

事象と記事の主題・背景関係の分析に基づくニュース記事の理解支援

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、ニュース記事中で記述される事象間の関係の分析に基づき、与えられたニュース記事を理解するのに必要となる過去のニュース記事を検索する手法を提案する。ニュース記事は、一般的に、ニュース記事の主題となる事象の記述（主題事象）、および、主題事象の背景となる事象の記述（背景事象）から構成されている。背景事象は、主題事象以前に生じた事象記述であり、この意味で、背景事象と主題事象との間には、時間的展開関係や事象としての包含関係が存在する。与えられたニュース記事の主題事象の位置付けを理解するためには、そのニュース記事中の背景事象や、その背景事象を主題事象とした過去のニュース記事を理解することが必要となる。本稿では、どの事象がどの記事で主題または背景として記述されているかに着目し、記事集合と事象集合との間の言及構造に基づき、理解するのに必要となる事象・記事の重要度を計算し、記事をランキングする手法を提案する。提案手法は、事象が記事の主題あるいは背景として言及されているという2種類の関係を利用し、この関係構造を2部グラフによって表現し、相互再帰的なアルゴリズムによって記事と事象の重要度スコアを計算する。実験では、ニューヨーク・タイムズの英語ニュース記事アーカイブを用いて提案手法を評価し、その結果について考察する。

キーワード ニュース記事, 理解容易性

1. はじめに

我々が今日目にするニュース記事の中には、以前から続く話題の一部であるような事象（出来事）を主題とするものが多く存在する。このような記事では、その事象を含む話題に属する他の事象が言及されることがある。例えば図1のニュース記事は、ギリシャにおける国民投票について報じたものである。このニュースはギリシャの金融危機という話題の一部である事象についての報道と見なせる。さらに、記事本文中にはギリシャが債権国から緊縮策の実施を求められたという他の事象についての言及が見られる。

これらの背景的事象についての記述があることによって、読者はニュースをとりまく話題の存在を知ることができる。しかしそれらの事象への言及は、必ずしもその話題を理解する上で十分な情報を提供するものではない。実際には、これらの言及は関連する話題の存在を示唆する程度のものであることも多い。このため、読者が事前にその話題について知らない場合、読んでいる記事の主題事象が一連の話題の中でどのような意味をもつのかを理解することが困難となることがある。本研究では、これをニュースの位置付けを理解するための情報が不足している状態であると捉える。

本研究において、ニュースの位置付けの理解のために役立つと考える情報は主に次のようなものである。

- ニュースが含まれる話題の中の主要な事象
- 話題の中の主要な事象と、ニュースの内容との関係性

Greek Prime Minister Calls for Referendum on Bailout Terms^(注1)

Friday, June 26, 2015

In an unexpected move, Prime Minister Alexis Tsipras went on national television early Saturday to call for a referendum on July 5, so that Greek citizens can decide whether to accept or reject the terms of a bailout deal proposed by the country's creditors.

...

greek financial crisis

All News Videos Images Shopping More Search tools

About 189,000 results (0.57 seconds)



Greece's economic crisis goes on, like an odyssey witho...

The Guardian - Jan 4, 2016
Greek prime minister Alexis Tsipras has predicted that 2016 will mark 'a final exit from economic crisis'. Photograph: Thierry Charlier/AFP/Getty ...

Greece Not to Agree to Creditors Unreasonable Austerity Demands ...
Sputnik International - Jan 3, 2016

Explore in depth (80 more articles)



Hope and fear in the endless Greek crisis

Financial Times - Dec 22, 2015
The Greek economic crisis has blighted the country and the eurozone for six years. The election last January, which brought Alexis Tsipras and ...

Economic crisis affects Greeks' health
Fulton News - Dec 20, 2015

Understanding Greece's Year in Crisis
In-Depth - Huffington Post - Dec 21, 2015



Explore in depth (487 more articles)

図1: 背景事象に言及するニュース記事と、それに関する検索結果の例

このうち後者については、前者の情報があれば、通常のニュース検索エンジンに入力するクエリを工夫することで、ある程度

(注1): <http://www.nytimes.com/2015/06/27/business/international/greek-debt-talks-enter-final-stages.html>

