ニュースからのトピックに関するストーリーラインの生成

谷口 祐太郎　　小林 哲則　　林 良彦
早稲田大学 理工学術院
y-taniguchi@suou.waseda.jp

1 はじめに

ニュースコーパスからトピック (知りたい事柄) に関連するテーマを抽出し，そのテーマに関連する文が時系列順に並んだ文集合 (ストーリーライン) を出力するシステムを提案する．

ある事柄についての情報を取得しようとしたとき，これに関するニュース記事群をそのまま読むと，事柄についての様々なテーマの情報が混在しており，量も多い．一方で,「Wikipedia」や「NAVER まとめ」などの Web サービスを利用した場合，編集者の主観が含まれる情報を見ることになるという問題がある．したがって，知りたい事柄について，ニュースから体系的かつ効率的に無加工の情報を取得する機能が望まれる．

ニュースからの体系的・効率的な情報取得の手法としては，ニュース検索結果を記事単位でクラスタリング [2][3] した結果を見ることが挙げられる．しかし，適切な文集合が提供されるなら，この方が記事集合を見るよりも効率的だと考える．

ニュースを体系化するために，ニュース記事集合全体からテーマを抽出し，テーマごとの記事集合を図示するという手法 [1][4] もあるが，本研究では，ユーザが指定するトピックを入力とすることで，ユーザの知りたい事柄を出力に反映させる．

複数の文からなる文書に対して，あるキーワードに基づいて重要文抽出を行う [5] 試みはあるが，出力される文集合は 1 つのみである．知りたい事柄についての情報の中には様々なテーマがあるため，テーマを自動で抽出し，文集合を複数出力することが望ましい．

以上より，ユーザが指定するトピックに応じた記事集合からテーマに沿った文集合を複数提示することは有用である．

2 提案システム

提案システムを図 1，図 2 に示す．ここでは，トピックとして「オバマ」が与えられた場合を例としている．

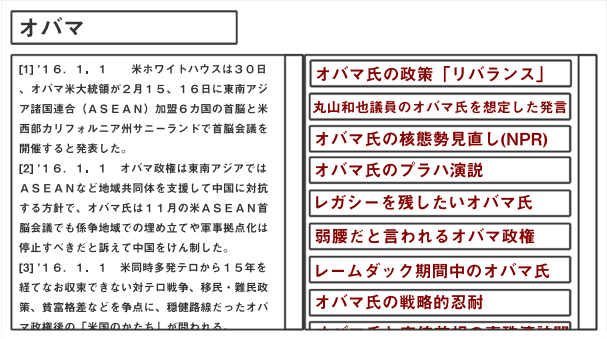

図 1: 提案システム (階層 1)

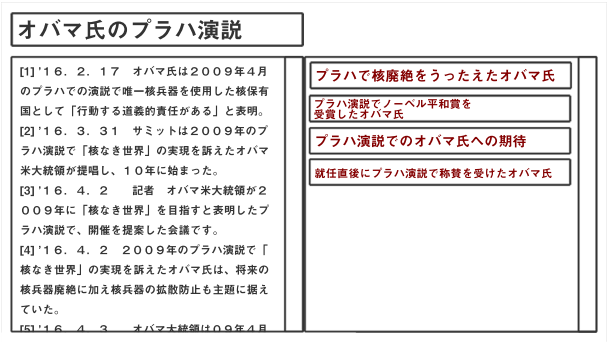

図 2: 提案システム (階層 2)

以下に，想定する使用手順を示す．

1. 初期クエリとして「オバマ」を与える．(図 1)
2. 左部分に，ニュースコーパスから検索された,「オバマ」に関連する文が時系列順に表示される．右部分に,「オバマ」に関連する**文集合から取得されたテーマが表示**され，**各テーマにはタイトルがつけられている**．(図 1) 本稿では，このタイトルを「テーマタイトル」を呼ぶ．
3. 右部分のテーマタイトルのうち一つ，例えば「オバマ氏のプラハ演説」を選択すると，左部分には，前段階から抽出された「オバマ」に関連する文集合からさらに抽出された,「オバマ氏のプラハ演説」に**関連する文**が，右部分にはその文集合からさらに取得されたテーマが表示される．(図 2)

