ニュース記事の定型性を利用した話題要約の検討

山田一郎　　柴田正啓
ＮＨＫ放送技術研究所　マルチメディアサービス
〒157-8510　東京都世田谷区砧 1-10-11
Tel: 03-5494-3137
E-mail:　{ ichiro, shibata }@strl.nhk.or.jp

あらまし

本稿では、大量の日本語ニュース記事を、要約する手法について述べる。ニュース記事には、報道する話題に特有の単語や句が用いられることが多い。このような表現は、その話題に関係した出来事の経過、経緯を説明するのに不可欠な要素と考えられる。そこで、ニュース記事に含まれる係り受け関係にある２つの単語と助詞の組み合わせからなる表現の出現頻度を統計的に評価することにより、話題特有の表現を選び出し、話題に関係した複数のニュース記事の要約を生成する手法を開発した。この手法の有効性を、実際のニュース記事を使った実験によって検証し、話題を特徴づける属性値などの知識をあらかじめ与えることなく、話題に関係する主要な出来事を含んだ要約文を自動生成できることを確認した。

Multiple Text Summarization using Fixed Expressions in News Articles

Ichiro Yamada　　Masahiro Shibata
Japan Broadcasting Corporation
Science & Technical Research Laboratories
1-10-11 Kinuta, Setagaya-ku, Tokyo 157-8510, JAPAN
Tel: 03-5494-3137
E-mail:　{ ichiro, shibata }@strl.nhk.or.jp

Abstract

In this paper, we propose a new method of summarizing multiple Japanese news articles. News articles are often described using fixed words or fixed syntactic constructions that are distinctive to the topic. These fixed expressions are considered to be elements that are essential to that topic. Our proposing method examines dependency relations between two words included in news articles, and extracts these fixed expressions. This method can generate summary sentences composed of these extracted fixed expressions that included main events without providing preliminary knowledge such as what events are important to the topic. An experimental analysis of actual news articles shows effectiveness of this method.

1　はじめに

近年、放送局では大量のニュース記事データを電子化して蓄積するようになり、これらの効率的な管理、活用が急務となっている。そこで、我々はニュース記事の内容を自動解析する研究を進めている。これまでに、日本語のニュース記事からインデックスとして利用できる話題を自動抽出する手法を提案した[1]。ここでは、ある期間、人口に膾炙する社会事象を「話題」と定義し、抽出を行った。

話題はいくつかの「出来事」の連鎖によって構成され、毎日のニュースはこの連鎖を追跡・報道し視聴者に話題の全体像を提示する。今回、一つの話題を構成するニュース記事集合を分析して、上記の出来事連鎖を抽出し、出来事を自然言語の文によって要約して提示する手法を提案する。

従来、複数ニュース記事を対象とした内容分析に関する研究として、McKeown らはテンプレートを利用して、そのスロットを埋めていく手法を提案している[2]。しかしこの手法では、あらかじめ手作業によりニュースの種類ごとのテンプレートを決めておく必要があり、さらにテンプレートのスロットごとに抽出ルールを生成しなければならない。この作業には大変な労力を要する。

また Radev らは、一つの話題を構成する記事集合をあらかじめクラスタリングして要約を生成する手法を提案している[3]。この手法では、クラスタリングにより、話題に含まれる出来事を記事集合のクラスタとして抽出し、クラスタごとに各記事の重要度を計算して重要度の高い記事を抜き出し、最後に日付順に並べて要約としている。しかし、記事中の重要な一つの出来事が、一つのクラスタを形成しなければならないため、重要な出来事の記事数が少ない話題には適さない。

我々は、これまでに、一つの話題を構成するニュース記事を時系列に並べ、そこに出現する単語の変化によってニュース記事の内容の変化点を統計的に抽出し、その変化点間の記事の重要度を計算して、変化点間の最重要記事を提示する要約手法を提案した[4]。しかし、この手法では、同時期に複数の出来事が起きるような話題は処理できない。

