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

三 上 真[†] 増 山 繁[†] 中 川 聖一^{††}

ニュース原稿を1文ごとにそれぞれ要約する手法について報告する。1文が長く、1記事中の文数の少ないニュース原稿に対して文を抽出単位とする要約手法を用いることは、大きく情報が欠落する可能性があり、適切でない。そこで、本要約手法では修飾部および比較的冗長と考えられる部分を削除することにより、1文ごとの要約を行う。また、1文を部分的に削除する際に構文構造が破壊されることを防ぐため、ニュース文要約に特化した簡易構文解析手法を利用している。字幕文は、画面上を一方的に流されるという性質から、適切な長さに要約されている必要があり、読みやすく、原稿の情報が正確に伝わり、冗長さが解消されている必要がある。このため、被験者32名に対し、本手法による要約文についてのアンケートを行うことにより、自然さ、忠実さ、非冗長さの3つの視点から評価を行った。その結果、理想的な要約を100%とした場合で、自然さ81.5%、忠実さ74.3%、非冗長さ83.3%という評価値を得た。

キーワード: 要約作成, 構文解析, ニュース文, 修飾語省略

A Summarization Method by Reducing Redundancy of Each Sentence for Making Captions of Newscasting

MAKOTO MIKAMI [†], SHIGERU MASUYAMA[†] and SEIICHI NAKAGAWA ^{††}

We propose and evaluate a method for summarizing each sentence in TV news texts written in Japanese. It is not appropriate to select important sentences for abstracting news texts, because a news text consists of only a few and long sentences. Then, we try to reduce redundant parts, which consist of modifier etc., of each sentence. We use a simple parsing method specialized for news texts so that the syntactical structure is not destroyed. As audiences cannot read repeatedly, a summary must be shortened moderately. It must also be easy to read, containing important information, and reduced its redundancy. Therefore, we evaluate this summarizing method by obtaining information by means of questionnaires to 32 examinees.

KeyWords: *Summarization, Parsing, TV news text, Deletion of modifiers*

[†] 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

^{††} 豊橋技術科学大学 情報工学系, Department of Information and Computer Sciences,
Toyohashi University of Technology

1 はじめに

日本のテレビ番組における字幕付き放送の割合は10%程度と低く、近年、字幕放送率向上を目指し、自然言語処理技術を応用した効率的な字幕生成が切望されている(江原, 沢村, 若尾, 阿部, 白井 1997). 番組の音声情報を字幕化するには、文章を適度な長さに要約する必要があるため、本研究では、ニュース原稿(テキスト)を入力とした、字幕生成のための自動要約を試みた。本要約手法では、ニュース文の特徴を利用し、1文ごとの要約を行っている。

テキスト自動要約研究の多くは、テキスト中の文もしくは文のまとまりを1単位とし、何らかの情報に基づき重要度を決定、抽出することで要約を行う。このような要約手法は、文献検索において原文の大意を把握するための補助などに用いられ、成果を上げている(住田, 知野, 小野, 三池 1995). ニュース番組における字幕生成では、ニュース原稿の第1文(全体の概要を述べる場合が多い)を抽出することによる要約が考えられるが、画面に表示されるVTRなどとの対応を考慮に入れると、必ずしも十分でない。文単位の抽出においては、照応や文の結束性を保つため、採用文の前文も採用するなどの対策が講じられているが(Paice 1990)、不要な文まで芋蔓式に採用してしまう場合もあり、結束性と首尾一貫性をより高めるには後編集を行う必要があるなど、その困難さも同時に報告されている(山本, 増山, 内藤 1995). また、与えられたテキストから必要な情報を抜き出す手法として、情報抽出研究が注目されている(Cowie and Lehnert 1996). この手法は、領域が限定された記事に対しては有効である。しかし、与えられたニュース原稿には「事件」、「政治」といった領域を限定する情報が与えられておらず、字幕文生成への情報抽出手法の適用は難しいと考えられる。

ニュース文は新聞記事に比べ、1文中の文字数が多く、1記事あたりの文数が少ないという特徴を持つ(若尾, 江原, 村木, 白井 1997b). このため、字幕用の要約文を生成するために、文を単位とした抽出を行うと、採用される情報に大きな偏りが生ずるという問題がある。若尾ら(若尾, 江原, 白井 1998)は、自動短文分割後、重要文を抽出することによるニュース文の自動要約を行っている。これに対し、本手法は、ニュース原稿における各文はそれぞれ同様に重要であり、画面との対応や記事全体での結束性を重視するという立場から、ニュース文の構文構造を利用し、文中の修飾語句等を削除することによる、1文ごとの要約を行っている。1文の一部を抜き出すことで、より自然な文章を生成するには、残存部に係る部分の削除を避けなければならない。本手法では、ニュース文の各文における最後尾の動詞は重要であると仮定し、これに係ると考えられる部分を残すことにより、不自然な要約文の生成を防いでいる。また、本研究は、言い替えによる要約(山崎, 三上, 増山, 中川 1998)を後処理に適用し、最終的な字幕文を生成することを想定しているが、本論文では両手法を併用せず、本要約手法の分析に焦点を絞った。本要約手法についての背景、目的等は、2節でも詳述する。

自動要約研究においては、正しい要約を唯一に定義することが困難なことから、その評価についても様々な手法が用いられる。その一つに、人間の被験者の生成した要約文と、システム

