ニュース記事の理解のための背景・前提事象の抽出と分析

田中祥太郎† ヤトフトアダム† 田中 克己†

† 京都大学大学院情報学研究科 〒 606-8501 京都府京都市左京区吉田本町
E-mail: †{stanaka,adam,tanaka}@dl.kuis.kyoto-u.ac.jp

あらまし 本稿では，ニュース記事から，その主題事象の記述と主題事象に対する背景・前提事象の記述を分類して抽出する手法を提案する．ニュース記事には，その記事で新たに報道される中心的な出来事について述べた主題事象記述と，その出来事の経緯や位置付けを示す過去の出来事について述べた背景・前提事象記述が含まれる．背景・前提事象記述は，主題事象の経緯や位置付けを理解するために必要な付加的情報を示すと考えられる．これらの記述を分類して抽出することは，ニュースで報道される事象間の関連や，記事の理解容易性を分析する上で役立つと考えられる．提案手法では，ニュース記事に特有の時間表現や時制などの特徴を利用することで，主題事象記述と背景・前提事象記述を分類することを試みる．また，実際に英語のニュース記事を用いて提案手法を評価するとともに，提案手法を応用してニュースを分析し，その結果について考察する．

キーワード ニュース記事，背景情報，理解容易性

1. はじめに

近年のニュースメディアの発達により，日々発行されるニュース記事の数は増加の一途をたどっている．また Web の普及に伴い，特定のトピックに関する記事を主に扱う専門的なニュースメディアも多くなっている．このような記事数やメディア数の増加に伴い，ニュース記事の内容もより詳細で専門的なものとなっている．

これらの記事の中には，しばしばその主題となっている出来事に加え，その出来事の経緯や位置付けを示す過去の出来事についても述べたものがある．このような過去の出来事についての記述は，主題となっている出来事を理解する上で重要であるとみなされたために，記事の著者によって加えられたと考えられる．よって，このような記述を抽出することは，ニュース記事の理解容易性を分析し，記事に対して有用な情報を補完する上で役立つと考えられる．しかしながら，ニュース記事中で過去の出来事についての記述が他の記述と明示的に分けて記されることは少ないため，これらを分割して抽出することは難しい．

そこで本研究では，ニュース記事から主題事象記述と背景・前提事象記述を分類して抽出する手法を提案する．ここで，主題事象記述とはニュースの主題となっている出来事についての記述，背景・前提事象記述とは主題となっている出来事の経緯や位置づけを示す過去の出来事についての記述である．提案手法では，ニュース記事を入力とし，記事本文を文や段落などの単位に分割し，さらに各記述が主題事象記述であるのか背景・前提事象記述であるのかを，記事から得られる特徴を用いて機械学習を行うことで判別する．

図 1 に主題事象記述と背景・前提事象記述が共に含まれるニュース記事の例を示す．図 1 の記事の本文中では，下線で示した部分が背景・前提事象記述，他の部分が主題事象記述と考えられる．また，表 1 に図 1 の記事の本文を文単位に分割し，各文に対して主題事象記述または背景・前提事象記述のラベルを付与した例を示す．提案手法は，表 1 のような出力を得ることを目標としている．

Egypt opens Rafah border crossing sporadically during Gaza crisis(注1)

Tuesday, July 15, 2014

For the last few days, Egypt has sporadically opened and closed the Rafah Border Crossing with the Gaza Strip. On Thursday, Egypt opened the Rafah Border Crossing for a limited time to transfer only 11 wounded, out of dozens of Gaza residents injured during the Israeli-Palestinian conflict, which has intensified significantly lately.

...

”What is happening in Gaza is disgusting”, MENA quoted Egyptian Foreign Ministry spokesman Badr Abdelatty, ”Israel has to stop its assault on Gaza.”

Mohamed Morsi during his tenure as President of Egypt strengthened the connection between Egypt and the Gaza Strip. At the end of 2012, his government aided the coordination between the two sides and advanced a peace.

...

