トピックを構成する下位トピックの注目度に着目した複数発信者意図比較システム

青木　伸也† 湯本　高行†† 角谷　和俊††† 新居　学†† 高橋　豊††

† 兵庫県立大学大学院工学研究科　〒 671-2201 兵庫県姫路市書写 2167
†† 兵庫県立大学大学院工学研究科　〒 671-2201 兵庫県姫路市書写 2167
††† 兵庫県立大学環境人間学部　〒 670-0092 兵庫県姫路市新在家本町 1-1-12
E-mail: †er08h001@steng.u-hyogo.ac.jp, ††{yumoto,nii,takahasi}@eng.u-hyogo.ac.jp,
†††sumiya@shse.u-hyogo.ac.jp

あらまし　同じトピックに関するニュース記事でも発信者によって言及する事柄（下位トピック）に違いがある．これは発信者によって注目する下位トピックが異なるからであり，どの下位トピックについて書き，どの下位トピックについて書かないかは発信者の意図で異なる．ユーザは発信者の意図を知ることにより，例えば，ある情報に偏りがあるかなどの判断が可能になると考えられる．そこで，本論文では，トピックに関する記事全体から下位トピックを抽出し，同じトピックについて記事を書いた複数の発信者と意図を比較できるように提示するシステムを提案する．システムでは選択したトピックについて各下位トピックごとに複数の発信者の注目度を比較することができる．実験では，下位トピックへの記事の分類精度を確認する実験を行い，全体で 0.6 以上という精度を得た．

キーワード　発信者意図，ニュース分析，下位トピック

System to Compare Publishers' Intention Focused on Attention of Subordinate Topic Composing Topic

Shinya AOKI†, Takayuki YUMOTO††, Kazutoshi SUMIYA†††, Manabu NII††, and Yutaka TAKAHASHI††

† Graduate School of Engineering, University of Hyogo　2167 Shosha, Himeji, Hyogo, 671-2280, Japan
†† Graduate School of Engineering, University of Hyogo　2167 Shosha, Himeji, Hyogo, 671-2280, Japan
††† School of Human Science and Environment, University of Hyogo　1-1-12 Shinzaike-honcho, Himeji, Hyogo, 670-0092, Japan
E-mail: †er08h001@steng.u-hyogo.ac.jp, ††{yumoto,nii,takahasi}@eng.u-hyogo.ac.jp,
†††sumiya@shse.u-hyogo.ac.jp

Abstract　When the several authors report the same news topic, which facts are reported is often different by the authors. We call these facts subordinate topics. Each author has his own intention about the news topic, and this intention causes the differences that each subordinate topic is focused or not. If users understand the author's intention before they read his article, they can judge whether his article is biased or not. In this paper, we propose the system to extract subordinate topics of a given topic from news articles and compare intentions of multiple authors. By using the proposed system, we can compare degree of notice about subordinate topic of the multiple authors for each topic. We evaluated precision of classifying new articles into subordinate topics, and the average precision was more than 0.6.

Key words　Author's intention, News analysis, Subordinate topic

1. はじめに

世の中での重要な出来事は新聞やオンラインニュースなどニュース記事として伝えられるが，これらニュース記事の中から同じトピックに関する記事を集めても，記事ごとに言及されている内容は異なる．これは，記事ごとに新聞社など，発信者が違うからであり，発信者ごとにその意図が異なるからである．発信者の意図を知らないで記事を読んだ場合，読み手はそのトピックに対して記事の内容通りのイメージを持つと考えられる．記事の内容がそのトピックに関して偏ったものであった場合，そのトピックに関する読み手のイメージは公平なものではなくなる恐れがある．

例えば，ある殺人事件に関するニュース記事がある場合，“犯人” や “被害者”，“使用された凶器” など複数の注目すべき事柄があると考えられる．犯人に注目した発信者の記事を読んだ場合，犯罪を犯す人物のイメージが印象に残ったり，凶器に注目した発信者の記事を読んだ場合，犯罪に使用される道具としてのイメージが印象に残ったりする．

このような問題は，発信者の意図を知ることができれば解決できると考えられる．なぜなら，発信者が読み手にどのように伝えたいかということを知ることで，読み手は記事を読む前に心構えができるからである．