理解に役立つ情報を発見することが可能であると考えられる。しかしながら前者については、既存のニュース検索エンジンを用いても発見が容易ではない場合がある。例えばあるニュースの背景として存在する話題について検索する場合、ユーザはまず記事の見出しや本文から、話題を表現すると考えられるいくつかの単語を選び、それをクエリとして入力する。しかし、ユーザがその話題について十分な知識を持たない場合、話題を表現する適切なクエリを作成するのは難しい。また、仮に適切なクエリを作成できた場合でも、検索結果に含まれる膨大な記事の中から、話題の中で主要な事象について述べている記事を発見するためには多大な労力を要する。これは、ユーザがどの事象が重要かを知るには、各記事を逐一閲覧しなければならないためである。例えば先のギリシャの金融危機に関するニュースの場合、“greek financial crisis”でニュース検索を行うと、図1のように膨大な検索結果が得られる。さらに、多くのニュース検索エンジンは記事を時系列順に表示するため、有用な記事を早期に発見することが困難となる。一部の検索エンジンは関連度順に記事を表示することも可能であるが、これらの関連度は単純な文書類似度に基づくものであることが多く、話題の中で重要な事象について述べた記事が必ずしも上位にランキングされるとは限らない。

そこで本研究では、これらの問題を解決するため、与えられた記事に対し、ユーザがその記事を理解する助けとなるような記事をランキングして出力する手法を提案する。提案手法において上位にランキングする理解に役立つ記事とは、ニュースが属する話題の中の主要な事象について記述した記事である。ここで話題の中の主要な事象とは、話題に属する多くの事象の説明のために言及される事象とする。

提案手法では、記事のランキングにおいて、記事とそれに含まれる事象の関係性に着目する。従来の研究では、ニュース記事は単一の事象について記述した文書として捉えられることが多かった。しかし実際には、単一の記事において複数の異なる事象が言及されることは珍しくない。また逆に、単一の事象が複数の異なる記事から言及されることもある。これらを踏まえ、本研究では記事と事象は多対多の関係にあるものとして捉える。提案手法では、記事と事象の関係を、ある事象がある記事の主題であるという関係と、ある事象がある記事の背景であるという2種類に分類する。その上で、これらの考えをモデル化するため、記事と事象の関係を2部グラフ構造によって表現する。さらに、この2部グラフに対してHITSアルゴリズムを応用した相互再帰的な重要度計算アルゴリズムを適用し、記事をランキングする。

本稿の構成は以下のとおりである。第1.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、各種の定義事項について述べる。第4.章では、記事のランキングのための重要度計算手法について述べる。第5.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、本稿のまとめと、今後の課題について述べる。

2. 関連研究

本研究の目的はニュース記事の理解を支援することである。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の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。瀬本ら[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[3]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。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。

本研究の提案手法では、ニュース記事から事象を構造化して抽出する既存手法を応用する。自然言語による文書からの事象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。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。Yangら[15]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[4]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。

本研究の提案手法では、ニュース記事と事象の関係を主題関係または背景関係に分類する。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間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、ニュース記事中の文や段落といった意味的まとまりが、ニュースの主題についての記述であるか、あるいはニュースの背景についての記述であるかを分類する手法を提案した。また著者ら[13]は、ニュース記事中の歴史的出来事への言及を、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。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。

本研究の提案手法では、記事と事象の関係を2部グラフ構造として表現し、グラフに対して相互再帰的な計算手法を適用することで記事と事象の重要度を計算する。グラフ上のノードをリンク関係に基づいてスコアリング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。Brinら[2]は、Webページ間のリンク関係に基づいてWeb検索エンジンの検索結果をランキングする手法を提案した。Kleinbergら[7]は、Webページに対してHub度とAuthority度という2種類の重要度評価指標を導入し、これらを相互再帰的な計算手法によって求めるためのHITSアルゴリズムを提案した。Dengら[5]は、HITSアルゴリズムを一般的な2部グラフに拡張した手法を提案した。

3. 定 義

3.1 事象と記事

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である。事象には、例えば“2015 年 11 月 13 日にフランスのパリで同時多発テロ事件が発生した”といったようなものが考えられる。

ニュース記事やニュースメディアに関する既存研究では、ニュース記事を単一の出来事について記述した文書として定義することがある。これらの研究では、事象と記事は一対一に対応すると仮定される。多くの場合、この定義は問題を抽象化する上で有効である。

しかし、実際のニュース記事の全てが必ずしも単一の事象を扱っているとは限らない。例えば図 2 の記事は、2014 年 8 月にロシアが実施した軍事演習について報道しているが、記事本文の下線部では同年 3 月にロシアが実施した別の軍事演習等の、過去の他の事象についても言及している。これらの例を鑑みると、特に複数の記事にまたがるような話題とそれに含まれる事象を考える場合、記事と事象は区別して扱う必要があると思われる。よって、本研究では事象と記事を区別して扱う。

