社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2005－NL－170 (19)
2005／11／22

ニュース要約の実態調査と要約モデルの検討

田中英輝 熊野正 西脇正通 伊藤崇之
NHK 放送技術研究所
tanaka.h-ja@nhk.or.jp

あらまし

本稿では、放送局の専門家による、日本語ニュース要約の実態についての聞き取り調査結果を報告する。また、約 27,000 件の要約原稿と元原稿の自動照合実験を行い、聞き取り調査の結果を定量的に考察する。これにより、専門家の要約は典型的な抜粋型であり、ニュース特有の構造を最大限に利用している事を示す。次にこれらの知見に基づいた要約の作業モデルを提案する。またこのモデルで重要となる、ニュースの書き出しの部分と、本体の間の照応関係が多岐にわたることをいくつかの事例を用いて説明する。

The Investigation and Modeling of Manual Summarization of Japanese Broadcast News

Hideki Tanaka, Tadashi Kumano, Masamichi Nishiwaki and Takayuki Itoh
Science and Technical Research Laboratories of NHK

Abstract

We describe our analysis and modeling of the summarization process of Japanese broadcast news. We have studied the entire manual summarization process of the Japan Broadcasting Corporation (NHK). The staff of NHK has been making manual summarizations of news text on a daily basis since December 2000. We interviewed these professional abstractors and obtained a considerable amount of news summaries. We matched the summary with the original text, investigated the news text structure, and thereby analyzed the manual summarization process. We then developed a summarization model on which we intend to build a summarization system.

1 はじめに

NHK では、短いニュースを衛星、地上デジタル放送の文字画面、ホームページでのニュースサービスなどさまざまなメディアで提供している。このようなサービスは、通常のラジオ、テレビニュースの原稿を人手で短縮した原稿を元にして行われている。今後も視聴機器の多様化が進むと思われ、上述のような短いニュースの需要は増加すると思われる。このような背景の元、著者らは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行っている。これまでに(田中 05)(Tanaka et al. 05)で要約ニュース作成手順の調査、短いニュースと通常の原稿の照合実験などを報告した。本稿ではデータを増加して同様の実験を行い、要約作業を定量的に分析する。またこれらに基づいた要約者の作業モデルを提案する。これは自動要約のための計算モデルでもある。また、このモデルを使った自動要約で特に重要となる元ニュース内の照応現象を、例を用いて説明し、最後に今後の課題を述べる。

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、これを使ってラジオやテレビなどのニュース番組を放送している。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 節で述べた短いニュースのサービスには、この原稿を人手で短縮したものを利用している。本稿ではこの短縮した原稿を「要約」または「要約原稿」と呼ぶ。

2.2 聞き取り調査

要約原稿を日々作成している要約者への聞き取り調査を行った。結果は以下の通りである。

要約者

要約原稿の作成を担当するのは退職した記者であり、元原稿の作成記者と別である。いずれも NHK のニュースを熟知している。

文字数

本稿で分析対象とした要約原稿は、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 文字以内に元原稿を短縮したものである。この字数は主目的のサービス画面のデザインを元に決められており、変化することがある[1]。文字数の許す範囲でできるだけ情報を盛り込む努力をする。

時間の制限

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。元原稿の書き出しを中心に据えて、その他の文との関係を分析し、その後、書き出し部分の単語の削除、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、要約者は同様の手法を用いていることがわかった。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た。

3 要約と元原稿の照合

3.1 JM 法

前節の聞き取り調査で得た要約手法の実態を定量的に調査するため、要約原稿を入手して元原稿との対応付けを行った。ここで利用したのは 2003 年 11 月から 2004 年 9 月までの要約原稿で、4 文以下の 26,777 を対象とした。

要約原稿は原則的に元原稿から作成されているが、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要約と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

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

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[3]以上離れた原文の文にある 2 単語と対応する場合」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。

3.2 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。具体的な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。