本論文ではトピックを分割する事柄を “下位トピック” と呼ぶ．下位トピックに関する発信者の注目度である “下位トピック注目度” をそれぞれの下位トピックに対して求めることで発信者の意図を表現することができると考えている．ある発信者に関して各下位トピック注目度を求めるには，そのトピックに関する，その発信者の文書が多数必要であると考えられる．その発信者の文書集合全体をみることで，発信者の各下位トピック注目度が分かる．また，発信者の意図を示すには，他の複数の発信者との比較が必要であると考えられる．1 つの発信者の文書のみからではトピックの全体像が分からず，下位トピックに分割することができないからである．

そこで本論文では，複数発信者の各下位トピックに対する下位トピック注目度を提示することで発信者の意図を比較することのできるシステムを提案する．本提案システムを使用することで，ユーザは各発信者がどの下位トピックに注目しているかを知ることができる．また，異なる意図をもった他の発信者を選択して，同じトピックに関する記事を読むことができる．本提案システムを利用することで，ユーザはそのトピックに関して公平な考えを持つことができると考えている．

以下，2 節では本提案システムの概要と関連研究について述べ，3 節で提案する発信者意図の抽出手法を述べる．実験について 4 節で述べ，最後に 5 節でまとめる．

2. 発信者意図比較システムの概要と関連研究

2.1 提案システムの概要

本提案システムはあるトピックに関して複数の発信者の発信者意図を比較できるように提示するものである．システムが提示する情報は以下の通りである．

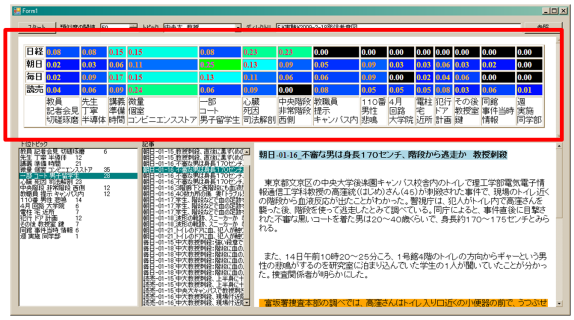

図 1　発信者意図比較システム
Fig. 1　System to compare author intentions

- 分割した各下位トピックを表す下位トピックキーワード
- 各発信者の各下位トピックに対する下位トピック注目度
- 各下位トピックに分類されたニュース記事
- 各発信者のニュース記事
- 各ニュース記事がどの下位トピックに分類されたかを示すラベル

これらをの情報を提示することで，発信者意図の比較が可能となるだけでなく，任意の下位トピックを選択して，その下位トピックに分類されたニュース記事を閲覧することも容易となる．下位トピック注目度だけではなく各下位トピックに分類されたニュース記事を閲覧することでより適切な判断が可能と考えている．

本提案システムでは，あるトピックに関するニュース記事集合をクラスタリングすることにより，複数の下位トピックに関するニュース記事集合へ分割し，下位トピック注目度を計算する．実際の処理では，各ニュース記事を段落で区切り，段落集合としてからクラスタリングを行う．そのため，段落単位で各下位トピックへ分類されることになり，ニュース記事自体は複数の下位トピックへ属することになる．システムでは段落を別々に分けて提示するとニュース記事の閲覧が困難になると考え，記事単位で提示する．その際，ニュース記事の各段落がどの下位トピックに属するかという情報も提示する．

システムの実行例を図 1 に示す．システム上部の表が各発信者の各下位トピックについての下位トピック注目度を示したものである．各セルが同じ行の発信者の同じ列の下位トピックに対する下位トピック注目度を示している．下位トピック注目度の大きさによってセルの色を変えることで比較しやすくしている．

2.2 下位トピック

我々はトピックは階層化されているものと考え，下位トピックはトピックを 1 つ下の階層で分割するトピックであると考える．そのため，下位トピックそれ自体をトピックとし，さらに下位トピックに分割することが可能であると考えている．図 2 にトピックと下位トピックの関係を示す．

実際の処理では，ニュース記事の各段落の集合である段落集合をクラスタリングすることにより下位トピックへの分割を行う．すなわち，クラスタリング結果の各クラスタを下位トピックとし，各クラスタに属する段落の集合を対応する下位トピッ