3.2 主題・背景関係

ある記事中である事象が言及されるとき、そこには様々な理由が考えられる。そのような理由の中で最も一般的であると思われるのは、記事自体がその事象を主題として報道するために、事象について記述している場合である。しかし、記事が事象に言及する理由はこのひとつのみに限らない。例えばある記事で主題となっている事象に対し、それを引き起こす原因となった別の事象が存在する場合、そのような原因事象についての記述が記事に含まれることがある。また、事象が生起するまでの経緯を説明するために同じ話題に属する一連の事象に言及することや、過去の事例と比較するために類似する事象に言及することが考えられる。

これらを踏まえ、本研究では記事中に出現する事象を次の 2 種類に分類する。

主題事象 記事における事実の報道の主題となる事象

背景事象 記事中で言及されており主題事象とは異なる事象

図 2 の例において、8 月のロシア軍の演習は主題事象と考えられる。対して 3 月のロシア軍の演習は背景事象と考えられる。これらを含めた事象の分類を図 2 の表にまとめる。

主題事象あるいは背景事象という呼称は、記事から見た場合の記事と事象との関係に基づくものである。よって、ある記事の主題事象である事象が、別の記事の背景事象となる場合もある。事象が記事の主題事象となると、事象と記事は主題関係にあると言う。また、事象が記事の背景事象となると、事象と記事は背景関係にあると言う。

ひとつのニュース記事は必ず一件以上の主題事象を含むものとする。また、あるニュース記事に対して、背景事象は全く含まれないか、一件以上含まれるかのいずれかである。

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. ...

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	背景事象	Russia held ...
ウクライナ軍の侵攻	背景事象	Meanwhile, fighting ...

図 2: ニュース記事とそれに記述される複数の事象

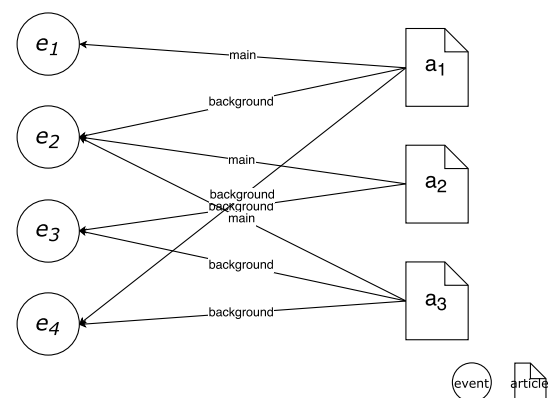


図 3: 事象と記事の関係を表現する 2 部グラフ

3.3 事象・記事関係グラフ

複数の記事の集合を考えた場合、その各記事中で言及される事象の集合を求めることができる。さらに、記事と事象の多対多関係を考えると、これらの成す構造は 2 部グラフ構造として表現することができる。2 部グラフにおいて、事象と記事は異なる種類のノードとして表現する。また、事象と記事の関係はそれぞれに対応するノード間を結ぶリンクとして表現する。事象と記事との関係としては、主題関係と背景関係の 2 種類を考えるため、2 部グラフには 2 種類のリンクが存在する。2 部グラフにおいて、事象を表すノードを事象ノード、記事を表すノードを記事ノード、主題関係をおよび背景関係を表すリンクをそれぞれ主題リンクおよび背景リンクと呼ぶ。

(注2): <http://en.wikinews.org/wiki/>

4. 提案手法

提案手法における入力と出力は以下のとおりである．

入力 単一のニュース記事

出力 ランキングされたニュース記事集合

手法の概要は以下の通りである．以降では手法の各ステップの詳細について述べる．

- (1) 候補記事集合の検索
- (2) 事象集合の抽出
- (3) 主題・背景関係の分類および2部グラフの構築
- (4) 重要度計算による記事のランキング

4.1 候補記事集合の検索

あるニュース記事を読んで、その理解に役立つ他のニュース記事を発見したい場合、既存のニュース検索エンジンを利用することが考えられる．このとき、ユーザは記事からいくつかの重要と思われる語を選び、それらをクエリとしてニュース検索エンジンに入力し、検索結果から記事を選択して閲覧する．仮にこの方法を用いた場合、検索結果として出力される記事の数が十分に多ければ、全ての検索結果を逐一閲覧することで、最終的に理解に役立つニュース記事を発見することは可能であると考えられる．

そこで提案手法では、まず有用な記事を十分多く含むような候補記事集合を得るため、記事の bag-of-words 表現を用いて類似検索を行う．Bag-of-words 表現における記事の各語の重みの計算は、文書検索において広く用いられる TF-IDF スコアを採用する．

