

記事アーカイブを用いた歴史的事象の言及分析

田中祥太郎[†] アダムヤトフト^{††} 田中 克己^{††}

[†] 京都大学工学部情報学科 〒 606-8501 京都府京都市左京区吉田本町

^{††} 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 新聞などのニュースメディアでは、日々新たな出来事が報道される。これらの出来事の中には、他の出来事についての記事中で、「歴史上の出来事」として参照されるものがある。しかし、出来事がいつから「歴史上の出来事」として認識されるようになったかを知るのは難しい。そこで本研究では、ニュース記事中の事象への言及に注目し、事象への言及をいくつかのクラスに分類する。この各クラスごとの言及件数の変化を見ることで、事象がいつから歴史的事象として参照されるようになったかを分析する。

キーワード 時系列データマイニング、ニュースアーカイブ、歴史的事象

1. はじめに

ニュースでは、日々新たな出来事が報道される。これらの出来事の中には、長期間にわたって様々な記事で言及され続けるものもある。例として、2001年9月にアメリカで起きた同時多発テロ事件に言及している記事数の推移を図1に示す。これを見ると、事件発生当時や、事件から1年が経過した2002年9月以外にも、2004年3月頃に事件が多く言及されていることがわかる。これには2004年3月にスペインで起きた列車爆破テロ事件や、イラク戦争などが関連していると考えられる。このように、社会に大きな影響を与えた出来事は、他の出来事についての報道の中で言及されることが多い。

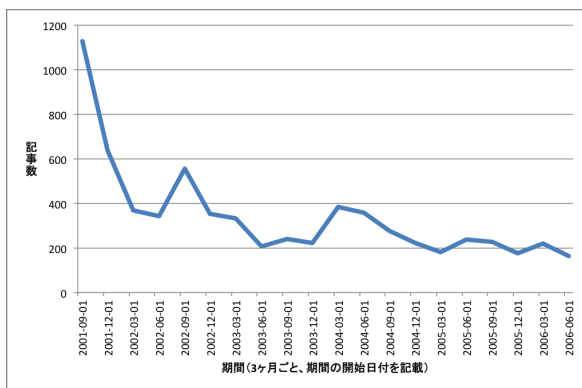


図1 同時多発テロ事件に言及している記事数の推移

ある出来事がその終了後も参照されるのは、現在の社会状況を説明するために過去の出来事が引用されているためであると考えられる。このように過去との関連から現在を説明する行為は、しばしば「歴史に学ぶ」と表現される。20世紀を代表する歴史家の一人である E. H. Carr は、その著書「歴史とは何か」において、「歴史とは現在と過去との対話である」と述べている。このように、歴史を考える上では、過去の客観的事実のみではなく、現在において過去がどのような意味をもつかを考えることが重要である。そこで本研究では、現在を説明する上で重要

な過去の出来事を、「歴史上の出来事」として定義する。過去のある期間に生じた出来事を事象と呼ぶ。また、ある事象がそれ自身と異なる事象についての記述中で言及されている場合、言及されている事象が「歴史上の出来事」すなわち歴史的事象として認識されていると考える。

ある事象が、生じた時から既に歴史的事象として認識されていたとは考えにくい。例として、同時多発テロ事件に言及している2つのニュース記事から抜粋した記述を図2および図3に示す。図2は事件発生直後にその被害状況について述べた記述である。この記述において、事件が歴史的事象として認識されているとはいえない。図3は2004年のスペイン列車爆破テロ事件についての記述である。この記述では、スペインでのテロ事件について述べる上で、過去の類似する事象であるアメリカでのテロ事件を歴史的事象として参照していると言える。このように、事象はその生起から時間が経過するにつれて歴史的事象として認識されるようになると思われる。しかしながら、事象がその生起後どれくらいの期間が経つと歴史的事象として認識されるようになるのかを知るのは難しい。

A Times reporter, Terry Pristin, said that as a result of **the twin tower attacks**, Bellevue Hospital had admitted 30 people for minor injuries, 5 for major injuries, and that one person was dead on arrival. The total included four children, but it was not known if they were among the injured or the dead. (2001-09-11)

図2 同時多発テロ事件に言及している記述1

Whether the bombers came from the Basque terrorist group ETA, as the Spanish government initially presumed, Al Qaeda or elsewhere, comparisons to **the attacks of Sept. 11, 2001**, are inevitable and appropriate. Neither Spain nor America stand alone. (2004-03-11)

図3 同時多発テロ事件に言及している記述2

そこで本研究では、ニュースにおける事象への言及に注目し、

その特徴の時間的変化を見ることで、事象が歴史的な事象として認識されるようになったかを分析する。本研究では、広く知られた名前をもつような事象を扱い、その名前を含むニュース記事の一部を事象への言及として扱う。事象が歴史的な事象として認識されているかを判定するため、事象への言及をいくつかのクラスに分類することを考える。この分類に寄与するような事象への言及の特徴を数値化し、計算された特徴量を用いて機械学習による分類を行う手法を提案する。