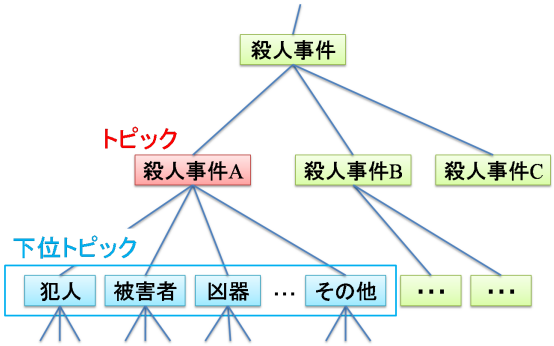

図 2　トピックと下位トピックの関係
Fig. 2　Relation between topic and subordinate topic

クに属する段落の集合とする．以降では下位トピックの要素は段落であるとして説明する．

下位トピックには次のような特徴があると考えられる．

- 下位トピック数は多すぎない．下位トピックはトピックを 1 つの下の階層で分割するものであるので，その数が多すぎることはない．また，下位トピック数が多くなると発信者意図の比較が困難となるため，システムを利用するユーザにとっても望ましくないと考えられる．できるだけ下位トピック数は少ないほうが適切である．
- 下位トピックはトピックを概ね平等に分割するものである．各下位トピック間で含まれる段落数が大きく異なるのは不適切であると考えられる．しかし，すべての下位トピックについての段落数が等しくなる可能性は低いと考えられ，ある程度の差は許容するべきであると考える．
- 下位トピック間で含まれる段落は重複しない．下位トピックはトピックを分割するものであるので，下位トピック間ではトピックは独立していると考えられる．ただし，ニュース記事を単位としたみた場合，異なる下位トピックに分類された段落を持つ可能性があるので，複数の下位トピックに属することもある．
- 下位トピックのみではトピックに含まれる段落をすべて分類できないこともある．トピックにはどの下位トピックにも含まれないが，1 階層下のトピックとするには適切でない段落も含まれていると考えられる．そのため，どの下位トピックにも分類されない記事を分類する下位トピック「その他」が必要であると考えられる．

2.3　関 連 研 究

本論文では，下位トピックに注目して発信者の意図を表現し，発信者間の意図を比較するシステムを提案するが，同一トピックに関するニュース記事間の違いに注目したものとして以下の研究があげられる．

北山らは映像ニュースからテキストニュースを検索する比較ニュース検索方式を提案している [1]．北山らは，異なるメディア間 (映像とテキスト) ではニュースの構成が違うことから，異メディアのニュースは相補的な関係にあると考えているが，本論文でも，発信者意図の異なるニュース記事はお互いに相補的な関係にあるとし，その利用のためには発信者意図を比較提示することが必要だと考えている．

灘本らは，同一トピックに関するニュースでも発信国間で視点が異なることに注目して，B-CWB(Billingual Comparative Web Browser) を提案している [2]．B-CWB は言語の異なる 2 つの類似 Web ページを発見し，同時に比較提示するブラウザで，ユーザは B-CWB を用いて，2 カ国のニュース記事の比較を行うことができる．また，段落を単位として差異情報の抽出手法も提案しており，ユーザは容易に 2 つのニュース間の差異を発見することができる．B-CWB では 2 つのニュース記事間の差異を提示することを目的としているが，本論文では，より多くの記事からでないと発信者の意図やその違いを表現することはできないと考え，発信者意図比較システムを提案する．

吉岡らはニュース発信者の視点に注目し，トピックの全体像を構築する手法を提案している [3]．吉岡らはまず，ニュース記事からトピックの側面と呼ばれる本論文での下位トピックに相当するものを抽出する．トピックの側面はキーワード集合で構成される．複数の記事から抽出した類似するトピックの側面を統合する際，各キーワード集合の和集合を統合されたトピックの側面としているが，この手法ではニュース記事を多くした場合にトピックの側面を構成するキーワードが多くなるとともに，トピックの側面自体も多くなりすぎ，把握しづらくなる問題があると考えられる．本提案システムでは，トピックの記事集合の大きさを考慮し，下位トピックを抽出するので，そのような問題はないと考えている．

3.　手　　法

本提案システムではユーザによって指定されたトピックのニュース記事を自動収集する機能は未実装であるので，あるトピックに関するニュース記事集合を与えられたところからの処理の手順を説明する．与えられるニュース記事は段落わけされており，執筆者名や執筆日時を除いたものとする．