4.2 事象集合の抽出

一般に、ニュース記事中には“○○月日に 国と□□国が会談した”といった具体的な事象の記述が多く含まれる．これらの事象記述は、次のような情報を用いて構造化することができる．

- 事象の生起日時
- 事象の生起場所
- 事象の種類
- 事象に直接関わるエンティティ

このように事象を構造化することで、異なる記事中で言及されている事象の同一性を判定することが可能となる．

自然言語テキストから事象を構造化して抽出する手法としては、Smith ら [12] による手法が代表的である．提案手法では、Smith らの手法によって抽出・構造化された事象記述を簡易的なルールによりグループ化し、事象として扱う．グループ化の詳細については第 5. 章で述べる．

4.3 主題・背景関係の分類および2部グラフの構築

ニュース記事中に記述される事象が、その記事の主題事象であるか、あるいは背景事象であるかを区別することは単純な方

法では難しい．これらを判別するには、記事に含まれる事象記述の特徴を、記事全体の内容の特徴と比較することが必要となる．

著者ら [14] は、事象記述と記事内容を比較することで得られる相対的な特徴を利用し、記事中の各事象記述を、機械学習を利用して主題事象記述あるいは背景事象記述に分類する手法を提案した．著者らはこの手法において、各事象記述に対して次のような特徴を利用した．

- 事象記述の bag-of-words 表現
- 事象記述中の語の品詞タグ
- 記事本文中での事象記述の相対的な出現位置
- 事象記述中の時間表現と記事発行日付との時間的差異

また分類には機械学習による2値分類アルゴリズムとして一般的なサポートベクタマシン (SVM) を使用した．

本研究の提案手法では、この分類手法を応用して記事と事象の関係を主題または背景に分類する．提案手法では、上記と同様の特徴を用いて事象への言及を含むパラグラフのベクトル表現を生成し、各パラグラフを機械学習によって主題事象記述あるいは背景事象記述に分類する．これにより得られたパラグラフのラベルを、記事と事象の関係ラベルとして用いる．以上の手順により得られた関係ラベルおよび記事集合・事象集合を要素として、2部グラフを構築する．

4.4 重要度計算による記事のランキング

提案手法では、記事のランキングを得るため、2部グラフ上の記事ノードの重要度スコアを HITS に類似した相互再帰的計算により求める．

4.4.1 HITS および Generalized co-HITS

2部グラフ上での相互再帰的計算手法としては、Deng ら [5] による generalized co-HITS アルゴリズムが知られている．Generalized co-HITS アルゴリズムは、Kleinberg ら [7] による HITS アルゴリズムを2部グラフに対して拡張したものである．

HITS アルゴリズムは Web ページの authority 度および hub 度をページ間のハイパーリンク構造に基づいて計算する手法である．ここで、authority 度および hub 度は、Web ページの重要度を示す2種類の評価指標であり、以下のような相互再帰的定義によって説明される．

- 多くの重要な authority にリンクするページは重要な hub
- 多くの重要な hub にリンクされるページは重要な authority

この定義に基づき、HITS アルゴリズムは Web ページ p の hub 度 $hub(p)$ および authority 度 $auth(p)$ を次のように計算する．Web ページ集合 P とページ間リンクの集合 $L: P \times P$ を考える．ここで $(p, q) \in L$ はページ p からページ q へのハイパーリンクが存在することを意味する．

$$auth(p) = \sum_{q \in P_p^h} hub(q) \quad \text{s.t.} \quad P_p^h = \{q \mid (q, p) \in L\} \subset P$$

$$hub(p) = \sum_{q \in P_p^a} auth(q) \quad \text{s.t.} \quad P_p^a = \{q \mid (p, q) \in L\} \subset P$$

上式からわかるように、HITS アルゴリズムは各 Web ページが authority 度および hub 度の両方をもつという仮定に基づいている。

Generalized co-HITS アルゴリズムでは、HITS アルゴリズムを 2 部グラフに拡張する。2 部グラフ $G = (U \cup V, E)$ を考える。ここで U および V はノード集合であり、 $U \cap V = \emptyset$ とする。また、 $E : U \times V$ はリンク集合である。ノード $u \in U$ のスコア $s(u)$ および $v \in V$ のスコア $s(v)$ は次のように計算する。

$$\begin{aligned} s(u) &= \sum_{v \in V_u} s(v) & \text{s.t.} \quad V_u &= \{v \mid (u, v) \in E\} \subset V \\ s(v) &= \sum_{u \in U_v} s(u) & \text{s.t.} \quad U_v &= \{u \mid (u, v) \in E\} \subset U \end{aligned}$$

