自然言語処理　114－7
（1996. 7. 18）

冗長度削減による関連新聞記事の要約

船坂 貴浩†，山本 和英‡，増山 繁†

†豊橋技術科学大学 知識情報工学系
‡*ATR*音声翻訳通信研究所

豊橋技術科学大学 知識情報工学系 増山研究室
〒441 愛知県豊橋市天伯町字雲雀ヶ丘 1-1
Tel. 0532-47-0111 (内線 5737)
E-mail: funasaka@smlab.tutkie.tut.ac.jp

あらまし

近年，新聞記事は機械可読の形でも提供されるようになり，容易に検索を行なうことができるようになった．しかし，検索の対象が長期に及ぶ事件や政治問題などの場合，検索の結果数多くの記事が現れ，それらすべてに目を通すには多大な時間を要する．そこで，これらの記事をまとめて要約する技法を開発することは極めて有用である．検索された記事は関連したものであるから，ある記事で述べられていることを別の記事で述べていることが多い．また，新聞記事には事実文と推量文があるが，前者の方が重要である．本研究では，関連新聞記事を冗長な部分と推量文を削除することにより要約する手法を提案する．

キーワード　要約，新聞記事，冗長

Relevant Newspaper Articles Summarization by Redundancy Reduction

Takahiro Funasaka†, Kazuhide Yamamoto‡, Shigeru Masuyama†

† Department of Knowledge-based Information Engineering, Toyohashi University of Technology
‡ ATR Interpreting Telecommunication Laboratories

Department of Knowledge-based Information Engineering, Toyohashi University of Technology,
Toyohashi-shi, 441 Japan
Tel. 0532-47-0111 (Ext. 5737)
E-mail: funasaka@smlab.tutkie.tut.ac.jp

Abstract

In recent years, we can retrieve newspaper articles easily, because they are provided in machine readable forms. However, if we retrieve articles on long term cases and political issues, many articles are retrieved and it takes much time to read them. Thus developing relevant articles summarization method is very useful. In this study, we propose a method of summarizing relevant newspaper articles by eliminating redundancy parts and guess sentences based on the observations that fact sentences are more important than guess sentences and that relevant articles have redundant parts.

key words　summarization, newspaper articles, redundancy reduction

1 はじめに

近年，新聞記事は機械可読の形でも提供されるようになり，容易に検索を行なうことができるようになった．しかし，検索の対象が長期に及ぶ事件や政治問題などの場合，検索の結果数多くの記事が現れ，それらすべてに目を通すには多大な時間を要する．そこで，これらの記事をまとめて要約する技法を開発することは極めて有用である．しかしながら，従来，要約の研究は単独の文章を対象とするものが多く [2, 3]，複数の文章に対して要約する研究はほとんどなされていない [4, 5]．

複数の新聞記事の要約に関する研究として，[4] では，関連した二つの新聞記事について一方を元記事，他方を要約対象記事とし，後者で連体修飾句，節照合処理，言い替えにより重複部分の削除を行なっている．一方，本研究では，二つ以上の記事についてそれらの記事全体を要約対象として新聞記事に現れる表現を手がかりに重複部分と文の削除を行なった．言い替えの処理については本研究でも行なったが，これは括弧の処理の一つとして行なった．[5] では，ある話題の複数の記事に対して重要と考えられる項目をテンプレートとし，それを用いて抽出した各記事のデータから要約を作り出している．一方，本研究では，新聞記事に現れる表現を手がかりに重複部分と文を削除することにより要約とした．

ところで，正しい絞り込みによって検索された記事は関連したものであるから，ある記事で述べている事柄を別の記事でも述べていることが多く，このような部分は冗長である．また，新聞記事は冒頭の部分で記事の概要を述べていることから，記事の前半部分が重要である．さらに新聞記事には事実文と推量文があるが，前者のほうが重要である．

そこで，本研究では，文中の重複部分と不要な文を削除することにより関連した複数の新聞記事の要約を行なった．これらの処理は，形態素解析のみによって得られる表層的な情報を用いて，計算機上で行なった．また，入力となる新聞記事はある事柄について検索を行なった結果現れた関連した記事であるとする．

2 要約手法

要約の対象とする新聞記事はすでに与えられているものとする．処理の順序は 2.1 節の処理を最初に行なった後，2.2〜2.5 節の処理を行なう．本手法は，26 種の関連記事をもとに決定した．