複数の関連する話題から構成される「話題分野」には、話題を構成する出来事の連鎖に共通なモデルが存在する。例えば「選挙」の分野に属する多くの話題には、「選挙の公示」「立候補の表明」「選挙活動」「投票」「開票」といった出来事が出現する。モデルを構成する各項は、出来事に関する一種のフレームであり、それぞれ複数のスロットを持つ。例えば、上記「投票」では、「選挙名」「立候補者名」「投票率」「投票日」などのスロットがある。各話題中のニュース記事で、これらのスロットの値を伝える場合、決まった表現が利用されることが多い。この特徴を、「定型性」と定義する。この定型性は、ニュース記事が属する分野を特徴付けるため、話題要約に有効な性質となる。

本稿では、この定型性を、係り受け関係にある２つの単語と助詞の組み合わせを指標にして評価することにより、一つの話題を構成するニュース記事集合を要約する手法を提案し、実際のニュース記事を使った実験を示す。

2　話題を構成するニュース記事の定型性

NHK の放送用読み原稿として利用されるニュース記事を処理対象とする。このニュース記事は、１日当たり約 200 記事が作成されている。我々が所有するデータベースには、10 年分のニュース約 33 万記事（200 万文）もの大量のテキストデータが蓄積されている。

本手法では、異なる話題分野ごとに定型性を評価しなければならない。そこで、ニュース記事を関連する記事集合にクラスタリングして、クラスタごとに、定型性を評価する。クラスタリングは、我々が従来から提案している手法を利用した[1]。この手法では、適合率 92.2%、再現率 93.6%と高精度でクラスタリング処理ができる。

以下に、ニュース記事の特徴を紹介する。本手法は、このニュース記事の特徴を利用して要約を行う。

2．1　ニュース記事の特徴

ニュース記事の第一文は“リード”と呼ばれ、５W１Hなどの内容が具体的に記述されている[5]。そのため、リードはニュース内容の全貌を説明する事が多く、これに対して、第二文以降は情報抽出処理においてノイズとなりうる要素が多い[6]。本手法では、記事の第一文のみを利用する。

リードでは、主観的な表現、あいまいな形容詞、副詞は使われない。例えば、「良い天気」といった表現は避けられ、代わりに「雲ひとつ無い天気」という表現が使われる。

また、ニュース記事では、その中で使う動詞にも一定のルールがある。ニュースは聞き返しができないため、その記事中では漢語表現などの難しい言葉は避けられる。例えば「判明する」の代わりに「わかる」に、「示唆する」の代わりに「ほのめかす」が使われる。さらに、汎用的な動詞表現も避けられる。例えば、「行われる」という動詞表現はできるだけ避けられ、主格が「議論」の場合は「行われる」の代わりに「交わされる」、「会合」の場合は「開かれる」が使われる。ニュース記事には、このような一定の制約があるため、統計処理を利用した情報抽出に適している。

さらに、前章でも述べた通り、同じ話題分野には、類似内容の出来事が繰り返し出現する共通のモデルが存在する。そのため、対応するニュース記事では、定型的な表現が多く出現するという特徴を持つ。例えば、以下は国会審議における法案可決時の定型的な表現「～法案は、～で、～賛成多数で可決され、～に送ら

れました。」の例である。

アメリカなどの軍事行動に自衛隊が支援するためのテロ対策特別法案が、きょうの衆議院本会議で、与党三党などの賛成多数で可決され、参議院に送られました。

そこで、この定型部分を含むニュース記事を、特定の話題分野における共通なモデルを構成する出来事を表していると判断する。この定型部分を含むニュース記事から要約文を生成し、それらを集めて話題全体の要約を生成する。この際、異なる話題分野ごとに単語や統語構造の定型性を評価しなければならない。本手法では、ニュース記事集合の定型部分抽出のために、２つの単語の係り受け関係（以後、助詞も含めて３項組と呼ぶ）の定型性に注目する。