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。

---

[1] 2004 年 11 月からはデジタル放送の画面の変更によって 150 文字程度への要約を行っている。

[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

Summary
住宅金融公庫は、個人が住宅を購入するために融資を受ける際の基準金利を、これまでの２．６％から０．０５％引き下げ、２．５５％とすることを決めました。
新しい金利は、１９日の受け付け分から適用されます。

Original
住宅金融公庫は、個人が住宅を購入するために融資を受ける際の基準金利をきょうの受け付け分からこまでの二点六％から、二点五五％に引き下げることを決めました。
これは、先月に比べて長期金利が若干低下し、住宅金融公庫の財源となる財政融資資金の貸し付け金利が引き下げられた事に伴うものです。
これにより個人が住宅購入資金の融資を受ける際の基準金利は、これまでの二点六％から〇点〇五％引き下げられ、二点五五％となり、二千万円を三十五年返済の固定金利で借りた場合、返済総額は、これまでよりも十一万円減る計算になります。
新しい金利は、きょうの受け付け分から適用されます。

図 1　要約と元原稿の照合

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。

4　照合結果

4.1 記事照合

照合された原稿ペアの基本的な統計量を表 1 に示す。また、原稿ペアの例を図 1 に示す。図中の下線は対応を示す。要約原稿の長さは平均 105.3 文字となった。105 文字の制限を精一杯利用するという要約者の発言から、これに近い数字を予想していたとおり、ほぼ一致した。すなわち文字数からみた要約率は 21.6%であることがわかる。なお、平均が 105 文字を超えたのは、入手した要約原稿の一部に要約でない長い原稿が入っていたためである。

表 1　照合原稿ペアの特徴

	元原稿	要約原稿
記事数	26,277	
平均文数	5.81	1.59
平均原稿長(文字)	487.4	105.3
第 1 文平均長(文字)	98.3	80.6

[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。

記事の照合正解率を評価するために 2004 年 1 月の 272 記事ペアを調査したところ、268 ペアが正しい照合で(98.5%)、4 ペアは誤りであった。誤った照合は、いずれも要約と元原稿で同じニュースだと判定できたが、細かな内容に差があった。例えば、けが人の数の相違、「身元の確認を急いでいます」という表現に対して「身元が判明しました」などの差である。これらは、同一ニュースで、時刻によって変化した部分の不一致である。これらの差は小さいとはいえ、内容は違うと判断して誤りとした。いずれにせよ、記事の照合は十分な精度だと考えている。
単語(形態素)の照合は本当の正解がないため、精度を測ることはできないが、観察の結果ほぼ正しいと思われた。そこで、以下単語の照合状態が正しいとしての観察結果を報告する。

4.2 形態素対応率

要約原稿の全形態素のうち、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。全記事の平均形態素対応率は 96.4%であった。また表 2 に示すように、形態素対応率が 100%となった原稿の相対度数は 0.265、90%以上となった相対度数は 0.957 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型の要約であることが確認できた。

4.3 形態素採用率

元原稿のどの部分の形態素が要約に利用されやすいかを調査するため、元原稿の各文について要約に採用された形態素の割合、形態素採用率、を

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	0.957

計算した。
まず図 2 に元原稿を文数で分類した場合の相対度数を示す。これから、大多数の元原稿(88.0%)は 4 文ないし 8 文からなることがわかる。今回はこの部分と 9 文ないし 13 文の元原稿を対象に形態素採用率を計算した。
図 3、4 に結果を示す。図 3 は 4 文ないし 8 文からなる元原稿、図 4 は 9 文ないし 13 文からなる元原稿の結果である。グラフの横軸は文番号、縦軸は平均形態素採用率(%)を示す。例えば図 3 の「4 文ニュース」で示したグラフでは、4 文からなるニュースの第 1 文からは平均 72.4%の形態素が要約原稿に採用され、2 文から 4 文までは 10%程度が採用されたことがわかる。