上式は、hub 度のみをもつノードと authority 度のみをもつノードの 2 種類のノード集合からなる 2 部グラフにおいて、各ノードの重要度を計算していると考えることができる。

なお、Generalized co-HITS アルゴリズムにおいては、2 部グラフへの拡張に加えて隣接行列の各値の重みを正規化しているが、本研究ではこの点については考慮しない。

4.4.2 事象—記事関係グラフに対する HITS の仮説の適用

ニュース記事を事象へのリンクを有する文書として考えた場合、事象の重要度を HITS アルゴリズムにおける authority 度、記事の重要度を hub 度に対応づけてとらえることができる。本研究では、2 部グラフ上のリンクとして主題リンクおよび背景リンクの 2 種類を考える。HITS アルゴリズムの仮説を本研究に対して形式的に適用した場合、以下の 4 つの仮説が導かれる。

- (i) 多くの重要な事象を主題として言及する記事は重要
- (ii) 多くの重要な事象を背景として言及する記事は重要
- (iii) 多くの重要な記事で主題として言及される事象は重要
- (iv) 多くの重要な記事で背景として言及される事象は重要

以降では、実際のニュース記事における記事と事象の関係に基づき、これらの仮説のうち幾つかを選択あるいは修正して用いることで、2 種類の計算アルゴリズムを導く。

4.4.3 アルゴリズム 1: 重要度伝播における非対称性の導入

アルゴリズム 1 では、事象と記事の間の重要度伝播方向が、主題リンクと背景リンクで異なるという考え方を導入する。この考え方にに基づき、各仮説の妥当性を検証する。

まず仮説 (i) については、例えば大規模なテロ事件のような重要度が大きい事象に対して、その事象を主題として報道する記事は当然ユーザにとって重要と考えられる。よって、仮説 (i) は妥当といえる。

次に仮説 (ii) については、例えばさまざまな重要な事象を原因として言及されているからといって、その記事自体が重要であるとは限らない。よって、仮説 (ii) は妥当とは言えない。

さらに仮説 (iii) については、事象の重要性はそれが多くのニュースで報道されるからといって重要であるとは限らない。よって、仮説 (iii) は妥当とは言えない。

最後に仮説 (iv) については、多くの重要な記事において原因

となっているような事象は重要な事象と考えられる。よって、仮説 (iv) は妥当といえる。

これらの考察に基づき、アルゴリズム 1 では、記事および事象の重要度を以下のように計算する。対象とする記事ノード集合を A 、事象ノード集合を E 、主題リンク集合を L^m 、背景リンク集合を L^b と表す。 $e \in E$ の重要度 $s(e)$ および $a \in A$ の重要度 $s(a)$ は次のように計算する。

$$\begin{aligned} s(e) &= \sum_{a \in A_e^b} s(a) & \text{s.t.} \quad A_e^b &= \{a \mid (a, e) \in L^b\} \subset A \\ s(a) &= \sum_{e \in E_a^m} s(e) & \text{s.t.} \quad E_a^m &= \{e \mid (a, e) \in L^m\} \subset E \end{aligned}$$

上式はすなわち、事象の重要度を記事の重要度に伝播する際には主題リンクを、記事の重要度を事象の重要度に伝播する際には背景リンクをたどることを意味する。

4.4.4 アルゴリズム 2: 事象に対する主題・背景重要度の導入

アルゴリズム 2 では、事象が主題としての重要度と背景としての重要度の 2 種類の重要度を別個に有するという考え方を導入する。この考え方にもとに基づき、上記の仮説を修正する。

事象の中には、多くの記事において背景事象として参照されながら、主題として参照されることが少ないようなものが存在する。またその逆に、多くの記事において主題事象として参照されながら、背景として参照されることが少ないようなものも存在する。これらのことから、事象の主題としての重要度と背景としての重要度は必ずしも一致しない場合があると考えられる。

これに対し、対象記事の理解に役立つ情報を含むかという観点から見た場合、記事の重要度は一種類のスコアとして計算されることが望ましい。より具体的には、重要な主題事象と重要な背景事象をともに含むような記事が、より有用な記事としてランキングされるべきと考えられる。

