要約ニュースコーパス構築に向けた基礎検討

田中 英輝[1] 後藤 功雄[2] 伊藤 崇之[1]
[1]NHK放送技術研究所 [2]ATR音声言語コミュニケーション研究所
{tanaka.h-ja, itou.t-gq}@nhk.or.jp, isao.goto@atr.jp

1 はじめに

NHKでは、短いニュースをいろいろなメディアで提供している。テレビやラジオのニュース中で多くの項目を短くまとめて紹介するコーナー[1]や、衛星、地上デジタル放送の画面中でニュースの大意を 105 文字で提供するサービスなどである。
このようなサービスは、通常のニュース原稿を要約した原稿を元にして行われている。今後もメディア、視聴機器の多様化は進むと考えられ、上述のような短いニュースの需要はさらに増えるものと予想される。
そこで著者らは、通常のニュース原稿から、短いニュース原稿を作成する過程を支援する自動要約の研究を開始した。今回、研究の第一歩として、要約者への聞き取り調査、要約原稿の収集と分析、元原稿との照合実験による要約過程の分析を行ったので結果を報告する。

2 要約原稿の作成手順

2.1 元原稿と要約原稿

NHK では記者が取材結果を元に原稿を作成し、これを使って通常のラジオやテレビなどの放送を行っている。この原稿は、テレビ、ラジオなどのサービスに応じて適宜変更して使う汎用的な性格を持つ。本稿では、この原稿を「元原稿」と呼ぶ。
1 章で述べた短いニュースのサービスは、この原稿を人手で短くした原稿を元に行っている。こちらも、音声、文字といったサービス形態に応じて変更して使う汎用的な性格を持っており、本稿ではこの原稿を「要約原稿」と呼ぶ。

2.2 聞き取り調査

要約原稿を作成する専門家への聞き取り調査を行った。概要は以下の通りである。

作成者

要約原稿の作成を担当する記者は元原稿の作成記者と別である。

文字数と時間の制限

現在はデジタル放送の文字サービスに直接利用すること考慮して、105 文字以内で要約を作成している。この字数は現在のデジタル放送画面のデザインとして決められている。また、要約原稿は放送の直前に作成することが多く、短時間の作業となることが殆どである。

要約方法

元原稿以外の情報は原則として利用しない。また、元原稿を最初から最後まで丁寧には読まない。元原稿の第 1 文がニュースの要約を表すことが多いため、これを中心に据えて、その他の文との関係を分析し、その後、第 1 文を構成する単語の削除、言い換え、追加という手段で要約する。また、元原稿の最終文の情報を採用することが多い。

要約の専門家は、要約作成のために独特の能動的な読み方をすることが報告されており（Mani 01）、この聞き取り調査でも、同様の傾向が現れていたと考える。また、(Jing 99)が指摘する、元原稿の Cut-and-Paste によって要約原稿を作成している傾向も伺える。

3 要約原稿データ

聞き取り調査の結果を具体的に検証するため、2003 年 11 月から 2004 年 6 月までの要約原稿 18,777 記事を入手して分析した。
4 章で述べるように、著者らは要約原稿と元原稿を照合して（要約原稿、元原稿）のペアデータも作成している。照合された対応元原稿の統計情報と合わせて、要約原稿の統計情報を表 1 に示す。

3.1 平均長と要約率

表 1 より、文字数で見た平均要約率は 22.5%である。また、要約原稿の平均長は 109.9 文字であった。なお、要約原稿の文数最頻値は 2 であり、文数 1 から 4 までの要約原稿の累積相対度数は 0.99 であった。そこで、文数 1 から 4 の要約原稿の平均長を計算した結果、105.4 文字となった。聞き取り調査の通り、大半の要約原稿は 105 文字に合わせて作成されていることが確認できた。また、元原稿の第 1 文の平均長は 94.9 文字となった。要約原稿は 105 文字で作成すること、元原稿の第 1 文を使うことから、単純には第 1 文に 10 文字追加すれば要約原稿長が得られる。次章ではこの実態を詳細に調査する。

表 1 要約原稿と元原稿の特徴

	元原稿	要約原稿
原稿数	18,777	
平均文数	5.13	1.63
平均原稿長（文字）	487.7	109.9
第 1 文平均長	94.9	81.3

[1] 例えばBS50 ニュース

4 要約原稿と元原稿の照合

4.1 JM 法

要約原稿と元原稿の対応をさらに詳細に調査するため、既に収集済みのニュース原稿（元原稿）データベースとの照合を行った。
要約原稿は原則的に元原稿から作成されているが、今回入手した要約原稿には両者の対応関係を示す情報がない。このため、元原稿を推定する必要がある。また調査のためには元原稿の照合にとどまらず、文、単語といった詳細な照合情報が欲しい。著者らは、これらの要求を満たす手法として(Jing 99)の研究に着目し、これを応用することにした。この論文で Jing らは Ziff-Davis コーパスを対象に、要約と原文の単語の対応関係を調査するため、確率モデルを使った単語対応推定手法を提案している。概要は次の通りである。