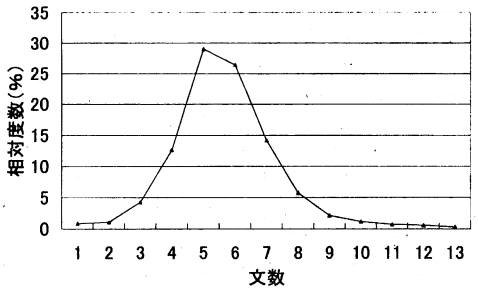

図 2　文数と原稿の相対度数

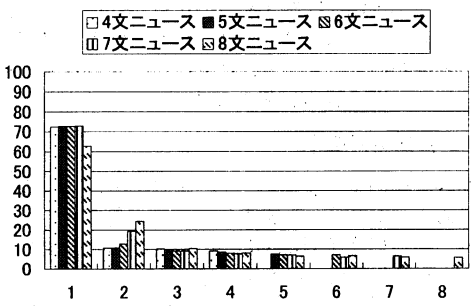

図 3　元原稿各文の形態素出身率(4－8)

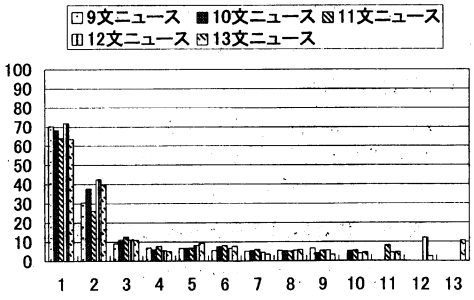

図 4　元原稿各文の形態素出身率(9－13)

これらのグラフから以下の特徴がわかる。

第 1 文の利用

どの文数の元原稿でも、第 1 文からの採用は 70%程度と顕著に高い。これは、要約者の発言を裏付けるものである。また、元原稿の文数が増えると第 2 文からの採用が増える傾向がある。

最終文の利用

図 3 の 4 文ないし 8 文ニュースのグラフでは、以前の報告と同様、最終文からの採用が増加する傾向は見られなかった。しかし図 4 の 11 文、12 文、13 文ニュースでは最終文からの採用がその前の文に比べて顕著に増えていた。13 文からなる元原稿で、最終文の採用の高い 10 件を調査したところ、6 件が台風の動きを伝えるニュースであり、すべて「気象庁の今後の動きを警戒する呼びかけ」であった。以下に例を示す。この例では、要約の主要部分は元原稿第 1 文と最終の第 13 文からできていることがわかる。

要約：「台風 16 号は東北地方を暴風域に巻き込みながら日本海を北東に進んでいます。31 日昼前に東北北部から北海道南部に近づく見込みで、気象庁は北日本を中心に強い風雨と高波に警戒するよう呼びかけています。」

元原稿 1 文：「台風 16 号は青森県沖の日本海を北東に進んでいます。」
元原稿 13 文：「気象庁は北日本を中心に強い風雨と高波に警戒するよう呼びかけています。」

このほか、11 文ニュース、12 文ニュースにも同様の台風のニュースが見られた。文数の多い原稿は図 2 に示したように全体の頻度が少ない上、内容に偏りが見られる。このためこれらが類似の構成を取ることから、最終文からの採用が顕著に観察できたものと考える。
さらに 5 文からなる文で、最終文からの採用の多い順に観察したところ、スポーツニュースが多かった。これらの最終文は(5 文目)試合結果、予定などを報じていた。要約に採用された最終文の例を示す。

「試合はメッツが 7 対 5 で勝ちました。」
「大会は 3 月 23 日に開幕します。」(選抜野球)
「試合はドジャーズが 13 対 4 で勝って石井投手が勝ち投手になり今シーズンの成績は 4 勝 1 敗となりました。」

以上のように、全体を平均すると最終文の採用はそれほど目立たないが、内容によってかなり固定的に発生していると考えられる。

5 要約作業のモデル化

前節で要約原稿と元原稿の第1文第2文との高い形態素の重なりが見られることを報告した。また、スポーツや台風などのジャンルによっては最終文との高い重なりが見られることもわかった。
ここではこれらの知見と要約者への聞き取り調査の結果を合わせて、要約者の作業過程を表すモデルを提案する。以下では、まずニュースの典型構造について説明して、要約者の作業モデルを提案する。

