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

江原 暉将　沢村 英治　若尾 孝博　阿部 芳春　白井 克彦
NHK/TAO　TAO　TAO　三菱電機/TAO　早稲田大学/TAO
eharate@strl.nhk.or.jp

1　はじめに

聴覚障害者がテレビ放送を楽しむことができるように、番組中の音声を元に字幕を作成し、テレビ放送に多重して放送する字幕つきテレビ放送（以下、字幕放送という）が行われている。このような字幕放送は、欧米、特に米国では、７０％以上の番組に付加されているが、日本では、１０％程度と、少ない状況にある。こうした低い字幕放送率の原因の一つは、字幕の付与に人手がかかり、コストが高いことにある。このような現状を打開するための一手段として、郵政省は通信・放送機構（Telecommunications Advancement Organization of Japan）のプロジェクトである「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」を平成８年度から５年の期間で、２億円／年の予算をかけて発足させた。本プロジェクトの研究目的は、音声処理技術や自然言語処理技術を用いて、字幕作成を効率的に行う技術手段を得ることにある。

本文では、発足したプロジェクトの研究項目、研究計画を述べるとともに、これまでに行った予備的実験の結果についても触れる。

2　具体期な研究項目

本プロジェクトの具体期な研究項目は、以下のとおりである。

- 自動要約技術の研究
- 自動同期技術の研究
- 字幕作成システム技術の研究

本プロジェクトの研究成果を利用することで、図1に示すような字幕放送制作システムが得られる。なお、本プロジェクトでは、テレビ放送番組全般を対象にしているが、第１のターゲットはニュースを中心とする情報番組である。そこで、本文では、もっぱらニュース番組を対象にして記述する。

自動要約技術　ニュース番組のかなりの部分には、あらかじめ、原稿が存在し、アナウンサーがその原稿を読み上げる方式のものが多い。そのため、原稿から字幕を作成する方法が有効である。しかし、アナウンサーの発声する音声の話速は、1分当たり400文字程度であり、かなり高速である。そのため、アナウンス音声をすべて字幕とすると読み切れない可能性がある。そこで、ニュース原稿を要約して字幕とする方法が考えられる。自動要約技術の研究では、このような要約を自動的または半自動的に行って、字幕作成作業の効率化を図ることを目的としている。

自動同期技術　出来上がった字幕データはアナウンスに合わせてタイミング良く送出しなければならない。現在の字幕放送では、この同期の作業を人手で行っている。自動同期技術は、音声認識技術で研究された手法を利用してこの部分を自動化しようとするものであり、人手による同期作業を支援することを目的としている。

字幕作成システム技術　自動要約技術と自動同期技術を結合して、さらに、字幕データ多重化技術も利用して、システムインテグレーションをする必要がある。字幕作成システム技術の研究では、このインテグレーションを研究するとともに、様々な角度から、字幕放送制作システムを検討し、最適なシステムとすることを目的としている。

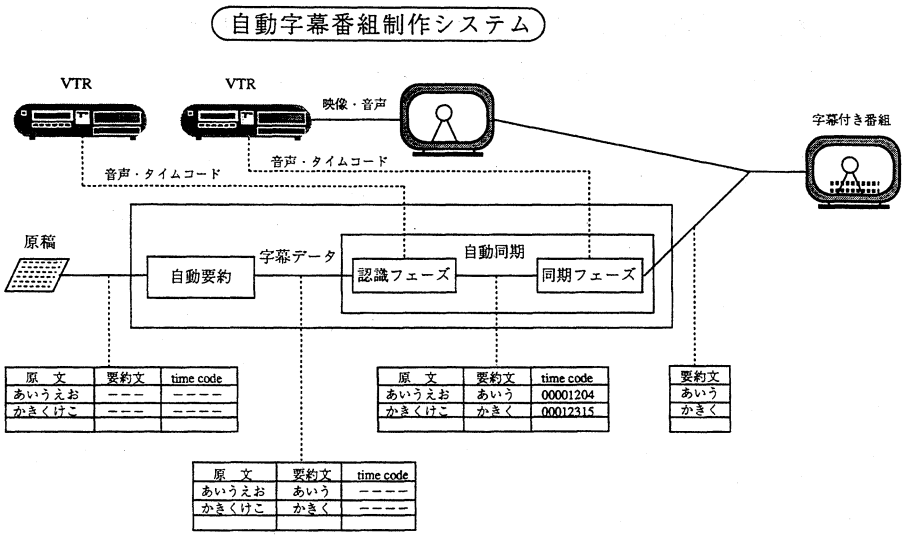

図 1: 字幕放送制作システム概念図

3 研究計画

本プロジェクトの研究計画は前期（前の3年）と後期（後の2年）の2期にわけられる。前期では、各研究項目の要素技術を研究するとともに、それらを結合して、字幕作成システムのプロトタイプを作成する。さらに、このプロトタイプを用いて、実際に字幕を作成・提示し、その機能・性能を評価する。この評価結果を踏まえて、後期では、必要に応じて、要素技術やシステムの改良を行うとともに、実用化に向けてのブラシアップを行う。