２．２　係り受け関係を利用した定型性評価

ここでは、まず最大エントロピー法による構文解析[8]を利用して、係り受け関係を持つ２つの単語と助詞（直接係る場合はφ）の３項組を抽出する。そして、特定の話題分野において特徴的な単語間の係り受け関係を抽出する。この処理では、特定の話題分野における３項組の特異性を評価するために、観測値と期待値がどの程度一致しているかを測る指標である χ^2 値を利用した。母集団を 10 年分のニュース記事(330,066 文)のうちの国会審議に関する記事(9,227 文)とした。３項組(w_1,w_2,w_3)の出現頻度を $n(w_1,w_2,w_3)$、その期待値を $e(w_1,w_2,w_3)$としたとき、$\chi^2(w_1,w_2,w_3)$は次の式で与えられる。

$$\chi^2(w_1,w_2,w_3)=\frac{(n(w_1,w_2,w_3)-e(w_1,w_2,w_3))^2}{e(w_1,w_2,w_3)}$$

観測値 $n(w_1,w_2,w_3)$が期待値 $e(w_1,w_2,w_3)$より小さい場合は $\chi^2(w_1,w_2,w_3)=0$ とした。このとき、単語の属性が人名、組織名、地名、数値名である場合は、抽象化した属性名を利用し、例えば「自民党の政策」と「社会党の政策」は、共に“組織名“の政策」として χ^2 値を計算する。χ^2 値が大きい３項組ほど、その話題に偏って出現していると言える。

また、記事中に頻繁に出現する３項組は、その内容を特定するための分別能力に乏しい。例えば、衆議院総選挙の話題では、「衆議院の総選挙」という３項組は、ほとんどのニュース記事で出現するため、この話題を対象とした内容解析処理では不要な要素となる。そこで、そのような３項組の値を制限するために、*IDF* 値を利用した。対象とする話題を構成するニュース記事の総数を N、ニュース記事中で３項組(w_1,w_2,w_3)が出現した記事数を $DF(w_1,w_2,w_3)$としたとき、$IDF(w_1,w_2,w_3)$は次の式で与えられる。

$$IDF(w_1,w_2,w_3)=\log\frac{N}{DF(w_1,w_2,w_3)}$$

さらに、品詞の組み合わせにより、定型性評価の重み付けに制限を与える。品詞による制限値 $C(w_1,w_2,w_3)$は、（名詞、助詞、動詞）の組み合わせを最重要とし、表１に示す値とした。

表 1. 品詞による重み付け

w_1,w_2,w_3	$C(w_1,w_2,w_3)$
名詞, 助詞, 動詞	1.0
名詞, 助詞, 名詞	0.2
動詞, φ, 動詞	0.1
その他の組み合わせ	0.05

χ^2 値、*IDF* 値、さらに品詞による制限値を相乗的に利用することにより、話題の基本的な構成要素を抽出するための３項組の定型値 $weight(w_1,w_2,w_3)$を以下のように定義した。

$$weight(w_1,w_2,w_3)=\chi^2(w_1,w_2,w_3)\times IDF(w_1,w_2,w_3)\times C(w_1,w_2,w_3)$$

この値が大きいほど、対象とする特定の話題における定型的な表現と考えられる。

２．３　定型値の例

表２に「テロ対策基本法案の国会審議」に関するニュース記事に出現した３項組の定型値計算結果の上位 20 組を示す。「賛成多数で可決される」「参議院に送られる」といった、国会審議に関するニュース記事の型にはまった表現が上位にある。逆に「経済の問題」といった国会審議に特有の表現でない３項組の $weight(w_1,w_2,w_3)$の値は 0 であった。