本提案システムでは，次の 4 段階に分けて処理を行う．

（1）　語の抽出
各記事からトピックを下位トピックに分割するために必要な語のみ抽出する．すなわち，トピック全体で多すぎる語や少なすぎる語は除き，中頻度の語のみを抽出する．また，抽出した語のみを用いて各記事の各段落に関する特徴ベクトルを作成する．

（2）　下位トピックへの分割
トピックの下位トピックへの分割は各段落に作成した特徴ベクトルをクラスタリングすることにより行う．クラスタリング結果の 1 つのクラスタが 1 つの下位トピックに相当し，クラスタに属する段落を対応する下位トピックに属する段落とする．

（3）　下位トピック注目度の計算
各発信者の各下位トピックに対する下位トピック注目度を，分類された段落数を基に計算する．

（4）　発信者意図の比較提示
下位トピック注目度を用いて各発信者間の発信者意図を比較しやすいように提示する．

3.1 語の抽出

3.1.1 一般名詞，複合語の抽出

各記事からの語抽出には形態素解析システム MeCab を用いる [4]．MeCab を用いて文章を単語に分割し，その結果から重要でないと考えられる以下の単語を除く名詞を使用する．これを一般名詞と呼ぶことにする．

- 品詞情報に “非自立”，“代名詞”，“数詞” を含む名詞
- ひらがなのみの名詞
- 半角記号

また，名詞が連続して構成する複合語も抽出するため，一般名詞が連続する語をすべて抽出する．ただし，品詞情報に “接尾” を含む名詞が先頭に来る場合は除く．例えば「発信者意図比較システム」という語は「発信者」,「意図」,「比較」,「システム」のように 4 単語に分割されるが,「発信者意図」,「発信者意図比較」のように一般名詞の連続するものをすべて抽出する．この例の場合，4 単語以外に 6 つの複合語を抽出する．

3.1.2 中頻度語の抽出

下位トピックへの分割は段落の特徴ベクトルをクラスタリングすることによって行うため，高頻度語や低頻度語を特徴ベクトルの要素とすると適切にトピック分割が行われないと考えられる．高頻度語は 2 つの段落を区別することに役立たず，低頻度語は 2 つの段落を過剰に区別してしまう．そのため，中頻度語を 3.1.1 節で抽出した語の段落頻度を基に決定し，トピックの下位トピックへの分割に使用する語とする．

ある段落を p，抽出した語集合中のある語を t とし，語 t の段落 p 中での出現数を $TF(p,t)$ とすると，語 t の段落頻度 $PF(t)$(Paragraph Frequency) は次のように表わすことができる．

$$PF(t) = |\{p|p \in P_{all}, TF(p,t) \neq 0\}| \tag{1}$$

ここで，P_{all} はトピック全体の段落集合を表わし，$|A|$ は集合 A の大きさを表す．

以下の式を満足する語を中頻度語として抽出する．

$$\theta_{min} \leqq PF(t) \leqq \theta_{max} \tag{2}$$

θ_{max}，θ_{min} は中頻度となる範囲を決定するための閾値で，あらかじめ設定しておくパラメータである．

以降は，抽出した中頻度語のみを使用し，処理を行う．

3.1.3 段落の特徴ベクトル

抽出した中頻度語を基に各段落の特徴ベクトルを作成する．段落 p の特徴ベクトルを fv_p とすると，特徴ベクトル fv_p の中頻度語 t に対する値 $fv_p(t)$ は次のようにする．

$$fv_p(t) = TF(p,t) \tag{3}$$

ただし，3.1.1 節で抽出した複合語の，出現箇所の重複を解消するため，語 t_i が語 t_j を完全に含んでいる場合の特徴ベクトルの値は次のようにする．

$$fv_p(t_j) = TF(p,t_j) - TF(p,t_i) \tag{4}$$

3.2 トピックの下位トピックへの分割

下位トピックへの分割は，3.1.3 節で作成した各段落の特徴ベクトルをクラスタリングすることによって行う．クラスタリングした結果の各クラスタを各下位トピックに対応させ，各クラスタに代表的なキーワードを下位トピックを表すキーワードとする．

