放送ニュース要約のための表現置換

田中 英輝　木下 明徳　後藤 功雄　熊野 正　加藤 直人
NHK 放送技術研究所
{tanaka.h-ja, kinoshita.a-ek, goto.i-es, kumano.t-eq, katou.n-ga}@nhk.or.jp

1 はじめに

NHK では短いニュースをデジタル放送やインターネットの文字ニュースの中で提供している。これらのサービスは通常のニュース原稿を人手で要約して行われている。著者らはこの手間の軽減を目的としたニュースの自動要約・支援の研究を行っている。先の報告（田中 2005）では、要約者の作業過程を観察し、これに基づいた要約作業モデルを報告した。このモデルには、リード文の一部の表現を、本記の表現によって変更する機能を含んでいた。本稿ではこの変更を自動化する試みとして、文字列照合アルゴリズムに基づいた手法を提案する。以下、第 2 節ではニュースの構造と置換による要約の説明を行い、第 3 節で手法の概要を説明する。続く第 4、5 節で実験、および結果を説明して、本稿をまとめる。

2 ニュースの構造と置換による要約

典型的なニュース記事は、全体の概要である「リード」とこれを詳述する「本記」からなる。また補足情報は「追記」されることがあるが、本稿では簡単のためリードと本記の 2 種類からなるとする。
人手の要約ニュースとその原文ニュースを観察したところ、要約の骨格には原文ニュースのリード文が使われ、その一部が本記の表現で置換、挿入されることが明らかになった。リード文だけでは具体性が不足するのが主な原因である。
例えば、次のリード文

「東京の上野動物園のパンダリンリンの二世誕生を目指すためメキシコからメスのパンダが到着しました。」

に下記の挿入（括弧内）と置換（取消線部分）操作を行うことで人手の要約を復元できる。

「東京の上野動物園（のオス）のパンダ（、）リンリンの ~~二世~~（2 世）誕生を目指すため（、）メキシコから（借りた 16 歳の）メスのパンダ（、シュアンシュアン）が（3 日夜上野動物園に）到着しました。」

NHK の要約はきわめて抜粋の性格が強く（田中 2005）括弧の中の情報は、ほぼ本記に見いだすことができる。
以上の観察から、著者らはニュースの自動要約、要約支援で、このようなリードに対する置換や挿入機能を使うことを検討した。以下はリード文に対する、挿入と置換をユーザに自動提案するシステムを作ることを想定した報告である。なお、今後は置換と挿入の両方を合わせて「編集」と呼ぶ。

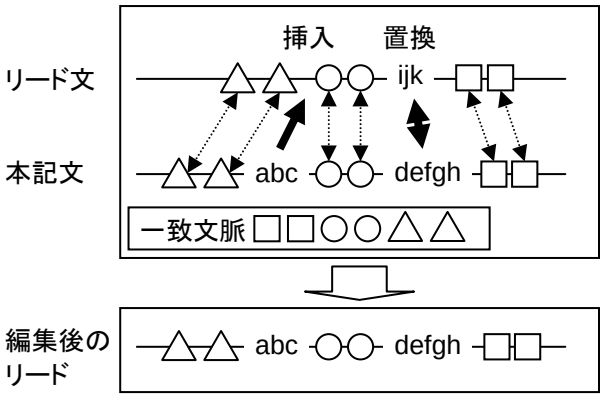


図 1 文字列照合アルゴリズムによる編集

3 表現編集手法の概要

どのようなリードの編集が効果的だろうか。また、できるだけ頑健で高性能な手法を見いだしたい。このような疑問、要請に答える最初の試みとして、今回、編集距離による文字列照合（アラインメント）アルゴリズムを使った編集を検討した。以下その動機を説明する。
リードに対する編集を行う際には、「どこに」、「どういう」編集を行うかを決定しなくてはならない。著者らはこの決定に「共通の文字列に挟まれた表現は同一の意味を表す」という「言い換え」や「同一表現抽出」の研究でよく用いられる仮定を採用した。これは、前節の例で示したようなリードの編集箇所を観察すると、リードと本記の周辺の文字列が同一である例が見られたことによる。
また、この仮定に基づいた編集を実現する手段として、編集距離を用いた文字列照合アルゴリズムに着目した。これは二つの文字列の一致箇所を効率良く照合するもので、あらかじめ決められた挿入、削除、置換の各操作のコストの合計が最小になるように文字列の一致箇所を認定する。なお、本稿ではリード文に対する操作を考えるので、挿入操作はリードに対する操作であると約束する。また今回は、リードの削減を対象外とするため、削除は考慮しない。
文字列照合アルゴリズムをリード文と本記文の二文に適用すると図 1（上）のような一致、挿入、置換の各状態が決まる。
ここで、一致文字列に挟まれた置換と挿入を候補として採用する。すなわち、このアルゴリズムによって先の仮定を満足する編集が得られることになる（図 1 下）。なお本稿では、挿入、置換候補の両側にある一致文字列を文脈と呼ぶ。実際の実験では、