表 2.　３項組の定型値計算結果（上位 20 組）

weight	3項組
6038.2	国会/に/提出する
5517.8	賛成多数/で/可決される
3686.2	参議院/に/送られる
2153.9	参議院本会議/で/可決される
2032.7	衆議院本会議/で/可決される
1313.5	考え/を/示す
1254.4	考え/を/強調する
1140.5	衆議院/を/通過する
1026.7	国会対策委員長/が/会談する
992.6	審議/が/始まる
917.5	政府/は/提出する
885.9	成立/を/目指す
882.6	衆議院/に/提出する
779.6	政府/が/提出する
529.7	特別委員会/を/設置する
528.9	本会議/で/行われる
527.0	所信表明演説/に/対する
502.5	出席/を/求める
500.0	修正/を/めぐる
487.1	修正/を/行う

3　ニュース記事の要約処理

２章では、話題分野ごとの定型的な表現を評価する手法を説明した。この値を利用して、当該分野の特定の話題を構成する複数ニュース記事を要約する。この処理では、まず、一つの話題を構成する複数ニュース記事を対象として、含まれる３項組の定型値により、各記事からそれぞれ定型的な文（以後、定型文）を生成し、その重要度を評価する。この処理は話題を構成する全てのニュース記事に対して行う。次に、生成した定型文に主語や目的語を補う。この定型文の内容が、確定している事項か否かを判定し、確定事項のみを抽出する。最後に、内容が重複するものを除去し、重要度が高いニュース記事集合から生成された定型文集合を、全体の要約として提示する。以下に各処理を説明する。

3．1　3項組の定型値を利用した定型文生成

３項組の定型値を利用して、ニュース記事から、定型的な文を生成する。３項組が少しでもそのニュース記事の属する話題分野に依存する場合は、その定型値は0より大きな値をとる。そこで本実験では、定型値が0より大きい組を抽出し、共通する項を持つ３項組を、出現順に統合して文を生成した(図１参照)。また、このとき、統合した３項組が持つ定型値の合計を、重要度の指標となる「文の定型値」とした。

3．2　主語、目的語補完

前節の処理では、一つのニュース記事から、その記事が属する話題分野における定型的な係り受け関係のみを抽出して統合している。しかし、話題ごとに異なる要素は、定型値は小さくなり、前節の処理では抽出できない。つまり、文の主語や目的語などの、要約において重要な要素が抜けてしまう。そこで、前節で生成された文に含まれる動詞に対して、その主語と目的語を補完する必要が生じる。この処理では、構文解析結果を利用した。この結果生成された文を、定型文と呼ぶ。定型値が一定以上の307個の定型文を対象に実験を行った結果、表３に示す通り、主格と場所格の補完が行われた。

表 3．主格・場所格の補完結果

	補完前	補完後
主格の存在率	23.1%	67.1%
場所格の存在率	68.2%	77.6%

図１に定型文生成例を示す。与えられたニュース記事から４つの定型的な３項組が抽出され、共通項の「可決される」「賛成多数」を持つ３項組を順に統合していくことにより、「衆議院本会議で与党三党などの賛成多数で可決されて成立する」という文が生成できる。最後に、動詞「可決される」に係る句の「テロ対策特別法が」を補完している。また、この例における文の定型値は、統合した３項組の定型値の総和（7765.3）となる。処理対象とする話題を構成する全てのニュース記事に対してこの処理を行うため、各記事ごとに図１に示すような定型文（NULLも含む）が生成され、文の定型値が与えられる。