実験では、過去に実際に生じた事象の中から、ニュースにおいて特に多く言及されている事象を取り上げる。注目する事象に対する言及をニュースアーカイブから抽出し、提案手法を用いてこれらを分類し、クラスごとの言及件数の変化を見る、ニュースアーカイブから事象への言及を抽出する際には、一つの事象に対して複数の名称が存在する場合を考慮し、Web上の情報を利用して事象に対する名称を収集する。

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、事象とそれへの言及、および言及中の特徴語の定義について述べる。第4章では、事象への言及の分類と、分類に利用する言及の特徴について述べる。第5章では、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

本章では、本研究と関連のある研究を紹介し、本研究の特徴を明確にする。

2.1 時系列文書データからの知識抽出に関する研究

Michel ら [6] 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。彼らは作成したコーパスを利用して1800年から2000年の間に発生した言語や文化に関わる現象について大規模な調査を行い、その調査結果が様々な分野で応用できることを示した。

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。彼らはニュース記事中に現れる人名を抽出し、ある話題中である人物に対する注目が増加している期間、およびニュース全体で特定の人物に対する注目が持続している期間を測る手法を提案した。

Yang ら [7] 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。彼らは出来事をキーワードの集合によって表現し、ニュース記事の内容およびタイムスタンプを利用して記事集合から明示的な名称が存在しない出来事を発見できることを示した。

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。彼らは出来事をニュース記事の集合によって表現し、新たに発行されたニュース記事が、既知の出来事についてのものか、または未知の新たな出来事についてのものかを判定した。

[6] および [3] は、大規模な時系列文書アーカイブを用いてあるトピックに関する長期的な特徴変化を分析している点で本研究と共通するが、必ずしも個々の出来事に着目していない点では本研究と異なる。また [7] および [1] は、ニュース記事中の出来事を何らかの形で表現し、出来事についての情報を抽出している点で本研究と共通するが、出来事への明示的な言及を扱っていない点では本研究と異なる。

2.2 コンピュータによる歴史学に関する研究

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。彼らは数か国のメディアによって発行された記事を含む大規模なニュースアーカイブを利用し、統計的な手法を適用することで、社会的記憶の時間的および地理的な特徴を抽出した。

高橋ら [8] は、Web 百科事典 Wikipedia の中から歴史的観点から重要な項目を発見する手法を提案した。彼らは歴史上の出来事や人物について、その時間的または地理的な重要度を、Wikipedia のリンク構造に基づいて反復計算アルゴリズムを適用することで計算した。

Jatowt ら [5] は、時間情報をもつ文書に対し、その文書の内容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。彼らは文書中に現れる時間属性をもつキーワードを用いて、注目時間が明示されていない文書に対してもその注目時間を推定できることを示した。

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。彼らは歴史上の出来事についての記述の特徴が、その出来事が起こった直後に書かれた文書と、時間が経過した後にその出来事を振り返って書かれた文書ではどのように違うかについて考察した。

[2] および [8] は、歴史に関わるデータを用いて歴史上の出来事や人物を分析するという点で本研究と共通するが、出来事への言及の意味的特徴を利用していない点では本研究と異なる。また [5] および [4] は、文書の時間的または歴史的な特徴に注目している点で本研究と共通するが、特定の出来事について大量の文書集合を用いた定量的な分析を行っていない点では本研究と異なる。

3. 事象とそれへの言及

本章では、本研究における事象とそれへの言及、および言及中の特徴語の定義について述べる。

3.1 事象

本研究では、過去のある期間に生じた出来事を事象として定義する。

事象は人物や場所などと異なり、実体を持たないため、ある事象と他の事象の区別が難しい場合がある。そこで本研究では、事象への明示的な言及に注目する。また、本研究では事象への言及の特徴の時間的な推移を分析するため、事象の生起期間を知ることができないなければならない。

以上のような理由から、本研究では次の2つの条件をみたすような事象を扱う。

- ・ 事象に対する一般に知られた名称が存在する
- ・ 事象の生起期間が既知である

ここで事象の名称とは、事象を指す単語または単語の集合をいう。また事象の生起期間とは、事象が実際に起こっていた期間を指し、出来事の名称がニュースに現れる期間を指すわけではない。例えば地震などのように、ごく短時間の間のみ生起していた事象については、その事象が生起した日のみをその生起期間として扱う。

事象の名称は、必ず一意であるとは限らない。ある事象に対して複数の異なる名称が存在する場合、それらの名称全てをその事象に対する名称とする。事象に対する名称を収集する方法については第5章で詳しく述べる。

表1に本研究で扱うような事象の例を、その名称および生起期間とともに示す。

表1 事象とその名称および生起期間の例

事象	生起期間	名称
同時多発テロ事件	2001-09-11	September 11 attacks, 2001 terrorist attacks, Terrorist attacks on New York and Washington
アトランタ ・オリンピック	1996-07-19 - 1996-08-04	Atlanta Olympics, Atlanta games, 1996 Summer Olympics
スマトラ地震	2004-12-26	Sumatra earthquake, Boxing Day earthquake, 2004 Indian Ocean earthquake