平成8年度と9年度は、要素技術の研究として、以下に示すものを研究しており、また研究する予定である。

- 自動要約技術
  - 短文分割システム
  - 重要部切り出しシステム
  - つなぎ部分調整システム
- 自動同期技術
  - 読みつけ・音声モデル合成システム
  - 最尤照合システム
  - 音声データベース
- 字幕作成システム技術
  - 字幕作成統合化シミュレーションシステム

4 自動要約技術

従来の自動要約の研究は、大きく3種類に分類できる。第1の方法は、言語理解とそれに続く言語生成による要約である。この方法は、理想的ではあるが、ニュース番組に対する実用レベルの要約は現時点では望めない。第2の方法は、文間の関係や段落構造などを利用して、要約するものである。この方法は、新聞の論説記事に適用した事例は見られるが[3]、テレビニュースの記事は後述するように新聞と異なり、文の数が少なく、段落構造も見られないため、本方法も適用が難しい。また、テレビニュースには見出しがないため、見出しを利用する方法も適用できない。第3の方法は、重要語が集中する部分を重要部分と判断し、その部分を切り出すことで要約を行うものであり、広範な文章にロバストに適用できる。そのため、少なくとも現時点では、第3の方法を中心に研究を行っている。

新聞記事とテレビニュース記事の構成の違いを示すために、1991年の日経新聞記事とNHKテレビニュース記事から各々約1000記事を抽出して、両者を比較した。その結果を図2と 図3に示す。これらの図から、テレビニュース文の特徴として、

- 1記事を構成する文数が少ない。

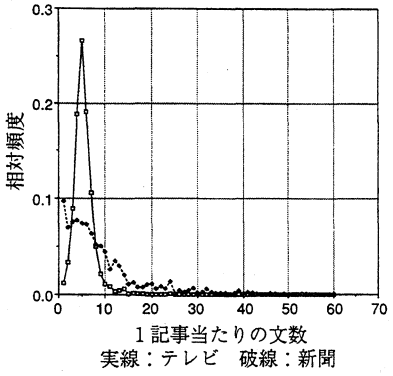

図 2: 1記事を構成する文数

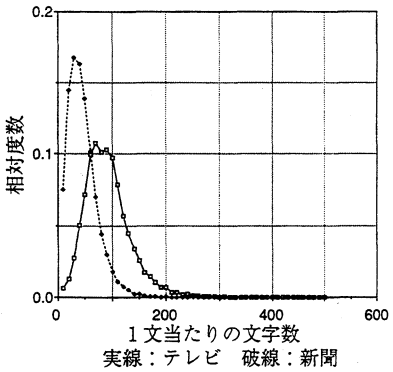

図 3: 1文の長さ

・ 1文の長さが長い。

という性質があることが分かる。このような性質があるために、重要部分を含む文を重要文として単純に抽出するときわめて粗い要約になってしまう。そこで、われわれは、重要部分を抽出する前に自動短文分割 [1] を適用する方法を検討している。自動短文分割によって、長文からなるニュース文をいくつかの短文に自動的に分割することができ、1記事を構成する文数を増加させることができる。その結果、きめ細かい要約ができる可能性がある。

次に、要約の方法ととしては、前述したように、重要語が集中する部分を重要部分として切り出す方向で研究している。そのためには、何が重要語であるかの判定をしなければならない。そのような判定を統計的に行う方法として、従来から、キーワード密度法と TF-IDF 法が提案されている。われわれも、まずこの2つの方法を利用して予備的実験を行った。1992 年から 1995 年にかけてのテレビニュース記事を各年度 3,000 件、全部で 12,000 件抽出し、そのうち、各年度 500 件を学習データとして、形態素解析し、各語に重要度を付加した。つぎに、残りの 2,500 件を試験データとして両手法を評価した。テレビニュース記事は、前述したように文数が少なく、記事の先頭の文で概要を述べるものが多い。そこで、先頭の文が最も重要であると仮定して、評価を行った。これによって 10,000 記事という大量の記事に対する評価が可能となった。