これらの考察に基づき、記事および事象の重要度を以下のように計算する。対象とする記事ノード集合を A 、事象ノード集合を E 、全リンク集合を L 、主題リンク集合を L^m 、背景リンク集合を L^b と表す。 $e \in E$ の主題としての重要度 $s^m(e)$ および背景としての重要度 $s^b(e)$ 、さらに $a \in A$ の重要度 $s(a)$ は次のように計算する。

$$\begin{aligned} s^m(e) &= \sum_{a \in A_e^m} s(a) & \text{s.t.} \quad A_e^m &= \{a \mid (a, e) \in L^m\} \subset A \\ s^b(e) &= \sum_{a \in A_e^b} s(a) & \text{s.t.} \quad A_e^b &= \{a \mid (a, e) \in L^b\} \subset A \\ s(a) &= \sum_{e \in E_a} s^m(e) + s^b(e) & \text{s.t.} \quad E_a &= \{e \mid (a, e) \in L\} \subset E \end{aligned}$$

上式はすなわち、事象の主題としての重要度および背景としての重要度を、それぞれ主題リンクおよび背景リンクに基づいて計算し、記事の重要度についてはこれらの合計を計算することを意味する。

5. 実験

提案手法の評価のため、実際のニュース記事を用いた実験を行った。実験に必要なプログラムは、全てプログラミング言語 Python^(注1) を用いて実装した。

5.1 実験概要

5.1.1 ニュース記事データ

実験には The New York Times Annotated Corpus (NYT Corpus)^(注2) のニュース記事を利用した。NYT Corpus は、ニューヨーク・タイムズ^(注3) に 1987 年から 2007 年の間に掲載された、総計 1,855,658 件の新聞記事からなるニュースアーカイブである。

NYT Corpus には、一般的なニュース記事だけでなく、社説やコラムなども含めた様々なコンテンツが含まれている。これらの内容の種別は、各記事に付与された属性情報および記事タイトルなどからある程度判別することが可能となっている。実験では、対象を一般的なニュース記事のみに限るため、記事のカテゴリを表す属性 ‘taxonomicClassifiers’ や、通常の記事以外に設定される属性 ‘typesOfMaterial’、記事のタイトル等をもとに、以下の条件に全て合致する記事のみを対象とした。

- 発行日付・タイトル・本文が付与されている
- ‘taxonomicClassifiers’ に ‘Top/News’ を含む
- ‘taxonomicClassifiers’ に ‘Top/Opinion’ を含まない
- ‘typesOfMaterial’ の値が設定されていない
- タイトルに特定の特集記事を示す文字列を含まない

記事の発行日付は、1987-1991 年の 5 年間に収まるものに限定した。評価対象の記事は、主に背景情報を有するような記事を中心とし、計 25 件を選出した。

5.1.2 候補記事集合の取得

提案手法では、選択した記事に対して類似検索より候補記事集合を動的に取得する必要がある。この処理の高速化を図るため、予めコーパスに含まれる全てのニュース記事を検索エンジンライブラリ Elasticsearch^(注4) を用いてインデックスした。類似検索には、Elasticsearch の提供する機能である More Like This Query^(注5) を利用した。More Like This Query は、Elasticsearch にインデックスされた文書を入力とし、その文書に類似する文書を検索するための機能である。More Like This Query では、入力として与えられた文書から、TF-IDF 計算に基づいてその内容を表現する主要な単語を選択し、これらをクエリとしてインデックスされた文書を検索することで、類似文書を出力する。

候補記事集合の検索対象は、入力記事よりも発行日付が古い

表 1: 主題・背景関係分類の精度評価

データセット		評価結果
主題記述数	背景記述数	精度平均値
107	121	0.728

ものみに限った。これはユーザが新しいニュースに対して過去の関連記事を検索することを想定したためである。予め十分な数の候補記事を取得するため、取得件数は類似度上位 50 件とした。

5.1.3 事象の抽出と構造化

事象の抽出には、Smith らの手法の Python 実装である PETRARCH^(注6) を利用した。PETRARCH は、自然言語テキストを入力として、その中に記述された事象情報を、日付・ソースエンティティ・ターゲットエンティティ・アクションの組として出力する。エンティティおよびアクションの辞書はライブラリに標準添付されているものを利用した。

提案手法では、異なる記事に出現する事象の同一性を判定する必要がある。実験では、ソースエンティティ・ターゲットエンティティ・アクションが同一であり、かつ日付の日数差が 30 日以内であるような事象を同一事象として判定した。これは、ニュース記事における事象記述の生起日付が曖昧であるような場合を考慮したためである。

5.1.4 事象—記事関係の分類

事象—記事関係の主題あるいは背景への分類には、著者らが [14] において評価のために作成した実装を改良して利用した。SVM の実装は、Python の機械学習ライブラリである scikit-learn^(注7) が提供するものによって提供されているものを用いた。使用した特徴は以下の通りである。

- 出現語
- 出現語の品詞情報
- 出現エンティティ
- 記事本文に対する事象記述の相対的位置
- 出現語の品詞情報
- 出現時間表現の示す日時と記事発行日付の時間的距離