4. ユーザは，興味のあるテーマを選択し，そのテーマに関連する文集合を読むことで，トピック（= 初期クエリ）に関する様々なテーマに基づいた情報を取得することができる．

上の使用手順で太字で示した部分から考えると，「1. 文集合からのテーマ取得」，「2. ユーザがテーマを選択するきっかけとなるテーマタイトル生成」，「3. 前の文集合からの適切な文集合抽出」が，この提案システムの必要事項である．

3 テーマ取得・文集合抽出

3.1 テーマ取得・文集合抽出の手法

2 章で示した必要事項のうち，「1. テーマ取得」，「3. 文集合抽出」を実現するための手法を提案する．提案手法を図 3 に示す．重要語を取得することで，適切なテーマ取得を行い，さらにその重要語を検索式に追加することで，適切な文集合を抽出するという狙いである．文の集合をもとに，重要語をつなげた検索式をユーザに提示する研究としては，松生ら [6] の研究が挙げられる．ただし，これは Web ページを対象として，ユーザが自身の検索する目的を明らかにするためにキーワード式を提示するという動機である．それに対して本研究は，ユーザが知りたい事柄に対して，様々な側面から情報を取得するということを狙っている．

本手法は，ユーザのテーマ選択に応じて文集合が生成されるのではなく，あらかじめ一定階層数 d までの文集合を生成しておいて，ユーザの選択に応じてその文集合を表示するという想定である．

擬似コード 1 で示したアルゴリズムによって，検索式とこの検索式により得られる文集合のタプルの集合を生成する．検索式は，初期クエリを含むキーワードの集合で，各要素は AND 結合される．

文集合からの重要語の抽出方法　以下の式 (1) で重要度を計算し，上位 N 語を重要語とする．

$$R_w = \frac{c_w}{c_{all}} \cdot \frac{C_{all}}{C_w} \tag{1}$$

ただし，式 (1) の R_w は，単語 w の重要度（文集合における w の出現頻度を，コーパス全体における w の出現頻度で正規化したもの），c_w は文集合における w の出現回数，c_{all} は文集合における全単語の出現回数，C_w はコーパス全体における w の出現回数，C_{all} はコーパス全体における全単語の出現回数を示す．

c_w，c_{all} において，文集合の検索式に含まれる任意のキーワード（初期クエリ含む）に対して，距離（原文における単語間の距離）が 4 以上，かつ，間に助詞が含まれる，という場合のみ単語をカウントする．

Algorithm 1 テーマに沿った複数の文集合の抽出方法

Input: ニュース文集合 $\mathcal{S}$，初期クエリ q，階層数 d，重要語数 N，閾値 t(最少文数)

$First \leftarrow \{s | s$ は $\mathcal{S}$ のうち q を含む文 $\}$
$OutSets \leftarrow \phi$
$NewSets \leftarrow \{(q, First)\}$
$OutSets \leftarrow OutSets \bigcup NewSet$
for from 0 to d-1 **do**
  $PreviousSets \leftarrow NewSets$
  $NewSets \leftarrow \phi$
  for all set in $PreviousSets$ **do**
    $Temp \leftarrow Algorithm2(set[0], set[1], N, t)$
    $NewSets \leftarrow NewSets \bigcup Temp$
  end for
  $OutSets \leftarrow OutSets \bigcup NewSets$
end for

Output: $OutSets$

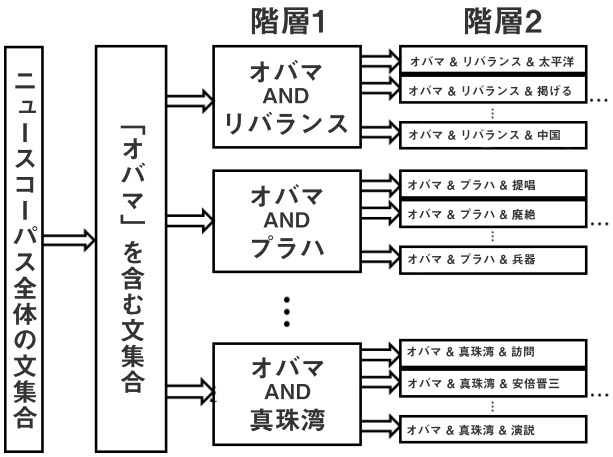

図 3: 文集合抽出の提案手法 (例:オバマ)

