自然言語処理 122－13
（1997．11．20）

テレビニュース番組の字幕に見られる要約の手法

若尾 孝博
通信・放送機構 (TAO)
渋谷上原リサーチセンター
wakao@shibuya.tao.or.jp

江原 暉将
NHK／TAO

白井 克彦
早稲田大学／TAO

概要

テキストの要約についての研究が盛んになりつつあるが、本研究では、現在数少ない字幕付きテレビニュース番組の１つであるＮＨＫの「手話ニュース８４５」を題材として、そこで使われている字幕のための要約の手法について分析した。「手話ニュース８４５」では、ニュース原稿は読まれると同時に、要約された形で字幕が映像に付与されている。この読みの部分を書き起こしたものと、それに対応する字幕を比較し、字幕で使われてる要約手法の分析を試みた。要約の手法は５種類に分類し、それらの使用頻度や、手法により削除される文字数などを測った。これにより、字幕作成の為にどの要約の手法が頻繁に使われるか、また、どれだけの文字数を削減出来るかが明らかになった。

Summarization Methods Used for Captions in TV News Programs

Takahiro Wakao
TAO of Japan
wakao@shibuya.tao.or.jp

Terumasa Ehara
NHK / TAO

Katsuhiko Shirai
Waseda University / TAO

Abstract

We analyze the methods which are used in closed captions in “Shuwa News 845” (sign language news). The program is broadcast by NHK for the deaf and hard-of-hearing people and one of the few TV news programs in Japan which come with captions. The captions in the program are summarization of what is said. We examine what the methods are used for the summarization, how frequently they are used, and how many characters are deleted by the methods.

1. はじめに

近年大量のテキストが電子化され入手が容易となり、テキストを要約する技術に関する研究が盛んになって来ている。本研究は、通信・放送機構 渋谷上原リサーチセンターで進められている「視聴覚障害者向けの放送ソフト制作技術研究開発プロジェクト」（略して「放送ソフトプロジェクト」） での自動要約技術に関する研究の一環であり（[1]、[2]、[3]）、自動的にテレビニュース原稿を要約する手法について研究を進めている。ここでは、現在数少ない字幕付きテレビニュース番組の１つである、ＮＨＫの「手話ニュース８４５」を題材として、そこで使われている字幕のための要約の手法を分析した。

「手話ニュース８４５」は、聴覚障害者向けのニュース番組であり、１５分の放送時間内に、その日の主なニュースを簡潔に網羅している。（毎週月曜～金曜、教育午後 8:45 ～ 9:00）

「手話ニュース８４５」（以後は単に「手話ニュース」とする）では、ニュース原稿は声を出して読まれると同時に手話または字幕、時には両方が付与される。字幕は文字数に制限があり、横表示の場合１行１５文字、２行まで、縦表示の場合は、１行１１文字最大２行までとなる。この為、字幕は原則として、読まれる原稿を要約したものとなっている。

この「要約された字幕」に注目し、そこで使われている要約の手法を分類した。要約の手法は、専ら表層文字列からの情報で可能な手法に限定し、文脈を理解した上での要約は、今回要約の手法として含めていない。これは、今後実行する予定である機械での自動要約（元テキストの７０%程度）を考慮してのことである。

以下では、まず原稿と字幕の特徴を示し、そして、要約の手法を分類した結果を例文とともに列挙する。その分類に基づいて、手法の使用頻度および手法を使って削減される文字数について調査した結果を示す。

2. 原稿と字幕

原稿は、ＮＨＫ「手話ニュース」の１９９７年５月１４、２３、２６日、および７月１７日放送分を見て、書き起こした。書き起こしたものは、アナウンサーの読む音声部分とそれに付与された字幕の部分である。

音声部分と字幕部分の関係は、メインのニュースとその他のサブニュースでは、多少違いがあるが、文字数において、音声部分を１００とすると、メインニュースでは、その字幕は６０～７０であり、サブニュースでは、４０～５０の割合となっている。ただし、これらの割合は、ニュースの内容によって左右され、固有名詞の多く現れるニュースでは、字幕部分でもあまり要約・省略が行われない傾向が見られた。