2.1 新聞記事の第一段落以外の削除

新聞記事は，冒頭で記事の概要を述べている．また，段落が内容の単位になることから，記事の第一段落のみを残し，第二段落以下を削除する．本研究では，各記事の第一段落について要約を行なうものとし，この部分を入力とする．

2.2 推量文の削除

文末が推量表現である文は未知，未確認のことを述べている [1] ので，事実を伝えるには必要がない．これらの文を文末表現で判断し，削除する．文の削除対象とした文末表現を表 1に列挙する．ただし，文末表現が「向かう」の場合は，「〜う」に該当するが推量を表していないので，この処理を行なわない．

2.3 人物の説明の削除

新聞記事では，次のように，人名の前後でその人物の住所や職業が述べられる．

沖縄県沖縄市○○○、観光会社社長、○○○○容疑者（４１） [日経新聞 27/Feb/1992]
前道議の○○○○容疑者（５６）＝渡島管内○○町○○○＝ [日経新聞 20/Feb/1992]

表 1: 文の削除対象となった文末表現

～う	～可能性が大きい	～可能性もある	～かもしれない
～情勢だ	～そうだ	～ちがいない	～と思われる
～はずだ	～見通し	～見通しだ	～みられる
～模様だ	～ようだ	～予想される	～らしい

このような部分が繰り返し現れた時は，職業，住所の部分を削除する．方法としては，同じ人名が繰り返し現れた時に，その人名から名詞または助詞「の」または「、」の続く限り文頭の方にたどり，その部分を削除する．また，人名の後に説明が来る場合は「＝」で囲まれているので，その部分を削除する．

2.4 括弧の処理

新聞記事における括弧の使われ方は，言い替えと括弧の前の語の説明がある．

2.4.1 言い替え

言い替えは，括弧の前の語を括弧内にある語で言い替えている部分が繰り返し (二度以上) 現れた時に括弧内の語に置き替える処理である．次のような例の場合，「石油輸出国機構」を「ＯＰＥＣ」と言い替えているので，繰り返し現れた時は「ＯＰＥＣ」のみに置き換える．この処理を行なう条件は，(1) 括弧内の語がその前の語より短い，かつ，(2) 括弧内の語が一単語である，の二つである．但し，未知語も一単語とみなす．

石油輸出国機構（ＯＰＥＣ） → ＯＰＥＣ

2.4.2 括弧の前の語の説明

-1z 括弧の前の語を説明する部分が繰り返し (二度以上) 現れた時は，括弧をその中の語と共に削除する．次のような場合，「本社神戸市」が「ダイエー」を説明しているので，繰り返し現れた時は「（本社神戸市）」を削除し，「ダイエー」のみとする．この処理は，言い替えの処理が行なわれなかったものについて行なわれ，括弧の中の語が前に現れた括弧の中の語と完全に一致している，または，括弧の中の語が前に現れた括弧の中の語に連続した文字列として含まれる場合行なう．また，括弧の中の語に「当時」(年齢等に用いられる) が含まれている場合は，この語は除外して処理を行なう．

ダイエー（本社神戸市） → ダイエー

2.5 導入部の削除

ある事柄についての一連の記事では，古い記事で述べられていることを新しい記事で再び述べている部分がある．このような部分は，単独の記事であればそれまでの経過を知らせるために必要であるが，古い記事がある場合は冗長である．処理の対象となる部分を導入部とし，記事の第一文の文頭から次に示す表現までの部分とする．

～したが，～問題で，～事件で，～事故で，～していたが，～について

また，文頭から「名詞＋は」までの部分も処理の対象となる導入部とする．

導入部の削除の方法としては，導入部の名詞と古い記事の一文ごとの名詞を比較し，導入部の名詞のうち７割以上が一致すれば，その導入部は冗長であるものとみなし，削除する．ただし，「名詞＋は」の部分は文の主題を示しているので，導入部を削除する際，その部分は削除しない．次に示す例では，下線部が削除される．

フィリピン南部のミンダナオ島で四日未明、フィリピン国軍の一部将兵が反乱を起こし、陸軍基地や主要都市の一部を占拠した。[日経新聞 4/Oct/1990]

フィリピン南部のミンダナオ島で四日未明発生した国軍の一部将兵による反乱は、その後反乱軍とアキノ政権軍のにらみ合いが続いている。[日経新聞 5/Oct/1990]