図 1 背景・前提事象記述を含むニュース記事（一部）

表 1 図 1 の記事における主題事象記述と背景・前提事象記述

記述（抜粋）	分類
For the last few days ...	主題事象記述
On Thursday, Egypt opened the ...	主題事象記述
”What is happening in Gaza ...	主題事象記述
Mohamed Morsi during his tenure ...	背景・前提事象記述
At the end of 2012 ...	背景・前提事象記述

(注1)：http://en.wikinews.org/wiki/Egypt_opens_Rafah_border_crossing_sporadically_during_Gaza_crisis

本稿の構成は以下のとおりである．本章では，研究の背景，およびその目的について述べた．第 2 章では，関連研究を紹介し，本研究の特徴を明確にする．第 3 章では，事象と事象記述，主題事象記述と背景・前提事象記述などの定義について述べる．第 4 章では，ニュース記事中の事象記述を，その特徴を用いて分類する手法について述べる．第 5 章では，提案手法の評価のための実験を行い，その結果に対する考察を述べる．第 6 章では，提案手法を用いてニュース記事データを分析し，その結果に対する考察を述べる．第 7 章では，本稿のまとめと，今後の課題について述べる．

2. 関 連 研 究

ニュース記事から実際の社会における出来事についての記述を抽出し，また記事と出来事の関係を分析する取り組みとしては，以下のような関連研究が挙げられる．著者ら [14] は，ニュース記事中の歴史的事象への言及を，その言及理由に基づいていくつかのクラスに分類する手法を提案した．Yang ら [13] は，文書クラスタリングアルゴリズムを用いて，ニュース記事から出来事を発見する手法を提案した．Cybulska ら [5] は，歴史上の出来事についての記述が，ニュース記事や百科事典などの文書の種類によって異なることを指摘し，これらの差異を説明するモデルを提案した．

ニュース記事などの実社会情報に基づく文書の理解を支援する取り組みとしては以下のような関連研究が挙げられる．Rennison ら [10] は，ニュース記事間の関連性に基づいて各ニュース記事を特徴空間にマッピングし，記事間の関係を可視化する手法を提案した．灘本ら [9] は，あるニュース記事に対し，そのメイントピックとサブトピックをそれぞれ抽出し，これらの特徴に基づいて関連する過去のニュース記事を取得する手法を提案した．Cui ら [4] は，ニュースのもつ多様な特徴を利用し，ニュース系列における特徴の変化を可視化する手法を提案した．

ニュース記事などの文書からの特徴抽出に関連する取り組みとしては，以下のような関連研究が挙げられる．Smith ら [12] は，日付や場所などキーワードを用いて非構造化文書からイベントを抽出する手法を提案した．Jatowt ら [6] は，文書中の時間表現やエンティティなどの特徴を用いることで，その文書において注目されている時間軸上の期間を推定する手法を提案した．Milne ら [8] は，文書中の要素から Wikipedia 項目へのリンクを生成する手法を提案した．

ニュース記事などの文書に対する関連情報の発見に関する取り組みとしては，以下のような関連研究が挙げられる．Shahaf ら [11] は，ニュース記事にまたがるトピックに基づき，ニュース記事間の関係を発見し，記事集合を構造化する手法を提案した．Allan ら [1] は，ニュース記事から新たな出来事を発見し，その経過を追うためのモデルとアルゴリズムを提案した．高橋ら [15] は，人物や組織などのエンティティの時間的および空間的なインパクトを計算する手法を提案した．

ニュース記事などの時系列文書を定量的に分析することで社会学的な知見を得る取り組みとしては，以下のような関連研究が挙げられる．Michel ら [7] は，これまでに世界中で出版された全ての本の約 4%を含むコーパスを作成し，社会文化のトレンドの変化を定量的に分析した．Cook ら [3] は，ニュースアーカイブを用いて，ニュースで特定の人物が話題になっている期間の平均の長さが 20 世紀の 100 年間に変化したかどうかを調査した．Yeung ら [2] は，ニュースアーカイブを用いて過去の出来事についての社会的な記憶を分析した．

Russia stages military exercises as Ukrainian forces advance(注2)

Wednesday, August 6, 2014

On Monday, as Ukrainian forces continue to advance into rebel held territory, Russia announced it is to hold military exercises near its border with Ukraine.