3.2 事象への言及

本研究では、ある事象の名称を含む記述を、その事象への言及として定義する。

事象への言及は様々な文書に現れるが、本研究では特にニュース記事における事象への言及を対象とする。ニュース記事の本文が事象の名称を含む場合、その本文のうち事象の名称を含む一部分を、その事象への言及として扱う。

ここでニュース記事本文全体ではなくその一部のみを対象とするのは、次のような理由による。まず、記事ごとに本文の長さが異なるため、本文全体を言及として抽出した場合、言及の粒度にばらつきが生じてしまう。また、複数のトピックを含むような記事の場合、事象の名称が現れる箇所から離れた位置にある内容は、言及されている事象と無関係である可能性が高い。これらの理由を考慮し、ニュース記事本文から事象の名称を含む一部分を抽出し、その事象への言及とする。また、ひとつの記事に事象の名称が複数回出現する場合、各部分を個別の言及として扱う。

具体的にどの程度の長さの記述を事象への言及とするかについては第5章で詳しく述べる。

3.3 言及中の特徴語

本研究では、事象への言及の分類を考える。この分類には、事象への言及の特徴を利用する。特徴量の計算には、言及中の特徴語を用いる。ここでは、特徴量の計算に利用する時間表現、

固有名詞、および普通名詞について述べる。

(1) 時間表現

時間表現とは、事象への言及中に現れる日付または期間を指す語の中で、特定の日付または期間を指すものである。時間表現には、例えば“2001” “Sept. 11” “last week”といった語が挙げられる。時間表現には、“every day”のように特定の期間を指さない語は含まない。また、“winter”のように複数の連続しない日付または期間を指す語も含まない。

(2) 固有名詞

固有名詞は、事象への言及中に現れる、特定の人物、組織、場所などのエンティティを指す語である。固有名詞には、例えば“George W. Bush”, “C.I.A.”, “New York”といった語が挙げられる。

(3) 普通名詞

普通名詞は、事象への言及中に現れる固有名詞以外の名詞のうち、ストップワードを除いたものである。

4. 事象への言及の特徴と分類

本章では、事象への言及の分類と、分類に利用する言及の特徴について述べる。

4.1 言及の分類

ニュースにおいて、過去に生起した事象が言及される場合、その理由には様々なものが考えられる。例えば、社会的に大きな影響を与えた事件や、被害の大きい自然災害などは、その発生からちょうど1年が経過した節目の時期に言及されることが多い。また、殺人事件の容疑者が逮捕された場合や、その裁判が行われる場合には、その事件が多く言及される。他に、オリンピックなどの定期的に発生する事象や、地震など同種の事象がいくつもある事象は、類似する事象や同種の事象が発生するたびに言及される。

このように、ニュースにおいて過去に発生した事象が言及されているとき、その言及の内容の主題となっている事象が必ずしも言及されている事象そのものであるとは限らない。ここで、言及されている事象を言及事象、言及の主題となっている事象を主題事象と呼ぶ。ニュースにおいて事象が言及される様々な理由は、言及事象と主題事象との関係に基づいて分類することができる。

例えば同時多発テロ事件を事象 e_1 、イラク戦争を事象 e_2 、ロンドンでのバス爆破テロ事件を事象 e_3 とする。また、 e_1 への言及 r_1, r_2, r_3 を、それぞれ同時多発テロ事件についての記述、イラク戦争についての記述、ロンドンでのバス爆破テロ事件についての記述とする。すなわち、 r_1, r_2, r_3 の言及事象は全て e_1 であり、主題事象はそれぞれ e_1, e_2, e_3 である。

r_1 において e_1 が言及されている理由は、 e_1 自体について述べるためである。また、 r_2 において e_1 が言及されている理由は、 r_2 の主題事象である e_2 と e_1 の間に因果関係が存在するためである。さらに、 r_3 において e_1 が言及されている理由は、 r_3 の主題事象である e_2 と e_1 が類似するためである。

図4, 5, 6に言及 r_1, r_2, r_3 の例を示す。また、図7に言及と事象の関係を示す。

The smooth banking of the second plane to strike the towers supports this point of view, Mr. Nance said. He added that precisely controlling a large jet near the ground, necessary for **the Pentagon attack**, also required advanced skill. “There’s no way an amateur could have, with any degree of reliability, done what was done yesterday,” Mr. Nance said. (2001-09-13)

図4 言及 r_1 (事象自体の記述) の例

Senior intelligence and law enforcement officials say that they have found no evidence to prove that Iraq had any involvement in **the Sept. 11 attacks**. They add that they have not received any corroborating evidence to support an initial report from Czech intelligence that Mohamed Atta, one of the Sept. 11 hijackers, met an Iraqi intelligence officer in Prague in April 2001. (2002-09-10)