が生成した要約文を比較する評価法があるが、複数の被験者の要約が高い割合で一致することは難しいと考え(奥村 難波 1998)、システムによる要約文を被験者に数値で評価させる手法をとった。同様の手法による評価を山本ら(山本他 1995)が行っているが、数値のみで評価した場合、被験者が不適切と判断した箇所を特定するのが難しいという問題がある。山本らは被験者に対し、質問項目以外に感想を求めており、それを分析することで要約の不適切さの原因やその改善を検討している。本論文においては、要約が不適切な箇所をより特定し、分析を行うことを考え、実施したアンケートでは数値による評価に加え、要約が不適切と思われる箇所を被験者に指摘させた。自動要約の評価法に関しては、他に、要約を利用したタスクの達成率を見ることにより、間接的に要約文の評価を行うものがある。住田ら(住田他 1995)は、抄録文の文書集合から、設問に対応する文書を選択するというタスクを被験者に与え、選択された文書数と正解の文書数から再現率を求めている。しかし、本論文では字幕文生成の要約のため、適切なサブタスクを設定することが難しく、また、被験者の持つ知識の差を考慮した場合、その評価が難しいと予想されるため、用いなかった。

以下、2節でニュース文要約の目的、手法およびニュース原稿の特徴等について述べ、3節から5節で、提案する1文ごとの自動要約手法について述べる。6節では、アンケート調査に基づき、本手法を評価する。7節では、自動要約実験およびアンケート調査によって明らかになった、本要約手法の問題点等を考察する。なお、入力コーパスとして、NHK放送技術研究所との共同研究のため提供された、NHK 汎用原稿データベースを使用した。

2 ニュース文要約

自動要約文は、多くの場合、文書のダイジェスト情報を把握することや、関連記事群の鳥瞰情報を得る等の目的で用いられる。このような目的で用いられる要約結果は、できる限り文字数が少なく、かつ、入力文の特徴もしくは必要とする情報を圧縮し、適切に抽出していることが求められる。

これに対し、本研究における要約は、ニュース字幕の自動生成を目的とする。字幕の利用者として聴覚障害者等を想定した場合、本来ならニュース原稿の内容を全て字幕にするのが望ましい。しかし、より読みやすくするためには、適切な長さに要約する必要がある(冗長さの解消)。実際、テレビ番組で使用されている字幕は文字数が制限されており(若尾, 江原, 白井 1997a)、宮坂は、原稿を要約した字幕が、全文を用いた字幕より読みやすいことをアンケート結果から分析、報告している(宮坂 1998)。また、画面は時系列的に変化するため、利用者が読み直すことなく理解できる、より読みやすい(文として自然である)文章であることが字幕文には求められる(自然さの確保)。さらに、要約による入力原稿重要部の欠落は極力避ける必要があり(忠実さの確保)、生成された字幕は、できる限り画面の内容と同調していることが望ましい。

このような理由から、本要約手法では、ニュース原稿の文ごとの短縮を目標としている(入

力原稿は画面の内容と同調していると仮定)。入力中の1文を全て削除した場合、画面に字幕が出力されない時間が生じ、字幕利用者に不安を与える可能性があるため、文そのものの削除は原則的に行っていない。より質の良い出力を得るには、記事全体を理解し、文脈を理解した上で要約を行うことが求められるが、現時点でそのような要約過程の全てを自動化するのは難しい。そこで、本研究では、まず、表層的な情報から連体修飾句などを冗長部と認定し、削除することによるニュース文要約を試みる。また、従来の自動要約には取り入れられていなかった構文解析を利用することで、文内の部分的な削除による構文構造の破壊に対処した。構文解析は、解析誤りによる要約文の品質低下への虞などから従来用いられていなかったと考えられ、本手法でも厳密な解析結果の利用は避けるべきと判断し、ニュース文要約に特化した簡易構文解析手法を考案した。

ニュース文の要約手法として、言い替えによる要約(山崎他 1998)があるが、現段階の要約率は90%程度であり、単独で用いたのでは必ずしも十分でない。本要約手法は、後処理として言い替えによる要約を適用することを想定し、実用化するには両手法を併用して要約率70%程度の達成を目標としている。このため、本要約手法は、語尾の冗長な表現などには対処していない。また、本研究は現在、実用的な要約システム構築のための基礎研究の段階にあり、本要約手法自身の持つ性質、効果、限界を明確にするため、本論文では言い替えによる要約(山崎他 1998)を併用しなかった。言い替えによる要約は江原ら(江原他 1997)の一連の要約研究においてもほぼ同様の手法が提案されており、これを本要約手法の後処理に用いることも原理的に可能である。

また、本研究は対象をニュース文に限定し、要約手法を検討した。以下に、ニュース文の主な特徴をまとめる。

ニュース原稿の特徴

与えられたニュース原稿の特徴を以下に列挙する。なお、これらのうちの一部は若尾ら(若尾他 1997b)によって既に指摘されている。

- (1) 各記事ごとに、日付と原稿名が与えられている(ただし、原稿名が記事の要約にはなっていない場合も多い)
 - (2) 段落構造が無い(若尾他 1997b)
 - (3) 最初の1文が全体の要約文になっている(若尾他 1997b)
 - (4) 新聞記事に比べ、1文が長く、1記事内の文数が少ない(若尾他 1997b)
 - (5) 複文((主・)述を含む節が文全体の主・述部または修飾部等になっている文)や、連用中止法による並列構造が多い
 - (6) 新聞記事に比べ、話し言葉に近く、助詞の欠落や変則的な読点の使用が見られる
- 本研究では、これらの特徴を踏まえ、要約手法を検討した。