...

Russia held military exercises in the area in March; last week NATO General Philip Breedlove said Russia still has 12,000 troops adjacent to Ukraine, although Russia has claimed its forces have withdrawn from the area.

...

Meanwhile, fighting continued in Eastern Ukraine as government forces advance on the cities of Donetsk and Luhansk, the principal cities left in rebel hands.

...

図 2　複数の事象についての記述を含むニュース記事（一部）

表 2　図 2 の記事に出現する各事象と対応する記述

事象	分類	記述（抜粋）
8 月のロシア軍の演習	主題事象	On Monday ...
3 月のロシア軍の演習	背景・前提事象	Russia held ...
ウクライナ軍の侵攻	背景・前提事象	Meanwhile, fighting ...

3. 定　　義

3.1 事象とその記述

本研究における事象とは，実社会において生起し，その生起場所および生起期間が特定できる出来事とする．また，事象記述とは，事象そのものやその関連情報についての記述とする．

図 2 に複数の事象についての記述を含むニュース記事の一部を示す．また，表 2 に図 2 の記事に出現する各事象と対応する記述を示す．

3.2 主題事象とその記述

ニュースでは通常ひとつ以上の事象が取り上げられる．これらの事象のうち，特にそのニュースの主題となっているものを主題事象と呼ぶ．さらに，ニュース記事本文中の主題事象についての記述を主題事象記述と呼ぶ．

例えば図 2 の記事における主題事象は，8 月にウクライナ国境付近で行われるロシア軍の軍事演習と考えられる．また，“On Monday” から始まる一文は主題事象記述のひとつと考えられる．

(注2)：http://en.wikinews.org/wiki/Russia_stages_military_exercises_as_Ukrainian_forces_advance

3.3 背景・前提事象とその記述

ニュースでは，主題事象の他に，主題事象の経緯や位置付けを示す過去の事象が取り上げられる場合がある．これらの事象を背景・前提事象と呼ぶ．さらに，ニュース記事本文における背景・前提事象についての記述を背景・前提事象記述と呼ぶ．

例えば図 2 の記事における背景・前提事象は，3 月にウクライナ国境付近で行われたロシア軍の軍事演習，および 8 月のウクライナ軍の侵攻と考えられる．また，“Russia held” から始まる一文は前者に対する，“Meanwhile, fighting” から始まる一文は後者に対する背景・前提事象記述のひとつと考えられる．

4. 提 案 手 法

4.1 分類手法とクラス

事象記述の分類には，機械学習による分類手法として広く使われているサポートベクターマシン (SVM) を用いる．サポートベクタマシンは代表的な 2 クラス分類器であり，高い認識性能を実現しやすいことが知られている．SVM の基本的なアルゴリズムは，特徴空間上の点集合を 2 クラスに分割する超平面のうち，各点との距離が最大となるものを分離平面とするというものである．SVM による分類では，非線形分類を可能にするためカーネル法が利用される場合があるが，今回は各特徴の有効性に着目した評価を行うため，単純な線形カーネルを用いる．

提案手法では，まず自然言語処理技術を用いてニュース記事本文を複数の記述に分割する．次に，各記述から数種類の特徴を抽出し，記述を特徴空間に写像する．

分類のクラスは，主題事象記述，背景・前提事象記述の 2 クラスとする．

4.2 分類に用いる特徴

4.2.1 出 現 語

一般に，テキストを高次元ベクトル空間に写像する際には，テキスト集合から得られる各特徴語に次元を割り当てる手法が利用される．そこで，記述集合から得られる各語を特徴のひとつとして用いる．記述集合から得られる各語を正規化し，ストップワードを除くことで，特徴語集合を得る．この特徴語集合中の各語に対し，それぞれひとつの次元を割り当てる．

記述 d の語 w に対応する特徴量 $\mathrm{f}_1(d,w)$ は次のように計算する．ここで W_d は記述 d に出現する語の集合である．

$$\mathrm{f}_1(d,w)=\begin{cases}1 & w\in W_d\\ 0 & \text{otherwise}\end{cases}$$