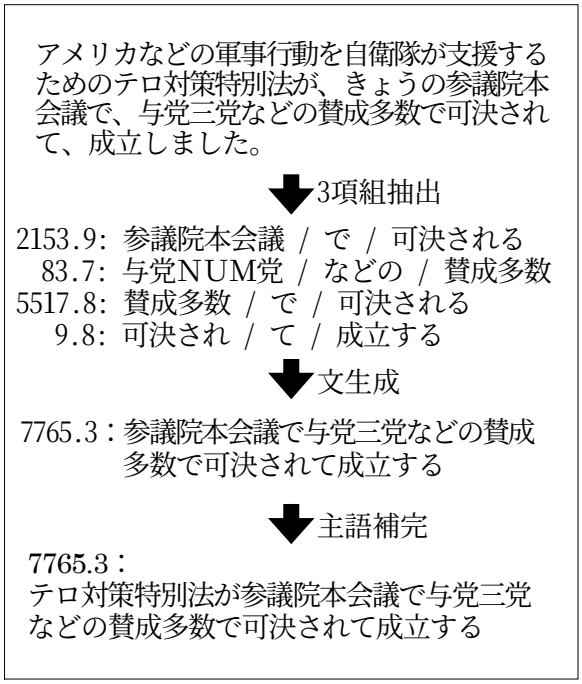

図 1．ニュース記事からの定型文抽出処理例

3．3　動詞の確定・未確定の判定

ニュース記事には未確定な出来事が含まれる。たとえ、文の定型値が大きい定型文であっても、そこに含まれる出来事が未確定な事柄である場合は要約文として適さない。例えば、次のニュース記事には「開く」「支援する」「求める」「行う」「決める」の５つの動詞が出現している。

> 参議院の外交防衛委員会が、きょう開かれ、アメリカなどの軍事行動に自衛隊が支援するための「テロ対策特別法案」の審議日程について、来週二十三日と二十四日の二日間、小泉総理大臣とすべての閣僚の出席を求めて、質疑を行うことを決めました。

ここで、「開く」「決める」は既に実施された事実を述べた確定事項だが、「支援する」「求める」「行う」は、実施中、もしくは、これから実施される予定の未確定事項である。そこで、ニュース記事に含まれる全ての動詞を対象に、確定・未確定の判定処理を行い、確定と判定された動詞を文末に持つ定型文のみを抽出した。

また、この処理では、事態の確実性を表す名詞[7]（「こと」「考え」「方針」「意向」「見通し」など）以外の名詞を修飾する動詞を、文の主題とは無関係と判断し、確定・未確定の判定処理の対象から除いた。上記の例では、「支援する」の判定処理は行わない。

確定・未確定の判定処理は、動詞の時制を利用する。動詞が「過去形」の場合は確定、「現在形」の場合は未確定とした。過去にあった出来事を表す動詞は、通常、

過去を表す助動詞「た」を伴うことにより過去形となる。しかし、動詞が複数出現する複文では、最後に出現する動詞以外は、過去を表す助動詞が省略されてしまう。また、条件文でも例外が生じる。そこで、以下の場合は、例外処理を行い判断した。

- 連用修飾節の動詞の場合
  係り先の連用節と同じ時制として判定する
  例：「・・・と述べ、・・・ことを示しました。」
  →「述べ」は「示しました」と同じ時制「過去」として「確定」と判定
- 条件を表す節中の動詞の場合
  過去を表す助動詞「た」を伴っても未確定とする
  例：「日本に武力攻撃が加えられた場合は、・・・」
  →「加えられた」は「未確定」と判定

この処理を無作為抽出した331個の国会審議に関するニュース記事に対して行い、手作業による結果と比較検証した。その結果を表4に示す。出現した929個の動詞中、810個(87.2%)の動詞に対して正解が与えられ、ある程度、良好な結果が得られている。確定事項を未確定と誤判定してしまった原因の多くは、連用修飾節における係り受け解析の失敗によるものであった。

表 4． 確定・未確定の判定結果

	確定事項	未確定事項
確定と判定	354(95.7%)	16(4.3%)
未確定と判定	103(18.4%)	456(81.6%)

3．4 定型文の削除

一つの話題を構成するニュース記事集合には、同一内容について述べたニュース記事も数多く存在する。そのため、類似内容の定型文も複数抽出してしまう。そこで重複する定型文を削除する処理を行う。この処理では、以下の2つの条件を満たす場合に重複した定型文と判断し、定型値が低い文を削除する。

- 一定値（本実験では 0）より大きい定型値を持つ3項組の係り受け関係で、その内容に不整合（2項が同じで1項のみ異なる組み合わせ）が存在しない
- 共通である3項組の定型値の合計が一定値以上（本実験では、{min(2文の定型値)/2}以上）