	割合（文字数で）
字幕（メイン）	60 - 70 %
字幕（サブニュース）	40 - 50 %

表 1 手話ニュースの音声と字幕部分の比較

「手話ニュース」における字幕の特徴であるが、まず、前述の通り文字数に制限がある。基本として複雑な構造をした文は無く、単文である。また、修飾語は出来るだけ省略、簡略化されており、修飾が重なる時は、分解して単文としている。それに固有名詞については、キーワードでなければ、省略出来るものは省略するか、または、簡単な言い方にしている。

また、「手話ニュース」と ＮＨＫ午後７時のニュースとを比べてみると、読みの速度においては、７時のニュースでは、１分間に３５０～３６０文字であるが、「手話ニュース」では、２６０～２８０文字であった。

	読みの速度 (文字／分)	文字数での 比較
7時のニュース	350 - 360	100 %
手話ニュース	260 - 280	60 - 80 %

表 2　7時のニュースと手話ニュースの比較

1ニュース当りの文字数では、7時のニュースを100とすると、「手話ニュース」は、60～80となっていた。しかし、ニュースによっては、「手話ニュース」での報道のほうが詳細であり、 文字数が7時のニュースよりも多いニュースもみられた。これは、「手話ニュース」の編集が、独自の判断で行われているからであるとのことである。

3. 要約の手法

次に「手話ニュース」の字幕に使われた要約の種類とその例を示す。要約の中は、原文の意味を解釈して行われるものもあるが、ここでは主に表層の情報を手がかりに行われていると思われる要約について5種類にまとめた。

1.　文末を削除、言い換え

1.1　サ変動詞はサ変名詞に変え、その後来る語句は削除(「へ」が付けられることもある)

「‥ 疑いで逮捕したものです。」
→「‥ 疑いで逮捕」

「7月中に解散します。」→「7月中に解散へ」

1.2　動詞を終止形にして、後は削除

「行政処分を行うことになると述べました。」
→「行政処分を行う」

1.3　丁寧助詞「ます、まし」は削除する

「‥ との取り引きを見合わせました。」
→「‥ 見合わせ」

「‥ 余震が相次ぎました。」
→「‥ 余震が相次だ」

1.3　完了助詞＋丁寧助詞「ます、まし」は削除(過去の助詞「た」追加することもある)

「‥ 指導力にかかっています。」
→「‥ 指導力にかかる」

1.5　名詞＋断定助詞「です」、「でした」は削除し名詞だけにする

「‥ 左志小学校です。」→「‥ 左志小学校」

1.6　文末の名詞＋格助詞「に」は「に」で止める

「‥ 暫定使用になります。」
→「‥ 暫定使用に」

1.7　簡潔な否定の表現へ

「‥ を認めていませんでした。」
→「‥ を認めず」

1.8　報告文の文末に来る報告動詞の部分は削除

「‥ 売ろうとしていたと言うことです。」
→「‥ 売ろうとしていた」

「‥ べきだと述べました。」→「‥ べきだ」

2.　文の部分を残す

2.1　名詞語句を抽出

「逮捕されたのは、株式売買の責任者だった松木新平元常務」→
「逮捕　松木新平元常務(株式売買担当)」

2.2　文の一部(名詞語句以外を含む)をそのまま残し、あとは削除

「特捜部では、不正な利益提供は株主総会を乗り切るため行われたものとみています。」→
「不正な利益提供は株主総会を乗り切るため」

2.3 「･･･」を文末に付けて文末の表現を削除

「野村証券にはペナルティを払ってもらわなければならないと述べました。」→

「野村証券にはペナルティを･･･」

3. 意味を変えずに別の語句・表現で言い換え

3.1 意味をとり簡潔な表現に言い換え

「･･ 震度6弱の地震に襲われています。」

→「･･ 震度6弱を観測」

「珍しいことです。」→「異例」

3.2 簡潔な同意語に言い換え

「橋本総理大臣」→「橋本首相」

「朝鮮民主主義人民共和国 」→「北朝鮮」

3.3 短縮形を使う

「野村証券」→ 「野村」（組織名）

「駐留軍用地特別措置法」

→「特措法」(法律名)