削除の対象となる導入部は，古い記事で現れている名詞を含んでいるはずである．そのような名詞は，古い記事でも，ある一文中にまとまって現れているので，名詞を抽出する単位を一文とする．

3 実行例

実行例には，日本経済新聞 CD-ROM92 年版の 2 月 27 日，3 月 19 日，4 月 11 日の記事を使用した．

この例では，人物の説明の削除，括弧の前の語の説明，導入部の削除が行なわれている．人物の説明の削除では，「○○○○」の住所，職業の部分が削除されている．括弧の前の語の説明では，「○○○○容疑者（４１）」の部分が繰り返し現れているので，「（４１）」の部分が削除されている．導入部の削除では，第３記事の「大手スーパー、ダイエーに対する恐喝未遂事件で」の部分が削除されている．

3.1 実行例の原文

「沖縄の社長逮捕、ダイエー脅迫――大阪府警、そごう爆破追及」

大手スーパー「ダイエー」（本社神戸市、中内功会長）に、今年一月から二月にかけて、現金百五十万ドル（約二億円）を要求する脅迫状が送りつけられる事件があり、大阪府警捜査一課は二十六日、沖縄県沖縄市○○○、観光会社社長、○○○○容疑者（４１）を、恐喝未遂の疑いで逮捕した。[27/Feb/1992]

「○○被告、ダイエー恐喝未遂起訴――地検、そごう爆破も追及」

「そごう大阪店」（大阪市）で起こった爆破事件を材料にして、大手スーパー「ダイエー」（本社神戸市）が百五十万ドル（約二億円）を要求された事件で、大阪地検は十八日、沖縄県沖縄市○○○、観光会社社長、○○○○容疑者（４１）を恐喝未遂の罪で起訴した。

○○被告は起訴事実を否認している。

○○被告はダイエーに送り付けた脅迫状で、一月十七日発生のそごう爆破事件への関与をほのめかしており、同地検は引き続き、爆破事件についても追及する。[19/Mar/1992]

「大阪府警、ダイエー脅迫の観光会社社長、そごう爆破で再逮捕――経営不振で借金苦」

大手スーパー、ダイエーに対する恐喝未遂事件で逮捕、起訴された沖縄県沖縄市○○○、観光会社社長、○○○○被告（４１）が十一日までに、大阪・ミナミの百貨店「そごう大阪店」で今年一月に起きた爆破事件についても犯行を自供、大阪府警捜査一課の南署捜査本部は同日、○○容疑者を爆発物取締罰則違反の疑いで再逮捕した。

○○容疑者は「会社の経営が行き詰まり、生活が苦しくなった。

家の新築費も欲しくて、金に困って犯行を思い付いた」と供述している。[11/Apl/1992]

3.2 実行例の出力

大手スーパー「ダイエー」（本社神戸市、中内功会長）に、今年一月から二月にかけて、現金百五十万ドル（約二億円）を要求する脅迫状が送りつけられる事件があり、大阪府警捜査一課は二十六日、沖縄県沖縄市○○○、観光会社社長、○○○○容疑者（４１）を、恐喝未遂の疑いで逮捕した。

「そごう大阪店」（大阪市）で起こった爆破事件を材料にして、大手スーパー「ダイエー」（本社神戸市）が百五十万ドルを要求された事件で、大阪地検は十八日、○○○○容疑者を恐喝未遂の罪で起訴した。

○○被告は起訴事実を否認している。

○○被告はダイエーに送り付けた脅迫状で、一月十七日発生のそごう爆破事件への関与をほのめかしており、同地検は引き続き、爆破事件についても追及する。

逮捕、起訴された○○○○被告が十一日までに、大阪・ミナミの百貨店「そごう大阪店」で今年一月に起きた爆破事件についても犯行を自供、大阪府警捜査一課の南署捜査本部は同日、○○容疑者を爆発物取締罰則違反の疑いで再逮捕した。
○○容疑者は「会社の経営が行き詰まり、生活が苦しくなった。
家の新築費も欲しくて、金に困って犯行を思い付いた」と供述している。

4 実験結果

実験にはPerl言語を用い，SUN SPARC station 1上に実装して行なった．実験に使用したデータは日本経済新聞CD-ROM90年版から取り出した．記事の種類は23種類で，それぞれ2～5記事で構成されている．実験は，未知の関連記事に対して行なった．

