文単純化のための文字数圧縮規則

福島 孝博[1]
通信・放送機構（ＴＡＯ）
渋谷上原リサーチセンター
fukushima@shibuya.tao.go.jp

江原 暉将
ＮＨＫ放送放送技術研究所
／ＴＡＯ

白井 克彦
早稲田大学 ／ＴＡＯ

1. はじめに

自動要約の研究では、従来からテキストの重要部分を特定し、抽出することが行われて来た。その中でも、特に文を基本単位とする重要文抽出（sentence extraction）については、多くの研究がなされて来た。本研究では、重要部分を特定するのではなく、不要部分に焦点を置き、文中の不要部分を選び出し、削除することにより、文字数を削減し、文の単純化（sentence simplification）を行う手法について述べる。この文単純化の手法は、文中の文字数を削減（今後は「文字数圧縮」と呼ぶ）することにより実現され、そのために用いる規則は、文字列や品詞などの表層から得られる情報だけを用いており構文解析は行っていない。

また、本研究は、通信・放送機構（Telecommunications Advancement Organization of Japan 略称 TAO）の進める「視聴覚障害者向け放送ソフト制作技術研究開発プロジェクト」(http://www.shibuya.tao.go.jp) の一環として行われているものである。このプロジェクトの研究目的は、自然言語処理技術や音声認識の技術を用いて、字幕作成を効率的に行う技術を得ることにある（[1] [2] [3] ）。自動要約は、テレビ字幕を制作するために必要な要素技術の１つとして研究されており、テレビニュース番組の字幕用の文字数圧縮規則ついては、概要のみを発表している。

本文では、文字数圧縮規則の作成の過程を述べ、得られた規則について概要に留まらず詳細に説明するとともに、テレビニュース記事および新聞記事を使った文字数圧縮規則の評価の結果について報告する。

2. 背景

現在のテレビニュース番組では、ニュースキャスターの喋るスピードは、文字になおすと１分間に約３５０～３７０文字を読んでいる。一方、現在の字幕放送で使われている字幕の経験則では、１行最大１５文字、一度に最大２行までを、６秒間表示するのが限度であるとされている。これは、１分間に換算すると、３００文字であり、これが字幕の表示文字数の限度となる。つまり、ニュース番組の場合、ニュースキャスターの読むスピードが字幕の文字数の限度を越えており、何らかの方法で文字数を削減する必要が生じてくる。このため、文字数圧縮をする自動要約、文の単純化が必要となる。

また、テレビニュース用の字幕は、要約の種類としては、informative と言われるものになり、原文と要約文を照らし合わせることが出来ない形での要約である（[4] ）。このため、要約されたテキストの信用性を高いものにする必要があり、あまり多くの部分を削除できなくなり、文字数を圧縮する文単純化に向いている。

その文字数圧縮規則を得るために、まず、現在、実際にテレビニュース番組で使われている字幕について調査をした。現在放送されている数少ない字幕付きニュース番組として、ＮＨＫ教育テレビの「手話ニュース845」がある。この番組では、その日の主なニュースに字幕と手話が付与されている。字幕は、音声として読まれる原稿と比べると文字数や、文の構造などに制限があり、結果として音声原稿の要約となっている。この音声原稿と字幕に注目をして、そこで使われている要約の方

[1] 99年１月より苗字が「若尾」から「福島」となった。

法を調べた（[5]）。

この調査の結果、文字数を削減する為に使われている規則は、以下のように、大きく分けて５つのグループに分類出来る。

1. 文末の削除、言い換え
2. 文の一部を残す
3. 別の語句で言い換え
4. 文頭の語句の削除
5. 日時表現の削除

これらの規則は、２を除いて、基本的に文字列の表層からの情報をもとにして実行可能なものである。

3. 文字数圧縮規則

「手話ニュース845」の字幕を分析した結果を基にして拡充を行い、自動要約のシステムに取り入れた。拡充に当たっては、ＮＨＫ放送データベースから選んだテレビニュース原稿数百にを分析して、人手で規則を追加した。

以下にそれらの規則について、例を付けて説明を行う。

尚、以下の文字圧縮規則は、山崎らの研究（[6]）で発表されているものと重なるものもあるが、現在、我々の自動要約システムで実装されている規則を説明する。

3.1. 文末処理

- 文末の動詞がサ変動詞（但し、否定の表現は含まない）の場合は、そのサ変動詞以降を全て削除する。
  例 「増加しました。」 → 「増加。」
- 文末の動詞がサ変名詞＋「を」＋「する」（但し、否定の表現は含まない）場合は、そのサ変動詞以降を全て削除する。
  例 「宣誓をしました。」 → 「宣誓。」
- 丁寧助詞の「ます」が文末にある場合は、削除して適当な文末にする。
  例 「…なりました。」 → 「…なった。」
  「…訪れます。」 → 「…訪れる。」
- 特定の文末表現は、その表現を削除。
  例 「ということです」、「としています」、「ことにしています」、
  「ことになりました」など
  例 「…を求めていくことにしています」
  → 「…を求めていく」
- 特定の文末表現は、短い語句に言い換えている。
  例 「調べています。」又は
  「調べを進めています。」→ 「調査」
  「…ことがわかりました。」→
  「…ことが判明。」
- 特定の文末表現は、一部を残して削除する。
  例 「考えを示しました。」→「考え。」
  「見解を明らかにしました。」→「見解。」
- 文ば名詞性語句 ＋ 断定の助動詞「です」又は「でした」で終る場合、「です」、「でした」を削除する。
  例 「状況です。」→ 「状況。」
  「５０歳ぐらいでした。」→
  「５０歳ぐらい。」

