重複部・冗長部削除による複数記事要約手法

大竹　清敬[†]　船坂　貴浩[†]
増山　繁[†]　山本　和英[††]

複数の関連記事に対する要約手法について述べる．記事の第一段落を用いて，その重複部・冗長部を削除することにより複数の関連記事をどの程度要約できるかを明確にすることを目的とする．さらに，重複部・冗長部を特定，削除する処理をヒューリスティックスにより実現する手法を提案する．まず，新聞記事における推量文の一部は重要度が低いと考えられ，これを文末表現ならびに手掛り語で特定し，削除する．次に，詳細な住所の表現は記事の概要を把握するためには不必要であり，これも削除する．さらに，導入部と呼ぶ部分を定義し，導入部内の名詞と動詞が他記事の文に含まれるならば導入部は重複しているとし，削除する．また，頻繁に出現する人名・地名に関する説明語句，括弧を用いた表現について，他記事との重複を調べる．重複している部分は，1 つを残し他は削除する．提案手法を計算機に実装し，実験を行った．その結果，27 記事群に対して各記事の第一段落を平均要約率 82.1%で要約することができた．さらに，実験結果のうち 6 記事群を用いて評価者 11 人に対してアンケートを行い評価した．アンケートの内容は，要約文章において冗長に感じる箇所，ならびに削除部分を含めた元記事において重要と考えられるが削除されている箇所を指摘する，である．アンケート調査の結果，本手法による要約がおおむね自然であることを確認した．また，本手法によって削除された部分がおおむね妥当であることが明らかになった．

キーワード:　要約，複数文章，重複表現

Multiple Articles Summarization by Deleting Overlapped and Verbose Parts

Kiyonori Ohtake [†] , Takahiro Funasaka [†] , Shigeru Masuyama [†]
and Kazuhide Yamamoto [††]

In this paper, we attempt to summarize multiple Japanese articles into one document. Our summarization method deletes verbose parts and overlapped parts in the input texts. We defined an introduction part to detect overlapped part, and if nouns and verbs in an introduction part were included in some other article, then the introduction part is considered to be an overlapped part. This paper focuses on the following five points to sum up: guess sentences, noun modifiers, expressions using parenthesis, detailed expressions of address and the introduction part. We have implemented a prototype system of summarization and experimented on the system using 27 groups of articles. As a result of the experiments 82.1% compression ratio on the average was achieved. In addition, we evaluated this method using 6 groups of articles by obtaining information by means of questionnaires to 11 examinees. The result of evaluation showed us that the summarizations were almost always natural and deleted parts were also appropriate.

自然言語処理　Vol. 6　No. 3　July 1999

KeyWords: *Summarization, Multiple texts, Overlapped Expressions*

1　はじめに

複数の関連記事に対する要約手法について述べる．近年，新聞記事は機械可読の形でも提供され，容易に検索することができるようになった．その一方で，検索の対象が長期に及ぶ事件などの場合，検索結果が膨大となり，全ての記事に目を通すためには多大な時間を要する．そのため，これら複数の関連記事から要約を自動生成する手法は重要である．そこで，本研究では複数の関連記事を自動要約することを目的とする．

自動要約・抄録に関する研究は古くから存在する (奥村・難波 1998) が，それらの多くは単一の文書を対象としている．要約対象の文書が複数存在し，対象文書間で重複した記述がある場合，単一文書を対象とした要約を各々の文書に適用しただけでは重複した内容を持つ可能性があり，これに対処しなければならない．

対象とする新聞記事は特殊な表現上の構成をもっており (平井 1984)，各記事の見出しを並べると一連の記事の概要をある程度把握することができる．さらに詳細な情報を得るためには，記事の本文に目を通さなければならない．ところが，新聞記事の構成から，各記事の第一段落には記事の要約が記述されていることが多い．これを並べると一連の記事の十分な要約になる可能性がある．しかし，各記事は単独で読まれることを想定して記述されているため，各記事の第一段落の羅列は，重複部分が多くなり，冗長な印象を与えるため読みにくい．そこで，複数の記事を 1 つの対象とし，その中で重複した部分を特定，削除し，要約を生成する必要がある．

本論文で提案する手法は複数関連記事全体から判断して，重要性が低い部分を削除することによって要約を作成する．重要性が低い部分を以下に示す冗長部と重複部の 2 つに分けて考える．なお，本論文で述べる手法が取り扱う具体的な冗長部，重複部は 3節にて説明する．

冗長部：　単一記事内で重要でないと考えられる部分．

重複部：　記事間で重複した内容となっている部分．

従来の単一文書を対象とした削除による要約手法は，換言すると冗長部を削除する手法であるといえる．重複部は，複数文書をまとめて要約する場合に考慮すべき部分である．

本研究において目標とする要約が満たすべき要件は

- それぞれの単一記事において冗長部を含まないこと，
- 記事全体を通して重複部を含まないこと，
- 要約を読むだけで一連の記事の概要を理解できること，
- そのために各記事の要約は時間順に並べられていること，
- ただし，各記事の要約は見出しの羅列より詳しい情報を持つこと，

† 豊橋技術科学大学 知識情報工学系, Department of Knowledge-based Information Engineering, Toyohashi University of Technology

†† ATR 音声翻訳通信研究所, ATR Interpreting Telecommunications Research Laboratories

である．

本研究では，時間順に並べた各記事の第一段落に対して要約手法を適用し，記事全体の要約を生成する．したがって，本手法により生成される要約は，見出しの羅列よりも詳しいが第一段落の羅列よりは短かい要約である．以上により，事件等の出来事に関する一連の流れが読みとれると考える．