4.2.2 品 詞 情 報

主題事象記述と背景・前提事象記述の間には，文法的特徴の差異が存在する可能性がある．例えば，主題事象記述では現在時制，背景・前提事象記述では過去時制や完了時制が多く用いられるといったことが考えられる．そこで，記述を形態素解析した際に得られる品詞タグを特徴のひとつとして用いる．形態素解析の結果として得られる可能性のある品詞タグにはそれぞれ一つの次元を割り当てる．

記述 d の品詞タグ p に対応する特徴量 $\mathrm{f}_2(d,p)$ は次のように計算する．ここで P_d は記述 d に出現する品詞タグの集合である．

$$\mathrm{f}_2(d,p)=\begin{cases}1 & p\in P_d\\ 0 & \text{otherwise}\end{cases}$$

4.2.3 タイトル類似度

一般に，ニュース記事のタイトルは主題事象の簡潔な説明となっていることが多いと考えられる．このため，主題事象記述には記事タイトルと共通する語が多く出現する可能性がある．そこで，記述と記事タイトルの類似度を計算し，特徴のひとつとして用いる．

記述 d のタイトル類似度に対応する特徴量 $\mathrm{f}_3(d)$ は次のように計算する．ここで t は記述 d が属する記事のタイトル，W_d および W_t はそれぞれ d および t に出現する語の集合である．

$$\mathrm{f}_3(d)=\frac{|W_d\cap W_t|}{|W_d||W_t|}$$

4.2.4 時間表現スコア

背景・前提事象記述には，過去の時間・時点を示す時間表現がより多く出現する可能性がある．ここで時間表現とは，“先週” や “2015 年” などの特定の時間・時点を示す表現である．これらの時間表現が示す時間・時点と記事の発行日付を比較すれば，ある記述が主題事象記述であるか背景・前提事象記述であるかを判別しやすいと考えられる．そこで，記述中の時間表現が示す時間・時点とと記事の発行日付との差異を示すスコアを計算し，特徴のひとつとして用いる．

時間表現スコアの計算では，まず記述中の時間表現の示す日付を求め，これと記事の発行日付との日数差を求める．さらにこの日数に対して対数をとり，基準日付からの日数差が大きい日付同士ではあまり大きな差がつかないようにする．

基準日付 τ_0 に対する日付 τ のスコア $\mathrm{s}(\tau,\tau_0)$ は次のように計算する．ここで $\tau-\tau_0$ は τ_0 から τ の日数差である．

$$\mathrm{s}(\tau,\tau_0)=\begin{cases}\log_2(\tau-\tau_0+1) & \tau-\tau_0>0\\ 0 & \text{otherwise}\end{cases}$$

記述 d の特徴量としては，記述から得られる各時間表現に対して計算されたスコアの最大値 $\mathrm{f}_{4\max}(d)$ および最小値 $\mathrm{f}_{4\min}(d)$ を用いる．ここで T_d は記述 d 中に含まれる時間表現に対応する日付の集合，τ_{article} は d の属する記事の発行日付である．

$$\Delta_d=\{\mathrm{s}(\tau,\tau_{\text{article}})|\tau\in T_d\}$$
$$\mathrm{f}_{4\max}(d)=\max(\Delta_d)$$
$$\mathrm{f}_{4\min}(d)=\min(\Delta_d)$$

4.2.5 抽象時間表現

背景・前提事象記述には，過去のの時間・時点を示す抽象時間表現がより多く出現する可能性がある．ここで抽象時間表現とは，“以前” や “数年前” などの具体的でない時間・時点を示す表現である．これらの抽象時間表現に対しては，それが示す具体的な時間・時点を求めることはできないが，過去・現在・未来のいずれを示すかを特定することができる．そこで，これらの時間表現に対して付与された参照時間タグを特徴のひとつ