図5 言及 r_2 (因果関係のある事象からの参照) の例

British officials have been forecasting a major terror strike on London since **the Sept. 11 attacks**, particularly since Britain sided with the United States in the war in Iraq. Mr. Blair, speaking before he left Scotland for London, said, “Just as it is reasonably clear that this is a terrorist attack or a series of terrorist attacks, it is clear that it is designed and aimed to coincide with the opening of the G8. (2005-07-08)

図6 言及 r_3 (類似する事象からの参照) の例

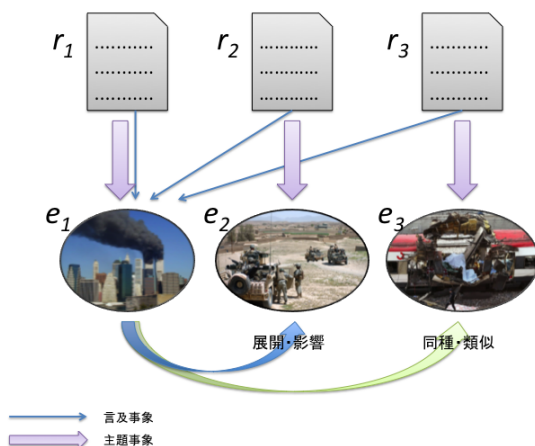


図7 言及と事象との関係

言及における言及事象と主題事象との関係を考えることで、事象への言及はいくつかのクラスに分類される。第1章で述べたように、過去の事象が現在起きている事象を説明するために引用されている場合、その事象は歴史的事象として社会に認識されていると考えられる。よって、言及事象と主題事象が異なるような言及は、言及事象が歴史的事象として認識されていることを示す手がかりとなる。

先に例として挙げた分類は「事象自体の記述」「因果関係のある事象」「類似する事象からの参照」であり、ニュースにおいてはこの3つのクラスのいずれかに分類される言及が特に多く

見られる。「事象自体の記述」は、単に事象自体の報道中で事象名が現れているだけであり、言及事象が歴史的事象としての参照されているとは言い難い。「因果関係のある事象からの参照」は、同一ではないが依存する事象についての報道中での参照であり、「歴史的事象としての参照」である可能性がある。「類似する事象からの参照」は、過去の事例との対比のための参照であり、「歴史に学ぶ」という見方から、「歴史的事象としての参照」であると言える。この各クラスの言及数の時間的な推移を調べれば、ニュースにおいて何度も言及される事象については、その事象がいつから歴史的事象として認識されるようになったかを推測することに役立つと考えられる。以降ではこれらの3つのクラスと、そのいずれにも属しない「その他」の計4つのクラスについて考える。

4.2 事象への言及の特徴

事象への言及の分類を困難にしている理由の一つに、言及における主題事象の特定が難しいことが挙げられる。例えば同時多発テロ事件への言及を考えた場合、言及における主題事象はイラク戦争やロンドンでのバス爆破テロ事件である場合がある。しかし、これらの主題事象の名称が記事中に必ず現れるとは限らない。また、主題事象の名称が記事中に現れる場合でも、事象間の関係をその名称のみから判定することは難しい。

この問題に対処するため、本研究では言及事象と主題事象の関係の推定に有効と考えられるいくつかの特徴を定義する。これらの特徴に対する特徴量を計算によって求め、これを用いて事象への言及を機械学習により分類することを試みる。以下に本研究で考える特徴について述べる。

• 時間表現と事象の関連度

言及中に現れる“last year”などの時間表現に対し、その時間表現が指す時間軸上の期間が存在する。すなわち、ある時間表現 w に対し、その時間表現の注目期間 $[t_w^{\text{begin}}, t_w^{\text{end}}]$ が存在する。また第3章で述べたとおり、事象 e に対してその生起期間 $[t_e^{\text{begin}}, t_e^{\text{end}}]$ が存在する。このように考えた場合の w の注目期間と e の生起期間の近さを、 w と e の関連度と定義する。

事象 e への言及 r において、 r 中の各時間表現と e との関連度が高いとき、 r における言及事象と主題事象の関連度が高いと考えられる。

• 固有名詞と事象の関連度

言及中に現れる“New York”などの固有名詞は、人物・組織・場所などのエンティティを指す。ある固有名詞 w に対し、 w と事象 e との関連度を考えることができる。例えば同時多発テロ事件を考えた場合、事件の発生場所の一つを指す固有名詞“New York”は、“Japan”などの固有名詞に比べて事件との関連度が高いと考えられる。

事象 e への言及 r において、 r 中の各固有名詞と e との関連度が高いとき、 r における言及事象と主題事象の関連度が高いと考えられる。

• 普通名詞と事象の関連度

言及中に現れる“terror”などの普通名詞は、言及がどのような属性を持つ事象について述べているのかを表す。ある普通名詞 w に対し、固有名詞の場合と同様に w と事象 e との関連度

