パターンと重要語に基づく関連記事からの話題抽出

池田成宏　松尾義博　林良彦
日本電信電話株式会社　NTTサイバースペース研究所
{ikeda.naruhiro, matsuo.yoshihiro, hayashi.yoshihiko}@lab.ntt.co.jp

1 はじめに

ネットワークへの高速なアクセス手段の普及に伴い、映像・音声を含むマルチメディアコンテンツの流通が活発になりつつある。マルチメディアコンテンツが増大すれば、マルチメディアコンテンツについても Web の検索と同様に内容に基づく検索が必要となる。そのため、我々は映像コンテンツ中の音声を音声認識[1]し、メタデータ化してからインデックスを作成し、内容に基づく検索を可能とする音声インデクシングシステム[2]の開発を行っており、現在はニュース映像を対象としている。

ニュース検索の場合、一般に検索キーワードに対して複数の話題の記事が検索される。このような場合、単に検索結果を提示するのではなく、検索結果を話題毎に分類し、話題を付与して検索結果を提示した方が、ユーザは所望のニュースを見つけやすい。

これまでにニュース記事の話題を抽出する方法として、ニュース記事をクラスタリングし、クラスタの代表記事から単語重要度に基づいて記事中の名詞句を話題として抽出する方法[3]や、係り受け解析を行って特徴的な係り受け関係を抽出して話題を生成する方法[4]が提案されている。

本稿では、同じ話題に関する記事を関連記事とみなし、複数の関連記事の中から話題を述べている箇所を抽出する話題抽出方法を提案する。提案手法では、話題の候補となりえる単語列を定義した話題パターンにマッチする話題候補を各記事から抽出し、話題候補中の単語の重要度などに基づいて各話題候補のスコアを計算して、最大スコアの話題候補を話題とする。本手法では、話題パターンを自由に設定できるため、話題として抽出する表現を柔軟に設定することができる。

なお、本稿では記事群からの関連記事抽出は対象としていない。

2 関連記事からの話題抽出

図1は本稿で提案する関連記事からの話題抽出方法の概要を示している。関連記事が与えられると、まず各記事から話題パターンにマッチする単語列を話題候補として抽出する。次に、話題パターンに設定されているスコアと話題候補中の単語の tfidf 値などを用いて各話題候補のスコアを計算し、スコアが最大となる話題候補を関連記事の話題として選択する。

2.1 話題候補抽出

まず、各記事の中から話題パターンにマッチする単語列を話題候補として抽出する。話題パターンは表1のような正規表現の表記方法を取り入れた表記で表される。下線つきの単語にマッチする単語列が話題候補として抽出される。

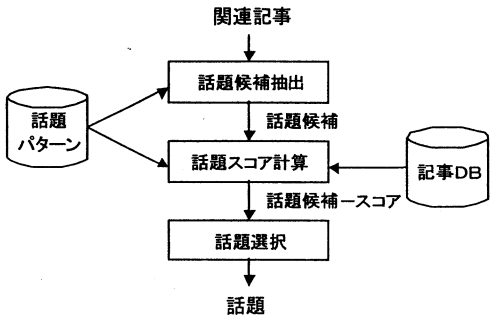

図1. 関連記事からの話題抽出方法

表1. 話題パターンの例

話題パターン1	パターン	(^ | [:^(名詞 | 接尾辞):]) ([:接頭辞:]*[:名詞:]+[:接尾辞:]*)+ ([:助詞:]*[の:助詞:] [:接頭辞:]*[:名詞:]+[:接尾辞:]*)* ($ | [:^(接頭辞 | 名詞):])
	例	日米半導体協定, 市民への情報公開
	スコア	1.0
話題パターン2	パターン	(^ | [:^(名詞 | 接尾辞):]) ([:接頭辞:]*[:名詞:]+[:接尾辞:]*)+ ([:助詞:]*[の:助詞:] [:接頭辞:]*[:名詞:]+[:接尾辞:]*)* [(事故 | 問題 | 事件):名詞:] ($ | [:^(接頭辞 | 名詞):])
	例	原発事故, NTTの分割問題
	スコア	1.5
話題パターン3	パターン	(^ | [:^(名詞 | 接尾辞):]) ([:接頭辞:]*[:名詞:]+[:接尾辞:]* [:助詞:]*)+ ([:動詞:]*[:動詞/連体形:] | [:動詞:]+[:助動詞:]*[:助動詞/連体形:]) [:接頭辞:]*[:名詞:]+[:接尾辞:]* ($ | [:^(接頭辞 | 名詞):])
	例	偽表示の食品が販売されていた問題
	スコア	0.6

※【表記:品詞:標準形】で単語を表し、省略された要素は何にでもマッチする。また、^、$はそれぞれ文頭、文末を表す。ただし、^(…)は括弧内以外のものを表す。なお、下線つきの単語が話題候補として抽出される。

話題パターン1は、連続する名詞・接辞か、名詞・接辞が（助詞＋）「の」で連結された表現を抽出する。例えば、「日米半導体協定」や「市民への情報公開」が話題パターン1にマッチする。なお、パターン1の先頭、末尾の下線なしの部分は、連続する名詞・接辞を抽出するためのものである。