例えば、抽出された定型文の「衆議院本会議で可決される(定型値 2153.9)」と「衆議院本会議で与党三党などの賛成多数で可決され参議院に送られる(定型値 11602.4)」は上記の条件を満たすため、文の定型値が低い前者は削除される。

また、定型文の中で、文末の動詞が「発表語」で、その前に「こと」以外の「事態の確実性を表す名詞」がある場合は、その前に述べられた行為の確定性が低いことが判っている[7]。そこで本手法では、「考えを表明する」などが含まれる定型文を削除した。

最後に、その文の定型値が一定値（本実験では500）以下の定型文を削除することにより、要約結果とした。

表 5． 話題「テロ対策基本法案の国会審議」に関するニュース記事の要約結果

日付	要約結果	文の定型値
2001/10/4	衆議院議院運営委員会は理事会で「テロ対策特別法案」などを審議するため特別委員会を設置することを決める	642.3
2001/10/5	政府は「テロ対策特別法案」を決定し国会に提出する	7041.4
2001/10/5	法案をまとめ衆議院に提出する	1093.0
2001/10/8	早期成立に向けて協力を求めることを決める	963.3
2001/10/9	特別委員会は出席を求めて質疑を行なうことで与野党が合意する	1164.5
2001/10/10	十九日までに法案の成立を目指すことで一致する	922.7
2001/10/10	参考人質疑を十五日には一般質疑を行なうことを決める	1215.2
2001/10/15	小泉総理大臣は野党各党の党首と個別に会談し早期成立に向けて協力を要請する	940.7
2001/10/15	修正を行なう考えを示して賛成を求める	1644.6
2001/10/16	与党三党などの賛成多数で可決される	5601.5
2001/10/16	テロ対策特別法案が衆議院の特別委員会で可決されたことについて小泉総理大臣は総理大臣官邸で記者団に対し述べる	631.6
2001/10/17	社民党は衆議院本会議でテロ対策特別法案が採決されるのを前に国会内で反対集会を開き土井党首は述べる	599.7
2001/10/18	「テロ対策特別法案」が衆議院本会議で与党三党などの賛成多数で可決され参議院に送られる	11602.4
2001/10/18	テロ対策特別法案が衆議院を通過したことについて中谷防衛庁長官は記者団に対し述べる	1568.6
2001/10/18	小泉総理大臣は「テロ対策特別法案」が衆議院本会議で可決されたことについて述べる	2039.4
2001/10/19	合同理事会が開かれ審議日程について出席を求めて質疑を行うことを決める	784.4
2001/10/26	「テロ対策特別法案」は参議院外交防衛委員会で与党三党の賛成多数で可決される	5588.3
2001/10/28	自民党の山崎幹事長は記者団に対し述べる	675.0
2001/10/29	テロ対策特別法が参議院本会議で与党三党などの賛成多数で可決されて成立する	7765.3

全ての定型文に対して、その重要性を表す定型値を与えたため、ニュース記事の要約率は、このしきい値を変化させることにより、容易に変更できる。

３．５　実験

話題「テロ対策基本法案の国会審議」を構成するニュース212記事を要約した結果を表5に示す。法案の国会への提出、参考人質疑、衆議院本会議の可決、参議院外交防衛委員会での可決、参議院本会議での可決成立など、主要と考えられる要素が、適切な短文で抽出されている。

このニュースの出現数の推移を図２に示す。衆議院における審議が行われ可決した10月中旬には、ニュース記事が多く出現しており、この時期に世間で騒がれていたことがわかる。しかし、このニュース記事の出現数と、話題における出来事の重要さは、必ずしも一致しない。法案が参議院で可決され成立した10月26日には、「テロ対策法案の審議」に関するニュースは6個しか出現していない。このような重要な出来事に対するニュース記事が少ない話題は、クラスタリングや、単語の変化点を用いる従来手法では、要約することが難しい。しかし、本手法では、表４に示すとおり、参議院本会議での可決のニュースにも高い定型値が与えられ、抽出されている。

