

Wikipedia を用いた事象の日付情報の推定

島田 裕司[†] 浅野 泰仁[†] 吉川 正俊[†]

[†] 京都大学大学院情報学研究科

〒 606-8501 京都市左京区吉田本町

E-mail: †h.shimada@db.soc.i.kyoto-u.ac.jp, ††{asano,yoshikawa}@i.kyoto-u.ac.jp

あらまし Wikipedia などの誰でも容易に記述できる Web 上の情報は、誤った記述が含まれていたり、記述が不十分であることがある。特に日付情報に関する不整合、誤りは多く、気づかれづらい。このような不整合を自動的に発見するためには、各事象の発生した日時を推定することが必要である。「2009/2/20」「翌日」「2 日後」「3 月 4 日」など、日付が明示的に記述された事象の発生した年月日を推定するといった研究はなされている。本研究では、Wikipedia の百科事典としての性質、および文書構造を用いることによって、日付が明示的に記述されていない事象について発生した日時を推定することを目的とする。

キーワード 情報抽出

Estimating date information of events in Wikipedia

Hiroshi SHIMADA[†], Yasuhito ASANO[†], and Masatoshi YOSHIKAWA[†]

[†] Graduate School of Informatics, Kyoto University

Yoshida-Honmachi, Sakyo-ku, Kyoto, 606-8501 Japan

E-mail: †h.shimada@db.soc.i.kyoto-u.ac.jp, ††{asano,yoshikawa}@i.kyoto-u.ac.jp

Abstract Everyone can easily describe contents on the web, like Wikipedia. Such contents include wrong information or lacked information. Especially there are a lot of inconsistencies of date information, and they are hard to be noticed by humans. To find out these inconsistencies automatically, we require to estimate the date information of event. There are several researches for estimating the date information of events whose date expressions are described explicitly, such as "2009/2/20", "the next day", "2 days later", or "March 4th". We propose a method for estimating the date information of events which whose date expressions are not described explicitly.

Key words Information Extraction

1. はじめに

現在、Wikipedia をはじめとする Web2.0 系コンテンツの普及により、個人でも容易に情報を発信することが可能になっている。これらの情報の量は膨大であり、有用なものも多く存在するが、誤った情報や、不十分な情報が含まれていることも多く、コンテンツの信憑性には問題がある。しかし、2008 年 9 月に 1000 人を対象に行ったアンケート調査^(注1)では、Wikipedia の記事について約 8 割のユーザが 6 割以上の記事を信用できると回答し、約 6 割のユーザが 8 割以上の記事を信用できると回答しており、コンテンツを鵜呑みにしてしまっている可能性がある。特に日付情報に関する不整合、誤りは多く、気づかれづらい。例えば、Wikipedia の 2009 年 1 月 8 日現在の「安倍

晋三」の記事では、首相就任後 2007 年 5 月初旬に初訪米との記述があるが、実際には安倍晋三は 2006 年 9 月 26 日に首相に就任し、2007 年 4 月に訪米している。また、同じ記事内で、ディック・チェイニー米副大統領の来日が 2007 年 2 月 12 日となっているが、実際には来日したのは 2007 年 2 月 20 日である。さらに 7 月 3 日に久間防衛相の原爆投下をめぐる発言が問題化との記述もあるが、実際に発言したのは 6 月 30 日であり、7 月 1 日には本人がテレビの報道番組で発言に関する批判意見に対して反論していることから、すでに問題化しているととれる。また、完全な誤りではないが、読者に誤解を生じさせる可能性のある不整合の例も多い。「文部科学大臣」の記事では、町村信孝元文部科学大臣の画像が掲載されており、キャプションは「初代文部科学大臣町村信孝」となっている。掲載された画像が町村信孝の写真であり、町村信孝が初代文部科学大臣であることに誤りはない。しかし、この画像は、外務大臣時代の町村

(注1): netmile.co.jp

信孝の写真であり、キャプションから想像されるような「初代文部科学大臣時代の町村信孝」の写真ではない。このような例も本論文では不整合の一種として扱うことにする。この例に関しては、上記のアンケート結果で、約 6 割のユーザが、この写真を載せるのが良いが、「写真は外務大臣当時のもの」といった説明を補うべきと回答している。また、記事を信用するかどうか、Wikipedia の記事は多数のユーザによって様々に編集されているという事実を提示した後にアンケートをとりなおしたところ、8 割以上の記事を信用するとしたユーザは 3 割程度に減少し、6 割以上の記事を信用するとしたユーザは 6 割程度に減少した。これらのことから、日付情報の不整合を適切に提示することは、Wikipedia が持つ大量の情報を有効に活用するために必要であると考えられる。