一方、話題パターン2は話題パターン1と同様な表現にマッチするが、末尾の単語が「事故」、「問題」、「事件」のいずれかの表現に限られる。これらの単語を末尾に含む表現は話題になりやすいと考えられるので、スコアを話題パターン1よりも大きい1.5としている。このように、本手法では話題らしいと思われる表現に対してバイアスをかけられるようになっている。

話題パターン3は格要素を含む動詞連体修飾句とそれに修飾されている名詞にマッチし、「偽表示の食品が販売されていた問題」などを抽出することができる。なお、話題パターン3は話題パターン1、2に比べて不要な修飾語句を含む話題候補が抽出される可能性が高いため、話題パターンのスコアを小さくしている。

以上に述べた話題パターン1～3によって記事から話題候補を抽出した例を図2に示す。話題パターン1によって実線の語句、話題パターン2によって破線の語句、そして話題パターン3によって長破線の語句が話題候補として抽出される。例えば、話題パターン1により「ＮＴＴ」や「ＮＴＴの分割問題」が抽出され、話題パターン2で「ＮＴＴの分割問題」が抽出されている。このように複数の話題パターンにマッチする表現は、それぞれ異なる話題候補として抽出される。そして、話題パターン3により「ＮＴＴの分割問題の決着について検討する与党三党」などが抽出されている。

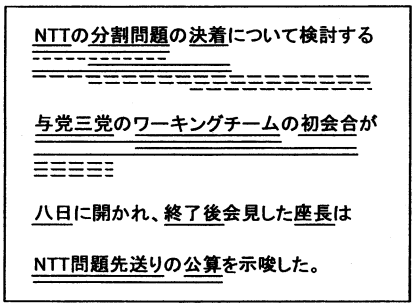

図2. 話題候補抽出の例

2.2 話題スコア計算

次に、話題候補のスコアの計算方法について説明する。話題スコアは、話題パターンのスコアと話題候補の重要度を基に計算する。関連記事群D中の記事dから話題パターンpにより抽出された話題候補tの話題スコア$ts(t,d,p)$は以下のように計算する。

$$ts(t,d,p) = ps(p) \times \sum_{e \in D} \left(sim(d,e) \times s(t,e) \right) \quad (1)$$

ここで、$ps(p)$は話題パターンpのスコア、$sim(d,e)$は記事dと記事eの類似度、そして

$s(t,e)$は記事eにおける話題候補tの重要度を表している。また、話題スコア$ts(t,d,p)$は、記事eにおける話題候補tの重要度を記事dと記事eの類似度の重みつきで加算し、さらに話題パターンのスコアで重みづけした値となっている。

(1)式中の記事dと記事eの類似度$sim(d,e)$、記事eにおける話題候補tの重要度$s(t,e)$は次のように計算する。

$$sim(d,e)=\frac{\sum_{w\in d} tf(w,d)\times tf(w,e)}{(\sum_{w\in d} tf(w,d)^2)^{1/2}\times(\sum_{w\in e} tf(w,e)^2)^{1/2}} \quad (2)$$

$$s(t,e)=\sum_{w\in t}\{tfidf(w,e)-\alpha\times c(w,e)\} \quad (3)$$

$$c(w,e)=\begin{cases}0 & (tf(w,e)>0\text{の場合})\\ idf(w) & (tf(w,e)=0\text{の場合})\end{cases} \quad (4)$$

ここで、αは定数、$c(w,e)$は記事e中に単語wを含まない場合のペナルティである。ペナルティがなければ、語数が多い話題候補ほど話題スコアが大きくなってしまい、不要な語句が付いた話題候補が話題となってしまう。そのため、他記事には出現しない単語に対してペナルティを課し、不要な語句を含む話題候補のスコアが低くなるようにしている。

以上の(1)−(4)式により全話題候補の話題スコアを計算し、話題スコアが最大となる話題候補を関連記事群の話題とみなす。

なお、複数の記事から話題候補を抽出すると、各記事から同じ話題候補が抽出されるが、話題スコアの計算式では話題候補が抽出された記事が異なれば話題スコアの値が変わるため、異なる話題候補として話題スコアを計算する。

3 実験

提案手法を新聞記事に対して適用し、関連記事からの話題抽出を試みた。毎日新聞1996年3、4月の記事のうち、掲載面が1面、2面、3面、国際、経済、社会の記事を用いて実験を行った。

3.1 実験1　～パラメータ決定～

まず、話題スコア計算式(3)中のパラメータαによる話題抽出精度の変化を調べた。

毎日新聞1996年3月分の2,277記事から人手で34グループの関連記事(平均11.6記事)を抽出し、さらに各関連記事から人手で話題を抽出した。そして、これらを参考にして、表1の例のように「問題」、「事件」など詳細に話題候補を規定するほどスコアが大きくなるような話題パターンを6個作成し、話題抽出を行った。ただし、話題候補を記事の全文から抽出すると話題候補が非常に多くなり、話題抽出・スコア計算に時間がかかるため、記事の見出しと第1文のみを対象とした。