具体的な要約例として付録 Aを挙げる．この要約例は本論文の 3節で説明する手法を適用して作成した．この要約例には重複部が多く存在し，それらが本要約手法によって削除された．重複部の削除は，それが正しく特定されている限り適切であると考えることができる．なぜならば，重複部分が既知の情報しか持たず，重要性が低いことは明らかだからである．また，実際の評価においても，要約例 Aについて本手法による削除が不適切とされた部分はなかった．冗長部の特定は重要性の指針を含むことであり，要約に対する視点，要求する要約率などにより変化するので，評価もゆれることが考えられる．これは従来の単一文書に対する要約評価においても同様に問題とされていることである．したがって，付録 Aに挙げた要約例も重複部の削除に関しては妥当であると言えるが，冗長部の削除については，その特定が不十分であり，削除が不適切である部分が存在すると言える．しかしながら，付録 Aに挙げた要約例は，実際のところ，記事の概要を把握するためには十分な要約になっている．評価においても，削除が不適切であると指摘された部分はなく，冗長であると指摘された部分を数ヶ所含んだ要約である．

新聞記事検索時などにおいて，利用者が関連する一連の記事の要約を求めることは，関連記事数が多ければ多いほど頻繁に起こると想定できる．このとき，本研究が目的とする要約によって，関連記事群全体の概要を知ることができれば，次の検索への重要な情報提供が可能となる．また，見出しの羅列のみでは情報量として不十分であるが，第一段落の羅列では文書量が多すぎる場合に，適切な情報を適切な文書量で提供できると考えられる．換言すれば，段階的情報（要約）提示の一部を担うことが可能となる．したがって，本研究において目標とする要約が満たすべき要件として，重複部・冗長部を含まないのみならず，一連の記事を時間順に並べることが挙げられていることは妥当である．

冗長部はどのような記事にも含まれる可能性があるが，重複部は記事の文体によっては特定することが困難となる場合がある．逆に，重複部が存在する場合，複数関連記事要約の観点からそれを削除することは妥当である．一般的に新聞記事の記述の方法から，長い時間経過を伴う一連の関連記事の場合には重複部が多く存在することが予想できる．そのような記事群は一連の事件や政治的出来事に関する場合が多い．また，このような関連記事に対する要約の需要は多く，本論文で示す重複部・冗長部の削除による要約は十分に実用性があると考える．実際に，要約例 Aはある事件について述べられている一連の記事群であるが，これは既に述べた効果を持ち，おおむね本研究の目指す要約であると言える．

本論文では上記の処理がヒューリスティックスにより実現可能であることを示し，そのため

の手法を提案する．そしてこの手法を実装し，評価実験を通して手法の有効性を確認する．

以下では，2節にて本研究に関連する研究について触れ，3節では，本論文で提案する要約手法について述べる．4節では3節で述べた手法を用いて行った実験とアンケート評価について示す．そして，5節で評価結果について議論し，最後に本論文のまとめを示す．

2　関連研究

複数文書を対象とした要約の研究として (Mani and Bloedorn 1997; 山本, 増山, 内藤 1996; McKeown and Radev 1995; 柴田, 上田, 池田 1997, など) がある．

Mani and Bloedorn の手法は，文書をグラフを用いて表現し，活性伝播によって文書の話題と関係する部分グラフを抽出する．そして，複数文書に対して各文書の部分グラフを照合することにより，文書間の類似箇所と相違箇所を抽出し，この結果に基づき文を選択する．

柴田らは形態素の出現頻度を利用し，重複文を同定している．こうして特定した重複文の一方を用いて要約を生成する．

従来は，要約を目標に置きながらも，実際には統計的手法などにより重要文を決定し，その選択結果である抄録を生成する研究が多い．しかし，山本らは表層情報のみを用いたヒューリスティックスによる手法を提案し，文中の重複部分を削除している．山本らは関連した文書を利用して，後続記事の重複部分の除去による要約を研究目的としており，複数の関連記事全体の要約は考慮していない．そのため，本手法とは対象とする問題が異なる．

本手法も山本らと同様にヒューリスティックスのみによる手法であるが，山本らは節単位で比較，削除を行うのに対し，本手法は新聞記事において頻繁にみられる表現を手掛りに重複部分を特定し，削除する．

また，要約生成のために構文解析を用いる手法 (三上, 山崎, 増山, 中川 1998; 三上, 増山, 中川 1999) も存在する．一般に，構文解析器は形態素解析器に比べ解析誤り率が高く，処理時間もより多く必要とする．そのため，構文解析結果を利用して要約生成する場合，処理時間を費やしたにもかかわらず，要約文章が不自然になる可能性がある．したがって，ヒューリスティックスのみで十分な要約が生成できるとする本手法では，処理時間の短縮と構文解析誤りによる影響の排除のため，構文解析器を使用しない．

本手法の特徴は，形態素解析結果のみを用いるヒューリスティックスによって構成されていることである．関連した記事間において，重複・冗長部分の特定にはヒューリスティックスで十分対応可能であり，それによる要約文章が十分に自然であると考えた．そのため，本手法では従来提案されている統計的手法などによる重要部分の特定は行わない．しかし，単一文書を対象として従来提案されてきた要約手法と本手法を組み合わせて用いることにより，さらに文書量が少ない要約を生成できると考える．

3　要約手法

本手法の概要は以下の通りである．新聞記事が表現上の特有の構造を持っていることから，各記事の第一段落を要約の対象とし，記事を時間順に並べる．これにより，利用者が，一連の事件に関する全体像を見出しの羅列以上に容易に把握することが期待できる．冗長部として，推量文を文末表現で特定し，削除する．また，新聞記事において詳細な住所の表現が頻出するが，記事の全体像を把握する上では重要度は低いと考えられるので，これも冗長部として削除する．重複部としては，導入部と呼ぶ部分を定義する．導入部の名詞と動詞がそれ以前の記事の文に含まれるならば，それは既知の情報であるため，削除する．また，頻繁に出現する人名・地名に関する説明語句，括弧による言い換え，それぞれの文，ならびに括弧を用いた説明語句についてそれぞれ記事内と記事間における重複を調べる．重複している部分は，1 つを残し，他は削除する．以上の処理を各関連記事の第一段落を並べたものに適用し，最終的に残った文章が要約文章となる．