実験の方法は以下のとおりである。キーワード密度法または TF-IDF 法で得られた語の重要度を用いて、記事を構成する各文の重要度を計算し、その重要度によって文を順位づけする。そして、最も重要であると判定された文が先頭の文であるかどうか、または、重要度が第1位または第2位に判定された文の中に先頭の文が含まれるかどうかで手法の精度を評価した。なお、キーワード密度法での重要語の定義は、当該記事中に2回以上出現する内容語とした。評価結果を表 1に示す。この表から、キーワード密度法の方が TF-IDF 法よりも精度が高いことが分かる。本実験の詳細については、[2] で発表する予定である。

表 1: 重要文抽出精度

手法	第1位 (%)	1位または2位 (%)
キーワード密度法	68.86	88.95
TF-IDF 法	54.02	80.67

5 自動同期技術

自動同期技術では、ニュース音声と送出する字幕とを自動的に同期させなければならない。そのために、ニュース原稿の内容を音声合成で用いられている読みつけ技術を用いて、発音記号列に変換し、実際のニュース音声（以後、「実音声」と呼ぶ）と比較することでタイミングを検出する方

法を基に研究を進めている。また、自動同期技術研究のための基礎データとなるニュース音声データベースを作成している。平成８年度は、シミュレーションによってニュース音声データを収集した。つまり、実際の放送に用いられた音声ではなく、ニュース原稿をアナウンサーが、別途スタジオで読みあげたものをデータとして収集した。その結果、男女合計２０名、トータルで約 7.5 時間のニュース音声が収集できた。平成９年度以降は、シミュレーションに加えて、実際の放送で発声されたリアルデータも収集する方向で計画している。リアルデータとしては、ラジオとテレビの２種類のデータを収集する予定である。

これまでに収集したニュース音声データを用いて、同期点検出の予備的実験を行った。収集した音声データのうち、男女各４名の発声による合計３時間のデータを学習データとして、音声モデルを構成した。このデータを用いて、各話者毎に異なる２ループ４混合分布の音韻 HMM を学習した。その結果得られた音韻 HMM を基に、ニュース記事を読みつけた発音記号列から、ワード列ペアモデルを構成する。ワード列ペアモデルとは図 4に示すように、同期点の前後の単語列をつないでモデルを構成するものである。こうして構成された

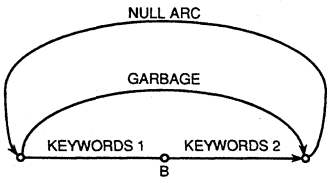

図 4: ワード列ペアモデル

ワード列ペアモデルに実音声を加えると、同期点以外では、garbage のループをたどるが、同期点前後では、ワード列ペアからなる経路の部分を通る。すなわちこの部分で尤度が増加する。そこでワード列ペアの中点（Ｂ点）の尤度を観測しており、その点の尤度があるしきい値を越えたことで同期点と判定する。このようにして同期点を検出した。試験データとして学習データに含まれない音声データから、21 のワード列ペアを選び、モデルを構成した。そして、男女各１名の発声による合計 14 分の実音声を加えて実験した。その結果を表 2に示す。しきい値を下げると検出率は良くなるが、False Alarm が頻出するようになる。本実験の詳細については機会を改めて、発表する予定である。

表 2: 同期点検出結果

検出しきい値	検出率 (%)	沸き出し率 (FA/KW/Hour)
-10	69.05	2.78
-20	76.19	9.17
-30	85.71	39.72
-40	90.48	131.93
-50	92.86	409.97
-60	97.62	975.75
-70	97.62	1774.30
-80	97.62	2867.55
-90	97.62	4118.56
-100	97.62	5403.45

6 おわりに

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の概要について述べた。今後、要素技術を確立するとともに、システムインテグレーションを行い、できるだけ早期に、実際の字幕放送制作現場に研究成果を適用したい。

参考文献

[1] 金淵培, 江原暉将. 日英機械翻訳のための日本語長文自動短文分割と主語の補完. 情報処理学会論文誌, Vol. 35, No. 6, pp. 1018–1028, June 1994.

[2] 若尾孝博ほか. 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. 情報処理学会研究会資料 NL-119-??, 情報処理学会, 1997(発表予定).

[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. 情報処理学会研究報告 NL-99-3, 情報処理学会, 1994.