5.1 ニュースの構造

ニュース原稿は基本的にリード文、本記、追記の3つの部分からなる。それぞれの役割は以下の通りである(井上 81)。

リード
「見だし」とも言われる部分であり、ニュースの最も重要な内容を簡潔に記述する。この部分では、具体名を使わず、抽象的な表現を使うことがある。たとえば社名の代わりに「大手生命保険会社」といった表現を使うことがある。政治家やスポーツ選手などについてはこの限りではない。

本記
リードの内容を詳述する。いわゆる 5W1H に相当する情報を補足する。リードで抽象的に述べられた内容をここで具体的に記述する。

追記
リードや本記で述べられていない情報を必要に応じて追加する。実際の原稿では今後の動きや関連情報などが記述されている。

リードと本記で内容を繰り返すのは、ニュース原稿がラジオでの読みを基準に書かれているためである。ラジオでは聞き直しができないため、まず、概要を伝えた上で内容を詳述する方法をとっている。

5.2 提案モデル

要約者が原稿の書き出し部分を利用することは2節で述べた。観察によると、利用しているのはリードであり、その文構造をほぼそのまま生かして要約する場合が多い。また、最終文の採用は、追記の採用であることもわかった。
そこでまずリード、本記、追記を認定し、その後リードを編集する2段階の要約モデルを考えた。具体的な内容を図5に示す。これは要約者の作業モデルではあるが、基本的な要約の計算モデルとも考えている。各機能を自動化できればニュース原稿は自動的に要約できると考えている。

(I) 解析過程
  (a) 構造解析
    ニュース構造を分析し、リード、本記、追記の部分を認定する
  (b) 照応解析
    リードの内容と本記の対応を分析する

(II) 編集過程
  (a) リード文を要約文の起点とする(以下、起点文と呼ぶ)
  (b)下記の操作を字数制限に合わせて選択し、起点文を編集する
    (1) 起点文の短縮
    (2) 起点文の表現を本記の表現で置換
    (3) 起点文に本記の表現を追加
    (4) 起点文に追記文を(短縮して)追加

図 5 要約モデル

各部の自動化は今後検討していく予定である。特に置換には、リードと本記の表現対応を把握する必要がある。これを行うのが図5の(I)(b)の照応解析であるが、現実のリードと本記の置換(表現対応)を見たところ、一般の照応解析の対象より広い問題を扱う必要があることがわかってきた。著者らはリードと本記の表現対応を網羅的に調査したいと考えており、すでに把握している事例の一部を次節で示す。

5.3 照応の実態

下記はそれぞれ、リード文(の一部)と要約原稿(の一部)の対である。リードの表現が本記の表現で置換されたと思われる事例を列挙した。ただし、狭義の置換でなくリードに表現が挿入されている場合も含めていることに注意されたい[6]。対応個所には下線を施した。

・表現の詳細化
  リード:「南極・昭和基地で、1年間活動をしてきた観測隊が基地の運営を次の観測隊に引き次ぐ越冬交代式が行われました。」
  要約:「南極・昭和基地で、1年間活動してきた44次観測隊が次の45次観測隊に引き継ぐ越冬交代式が行われました。」

観測隊が44次、45次観測隊であることが本記にあり、これを要約に反映している。

・発言の採用
  リード:「アメリカのケリー国務次官補は、今日韓国を訪問し、北朝鮮の核開発問題を巡る

[6] リードのヌルポジションが本記の表現で置換されたと解釈した。

次の 6 か国協議が今月中に開かれる可能性について楽観的に考えていると述べました。」
要約：「アメリカのケリー国務次官補は、6 か国協議の早期開催に向けて調整するため 1 日、韓国を訪れました。空港で記者団に対し、『次の 6 か国協議を 2 月中にも開くことができるかもしれないと多少楽観的に考えている』と述べました。」