として用いる．

記述 d に対し，過去・現在・未来のそれぞれを示す抽象時間表現の有無を示す特徴量 $\mathrm{f}_{5\mathrm{p}}(d)$，$\mathrm{f}_{5\mathrm{c}}(d)$，$\mathrm{f}_{5\mathrm{f}}(d)$ は次のように計算する．

$$\mathrm{f}_{5\mathrm{p}}(d) = \begin{cases} 1 & \text{if } d \text{ references to the past} \\ 0 & \text{otherwise} \end{cases}$$

$$\mathrm{f}_{5\mathrm{c}}(d) = \begin{cases} 1 & \text{if } d \text{ references to the current} \\ 0 & \text{otherwise} \end{cases}$$

$$\mathrm{f}_{5\mathrm{f}}(d) = \begin{cases} 1 & \text{if } d \text{ references to the future} \\ 0 & \text{otherwise} \end{cases}$$

4.2.6 記述出現位置

主題事象記述と背景・前提事象記述の間には，記事本文中での位置に差異が存在する可能性がある．例えば主題事象記述は記事の冒頭に近い部分，背景・前提事象記述は末尾に近い部分に置かれるといったことが考えられる．そこで，記事本文中での記述の相対的な出現位置を計算し，特徴のひとつとして用いる．

出現位置の計算では，まず記述が本文中の何文目に出現するかを求め，次にこの値を記事本文中の全文数で割る．

記述 d に対し，記事中における出現位置を示す特徴量 $\mathrm{f}_6(d)$ は次のように計算する．ここで D は d の属する記事に含まれる記述の集合，$\#d$ は d が記事中の何番目の記述かを示す序数 $(0, 1, ...)$ である．

$$\mathrm{f}_6(d) = \frac{\#d}{|D| - 1}$$

5. 評価実験

5.1 ニュース記事データの準備

実験には，Wikinews 英語版 (注1) のニュース記事データを用いた．Wikinews は Wikimedia 財団の運営するニュースメディアであり，2014 年 11 月時点で月間 PV は 4,400 万回以上，月間編集回数は 1,100 回以上となっている (注2)．また，オンライン百科事典 Wikipedia(注3) と同様に，利用者の誰もが記事を投稿・編集できるという特色をもっている．

ニュース記事データの収集には，Wikinews 英語版の API(注4) を利用した．収集に用いたプログラムはプログラミング言語 Python(注5) を用いて記述した．得られたニュース記事データの概要を 3 に示す．

5.2 事象記述の抽出

事象記述の抽出は，ニュース記事データからの本文抽出と本文の記述分割の二段階に分けて行った．

Wikinews の各記事は MediaWiki 記法によって記述されているため，まずこれを HTML 形式に変換した．次に，Python の HTML/XML 処理ライブラリである lxml(注6) を用い，ニュース記事本文を表すと考えられる特定のタグのみを指定して本文を抽出した．さらに，各記事の下部に存在する関連記事リストなど，ニュース記事本文とみなせない部分を除去し，改行や空白を整形して，ニュース記事本文を得た．

文の構成単位としての記述の粒度としては，文や段落などが考えられる．今回の実験では，記述を排他的に分類できる点を重視し，個々の文を記述とした．文分割には Python の自然言語処理ライブラリである NLTK(注7) を用いた．

表 3 ニュース記事データの概要

記事数	20506
学習・評価に用いた記事数	50
記述数	289396
学習・評価に用いた記述数（合計）	662
学習・評価に用いた記述数（主題事象記述）	446
学習・評価に用いた記述数（背景・前提事象記述）	216

表 4 提案手法の評価結果

手法	特徴	前後の記述	精度	
			平均	標準偏差
提案手法	全て使用	使用	0.745	0.067
ベースライン	出現語のみ	不使用	0.648	0.105
ランダム	-	-	0.673	-

5.3 訓練・評価用データの作成

事象記述の分類にあたっては，まず機械学習の訓練・評価に用いる正解データを作成した．予め取得した Wikinews 英語版の記事の中からランダムに 50 件を選び，これらの記事から記述を抽出した．さらに，これらの記述を主題事象記述，背景・前提事象記述の 2 クラスに排他的に分類した．分類作業は 1 人の人間による手作業で行った．

5.4 事象記述の分類

分類手法は第 4. 章で述べたとおりである．分類には Python の機械学習ライブラリである scikit-learn(注8) を用いた．

5.5 分類精度の評価

評価は，5 分割交差検定によって行った．評価結果を表 4 に示す．なおベースラインとして，出現語のみを特徴として用い，前後の記述を使用しなかった場合の結果を併記する．また，全ての記述をランダムに分類した場合に得られる精度の平均値を示す．