3.1　前提条件

本手法は表層情報である形態素解析結果のみを用いて要約を生成する．そのため，各記事をあらかじめ形態素解析しておくことが必要となる．また，本手法が利用される状況として，新聞記事検索の結果に対して利用することを想定する．利用者は各記事の見出しなどを参照しつつ最終的な関連記事を決定し，関連記事群を指定する．この際，一般的に複数記事の要約文書の記述順序なども問題となるが，要約システムは与えられた記事を時間順に整列し，これを記述順序として要約処理を行う．なぜならば，実際に記事が書かれた順番が記述順序であり，こうすることにより一連の記事に関する流れを容易に把握できると考えたからである．要約処理が終了すると，時間順に整列された順番で各記事を出力するものとする．

本手法では，各記事の第一段落に対して，その重複部・冗長部を特定し，削除することによる要約を目指すので，要約率の制御は行わない．

3.2　要約処理手順の概要

本要約手法は，冗長部処理と重複部処理に大別される．冗長部処理は推量文処理と住所表現処理から構成される．重複部処理は導入部処理，括弧の処理，人物・地名の説明処理，重複文処理から構成される．要約手法の処理手順の概略は以下のとおりである．

(1)　各記事を形態素解析する．
(2)　入力された各記事から第 1 段落を抽出し，時間順に整列する．
(3)　推量文処理.
(4)　重複文処理.

(5)　住所表現処理.
(6)　人物・地名の説明処理.
(7)　括弧の処理.
(8)　導入部処理.
(9)　その他の処理.
(10)　結果を出力する.

以下，順に各処理について説明する.

3.3　推量文処理

推量文とは，記述した

コトと現実とに不一致があり得ることを示す (城田 1998)

文である．文末が推量表現である文は，記述内容と事実とに不一致があり得ることを示している．したがって，新聞記事を対象とする場合においては，推量文を要約文書に含める重要性は低いと言える．そのため，まず，各文の文末表現を判断し，推量表現であれば削除対象としてその 1 文全てを仮冗長部とする．この削除対象と判断する文末表現を表 1に列挙する．これらの文末表現は，日本経済新聞の 1990 年から 1992 年の記事を参考に収集した．そして，以下の条件を全て満たした場合に限り仮冗長部を削除する.

(1)　「~ によると」,「~ ため」という根拠を示す表現を含まない.
(2)　「だが」,「しかし」,「が」といった逆接の接続詞あるいは接続助詞を含まない.
(3)　「ただ」,「ものの」といった条件，譲歩を示す表現を含まない.

表 1　文の削除対象とする文末表現

~ 可能性が大きい	~ 可能性もある	~ 可能性も出てきた	~ そう
~ 情勢だ	~ かもしれない	~ そうだ	~ ちがいない
~ という	~ と思われる	~ はずだ	~ 微妙
~ 微妙だ	~ 微妙である	~ 見通し	~ 見通しだ
~ 見通しである	~ みられる	~ 見られている	~ 模様だ
~ よう	~ ようだ	~ 予想される	~ らしい
~ ろう	~ そうもない		

3.4　重複文処理

新聞記事では極く稀にほぼ同一内容の文が出現する場合がある．特に記事が掲載されてから時間経過が大きいほど，その可能性が大きくなる．一例を以下に示す.

[日本経済新聞 1993/1/26 から抜粋]

米国が日本と共同開発中のＦＳＸの関連技術を入手できると定めた政府間合意に基づき，技術に付随する「試供品」としてハードウエアを輸出する．日本企業が独自開発した本格的な軍用機材を対米供与する初の事例になる．米空軍は将来の戦闘機開発で同レーダーの採用を検討するとみられ，同社は引き合いがあれば積極的に対応する方針だ．

[日本経済新聞 1993/8/4 から抜粋]

米国が日米共同開発中のＦＳＸの関連武器技術を入手できると定めた政府間合意に基づき，技術に付随する参考品の形で供与した．日本企業が独自開発した本格的な軍用機材を対米供与する初の事例となる．

これらの 2 記事間には全く同一内容の文があり，重複しているため削除する．このような重複部分を特定する手法は複数可能である．本手法では次に示す処理を行い，同一内容である重複文を削除する．文 $S1$ と $S2$ が与えられたとき，$S1$ に含まれる名詞数を $n(S1)$ とし，同様に $S2$ に含まれる名詞数を $n(S2)$ とし，さらに，$S1$ と $S2$ に共通に含まれる名詞の数を $m(S1, S2)$ として

$$\frac{m(S1,S2)}{n(S1)} > \alpha \quad \text{かつ} \quad \frac{m(S1,S2)}{n(S2)} > \alpha$$

という条件が成立する場合，$S2$ を削除する．本手法では後述の理由により $\alpha = 0.8$ とする．

本手法では重複文処理として，以上の処理を一連の記事内の第一文を除く全ての文の組み合わせに対して行う簡便な手法をとった．この処理では，ほとんど同一内容の文のみを削除する．一文が複数の文に分かれるなどの理由により，$S1$ が $S2$ に含まれる場合は，$S1$ を削除すべきと考えられるが，削除後の記事が不自然となる場合があるので，本論文では上述の手法とした．この手法を先の例に適用すると第 2 記事の「日本企業が独自開発した本格的な軍用機材を対米供与する初の事例となる.」が削除される．また，第 1 記事の「米国が日本と~ 」と第 2 記事の「米国が日米共同開発中の~ 」はほぼ同一内容であると判断できるが,「試供品」→「参考品」などの違いがあり，本手法では削除されない．α の値を小さくすることにより，削除も可能となるが，削除すべきでない文も削除する場合があることが観察されたので，本手法では，$\alpha = 0.8$ としている．

3.5　住所表現処理

新聞記事においては詳細な住所の表現が頻出する．例えば

[日本経済新聞 1992/2/13 より抜粋]

放棄させていた北海道○○町△町一二三，廃品回収業

[日本経済新聞 1990/1/26 より抜粋]

東京都江戸川区西葛西七丁目の都道交差点で

などの表現がある．この場合，記事の概要を把握する上では詳細な住所の記述は冗長である．このような住所表現をパターンマッチングで特定し，削除する．ただし，住所表現の次の形態素が読点の場合はその住所表現全体および後続の読点を削除するが，住所表現の次の形態素が「の」の場合は住所表現の先頭から「都，道，府，県，管内，市」のいずれかの形態素までを残し，それ以降の住所表現を削除する．上の例では，それぞれの部分を削除すると