以上の手順に次の変更を加えた。

照合単位

アルゴリズムの照合単位に採用したのは形態素と文節である。これらはいずれも言語的に意味のある単位であるため、編集による表現の接続を滑らかにする効果が期待できる。

文脈条件

編集候補の両端文脈が 1 照合単位以上一致していれば編集候補とする。ただし、事前の観察の結果、照合単位が形態素の場合には、一文字の文脈の一致をすべて採用すると、不適切な編集が多数発生した。このため、両端の文脈のどちらかが 2 文字以上一致するという条件を付加した。

4 表現編集実験

表現編集実験の詳細を説明する。本実験のために評価用システムを作成し、システムの提示した結果を 1 名で主観評価した。

編集実験に使ったのは 2004 年 1 月 19 日のニュース記事 23 本である。以下、具体的な手順を示す。

リード文指定

評価者はシステムに表示されたニュースのリード文を選択する。なお、本実験については、すべて第 1 文だけがリード文であった。

編集候補提示

システムはリードと本記の各文との間で文字列照合アルゴリズムを適用して、置換と挿入候補を認定して評価者に提示する。リードと本記文との間で照合を行うと、複数の編集候補が得られることがある。この場合、一ヶ所のみ挿入もしくは置換したリード文を、すべて作成して評価者に提示した。図 1（下）を例にすると “abc” を挿入したリード文と”ijk” を”defgh”に置換したリードを別に提示した。

また、実験は照合単位を形態素とした場合と文節とした場合で行った。それぞれ形態素法、文節法と呼ぶ。

評価

編集結果を 5 段階評価した。評価は、要約支援システムを想定して、ユーザに置換を提示する価値があるかどうかの 5 段階である。具体的には以下の通りである。

5　置換候補とすべき
4　かなりの確度で置換候補とすべき
3　ある程度、置換候補にしたほうがよい
2　置換候補にしない方がよい
1　置換候補にしてはいけない

置換候補とすべきかどうかの判断は、次の条件を考慮して決めた。

表 1 全体結果

	形態素法	文節法
記事数	23	23
編集発生記事数	23	19
全文数	136	112
編集発生文数	103	38
編集発生数	353	58
平均評価値	2.17	3.69

表 2 挿入の結果

	形態素法	文節法
挿入発生数	76	7
挿入表現平均長（文字）	8.92	27.9
挿入平均評価点	2.89	4.00

- 文法的に正しい表現であること
- 意味が矛盾しない表現であること
- 具体化、抽象化の効果を持つこと

最後の「具体化、抽象化の効果」の条件を考慮したのは、編集操作によって、何らかの効果が得られたことを評価するためである。この項目は同一表現抽出であれば不要だが、本研究では重要な項目となる。

5 結果

5.1 実例と考察

以下に実例を示す。それぞれ下線部が挿入表現で、置換の場合は括弧に置換元を示す。

形態素法（第 3 文による挿入　5 点）
政府は今月の月例経済報告で全体的な景気判断をこれまでの「持ち直している」から「着実に回復している」に上方修正し 3 年ぶりに「回復」という表現を使って日本の景気が回復に向かっていることをはっきりと打ち出しました。

形態素法（第 4 文による置換　1 点）
政府は今月の月例経済報告で全体的な景気判断を「着実に回復している」に上方修正し 3 年ぶりに「回復」という表現を使ったのは平成 13 年 1 月以来 3 年ぶり（て日本）の景気が回復に向かっていることをはっきりと打ち出しました。

文節法（第 3 文による置換　5 点）
今月〇日〇〇市で行われた成人式で、新成人の一部が、大声を上げたり、壇上に上がるなどして職員から何度も制止されたほか、市民憲章が書かれた（新成人が）垂れ幕をはずすなどの騒ぎを起こした問題で、きょう騒ぎを起こした男性 10 人とその保護者が市長に謝罪しました。

全体の結果を表 1 に示す。この表を挿入、置換それぞれに分類した結果を表 2、3 に示す。また形態素法、文節法の評価点の分布を図 2、図 3 に示す。

形態素法と文節法には下記の相違が見られる。

表 3 置換の結果

	形態素法	文節法
置換発生数	277	51
置換元表現平均長（文字）	9.04	14.6
置換後平均長（文字）	8.61	26.2
置換平均評価点	1.98	3.65

