し，さらに格要素を修飾している部分の一つでも修飾部が一致した場合，その文は類似文だと判断する．

また，対象としている国会会議録の性質を考慮して，解析には，以下の制約をつけている．

I. 動詞「聞く，伺う，承る，求める」が「ヲ格」を持ち，かつその格の要素が「見解」かまたはその類義語だった場合，他の動詞と照応を行なわない，

II. 動詞「終わる」が「ヲ格」を持ち，かつその格の要素が「質問」かまたはその類義語だった場合，他の動詞と照応を行なわない，

III. 動詞「ある」が「ガ格」を持ち，かつその格の要素が「お尋ね」だった場合，他の動詞と照応を行なわない，

IV. 議長の発言はすべて格要素の照応を行なわない．

3. 類似文判定実験

3.1 実験と結果

2001 年度〜2003 年度に行なわれた 20 会議録中の類似文の組合せを人手で抽出し，結果として 211 種類の組合せを得た．これまでに述べた手法で類似文の抽出を行なうが，正解 211 組でもっとも低い類似度の値（　0.26）をコサイン尺度での閾値とする．この時の格文法を組み合わせた結果は，抽出組 204，抽出組中の正解 133 であり，再現率 0.630，適合率 0.651，F 尺度 0.641 という結果であった．類似文と判定された組の例を以下に示す．

- 去る九月十一日の米国における同時多発テロにより犠牲になられた方々、被害に遭われた方々、御家族初め関係者の方々に、改めて、心からお悔やみとお見舞いを申し上げます。
- 犠牲となられた方々は六千人とも言われ、心から哀悼の意をささげるとともに、行方不明者の御家族の方々にはお見舞いを申し上げたいと存じます。

次に，コサイン尺度での閾値を 0.3 から 0.1 刻みで変化させ，格文法を用いて類似文を判定した時の再現率，適合率，F 尺度の値を図 2 に示す．

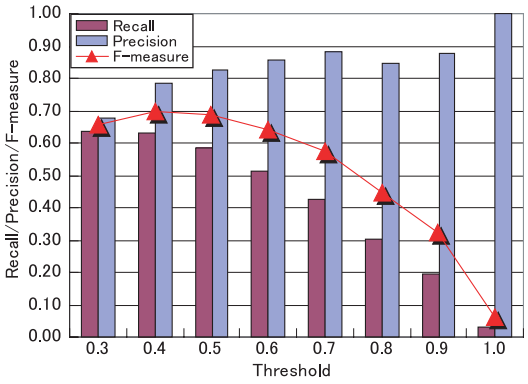

図 2: 閾値を変化させた時の提案手法の結果

3.2 考察

簡単な手法ではあるが，比較的良好な結果を得ることが出来た．国会での発言という性質から，原稿を用意して答弁をするため，比較的似通った文が多いのではないかと考えられる．また，相手の発話した内容をそのまま鸚鵡返しする答弁も多かったが，これらの情報を有効に利用することが出来れば，質問-回答の対応付けが精度よく出来るかも知れない．

格文法の処理を組み合わせずにコサイン尺度のみで類似文の判定を行なうと，再現率は高いままであるが，適合率は 0.1%以下になる．これは，類義語辞典を用いたために，類似文でない文も類似文だと判定されたためである．ただし，格文法の処理のためにより多くの候補を残したことがよい結果を生んでいる可能性もあり，それらについても調査する必要がある．

4. おわりに

類似文を得るための一般的な手法であるコサイン尺度に，格文法の概念を組み合わせることにより，精度のよい結果を得ることが出来た．今後は，高度に言い換えられた文（「今は何時ですか.」と「時間が知りたい.」など）に対しての手法を考えるとともに，類似と判定された文の組から，互いの文の情報を保持しながら一文に再構成することを試みていく．また，今回は制約を用いて実験を行なったが，そういった国会会議録の特徴を計算機上で自動学習できれば，さらに良い結果が出せると考えられる．

参考文献

[1] 国会会議録検索システム，国立国会図書館，http://kokkai.ndl.go.jp/
[2] 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.107-114，2003.9.
[3] 松本裕治，北内啓，山下達雄，平野善隆，松田寛，高岡一馬，浅原正幸：形態素解析システム「茶筌」version2.3.3 使用説明書，松本研究室，2003.
[4] 大野晋，浜西正人:角川類語新辞典，角川書店，1981.
[5] 日本電子化辞書研究所：EDR 日本語単語辞書 version2.0，1998.
[6] 情報処理振興事業協会：計算機用日本語基本辞書 IPAL，1993.