E-031

複数語句から象徴語句への換言可能性に関する考察
A Consideration on Paraphrasing Several Phrases into a Symbolic Phrase

渋木 英潔 † 荒木 健治 †† 桃内 佳雄 †† 栃内 香次 ‡‡
Hideyuki Shibuki Kenji Araki Yoshio Momouchi Koji Tochinai

1. まえがき

今日では，インターネットの普及等により大量の文書にアクセスすることが可能となった．しかしながら，人間の情報処理能力には限界があるため，それらの文書全てに目を通すことは容易ではない．それゆえ，自動要約の重要性が一層高まっていると考えられる．なお，本研究は単一文書を対象とした報知的な要約を目的としている．

従来の単一文書要約は，文単位で重要度を付与して抜粋する重要文抽出型の要約 [1, 2] と，一文ごとに要約を行う文内要約 [3] に関する研究が多く行われている．これらの手法は文単位での処理であるため，断片的な情報を示す文が並べられた出力となり，文間の関係性などが不明瞭となりやすい．それゆえ，我々は文未満の単位で重要箇所を抽出し自然な文となるよう再構築して出力する要約を目指している [4]．

このような要約においては，要約対象となる文中に存在しない語句を加えて再構築した方が，要約された文間の関係性を明確にできる可能性がある．しかしながら，要約文間の関係性を象徴するような語句は，原文中の語句と意味的に強い関連がある一方で，表層的には原文中に直接出現することはあまりないと考えられる．一般に文生成には深いレベルの解析が必要と考えられるが，現在の意味解析や文脈解析は精度的に改善の余地が存在する．また，文間の関係性を焦点として Web 上の情報を用いた研究には文献 [5] などが存在し，Web 情報を活用した文生成の可能性を示唆している．したがって，構文解析までの結果と Web 情報を利用して，比較的浅いレベルの処理で解決することを目的としている．本稿では，形態素解析と Web 検索の結果を活用して，要約対象となる文間の関係性を象徴する語句の生成可能性に関する考察を行う．

2 節では，要約対象文に含まれる語句から関係性を象徴する語句を抽出するまでの処理を記述する．**3** 節では，抽出された象徴語句に対する考察を行う．**4** 節は今後の予定である．

2. 処理

本手法は，Web 上に存在する文は可読性の点で自然な文であり，それらの文中に含まれる語句の関係が明瞭であるような語句を含んでいるという仮定に基づいて行われる．象徴語句を抽出するまでの流れを図 1 に示す．本稿では，入力として平成 18 年の小樽市議会 [6] の議事録を用いた．

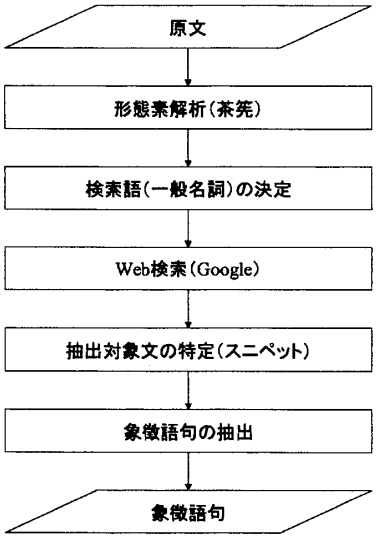

図 1: 象徴語句抽出の流れ

まず，要約対象となる文書に対して ChaSen[7] を用いて形態素解析を行う．次に，Web 検索のためのキーワードとして，形態素解析の結果から品詞が「名詞–一般」または「名詞–サ変接続」となる単語を抽出する．本手法では，原文を一度語句単位に分解した後，要約文として再構築するアプローチをとっているため，一度の検索で用いる語句は語数に基づくこととした．本稿では，文書の最初から 10 語区切りで検索語をまとめた．

本稿での Web 検索は Google SOAP Search API[8] を用いて行った．続いて，検索された文書から象徴語句を含む文を特定する必要があるが，特定にはスニペットを利用した．スニペット中の文には，全ての検索語が含まれているとは限らない．しかしながら，本稿での目的は，検索対象となった語句全ての関係性を明確にすることではなく，入力文書に含まれていない新規な象徴語句を発見することである．スニペット中の文にも新規な語句は含まれており，象徴語句を発見する目的においては問題ないと判断した．

最後に，特定された文から象徴語句の抽出を行う．現段階では最も単純な処理として，スニペットの最初の一文に対して ChaSen で形態素解析を行い，「名詞–一般」または「名詞–サ変接続」となる単語の中で検索語に含まれていない最初の語を抽出した．

3. 考察

入力された文章と，検索で用いた語句および特定されたスニペット文の例を表 1 に示す．入力文章の上部には

†北海学園大学ハイテク・リサーチ・センター, High-Tech Research Center, Hokkai-Gakuen University
††北海道大学大学院情報科学研究科, Graduate School of Information Science and Technology, Hokkaido University
‡北海学園大学工学部, Faculty of Engineering, Hokkai-Gakuen University
‡‡北海学園大学経営学部, Faculty of Business Administration, Hokkai-Gakuen University