3.2.1 段落のクラスタリング

各段落のベクトルを要素に群平均法でクラスタリングを行う．ベクトル間の距離にはコサイン類似度を用いる．ベクトル $v1$，$v2$ 間のコサイン類似度 $cosine(v1,v2)$ は以下の式で表わされる．

$$cosine(v1,v2) = \frac{v1 \cdot v2}{|v1||v2|} \tag{5}$$

$v1 \cdot v2$ は $v1$ と $v2$ の内積を表し，$|v|$ はベクトル v の大きさを表す．

段落の特徴ベクトルは 3.1.2 節で抽出した中頻度語のみを用いて作成されているため，すべての次元で値が 0 となる特徴ベクトルを持つ段落が存在する可能性がある．式 (5) からも分かるように，すべての次元で値が 0 となるベクトルとの間では適切にコサイン類似度を計算することができず，クラスタリングを行うことができない．そのような段落はどの下位トピックにも属さず,「その他」に属するものとしてクラスタリングを行う前に除外しておく．

群平均法は階層型クラスタリングのアルゴリズムである．階層型クラスタリングではデンドログラムが作られ，任意のコサイン類似度の閾値でクラスタ群を得ることができる．つまり，適切な下位トピックへの分割を行うためには，最適なコサイン類似度の閾値 θ_{sim} を決定する必要がある．しかし，コサイン類似度の閾値 θ_{sim} の決定方法は未開発であり，今後の課題である．

閾値 θ_{sim} で得られたクラスタ群を下位トピックに対応させると，下位トピック s_i の段落集合 $P_{sub}(s_i)$ は次のように対応するクラスタ C_i に属する段落の集合として得られる．

$$P_{sub}(s_i) = \{p|p \in C_i\} \tag{6}$$

3.2.2 下位トピックキーワード

各下位トピックにはその下位トピックを代表する特徴的な語を下位トピックキーワードとして付与する．下位トピックキーワードは $TF \cdot IDF$ の要領で，各語の各下位トピックにおける $PF(s,t) \cdot ISF(t)$ 値を用いて取得する．$PF(s,t)$，$ISF(t)$ は式 (7)，式 (8) で計算される値で，$PF(s,t)$ は下位トピック s の段落集合 $P_{sub}(s)$ 中での語 t の段落頻度，$ISF(t)$ は語 t の下位トピック頻度の逆数を表す．

$$PF(s,t) = |\{p|p \in P_{sub}(s), TF(p,t) \neq 0)\}| \tag{7}$$

$$ISF(t) = \frac{1}{|\{s|s \in S, p \in P_{sub}(s), TF(p,t) \neq 0\}|} \tag{8}$$

ここで S は下位トピック集合である．

下位トピック s_i にはそれぞれの $PF(s_i,t) \cdot ISF(t)$ 値の大きい語を上から k 語選択し，下位トピックキーワードとして与える．k はあらかじめ設定しておくパラメータである．

3.3 下位トピック注目度の計算

下位トピック注目度は発信者が各下位トピックにどの程度注目しているかを表す指標である．そのため，発信者はすべての下位トピックに平等に注目することはできても，すべての下位トピックを最大限注目することはできないと考えられる．そのため，各発信者は持ち点 1 を各下位トピックに振り分けるものと考え，それが下位トピック注目度であるとする．また，ある下位トピックについてのその発信者の文章が多いほど下位トピック注目度が大きくなることは自然と考えられるので，各下位トピックに割り振られた段落数を基に下位トピック注目度を計算するのが適切である．これらを合わせ，発信者 a の下位トピック s に対する下位トピック注目度 $attention(a,s)$ を次のように求める．

$$attention(a,s) = \frac{|\{p|p \in P_{sub}(s), p \in P_{auth}(a)\}|}{|P_{auth}(a)|} \qquad (9)$$