3.2. 文頭処理

- 特定の文頭表現（特に接続詞）は、削除する。
  例 「一方」「その一方で」
  「このあと」「この結果」
  「これまでの調べによりますと」など

3.3. 文中の特定表現

- 意味を変えずに省略可能な表現は、省略形にする。
  例 「総理大臣」 → 「首相」
  「最高裁判所」→ 「最高裁」
  「きょう」 → 「今日」
  「日本訪問」 → 「訪日」
  「アメリカ軍」→ 「米軍」
- 特定の品詞に続いて省略可能な表現は、省略形にする。
  例 地名＋「警察署」→ 地名＋「署」
  「鴨川警察署」→ 「鴨川署」
  地名＋「地方検察庁」→ 地名＋「地検」
  「大阪地方検察庁」→ 「大阪地検」
- 省略形と省略をしない長い形での語句がペアで出てくる場合は、省略形だけを残す。
  例 「ＥＵ＝ヨーロッパ連合」→「ＥＵ」
  「アセアン＝東南アジア諸国連合」
  → 「アセアン」
  「大規模小売店舗法＝大店法」
  → 「大店法」
- 括弧でくくられたカタカナ文字列は、削除する。

例「大洗漁港（オオライギョコウ）」 → 「大洗漁港」

- 括弧でくくられた数字列は、削除する。
  例「…容疑者（４９）」 → 「…容疑者」

3.4. その他

- 文頭が「問い合わせ先」で始まり、その後が電話番号だけの場合は、削除する。
- 「ありません」は「ない」と言い換える。

これらの規則は、正規表現的に記述されており、規則実行する部分とは切り離して、改良、管理を容易にしている。

規則中では、文字列そのもの、品詞、また、文字列や品詞をグループにまとめたもの、正規表現で使う「*」（ワイルドカード）などが使える。規則数は、全部で約２００であり、内訳は表 1の通りである。

分類	規則数
文末処理	93
文頭処理	42
文中の特定表現	71
その他	2

表 1　文字数圧縮規則内訳

4. 評価結果

文字数圧縮規則をテレビニュース番組原稿及び新聞記事を対象にして、どれだけ文字数が削減出来るかを調べた。

テレビニュース原稿は、ＮＨＫ放送データベースから選んだものであり、１９９２年の原稿１０００記事である。尚、これらの記事は、文字数圧縮規則の拡充の際に用いた記事とは、異なるものである。一方、新聞記事は、毎日新聞（１９９５年版 [7] ）から１０００記事を選んだ。

文字数圧縮に際しては、同じ文字数圧縮の規則をＴＶニュース原稿、新聞記事の双方に適用した。その一例を論末に付録として載せている。

また、評価の尺度は、次の式で計算される圧縮率を用いた。

圧縮率 ＝ 文字数圧縮後の文字数 ／ 元記事の文字数

評価の結果は、表 2及び図 1の通りである。

	元原稿文字数	圧縮後文字数	圧縮率
ＴＶニュース	528,648	491,825	0.9303
新聞記事	546,614	536,557	0.9816

表 2　文字数圧縮の評価結果

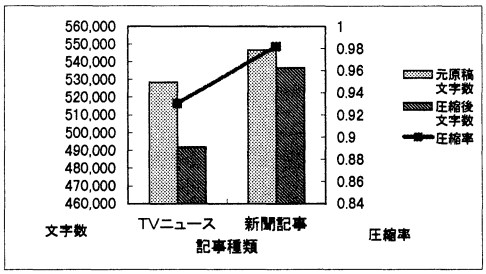

図 1　文字数圧縮結果（グラフ）

この結果から、現在の文字数圧縮規則は、テレビニュース原稿では、圧縮の効果があるが、新聞記事では、余り効果がないことが判明した。この理由は、テレビニュース原稿と新聞記事のスタイルの違いにある。相違点をあげてみると：

- 新聞記事では、丁寧さを示す「ます」「ました」があまり使われていない。一方、テレビニュース原稿は、基本的に「です、ます」調である。
- テレビニュース原稿では、文末に「…と言うことになっています」などの特有の表現が使われているとこが多いが、新聞記事では、このようなことはない。

新聞記事は、簡潔な表現がはじめから使われているのに対して、テレビニュース原稿では、必ずしもそうではなく、独特の表現が使われている。現在の文字数圧縮規則は、このテレビニュース原稿独特の表現を削除または、簡潔な表現に言い換えることにより、文字数を減らしているため、文字数圧縮の効果がみられることになる。