また，各特徴の有効性の評価のため，それぞれの特徴を除いた場合の分類精度を評価した．各特徴を除いた場合の精度を表 5 に示す．

5.6 評価結果に対する考察

表 4 より，提案手法がベースラインに比べて高い精度を実現している．このことから，提案手法の優位性が示されたと言えるなお，ベースラインはランダムに分類した場合を下回ってお

(注1)：http://en.wikinews.org/
(注2)：http://stats.wikimedia.org/wikinews/EN/SummaryEN.htm
(注3)：http://en.wikipedia.org/
(注4)：http://en.wikinews.org/w/api.php
(注5)：http://www.python.org/
(注6)：http://lxml.de/
(注7)：http://www.nltk.org/
(注8)：http://scikit-learn.org/

表 5　各特徴を除いた場合の精度変化

除いた特徴	前後の記述を使用		前後の記述を不使用	
	平均	標準偏差	平均	標準偏差
（全て使用）	0.745	0.067	0.725	0.119
出現語	0.734	0.056	0.748	0.079
品詞情報	0.740	0.118	0.714	0.123
タイトル類似度	0.743	0.067	0.725	0.119
時間表現スコア	0.716	0.073	0.690	0.088
抽象時間表現	0.745	0.067	0.730	0.108
記述出現位置	0.734	0.072	0.693	0.073

り，出現語のみを特徴とするような手法では有効に分類することが困難であると推定される．

表 5 より，出現語を特徴から除いた場合でも精度には大きな差が見られない．このことから，記述中の個々の語は，記述が主題事象記述か背景・前提事象記述かを判別することにはそれほど有用でないと推定される．この理由としては，記述の単位を文としているため，ひとつの記述に含まれる特徴語がそもそも少ないことや，事象を表現する語は事象ごとに異なり，各語の意味と記述の性質が無関係であることなどが考えられる．

品詞情報を特徴から除いた場合，精度の平均値には大きく変化しないが，標準偏差は大きくなっている．このことから，品詞は分類の安定性に寄与していると推定される．

時間表現スコアを特徴から除いた場合，精度の平均値が大きく低下している．このことから，時間表現スコアは用いた特徴の中で最も有効性が高いと推定される．この理由としては，主題事象記述では事象の詳細な説明のため時間表現が多用されること，背景・前提事象記述では主題事象との区別を明確にするために具体的な時間表現が用いられることなどが考えられる．

タイトル類似度，抽象時間表現，記述出現位置のそれぞれを除いた場合でも，精度の向上にはそれほど寄与しないと推定される．

また，前後の記述を使用した場合，全体的に精度の平均値が高く，標準偏差が小さくなっている．このことから，前後の記述を用いることは精度とその安定性の向上に寄与していると推定される．

6. ニュース記事データの分析

6.1 記事の総記述量と背景・前提事象記述の占める割合

提案手法を用いた記事の分析の一例として，記事の総記述数と，記事中で背景・前提事象記述が占める割合との関係についての分析を行った．結果を図 3 に示す．

図 3 より，記事中で背景・前提事象記述が占める割合は，記事の総記述数が 11-20 件の場合に最も大きくなっている．また総記述数が 11 件以上の場合でみると，総記述数が増加するにつれてと背景・前提事象記述が占める割合は減少している．このことから，背景・前提事象記述の量は文章の総記述量に比例して増加するわけではないと推定される．

6.2 背景・前提事象記述の割合と記事の理解容易性

提案手法の応用として，ニュース記事に対して適切な背景・前提事象記述を補完し，ユーザが記事を理解することを支援することが挙げられる．そこで提案手法の応用に向けた分析の一例として，記事における背景・前提事象記述の割合と，記事の理解容易性との関係についての分析を行った．