を考えることができる。例えば同時多発テロ事件を考えた場合，“terror”などの語は事件との関連度が高いと考えられる。

事象 e への言及 r において、 r 中の各普通名詞と e との関連度が高いとき、 r における言及事象と主題事象の関連度が高いと考えられる。

- 記事日付と事象生起期間の近さ

事象 e への言及 r について、 r が含まれる記事の発行日付と事象 e の生起期間の時間的距離を定義することができる。この距離を用いて記事日付と事象生起期間の近さを計算し、特徴量として採用する。

- 記事における事象への言及の出現位置

事象 e への言及 r について、記事本文中での r の相対的な出現位置を定義することができる。記事における事象への言及の出現位置は、記事における言及事象の扱いを知り手がかりであり、事象への言及の性質を示す特徴であると考えられるため、この値を特徴量として採用する。

4.3 言及の特徴量の計算と分類への利用

言及の特徴量の計算には、言及中の特徴語を用いる。

e を事象とする。 e への言及全体の集合を R_e とする。 $r \in R_e$ に含まれる時間表現の集合、固有名詞の集合、普通名詞の集合を、それぞれ $TimeExpressionSet_r$, $ProperNounSet_r$, $CommonNounSet_r$ とする。また、ニュース記事全体の集合を D とし、ニュース記事 $d \in D$ の本文中に単語 w が含まれることを $w \in d$ と表す。

時間軸上の e の生起期間を、 $[t_e^{\text{begin}}, t_e^{\text{end}}]$ と表す。

- 普通名詞および固有名詞と事象の関連度の計算

ある普通名詞または固有名詞 w と、事象 e との関連度 $\text{Rel}_{\text{word}}(w, e)$ は次のように計算する。

e への言及の全体において、語 w が出現する回数が多いほど、 w と事象 e との関連度は高いと考えられる。また、 w のニュース記事全体における出現確率が低いほど、 w と事象 e との関連度は高いと考えられる。

そこで、 $\text{Rel}_{\text{word}}(w, e)$ の計算には tf-idf の考え方を利用する。tf-idf は、単語 i の文書 j における重みを、文書 j における単語の出現頻度と、全文書集合における単語 i の逆文書頻度との積として計算する手法である。この考え方に基づき、語 w と事象 e との関連度 $\text{Rel}_{\text{word}}(w, e)$ を、事象への言及全体の集合 R_e をひとつの文書として考えた場合の語 w の tf-idf 値とする。

すなわち、

$$\text{Rel}_{\text{word}}(w, e) = \text{tf}_{w, R_e} \cdot \text{idf}_w$$

$$\text{tf}_{w, R_e} = \frac{\sum_{r \in R_e} n_{w, r}}{\sum_{w \in W} \sum_{r \in R_e} n_{w, r}}$$

$$\text{idf}_w = \log \frac{|D|}{|\{d | w \in d\}|}$$

ここで、 $n_{w, r}$ は言及 r における語 w の出現回数である。また、 W は各言及から抽出された普通名詞全体の集合または固有名詞全体の集合である。

- 時間表現と事象の関連度の計算

ある時間表現 w と e の関連度 $\text{Rel}_{\text{time}}(w, e)$ は次のように計算する。

w の指す時間軸上の期間を $[t_w^{\text{begin}}, t_w^{\text{end}}]$ とする。 e の生起期間と w の指す時間軸上の期間との距離が近いほど、時間表現 w と e との関連度は高いと考えられる。そこで、 w と事象 e との関連度 $\text{Rel}_{\text{time}}(w, e)$ を、 e の生起期間と w の指す時間軸上の期間との平均距離の逆数とする。

すなわち、

$$\text{Rel}_{\text{time}}(w, e) = \frac{1}{|t_w - t_e|}$$

$$t_w = \frac{t_w^{\text{begin}} + t_w^{\text{end}}}{2}$$

$$t_e = \frac{t_e^{\text{begin}} + t_e^{\text{end}}}{2}$$

- 事象への言及の各関連度の計算

事象 e と言及 $r \in R_e$ との各関連度は次のように計算する。ここで、 r 中の時間表現に基づく e と r の関連度、 r 中の固有名詞に基づく e と r の関連度、 r 中の普通名詞に基づく e と r の関連度を、それぞれ時間類似度 $\text{TimeSimilarity}(e, r)$ 、エンティティ類似度 $\text{EntitySimilarity}(e, r)$ 、属性類似度 $\text{AttributeSimilarity}(e, r)$ と呼ぶ。これは、これらの各特徴量が、言及中の言及事象と主題事象の類似性を表すと考えられるためである。

言及中に事象 e との関連度が高い語が多く含まれているほど、各類似度は高くなるべきである。よって、まずは言及中の各特徴語の関連度の合計によって各類似度を計算することが考えられる。しかしこの方法では言及の語数にばらつきがある場合、語数の多い言及の類似度が高くなってしまふ。また、各特徴語の関連度の平均値を用いた場合、関連度の高い特徴語が1語だけ含まれているような言及の類似度が高くなってしまふ。そこで、各類似度は、各特徴語の関連度の合計を、各特徴語の数の対数で割った値とする。

