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

山崎 邦子† 三上 真† 増山 繁† 中川 聖一‡
kuniko@smlab.tutkie.tut.ac.jp, mikami@smlab.tutkie.tut.ac.jp,
masuyama@tutkie.tut.ac.jp, nakagawa@slp.tutics.tut.ac.jp
†豊橋技術科学大学 知識情報工学系
‡豊橋技術科学大学 情報工学系

1 はじめに

現在，日本におけるテレビの字幕付き番組の放映割合は10%程度と少ないが [1]，高齢化社会の到来への対応，高度福祉社会の実現をめざし，字幕の普及が強く期待されている．そのため，字幕を自動的に作成する手法の確立が急務の課題となってきた．本研究ではニュース番組における字幕生成のための一要約法を提案する．

日本語文章を対象とした計算機による要約の研究としては [2, 3, 4] がある．[2] は，論説文を対象としており，文章から要約に含ませるべき重要文を見つけ出す手法で，[3] は，新聞記事を対象とし，重要文を抜き出す手法である．重要部を文単位で抜き出す場合，長い文の多いニュース文に対しては情報の大きな欠落が予想され，また，長い文が字幕に流れることは好ましくない．ニュース原稿には段落という概念がないため，[4] も適用できない．

テレビニュース番組の字幕の要約手法を研究したものに [1, 5, 6] があり，特に [6] で提案されている手法は，本研究の手法の一部と一致するところもある．しかし，いずれも予備実験の段階であり，自動要約には至っていない．

また，言い替えによる要約手法としては，物語文を対象とした [7] がある．エピソードを削ることなく，言い替えにより字数を減らすという考え方は同じであるが，言い替えの対象を動作に限定している点で，本手法とは異なる．このような自動要約の現状を踏まえ，本研究では，ニュース文に特化した，語句の言い替えによる要約手法を提案する．

ニュース文は，話し言葉であるため，冗長な表現が多い．それに対し，字幕は体言止めや漢字熟語も多用でき，情報を減らさずに短く表現することが可能である．また，ニュース文においては，文章全体にわたって一様な，一文ごとの要約が望ましい．これらより，特定の表現とそれに対する言い替え (これを規則と呼ぶ) を予めテーブル (以下，換言テーブル) に作成しておき，これと原稿とをマッチングさせ，短縮することを基本方針とし，語句の言い替えを規則の形で表す換言テーブルを作成した．なお，JUMAN[8] を用いて形態素解析することで，語尾の変化に対処し，品詞の情報により換言テーブル中の規則をまとめた．

2 要約手法

本研究で語句の言い替えのために用いた手法は，大きく以下のように分けられる．

- 原文を直接言い替え
- 形態素解析を利用した言い替え
- ニュース文のパターンに基づく言い替え

以下，これらについて順次説明する．

2.1 原文を直接言い替え

ニュース文に頻出する冗長な表現をまとめて換言テーブルを作成し，原文のままの表現とマッチングする．換言テーブルは特定の表現とそれに対する言い替えの規則から成る．語句を言い替える基本方針を以下に列挙する．

- 冗長な表現を短縮し，略語が存在する語は略語に言い替える
- 無くても理解できる部分は削除する
- 強調表現や文の接続を表す語句なども削除する

表1に換言テーブルの一部を示す．テーブルはNHK 汎用データベース7日分 (632記事) から人手で作成した．規則数は約500個である．

表 1: 換言テーブル

言い替え前の表現	言い替え後の表現
から逃げ出しました	から逃亡
良い天気に恵まれ、	晴天で、
総理大臣	首相
になりそうです。	に。
の疑いがもたれています。	の疑い。
進出を果たしました	進出
念願の	(削除)
このため、	(削除)
⋮	⋮

2.2 形態素解析の利用

2.1節で述べたように，文字列を置き換えただけでは，語尾の変化等に対応できない．そこで，形態素解析の出力である動詞の終止形及び品詞を利用した．語尾の変化に対処し，語の品詞を利用することで，「(サ変名詞) しました」から「(サ変名詞)」への言い替えなどを統一して扱った．また，前後の品詞を見ることにより，不適切な言い替えを防いだ．

形態素解析を利用した言い替えは，以下のように分類される．

- 動詞の終止形
- 接続詞
- 文末のサ変名詞
- 文中のサ変名詞
- 読点の前のサ変名詞

以下，それぞれについて詳述する．

動詞の終止形

動詞の次が動詞性接尾辞 [8] だけならば，動詞の終止形に換言テーブルをマッチングさせる．例えば，「許されます」や「許されています」を「許可」に言い替える場合，「許す→許可」という一つの規則にまとめた．以下に示すようなものの終止形が「許す」で，これが表 2のような終止形の換言テーブル (項目数２６) にマッチし，「許可」に言い替えられる．