七千万円の債権を放棄させていた廃品回収業

東京都の都道交差点で

となる．

3.6　人名・地名の説明処理

新聞記事では，次のように人名の前後でその人物について説明している部分がある．

[日本経済新聞 1992/2/20 より抜粋]

前道議の ○○○○容疑者（５６）＝渡島管内△△町△△△＝

[日本経済新聞 1991/4/9 より抜粋]

早大三年, ○○○○さん（２０）ら四人は

複数の関連記事を並べた場合，このような部分は頻繁に出現する．このことから，同一の人名が 2 度以上出現する時，それらの説明の部分は冗長となり，要約文章に含める重要性が低いため削除する．

説明の部分とは，人名にかかる連体修飾語句と，人名の後方にある括弧を用いた年齢の表現，ならびに，＝で囲まれた部分である．連体修飾語句は人名の形態素から前方へ名詞，または助詞「の」またはその前も名詞である「,」の続く限りたどり，到達できる部分によって特定する．本手法ではこれらの特定した説明部分を削除する．

また，地名の連体修飾語句も同様に削除する．例を以下に示す．

施行主体の 広島市は　[日本経済新聞 1991/3/15 より抜粋]

ただし，連体修飾語句の認定が人名の場合と異なり，地名の形態素から前方へ，普通名詞の続く限りたどり，到達できる部分までとする．また，助詞「の」が地名の直前に存在し，「名詞の地名」となる形式の場合，1 度目の出現時に「名詞」を登録しておく．そして，同一地名について再び同一の名詞を伴なって「名詞の地名」という形式が出現した場合，「名詞の」を削除し，「地名」とする．

3.7　括弧の処理

新聞記事に出現する括弧を用いた表現のうち，括弧の処理では，以下の形式を扱う．

A（B）

このような表現があった場合，以下のいずれかの表現に置き換えることが可能である．

(1)　　B
(2)　　A
(3)　　A（B）

(1) は語句の言い換えの場合で，括弧の前の語句が長いほど頻繁に起こる．例えば，複数の記事を並べた中に「石油輸出国機構（ＯＰＥＣ）」が 2 度以上出現する場合がある．このような場合，括弧の前の語は冗長であり，2 回目以降は以下のように置き換えることが可能である．

石油輸出国機構（ＯＰＥＣ） → ＯＰＥＣ

この置き換えが適用できる条件は

- 括弧内の語がその前の語より短かい，かつ
- 括弧内の語が 1 形態素からなる

である．以上の 2 つの条件を満たす場合についてのみ，括弧の前の語句を括弧内の語句へ言い換える．

次に，(2) は括弧の中の語句が付加的な情報の場合で，例えば,「ダイエー（本社神戸市）」が，複数の記事を並べた中に 2 度以上出現する場合，括弧内の語句による説明は冗長なので，以下のように置き換える．

ダイエー（本社神戸市） → ダイエー

この置き換えが適用できる条件は

- 括弧内の語が前に出現した括弧内の語と完全に一致している，または
- 括弧内の語が前に出現した括弧内の語に連続した文字列として含まれる

である．以上のいずれかの条件をみたす場合は，括弧内の語句を括弧と共に削除する．

(3) は (1) および (2) の適用条件に合致しない場合であり，この場合は特に削除を行わない．具体的には（B）がAと無関係と考えられる場合に相当する．例えば,

した.（関連記事１面に）

などがこれに該当する．

また，複数記事を要約する場合において，既に説明した形式とは異なる括弧表現のうち，要約文章に含める重要性が低いものがある．例えば，記事冒頭の記者に関する表現と，記事最後の関連記事参照のための表現がある．記者に関する情報は一連の関連記事の大意を把握するための情報としては重要性が低い．また，関連記事参照のための情報は，紙面という媒体の場合に有効な情報であり，利用者が選択した記事群を要約する場合には，重要性は低いと考えた．具体例を以下に示す．

【パリ７日＝○○△△】
（関連記事１面に）

などがこれに該当し，本手法ではこれらの表現を削除する．

3.8　導入部の処理

複数の関連記事を要約する場合，各々の記事は単独に読まれることを想定しているため，記事の前提条件，あるいは，時間経過に関する記述がある．そのため，単一文書の要約手法を各記事に適用し，それらの文書を並べるだけでは，要約文書として冗長な部分を数多く残してしまう可能性がある．そこで，記事の前提条件および時間経過に関する部分は，一連の記事に 1 つ存在すれば十分であるため，それらの重複部分を特定し，2 回目以降を削除する．

時間順に並べた一連の記事 $A_1, A_2, \cdots, A_n$ において，古い記事で記述された内容を新しい記事で再び記述している部分がある．このような部分は，単独の記事としてであれば，経過を知るためなどから必要である．しかし一連の関連記事を要約する立場から考えた場合，それらはすでに既知の事実であり，重複した内容を持つ．このような部分は一般に記事の冒頭にみられ，本研究では導入部と定義する．導入部の定義を以下に示す．

(1)　記事の第 1 文に存在するものとする．

(2)　文頭から次の表現[1]までの部分

~ したが，~ 問題で，~ 事件で，~ 事故で，

~ していたが，~ について

ならびに，「名詞+は」の前までの部分のうち最初に出現する表現までとする．

(3)　該当部分が存在しない記事には，導入部が存在しないものとする．

時間順に並べた一連の記事 $A_1, A_2, \cdots, A_n$ において，記事 $A_i\,(i \geq 2)$ の導入部は冗長である場合がある．なぜなら，記事 $A_1, A_2, \cdots, A_{i-1}$ において導入部と同一の内容が既に述べられている可能性があるからである．

本手法において，導入部が重複していると判断するための条件を以下に示す．

- 「~ 事件で」，「~ 事故で」を含む導入部の場合，導入部に含まれる名詞，動詞の終止形のうち 3 割以上が $A_1, A_2, \cdots, A_{i-1}$ 中のある文に含まれる．
- それ以外の導入部の場合，導入部に含まれる名詞，動詞の終止形のうち 6 割以上が $A_1, A_2, \cdots, A_{i-1}$ 中のある文に含まれる．