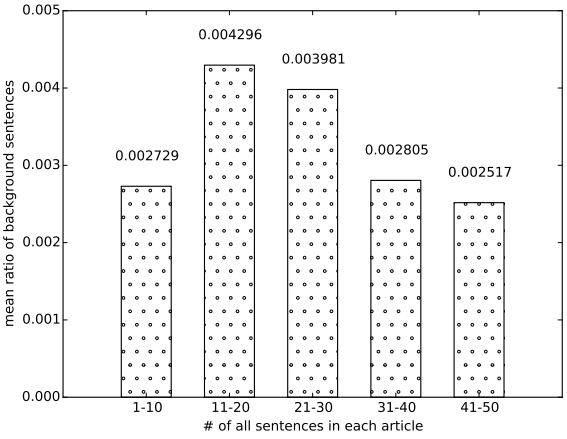

図 3　記事の総記述量と背景・前提事象記述の占める割合

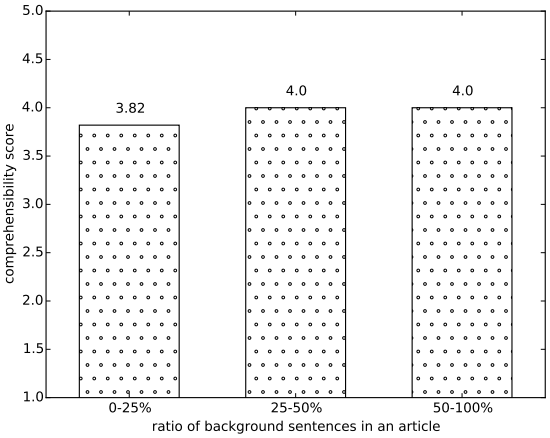

図 4　背景・前提事象記述の割合と理解容易性

分析実験は次の手順で行った．まず評価実験で用いた正解データを全て訓練データとして用い，提案手法による学習を行い，分類器を得た．次に評価データとして，Wikinews 英語版のニュース記事データからランダムに 30 件の記事を選択した．この際，訓練データとして用いたニュース記事データは除外した．さらに，評価データから提案手法によって背景・前提事象記述を抽出した．また，記事の評価者としては 6 人の被験者を選び，各記事の理解容易性を 1-5 点の 5 段階で評価させた．被験者は全て情報学の研究に従事する 20 歳代の学生であり，日常的に英語を使用している．記事中で背景・前提事象記述が占める割合が，それぞれ 25%未満，25%以上 50%未満，50%以上の場合における理解容易性スコアの平均値を図 4 に示す．

図 4 より，背景・前提事象記述が 25%未満の場合より 25%以上 50%未満の場合のほうがやや理解容易性スコアの平均値が高い．しかしながら，25%以上 50%未満の場合と 50%以上の場合では理解容易性スコアに差が見られない．このことから，背

景・前提事象記述は一定の割合より多くなっても，それに比例して理解容易性が向上するわけではないと推定される．

7. おわりに

本稿では，機械学習を用いてニュース記事中の記述を主題事象記述，背景・前提事象記述，その他の記述に分類する手法を提案した．また，提案手法を用いてニュース記事データを分析することを試みた．

提案手法の評価実験では，ベースラインに対する提案手法の優位性が確認できた．また，各特徴の有効性を詳細に分析した結果，時間表現スコアが特に有用であることがわかった．

今後の課題としては，提案手法の精度の改善および適用範囲の拡張が挙げられる．今回の実験では提案手法の優位性が認められたが，その精度は実用的に際して十分とは言えない．そこで，今後はより多くの種類の特徴を用いて提案手法を改良することが考えられる．その場合，事象記述のみから抽出できる特徴だけでなく，事象記述と記事中の他の部分の関係を示すような特徴に着目することが考えられる．

適用範囲の拡張としては，文ではなく段落を記述の単位とすることが考えられる．本稿では各記述を排他的に分類できることを重視し，文単位での分類を行ったが，抽出できる特徴数が少なかったため，十分な精度が実現できなかった可能性がある．多くのニュース記事では文ではなく段落がひとつの意味的まとまりであるため，今後は段落を記述の単位とすることが考えられる．段落を単位とした場合は，排他的な分類が困難となる可能性があるため，非排他的な分類手法についても検討したい．