ここで $P_{auth}(a)$ は発信者 a の段落集合，$P_{sub}(s)$ は下位トピック s の段落集合である．

下位トピック注目度は発信者に属する段落数で正規化された値となっているため，発信者ごとに段落数が違っても比較することができる．

3.4 発信者意図の提示

3.3 節で計算した下位トピック注目度を各発信者間，各下位トピック間での比較が容易となるように提示する．提示する情報が多いので，凝った提示方法ではなく，シンプルな表形式での提示とする．発信者意図を表す表を図 3 に示す．行に発信者，列に下位トピックをとり，発信者の行と下位トピックの列が交わるセルが下位トピック注目度を表す．セル中には下位トピック注目度を表示し，セル全体を下位トピック注目度の大きさに応じた色で塗りつぶす．セルの色は下位トピック注目度が小さいものから，青→水色→緑→黄色→赤と大きくなるにつれて変化させる．下位トピック注目度は各発信者での総和が 1 となる値であるため，1 つの下位トピックに対する値はそれほど大きくないと考えられる．そのため，下位トピック注目度が 0.5 以上の場合に赤となるようにする．また，下位トピック注目度が 0 となる場合には，黒で塗りつぶす．このようにセルに色を付けることで，一目で発信者同士の違いを見ることができる．

4. 実　　験

4.1 パラメータ θ_{min}，θ_{max} の推定実験

中頻度の語を判定する閾値である θ_{min}，θ_{max} を決定するため，表 1 に示す 4 つのトピックに関するニュース記事を Web から収集し，それを用いて実験を行った．表 1 には記事数を示している．実験を行う前に収集した記事から，執筆者名や日時，一緒に掲載されていた写真の説明文などは除いておいた．実験手順は次の通りである．

（1） 各ニュース記事を読み，下位トピックを表すキーワードとして適切だと考えられる語を抜き出す．

（2） 抜き出した語の PF 値を計算する．

（3） PF 値の分布を見て θ_{min}，θ_{max} を決定する．

表 1　実験に用いたニュース記事
Table 1　News articles

	発信者 (記事数)				
トピック	朝日新聞	産経新聞	日経新聞	毎日新聞	読売新聞
インフルエンザ	13	22	13	16	11
トヨタ	12	12	15	20	11
西松建設	10	0	13	10	14
中央大教授刺殺	13	0	5	9	13

表 2　トピック「インフルエンザ」に関する結果
Table 2　Experimental result of the topic “influenza”

下位トピック	段落数	正解段落数	精度
企業 備蓄 センター	11	8	0.73
ソ連型 耐性ウイルス 都道府県	48	38	0.79
女児 山西省 中国	29	20	0.69
指導 女性患者 病院職員	64	46	0.72
参拝 認知症 客	14	8	0.57
死者数 済南市 衛生省	6	5	0.83
インフルエンザワクチン 発症者 隔離	12	6	0.50
安全性 接種後 副作用	20	12	0.60
先 南棟	1	1	1.00
欠席 閉鎖 施設	9	6	0.67
スペイン 遺伝子 鼻	4	4	1.00
流行注意報 乳幼児 発令	10	7	0.70
年末年始 マスク姿 子供	10	10	1.00
湿度 加湿器 不十分	9	9	1.00
規模 危険 行動	6	5	0.83
不 タイプ 主流	2	0	0
自粛 ハクチョウ 餌付け	6	5	0.83
全体	261	190	0.73

この結果 $\theta_{min} = 0.01$，$\theta_{max} = 0.1$ が妥当であると判断した．次節の実験ではこれらのパラメータを用いた．

4.2 下位トピックへの分類精度の確認実験

4.2.1 実験方法と結果

表 1 に示した 4 つのトピックについて，段落の下位トピックへの分類精度を確認する実験を行った．パラメータには $\theta_{sim} = 0.05$，$k = 3$ を用いた．4 つのトピックそれぞれについて以下の手順で実験を行った．被験者は 2 人で「インフルエンザ」,「トヨタ」と「西松建設」,「中央大教授刺殺」に分け，実験を行った．

（1） 本提案システムで下位トピックへの分割処理までを行う．

（2） 下位トピックに分類された段落を読み，適切に分類されているかを被験者に判断してもらう．

（3） 下位トピックごとに次の式で精度を計算する．

$$\text{精度} = \frac{\text{適切に分類された段落数}}{\text{下位トピックに分類された段落数}} \qquad (10)$$

結果を表 2〜表 5 に示す．

表 2〜表 5 のいずれも段落の分類精度はトピック全体で 0.6 以上であり，分類精度は概ね良いといえる．しかし，各下位トピックに注目した場合に，次のような問題点や改善すべき点が