4.1 推量文の削除の結果

推量文の削除の実験結果を次に示す．再現率，適合率を次のように計算する．削除すべき文とは，推量文であると筆者が判断した文とする．

$$\text{再現率} = \frac{\text{正しく削除した文の数}}{\text{削除すべき文の数}} \times 100(\%)$$

$$\text{適合率} = \frac{\text{正しく削除した文の数}}{\text{削除した文の数}} \times 100(\%)$$

再現率，適合率の結果次のとおりである．

$$\text{再現率} = \frac{13}{16} \times 100 = 81(\%)$$

$$\text{適合率} = \frac{13}{13} \times 100 = 100(\%)$$

4.2 人物の説明の削除の結果

人物の説明の削除の実験結果を次に示す．再現率，適合率を次のように計算する．削除すべき部分とは，人物の職業や住所を述べている部分が前述されていると筆者が判断した部分とする．

「削除すべき部分」の一部分だけが削除された時は正しく削除できなかったものとする．例えば，「abcd」が削除すべき部分で，「ab」や「cde」を削除した時は正しく削除できなかったものとする．

$$\text{再現率} = \frac{\text{正しく削除した部分の数}}{\text{削除すべき部分の数}} \times 100(\%)$$

$$\text{適合率} = \frac{\text{正しく削除した部分の数}}{\text{削除した部分の数}} \times 100(\%)$$

再現率，適合率の結果は次のとおりである．

$$\text{再現率} = \frac{10}{10} \times 100 = 100(\%)$$

$$\text{適合率} = \frac{10}{10} \times 100 = 100(\%)$$

4.3 括弧の処理の結果

4.3.1 言い替え

言い替えの実験結果を次に示す．再現率，適合率を次のように計算する．言い替えるべき部分の数とは，括弧の前の語を括弧内の語で置き換えて良いと筆者が判断した部分とする．

$$\text{再現率} = \frac{\text{正しく言い替えた部分の数}}{\text{言い替えるべき部分の数}} \times 100(\%)$$

$$\text{適合率} = \frac{\text{正しく言い替えた部分の数}}{\text{言い替えた部分の数}} \times 100(\%)$$

再現率，適合率の結果は次のとおりである．

$$\text{再現率} = \frac{17}{19} \times 100 = 89(\%)$$

$$\text{適合率} = \frac{17}{18} \times 100 = 94(\%)$$

4.3.2 括弧の前の語の説明

括弧の前の語の説明の削除の実験結果を次に示す．再現率，適合率を次のように計算する．削除すべき部分とは，括弧の内容が前述されていると筆者が判断した部分とする．

$$\text{再現率} = \frac{\text{正しく削除した部分の数}}{\text{削除すべき部分の数}} \times 100(\%)$$

$$\text{適合率} = \frac{\text{正しく削除した部分の数}}{\text{削除した部分の数}} \times 100(\%)$$

再現率，適合率の結果は次のとおりである．

$$\text{再現率} = \frac{10}{10} \times 100 = 100(\%)$$

$$\text{適合率} = \frac{10}{11} \times 100 = 91(\%)$$

4.4 導入部の削除の結果

導入部の削除についての実験結果を示す．再現率，適合率を次のように計算する．ここで削除すべき導入部とは，記事の第一文の文頭から次に示す表現までの部分と文頭から「名詞＋は」までの部分で，筆者が読んで冗長なので，削除すべきと判断した部分とする．

～したが，～問題で，～事件で，～事故で，～していたが，～について

$$\text{再現率} = \frac{\text{正しく削除した導入部の数}}{\text{削除すべき導入部の数}} \times 100(\%)$$

$$\text{適合率} = \frac{\text{正しく削除した導入部の数}}{\text{削除した導入部の数}} \times 100(\%)$$

再現率，適合率の結果は次のとおりである．

$$\text{再現率} = \frac{26}{30} \times 100 = 87(\%)$$

$$\text{適合率} = \frac{26}{26} \times 100 = 100(\%)$$

4.5 削減率

本手法では，まず，第一段落以外を削除している．そこで，それから更にどの程度削除されたかを削減率として表す．削減率は，次のように計算する．

$$\text{削減率} = 100 - \frac{\text{出力結果の文字数}}{\text{全記事の第一段落の文字数}} \times 100(\%)$$