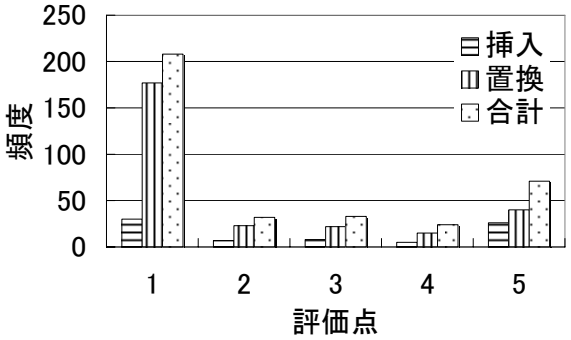


図 2 形態素法の評価点分布

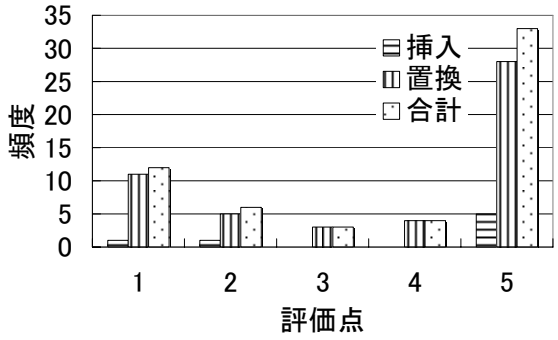


図 3 文節法の評価点分布

○ 編集発生数

表 1 から分かるように、形態素法では 353 回、文節法では 58 回であり、圧倒的に形態素法での発生が多い。文節は形態素より長い、このため、両端文脈の一致が形態素に比べて起こりにくくなることが原因である。

○ 評価点

表 1 の評価値平均を見ると、形態素法は 2.17 で文節法は 3.69 である。さらに図 2、3 の得点分布を見ると、形態素法の得点の最頻値は 1 で文節法の最頻値は 5 である。形態素法でユーザに提示できるのは一部だけであり、文節法は逆に大半をユーザに提示することができることがわかる。これも、文節法の文脈長が長いことが主要因だと思われる。

○ 置換と挿入

編集には挿入と置換がある。そこで、形態素法と文節法でそれらがどのように違うかを分類した。結果を表 2 と表 3 に示す。いずれの場合でも挿入の発生回数は置換よりも少ない。

また挿入の評価点は置換よりも高い。特に、全体の評価値が低い形態素法の場合、挿入だけの平均点は 3 の「ある程度ユーザに提示できるレベル」に近づいていた。正確な理由はよく分からないが、置換の場合は、置換元の表現との意味的な同一性を考慮した上で置換後の自然性を考慮するのに対して、挿入の場合は、前者が不要な点があると考えられる。

○ 編集の長さ

文節法で得られる置換、挿入表現は形態素法で得られるものより長い。表 2、3 を見ると、形態素法で得られる挿入表現の平均長が 8.92 文字であるのに対して文節法では 27.9 となっている。これも文節法では一致文脈が発生しにくい効果だと考えられる。

5. 2 リード文の類似文による編集

文節法では比較的高い評価点を得た。しかしユーザに編集を提示する応用を考えた場合に、文節法だけでは編集が起こりにくいこと、置換や挿入の発生単位が長いため編集後のリードの長さを制御しにくい問題がある。そこで、形態素法で得られる編集候補の中から、ユーザに提示する価値のあるものを選択することを考えた。

今回、この一手法としてリード文と本記の各文の類似度を測り、これが高い編集を採用する手法を検討した。置換や挿入を行うのであれば、似た文同士の方が適切になる可能性が高いと考えたからである。ただし、この方法はリードと一つの本記文との間の類似度を評価するので、その中の個々の編集の評価はできない。文間類似度が高ければ、そこで起こる編集はすべて採用することになる。

以下では前節の形態素法による編集結果を対象に、文間類似度と評価点の関係を調べた結果を示す。

類似度

類似度の測定手法による違いも検討するため、以下の 3 手法を検討した。

1) 形態素類似度 : リード文と本記文のそれぞれの形態素頻度リスト間で計算した余弦値[1] 。いわゆる”bag of words”の余弦値
2) 形態素アラインメント類似度 : リード文と本記文の間の形態素間で編集距離による照合を行なった後、この 2 文を、共通部分を持つ集合と見なして計算した余弦値（Manning 99）
3) 係り受け類似度 : リード文と本記文の係り受け解析により（係元、係先）の対を抽出する。これらの頻度リスト間で計算した余弦値