A) 要約中の単語位置を（I）とする
B) 原文中の単語位置を文番号（S）と文内位置（W）の対（S,W）で表現する
C) 要約中の各単語が原文で出現する位置すべてを B) の形式で表現して、要約中の各単語の原文での出現状況を表すトレリスを作る
D) 要約中の単語を先頭から右に連続的に動的計画法で走査して、式（1）で示す単語照合確率が最大になるトレリス上の経路を照合結果とする

$$P=\prod_{i=1}^{n-1} P(I_{i+1}=(S_2,W_2)\,|\,I_i=(S_1,W_1)) \quad (1)$$

式(1)は、隣り合う要約中の単語 I_i, I_{i+1} が原文の (S_1,W_1)と(S_2,W_2) という位置に出現[2]する確率の連乗積で、要約と原文の単語照合確率を示す。
(Jing 99)らは確率の値を経験値として、文番号と文内位置に応じた 6 段階で与えている。最も高い確率は、要約中の隣接 2 単語が、原文の「同一文内の隣接 2 単語と対応する場合」の 1 である。次は「原文の同一文内で、要約と同順に出現する 2 単語と対応する場合」の 0.9 である。最小の確率は「一定文以上離れた原文の文にある 2 単語と対応する場合[3]」の 0.5 である。
本手法は原文と要約の間で、語順が交差する照合を許すことに注意されたい。本稿では便宜的に上記の手法をJM法[4]と呼ぶ。

4.2 原稿照合手順

要約原稿と元原稿のペアに対して(1)式で決まる最大単語照合確率は、原稿間の類似性尺度の一つと考えてよい。このため与えられた要約原稿に対して(1)を使って元原稿データベースを検索し、最大の単語照合確率を示す元原稿を選択すると、元原稿の選択と単語の照合が同時に可能となる。著者らはこの性質を利用して、以下の原稿照合を行った。手順は以下の通りである。

A) 元原稿の数値表現正規化
元原稿では漢数字が使われるが、要約原稿では算用数字が使われる。表記統一のため元原稿の数値を算用数字に自動変換した。
B) 形態素解析
形態素解析器を使って原稿を形態素に分割した。単語照合にはすべての形態素を利用した。
C) 探索範囲
要約原稿の作成日から過去 3 日間の元原稿を探索した。3 日という範囲は経験的な知見による。

また、形態素の照合結果を観察するためのブラウザを作成した。

なお、オリジナルの JM 法のままで原稿の照合を行うと、問題が生ずることに注意が必要である。JM 法では、要約中の単語が原文に出現していないとき、要約にその単語が現れないものとみなす。すなわち非出現単語のトレリスを確率 1 で飛び越すことになる。このため、要約中の単語が出現していない原文の方が、大きな単語照合確率を得る事になり、単語照合数が少ない不適切な原文を選ぶ問題が発生する[5]。
著者らはこの問題の簡易な解決手段として、飛び越しが起こる場合には、確率 0.55 を使う事にした。これは最低の確率 0.5 を与える単語の照合（一定文以上離れた文にある 2 単語との照合）よりは、単語の飛び越しが有利になるように確率を調節したもので、これでほぼ適切な元原稿を選択できるようになった。

5 照合結果から見た文対応傾向

対応結果を観察したところ、おおむね適切な原稿との対応、形態素対応が得られた。図 2 に原稿照合例を示す。JM 法には後に議論するような問題はあるが、ここでは、JM 法の出力は近似的に正しいと考えて対応の分析を行った。

表 2 形態素対応率と原稿の相対度数

形態素対応率	原稿の相対度数
100%	0.265
90%以上	0.970

[2] あるいは「対応する」と解釈するとよい。
[3] 著者らは平均文数を考慮して、この値を 2 とした。
[4] Jing and McKewown
[5] (Jing 99)の研究のように照合する原文が一つだけの場合には問題とならない。

5.1 要約原稿と元原稿の形態素対応率

要約原稿の全形態素中のうち、元原稿中に対応先が求められた割合（形態素対応率）を計算したところ、平均 96.4%であった。また、表 2 に示すように、形態素対応率 100%となった要約原稿の相対度数は 0.265、同じく形態素対応率 90%以上の相対度数は 0.970 に達した。これより、要約原稿の大半の形態素が元原稿に出現していることがわかる。聞き取り調査のとおり、原文を生かした抜粋に近い要約であることが確認された。