国会審議に関する34個の話題を対象に、その話題を構成する複数ニュース記事を要約した結果を検証した。その結果、衆議院／参議院本会議における法案可決に関する記述は、再現率90.7%で抽出され、良好な結果が得られた。抽出されなかった理由の71%が、動詞の確定・未確定判定の誤りにあり、残りが定型値のしきい値による問題であった。今後、過去を表す助動詞以外の要素も取り入れた確定・未確定判定処理が必要と考えられる。

また、法案可決の要約結果の22.4%は、場所格が抜けていたため、どこで可決されたか、という重要な要素が落ちていた。表４でも、10月16日に「与党三党の賛成多数で可決される」という定型文が生成されている。この定型文は、「衆議院の特別委員会で」という場所格が抜けている。ニュース記事では、冗長な部分を少なくするため、複文の構造を持つ場合は、動詞の共通の格は省略されてしまう。そこで、要約時に省略された場所格を補完する必要が生じる。これは、従来から提案されている主語補完の技術[9]を応用することにより解決可能と考えられる。

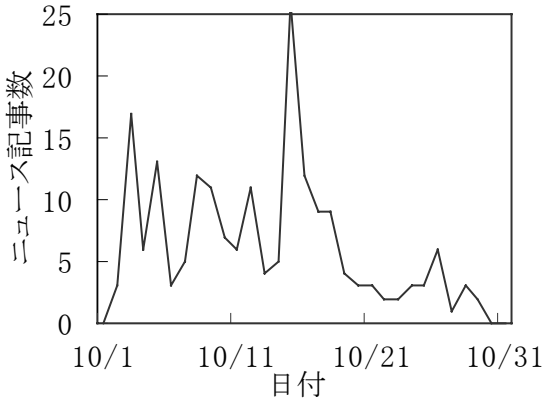

図２．「テロ対策法案の審議」に関するニュースの出現記事数

４　おわりに

本論文では、特定の話題を構成するニュース記事の解析を行い、ニュース記事の定型性を利用することによる要約手法を提案した。「法案を国会で審議する話題」に関するニュース記事を対象とした実験の結果、重要な出来事を自然言語の文により的確に提示でき、良好な要約結果が得られた。今回、国会審議に関する話題の要約を実験対象としたが、他の話題でも、基本要素をテンプレートで表現できるような話題に関するニュース記事であれば、本手法を適応することにより、要約可能と考えられる。

今後、本手法をニュース記事以外のテキストへ応用して、大量テキストデータの構成要素分析、さらには情報発見へと進めていく予定である。

【参考文献】

[1] I. Yamada:“Topic Event Detection using Japanese News Articles”, In Proc. of the NLPRS1999, 375-380(1999)

[2] McKeown and D. R. Radev: “Generating Summaries of Multiple News Articles”, In Proc. of the SIGIR-95(1995)

[3] D.R. Radev, H. Jing, and M. Budzikowska.: “Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies”, In Proc. of the ANLP/NAACL2000 Workshop on Automatic Summarization, pp.21-30,2000

[4] 山田ほか「ニュース原稿を利用した話題トラッキングの検討」情処学会第６１回全国大会講演論文集(3), pp193-194(2000)

[5] 奥秋義信「ニュース原稿の書き方～その理論と実際」岩崎放送出版社(1970)

[6] 加藤ほか「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回年次大会論文集,pp237-240(2000)

[7] 木田ほか「情報抽出のための文末表現分析」言語処理学会第６回年次大会論文集, pp304-307(2000)

[8] 江原「最大エントロピー法を用いた日本語文節間係り受け整合度の計算」言語処理学会第5回年次大会論文集,pp382-385(1999)

[9] Yeun-Bae Kim and Terumasa Ehara :”A Method of Partitioning of Long Japanese Sentence with Subject Resolution in J/E Machine Translation”, Proceedings of the 1994 International Conference on Computer Processing of Oriental Languages, pp467-473 (1994)