以上の条件を満たす導入部が削除される．ただし，「名詞+は」の部分は文の主題を示しており，削除しない．この「名詞+は」の「名詞」の部分は，「は」の直前の形態素から名詞，読点，なかぐろ点（・），接頭辞，接尾辞，および「名詞+と」のいずれかであるかぎり前方へたどり，特定する．また，導入部に含まれる括弧内の形態素ならびに導入部内の数詞は考慮しない．

上記の条件は，記事間の関連性が十分であると仮定できる場合は，「~ 事件で」，「~ 事故で」を含む導入部は名詞の一致度合が低くとも削除すべき場合が多いという事実に基づき割合を小さくした．一般に名詞の一致度合が低い場合，他の記事とは異なる名詞を用いて同一事象を指していることが多い．また，「~ 事件で」，「~ 事故で」以外の表現を含む導入部では形態素の一

1 これらの表現は日本経済新聞 1990 年から 1992 年の記事を参考に収集した

致割合を 6 割以上としている．その理由は，予備実験として割合を徐々に小さくする実験を行い，6 割より小さくなると，不自然になる場合が観察されたためである．

導入部処理は次のアルゴリズムに従う．ただし，ここで $A_1, A_2, \cdots, A_n$ は既に時間順に並べられているとする．

入力：記事 $A_1, A_2, \cdots, A_n$
出力：記事 $A_1, A_2, \cdots, A_n$
Step 1　$i := n$
Step 2　i が 1 より大きければ Step 2.1~ 2.4 を繰り返す
Step 2.1　記事 A_i の導入部の特定
Step 2.2　導入部がなければ Step 2.4 へ
Step 2.3　記事 $A_1, A_2, \cdots, A_{i-1}$ 内の各文と記事 A_i の導入部を比較し，削除のための条件を満足するならば，その導入部を削除
Step 2.4　$i := i - 1$，Step 2 へ
Step 3　終了

3.9　その他の処理

その他の処理では，以上の要約処理に含まれない処理を行う．処理の内容を以下に示す．

(1)　各記事の最後に「解説~ 面に」があればこれを削除する．

(2)　各記事を先頭から調べて行き，「数詞+日」が出現せずに「同日」という表現が出現した場合，それ以前に出現した「数詞+日」が削除されている．そのため，「同日」から前方にたどり，既に削除された部分の中から一番最初に出現する「数詞+日」を「同日」と置換する．

(2) のような省略の回復は他にも考えられるが，本手法では重要である日付の情報のみを扱う．

4　評価実験

本論文で提案する手法を計算機上に実装し，アンケートによる評価を行った．まず，本手法を CPU：PentiumII 300MHz，メモリ：128MB の PC/AT 互換機上に Perl 言語を用いて実装した．形態素解析器には形態素解析システム JUMAN3.5 を使用した．JUMAN の辞書ならびに設定には変更を加えず，システムの既定値のまま使用した．また，形態素解析に誤りが含まれていた場合にも，解析結果を修正せずに用いた．

実験には日本経済新聞の 1990 年と 1992 年の記事を使用した．なお，本研究の手法を構築するために 1990 年から 1992 年の記事を参考にしているが，参考に使用した記事とは異なる関連

記事を新たに抽出した．関連記事群はあらかじめ 27 記事群を抽出しておいた（平均記事数 4.7，最大 9，最小 3）．

4.1　実験結果

あらかじめ形態素解析しておいた記事を入力として実験した結果，各記事群を要約する時間は平均 0.8 秒（形態素解析時間を除く）であり，平均要約率は 82.1%であった．

4.2　評価方法

自然言語処理システムにおける評価の問題は機械翻訳の分野で古くから扱われている (King 1996)．自動要約の分野では，単一文書を対象とした研究の多くが，人間が生成した要約文章と自動要約結果を比較し，再現率および適合率を評価尺度とした評価が主に行われてきた (奥村・難波 1998)．

しかし，人間が生成した要約文章も，要約を行う人間の視点などによって要約結果が大きく異なることが十分に考えられる．つまり，原文に対して要約が唯一存在するわけではない．しかしながら，複数の人間が削除による要約を行った場合に，そのうちの大部分の人間が削除する部分を考えることはできる．そこで，自動要約結果の評価を複数の人間によって行うことは自然である．山本らは 18 人の被験者に対してアンケートを行っている (山本, 増山, 内藤 1995)．山本らのアンケートは要約結果全体に対して，その自然さ，内容の適切さ，および修飾句省略の適切さを問うものである．しかし，これでは適用した要約技法のうち，どれが有効なのかがわかりにくい．

本研究では要約文章と原文を比較し，(1) 要約文章中で削除すべき箇所，(2) 要約システムが削除した部分において削除すべきではない箇所を評価者に自由に指摘させるアンケートを行った．

4.3　アンケート調査

アンケートに使用した記事群は，実験のために用意した 27 記事群のうち要約率が 90%以下の記事群（21 記事群）から任意に 6 記事群を選択した．選択した記事群の概要を付録 Bに示す．選択した記事群の平均要約率は 74.5%である（最小：56.0%，最大：83.1%）．

調査対象としたのは大学工学部学生 11 人である．アンケート実施の前に，本手法の概要および想定状況等を説明し，元記事も第一段落のみであることを説明した．想定状況として

- 関連記事検索結果に対して使用すること，
- 記事は利用者が見出しなどをもとに決定すること，
- そのため，記事間の関連性は十分だと仮定できること，
- 利用者は見出しなどにより記事の大雑把な概要はある程度把握していること，よって，さ