類似度の 1）から 3）になるに従って構文的な情報の反映度合いが大きくなっている。

[1] Dice係数とJaccard係数による実験も行ったが結論に大きな差はなかった。

図 4 類似度による編集の選択（形態素）

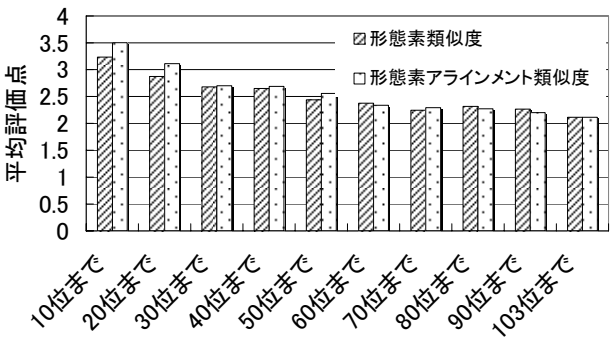
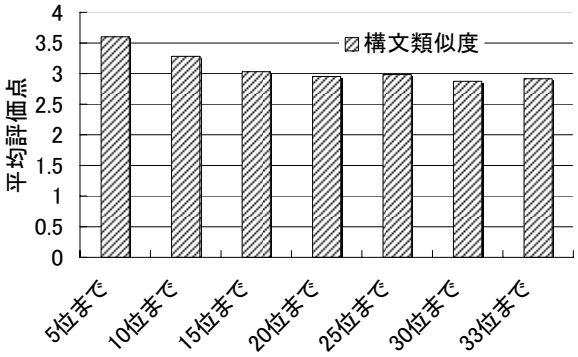


図 5 類似度による編集の選択（構文類似度）



データ処理法

表 1 に記載したとおり、全 23 記事の文数は 136 である。この内、リード文は 23 なので全本記文数は 113 である。形態素法で編集が一箇所でも発生した本記文数は 103 である。これらの本記文と対応するリードの間の類似度を測り、そのペアで発生した平均評価点を計算した。

結果

図 4 にリード文と本記文の間の類似度と評価点の関係を示す。横軸は、リードと本記文の類似度の上位 N（N = 10, 20, …）位までを表し、縦軸はその範囲で得られた平均評価点を表す。
たとえは、類似度が 1 位から 10 位までの、リードと本記文のペアで得られた平均評価値は、形態素類似度では 3.23、形態素アラインメント類似度では 3.50 である[2]。
グラフがほぼ単調減少していることから、類似度の高い順に編集を採用すれば、ある順位までは平均的に許容範囲となる。許容範囲を 3 点以上とすると、形態素類似度では上位 17 位、形態素アラインメント類似度では上位 21 位まで採用できる。単純な ”bag of words” よりもアラインメントを行った方が効果的であった。
さらに、構文類似度の結果を図 5 に示す。横軸が 33 位までになっているのは、34 位以下の文では類似度が 0 になったためである。構文類似度は、（係元、係先）対の一致が起こりにくいのが原因である。一方、1 位から 33 位までの平均評価値を計算すると 2.91 とほぼ許容範囲のままである。他の 2 手法に比べて、類似度を測りにくい問題はあるが、選択の効果は高いことが分かった。

[2] N位までの文の平均を計算するときに個々の編集の平均を取らず文の平均値の平均を取っている。このため全体の平均は表 1 と多少異なっている。

6 議論

形態素法の編集候補に文間類似度の評価を加えることで、ある程度採用できる見込みを得た。ただし上限値は、すなわち平均評価点が高い順に文を選ぶと、50 位程度となる。つまり構文類似度の結果にもまだ改善の余地がある。今後、より高度に構文情報を利用する手法や文脈の利用を検討したい。
本報告は、いわゆるカットアンドペースト操作に基づく要約（Jing 99）の一種であるが、この操作をニュースの特性を利用して、リードと本記に限定したものである。また、文字列照合に基づいた手法は（加藤 99）の要約知識獲得など多数の研究で採用されている。ただし、本稿の目的は文字数の削減ではなく、具体化、すなわち、むしろ増加にある。通常の要約と逆の操作に積極的に応用した点が異なる。さらに、「同一表現抽出」などとも興味深い関連があるが、これらは稿を改めて紹介したい。

7 おわりに

本報告ではニュースのリード文に表現を本記の表現で置換・挿入することで要約を生成する手法を提案し、実験を報告した。単純な文字列照合に基づいた手法であるが、類似した文を利用すると効果的であることを示した。今後は、手法の改良を進めるとともに、リード文や本記の自動認定を含めた全体システムに向けた研究を進める予定である。

参考文献

(Jing 99) Jing, H. and K. R. McKeown: “The Decomposition of Human-Written Summary Sentences”, pp. 129-136, proc.of SIGIR99, (1999)
(加藤 99) 加藤、浦谷: “局所的要約知識の自動獲得法”, 自然言語処理, Vol. 6, No.7, pp.73-92, (1999)
(Manning 99) C. Manning and H. Shütze, “Foundations of Statistical Natural Language Processing”, MIT press (1999)
(田中 05) 田中ほか: “ニュース要約の実態調査と要約モデルの検討” 情処 NL 研資, No.117, pp115-120, (2005)