5U-06

ニュース原稿を利用した話題トラッキングの検討

山田一郎　　金淵培　　柴田正啓
NHK放送技術研究所

1. はじめに

放送局では、現在、大量のニューステキストデータを蓄積している。ニュースには社会の最新の情勢や動向などの有益な情報が含まれているが、毎日更新されるため、人手により一貫性のあるインデックスを付加することは難しい。そこで、このようなデータを自動分類、自動インデキシングする技術が重要な課題となる。我々は、これまでにニューステキストデータから話題を自動抽出して、インデックスとして利用する手法を提案してきた[1]。しかし、この手法では一ヶ月ごとに期間を区切って話題を抽出しているため、長期にわたり継続するものは、複数の月の話題に分割されてしまう。そこで今回、複数月にまたがる話題のトラッキング処理を行い、一つの話題がいつ発生しどのように変化したかを表す話題の構成要素を抽出する実験を行ったので報告する。

2. 話題トラッキング

我々が提案した話題抽出手法[1]では、一ヶ月分のNHKニュース原稿の第1文を入力とし、まず、原稿に含まれる単語の重要度をその出現頻度の時期変化に注目して定義する。次に、この重要度を利用してニュース原稿のクラスタリングを行う。各クラスタが話題の候補となり、その重要度を評価して、クラスタを特徴付ける話題語を抽出、提示している。1999年10月の話題語抽出の結果上位8項目を表1に示す。話題の上位項目のクラスタリング結果では、適合率92.2%、再現率93.6%と良好な結果が得られている。

[1] ウラン燃料加工施設の臨界事故
[2] 中央アジアのキルギス
[3] 山陽新幹線北九州トンネル
[4] 政府関係の主要施設
[5] 西村前防衛政務次官の辞任問題
[6] 介護保険制度
[7] 燃料
[8] 茨城県

表1. 話題語抽出実験結果（1999年10月）

A Study on Topic Tracking Using News Corpus
Ichiro Yamada, Yeun-Bae Kim, Masahiro Shibata
Japan Broadcasting Corporation
ichiro@strl.nhk.or.jp

2.1 話題トラッキング

1999年1月～12月の期間で抽出された話題の上位項目を対象に、手作業により、その前後の月に関連した話題が存在するか否かを調査したところ、その29.2%で関連する話題が存在した。つまり、多くの話題に継続性があることがわかる。前後の月に抽出された話題との関連性が把握できれば、複数の月に分割された話題を一つの話題として認識できる。

抽出された話題は、その話題を構成するニュース原稿に含まれる単語を要素に、単語の重要度を要素の値に持つベクトルにより特徴付けられている。このベクトルと、隣接する月の話題が構成する特徴ベクトルとの類似度を以下の式で定義する。

$$\text{類似度} = \frac{\text{共通する要素の値の和}}{\text{2つの特徴ベクトルの要素の値の和} - \text{共通する要素の値の和}}$$

この値が一定のしきい値（本実験では0.2）より大きい時、関連性があると判断する。対象期間の全ての隣接する月の関連性の評価を行うことにより、話題のトラッキング処理が行われる。図1に1999年10月の話題「中央アジアのキルギス」からのトラッキング結果を示す。1999年1月～12月の話題に対するトラッキング結果を、人手による抽出結果と比較したところ、適合率73.5%、再現率53.2%であった。

2.2 話題トラッキング結果の内容分析

図1では、適切なトラッキング処理がされているが、図中の話題語だけでは、どの時期に何が起きているかわからない。そこで、トラッキング処理により関連付けられた話題を構成するニュース原稿を再度分析

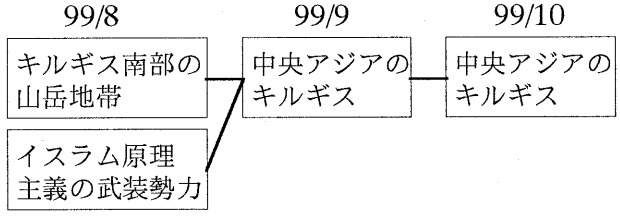

図1.「中央アジアのキルギス」（99年10月）からの話題トラッキング結果

し、長期にわたる話題の構成要素を抽出する。

まず、話題の内容は時系列で徐々に変化すると仮定して、その変化点を求める。そのために、話題を構成するニュース原稿を時間軸上に並べ、一定数の原稿に含まれる単語の変化を抽出する。ここでは、時系列に連続する10個のニュース原稿単位で処理を行い、その第1文を特徴付けるベクトルの集合と平均ベクトルとの距離を求めた。ベクトルの要素をその原稿に含まれる単語、要素の値をTFIDF値とした。結果を図2に示す。平均ベクトルとの距離が近い時はその話題の全体的な内容が、遠い時は局所的な内容が述べられていると推測できる。この局所的な内容を話題の構成要素と判断し、平均からの距離を示す関数の極小値を話題の変化点とした。(図2の点線部分)