5.2 要約原稿の形態素出身率と採用率

次に、要約原稿の各形態素と、元原稿の文との対応を下記の手順に従って分析した。

準備（元原稿の文数による分類）

要約原稿と元原稿のペアを、元原稿の文数でグループに分類した。元原稿の文数が違うデータを混在して分析すると、異なる最終文番号が混在して不都合が生ずる。これを防ぐために文数によって分類した。この分類ごとに以下の 2 手法で対応を計算した。

形態素出身率

要約原稿の全形態素を対応先の文番号で分類して、相対度数を計算した。要約原稿から元原稿を見て、要約原稿の形態素の出身地（元原稿の文番号）の相対度数を調査したことになり、この値を形態素出身率と呼ぶ。対応先のない形態素は NA と分類した。

形態素採用率

元原稿の各文の形態素で、要約原稿に採用された相対度数を計算した。こちらは、元原稿から要約を作成する場合に、元原稿の各文から何パーセントの形態素を採用したかを調査したことになり、形態素採用率と呼ぶ。

元原稿の文数の分布を見ると、最頻値（モード）は 5 文で、4 文から 8 文の区間の元原稿相対度数は 0.88 と高い集中が見られた。そこで典型例として、5 文、および 8 文からなる元原稿グループと、その要約原稿から得られた平均の出身率、採用率を図 1 に示す。グラフの凡例の末尾にある括弧付きの数値は、元原稿の文数である。このグラフから次のことが読み取れる。

形態素出身率の傾向

元原稿 5 文の結果を見た場合、形態素出身率は第 1 文が 65.1%と支配的で、第 2 文以下は 10%弱となっている。元原稿 8 文の結果を見ると、第 1 文が 45.5%、第 2 文が 14.4%となりその他は 10%以下であった。このように元原稿の文数が増えると元原稿第 1 文からの形態素出身率は減る傾向が見られる。すなわち、元原稿の文数が増えるにつれて、第 2 文以下の情報を利用する傾向があると予想される。

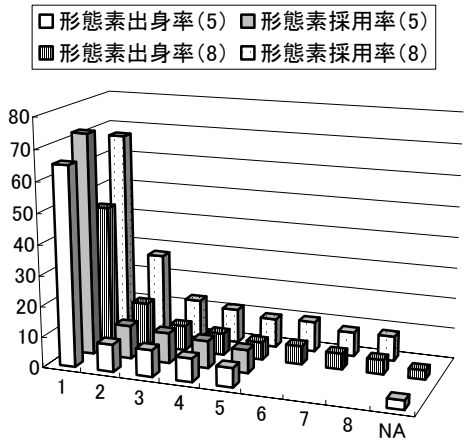

図 1 形態素平均出身率と平均採用率
（横軸は元原稿の文番号、縦軸は%）

形態素採用率の傾向

最大の形態素採用率を持つのは元原稿の第 1 文である。ただし、形態素出身率と違って、5 文から 8 文と元原稿の文数が増えても、元原稿第 1 文の形態素採用率は 70%程度で変わらなかった。他の文数でも調査したが、第 1 文の形態素採用率は 70%程度で一定する傾向がみられた。これに対して、文数が増加すると第 2 文以後の形態素採用率は上昇する。105 文字の要約を作るのに、第 2 文以後の形態素の採用が増えても、第 1 文の形態素採用率が一定なのは一見不思議である。今回の元原稿を調べたところ、元原稿の文数が多くなるにつれ、第 1 文の長さが減少していた。例えば、5 文元原稿の第 1 文は平均 99 文字なのに対して、8 文元原稿の第一文は 83 文字と減っていた。この減少傾向が一つの原因と思われる。

前節で元原稿第 1 文の平均長の観察から、元原稿第 1 文に 10 文字の追加で要約原稿長になることを述べたが、本節の観察から、実態はそれほど単純ではない事がわかった。

元原稿第 1 文が最重要な抜粋（extract）であり、ここから 70%程度の形態素を採用すること、元原稿の文数が増えても、第 1 文からの 70%程度の形態素を採用するが、文長が短くなるため、絶対的な採用数は減少すること、などがわかった。全体的な傾向として、元原稿の文数が大きくなるほど、複雑な編集になっているようである。

なお、聞き取り調査で最終文からの形態素採用が多いとの報告があった。図 2 はその例で、確かにこのような事例は観察されたが、本節の統計的調査では明確にならなかった。JM 法の照合には問題もあることから、引き続き調査したいと考えている。