表2にパラメータαを変化させたときの話題抽出精度を示す。抽出された話題が人手で抽出された話題と同内容であれば「正解(A)」、抽出された話題に数語の過不足はあるが話題を推定可能な場合に「ほぼ正解(B)」、そして抽出された話題が人手で抽出された話題からかけはなれている場合を「不正解(C)」としている。

表2によると、$\alpha=0.10$のときに正解(A)が79.4%で最も精度が高くなっていることがわかる。パラメータαは話題候補中の単語が他記事に含まれない場合のペナルティの係数である。そのため、αが大きければペナルティによる減点の影響が大きくなり、短い話題候補の話題スコアが大きくなりやすい。一方、αが小さければ、ペナルティによる減点の影響は少ない。そのため、語数が多い話題候補のスコアが大きくなりやすく、不要

表2. パラメータαによる話題抽出精度[%]の変化

α	0.025	0.05	0.075	0.10	0.125	0.15
正解(A)	73.5	76.5	76.5	79.4	73.5	64.7
ほぼ正解(B)	20.6	17.6	17.6	14.7	17.6	20.6
A+B	94.1	94.1	94.1	94.1	91.2	85.3
不正解(C)	5.9	5.9	5.9	5.6	8.8	14.7

表3. 抽出された話題の例

記事数	評価	記事例
4	A	[見出し] *新たな亀裂も――北海道・駒ケ岳噴火* [本文] 五十四年ぶりに五日夜、噴火した北海道駒ケ岳が、水蒸気噴火を続けていることが六日分かった。札幌管区・・・
3	B	[見出し] *反テロ中東首脳会議に対抗、国民会議を開催――イスラム原理組織* [本文] レバノンからの報道によると、イスラエルとの和平に反対するイスラム原理主義組織は十三日、エジプトの反テロ中東首脳会議に反対して「テロ首脳会議」をベイルートで開催。
7	C	[見出し] *航空3社の幅運賃制導入…値上げ路線に反発――北海道、宮崎など「再考して」* [本文] 日本航空、全日本空輸、日本エアシステムの航空三社が幅運賃制の導入により東京－札幌路線の値上げを・・・
		[見出し] *日航と全日空の新運賃を認可――運輸審議会* [本文] 運輸省の諮問機関、運輸審議会は五日、幅運賃に伴って申請されていた、日本航空と全日本空輸の運賃を・・・

な語句が付加した話題候補が抽出されやすい。このバランスがうまくとれているのが$\alpha = 0.10$のときということになる。

$\alpha = 0.10$のときの話題抽出例を表3に示す。提案手法により抽出された話題を実線で、人手により抽出された話題を点線で示している。

3.2 実験2 ～話題パターンのスコアの効果～

次に、(1)式での話題パターンのスコア$ps(p)$の有効性を調べるために、実験1で作成した話題パターンのスコアを一定（=1.0）にして前節で述べた関連記事群から話題抽出を行った。$\alpha = 0.1$のときの話題抽出精度を表4に示す。表2と表4を比較すると、話題パターンのスコアを一定にしたときには正解率が大きく低下しており、話題パターンのスコアによる効果が確認できる。

表4. 話題パターンのスコアを一定にしたときの話題抽出精度[%]

正解(A)	ほぼ正解(B)	A+B	不正解(C)
55.9	20.6	76.5	23.5

3.3 実験3 ～テストデータによる評価～

最後に、実験1で決定したパラメータで、4月分の記事の話題抽出精度を調べた。4月分の2,386記事から、人手で40グループの関連記事群(平均8.8記事)を抽出した。そして、提案手法により話題を抽出し、3段階で評価を行った。

テストデータに対する評価結果は表5のようになり、提案手法によって65%の精度で正解の話題が得られ、数語の誤りを許容すれば82.5%の精度で話題を抽出できることがわかる。話題抽出精度を上げるには話題パターンの改良・調整が必要と考えられる。

表5. 話題抽出精度[%]

正解(A)	ほぼ正解(B)	A+B	不正解(C)
65.0	17.5	82.5	17.5

4 まとめ

本稿では、話題パターンで定義される話題候補を複数の関連記事から抽出し、話題パターンに設定されたスコアと単語の重要度から話題候補のスコアを計算して、スコアが最大となる話題候補を話題として抽出する手法を提案した。実験により、提案手法によって比較的少数の記事からでも話題を抽出できることを確認した。

今後は、話題抽出精度の向上に向けた改良と、関連記事検索手法との組み合わせによる記事の自動分類・話題付与に取り組む予定である。

5 参考文献

[1]野田他：音声認識エンジンVoiceRexの開発，日本音響学会1999年秋季研究発表会，2-1-19，pp. 91-92，1999.

[2]林他：映像コンテンツのインデクシングのための音声・言語処理，情報処理学会全国大会，2003.

[3]山田他：ニュース記事を利用したトピック抽出の検討，言語処理学会 第5回年次大会，pp. 116-119，1999.

[4]山田他：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～，言語処理学会第7回年次大会，pp. 297-300，2001.