4 接続詞や文頭のつなぎの語句は削除

「しかし」、「一方」、「その一方で」、

「それによりますと」、「さらに」、「このため」、

「それだけに」、「また」、「このうち」、「その上で」

などの語句は削除

5 比較日時を示す語句は削除

「1997年」などの絶対日時ではなく、「今日」「昨日」

などの比較日時を示す語句は省略。

「今日全体会議を開き、中間報告を示しました。」

→「全体会議を開き、中間報告を示した」

4. 要約の頻度と程度

次に、書き起こされた音声部分と字幕部分を比べて、どの要約手法がどれだけの頻度で使われたかを、また要約の程度として、各要約手法によりどれだけの文字数が削減されたかを調査した。

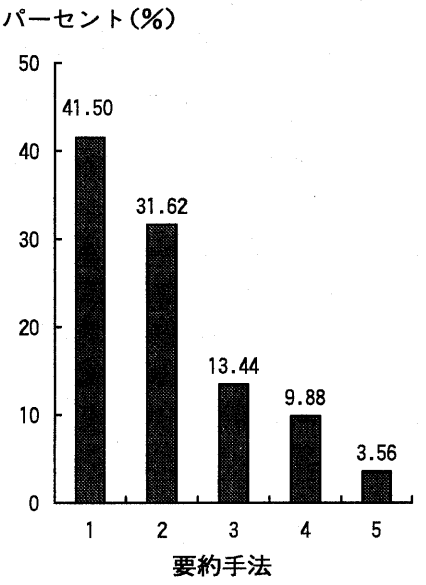

図 1 要約手法別の頻度での割合

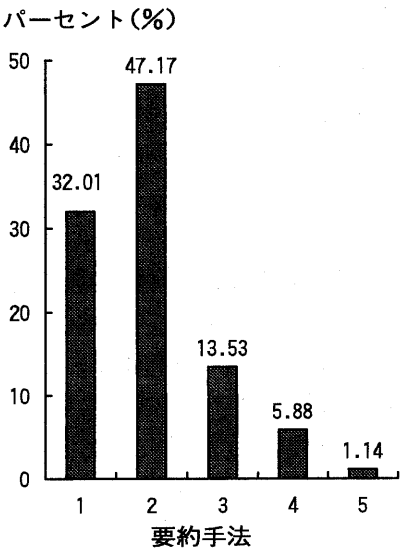

図 2 要約手法別の削減文字数での割合

要約手法	5月 14日	5月 23日	5月 26日	7月 17日	頻度合計	削減文字数 (合計)
1－1	11	8	9	6	34	203
1－2	3	3	0	3	9	69
1－3	6	12	9	1	28	59
1－4	1	4	4	1	10	26
1－5	3	3	6	2	14	32
1－6	2	1	0	0	3	18
1－7	1	0	1	0	2	12
1－8	0	1	1	3	5	33
2－1	5	2	3	1	11	94
2－2	15	15	16	11	57	490
2－3	5	1	4	2	12	82
3－1	5	8	6	1	20	139
3－2	2	2	2	0	6	22
3－3	3	5	0	0	8	30
4	7	10	6	2	25	83
5	2	1	4	2	9	20
合計	71	76	71	35	253	1412

表 3　要約手法別の使用頻度（4日分）と削減文字数

要約手法	内容	頻度	割合 %	削減文字数	割合 %
1	文末を削除、言い換え	105	41.50	452	32.01
2	文の部分を残す	80	31.62	666	47.17
3	意味を変えずに別の語句で言い換え	34	13.44	191	13.53
4	接続詞や文頭のつなぎの語句は削除	25	9.88	83	5.88
5	比較日時を示す語句は削除	9	3.56	20	1.41
合計		253	100.00	1412	100.00