すなわち、

$$\text{TimeSimilarity}(e, r)$$

$$= \text{Sim}(e, \text{TimeExpressionSet}_r, \text{Rel}_{\text{time}})$$

$$\text{EntitySimilarity}(e, r)$$

$$= \text{Sim}(e, \text{ProperNounSet}_r, \text{Rel}_{\text{word}})$$

$$\text{AttributeSimilarity}(e, r)$$

$$= \text{Sim}(e, \text{CommonNounSet}_r, \text{Rel}_{\text{word}})$$

$$\text{Sim}(e, S, f) = \frac{\sum_{w \in S} f(w, e)}{\log(1 + |S|)}$$

- 記事日付と事象生起期間の近さ

言及 r を含む記事の日付と事象生起期間の近さは、 r を含む記事の発行日付と事象生起期間の平均距離の逆数とする。すな

わち、言及 r を含む記事の発行日付を t_r とすると、記事日付と事象生起期間の近さは、

$$\frac{1}{|t_r - t_e|}$$

と計算する。

- 記事における事象への言及の出現位置

言及 r を含む記事の本文における r の出現位置は、本文において r が現れるまでに存在する文字数を、本文全体の文字数で割ったものとする。すなわち、 r が現れるまでに存在する文字数を c_r 、本文全体の文字数を c とすると、記事における事象への言及の出現位置は、

$$\frac{c_r}{c}$$

と計算する。

- 言及の各クラスへの分類

言及の各クラスへの分類には、決定木学習を利用する。分類器の中でも、3 クラス以上の多クラス分類の場合、決定木学習は SVM などの線形分類器に比べて精度が高いことが経験的に知られている。

決定木学習アルゴリズムは、ID3 を用いる。ID3 は、正解データをグループに分割する規則を選択する際、分割後のエントロピーが最も小さくなる規則を優先的に採用することで決定木を構成するアルゴリズムである。

5. 実 験

本章では、提案手法を用いて行った実験の概要と結果を示し、結果に対する考察を述べる。

5.1 概 要

実験は以下の手順で行った。実験に用いたプログラムは、プログラミング言語 Python^(注1) を用いて記述した。ニュース記事本文の文分割などの基本的な自然言語処理には、自然言語処理ライブラリ Natural Language Toolkit^(注2) を用いた。

(1) 対象とする事象の選出

1987 年から 2007 年の間に起きた事象の中から、ニュースで特に多く言及されている次の事象を実験の対象候補として選んだ。

- ベルリンの壁崩壊（1989 年、東西ドイツ）^(注3)
- アトランタ・オリンピック（1996 年、アメリカ）^(注4)
- 同時多発テロ事件（2001 年、アメリカ）^(注5)
- ハリケーン・カトリナ（2005 年、アメリカ）^(注6)

(2) 事象の生起期間および名称の収集

事象の名称および生起期間など、事象に関する基本的な情報の取得には、オンライン百科事典 Wikipedia 英語版^(注7) を用

いた。

各事象の生起期間は、Wikipedia における各事象の項目ページの Infobox^(注8) の情報を利用して特定した。また、各事象の項目ページにリダイレクトしている項目名を各事象の名称の候補として使用した。これらの項目名から、事象を指すと考えられる項目名を手作業によって選び、さらにニュースアーカイブから項目名を本文中に含む記事が 10 件以上見つかったものを名称として採用した。

(3) ニュースアーカイブからの事象への言及の収集

ニュース記事の収集には、The New York Times Annotated Corpus^(注9) を使用した。The New York Times Annotated Corpus は、アメリカ合衆国における主要な日刊新聞紙のひとつである The New York Times^(注10) によって提供されているニュースアーカイブであり、1987 年から 2007 年の間に発行された 180 万件以上のニュース記事を含んでいる。

実験では、各事象の生起から 5 年以内に発行された記事を対象とした。ただし、ハリケーン・カトリナについては、ニュースアーカイブに 2008 年以降の記事が含まれていないことから、事象の生起から 2 年以内に発行された記事を対象とした。

実験では、まず対象期間において事象の名称を含む記事を全て収集した。さらに、各記事を文に分割し、事象への名称を含む文とその前後 1 文を事象への言及として抽出した。

(4) 言及中の特徴語の抽出

言及からの特徴語抽出には、The Stanford Natural Language Processing Group^(注11) によって提供されている自然言語処理ライブラリを用いた。

時間表現および固有名詞の抽出には、Stanford Named Entity Recognizer (Stanford NER)^(注12) を用いた。Stanford NER によって抽出された固有表現から、日付・期間を示す“DATE”タグが付与された語を時間表現として、人物・組織・場所・その他の固有名詞を示す“PERSON”、“ORGANIZATION”、“LOCATION”、“MISC”タグが付与された語を固有名詞としてそれぞれ抽出した。また、“yesterday”などの相対的な時間表現を、言及を含む記事の発行日付を基準とした絶対時間に正規化するため、Stanford Temporal Tagger^(注13) を用いた。