日経	0.08	0.08	0.15	0.15	0.08	0.23	0.23	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
朝日	0.02	0.03	0.06	0.11	0.25	0.13	0.09	0.05	0.09	0.03	0.03	0.06	0.03	0.02	0.00
毎日	0.02	0.09	0.17	0.15	0.13	0.11	0.06	0.06	0.09	0.00	0.02	0.04	0.06	0.00	0.00
読売	0.04	0.06	0.09	0.24	0.06	0.09	0.00	0.08	0.05	0.05	0.05	0.08	0.03	0.06	0.01
	教員 記者会見 切磋琢磨	先生 丁寧 半導体	講義 準備 時間	微量 個室 コンビニエンスストア	一部 コート 男子留学生	心臓 死因 司法解剖	中央階段 非常階段 西側	教職員 提示 キャンパス内	110番 男性 悲鳴	4月 回路 大学院	電柱 宅 近所	犯行 ドア 計画	その後 教授室 鍵	同館 事件当時 情報	週 実施 同学部

図 3　発信者意図を提示する表
Fig. 3　Table expressing each authors' intention

表 3　トピック「トヨタ」に関する結果
Table 3　Experimental result of the topic "TOYOTA"

下位トピック	段落数	正解段落数	精度
合併 ライン 休止	29	11	0.38
要求 ベースアップ 交渉	21	17	0.81
生産台数 休業日 操業停止日	27	18	0.67
モデル レクサス エンジン	8	8	1.00
正社員 英国 契約	28	14	0.50
昇格 取締役 適任	43	30	0.70
新型車 公開 F1 世界選手権	14	8	0.57
ダイハツ工業 速報 確定	31	22	0.71
GP 山科忠チーム代表 撤退	8	6	0.75
業績予想 単体 予想	12	9	0.75
章男氏 拡大路線 企業	18	9	0.50
市民 税 法人	9	4	0.44
政府 国内生産台数 支援	8	6	0.75
大政奉還 創業家 状況	4	3	0.75
コメント 6 月末 圧縮	2	2	1.00
全体	262	158	0.60

表 4　トピック「西松建設」に関する結果
Table 4　Experimental result of the topic "Nishimatsu Construction Co., Ltd."

下位トピック	段落数	正解段落数	精度
業務上横領罪 高原和彦被告 8 月	54	21	0.39
融資 捜索 搬出	20	20	1.00
贈賄 バンコク都庁 バンコク	26	20	0.77
見通し 工作 可能性	8	3	0.375
経理 管理本部長 一身上	18	11	0.61
明確 経営責任 取締役会	17	4	0.24
企業献金 解散 新政治問題研究会	30	30	1.00
徹底 外国貿易法 外国為替	24	18	0.75
6 月 専務 側近	4	3	0.75
役員 業務上横領 委員会	3	2	0.67
受注工作 当局 違法	2	2	1.00
全体	206	134	0.65

あることが分かった．
問題点 1　1 つの下位トピックが複数の下位トピックに分割すべきと考えられる内容の段落から構成されている．
問題点 2　下位トピックキーワードがよくないため，同じ下位トピックについての段落が分類されていても下位トピックキーワードと一致するかが分からない．
問題点 3　分類された段落数の少ない下位トピックが多数存在するため，トピックがどのように分割されているかを把握することが困難である．
問題点 4　同じ下位トピックに分類すべき段落が 2 つに分割されている．

表 5　トピック「中央大教授刺殺」に関する結果
Table 5　Experimental result of the topic "murder case of chuo university"

下位トピック	段落数	正解段落数	精度
教員 記者会見 切磋琢磨	6	5	0.83
先生 丁寧 半導体	12	11	0.92
講義 準備 時間	21	16	0.76
微量 個室 コンビニエンスストア	35	9	0.26
一部 コード 男子留学生	28	16	0.57
心臓 死因 司法解剖	23	17	0.74
中央階段 非常階段 西側	12	12	1.00
教職員 提示 キャンパス内	12	6	0.50
110 番 男性 悲鳴	14	9	0.64
4 月 回路 大学院	6	4	0.67
電柱 宅 近所	7	7	1.00
犯行 ドア 計画	12	9	0.75
その後 教授室 鍵	7	6	0.86
同館 事件当時 情報	6	5	0.83
週 実施 同学部	1	0	0
全体	202	132	0.65