3 本要約手法の構成

本要約手法は以下に示す部分からなる。

簡易構文解析部 形態素解析器 JUMAN3.5 による形態素解析後, 構文解析器 KNP2.0b5 により入力文の文節情報を取り出し, 入力文を文節に切り分け, ニュース文要約に特化した簡易的な係り受け解析を行う。

削除文節選択部 修飾部など, 事実を伝える上で比較的冗長である文節を認定後, さらに簡易構文解析結果を利用し, 削除部を認定する。

4 簡易構文解析

1 文を部分的に削除することにより構文構造が破壊されることを防ぐためには, 構文解析を行う必要がある。しかし, 2節でも触れたように, ニュース文は新聞記事等に比べ口語体に近く, 現在の構文解析技術を用いて高精度な解析を行うことは難しい。このため, 不適切な削除を避ける目的においてより頑健であることを重視し, ある程度の曖昧性の残存を許した, ニュース文要約に特化した係り受け解析を行う。このような観点から, 簡易構文解析は, 厳密な解析を避け, KNP が出力する文節情報を基に, 適切な削除を補助する目的で行った。

本論文では, ニュース原稿における 1 文の最後の述部 (1 文中で最後に現れる用言, 以下単に述部と呼ぶ) を文全体の核と仮定し, 述部に係る文節の削除は, 文全体の自然さを破壊すると考える。このような不適切な削除を避けるため, 簡易構文解析では, 係り関係が連用である文節は述部に係ると認定するなど, より遠い用言に係ると推定される文節は, 全て述部に係ると認定する。ここで, 述部に係ると考えられる文節までで文を分割した各文節列を並列単位 (4.2 節参照) とし, 各並列単位の最後尾の文節 (以下, 連用文節と呼ぶ) を原則的に残すことで, 不適切な削除を避ける (「～は」などの文節も用言に係ると考え, 連用文節に含める)。連用文節の全てが, 必ずしも述部に係るわけではないが, より遠い用言に係ると推定される文節は, 多くの場合, 他の連用文節のいずれかに係ると推定される。

また, 削除部の認定においては, 冗長部に係る文節が残存することによる, 不適切な要約文の生成を避ける必要がある。このため, 簡易構文解析は, 冗長と認定された文節に係る文節を特定し, 不要となる部分が残存することを避ける。この場合も, 厳密な解析を避け, 連用文節以外の文節は, 同一並列単位内に含まれる後部のいずれかの文節に係ると推定することで対処する。

以下では, 簡易構文解析について述べる。

4.1 文節の認定

入力文を KNP によって解析した結果, KNP が 1 文節として出力した形態素列を, それぞれ 1 文節とする。

4.2 並列単位の認定

以下に示す規則に基づき、述部に係る文節を認定する。

KNP が出力した文節パターン情報 (文節の働きを示す情報) に基づき、

- 係り関係が<連用>の場合 (用言で、その活用形が連用形)、
- 係り関係が<同格未格¹>、<未格²>の場合、
- パターンが、「では」、「に」(外の関係³)、「で」(並列⁴)を示す場合

に、その文節は述部に係ると認定する。ただし、

- 時間を表す名詞の場合、
- 副詞の場合、
- 最後の形態素が「と」、「も」の場合、
- 直後の文節が動詞を含む場合

には、その文節は述部には係らないものとする。

以上の結果、述部に係ると認定された文節を連用文節とし、連用文節までで文を分割したそれぞれの文節列を、並列単位とする (ただし、形態素数が1の並列単位は認めず、その場合は直後の並列単位と統合する)。また、連用文節以外の文節は、同一並列単位内に含まれる後部のいずれかの文節に係ると認定する。

4.3 簡易構文解析結果例

簡易構文解析結果の例を以下に示す (付録 A 節で示す要約結果例の第 1 文)。認定された並列単位の区切りを、「||」で表す。また、各文節の区切りを、「|」で表す。

[例文 1]

政府 は、||
 きょう の | 閣議 で、||
 退職 公務員 など に | 支給 される | 恩給 を、| 今年 の | 四 月 から | 年額 | 二
 点 | 六六 パーセント | 引き上げる | 恩給法 の | 改正 案 を | 決定 し、||
 国会 に | 提出 する | こと に | して います 。

KNP により、例文中の「政府は」は未格、「閣議で」および「決定し」は連用と解析されたため、述部に係ると認定した。

この例においては、「退職公務員などに支給される恩給を、」は「引き上げる」に係ると解釈するのが正しいと考えられるが、KNP では、そのような解析結果を得ることができなかった。

1 「『(体言) など』(同格か未格かわからない)」(KNP rule-comment.txt).