これらの不整合を検出するためには、同一の事象に関して複数の時間情報を収集したり、不完全な時間情報を前後の文章などから補完する必要がある。時間情報の収集や補完を行う研究は多い。例えば、木村ら [1] の研究では、Web から時間の表現を収集し、その周辺の情報を用いることで人物の年表を作成している。奥村ら [2] の研究では、テキストに含まれる Who, When, Where, What, Why, How という 5W1H 情報に着目し出来事に至る経緯を発見するとともに、類似の情報に関する比較と、情報全体の鳥瞰情報を抽出する方法を提案している。鳥瞰情報の抽出とは、膨大な情報を適切に分類することで情報の集合全体の把握することである。

しかしこれらの手法では、もともと日付表現のない文章の時間情報を取得することはできない。そのため、上で挙げたような日付情報の不整合例の検出は困難である。そこで本論文では、Wiki コンテンツの 1 つである Wikipedia 上に記述されている日付表現記述のない事象について発生した日時を推定する手法を提案する。提案手法は、Wikipedia が百科事典として持っている特徴、また、リンク先の情報および段落の構成などの文書構造を用い、日付情報のない事象の発生日時を推定する。

推定の手順は、日付表現の抽出、日付表現の補完、日付表現記述を有する事象の発生日時の推定、日付表現記述を持たない事象の発生日時の推定の順となる。現在は、実際の Wikipedia の数個の記事を用いて提案手法を適用し、推定された発生日時の精度、再現率を計算し、提案手法の評価を行いつつ、手法の改良を続けている。

この手法が確立することで、上記のような不整合例の検出が可能になると考えられる。それだけでなく、各記事ごとにこの手法を適用し、得られた知識を各記事間でマージすることにより、Wikipedia 上に記述された膨大な数の事象に関して、もともと日付表現が記載されていた事象に関してもそうでない事象に関しても時間情報に関するデータベースを得ることができると期待される。また、各事象に関して複数の記事から得られる情報をマージすることにより、より精度の高い情報を得ることができると考えられる。ここから得られる知識は様々な分野で応用可能な知識になりうると期待される。

本稿の構成は以下の通りである。まず 2 章にて、日付情報を扱う関連研究について述べ、3 章にて提案手法を述べる。次に

4 章で実験方法及び結果について述べ、最後に 5 章でまとめと今後の課題について述べる。

2. 関連研究

日付表現を扱う研究としては、木村ら [1]、金田 [3]、Kim ら [4]、Schiffman ら [5] の人物の年表や伝記を生成する研究がある。

木村ら [1] は、ある特定の人物に関する Web 上の文書の集合から、人物の年表を生成する手法を提案している。西暦や元号など多様な表記をされる日付の表現を正規表現を用いて収集し、日付の表記法の統一を行っている。またその時間表現と共に記述されている人物に関する記述を収集し、人物に関する記述を時系列順に提示している。日付表現の補完については、年月日をそれぞれ 4 桁、2 桁、2 桁の数値で表し、8 桁の数値に変換する事で正規化を行っている。例えば、「2007 年 1 月 1 日」は「20070101」と正規化される。また「翌日」「2 日後」などのように、ある時点からの時間経過を指し示す日付表現など、その語だけでは年月日の正規化を行うことができない表現について補完対象の直前に出現する日付表現の情報を用いて補完することを提案している。すなわち「翌日」の直前に出現する日付表現が「2007 年 1 月 1 日」であれば「翌日」は「20070102」と正規化される。

金田 [3] は、文字列パターンのマッチングにより、年代情報を抽出し、年代インデックスを作成している。その後、キーワードと年代範囲を与え、検索結果を年代でソートすることで、年表を作成している。

Kim ら [4] の Artequakt では、Web から自動収集した情報を元に芸術家の伝記の自動生成を行っている。Artequakt は、WordNet などのオントロジーと予め記述したパターンを元に情報を抽出している。

Schiffman ら [5] は、ニュース記事の集合から人物の伝記を自動的に生成する手法を提案している。同格関係により人物の職業を導き出し、その職業と強い関連のある動詞が含まれる文を抽出している。

また、日付表現を扱う問題として固有表現抽出という問題がある。これはテキストなどから、固有表現を抽出分類する問題である。固有表現とは、人名、地名、組織名などの固有名詞や、時間、日付の表現、その他数量表現を指す。

奥村ら [2] は、5W1H 情報に着目し出来事に至る経緯、出来事に関する比較、情報全体の鳥瞰情報を抽出する方法を提案している。彼らは形態素解析、表層格の解析、固有名詞辞書を用いて、人名、組織名を Who 要素、地名を Where 要素として抽出している。また、パターンマッチングを用いて日時、場所、理由の表現を抽出し、When, Where, Why 要素としている。How 要素は動詞としている。