この例ではリードだけを要約の起点文とはしてはいない。リードの「楽観的に考えている」が本記中の発言に相当していることを認定して、利用している。このような発言の利用はかなり頻繁に起こっている。

・表現の具体化

リード：「今日午後兵庫県北部の浜坂町で体長が 1 メートルを超える熊が住宅に入り込みました。」
要約：「午後 2 時半頃、兵庫県北部の浜坂町の農家に体長が 1 メートルを超えるツキノワグマが入り込み、家にいた夫婦は逃げて無事でした。」

リードの「熊」が本記のツキノワグマと対応することを認定して、置換している。

6 今後の課題

著者らのモデルは、(Mani 1999)の"draft and revision" や (Jing 2000) の "extraction and cut-and-paste generation"などと同類のモデルだと考える。いずれも 2 段階で要約を作成するモデルであり、今後の自動処理実現の参考にしたい。著者らと同じく、放送ニュースの要約のためにリードを利用する研究に、加藤らの研究(加藤 00)がある。この手法は重要文抽出に基づくもので、リードを採用した上で、これと単語の重なりの少ない他の文を追加する。一方、表 1 からわかるように、元原稿の平均文長は 84 文字となるため、この手法単独では、著者らの想定する 105 文字以下の要約は困難である。何らかの「編集」が必要だと考えている。
今回、元原稿の最終文が採用されている要約を調査したところ、文数の多い元原稿で顕著な事がわかった。これは文数の多い原稿が一部の話題(ジャンル)に偏っていたために起こっていた。このことから、元原稿の話題によって個別の要約戦略が見つかる可能性が考えられるため、今後調査したい。

7 おわりに

NHK の要約ニュース作成過程に関する聞き取り調査を行い、要約原稿と元原稿の照合実験を行った。またこれらの結果から要約者の作業モデルを提案した。このモデルは元原稿の解析とリード文の編集の 2 段階からなる。またこの中で重要な役割を果たすリードと本記の表現照応の実態を報告した。

謝辞

本研究を進めるに当たり、ユージンソフトの脇隆三氏にはソフトウエア開発でお世話になった。また日頃からご指導頂く NHK 放送技術研究所の榎並和雅所長、渡辺敏英次長に感謝する。

参考文献

Jing, Hongyan and Kathleen R. McKeown. 1999. The Decomposition of Human-Written Summary Sentences. *The 22nd Annual International ACM SIGIR Conference*, pages 129-136, Berkeley.

Jing, Hongyan and Kathleen R. McKeown. 2000. Cut and Paste Based Text Summarization. *The 1st Meeting of the North American Chapter of the Association for Computational Linguistics*, pages 178-185, Seattle.

Mani, Inderjeet, Barbara Gates and Eric Bloedorn. 1999. Improving Summaries by Revising them, *The 37th Annual Meeting of the Association for Computational Linguisics*, pages 558-565, Maryland.

Mani, Inderjeet. 2001. *Automatic Summarization*. John Benjamins, Amsterdam/Philadelphia.

Marcu, Daniel. 1999. The automatic construction of large-scale corpora for summarization research. *The 22nd Annual International ACM SIGIR Conference*, pages 137-144, Berkeley.

Tanaka, Hideki et al. 2005. Analysis and Modeling of Manual Summarization of Japanese Broadcast News. *The 2nd Int'l Joint Conference of Natural Language Processing (Companion Volume)*, pages 52-56, Jeju, Korea

井上 1981. ニュース文章は変えうるか. 文研月報 12 月号, NHK 総合放送文化研究所, pages 12-21.

加藤、浦谷 2000. 放送ニュースを対象にした重要文抽出. 言語処理学会第 6 回年次大会, pages 237-240.

田中、他 2005. 要約ニュース構築に向けた基礎検討. 言語処理学会第 11 回年次大会, pages 632-635.