2 「未格、『～は』、『～すら』など」(KNP rule-comment.txt). KNP 内部での、副助詞だけで格助詞がない格要素の呼称。

3 被修飾語が修飾語 (動詞) の格要素にならない (佐久間 1994). 格要素になる場合を、「内の関係」と呼ぶ。

4 KNP では、「(外の関係)」および「(並列)」の特定を、構文解析の前に行う。

この場合、「引き上げる」が冗長と認定されても、「退職公務員などに支給される恩給を、」が要約文に残存してしまう。実際に、要約結果では「引き上げる」が冗長と認定されるが、簡易構文解析結果を利用することにより、同一並列単位内に含まれ、かつ、「引き上げる」よりも前方の全ての文節が、削除部として認定された(削除部の認定法については、5.3節で後述)。簡易構文解析では、「退職公務員などに支給される恩給を、」が「引き上げる」に係るのか「決定し」に係るのかは特定しないが、この例のように、不適切な残存を避けるために有用である。また、「決定し、」に係ると解釈するのが妥当と考えられる「閣議で、」は、簡易構文解析によって連用文節と認定されているが、「決定し、」も連用文節と認定されるため、適切な要約を生成することができる。

5 削除文節選択

ニュース文の中心的内容が影響を受けない部分を削除することにより、要約を行う。本要約手法では、連体修飾語、および、例示を表す語など、削除によって意味的に変化が生じにくいと考えられる語を伴う部分を冗長部と定義し、削除候補とする。削除候補は文節を最小単位として選択するが、同一文節中に重要と考えられる語を含む場合や、係り先の文節(直後の文節)の最初の形態素が形式名詞および形式名詞に準ずる名詞(その語単独ではあまり意味を持たない語、以下形式的表現と呼ぶ)である場合などは削除しないなど、例外処理も設けている。

以下、冗長部の認定、形式的表現、削除部の認定について述べる。

5.1 冗長部の認定

JUMAN による形態素解析結果において、活用形がタ形・基本形・連体形のいずれかに分類され、かつ、直後に名詞(形式名詞および副詞的名詞を除く)がある場合、そのような形態素を伴う文節は修飾部であると判断し、冗長部とした。また、例示などを示す表現を含む文節を、冗長部とした(「など」、「や」、「ほか」、「とともに」、「うち」、「として」、「結果」)。ただし、同一文節中に(1)原稿名および第1文中の名詞、(2)主要語(山本ら(山本他 1995)が定義、角川類語新辞典(大野 浜西 1981)の大分類番号が{0,5,7,8,9}である語および固有名詞)、(3)重要と考えられる表現(主格や目的格となり得る格助詞や、「ため」など)を含む場合、それらの情報も考慮して、冗長/非冗長を決定した。なお、重要/冗長を示す語は、あらかじめ人手でテーブルを作成し、その情報から認定する。

5.2 形式的表現

文中の修飾部を削除する場合、形式的表現に係るものを削除すると、意味が取れなくなる可能性がある。このため、係り先に形式的表現がある場合は、削除候補としない。形式的表現は、

あらかじめ人手で辞書を作成し、その辞書を用いることで認定する。現在形式的表現として 43 個の表現を登録している (「時期」、「見通し」、「構え」など)。以下の例文では、形式的表現に係る部分 (下線部) が強制的に採用された。

[例文 2] 六月までの上半期では 去年の同じ 時期に比べて...

5.3 削除部の認定

5.1節および5.2節に示す処理の結果、冗長部と認定され、削除の対象となる文節が選択される。このとき、削除される文節に係る部分が残存することによる構文構造の破壊を防ぐため、簡易構文解析において係り先である可能性があると判断された文節が削除される場合、係り元の文節も削除する。この結果、冗長部と認定された文節と同一並列単位内に含まれ、かつ、冗長部と認定された文節より前方にある文節は削除部と認定される。以下の例文では、「言われる」および「発生するなど」が冗長部と認定され、その文節に係る可能性のある文節を削除した結果、最終的に文中の [...] の部分が削除部として認定された。

[例文 3] 「首都圏最後の”水がめ”とも言われる」霞ヶ浦は、[毎年、夏場になると大量のアオコが発生するなど] 流域の都市化に伴って年々汚れが目立っていることから建設省では浄化対策として昭和五十年から土浦港を中心に、しゅんせつ工事を進めています。

6 評価

6.1 要約結果および評価方法

ニュース文要約においては、自然さの確保、忠実さの確保、冗長さの解消が重要であると考えられる (2節参照)。このため、被験者 (工学部学生および大学院生) 32 名に対して本手法による要約文に関するアンケートを行い、本要約手法の有効性を評価した。アンケートは、10 記事 (1993 年 1 月および 2 月のニュース原稿) を用い、それぞれの記事を入力とした際の自動要約結果および自動要約による削除部を明示した入力原稿を被験者に提示し、以下に示す項目について 0~5 までの整数で評価値を付与させることで行った。また、評価値が 5 以外の場合、要約の失敗箇所を明確化するため、評価値の他に自動要約結果の不適切と考えられる部分を指摘させた。

(1) 自然さ

各要約文章のみを独立した文章として読んだときに、自然かどうか

(2) 忠実度 (重要部欠落の指摘)

原文と要約文とを比較し、原文で重要と考えられる部分を抽出しているか (抽出されていない箇所がある場合、上位 5 箇所を指摘)

(3) 非冗長度 (冗長部残存の指摘)

表 1 要約結果

	原稿 1	原稿 2	原稿 3	原稿 4	原稿 5	原稿 6	原稿 7	原稿 8	原稿 9	原稿 10
要約率 (%)	82.3	76.7	82.8	77.7	85.1	79.5	71.9	80.0	66.7	75.4
原稿文字数	672	514	580	506	424	555	551	426	555	334

表 2 有効回答数

	原稿 1	原稿 2	原稿 3	原稿 4	原稿 5	原稿 6	原稿 7	原稿 8	原稿 9	原稿 10
自然さ	32	32	32	32	32	30	30	30	29	30
忠実度	32	32	32	32	31	30	30	29	29	30
非冗長度	32	32	32	32	32	30	30	29	29	29

原文と要約文とを比較し、原文で冗長と考えられる部分が残存していないか (残存している箇所がある場合、上位 5 箇所を指摘)