渡邉ら [6] は、資源として Wikipedia 利用して固有表現を獲得し、それを固有表現抽出の手掛かりとして利用することで解析精度の向上を行っている。Wikipedia の構造、アンカーによって構成されるグラフ構造を用いることで、出現する固有表現の分類を行っている。

2007年9月10日に第168回臨時国会が開催され、安倍首相は所信表明演説を行った。その中で安倍首相は「職責を全うする」という趣旨の決意を表明した。なお、この表明では自身の内閣を「政策実行内閣」と名づけ、「美しい国」という言葉は結びに一度使ったのみであった。午後には「(改正案を通すのは) 厳しいでしょうか」と辞任を示唆する発言を麻生幹事長が述べたが、麻生は「このタイミングで辞めるのはいかがなものか」と慰留していた。9月11日には妻の昭恵に対し「もうこれ以上、続けられないかもしれない」[24]と漏らしている。しかし、辞任する具体的な日程までは、昭恵には一切明かさなかった[24]。

2007年9月8日、アジア太平洋経
済協力首脳会議にて、アメリカ合衆
国大統領ジョージ・W・ブッシュと握
手を交わす安倍晋三

辞意表明

2007年9月12日午後2時(JST)、「内閣総理大臣及び自由民主党総裁を辞する」と退陣を表明する記者会見を行った。これにより同日予定されていた衆議院本会議の代表質問は中止となった。

2007年9月12日

図 1 日付表現の補完

固有表現抽出規則学習においては、Support Vector Machine [7]、最大エントロピー法 [8]、決定木学習 [9]、決定リスト [10] などの機械学習を用いる研究がなされている。

3. 提案手法

3.1 日付表現の抽出

記事中から日付表現の抽出を行う。日本語の表現を対象とし、正規表現を用いることにより日付を表す表現を抽出する。抽出する日付表現は、「**年**月**日」といった数字の後に年月日が出現するもの、「明治」、「大正」、「昭和」、「平成」などの和暦を含むもの、「当日」、「翌月」、「2日後」などといったある時点を基準として日付を表現するものとする。「**日間」などの期間を表す表現については開始・終了日時の推定が必要であるため今回は扱わない。

3.2 日付表現の補完

抽出した日付表現を西暦(4桁)、月(2桁)、日(2桁)の8桁の整数に正規化する。この際、「**月**日」や「**日」のように年、または月の情報が欠落している日付表現に対しては直前に出現する日付表現から補完する。また、「当日」、「翌月」、「2日後」などの日付表現は直前に出現する日付表現を基準に計算し、8桁の整数に正規化する。このとき、「**年」、「**年**月」という日付表現に関して月、および日の情報を補完するといったことはしない。補完の例を図1に示す。上の例では、「9月11日」に直前の「2007年9月10日」から不足している「2007年」の情報を補完し、「20070911」と正規化する。また、下の例では「当日」という表現を直前の「2007年9月12日」をもとに計算し、「20070912」と正規化する。

3.1, 3.2の手法は関連研究[1]で用いられている。

3.3 日付表現記述を有する事象の発生日時の推定

事象の発生日時は正規化された2つの日付「開始日時」と「終了日時」で表す。

単一の日付表現記述を持つ事象は、その日付表現が発生日時であると推定する。また、複数の日付表現記述を持つ事象は、ある期間発生していた事象であると推定し、それら日付表現の最も前の日付から最も後の日付までの範囲が事象の発生日時であると推定する。例を図2を用いて説明する。各文章を事象A,B,C,D,Eとする。この例では、日付表現記述を有する事象は、A,B,Dである。このうち、事象A,Bは単一の日付表現記述を有するため、発生日時をそれぞれ「20051031」～

「20051031」、「20060901」～「20060901」と推定する。事象Dは「20060920」と「20060926」の日付表現記述を有する。そこで、事象Dは「20060920」から「20060926」にわたって発生したと推定する。ただし、1つの事象に対し、複数の記述から複数の日付表現を取得した場合、このように推定を行うことで不整合を吸収してしまうことが考えられる。そのためこの推定は1つの事象に対し、単一の記述から複数の日付表現を取得した場合に行う。各事象に関する複数の記述による情報のマージは日付の推定とは別に行うものとする。情報のマージとは、同一の事象に関して得られた2つ以上の情報をまとめることである。例えば、同じ事象についてある記述では「20060920」、別の記述では「20060926」とされていた場合、これは事象の発生日時に関する不整合である。そのため、この事象の発生日時をそれぞれの記述ごとに保持する。ここでは「20060920」～「20060920」、「20060926」～「20060926」となる。その後で情報をマージする際に不整合を発見するということである。