3.2 文集合抽出の評価手法

2 章で示した必要事項のうち，「3. 文集合抽出」の手法を評価する．まず，それぞれの文集合について正解データを作成し，評価対象の手法で抽出したときの文集合を，その正解データと比較することで，テーマに関連した文を前段階の文集合から適切に抽出できているかどうかを評価する．各文集合の評価結果を平均することで，文集合抽出の手法自体を評価する．

以下に，文集合の正解データ作成手法と，評価手法を示す．ただし，評価する文集合を S_{eval}，その前段階

Algorithm 2 検索式と文集合のセット (q, S) からのテーマに沿った複数の文集合の抽出方法

Input: 検索式 Q，文集合 S，重要語数 N，閾値 t(最少文数)

$OutSets \leftarrow \phi$
$K \leftarrow \{k|k$ は S に含まれる単語のうち 3.1節で示した手法で重要語抽出したときの上位 N 語 $\}$
for all k in K **do**
  $Temp \leftarrow \{s|s$ は S のうち k を含む文 $\}$
  if $|Temp| \geq t$ **then**
    $NewQ \leftarrow Q \bigcup k$
    $TempSet \leftarrow (NewQ, Temp)$
    $OutSets \leftarrow OutSets \bigcup TempSet$
  end if
end for

Output: $OutSets$

の文集合を S_{prev}，システムが出力する S_{eval} を S_{evalo}，S_{eval} の正解を S_{eval_a} とする．

文集合の正解データ作成手法　はじめに，S_{evalo} に，検索式に含まれる，初期クエリと追加クエリをすべて使用したタイトルをつける．これは，2 章で示した 3 点の必要事項のうち，本研究では着手していない (2) の処理が行われたことを想定している．このタイトルを $T(S_{evalo})$ とする．例えば，検索式が「オバマ AND 所感 AND 広島」の場合,「オバマ氏が広島で述べた所感」といったタイトルが考えられる．次に，S_{prev} から，$T(S_{evalo})$ に関連し，かつ文単体で意味をなすような文のみを人手で抽出し，これを S_{eval_a} とする．

文集合の評価手法　S_{evalo} と S_{eval_a} と比較して，Precision，Recall，F 値を求める．本稿における評価条件を表 1 に示す．

3.3　文集合抽出の評価の予備実験

3.2 節で示した評価手法の有用性を確かめるため，評価の一致率を検証する予備実験を行った．表 1 で示した 100 文集合からさらに 5 文集合を選定し，これについて著者と 3 人の評価者で評価した．任意の 2 人の間の κ 係数を表 2 に，それぞれの評価者が評価したときの各文集合の Precision，Recall，F 値の平均値を表 3 に示す．表 2 に示した κ 係数によれば，評価者間の一致は低い．ただし，表 3 に示した Precision，Recall，F 値による評価結果によれば，評価結果自体のぶれはさほど大きくない．この結果はさらなる検証を要するが，提案する評価手法は，文集合抽出を評価するのに有用であり，評価者は 1 人でも十分であると考える．

E U AND 割り当て AND 反対

「ＥＵへの難民割り当て反対」
ＥＵはギリシャ、イタリアに集まった難民１６万人の割り当てを始めたが、ハンガリーは割り当てを拒絶している。

図 4: 類義語が使用されている例

マイナス金利 AND 預金 AND 保有

「個人預金にマイナス金利がつくことによる，預金を現金で保有する人の増加」
だが、第一生命経済研究所の熊野英生首席エコノミストは「マイナス金利政策が『自分の預金がどうなるかわからない』との不安心理に結びついているとしてもおかしくない」と指摘。

図 5: 人手の推測を要する例

3.4　文集合抽出の評価結果

3.3 節の予備実験により，1 人による評価でも十分であることを示したので，3.2 節で示した評価手法，表 1 で示した評価条件 (100 文集合) で，著者 1 人による評価を行った．文集合抽出の評価結果を表 4 に示す．一定の Precision が得られる一方で，Recall は低い．これは，抽出するべき文が抽出できない False Negative になる例が多いことを表す．この例を 3.5 節で示す．

3.5　文集合抽出の False Negative の例

類義語が使用されている例 (図 4)　「ＥＵへの難民割り当て反対」に関連する文であるから抽出すべきなのだが,「反対」ではなく,「拒絶」が使用されているため，本手法では抽出できなかった．