らに詳細な情報を得るために要約システムを使用するということ，

- 要約システムは記事の概要，すなわち見出しよりも詳しいが第一段落よりも短かいものを出力することを目的としていること

を説明した．また，本手法が重複部・冗長部を特定し，それらを削除することによって要約を行うことも説明した．

アンケートへは記事の日付，見出し，文章（要約/原文）を出力してある．調査は，(1) 要約結果のうち，さらに削除すべきだと思う部分を自由に指摘する．(2) 計算機が削除した部分のうち，削除すべきではないと思う部分を自由に指摘する．以上を 6 記事群それぞれについて自由に行う形式で調査を行った．評価にかける時間を限定せず自由に評価させた．

4.4 アンケート結果

指摘部分は完全に一致した場合のみを数えあげた．例えば，C_1 C_2 C_3 C_4 C_5 という文字列に対して，評価者 A が C_2 から C_4 を指摘し，評価者 B が C_1 から C_5 を指摘した場合，C_2から C_4 を共有していることになるが，両者は異なる部分として数え上げた．

まず，アンケートに用いた記事群に対して適用した本手法の内訳を以下に示す（() 内は総数に対する割合）[2]．

- 推量文処理：4 箇所（6.9%）
- 住所表現処理：15 箇所（25.9%）
- 人名・地名の説明処理：11 箇所（19.0%）
- 括弧の処理：9 箇所（15.5%）
- 導入部の処理：19 箇所（32.8%）

総数：58

[冗長さの指摘]

指摘箇所総数：101 箇所

内訳：

重複人数	1	2	3	4	5	6	7
頻度	56	16	11	6	6	3	3

平均重複人数：2.1

重複人数とは何人の評価者が同一の部分を指摘したかを示し，頻度は指摘された箇所の数を示す．

[削除不適切の指摘]

指摘箇所総数：13

2 今回の実験では重複文処理による削除は含まれていなかった

内訳：

重複人数	1	2	3
頻度	8	3	2

平均重複人数：1.5

指摘された 13 箇所の内訳（() 内はその重複人数）

- 推量文処理：4 箇所 (3) (2) (1) (2)
- 住所表現処理：2 箇所 (1) (1)
- 導入部処理：7 箇所 (1) (2) (3) (1) (1) (1) (1)

5　議論

5.1　手法の妥当性

まず，冗長さの指摘から考察する．ここで，重複人数が評価者の半数以上である 6 箇所についてみてみると，

(1)　記事冒頭の「~ とされた」までの部分 (1 箇所)
(2)　記事冒頭の「悪質な犯行で大きな社会的関心を呼んだ」という部分 (1 箇所)
(3)　他の記事の文と部分的に一致している部分 (1 箇所)
(4)　形態素解析誤りにより，人名の処理ができなかった部分 (1 箇所)
(5)　括弧の処理に関する部分 (2 箇所)

という構成になっている．

(1) は導入部として考慮すべき表現である．なぜならば,「~ とされた」はすでに過去にあった事実を述べており，導入部としての資格を十分に持っている．

(2) はいわゆる連体修飾節に該当し，表層情報のみを用いる本手法で対応することは困難である．

(3) は 3.4節にて述べたように，本手法では文の一部の重複に対して対応していないため，当然の結果である．これに表層情報のみを用いる手法で対応するためには山本らが提案する節照合処理 (山本他 1996) などが適用できる．

(4) は形態素解析結果において人名となるべき形態素が地名と解析されたために生じたものであり，現状の形態素解析器では頻出する誤りである．これを避けるために「地名+容疑者」 や「従業員，地名~ 」という明らかに人名を示す表現がその形態素の近隣にある場合には人名として処理するヒューリスティックスを適用すればよい．

(5) は該当箇所が 2 箇所あるが，これは 2 種類にわけることができる．1 つは括弧の中に括弧表現がある場合であり，システムがこのような入れ子の括弧表現に対応していなかったために生じた．これは直ちに対応可能である．もう一方は「関税貿易一般協定・多角的貿易交渉（ガッ

ト・ウルグアイ・ラウンド)」と「ガット・ウルグアイ・ラウンド（関税貿易一般協定・多角的貿易交渉)」と表現されていた場合に指摘されている．これは現状の手法では対処できないが，このような表現は言い換えに相当するため，より短かい表現である「ガット・ウルグアイ・ラウンド」に統一すべきであり，この実現は容易である．以上から，(2) と (3) を除いて，容易に修正できることがわかる．

次に削除不適切の指摘について検討する．以下に，削除不適切として指摘された 13 箇所すべてが，いずれの処理によるものかを示す．() 内の数字は重複人数を示す．

(1)　導入部処理「~ について」1 箇所（1）
(2)　導入部処理「名詞+は」 6 箇所（1）（1）（1）（1）（2）（3）
(3)　住所表現処理 2 箇所（1）（1）
(4)　推量文「~ ようだ」 1 箇所（3）
(5)　推量文「~ なろう」 2 箇所（2）（1）
(6)　推量文「~ そうだ」 1 箇所（2）

(1) は 1 箇所を 1 人が指摘している．これは「~ について」という部分が削除されたために，どのような話題についての記事なのかがわかりにくく感じた結果，削除不適切であると指摘したと推察する．また，アンケートにおける想定状況が十分に伝わっていなかったために，そのように感じた可能性もある．

(2) は 6 箇所について指摘されているが，関係する導入部は 2 箇所（以下，a., b. とする）であり，それぞれ，その中で 3 箇所ずつ指摘されている．a. は「名詞+は」の直前に「~ による」があり,「名詞+による」が欠如したため不自然に感じたと推察される．ここを指摘した人は 1 人であるが，必要性が十分に認められるならば，導入部に含まれる「名詞+による」を残して削除するようにすることは容易である．次に b. は「名詞+は」の直前に「名詞+の」があり,「名詞+は」のみでは情報が不十分であるとして指摘したと推察される．対処法として,「名詞+は」の直前に「名詞+の」がある場合は，これも含めて残すことがあげられる．

(3) の住所表現処理は一連の記事の冒頭の記事であり，住所表現のみを手掛りに削除したため不適切と感じたようである．この住所表現内には見出しに含まれる地名が存在し，削除すると不自然になる．したがって，一連の記事の冒頭記事において，住所表現内の形態素が見出しに含まれる場合はその住所表現を残すべきである．

