句単位の複数文要約に向けての基礎的検討

渋木 英潔† 荒木 健治‡ 桃内 佳雄*† 栃内 香次◇

†北海学園大学ハイテク・リサーチ・センター ‡北海道大学大学院情報科学研究科
*北海学園大学工学部 ◇北海学園大学経営学部

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．

自動要約は利用目的の観点から，原文の参照前に適切性の判断などに用いる指示的（indicative）な要約と，原文の代わりとして用いられる報知的（informative）な要約に分類される [1]．また，処理対象の観点からは「単一文書要約」と「複数文書要約」に大別できるが，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では，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 [2, 3] と，一文ごとに要約を行う文内要約 [4] 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．また，分野を特定して文未満の重要な要素を抽出し表形式で出力する研究 [5] も行われているが，不特定分野の要約においては可読性などの点から自然文としての出力が望ましい．それゆえ，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手法が期待される．

しかしながら，文未満の単位での要約には文生成の問題を解決する必要がある．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．したがって，本研究では，構文解析までの結果と Web 上の情報を利用して，比較的浅いレベルの処理で解決することを目的とする．本稿では，そのための基礎的検討を行う．

2 節では，目標とする要約システムの概要を述べる．**3** 節では，基本となるシステムを構築し，要約率に関する予備実験を行う．**4** 節はまとめと今後の予定である．

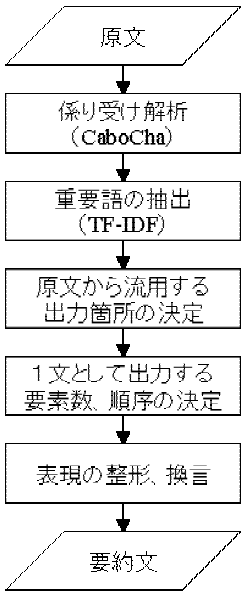

図 1: 手法の概要

2 手法の概要

提案手法における基本的な考え方は，単語をノードとするネットワーク構造として文書全体を表現し，出力する部分のネットワークを切り出した後，文として再構築するというものである．本手法による要約の流れを図 1 に示す．

最初に入力文書に対し係り受け解析を行う．係り受け解析には CaboCha[6] を使用した．文節境界は CaboCha の出力をそのまま用い，文節内の単語に対し以下の処理を行うことで未知語や過分割に対処した．文節の末尾から先頭に向かって単語を調べ，「助詞」,「助動詞」,「記号」以外の品詞をもつ単語が出現した箇所を自立語と付属語の境界とした．境界前の単語群を一つの自立語とし，境界直前の単語の品詞をその自立語の品詞とした．直前の語が未知語であった場合には，付属語があり，かつ，末尾の付属語の品詞が「助動詞」である場合には「動詞」，それ以外の場合には「名詞」とみなした．

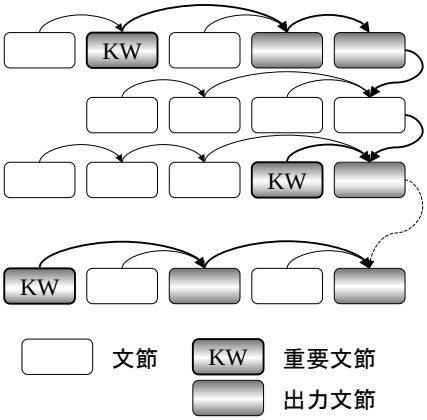

図 2: 出力箇所の決定の例

名詞と解析された単語の中から，重要と思われる語をキーワードとして抽出する．キーワードの抽出にはTF-IDF 法が一般に知られており [7]，本稿でも Web 上の検索結果を基にした TF-IDF 法を用いた．入力文書 d における単語 t の TF-IDF 値は以下の式 (1) で計算される．

$$tf{\cdot}idf(t,d) = tf(t,d) \cdot \log \frac{N}{df(t)+1} \qquad (1)$$

$tf(t,d)$ は文書 d 中に出現する単語 t の回数であり，$df(t)$ は単語 t をキーワードとして検索した時のヒット件数である．検索は Google SOAP Search API[8] を用いて行い，N の値を 80 億とした．また，分母が 0 にならないよう 1 を足している．

抽出したキーワードを基点として，要約文に出力する箇所を決定する．図 2 に出力箇所決定の例を示す．自立語をノードとし，係り受け解析の結果に基づいてノード間にリンクを張る．文書全体を一つのネットワークで表現するために，文末の文節は次の文末文節に係るとした．将来的には係り受け関係に加えて，要約上の意味的に関連のあるノード間にリンクを張ることで精度を高めたいと考えている．しかしながら，現段階では係り受け関係のみにリンクを張っているため，ツリー構造となっている．キーワードを含む文節から文末文節までリンクを辿り，その過程にあるノードを出力箇所とした．図 2 では陰影が付いたノードが出力箇所となる．

次に一文として出力する要素数を決定する．将来的には埋め込み文などに対して，順序を考慮した出力を行う必要があると考えられるが現段階では考慮していない．したがって，入力文書の先頭から順に出力箇所となるノードを出力していくが，どのノードを文末とするかが問題となる．条件の一つとして，原文における文境界を要約文の境界とするのが妥当である．しかしながら，図 2 の 1 文目には出力箇所が 3 文節しかなく，これをそのまま文として出力すると短すぎて不自然な表現になりやすい．予備実験で用いた文の平均文節数が 6.4 文節であり，一般的にもその程度の長さである文が読みやすいと考えられる．したがって，6 文節以上となる文末文節までを出力範囲とした．