許されます。
許されています。
許しています。
許します。
許しました。
} 許す。 ⇒ 許可。

表 2: 終止形の換言テーブル

言い替え前の表現	言い替え後の表現
許す	許可
求める	要求
調べる	調査
開く	開催
⋮	⋮

接続詞

語の品詞が接続詞と分かったものは，削除した．なお，同様の手法が [6] でも提唱されている．

文末のサ変名詞

サ変名詞の直後の動詞の終止形が「する」で，その次の形態素が動詞性接尾辞だけならば，サ変名詞だけ残し，それ以降は削除する．なお，同様の手法が [6] でも提唱されている．

(例) を予定している→を予定
が出席しました→が出席

サ変名詞の後に助動詞ぬ型 (否定) がある場合，「サ変名詞＋せず」に言い替える．

(例) が出席しませんでした→が出席せず

文中のサ変名詞

サ変名詞の直後の動詞の終止形が「する」で，その次が動詞性接尾辞だけで，その次が名詞ならば，「サ変名詞＋の＋名詞」に言い替える．

(例) 祈願するおはらい→祈願のおはらい
開催される予定→開催の予定

読点の前のサ変名詞

サ変名詞の直後の動詞の終止形が「する」で，その次が動詞性接尾辞か述語接続助詞だけで，文章が区切られているならば，「サ変名詞。…」とする．

(例) 要約前: …発表しており、橋本総理大臣は…
要約後: …発表。橋本総理大臣は…

2.3 ニュース文のパターンに基づく言い替え

ニュース文に頻繁に見られるパターンを分析した結果，以下の部分は言い替え可能であると判断した．

- 「〜によりますと」と「〜で調べたところ」
- 「この〜は」
- 「きょう」

- 会話文の前後の部分
- 文末の「です」や「でした」

以下，それぞれについて詳述する．

「～によりますと」と「～で調べたところ」

この表現は文頭によく現れるが，削除可能と判断した．～の部分から文頭まで除く．

（例） 要約前: 鴨川警察署の調べによりますと、田中容疑者は、…
要約後: 田中容疑者は、…

「この～は」

この表現も文頭によく見られる．～の部分が名詞，もしくは，特殊であるときは削除する．

（例） 要約前: この「世界福祉イニシアティブ」は、先月リヨンで…
要約後: 先月リヨンで…

「きょう」

一回目に出現した「きょう」は「今日」と漢字に直し，二回目からは削除する．その際，その次が助詞である場合には，それも削除する．なお，[6] では，一回目も削除としているが，本手法では一回目は必要と判断し，二回目から削除した．

（例） 要約前: 新進党の小沢党首は、きょう、細川元総理大臣と会談して、…きょうの会談では、細川元総理大臣が…
要約後: 新進党の小沢党首は、今日、細川元総理大臣と会談して、…会談では、細川元総理大臣が…

会話文の前後の部分

『「』の前が『～は(この後に句読点があってもなくても)』で，～部が名詞もしくは名詞性接尾辞である場合には，『～談』に言い替える．『」』の次が『と』で，その次の動詞の終止形が『話す』，『述べる』，『する』のいずれかで，残りが動詞性接尾辞のみであるならば，『」』に続く部分を削除する．それ以外は『「』より前はそのままで，『」』の後は『と。』とする．

（例） 要約前: 奥田文部大臣は、「現在、中学校の教育実習は、…」と述べました。
要約後: 奥田文部大臣談「現在、中学校の教育実習は、…」。

文末の「です」や「でした」

文末の「です」，「でした」を，その前が名詞であるならば削除する．なお，同様の手法が [6] でも提唱されている．

（例） 示したものです。→示したもの。

3 実験結果

これまで述べた手法を Perl 言語でコーディングした．要約率の計算方法は，以下の通りである．文字数には，句読点などの記号をすべて含む．

$$\text{要約率} = \frac{\text{要約後の文字数}}{\text{テキストの文字数}} \times 100(\%)$$

オープンテストの結果，要約率は 91.2%(1260 記事平均) であった．

4 考察

形態素解析を用いて語尾の変化等に対処したことにより，体言止め等を用いた字幕文らしい要約文を生成することができた．字幕文は表現の統一が取れているほうが好ましいが，このような要求はほぼ達成できた．また，要約による情報の削減はほとんどなく，情報の保存にも成功した．

オープンテストでは，文の途中で言い替えられた部分に比べ，文末の言い替えによる要約が多く，短い文からなる原稿に対して，より要約率が向上する傾向が見られた．これを踏まえ，[9] では，重要部認定による要約手法と本稿が提案する言い替えの手法を併用した実験を行っている．文頭の言い替えとして，「～によりますと」を削除するものがあるが，事故を報道するニュース原稿などでは，このような表現が頻繁に用いられ，高い要約率が得られている．全体として，文末の言い替えの割合が高くなるのは，文末の言い替えがどの記事でもほぼ確実に行われるためと推測できる．