6.2 節以下では、これらの質問および得られた回答について述べる。

なお、アンケートに用いた要約結果は、

- 適切な原稿名が付与されているもの
- 元原稿が文として読みやすいもの
- 入力の記事長が短すぎないもの
- KNP がエラー (「;; Cannot」で始まる行など) を出していないもの
- KNP が大きな解析誤り (係り受け解析の誤り) を起こしていないもの
- 要約率が 80% 前後のもの

の全てを満たすという条件の下で無作為に選んだ。また、事前に試行テストを行うことにより、回答時間を 1 記事当たり 10 分と設定して回答させた。

原稿 1 から 10 の要約率を表 1 に示す。要約率の算出は、

$$\text{要約率} = \frac{\text{要約文の全文字数}}{\text{入力文の全文字数}} \times 100(\%)$$

で行った。表 1 中の原稿文字数は、入力原稿の文字数 (記号、句読点等も 1 文字と数える) を表す。表 2 に、自然さ、忠実度、非冗長度の数値による評価に対する有効回答数を示す。また、付録に、原稿 5 および原稿 9 の要約結果を示す。

6.2 要約文の自然さ

要約文は字幕として表示されることを前提としているため、できるだけ読みやすい文章であることが求められる。このため、要約文が文章として自然であることが重要である。被験者が要約文の自然さを評価する際の判断基準を以下のように設定、提示し、要約文の自然さを尋ねた。

5 点: ほぼ自然である。つまり、このような文章を書く人間もいると考えられる。

表 3 要約文の自然さの評価

評価値	5	4	3	2	1	0	平均点
原稿 1(%)	37.5	59.4	3.1	0	0	0	4.34
原稿 2(%)	25.0	43.8	25.0	6.3	0	0	3.88
原稿 3(%)	34.4	53.1	12.5	0	0	0	4.22
原稿 4(%)	18.8	50.0	31.3	0	0	0	3.88
原稿 5(%)	87.5	9.4	3.1	0	0	0	4.84
原稿 6(%)	36.7	43.3	16.7	3.3	0	0	4.13
原稿 7(%)	20.0	43.3	36.7	0	0	0	3.83
原稿 8(%)	23.3	30.0	40.0	6.7	0	0	3.70
原稿 9(%)	0	31.0	48.3	17.2	3.4	0	3.07
原稿 10(%)	76.7	23.3	0	0	0	0	4.77
原稿 1..10(%)	36.2	38.8	21.4	3.2	0.3	0	4.07

0 点: 非常に不自然である。つまり、文もしくは文章全体としてのまとまりがなく、人間が用いないような表現が頻繁に見られる。

なお、入力原稿は自然であると仮定している。自然さの評価は、あくまで削除によって不自然になった場合のみに注目して行うよう指示した。また、自然さに関しては、文全体で不自然とを感じる場合もあり、被験者の判断が分かれることが予想されたため、文中の不適切な箇所の特定は難しいと判断し、行わなかった。

表 3 に要約文の自然さの評価結果を示す。表は、各原稿に対して、1 から 5 のそれぞれの評価値を与えた被験者の割合 (各割合は小数点以下第 2 位を四捨五入しており、足して 100% にならない場合がある) と、各原稿の評価値の平均値を示している (1..10 は 1 から 10 の全ての原稿で見た値)。全体の平均点は 4.07 と、良好な値を得た。特に原稿 5 は高い評価値を得ている。しかし、原稿 9 においては、「ほぼ自然」と判断した被験者は全くおらず、どのような記事に対しても一様に自然な要約を行うには、問題が残る。原稿 9 については、6.3 節で詳述する。

6.3 要約文の忠実度

要約文は、原文の重要部を適切に抽出している必要がある。重要部抽出の適切さを忠実度と表記し、以下のような判断基準を設けて忠実度を尋ねた。

5 点: このような抽出を行う人間もいると考えられる。

0 点: 重要部の欠落が非常に頻繁にみられる。

要約文の文章としての自然さが損なわれた場合でも、内容が伝われば、ニュース字幕としてある程度有用である。このような観点から、要約文の忠実度は、原則的に要約文の文章としての自然さを無視した上で評価するよう注意させた。また、4 点以下を与えた被験者には、欠落した重要部のうちで最も重要だと考えられる箇所から順に最大 5 箇所を指摘させた。

表 4 要約文の忠実度の評価

評価値	5	4	3	2	1	0	平均点
原稿 1(%)	12.5	65.6	21.9	0	0	0	3.91
原稿 2(%)	9.4	43.8	31.3	15.6	0	0	3.47
原稿 3(%)	15.6	59.4	21.9	3.1	0	0	3.88
原稿 4(%)	3.1	46.9	40.6	6.3	3.1	0	3.41
原稿 5(%)	54.8	25.8	16.1	3.2	0	0	4.32
原稿 6(%)	16.7	36.7	33.3	6.7	6.7	0	3.50
原稿 7(%)	3.3	36.7	50.0	3.3	6.7	0	3.27
原稿 8(%)	24.1	44.8	27.6	3.4	0	0	3.90
原稿 9(%)	0	31.0	51.7	13.8	3.4	0	3.10
原稿 10(%)	43.3	50.0	6.7	0	0	0	4.37
原稿 1..10(%)	18.2	44.3	30.0	5.5	2.0	0	3.71