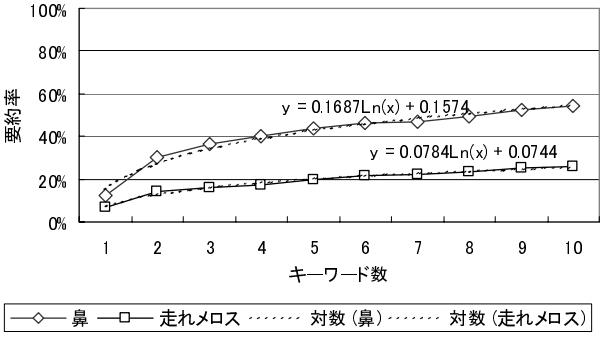

図 3: キーワード数と要約率

最後に，自然な表現となるように以下のように整形を行う．係り先のノードが出力対象となっている場合には原文の表現をそのまま使用し，出力対象でない場合には直後の出力単語に係るものとして整形する．現段階では，整形以前の処理における未解決の課題が多いため，整形処理は連体形と連用形への変化しかさせていない．係り先ノードの決定等も含めて今後の課題である．

3 要約率の予備実験

一般的な要約ではユーザが設定した要約率に基づいて要約が行われる．提案手法では，キーワードを基点にリンクを辿って出力範囲を決定するため，キーワード数と要約率の関係が明確でない．それゆえ，前節で述べたシステムを用いて，キーワード数と要約率の関係を調査した．

実験では，青空文庫 [9] から芥川龍之介の「鼻」と太宰治の「走れメロス」を用いた.「鼻」における TF-IDF による上位 10 語は順に,「内供」,「鼻」,「弟子」,「僧」,「事」,「顔」,「中童子」,「それ」,「自分」,「これ」であり,「走れメロス」では,「メロス」,「私」,「おまえ」,「セリヌンティウス」,「王」,「友」,「身代わり」,「それ」,「君」,「わし」であった．また，文書中の文数と文節数は,「鼻」が 159 文の 1,540 文節,「走れメロス」が 458 文の 2,393 文節であり，異なり名詞数は「鼻」が 434 語,「走れメロス」が 664 語であった．

図 3 は，基点となるキーワード数と出力された文の

要約率の関係を示したものである．また，対数曲線として回帰分析した結果を示す．図 3 からキーワード数と要約率の関係を対数曲線で近似できる可能性があると考えられる．しかしながら，その係数の予測には，より多くの文書を用いた調査が必要であり今後の課題である．

現在のシステムは開発の土台となるシステムであるため，その出力の多くは誤りを含んだものであり，精度を評価できる段階ではない．しかしながら，比較的良好であった結果をあげるならば，以下のような例があげられる．

原文 独りで食えば，鼻の先が鋺の中の飯へとどいてしまう．そこで内供は弟子の一人を膳の向うへ坐らせて，飯を食う間中，広さ一寸長さ二尺ばかりの板で，鼻を持上げていて貰う事にした．

要約 鼻の先がとどいてしまう内供は弟子の一人を坐らせて，食う間中，鼻を持上げていて貰う事にした．

原文 内供はその短くなった鼻を撫でながら，弟子の僧の出してくれる鏡を，極りが悪るそうにおずおず覗いて見た．

要約 内供は鼻を撫でながら，弟子の僧の出してくれる鏡を，見た．

最初の例では，二文で表現されている原文が適切な長さの一文に要約されている．二番目の例は，文内要約に相当するが，不自然な短さとなっていない．しかしながら，最後の読点は冗長である．この例に限らず，全体的に読点が冗長に含まれる傾向にあった．また，TF-IDF を用いているため頻出語がキーワードとして選出される傾向にあり，結果として要約文の殆どにキーワードが冗長的に含まれる傾向がみられた．今後，これらの問題を考慮して改善していく必要がある．

4 まとめ

本稿では，構文解析までの結果と Web 上の情報を利用して，文節単位の複数文要約を行う手法の概略を述べた．本手法は，Web の検索件数に基づいた TF-IDF 値を用いてキーワードを抽出し，キーワードを基点に係り受け関係を辿って出力範囲を決定する．その後，ノード数を考慮して一文あたりの出力要素数を決定し，直接の係り受け関係がない場合には係り先の品詞に応じて整形し出力する．また，キーワード数と要約率の関係を調査するための予備実験を行い，対数曲線で近似できる可能性を示唆した．今後，予備実験の結果を元に改善を続けていく予定である．

謝辞

本研究の一部は，北海学園大学ハイテク・リサーチ・センター研究費による補助のもとに行なわれた．

参考文献

[1] 奥村学, 難波英嗣: テキスト自動要約に関する研究動向 (巻頭言に代えて), 自然言語処理, Vol.6, No.6, pp.1–26 (1999).

[2] 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).

[3] 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).

[4] 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).

[5] A. Farzindar and G. Lapalme: Legal Texts Summarization by Exploration of the Thematic Structures and Argumentative Roles. In Proceedings of ACL-2004 Text Summarization Branches Out, pp.27–34 (2004).

[6] CaboCha/南瓜: 奈良先端科学技術大学大松本研究室, http://chasen.org/ taku/software/cabocha/

[7] 松村真宏, 大澤幸生, 石塚満: 語の活性度に基づくキーワード抽出法, 人工知能学会誌, Vol.17, No.4, pp.398-406 (2002).

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

[9] 青空文庫: http://www.aozora.gr.jp/