$$\text{最小削減率} = 0.0(\%)$$

$$\text{最大削減率} = 36.2(\%)$$

$$\text{平均削減率} = 14.5(\%)$$

また，筆者が重複部分と推量文を削除した時の削減率を次に示す．

$$\text{最小削減率} = 0.0(\%)$$

$$\text{最大削減率} = 42.8(\%)$$

$$\text{平均削減率} = 18.0(\%)$$

図 1 に，出力結果の削減率と筆者が重複部分と推量文を削除した時の削減率を示す．各点は，各関連記事を表す．

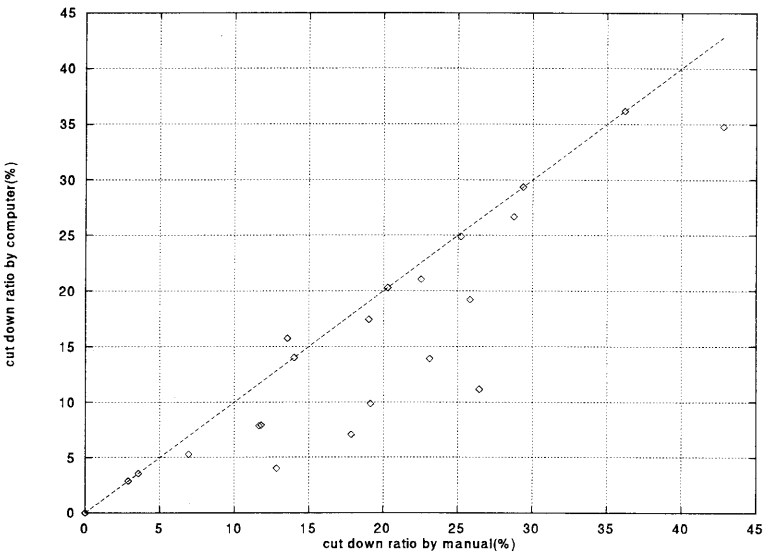

図 1: 要約手法と人手による削減率の比較

5 考察

実験例に示されるように，本稿で提案した要約手法により，文章を短縮することができた．出力例のような事件についての記事では，処理 2.3，2.4，2.5 により重複部分が削除されたが，政治問題のような記事では，第一段落を取

り出しただけで終わる例もみられた．このような例では，重複部分が少なく，また，重複部分は連体修飾語句にみられるため，本手法では削除できなかった．

推量文の削除については，再現率が81%となっているが，削除できなかった理由は，表1に示される文末表現以外のものがあったためである．

言い替えについては，再現率が89%となっているが，言い替えが行なえなかった理由は，括弧内の語が一単語で構成されていないためである．また，適合率が94%となっているが，間違って言い替えを行なった理由は，括弧の前の名詞の一部が言い替えの処理として置き換える部分だが，全部を置き換えたためである．

括弧の前の語の説明については，適合率が91%となっているが，間違って削除した理由は，同じ年齢の人物が2人現れたため，括弧内の語が一致し，削除したためである．

導入部の削除については，再現率が87%となっているが，削除すべき導入部が削除できなかった理由としては，別の単語を使った表記をしている，導入部の内容が2文にまたがっているということがあげられる．

6 まとめ

本稿で提案した要約手法により文章を短縮することができたが，冗長な部分を完全に削除することはできなかった．削除できなかった部分は，連体修飾語句に多く見られ，表層的な情報からこれらを正確に削除することは難しい．そこで，更に削減率を改善する方法として，構文解析を行ない文中の修飾語句を削除することがあげられる．

また，本手法で冗長な部分を多く削除できた記事は，事件についての関連記事であった．このような記事の冗長な部分は，人物の説明の削除，括弧の処理，導入部の削除の処理対象となることが多いためである．

謝辞

本研究で，日本経済新聞の一部記事について，本論文への引用許可をいただいた(株)日本経済新聞社に深謝する．

参考文献

[1] 国語学会編:" 国語学大辞典", 東京堂出版,1991.

[2] 田村俊哉, 田村直良:" 文章の表現形式に基づいた要約文章の生成について", 情処研報 NL92-1,1992.

[3] 山本和英, 増山繁, 内藤昭三:" 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,vol.2,No.1,pp.39-56,1995.

[4] K.Yamamoto,S.Masuyama and S.Naito:"An Empirical Study on Summarizing Multiple Texts of Japanese Newspaper Articles",Proc. of NLPRS'95,pp.461-466,1995.

[5] K.McKeown and D.R.Radev:"Generating Summaries of Multiple News Articles",In SIGIR'95,pp.74-82,1995.