表 4　要約手法（大別）の使用頻度での割合と削減文字数での割合

対象となった文の数は、すなわち、音声原稿文とそれに対応する字幕文のペアの数は、表５の通りである。

番組の日付	ペア数
５月１４日	５２
５月２３日	５１
５月２６日	６２
７月１７日	２７
合計	１９２

表 5　調査対象となったペア数

７月１７日は、手話ニュースの全部ではなく、メインニュース（４件）のみが調査対象となったため、ペア数が少なくなっている。

4.1.　要約手法の頻度

要約の手法別の使用頻度は、表３では詳細に、表４と図１では、大別した形で示されている。頻度から見ると、要約の手法としては、手法１の文末の処理が一番多く、次に手法２の文の部分を残すタイプの手法が多く使われていることが分かった。

4.2.　要約の程度

次に要約の程度を知るため、削減文字数、つまり、その手法を適用することによりどれだけの文字数が削減されるかを調べた。

頻度の場合と同じく、要約手法別の削減文字数が表３では詳細に、表４と 図２では、大別した形で示されている。

削減される文字数は、頻度の場合とは異なり、手法２の文の部分を残す手法が一番効果的であることが分かる。手法１（文末の処理）は使用頻度は多いが削減される文字数はあまり多くないと言える。

しかしながら、要約手法を個別にみてみると、手法１（文末の処理）の中にも削減文字数においてばらつきがあることが分かる。各要約手法を１回適用するとどれだけ文字数が削減されるかを手法別にみると表６のようになる。

手法の１－１（サ変動詞で終わる文の処理）は頻度も多く、削減される文字数も多い。この手法を１回適用すると約６文字が削減されている。これに対して手法の１－３は、手法１の中では、２番目に頻度が多いが、削減される文字数は、１回あたり約２文字と少ない。

要約手法２の場合は、３つの手法どれをとっても削減される文字数はかなり多いことが分かる。

要約手法	１回の削減文字数
１－１	5.97
１－２	7.67
１－３	2.11
１－４	2.60
１－５	2.29
１－６	6.00
１－７	6.00
１－８	6.60
２－１	8.55
２－２	8.60
２－３	6.83
３－１	6.95
３－２	3.67
３－３	3.75
４	3.32
５	2.22

表 6　要約手法１回で削減される文字数

5. まとめ

ＮＨＫの字幕付きテレビニュース番組である「手話ニュース８４５」を題材として字幕のための要約の手法を調査、分析した。字幕は音声部分の要約であり、そこで使われている要約の手法を５種類に分類した。この分類に基づき、各手法の使用頻度および削減文字数を調査した。

頻度からみると、文末を削除または言い換える手法が最も頻繁に使われていた。２番目に多く用いられていたのは、文の一部分を残す手法であった。

削減される文字数からみると、文の一部を残す手法が効果的であることが分かった。文末を処理する（手法１）方法は、頻度は高いが、削減する文字数は、それ程多くないことも判明した。

本研究は、通信･放送機構 渋谷上原リサーチセンターで進められている「放送ソフトプロジェクト」での自動要約技術に関する研究の一環であり、今後は、今回の研究をもとにして自動的にテレビニュース原稿を要約する、つまり、原稿文の文字数を削減するシステムを開発して行く予定である。

「放送ソフトプロジェクト」ではテレビニュース番組（例えば、ＮＨＫ午後７時のニュース）に字幕を付与するための研究を進めている。そこでの字幕は、原稿の要約したものを考えており、要約の目標は７０%程度で、文末を削除したり、言い換えたり、また文の一部を残す事で達成が可能ではないかと考えている。今回の研究で見つかった５種類の要約手法をどのように使って文字数を削減していくかが今後の課題である。

6. 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 １９９７年

［2］ 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、自然言語処理研究会 97-NL-119-6

［3］ Takahiro Wakao, Terumasa Ehara, Eiji Sawamura, Yoshiharu Abe, Katsuhiko Shirai, “Application of NLP technology to production of closed-caption TV programs in Japanese for the hearing impaired” in the proceedings of ACL 1997 Workshop, Natural Language Processing for Communication Aids, pp 55 - 58, 1997 Madrid, Spain.