普通名詞の抽出には Stanford Log-linear Part-Of-Speech Tagger (Stanford POS Tagger)^(注14) を用いた。Stanford POS Tagger によって、普通名詞を表す“NN”、“NNS”のタグが付与された語のうち、ストップワードを除いたものを普通名詞として抽出した。

(5) 各特徴量の計算

各特徴語の言及事象との関連度および言及の各特徴量の計算は、第 4 章において述べた手法に基づいて行った。

(注1) : <http://www.python.org/>

(注2) : <http://nltk.org/>

(注3) : http://en.wikipedia.org/wiki/Berlin_Wall

(注4) : http://en.wikipedia.org/wiki/1996_Summer_Olympics

(注5) : http://en.wikipedia.org/wiki/September_11_attacks

(注6) : http://en.wikipedia.org/wiki/Hurricane_Katrina

(注7) : <http://en.wikipedia.org/>

(注8) : <http://en.wikipedia.org/wiki/Help:Infobox>

(注9) : <http://catalog.ldc.upenn.edu/LDC2008T19>

(注10) : <http://www.nytimes.com/>

(注11) : <http://nlp.stanford.edu/>

(注12) : <http://nlp.stanford.edu/software/CRF-NER.shtml>

(注13) : <http://nlp.stanford.edu/software/sutime.shtml>

(注14) : <http://nlp.stanford.edu/software/tagger.shtml>

表 2 各事象の生起期間と記事収集期間

事象	生起期間	記事収集期間
ベルリンの壁崩壊	1989-11-09 - 1989-11-10	1989-11-01 - 1994-10-30
アトランタ・オリンピック	1996-07-19 - 1996-08-04	1996-07-01 - 2001-06-30
同時多発テロ事件	2001-09-11 - 2001-09-11	2001-09-01 - 2006-08-31
ハリケーン・カトリナ	2005-08-23 - 2005-08-31	2005-08-01 - 2007-07-30

表 3 各事象の言及数

事象	言及数
ベルリンの壁崩壊	2637
アトランタ・オリンピック	838
同時多発テロ事件	8522
ハリケーン・カトリナ	4683

表 4 同時多発テロ事件の名称と言及数

名称	言及数
sept 11 attacks	6648
9 11 attacks	598
2001 terrorist attacks	372
sept 11 2001 terrorist attacks	188
9 11 attack	117
terrorist attacks on the world trade center and the pentagon	113
attacks of 9 11	112
september 2001 attacks	90
9 11 terrorist attacks	81
terrorist attacks on new york and washington	68
sept 11 disaster	53
terrorist attacks of september 2001	28
pentagon attack	19
9 11 terror attacks	12
twin towers attack	12
september 11 attacks	11
計	8522

(6) 言及の分類とその評価

第 4 章で述べたとおり、言及の各クラスへの分類には決定木学習を用いた。まず手作業で言及の分類を行うことで、各クラスにつき 30 件、合計 120 件の正解データを作成した。次に正解データを訓練データと検証データに 1:1 の割合で分割し、訓練データによって得られた決定木を検証データによって評価した。この操作を 100 回行い、得られた中で最も精度の高い決定木を用いて正解データを分類した。また、正解データを用いて 5 分割交差検定を 100 回行い、分類の精度を評価した。決定木学習と交差検定には Python 上で動作する機械学習ライブラリである scikit-learn^(注15)を用いた。

5.2 結 果

表 2 に対象とした各事象の生起期間と記事収集期間を示す。

表 3 に対象とした各事象について抽出できた言及の数を示す。

表 4 に同時多発テロ事件の名称と、名称ごとの言及数を示す。

表 5 に同時多発テロ事件への言及中の各特徴語の平均出現数を示す。

表 6 に同時多発テロ事件との関連度が高い固有名詞および普通名詞を示す。

図 8 に、同時多発テロ事件を対象事象とした場合の分類され

表 5 言及中の各特徴語の平均出現数

特徴語	平均出現数 (個/言及)
時間表現	1.997
固有名詞 (人物)	1.753
固有名詞 (組織)	0.969
固有名詞 (場所)	1.472
固有名詞 (その他)	0.824
固有名詞合計	5.019
普通名詞	16.29

表 6 関連度の高い固有名詞および普通名詞

固有名詞	関連度	普通名詞	関連度
bush	0.428	attack	0.288
american	0.350	official	0.055
f b i	0.318	intelligence	0.042
united states	0.316	terrorism	0.031
world trade center	0.310	hijacker	0.026
al qaeda	0.298	commission	0.025
moussaoui	0.289	government	0.025
iraq	0.282	administration	0.024
congress	0.282	security	0.024
afghanistan	0.259	report	0.021

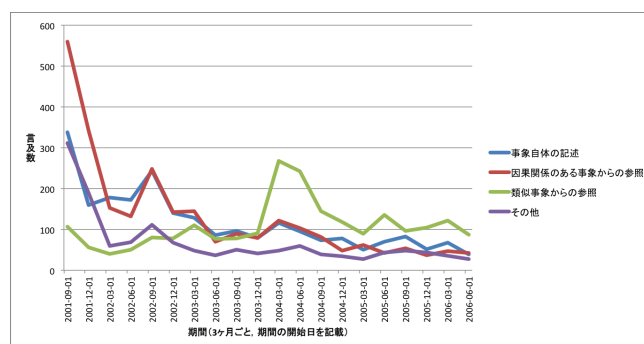


図 8 分類された言及の数の時間的推移 (同時多発テロ事件)