表 1: 検索語とスニペット文の例

人件費の削減比率	
入力文章	今年度は一律 7 パーセントカットを行っていました．確かにパーセンテージでのカットを行えば，多く給料をもらっている人は多く，少ない人は少なく削減されると思いますが，しかしながら入って間もない新入社員ですら，仕事をする前から 7 パーセントカットということになります．本来は経営責任が発生し，責任の重い人たちがその削減率を高めるのが本来であると思います．
検索語	カット，パーセンテージ，給料，人，削減，間，新入，社員，仕事，経営
スニペット	「業績不振 のため全社員の給料を一律 20 %カットしたいと思っています．」 「解雇 しやすければ，会社は，たしかにとりあえず多くの人を雇うだろう．」 「僕が新入社員の時書いた レポート （企画書） が随分使われていた見たいですね．」
新市立病院基本設計に伴う発注方式	
入力文章	プロポーザル方式をとるということですが，これについての説明をお願いいたします．また，公募型とおっしゃいますが，どのように公募をされるのか，その方法をお教えください．次に，選定委員を市職員部長職，外部委員を含め，10 名で構成されるということですが，この委員名を事前に公表しない理由をお知らせください．
検索語	プロポーザル，方式，説明，お願い，公募，方法，選定，委員，市，職員
スニペット	「それでは，事務局 より説明をお願いします．」 「市職員におかれましては，任命書 を配布しておりますので宜しくお願いいたします．」 「総務 部長は，三重県建築設計公募型プロポーザル方式実施要領第 2 条第四号に規定する選定委員会の審査に基づき，参加表明・基本提案書 ... 説明を求めることができる．」
市民協働	
入力文章	市民協働が信念だとお話しされていた市長ですが，病院新設に伴うことをはじめ，さまざまな政策において，市民との協働歩調がとりきれていないように思います．その市民との協働のシステムや行動を形にしたものが，まちづくり条例とか自治基本条例と呼ばれるものです．
検索語	市民，働，信念，お話し，市長，病院，新設，政策，歩調，システム
スニペット	「今後，市民の 皆さん の参画と協働によるボランティアの輪が更に広がるものと期待しております．」 「東金市に 中央 病院を新設し，大網と成東両病院を支援病院として運営することになっていた．」 「討議 資料『これが真実のお話し』の中で，Q1 で，『市長は大手民間病院を優遇して市民病院をつぶそうとしていると言う噂を聞き ..... 市としては初めての，市民との協働で行う広報広聴活動であることは理解できる． 」

ボールド体で入力文章の主題を示している．スニペットには検索結果が 1 位から 3 位におけるスニペットの最初の一文を示しており，抽出された象徴語句には下線を引いている．

現段階では要約まで実現していないため定量的な評価は困難であるが，表 1 に示すように，主題とある程度関係のある語句が抽出されていると考えられる．しかしながら，抽出された語句がそのまま要約に利用できるわけではなく，例えば，人件費の削減比率において「業績不振」から「費用削減の理由」を導き出すようにもう一段階処理を行う必要がある．また，市民協働の例では「皆さん」よりも「ボランティア」の方が関係性が高いと考えられるため，象徴語句の抽出アルゴリズムを改善する必要がある．これらは今後の課題である．

4. 今後の予定

象徴語句の抽出アルゴリズムの改善を行うとともに，スニペットから抽出された象徴語句を利用した要約アルゴリズムの構築を行う予定である．

参考文献

[1] G. L. Thione, M. Berg, L. Polanyi, C. Culy: Hybrid Text Summarization: Combining External Relevance Measures with Structural Analysis. In Proceedings of ACL-2004 Text Summarization Branches Out, pp.51–55 (2004).

[2] 桜井俊彦, 内海彰: 情報検索のためのクエリに基づく文書自動要約, 言語処理学会第 10 回年次大会, pp.265–268 (2004).

[3] 大竹清敬, 増山繁: 多重修飾に着目した文内要約：削除型換言, 言語処理学会第 7 会年次大会ワークショップ論文集, pp.59–64 (2001).

[4] 渋木 英潔, 荒木 健治, 桃内 佳雄, 栃内 香次: 句単位の複数文要約に向けての基礎的検討, 言語処理学会第 13 回年次大会発表論文集, pp.1149–1151 (2007).

[5] Yali Ge, Rafal Rzepka, Kenji Araki: Machines Having Other Ideas than Ours? - Evaluation of a System Based on Commonsensical Knowledge Retrieved from the Web, 言語処理学会第 12 回ワークショップ「感情・評価・態度と言語」論文集, pp.25–28 (2006).

[6] 小樽市議会会議録: http://www.city.otaru.hokkaido.jp/gikai/gijiroku.htm

[7] ChaSen/茶筅: 奈良先端科学技術大学大松本研究室, http://chasen.naist.jp/hiki/ChaSen/

[8] Google SOAP Search API: http://code.google.com/apis/soapsearch/