「～によりますと」に関しては，「(文頭) …参拝客が詰めかけ警察の調べによりますと初詣客は…」という表現が表れ，文頭から「によりますと」までがすべて削除される場合があったが，これは連用中止法にあらかじめ対処すれば回避できる．このように，言い替えの失敗のほとんどは，規則作成に使用する原稿の数を増やすことで対処可能である．また，インタビュー記事についてあまり良い要約率が得られなかったが，会話文は，言い替えにより不自然になる場合が

多く，規則を十分に作成できなかったことが原因であると考えられる．このように，ニュース文のさらに詳細な分析が重要である．

5 むすび

本研究では，言い替えによる要約手法を考案し，計算機実験を行った．[6] で設定されている要約率の目標値を参考に，70% 程度を目指した．しかし，情報の維持を考えると，言い替えだけではそれだけの要約率を得るのは難しいことが分かった．なお，重要部認定手法と本手法を併用することで，[9] では，この目標を達成した．今後の課題として，換言テーブルの充実と自動生成に言及し，研究報告とする．

• 換言テーブルの更なる充実

オープンテストの結果では，言い替え可能な部分がなお存在するため，更に換言テーブルを充実させる必要がある．また，本手法では「重傷を負いました」と「重傷でした」のような，内容は同じであっても表現が異なるものに対して対応できない．現状では，このような表現のゆれに対応できておらず，テーブル中の規則数を増やすことで対処していく方針である．

• 換言テーブルの自動作成

人手で換言テーブルを作成していたのでは，膨大な作業を要する．そのため，換言テーブルの自動作成を検討する必要がある．自動化の方法としては，予備的に以下の手法を検討している．

- EDR 電子化辞書 (日本語単語辞書)[10] の日本語概念説明を形態素解析する
- 入力の自立語の終止形と角川類語新辞典 [11] の小分類が同じである自立語の終止形を 2 つ以上含む日本語概念説明の単語見出しに言い替える

ここで，以下のような制限を加える．

- EDR 電子化辞書の頻度 (概念別頻度 / 単語別頻度) の分子が 0 であるものは，使用頻度の低い言葉であるので，言い替えとして不適切なため除く
- 単語見出しが入力の文字列の長さ以上であるものも除く

以上の手法を実装し，「軽いけが」 (今回作成した換言テーブルでは「軽傷」と換言) を入力として予備実験を行った．

その結果，「軽傷 (軽い傷)」と「軽症 (病気の症状が軽いこと)」と共に，「重傷」や「重病」など反対の意味である語も取り出された．これは，対義語である「軽い」と「重い」は小分類が同じであるためであるが，角川類語新辞典に対義語が掲載されているため，これを使用して対処する予定である．しかし，「軽傷」と「軽症」のうちどちらを選択するかなど，候補の中から適切な語を選び出す方法も検討課題として残る．また，入力の対象とする部分についても検討しなければならない．

謝辞

本研究を進めるにあたり，ニュース原稿を機械可読の形で提供いただき，その使用許可を頂いた NHK に深謝する．

参考文献

[1] 江原暉将, 沢村英治, 若尾孝博, 阿部芳春, 白井克彦：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 (1997).

[2] 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”, 情報処理学会研究報告, NL vol.2,No.1, pp. 39–56 (1995).

[3] 野本忠司, 松本裕治：人間の重要文判定に基づいた自動要約の試み, 情報処理学会研究報告, NL 120-11, pp. 71–76 (1997).

[4] 福本文代, 福本淳一, 鈴木良弥：文脈依存の度合を考慮した重要パラグラフの抽出, 情報処理学会研究報告, NL Vol.4 No.2, pp. 89–109 (1997).

[5] 若尾孝博, 江原暉将, 村木一至, 白井克彦：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 (1997).

[6] 若尾孝博, 江原暉将, 白井克彦：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 (1997).

[7] 近藤恵子, 奥村学：言い替えを使用した要約の手法, 情報処理学会研究報告, NL 116-20, pp. 137–142 (1996).

[8] 松本裕治, 黒橋禎夫, 山地治, 妙木裕, 長尾真：日本語形態素解析システム ＪＵＭＡＮ version3.1 使用説明書, 京都大学工学部長尾研究室 (1996).

[9] 三上真, 山崎邦子, 増山繁, 中川聖一：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 回年次大会併設ワークショップ (発表予定) (1998).

[10] EDR 電子化辞書 1.5 版使用説明書 (1996).

[11] 大野晋, 浜西正人：角川類語新辞典, 角川書店 (1981).