次に、各変化点間のニュース原稿から、その話題の構成要素を表す名詞句を生成する。この処理では、まず、各変化点間において単語のTFIDF値を再度計算し、ニュース原稿に含まれる単語のTFIDF値の和が最大のものを、その範囲の代表原稿として抽出する。この代表原稿から、話題の構成要素を表す名詞句を、以下の手順で生成する。

1. 「～の事件で」「～の問題で」といった、話題全体を説明する節を、処理の対象から除く
2. さ変動詞の中でTFIDF値が最大となる動詞の語幹を抽出する
3. 2.で抽出された語幹に係る「主格」「与格」「対象格」を抽出して名詞句を生成する
4. 定型表現の場合はさらに、変換規則を利用して名詞句を生成する

4では、文末の表現に着目している。その表記が「明らかにする」「考えを示す」「判る」「述べる」「話す」の場合に以下のような変換規則を適用した。

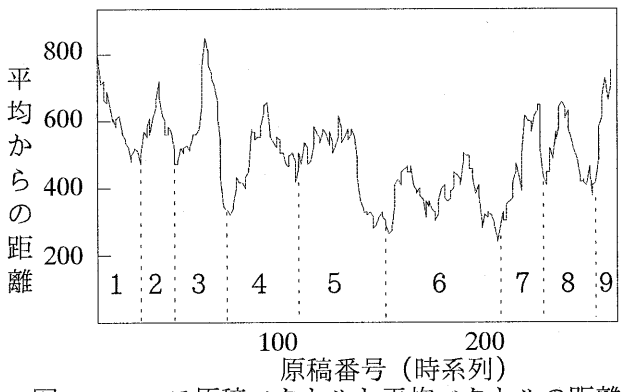

図2. ニュース原稿ベクトルと平均ベクトルの距離

変換規則例：

「主体」が「明らかにする」→「主体」の「表明」

今回、実験で利用した話題「中央アジアのキルギス」を構成する283原稿のうち、74原稿(26%)の第1文がこの文末表現に該当しており、ニュース原稿には定型表現が多いことが判る。

話題「中央アジアのキルギス」の構成要素を表す名詞句生成結果を表2に示す。事件後の外務省の対応から、武力勢力との交渉、人質の解放、帰国と、一応の話題構成要素は抽出されている。表中の話題タイトルは、前記の手順1の処理で対象から除かれた名詞句の中で最も重心に近いものを抽出している。

【話題タイトル】

中央アジアのキルギスで日本人の鉱山技師四人を含む七人が武装勢力に拉致されている事件

【話題の構成要素】

1. 外務省の川島事務次官の会見:外務省にオペレーションルームを設置 [99/8/23]
2. キルギス大統領府の当局者の表明:現地の住民の目撃 [99/8/28]
3. ウズベキスタンの反政府組織の通告 [99/9/4]
4. 人権活動家のアクノフ氏の表明:武力勢力側の合意 [99/9/15]
5. 武装勢力の最高幹部アフガニスタンに潜伏 [99/9/30]
6. 小渕総理大臣の会見 [99/10/17]
7. 加藤重信本部長代行の会見:日本人四人と通訳の無事解放[99/10/25]
8. キルギス政府のジャヌザコフ安全保障会議書記の会見 [99/10/25]
9. 四人の日本人技師成田空港に到着 [99/10/26]

※下線部は手順4の変換規則により生成

表2. 話題の構成要素となる名詞句生成結果

3. まとめ

本報告では、長期にわたり出現する話題のトラッキングを行い、話題の変化点を抽出し、さらに話題の構成要素を表す名詞句を生成する手法の検討を行った。統計的な手法のみでも、話題の変化点の一つの手掛かりとなる。今後は、本手法の評価を行うと共に、単語の意味を考慮した処理を追加して、より的確な構成要素の抽出を目指す。

【参考文献】

[1] Yamada, Kim, Shibata, Uratani: Topic Event Detection using Japanese News articles, In Proc. 5th NLPRS, pp.375-380(1999)