表4に要約文の忠実度の評価結果を示す。全体の平均点は3.71で、良好ではあるが、自然さ、非冗長度(6.4節参照)に比べて低い値となっている。また、自然さ、非冗長度の評価に比べ、1つの原稿に対する評価値にばらつきが目立ち、被験者によって重要部の認定に関する判断が異なることがうかがえる。

6.2節の自然さ同様、原稿9の忠実度が最も低いと判定された。削除するのが不適切と指摘された箇所特に数の多かったものは、「[日本を訪れる] 外国人」(14名)、「(文頭)[ところが]」(15名)、「[外国人労働者の多くが働いている] 建設現場」(10名)、「[バブル経済の崩壊が建設業界にも及んだ] 影響で」(28名)、「[売り上げにもやや] 陰りが出始めていますが」(18名)であった。特に、「[バブル...] 影響で」の部分は、重要と指摘した28名の被験者のうち22名が(指摘箇所の中で)最も重要と答えている。本要約手法では、あらかじめ登録した形式的表現に係る文節を優先採用しているため、「影響」を新たに形式的表現に登録することで、このような不適切な削除は防ぐことができる。また、6.2節で原稿9の自然さが低いと評価されたのは、通常修飾語句を伴う「影響」に係る連体修飾句の削除や、「ところが」という接続詞の削除による文の結束性の低下が、不自然な印象を与えたためと考えられる。また、「ところが」は、それ自身では何も情報を伝えないが、適切な要約を生成する上で不可欠であると被験者が判断したため指摘されたと考えられる。

6.4 要約文の非冗長度

要約文生成は、原文中の比較的冗長であると考えられる部分の削除によって達成される。冗長部が残存しない度合を非冗長度と表記し、被験者に以下のような判断基準を与えて非冗長度を尋ねた。

- 5点: 不要である部分はほとんどない。つまり、これ以上の削除は難しいと考えられる。
- 0点: 不要部の残存が非常に頻繁にみられる。

表 5 要約文の非冗長度の評価

評価値	5	4	3	2	1	0	平均点
原稿 1(%)	31.3	56.3	12.5	0	0	0	4.19
原稿 2(%)	18.8	65.6	12.5	0	0	3.1	3.94
原稿 3(%)	18.8	50.0	28.1	3.1	0	0	3.84
原稿 4(%)	37.5	31.3	31.3	0	0	0	4.06
原稿 5(%)	43.8	43.8	12.5	0	0	0	4.31
原稿 6(%)	36.7	50.0	13.3	0	0	0	4.23
原稿 7(%)	33.3	60.0	6.7	0	0	0	4.27
原稿 8(%)	37.9	41.4	20.7	0	0	0	4.17
原稿 9(%)	55.2	37.9	3.4	3.4	0	0	4.45
原稿 10(%)	31.0	58.6	10.3	0	0	0	4.21
原稿 1..10(%)	34.2	49.5	15.3	0.7	0	0.3	4.16

6.3節同様、要約文の非冗長度は、要約文の文章としての自然さを無視した上で、評価させた。また、4点以下を与えた被験者には、冗長な残存部のうちで最も冗長だと考えられる箇所から順に最大5箇所を指摘させた。

表5に要約文の非冗長度の評価結果を示す。全体の平均点は4.16で、自然さ、忠実度と比較して最も高い値を得た。非冗長度を3以下と判断した被験者は全体で16%程度と低く、要約結果は冗長さを解消していると言える。ただ、最も要約率が悪い原稿5に関しても非冗長度の値は高く、ニュース原稿自体がそれほど冗長ではないという被験者の判断がうかがえる。また、重要箇所の指摘に比べ、冗長箇所は全般に指摘が少なく、非冗長度の評価値を低いと判定した被験者にも同様の傾向がみられた。このことは、冗長と感ずることはあっても、いざ削除するとすると、人手であっても冗長箇所を特定することが難しいことを示していると考えられる。

原稿3は非冗長度が最も低いと判定された。しかし、原稿3の評価値を3以下と判定した被験者の冗長部の指摘箇所は一定せず、まちまちであった。このことから、冗長箇所の特定の難しさがうかがえる。原稿3の冗長部指摘に関して、最も一致した意見は、第1文を全て削除するという指摘だった(4名)。ニュース記事は多くの場合第1文が全体の概要になっているが、特に原稿3は第1文と第2文の内容がほとんど同じであり、これが冗長であるという印象を与えた原因であると考えられる。また、原稿2の非冗長度を0と判定した被験者がいるが、この被験者も原稿2の第2文を全て削除するという指摘をしている。

7 考察

ここでは、実験で明らかになった問題点や、有効であった点等を述べる。

本要約手法による冗長部認定法に関して、以下のような事項が観察された。

- 本要約手法では原則的に修飾部や例示などを表す部分を冗長部と認定しているが、アンケート結果より、その認定法が十分でないことが分かった。また、従来要約文において

は比較的不要とされてきた固有名詞への修飾部が重要と判断されるなど、重要な修飾部の認定の難しさが明らかとなった。以下、明らかになった(1)連体修飾部、(2)固有名詞への修飾、(3)例示を表す部分に関する問題点について述べる。

