ニュース記事からの話題構成要素抽出の検討
～国会審議に関する話題を対象として～

山田一郎　　金淵培　　柴田正啓　　浦谷則好
ＮＨＫ放送技術研究所
E-mail: { ichiro, kimyb, shibata, uratani }@strl.nhk.or.jp

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、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]。この処理では、ニュース記事を話題ごとに分類し、話題表示用の名詞句を記事の中から抽出している。今回、この抽出された話題を構成するニュース記事集合の分析により、いつ、何が起きたかという話題構成要素を抽出し、自然言語の文を合成して提示する手法を提案する。(以後、話題構成要素を表す自然言語文を、話題要素文と呼ぶ)

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、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が提案されている[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。

話題を構成する基本要素は、ニュース記事では、話題に特有の単語や統語構造を用いて表現されることが多い。そこで本手法では、ニュース記事中の単語や統語構造が、その記事に属する話題を特徴付ける性質を定型性と定義する。この単語や統語構造の定型性を評価することにより、あらかじめテンプレートを用意することなく、話題を構成するニュース記事集合の解析を行う。話題を構成するニュース記事集合の抽出には、我々が従来から提案しているクラスタリング手法[1]を適用することができる。今回、分類抽出された「ガイドライン関連法案の審議」に属するニュース記事集合の解析実験を行った。以下にその内容を報告する。

2 特定の話題を構成する複数ニュース記事解析

NHK の放送の読み原稿として利用されるニュース記事データベースは、１日当たり約２００記事が蓄積されている。各記事の第一文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[3]。そのため、本手法では、記事の第一文のみを利用する。

また、ニュース記事には、話題に依存した定型的な表現が多く出現する。例えば、以下は国会審議の話題における法案可決時の典型的な表現「～法案は、～で採決され、～賛成多数で可決されました。」の例である。

> 日米防衛協力の指針・いわゆるガイドラインの関連法案は、先程、参議院の特別委員会で採決され、自民、自由、公明の三党などの賛成多数で可決されました。

このような定型的な文は、その話題に必要不可欠な基本要素と考えられる。そこで、この定型部分を話題要素文の候補とする。本手法では、ニュース記事集合の定型部分抽出のために、単語間の係り受け関係の定型性に注目する。しかし、例で示したような１文全体の係り受け構造を学習する手法では、係り受け関係の組み合わせ数が大きくなり、スパースネスの問題が生じてしまう。そこで、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）のみを利用する。

2.1 係り受け関係を利用した定型性評価

ここでは、まず江原[5]による構文解析を利用して、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８年分(1992年～1999年)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。

また、話題を構成する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 2. 3項組の定型性評価結果（上位30組）

weight	3項組
5721.2	賛成多数 / で / 可決される
3865.1	参議院 / に / 送られる
3305.3	衆議院選挙 / に / 向ける
2922.8	次 / の / 衆議院選挙
2417.4	衆議院本会議 / で / 可決される
2346.0	参議院本会議 / で / 可決・成立する
2031.5	国会内 / で / 述べる
1468.4	政府 / は / 提出する
1078.9	考え / を / 示す
988.0	国会対策委員長会談 / が / 開かれる
918.1	参考人質疑 / が / 行われる
828.6	採決 / が / 行われる
822.9	国会対策委員長 / が / 会談する
810.0	修正案 / が / 可決される
799.3	予算案 / が / 通過する
786.9	衆議院 / を / 通過する
785.4	今 / の / 国会
753.1	成立 / を / 目指す
712.1	質疑 / が / 行われる
676.1	賛成多数 / で / 可決・成立する
654.4	総括質疑 / が / 始まる
593.8	審議 / が / 始まる
573.8	通常国会 / に / 提出される
570.3	衆議院予算委員会 / で / 関連する
562.2	施政方針演説 / に / 対する
547.9	考え / を / 強調する
535.7	成立 / を / 図る
487.3	調整 / が / 続く
484.3	衆議院 / の / 解散・総選挙
476.3	早期成立 / に / 向ける

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決まった表現と考えられる。表2に「ガイドライン関連法案の審議」に出現した3項組の定型性評価結果の上位30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。

この3項組を利用して、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分野に依存する場合は、3項組の定型値は0より大きな値をとる。そこで本実験では、その定型値が0より大きい3項組を抽出し、共通する項を持つ3項組を統合して文を生成した。このとき、3項組が持つ定型値の合計が、文の定型値となる。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」「送られる」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、賛

日米防衛協力の指針・いわゆるガイドライン関連法案は、きょうの衆議院本会議で、自民、自由両党と公明党・改革クラブの三会派による修正の上、賛成多数で可決され、参議院に送られました。

↓3項組抽出

2417.42: 衆議院本会議/で/可決される
5721.28: 賛成多数/で/可決される
323.851: 可決され//送られる
3865.17: 参議院/に/送られる

↓文生成

12327.7:
衆議院本会議で、賛成多数で可決され、参議院に送られる

図 1. ニュース記事からの定型文抽出処理例

成多数で可決され、参議院に送られる」という文が生成できる。

2.2 動詞の確定・未確定の判定

前節の処理により、特定の話題に属するニュース記事集合の定型的な表現を抽出できた。しかし、この定型的な表現中の、全ての動詞が確定した事柄であるとは限らない。例えば、次のニュース記事には「審議する」「開く」「求める」「行う」「入る」「決める」の6つの動詞が出現している。