(4) の推量文処理は「超伝導超大型粒子加速器（ＳＳＣ)」という固有名詞が含まれており，推量文の形式としては削除に該当する．しかし，これを削除しないためには，従来の単一文書に対する要約の際に用いられてきた文に対する重要度などを用いて，このような固有名詞を含む文は重要であると評価するなどの対応が必要である．また，この推量文は一連の記事の冒頭記事に含まれ，それ以降この記述に関連した記述がないため，削除するのは不適当であるとの指摘もあった．

(5) の推量文は 2 箇所指摘されているが，推量文としてシステムが削除した部分は 1 箇所のみである．この推量文は「~ おり，…なろう」という形式であり,「~ おり,」までの前半を指摘した人が 2 人，後半の「…なろう」までを含めて推量文を指摘した人が 1 人となっている．確かに前半の「~ おり」は事実を伝えているため，(4) と同様に重要性を評価する指針とあわせて今後検討する必要がある．

(6) の推量文処理は一連の記事の最終記事の最後の文を削除したものであり，一連の記事の最後の推量文は残すべきだとの意見があった．

以上から推量文処理は，文の形式からの判断はおおむね妥当であり，これ以上の改善のためには，文の重要度を判断する基準を導入する必要があると考える．

また，住所表現は冗長部であるため，評価者によって評価にゆれが生じる可能性がある．だが，今回の評価実験結果より複数関連記事の概要を把握する上では，住所表現の重要性はほとんど存在しないと結論づけられる．

以上の評価実験結果から，本論文で提案した手法はおおむね妥当である．

5.2　要約率と記事の関係

本手法を適用して得た記事群の要約率は，元の記事群に含まれる重複部・冗長部の割合に依存する．要約率が大きい記事は，本論文で定義した重複部・冗長部を含まない記事であり，それ以上に要約が困難な記事である．逆に，要約率が小さい記事は，重複部・冗長部を多く含む記事であり，それらが本手法により削除される．評価実験結果から，本手法が重要性の高い情報を削除せず重要性の低い情報のみを削除するという要約の要件を満たしていると言える．そのため，記事が重複部・冗長部を多く含む，含まないによらず，本手法によって妥当な要約を行うことができる．

また，どのような記事に対しても本手法を適用することは可能である．しかし，その要約率は既に述べたように，記事の内容により大きく変化する．これは，本手法の限界を意味するが，重複部削除という本手法の目指す処理は正しく機能している．重複部・冗長部を多く含む記事として，特定の事件・事故に関する記事を挙げることができる．評価実験においても，ある事件の記事群に対して 55%程度の要約率を達成することができた（付録 A）．逆に関連記事群によっては，重複部・冗長部をほとんど含まない記事群があり，要約率は 90%台後半にとどまる．評価実験において，我々が抽出した記事群の中に総務庁が毎月発表する完全失業者数に関する記事群があり，それに該当する．しかしながら，一般に，特定の事件・事故などについてその経過，概要などを求めることは比較的多いと予想できる．したがって，要約の要求が存在する記事の多くに対して，本手法は有効に機能すると言える．

6　おわりに

新聞の関連複数記事を 1 つの文書へと要約するために重複部・冗長部を削除する手法を提案し，実験を行った．その結果，新聞記事は重複部・冗長部の削除によって，要約率 80%程度に要約可能であることがわかった．また，アンケートによる評価の結果，本手法による削除はおおむね自然であり，本手法が削除する箇所はおおむね妥当であることがわかった．

また，重複部・冗長部の削除処理はヒューリスティックスで実現可能であり，その多くは本論文で提案した手法によって実現される．評価実験において，対処が困難な推量文表現も明らかになったが，これらに対しては従来用いられてきた重要文に関する指針 (奥村・難波 1998) が利用可能であると考える．重要文に関する指針をどのように本手法に反映させるかは今後の課題である．

謝辞 日本経済新聞の記事について，本論文への引用許可を頂いた（株）日本経済新聞社に深謝する．また，本研究の一部は文部省科学研究費特定領域研究 B（2）および（財）国際コミュニケーション基金の援助を受けて行った．

参考文献

平井昌夫 (1984). 何でもわかる文章の百科辞典. 三省堂.

King, M. (1996). “Evaluating Natural Language Processing System.” *Communications of the ACM*, **39** (1), 73–79.

Mani, I. and Bloedorn, E. (1997). “Multi-document Summarization by Graph Search and Matching.” In *Proceedings of the 14th National Conference on Artificial Intelligence*, pp. 622–628.

McKeown, K. and Radev, D. R. (1995). “Generating Summaries of Multiple News Articles.” In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 74–82.

三上真, 増山繁, 中川聖一 (1999). “ニュース番組における字幕生成のための文内短縮による要約.” 自然言語処理 特集号 掲載決定.

三上真, 山崎邦子, 増山繁, 中川聖一 (1998). “文中の重要部抽出と言い替えを併用した聴覚障害者用字幕生成のためのニュース文要約.” 言語処理学会 第 4 回 年次大会ワークショップ論文集, pp. 14–21.

奥村学 難波英嗣 (1998). “テキスト自動要約技術の現状と課題.” Research report IS-RR-98-0010I, 北陸先端技術科学大学院大学情報科学研究科 (http://www.jaist.ac.jp/~oku/okumura-j.html).

柴田昇吾, 上田隆也, 池田裕治 (1997). “複数文章の融合.” 情報処理学会研究報告 97-NL-120, pp. 77–82.

城田俊 (1998). 日本語形態論. ひつじ書房.

山本和英, 増山繁, 内藤昭三 (1995). “文章内構造を複合的に利用した論説文要約システム GREEN.” 自然言語処理, **2** (1), 39–55.

山本和英, 増山繁, 内藤昭三 (1996). “関連テキストを利用した重複表現削減による要約.” 電子情報通信学会論文誌, **J79-D-II** (11), 1968–1972.

付録

A　要約例

以下に 3節にて説明した手法による要約例を示す. 網かけ（　　）をしてある部分が要約システムによって削除された部分である. この要約の要約率は 56.0%である.