- (1) 本要約手法では、修飾節に含まれる重要語や、直後が形式的表現かどうかを調べることにより、修飾部の誤った削除を避けているが、忠実度の評価値を見ても、その精度が十分ではないことがわかる。しかし、修飾部に限っても、どの修飾部が重要であるかの判定は容易でない。人手による要約文においては、非内容的で付随的である「内の関係」の連体修飾語は削除される傾向にあることが指摘されているが(佐久間 1994)、アンケート結果では、「内の関係」の連体修飾語を重要部として指摘する例が少なくなかった。例えば、6.3節で挙げた「[日本を訪れる] 外国人」,「[外国人労働者の多くが働いている] 建設現場」などがそれにあたる。この場合、「外国人」や「建設現場」を形式的表現と解釈するのは不適切であると考えている。
- (2) また、一般に、固有名詞に係る修飾語句は削除可能とされるが、アンケート結果から、固有名詞に係る修飾部が必ずしも削除可能ではないことがわかった。原稿4「...[冷戦後の大幅な核軍縮を早期に実現するための前提条件になっている] START 1・第一次戦略兵器削減条約を批准しこれで、...」(4名)、原稿7「[手当てにあたった] ○○大学付属家畜病院の××講師は...」(3名)などが重要部として指摘されている。
- (3) 本要約手法では、例示を表す「など」を伴う部分は比較的冗長と判断するため、削除される場合が多い。しかし、アンケート結果では「など」を伴う部分が重要と判断される場合も多かった。例えば、原稿10は、忠実度において最も高い評価を得ているが、同原稿中で削除された、「...[医薬品や注射針など] 総額一千三百万円相当の緊急援助を...」(3名),「...[首都のルサカなど] 全土に広がっています。」(4名),「...[コレラの治療薬一万人分と、注射針二百箱など、] 合わせて一千三百万円相当の緊急援助物資を...」(6名)といった箇所が重要部として指摘されている。

これらの例で指摘されている部分は、付随的ではあるが、記事の背景を特定するなど、読み手の理解を助ける働きを持つため、被験者によって重要であると判断されたことが予想される。字幕文は画面上で一方的に流されるため、理解を助ける部分は読みやすさを保持する意味でも重要であると言える。このような重要な修飾部(および例示等を表す部分)を認定する新たな手法を検討する必要があるが、表層的な情報のみを用いた有効な対策は見つかっていない。

- 6.4節で述べたように、アンケート結果からは、人手によっても冗長箇所を特定すること

は必ずしも容易でないことがうかがえる。このため、本要約手法の指摘する冗長箇所を人手による字幕文生成の支援として補助的に用いることで、字幕作成者の負担を軽減することができると考えられる。

- 実施したアンケートでは、質問項目以外に被験者の意見・感想を求めた。その中に、「第1文の内容は、第2文以降で詳述されるため、第1文は大胆に省略すべき」という意見があったが、これは6.4節の分析結果と一致する。しかし、本要約手法は、音声認識処理および要約処理をリアルタイムに行う字幕生成に応用することを最終的な目標としており、第1文の要約時には第2文以降の情報は用いないことが前提になるため、このような第1文の冗長性を判定することはできない。

本要約手法は、画面との対応を考え、文間の重複部の削除を行っていないが、ニュースを字幕(文字)から得る場合には、重複する内容を冗長と感ずる場合が観察された。第2文以降の重複部の削除が適切か、検討する必要がある。

評価結果では、自然さ、忠実さ、非冗長さのそれぞれについて、良好な評価値が得られたが、以上に示すように、冗長部の認定について、より精度の高い手法を検討する必要があることが分かった。

本要約手法は、生成された要約文の自然さを重視し、構文構造を破壊するなどの不適切な削除を防ぐ目的で簡易構文解析手法を用いた。現状の技術では高い精度で厳密な構文解析を行うのが難しいため、簡易構文解析では、できるだけ表層的な情報を用いた、厳密でない係り受け解析を行っている。この結果、実施したアンケートでは、要約文の自然さについて良好な評価結果が得られた。このことから簡易構文解析結果はおおむね良好であると言えるが、前述の通り冗長部の認定は難しく、厳密で精度の高い構文解析が実現できたとしても、重要部の特定という、より難しい問題が残される。

また、前述の冗長部認定に関する考察に関連して、簡易構文解析は1文中の最後の用言(述部)を核とし、その必須格等を残すことにより自然な要約文を生成するが、連体修飾句が新情報を表す場合など、むしろ修飾部が重要である場合も考えられる。

8 おわりに

より自然な字幕文の生成を目標に、主に修飾部を削除することによる、ニュース文の1文ごとの自動要約を試みた。また、要約結果の有用性を評価するために、アンケート調査を行い、良好な評価値を得た。しかし、構文解析の失敗や、重要部の欠落もあり、前もって想定できない様々な表現を含む、幅広い入力原稿に対して一様に精度の高い要約を行うには、より高度な処理を行う必要がある。特に、限定修飾の認定は難しく、本要約手法では形式的表現をあらかじめ列挙することなどで対処した。より高い精度を実現するには、領域に依存しない知識の利用が必須であると考えられる。

また、修飾部および例示を示す箇所の削除だけでは、高い削減率を得るのは難しく、精度の高い構文解析が実現しても、必ずしも品質の高い要約は得られないため、今後は文脈解析等、より高度な解析の有用性、実現性を検討する必要がある。しかし、高度な解析は、前段階の処理で生じた解析誤りをさらに拡大するなど、問題も多くはらむことが予想され、その対処法を考える必要があろう。

9 謝辞

本研究でシソーラスに使用した「角川類語新辞典」(大野・浜西 1981)を機械可読辞書の形でご提供いただき、その使用許可をいただいた(株)角川書店、および、KNPに関する問い合わせに懇切にお答えいただいた京都大学大学院情報学研究科黒橋禎夫先生に深謝する。なお、本研究の一部は文部省科学研究費基盤研究(B)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。