> ガイドライン関連法案を審議している参議院の特別委員会は、きょう理事懇談会を開き、来月十日に、小渕総理大臣と全ての閣僚の出席を求めて総括質疑を行い、審議に入ることを決めました。

「開く」「決める」は既に実施した事実を述べた確定事項だが、「審議する」「求める」「行う」「入る」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。本手法では、既に実施された話題の構成要素を抽出することを目的とし、確定事項のみを対象とする。

この処理では、事態の確実性を表す名詞[4]（「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、「審議する」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。動詞の時制が「過去」を表す「タ形」の場合は確定、「ル形」の場合と時制が不明確な場合は未確定とした。ここで、以下の場合は、例外と判断する。

- 条件を表す名詞が存在する場合
  動詞が「タ形」でも、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
- 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

この処理を話題「ガイドライン関連法案の審議」を構成する331個の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3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3. 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%)

2.3 話題要素文抽出

定型性評価結果と動詞の確定・未確定の判定結果を利用して、話題の構成要素を抽出する。ここで、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[4]。そのため本手法では、「考えを表明する」などが含まれる定型文は抽出結果から除いた。

さらに、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 2417.4)」と「衆議院本会議で、賛成多数で可決され、参議院に送られる(定型値 12327.7)」は上記の条件を満たすため、文の定型値が低い「衆議院本会議で可決される」は削除される。

確定と判定された動詞を文末に持つ定型文で、その定型値が一定値（本実験では500）以上の文から、行為の確定性が低い文と、重複した定型文を削除することにより、話題要素文を抽出した。

話題「ガイドライン関連法案の審議」に関する331文のニュース記事から話題要素文を自動抽出した結果を表4に示す。衆議院本会議での趣旨説明、特別委員会の参考人質疑、衆議院本会議の可決、参議院特別委員会の可決、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。また、この話題に関するニュース記事の出現数の推移を図2に示す。ニュース記事が多く出現した4月27日は衆議院本会議で可決、5月24日は参議院特別委員会、本会議で可決されたことが一覧できる。

表 4. 話題要素文抽出結果

日付	話題構成要素	定型性
1999/2/12	理事会で、指針、いわゆるガイドラインの関連法案を審議する特別委員会を設置することを決める	645.2
1999/3/12	衆議院本会議で、趣旨説明と質疑が行なわれ、法案の早期成立に野党側の協力を求める	1361.9
1999/3/19	今の国会での成立を目指すことを確認する	1592.7
1999/3/29	国会対策委員長が、会談し、目指すことを確認する	832.4
1999/4/1	衆議院の特別委員会は、理事会で、法案の採決を行うよう求める	507.7
1999/4/7	衆議院の特別委員会で、参考人質疑が行なわれ、四人の参考人が、意見を述べる	1447.2
1999/4/26	修正案が可決される	810.0
1999/4/26	今の国会での成立に向けて協力を要請する	1218.4
1999/4/27	採決が行われ、賛成多数で可決される	6590.2
1999/4/27	衆議院本会議で、賛成多数で可決され、参議院に送られる	12327.7
1999/4/27	衆議院を通過したことについて、国会内で記者団に対し、述べる	3250.1
1999/4/28	参議院の特別委員会は、理事懇談会を開き、出席を求めて総括質疑を行い、審議に入ることを決める	1044.8
1999/5/9	参議院本会議で質疑が行われる	1093.8
1999/5/13	参議院の特別委員会で、参考人質疑が行われ、意見を述べる	936.4
1999/5/20	初会合を開き、今の国会で法案の成立を目指す方針を確認する	1979.1
1999/5/24	参議院の特別委員会で採決され、三党などの賛成多数で可決される	5989.3
1999/5/24	参議院本会議で採決され、三党などの賛成多数で可決され、成立する	6462.0

2.4 課題

本手法は構文解析結果を利用して定型文を生成しているため、構文解析で失敗すると、そのまま定型文抽出結果に影響する。表4の結果では、各定型文の主語となる要素があまり抽出されていない。これは、ニュース記事中では主語と動詞は離れて位置し、主語からの係り先は1つに制限しているため、動詞の主格抽出に失敗することが原因となっている。今後、動詞の主格抽出は別処理を行うなどの改善が必要と考えられる。

3 おわりに

本論文では、抽出された話題内の解析を行い、ニュース文の定型性を利用することにより話題要素文を生成する手法を提案した。実験では国会審議に関する話題を対象としたが、基本要素をテンプレートで表現できるような話題であれば、本手法を適応し、その構成要素を抽出できる。本手法は、さらに、従来手法の課題であったテンプレートを自動生成にも応用可能と考えられる。

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。

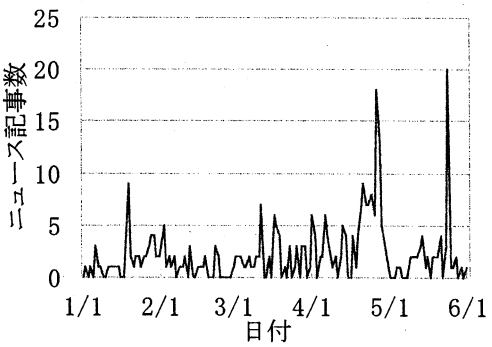

図 2. 「ガイドライン関連法案の審議」に関するニュースの出現記事数

【参考文献】

[1]I. Yamada:"Topic Event Detection using Japanese News Articles", NLPRS1999, 375-380(1999)

[2]McKeown and Radev: "Generating Summaries of Multiple News Articles", SIGIR-95(1995)

[3]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)

[4] 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)

[5] 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)