新千歳空港 ほぼ平常ダイヤへ
北海道の新千歳空港は１０日、雪の影響で１１５便が欠航しダイヤが大幅に乱れましたが、１１日は日本航空の午前８時１５分発名古屋行きの便が機材繰りのため欠航する以外は、始発便から平常通りのダイヤで運航する見込みです。

新千歳空港は平常ダイヤへ
北海道の新千歳空港はきのう雪の影響で１１５便が欠航しダイヤが大幅に乱れましたがけさは始発便から平常通りのダイヤで運航する見込みです。
新千歳空港はきのうの昼頃から局地的に強く降った雪の為滑走路の除雪作業が追いつかず３時間余りに亘って飛行機の離着陸ができなくなり１日に発着する国内線の半数近い１１５便が欠航しダイヤは最終便まで大幅に乱れました。
航空各社によりますときょうは日本航空の午前８時１５分発名古屋行きの便が機材繰りのため欠航する以外は平常通り運航する予定できょうは新千歳空港の空のダイヤに乱れは出ない見込みです。

図 2 要約原稿と元原稿の照合例

元原稿第 1 文の編集の過程をブラウザで観察したところ、長い連体修飾や冗長語句の削除など、よく指摘される編集操作が行われていた。この他、NHK の要約作業に特徴的と思われる現象を二つ指摘する。

A) 日付表現の変更
図 2 からもわかるように、元原稿は相対的に「きのう」「けさ」と書かれているが、要約は 10 日、11 日と絶対的表現で書かれている。要約原稿と元原稿ではメディアによるサービス時刻の違い、また継続期間の違いがあるためこのような変更を行っている。

B) 発言の採用
元原稿第 1 文で抽象的にまとめてある表現を、個人の発言の引用で置換して具体化する例がよく見られる。ある要約原稿では、元原稿第 1 文の「投機的な動きが見られるとした上で」とまとめた表現をそのまま使うのでなく、第 2 文にある発言の「『投機的な動きがやや目に付く』と述べて」を採用して具体化していた。

以上のような操作は、抜粋からアブストラクトを作成するための技術と考えることができ、今後、分類整理する予定である。

6 要約コーパス構築に向けて

今回の要約原稿と元原稿の照合結果をもとに、大規模な「要約コーパス」を作成したいと考えている。このコーパスは(Jing 99)や(Marcu 99)も有用性を指摘している（要約原稿、元原稿）のペアと文、形態素の対応情報からなるものである。これを使ってさまざまなコーパスベースの要約の研究(Mani 99)を進めたいと考えている。

たとえば、前章でアブストラクトを作成するための特徴を一部紹介したが、このような特徴も、正しい形態素対応が得られれば、自動抽出可能となる。(加藤 99)では、このような考え方で、ニュース原稿から文字放送ニュースへ変換するための、局所要約知識抽出の研究を行っている。

このような研究を進めるには、まず、JM 法の形態素照合能力の改善が必要だと考えている。JM 法は今回おおむね良好な形態素照合をしていたが、誤りと判断される対応もあった。元原稿の第 1 文がほぼ原稿の要約になっていることから想像できるように、元原稿内には表現の重複が多い。例えば図 2 の点線下線部分は 1 行目に出現している。このような場合に JM 法は、誤った照合をすることがある。

また、要約中に同じ単語が複数出現した場合に、元原稿中の一単語に誤って照合する事も多い。JM 法では、多対一の対応に関する制限がないため、このような問題が発生する。

著者らはすでにこのような問題点の改善方法も検討しており、これについては稿を改めて報告したい。

7 おわりに

NHK の要約原稿を収集して観察し、その特徴を示した。また要約原稿と元原稿との照合を行って、要約の作成過程を分析した。著者らは今後さらに要約原稿を収集して, 要約原稿と元原稿のペアからなる「大規模要約コーパス」を構築していく予定である。また、このコーパスをもとに、要約原稿の作成を支援するシステムの研究開発を行いたいと考えている。

参考文献

(Jing 99) Jing, H. and K. R. McKeown: “The Decomposition of Human-Written Summary Sentences”, pp. 129-136, proc.of SIGIR99, (1999)

(Mani 99) Mani, I and M. Maybury: “Advances in Automatic Text Summarization”, Section 2, The MIT Press, (1999)

(Mani 01) Mani, I: “Automatic Summarization”, John Benjamins, (2001)

(Marcu 99) Marcu, D.: “The Automatic Consturuction of Large-Scale Corpora for Summarization Research”, pp. 137-144, proc. of SIGIR99, (1999)

(加藤 99) 加藤、浦谷: “局所的要約知識の自動獲得法”, 自然言語処理, Vol. 6, No.7, pp.73-92, (1999)