付録

A 原稿 5

原稿名：恩給法改正案

政府は、きょう（五日）の閣議で、[退職公務員などに支給される恩給を、今年の四月から年額二点六六パーセント引き上げる]恩給法の改正案を決定し、国会に提出することになっています。

恩給は、[公務員給与の改定や]消費者物価の上昇などに伴って毎年引き上げられており、平成五年度についても、去年暮れの予算編成で、恩給年額と各種の最低保障額を二点六六パーセント引き上げることが決まっています。

具体的には、長期在職者の場合の普通恩給では最低保障額が、六十五歳未満で二万五百円上がって七十九万千百円に、六十五歳以上七十五歳未満で二万七千三百円上がって百五万四千八百円に、また七十五歳以上では、引上げ率がさらに上乘せられて、三万二千五百円上がって百六万円になります。

政府は、[こうした引き上げを盛り込んだ]恩給法の改正案を、きょうの閣議で決定した上で国会に提出することになっていますが、各党とも反対はないことから、改正案は今年度内に成立し、引き上げは今年四月から実施される見通しです。

B 原稿 9

原稿名：リレーニュース・外国人の地下足袋

[東京からは、建設現場などでの作業に欠かせない] 地下足袋が、外国人労働者の増加に伴って、[二十九センチや三十センチといった] 特大のものが売れているという話題です。

仕事を求めて[日本を訪れる] 外国人の数は昭和五十五年左右から急激に増え始め最近では全国各地で外国人労働者の姿を見掛けるようになりました。

[ところが外国人労働者の多くが働いている] 建設現場で困ったことが起きました。

[足場を気にしながらの作業に欠かせない] 地下足袋が、日本人用のものでは小さくて履けない、ということです。

[こうした注文をきいた] 東京・日本橋のメーカーでは、[十年ほど前から通常のものより一回りも二回りも大きい] 木型を新たに用意して、[二十九センチや] 三十センチという特大の地下足袋を作りはじめました。

この特大の地下足袋はビルの建設ラッシュも手伝って順調に売れ行きを伸ばしてきました。

[都内のある履物店には、イランなど] 中東出身の若者達が特大の地下足袋を求めて毎日のように訪れていますが[中にはパーティーや] ジョギング用にとスーツ姿の外国人も買っていくということです。

[最近は大バブル経済の崩壊が建設業界にも及んだ] 影響で、[特大の地下足袋の売り上げにもやや] 陰りが出始めていますが、業界では不景気になれば今度は公共工事が増えて、また地下足袋の売り上げが伸びるのではないかと期待しています。

参考文献

- Cowie, J. and Lehnert, W. (1996). "Information extraction." *Communications of the ACM*, **39** (1), 80-91.
- 江原暉将, 沢村英治, 若尾孝博, 阿部芳春, 白井克彦 (1997). "聴覚障害者のための字幕つきテレビ放送制作への自然言語処理の応用." 言語処理学会 第3回 年次大会発表論文集, pp. 489-492.
- 宮坂栄一 (1998). "人にやさしい放送のための技術." *NHK 技研 R & D No.52*, pp. 20-35.
- 大野晋, 浜西正人 (1981). 角川類語新辞典. 角川書店.
- 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.
- Paice, C. (1990). "Constructing literature abstracts by computer: Techniques and prospects." *Information Processing and Management*, **26** (1), 171-186.

佐久間まゆみ (1994). 要約文の表現類型. ひつじ書房.

住田一男, 知野哲朗, 小野顕司, 三池誠司 (1995). “文書構造解析に基づく自動抄録生成と検索提示機能としての評価.” 電子情報通信学会論文誌, **J78-D-II** (3), 511-519.

若尾孝博, 江原暉将, 白井克彦 (1997a). “テレビニュース番組の字幕に見られる要約の手法.” 情報処理学会研究報告 NL-122-13, pp. 83-89.

若尾孝博, 江原暉将, 村木一至, 白井克彦 (1997b). “テレビニュース番組電子化原稿を題材とした自動要約手法の大規模評価.” 情報処理学会研究報告 NL-119-6, pp. 31-36.

若尾孝博, 江原暉将, 白井克彦 (1998). “短文分割を利用したテレビ字幕用自動要約.” 情報処理学会研究報告 NL-126-9, pp. 63-68.

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), 39-56.

山崎邦子, 三上真, 増山繁, 中川聖一 (1998). “聴覚障害者用字幕生成のための言い替えによるニュース文要約.” 言語処理学会 第4回 年次大会発表論文集, pp. 646-649.

略歴

三上 真: 1999年 豊橋技術科学大学大学院修士課程修了. 現在, (株) 日立ソフトウェアエンジニアリング勤務. 在学中は, 自然言語処理, 特に TV ニュース字幕生成のための要約の研究に従事.

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, 及び, 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.

中川 聖一: 1976年 京都大学大学院博士課程修了. 同年 京都大学情報工学科助手. 1980年 豊橋技術科学大学情報工学系講師. 1983年 助教授. 1990年 教授. 1985~1986年 カーネギメロン大学客員研究員. 工博. 1977年 電子通信学会論文賞. 1988年度 IETE 最優秀論文賞. 著書「確率モデルによる音声認識」電子情報通信学会 (1988年), 「情報理論の基礎と応用」近代科学社 (1992年), 「パターン情報処理」丸善 (1999年) など.

(1998年9月30日 受付)

(1998年12月17日 再受付)

(1999年1月25日 採録)