表 7 最良の決定木による正解データの分類結果

正解\決定木	C1	C2	C3	C4
C1: 事象自体の記述	20	6	4	0
C2: 因果関係のある事象からの参照	2	21	4	3
C3: 類似する事象からの参照	0	3	27	0
C4: その他	6	5	1	18

Accuracy = 0.716

表 8 交差検定による評価結果

平均精度	標準偏差
0.428	0.148

た言及の数の時間的推移を示す。

表 7 に最良の決定木による正解データの分類結果を示す。

表 8 に交差検定による評価結果を示す。

5.3 考 察

- 対象事象の選択

表 3 を見ると、候補とした 4 つの事象のうち、同時多発テロ事件が最も言及数が多い。このため、今回は同時多発テロ事件

(注15) : <http://scikit-learn.org/>

を対象とし、特に詳細な分析を行った。

- 事象への言及の抽出の網羅性

表4を見ると、事象に対する複数の名称を考慮することで、発見できる記事の数が2割以上増加することがわかる。よって、Wikipediaの項目名のリダイレクトを用いた事象の名称収集は、事象への言及を発見する際の網羅性を向上させていると言える。

- 事象への言及中の特徴語数

表5を見ると、事象への言及における特徴語の平均出現数は、普通名詞が16.29個、固有名詞が人物・組織・場所・その他の合計で5.019個、時間表現が1.997個となっている。普通名詞および固有名詞については十分な数と言えるが、時間表現については十分ではない可能性がある。

- 言及中の特徴語の関連度計算

表6を見ると、提案手法により普通名詞および固有名詞の関連度を計算した場合、実際に事象との関連性が高いと考えられる語が上位となっていることがわかる。よって、提案手法による普通名詞および固有名詞の関連度計算は有効であると言える。

- 分類された言及の数の時間的推移

図8を見ると、言及事象として同時多発テロ事件を選んだ場合、事象自体の記述、因果関係のある事象からの参照、その他の分類が同様に時間経過による減少を示している。これに対し、類似する事象からの参照は時間経過により微増しており、また2004年頃にピークが見られる。後述の通り、類似する事象からの参照については比較的高い識別精度を実現している。このため、類似する事象からの参照というケースに注目した場合、2004年頃から事件を歴史的事象として参照することが増えていくと推測される。また、2004年3月期～6月期にピークが見られるのは、2004年3月11日に起きたスペインでの列車爆破事件やイラク戦争についての記事の中で、同時多発テロ事件が歴史的事象として参照されているからと考えられる。

- 言及の分類の精度

表7を見ると、最良の決定木を用いた場合、類似する事象からの参照については比較的高い分類精度が得られているが、それ以外のクラスにおける分類精度は高いとはいえない。また表8を見ると、平均精度は0.428と低い。さらに標準偏差は0.148となっており、正解データからの訓練データの選び方によって得られる精度にばらつきが見られる。

これらの結果から、時間表現と事象の関連度、固有名詞と事象の関連度、普通名詞と事象の関連度、記事日付と事象生起期間の近さ、記事における事象への言及の出現位置という5種類の特徴を用い、決定木によって言及の分類を行った場合、一部のクラスにおいては満足できる精度が実現できないことがわかる。原因としては、まず特徴の数が十分ではない可能性が考えられる。今回は先に述べた5つの特徴を利用したが、これは決して十分な種類数とはいえず、これらの特徴のみでは正確な分類ができないと考えられる。また先に述べたように、言及中の時間表現の平均数が少なく、時間類似度の計算が正確でないことも要因として考えられる。これにより、特に事象自体の記述と因果関係のある事象からの参照の判別が正確でなかった可能性がある。類似する事象からの参照において比較的精度が高

かったのは、言及中から抽出できた名詞の数が比較的多く、属性類似度の計算が他の特徴量の計算に比べて正確であったからと考えられる。さらに、正解セットの件数が120件と少ないため、機械学習の精度が下がってしまったことも考えられる。

6. おわりに

本稿では、過去に生じた事象が、それ自身と異なる事象についてのニュースの中で言及されることに着目し、事象がいつから歴史的事象として参照されるようになるのかを分析した。ニュースにおける事象への言及を分類する手法を提案し、過去に実際に発生した事象を取り上げて実験を行った。提案手法を評価した結果、クラスによって分類精度に差が見られることがわかった。

今後の課題としては、分類精度の向上のため、今回考案した以外の特徴を用いることが挙げられる。また、より有用なクラス設定を考案することも必要である。

謝辞 本研究の一部は、文科省科研費基盤(A)「ウェ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(さががけ)「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] 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] 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.
- [6] 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.
- [7] 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.
- [8] 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.