(1)　記事 1 [92 年 2 月 13 日]

見出し: 北海道警,「○○○○」から４億１０００万，恐喝の会社社長ら逮捕.

北海道警捜査四課と豊平署は十二日，廃バッテリー回収設備の工事をめぐって九〇年四月と十一月ごろ東証一部上場の化学会社○○○○（本社・札幌市，○○△社長）から現金三億四千万円を脅し取ったり，七千万円の債権を放棄させていた 北海道 ○○ 町 □町 一三四 ，廃品回収業「□□」社長，□□▼▼容疑者（５１）と ○○ 市 △ △町 八 ，無職●▽▽容疑者（５５）の二人を恐喝容疑で逮捕した.

(2)　記事 2 [92 年 2 月 18 日]

見出し: ○○○○恐喝事件で道警，■■前道議を逮捕.

東 証 一部 上場 の 高圧 ガス ， 産業 機器 メーカー 「 ○○○○ 」 （ 本社 札幌 市 ） 恐喝 事件 で 道警捜査四課と札幌・豊平署は十七日，恐喝の疑いで新たに ○○ 管内 ○○ 町 □□ 二 ，前道議■■◎◎容疑者（５６）を逮捕した. 同事件の逮捕者は三人となった.

(3)　記事 3 [92 年 2 月 20 日]

見出し: ■■前道議を送検，○○○○恐喝事件.

東 証 一部 上場 の 化学 会社 「 ○ ○○○ 」 （ 本社 札幌 市 ， 社長 ○○ △ 氏 ） が 廃 バッテリー 回収 設備 の 工事 を めぐって 現金 約 三億四千万 円 を 脅し取ら れたり 工事 代金 の 債権 （ 約 七千万 円 ） を 放棄 さ せ られて いた 事件 で 道警捜査四課と札幌豊平署は十九日，恐喝容疑で逮捕した 前 道議 の ■■◎◎容疑者 （ ５６ ） ＝ ○○ 管内 ○○ 町 □□ 二 ＝ を札幌地検に送検した. 同課は○○○○恐喝の中で■■容疑者が果たし

た役割などを本格的に追及，事件の解明を急ぐ.

(4)　記事 4 [92 年 2 月 23 日]

見出し: ○○○○恐喝，新たに会社社長逮捕.

東 証 一部 上場 の 化学 メーカー ，「○○○○ 」恐喝 事件 で 道警捜査四課と札幌・豊平署は二十二日，新たに ○○ 市 △△ 三 丁目 ，会社社長，●●□□容疑者（４３）を恐喝の疑いで逮捕した．同事件の逮捕者はこれで四人目となった.

(5)　記事 5 [92 年 3 月 5 日]

見出し: ○○○○恐喝，前道議ら３人起訴――札幌地検，余罪裏付け急ぐ.

東 証 一部 上場 の 高圧 ガス ，産業 機器 メーカー 「○ ○○○ 」（本社 札幌 市 ） 恐喝 事件 で 札幌地検は四日，恐喝罪で 北海道 □□ 郡 ○○ 町 △△ 二 ，前 北海道 議会 議員 ，■■◎◎ （５６ ），同 郡 ○○ 町 □□ □ 八四 ，廃品 回収 業 ，□□▼▼ （５１ ），○○ 市 △ △町 八 ノ 八 ，無職 ，●▽▽ （５５ ）の三容疑者を起訴した.

(6)　記事 6 [92 年 4 月 8 日]

見出し: ○○○○恐喝で札幌地検，●被告を追起訴.

東 証 一部 上場 の 高圧 ガス ，産業 機器 メーカー 「○○○○ 」（本社 札幌 市 ） 恐喝 事件 で 札幌地検は七日，先に恐喝罪で起訴していた ○○ 市 □ □町 八 ノ 八 ，無職 ●▽▽被告 （５５ ）を同罪で追起訴し，事件の捜査を終了した．被害総額は約七億四千万円となった.

B　評価実験のアンケートで用いた 6 記事群の概要

(1)　「○○○○」恐喝事件

構成記事数：6，要約率：56.0%

(2)　マドンナ写真集，税関で審査

構成記事数：4，要約率：77.5%

(3)　フィリピン・ミンダナオ島で国軍の一部将兵が反乱

構成記事数：5，要約率：82.5%

(4)　東京都の母子ひき逃げ事件

構成記事数：5，要約率：80.4%

(5)　大阪，奈良での連続放火事件

構成記事数：6，要約率：68.3%

(6)　日米首脳会談

構成記事数：7，要約率：83.1%

略歴

大竹 清敬: 1998 年 豊橋技術科学大学大学院修士課程修了．現在，同大学大学院 博士後期課程電子・情報工学専攻に在学中．自然言語処理，特に情報検索，自動要約の研究に従事．言語処理学会，情報処理学会，人工知能学会，各学生会員

船坂 貴浩: 1998 年 豊橋技術科学大学大学院 修士課程知識情報工学専攻 修了

増山 繁: 1977 年 京都大学工学部数理工学科卒業．1982 年 同大学院博士後期課程単位取得退学．1983 年 同修了 (工学博士)．1982 年 日本学術振興会奨励研究員．1984 年 京都大学工学部数理工学科助手．1989 年 豊橋技術科学大学知識情報工学系講師，1990 年 同助教授，1997 年 同教授．アルゴリズム工学，特に，並列グラフアルゴリズム等，自然言語処理，特に，テキスト自動要約等の研究に従事．言語処理学会，電子情報通信学会，情報処理学会等会員．

山本 和英: 1996 年豊橋技術科学大学大学院博士後期課程システム情報工学専攻修了．博士 (工学)．同年より ATR 音声翻訳通信研究所客員研究員，現在に至る．1998 年中国科学院自動化研究所国外訪問学者．要約処理，機械翻訳，韓国語及び中国語処理の研究に従事．1995 年 NLPRS'95 Best Paper Awards．情報処理学会，ACL 各会員．

(1998 年 9 月 30 日 受付)
(1998 年 12 月 17 日 再受付)
(1999 年 2 月 19 日 採録)