それぞれの問題点について，次節で考察する．

4.2.2　考　　察

問題点 1 は 1 つの下位トピックが複数の下位トピックに分割すべきと考えられる内容の段落から構成されていることである．実験結果の中で，分類精度が低かった下位トピックは複数の下位トピックに分けたほうが良いと考えられるものが多かった．表 3 の「合併 ライン 休止」や表 4 の「業務上横領罪 高原和彦被告 8 月」などがそうであった．この問題は クラスタリングを行う際の閾値 θ_{sim} を適切に決めることで改善することができると考えている．トピックの種類やニュース記事数の量によって適切な θ_{sim} は異なると考えているため，トピックに合わせて動的に θ_{sim} を決定する方法が必要である．

問題点 2 は下位トピックキーワードがよくないため，同じ下位トピックについての段落が分類されていても下位トピックキーワードと一致するかが分からないことである．表 2 の「指

導 女性患者 病院職員」や表 5 の「4 月 回路 大学院」などがそうであった.「指導 女性患者 病院」という下位トピックは東京都町田市の鶴川サナトリウム病院でのインフルエンザの集団感染に関する下位トピックであったが，下位トピックキーワードにはより具体的な「女性患者」や「病院職員」といった，間違ってはいないが適切とはいえない語が選ばれていた．また,「4 月 回路 大学院」という下位トピックなどでは「4 月」という時間に関する語が選ばれている．これらの語も不適切であると考えられる．この問題から，段落のクラスタリングには必要な語であっても，下位トピックキーワードとするには不適切な語が存在することが分かる．そのため，適切にクラスタリングを行うために抽出した中頻度語の中でもさらに下位トピックキーワードに適切な語と不適切な語に分ける必要があると考えられる．

問題点 3 は分類された段落数の少ない下位トピックが多数存在するため，トピックがどのように分割されているかを把握することが困難であることである．表 2 の「先 南棟」や「不 タイプ 主流」，表 3 の「コメント 6 月末 圧縮」など，分類された段落数が少なく下位トピックとして適切でないものが多数みられる．適切でない下位トピックが多くあるためトピックの全体が把握しづらくなっている．下位トピックとして適切でないものはまとめて「その他」としてしまうことで，下位トピック数が減り，全体を把握しやすくなると考えている．そのためには適切でないトピックを判断する方法が必要である．

問題点 4 は同じ下位トピックに分類すべき段落が 2 つに分割されていることである．3 の「昇格 取締役 適任」と「大政奉還 創業家 状況」などがそうである．この 2 つの下位トピックは両方とも創業家出身の豊田章男氏が社長に昇格するという内容の下位トピックであった．この問題についても，問題 1 についてと同様に，クラスタリングの際の閾値 θ_{sim} を適切に決定することができれば改善することができると考えている．

5. おわりに

本論文では，あるトピックに関しての記事の発信者の意図を，トピックを分割する各下位トピックの注目度によって表現する手法と，提案手法を用いて複数の発信者間で意図を比較するシステムを提案した．本提案システムを用いることで発信者がどの下位トピックに注目しているかを把握し，複数の発信者間で比較しながら記事を閲覧することができ，トピックについて偏りのない情報を得ることができると考えている．4.2 節で下位トピックへのの分類精度を確認する実験を行い，様々な課題が見つかった．今後はこれらの課題を 1 つずつ解決し，より有効なシステムとすることを目指す．

謝辞 本研究の一部は，NICT 委託研究「電気通信サービスにおける情報信憑性検証技術に関する研究開発」によるものです．ここに記して謝意を表すものとします．

文 献

[1] 北山, 角谷:“コンテンツ構成要素の順序特性に基づく比較ニュース検索方式”, 電子情報通信学会研究報告, **107**, 131, pp. 277–282 (2007).

[2] 灘本, 田中：“B-cwb: 類似コンテンツの視点差異情報を同時提示する多言語 web ブラウザ”, 日本データベース学会 Letters, **2**, 2, pp. 13–16 (2003).

[3] 吉岡, 湯本, 田中：“ニュースの視点の抽出によるマルチメディアニュースアーカイブの利用”, 電子情報通信学会研究報告, **2005**, 67, pp. 415–420 (2005).

[4] “Mecab: Yet another part-of-speech and morphological analyzer”, `http://mecab.sourceforge.net/`.