なお、機械学習による分類手法の詳細については [14] で詳細に述べている。このため本稿では、表 1 に簡単な評価結果のみを記載する。

5.1.5 グラフの構築と重要度計算

グラフ上での重要度計算においては、グラフ構造を事象および記事のスコアベクトルと事象—記事関係の隣接行列を構成する必要がある。これらは全て Python の行列演算ライブラリである NumPy^(注8) を用いて実装した。十分に収束した重要度スコアを得るため、重要度計算におけるスコア伝播の最大回数は 100 往復とした。

(注1): <https://www.python.org/>

(注2): <https://catalog.ldc.upenn.edu/LDC2008T19>

(注3): <http://www.nytimes.com/>

(注4): <https://www.elastic.co/products/elasticsearch>

(注5): <https://www.elastic.co/guide/en/elasticsearch/reference/current/query-dsl-mlt-query.html>

(注6): <http://petrarch.readthedocs.org/>

(注7): <http://scikit-learn.org/>

(注8): <http://www.numpy.org/>

表 2: 提案手法の nDCG@10 による評価結果

method	nDCG@10		
	all	short	long
asymmetric@50	0.581	0.623	0.542
multiplexing@50	0.560	0.603	0.519
asymmetric@10	0.690	0.706	0.674
multiplexing@10	0.679	0.691	0.669
baseline (date)	0.616	0.678	0.560
baseline (similarity)	0.667	0.665	0.668

5.1.6 評価

評価では、候補記事集合を類似順にソートしたものを次のようにして再ランキングすることで、6 種類のランキングを作成した。

asymmetric@50, multiplexing@50

上位 50 件を提案手法により計算されたスコア順にソート

asymmetric@10, multiplexing@10

上位 10 件を提案手法により計算されたスコア順にソート

baseline (date), baseline (similarity)

上位 50 件を類似度順または日付順にソート

ランキング評価のための評価指標としては、評価指標は nDCG@10 を採用した。対象記事 1 件に対する評価者数は 3 人とした。評価質問では、提示した記事が、対象記事の理解に関わる次のような情報を総合的にどの程度含んでいるかを 5 段階で評価させた。

- 対象記事で十分に説明されていない事象が
- 対象記事の内容と関係がある事象が
- 比較的重要な事象が

これらの評価実験には、専用に作成した Web インタフェースを用いた。このインタフェースでは、画面左側に対象とする記事のタイトル・日付・本文を表示し、画面右側にランキングされた各記事のタイトル・日付・本文の冒頭部分を表示した。評価者はランキングされた記事に付与されたアイコンをクリックすることで記事全体を閲覧することが可能となっている。実験では評価者に対し、ランキングされた各記事の全体を必ず 1 度以上閲覧することを求めた。

5.2 結果と考察

記事に対する nDCG@10 の評価値の平均を表 2 に示す。

2 より、提案手法によるランキングのうち asymmetric@10 および multiplexing@10 がいずれもベースラインよりも高い評価値を実現していることがわかる。このことから、提案手法の有効性が実証されたと言える。

一方、asymmetric@50 および multiplexing@50 の評価値はベースラインを下回る結果となった。これは、一部の対象記事に対しては関連する記事の全体数が少なく、候補記事集合中に関連度の著しく低い記事が含まれていたため、無関係な事象が

A Solidarity-Warsaw Meeting Is Said to Be 'Close at Hand'

Friday, December 16, 1988

Lech Walesa said today that Solidarity and the Communist Government were close to opening formal talks on Poland's future.

"The round table is very close at hand," the Solidarity leader told a news conference here.

He was referring to a Government invitation in September for the banned union to join negotiations on economic and political changes proposed by the Communist Party.

Mr. Walesa said the Government was moving closer to accepting Solidarity's main condition for attending the talks - publication of an official declaration of intent to consider legalizing the union.

#	asymmetric@50	baseline (similarity)
1	Poland's Bishops Name 5 Mediators in Labor Conflict	Ship Workers Strike, Defying Walesa
2	At Solidarity, A Quiet Stand On Changes	Walesa Sees a Legalized Union
3	Poland Announces December Closing of Lenin Shipyard	Walesa Threatens to Call a Strike Alert