ニュース記事データの分析実験については，より大規模で信頼性の高いデータセットを用いた実験や，様々なニュースメディアの差異の分析についても取り組みたい．

将来の展望としては，提案手法をニュース記事の検索・拡張・補完などに応用することが挙げられる．ニュース記事には，主題事象が同一であっても，主題事象記述のみを含む記事や背景・前提事象記述を多く含む記事など，様々なものが存在すると考えられる．例えばニュースの内容を手早く把握したい場合，主題事象記述のみを含む記事が有用であると考えられる．逆にニュースの内容について深く知りたい場合，背景・前提事象記述を多く含む記事が有用であると考えられる．しかしながら，従来のニュース検索ではこのような観点からの検索は困難である．このような場合，提案手法を応用すればより柔軟な検索が可能となると考えられる．また，大量のニュース記事から主題事象記述および背景・前提事象記述を自動的に分類して抽出することができれば，記事に対して有用な背景・前提事象情報を推薦することも可能になると考えられる．このように，提案手法の実際的な応用についても今後取り組みたい．

謝　辞

本研究の一部は，文科省科研費基盤 (A)「ウエブ検索の意図検出と多元的検索意図指標にもとづく検索方式の研究」(24240013, 研究代表者：田中克己)，戦略的創造研究推進事業 (さきがけ)「集合記憶の分析および歴史文書からの知識抽出手法の開発」(研究代表者：Adam Jatowt) によるものです．ここに記して謝意を表します．

文　献

[1] J. Allan, R. Papka, and V. Lavrenko. On-line new event detection and tracking. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 37–45. ACM, 1998.

[2] C.-m. Au Yeung and A. Jatowt. Studying how the past is remembered: towards computational history through large scale text mining. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pp. 1231–1240. ACM, 2011.

[3] J. Cook, A. Das Sarma, A. Fabrikant, and A. Tomkins. Your two weeks of fame and your grandmother's. In *Proceedings of the 21st international conference on World Wide Web*, pp. 919–928. ACM, 2012.

[4] W. Cui, H. Qu, H. Zhou, W. Zhang, and S. Skiena. Watch the story unfold with textwheel: Visualization of large-scale news streams. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 3(2):20, 2012.

[5] A. Cybulska and P. Vossen. Event models for historical perspectives: Determining relations between high and low level events in text, based on the classification of time, location and participants. In *LREC '10*, pp. 3355–3362, 2010.

[6] A. Jatowt, C.-M. Au Yeung, and K. Tanaka. Estimating document focus time. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, pp. 2273–2278. ACM, 2013.

[7] J.-B. Michel, Y. K. Shen, A. P. Aiden, A. Veres, M. K. Gray, J. P. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant, et al. Quantitative analysis of culture using millions of digitized books. *science*, 331(6014):176–182, 2011.

[8] D. Milne and I. H. Witten. An open-source toolkit for mining wikipedia. *Artificial Intelligence*, 194:222–239, 2013.

[9] A. Nadamoto and K. Tanaka. Time-based contextualized-news browser (t-cnb). In *Proceedings of the 13th international World Wide Web conference on Alternate track papers & posters*, pp. 458–459. ACM, 2004.

[10] E. Rennison. Galaxy of news: An approach to visualizing and understanding expansive news landscapes. In *Proceedings of the 7th annual ACM symposium on User interface software and technology*, pp. 3–12. ACM, 1994.

[11] D. Shahaf and C. Guestrin. Connecting the dots between news articles. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 623–632. ACM, 2010.

[12] D. A. Smith. Detecting and browsing events in unstructured text. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 73–80. ACM, 2002.

[13] Y. Yang, T. Pierce, and J. Carbonell. A study of retrospective and on-line event detection. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 28–36. ACM, 1998.

[14] 田中, ヤトフト, 田中. 記事アーカイブを用いた歴史的事象の言及分析. 第 6 回データ工学と情報マネジメントに関するフォーラム (DEIM), pp. F2–6, 2014.

[15] 高橋, 大島, 山本, 岩崎, 小山, 田中. インパクトを考慮した歴史エンティティの重要度計算手法. 情報処理学会論文誌, 52(12):3542–3557, 2011.