5. おわりに

本文では、文字数を圧縮することにより、文を単純化する手法を説明した。文字数圧縮のための規則を詳細に解説するとともに、その

規則を使って、テレビニュース及び新聞記事を題材として、圧縮の効果を調べた。

今後は、簡易構文解析を利用しての自動要約の研究([8] [9])を参考にし、文字より大きな単位である文節を基本単位とする自動要約の手法について研究を進めて行く予定である。

付録

[テレビニュース原稿（オリジナル）
太文字が圧縮の対象]

緊急時に障害者を施設や病院で受け入れる緊急一時保護制度などを充実してほしいと、東京・足立区の障害者の親のグループが、きょう東京都に**要請しました。**

要請したのは、東京・足立区の障害者の親二千人が作るグループ「なかま」の代表・鈴島妙子**（スズシマタエコ）**さんら七人です。

鈴島さんたちはきょう午前、東京都衛生局の母子保健課を訪れ、長野みさ子課長らに要請書を**手渡しました。**

それによりますと、障害者の親や家族が、病気やけがなどで介護ができなくなったときに、施設や病院で障害者を受け入れる東京都の緊急一時保護制度は、今のままでは受け入れ枠が狭いことや、手続きに時間がかかることなどから本当に必要なときに利用できないとして、足立区内に施設を作ることや、利用する時の手続きを簡単にすることなどを**求めています。**

また医療の面について、重度の障害者は一般の病院では診療を拒否されることが多く、急な病気や発作などが起きてもすぐに行けるところがないため、二十四時間体制の専門医療施設を地域に作ってほしいと**求めています。**

要請のなかで鈴島さんたちは、障害を持つ子供を受け入れる施設が見つからなかったために、母親が入院できないまま病気で亡くなった家庭もあったことなどを伝え、要望の緊急性を**訴えました。**

これに対して、長野母子保健課長は「難しい面が多いが、都としても努力はしていきたい。」と**答えていました。**

[圧縮の結果　圧縮率 0.913]

緊急時に障害者を施設や病院で受け入れる緊急一時保護制度などを充実してほしいと、東京・足立区の障害者の親のグループが、今日東京都に要請。

要請したのは、東京・足立区の障害者の親二千人が作るグループ「なかま」の代表・鈴島妙子さんら七人です。

鈴島さんたちは今日午前、東京都衛生局の母子保健課を訪れ、長野みさ子課長らに要請書を手渡した。

障害者の親や家族が、病気やけがなどで介護ができなくなったときに、施設や病院で障害者を受け入れる東京都の緊急一時保護制度は、今のままでは受け入れ枠が狭いことや、手続きに時間がかかることなどから本当に必要なときに利用できないとして、足立区内に施設を作ることや、利用する時の手続きを簡単にすることを要求。

医療の面について、重度の障害者は一般の病院では診療を拒否されることが多く、急な病気や発作などが起きてもすぐに行けるところがないため、二十四時間体制の専門医療施設を地域に作ってほしいと要求。

要請のなかで鈴島さんたちは、障害を持つ子供を受け入れる施設が見つからなかったために、母親が入院できないまま病気で亡くなった家庭もあったことを伝え、要望の緊急性を訴えた。

長野母子保健課長は「難しい面が多いが、都としても努力はしていきたい。」と答えていた。

参考文献

[1] 江原、沢村、若尾、阿部、白井（1996）「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pp 489-492.

[2] Wakao, T., Ehara, E., Sawamura, E., Abe, Y., Shirai, K. (1997) *Application of NLP technology to production of closed-caption TV programs in Japanese for the hearing impaired* In Proceedings of ACL 97 workshop, Natural Language Processing for Communication Aids, pp 55-58.

[3] Wakao, T., Ehara, E., Shirai, K. (1998) *Project for production of closed-caption TV programs for the hearing impaired,* In Proceedings of 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (Coling-ACL98), pp. 1340 – 1344.

[4] 奥村、難波（1998）「テキスト自動要約に関する研究動向」、「自然言語処理と情報提示技術」講習会資料、電子情報通信学会 言語理解とコミュニケーション研究会、pp. 1-24.

[5] 若尾、江原、白井（1998）「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、自然言語処理研究会、NL-122-13.

[6] 山崎、三上、増山、中川（1998）「聴覚障害者用字幕生成のための言い換えによるニュース文要約」言語処理学会第四回年次大会発表論文集、pp 646-649.

[7] 毎日新聞（1995）. CD-毎日新聞９５版、（株）毎日新聞社

[8] 三上、山崎、増山、中川（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」 言語処理学会第四回年次併設ワークショップ「テキスト要約の現状と将来」論文集, pp. 14-21.

[9] Grefenstette G. (1998) *Producing intelligent telegraphic text reduction to provide an audio scanning service for the blind,* In Working Notes of the AAAI Spring Symposium on Intelligent Text Summarization, pp. 111-117.