図 4: 提案手法による記事ランキングの例

多く抽出されたためだと考えられる。このことから、提案手法は候補記事集合としてある程度関連度の高い記事のみを取得した場合により有効に機能すると考えられる。

表 2 には、本文の長さに応じて記事集合を 2 つに分割したそれぞれのグループにおける評価値を記載している。表において、グループ short は本文の語数が 100 語未満、グループ long は本文の語数が 100 語以上の記事に対応するグループである。これを見ると、baseline(similarity) では short と long の評価値に大きな差がないのに対し、提案手法によるランキングでは short に対する結果に特に改善が見られる。short に属する記事は、本文が比較的短いためその内容の背景についての説明が少ないような記事であると考えられる。このため提案手法は、背景についての説明が充分でないような記事の理解を支援する上で特に有効であると思われる。

提案手法により有用な記事ランキングが得られた例を図 4 に示す。この記事は、1988 年当時のポーランドにおいて、共産主義政権と対立関係にあった「連帯」の代表ワレサが、政権との対話を表明した記事である。類似度によるランキング結果では、「Walesa」という語が過度に強調された記事が上位となっている。これに対して提案手法では、『「連帯」の違法な抗議行動を共産主義政権が静観」といった、重要な重要な背景事象について述べた記事が上位となっている。

(注 2): <http://www.nytimes.com/1988/12/16/world/a-solidarity-warsaw-meeting-is-said-to-be-close-at-hand.html>

6. おわりに

本稿では、ニュース記事中で記述される事象間の関係の分析に基づき、与えられたニュース記事を理解するのに必要となる過去のニュース記事を検索する手法を提案した。実験では、特定の記事において提案手法に一定の有効性が認められる結果が得られた。

今後の課題としては次のような事柄が挙げられる。まず、提案手法が有効に機能した対象記事が限定的であった原因として、抽出した事象の粒度に関する問題が存在すると考えられる。提案手法では、日付・ソースエンティティ・ターゲットエンティティ・アクションの組として表現される事象を対象としている。また、記事からの事象抽出には、既存研究の実装として提供されているライブラリを使用している。使用したライブラリは、自然言語テキストからの事象抽出に関する多くの先行研究における利用実績が存在している。しかしこのような既存研究において抽出対象とする事象と、多くのニュース記事の中で扱われている事象の間には、特にその粒度においてある程度の差異が存在する場合がある。一般的なニュース記事に記述される事象の多くは、テロ事件や政治家による声明など、日付やエンティティに基づいて同定可能なものである。しかしながら、事象の中には戦争や自然災害など、幾つかのより粒度の小さい事象の集合として表現されるような、より高次的な事象が存在する。通常的事象に加え、これらの高次的な事象について考えた場合、個々の事象間が階層的関係を有していると捉えることが妥当と思われる状況が存在すると考えられる。このような考え方を採用すると、例えば現在の提案手法において別個の事象として抽出されている事象がより高次的な単一の事象に属するような場合、事象の重要度計算において事象から事象への重要度を伝播させることが可能となる。よって、これらの階層的構造を利用することで、記事および事象の関係をより適切にモデル化することができれば、提案手法における記事および事象の重要度計算は大きく改善される可能性がある。今後は事象の粒度についてより深く考察し、提案手法のさらなる改善を目指したい。

その他の課題としては、事象と記事の間の主題・背景関係分類の精度向上や、相互再帰的計算における記事の経過時間による正規化等が挙げられる。また、現在よりも多角的な観点からの評価項目を加え、より大規模な評価実験を行うことも必要と考えられる。

将来の展望としては、研究の成果として得られた知見を、実際のニュース記事閲覧・検索サービスに応用することが考えられる。実験ではニュースアーカイブの記事を用いて評価を行ったが、実際のサービスにおいて要求されるリアルタイム性やインタフェースの観点からも改善を行いたい。

謝辞

本研究の一部は、文科省科研費基盤 (A) 「多元的検索要求に対応できるオンラインデータマイニング検索方式の研究」 (15H01718, 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] S. Brin and L. Page. The anatomy of a large-scale hyper-textual web search engine. *Computer networks and ISDN systems*, 30(1):107–117, 1998.
- [3] 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.
- [4] 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.
- [5] H. Deng, M. R. Lyu, and I. King. A generalized co-hits algorithm and its application to bipartite graphs. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 239–248. ACM, 2009.
- [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. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, 46(5):604–632, 1999.
- [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] S. Tanaka, A. Jatowt, and K. Tanaka. Analysis of mentions to historical events using news archives (in japanese). In *Proceedings of the 6th Forum on Data Engineering and Information Management*, pp. B4–2, 2014.
- [14] S. Tanaka, A. Jatowt, and K. Tanaka. Extraction and analysis of background events for improving comprehensibility of news articles (in japanese). In *Proceedings of the 7th Forum on Data Engineering and Information Management*, pp. F2–6, 2015.
- [15] 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.