人手の推測を要する例 (図 5)　自分の預金がどうなるかわからないと不安，つまり預金を銀行に預けるのは不安であるから，預金を現金で保有する人が増える，と評価者が推測したため，抽出すべきだとみなされたが，本手法では抽出できない．

追加された語がそれまでの語を補足する例 (図 6)　追加された重要語「北朝鮮」は,「オバマ氏の戦略的忍耐」

表 1: 文集合抽出の評価条件

使用コーパス	「CD-毎日新聞 2016 データ集」
評価対象初期クエリ	10 個 (「オバマ」「小池百合子」「琴奨菊」「マイナス金利」「増税」「北朝鮮」「アベノミクス」「ポケモンＧＯ」「東芝」「ＥＵ」)
評価対象文集合	100 個 (10 個の初期クエリに対して,「オバマ AND ○○ AND ××」といったように，追加クエリが 2 個である文集合，すなわち 2 階層目の文集合で，5 文以上含まれる文集合を各 10 個選定した．)

表 2: 各評価者のうち任意の 2 人の間の κ 係数

	評価者 1	評価者 2	評価者 3
著者	0.15	0.29	0.40
評価者 1	-	0.33	0.25
評価者 2	-	-	0.49

表 3: 各評価者による 5 文集合の評価結果

評価者	Precision	Recall	F 値
著者	0.81	0.55	0.66
評価者 1	0.83	0.54	0.66
評価者 2	0.88	0.51	0.64
評価者 3	0.87	0.59	0.70

を単に補足している．文中には「北朝鮮に対する」という情報が含まれていなくても，ただ「戦略的忍耐」といえば北朝鮮に対する戦略的忍耐であることがわかるため，抽出すべきであるが,「北朝鮮」の語が含まれていないので本手法では抽出されない．

4 まとめ・今後の計画

知りたい事柄 (トピック) を与えるとニュースコーパスからこのトピックに関連するテーマを抽出し，関連する文集合を提示するシステムを提案した．これを達成するための必要事項 (「テーマ取得」「タイトル生成」「文集合抽出」) を提示し，そのうちの「テーマ取得」「文集合抽出」について，段階的に重要語を検索式に追加する手法を試みた．この手法によって適切な文集合抽出が行われたかどうかについての評価手法を提案し，複数人評価による予備実験により評価手法の有用性を示した．小規模な文集合に対する評価を行い，適切な文集合が一定の Precision で得られることを示し，Recall を改善する課題を明確化した．

今後は，本稿では着手していない「タイトル生成」の手法を提案し，システム全体を実装する．システム全体の評価を評価手法の提案を含めて実施する．また，文集合抽出の Recall を向上させる．

表 4: 文集合抽出の評価結果

	Precision	Recall	F 値
100 文集合平均	0.898	0.669	0.737

オバマ AND 忍耐 AND 北朝鮮

「オバマ政権の北朝鮮に対する戦略的忍耐」

まずは米オバマ政権の、無策とも映る「戦略的忍耐」の転換を促すこと。

図 6: 追加された語がそれまでの語を補足する例

参考文献

[1] Philippe Laban，Marti A. Hearst，”newsLens: building and visualizing long-ranging news stories”，Proceedings of the Events and Stories in the News Workshop，pp. 1–9，2017.

[2] Mingjie Qian，Chengxiang Zhai，”Unsupervised Feature Selection for Multi-View Clustering on Text-Image Web News Data”，Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management，pp. 1963–1966，2014.

[3] Srinivas Vadrevu，*et al*，”Scalable Clustering of News Search Results”，Proceedings of the fourth ACM international conference on Web search and data mining，pp.675–684，2011.

[4] Zheng Xu，*et al*，”Knowle: A semantic link network based system for organizing large scale online news events”，Future Generation Computer Systems Vol.43，pp.40–50，2015.

[5] 砂山渡，谷内田正彦，”観点に基づいて重要文を抽出する展望台システムとそのサーチエンジンへの実装”，人工知能学会論文誌，vol.17，pp.14–22，2002.

[6] 松生泰典，*et al*，”検索結果の概要を表すキーワード式生成による質問修正支援”，電子情報通信学会データ工学ワークショップ（DEW2005），1C–i9，2005.