3.4 日付表現記述を持たない事象の発生日時の推定

Wikipediaは事象の情報が時系列に沿って記述されていることが多いという特徴がある。そこで我々は、事象は、直前に記述されている事象以降、直後に記述されている事象以前に発生したと考え、直前に出現する日付表現記述を有する事象の終了日時から直後に出現する日付表現記述を有する事象の開始日時の範囲が事象の発生日時であると推定する。例を図2を用いて説明する。各文章を事象A,B,C,D,Eとする。この例では、日付表現記述を持たない事象は、C,Eである。事象Cは事象B以降、事象D以前に発生したと考え、発生日時を「20060901」～「20060920」と推定する。事象Eは事象D以降に発生したと考え、発生日時を「20060926」～と推定する。

このように推定を行うため、3.3で推定する日時は、この期間中、事象の発生している期間であり、3.4で推定する日時は、この期間内のいずれかの日時に事象が発生したと考えられる期間である。

3.5 文章構造の利用

3.2, 3.4の手法は、記述間の時系列的なつながりをもとにしている。時系列的なつながりがあるとは、記述の出現順序が各事象の発生順序と等しいことを指している。すなわち、事象が発生順序にしたがって記述されているということである。しかし、すべてに記述間にこれらの関係があるわけではない。そこで、文章構造を利用する。同一の文章構造に含まれる記述間、

A	小泉政権末期の早い段階から自民党内の「ポスト小泉」の最有力候補の一人と言われ、2005年10月31日付で発足した第3次小泉改造内閣では内閣官房長官を務める。
B	2006年9月1日に自民党総裁戦への出馬を表明。憲法改正や教育改革、財政健全化に取り組む方針を示す。また、総裁選に当選し次第、所属する派閥の森派を離脱する考えを示した。
C	
D	小泉の総裁任期が満了となり、2006年9月20日、自由民主党総裁選挙で麻生太郎、谷垣禎一を大差で破って自由民主党総裁に選出、9月26日の臨時国会に於いて内閣総理大臣に指名される。戦後最年少で、戦後生まれとしては初めての内閣総理大臣であった。
E	

図 2 事象の発生日時の推定

	総抽出数	誤抽出
安倍 晋三	241	19

表 1 日付表現の抽出の結果

	総補完数	誤補完
安倍 晋三	58	8

表 2 日付表現の補完の結果

文章構造をまたいだ記述間で時系列的なつながりがあるかどうかについて考える．文章構造をまたいだ記述間で時系列的なつながりがある場合、それらの文章構造間には時系列的なつながりがあるとする．

まず、文章構造を利用しやすくするため、Mediawiki 形式の Wikipedia の記事を XML に変換する．また、各事象の発生日時は XML のタグの属性として表現する．図 3 に例を示す．例に出現する「00000000」、「99999999」という日付は情報を得られなかったことを意味している．

まず、文章構造のうち Section, Paragraph, ListItem, Definition に着目する．Section とは Wikipedia が採用している Mediawiki 記法で「==キーワード==」で表現される見出しを用いて表現される節であり、Paragraph は節内の段落である．ListItem は「* 項目」で表現される箇条書きであり、Definition は「; 定義語:定義」で表現される定義集合である．Subsection は Section を入れ子にすることで表現されている．また、Definition は記事内である語を挙げるその解説を行うというように、Section でタイトルを挙げて内容を記述するのと同様の使い方もなされている．3.2, 3.4 の手法、すなわち日付表現の補完、発生日時の推定をこれらの文章構造の内部の情報のみで、または外部の情報も用いて行うことで記述間の時系列的な関係を有効に利用する．実例を図 4 に示す．

4. 実 験

実際の日本語の Wikipedia の記事を用いて、提案手法を適用し、評価を行った．対象として、「安倍 晋三」の記事を使用した．また、記事に含まれる読点で区切られる各文章を 1 つの事象として扱った．評価基準として、各事象について以下の値を計算し、記事内に出現する事象の平均値を用いた．ただし、 T_1 は正解の時区間であり、 T_2 はシステムが出力する解の時区間である．正解は人手で作成した．

$$e = \frac{T_1 \cap T_2}{T_1 \cup T_2}$$

推定が行えなかった場合、推定した日時が矛盾している、すなわち開始日時が終了日時以後となっている場合は、評価値を 0 とする．開始日時もしくは終了日時の一方のみ推定ができた

場合は、もう一方の値を推定できた日時とし、同様の方法で評価値を計算する．また、日付情報の抽出、および日付表現の補完の結果は、事象の発生日時の推定に大きな影響を与えるため、これらの結果も評価した．

4.1 文章構造を利用しない場合

文章構造を利用せず、記事全体に対して提案手法を適用した．

日付表現の抽出の結果を表 1 に示す．これらの誤りは大きく 2 つに分けられる．1 つは「*日間」、「*年生」などのように言葉の 1 部を日付表現として抽出してしまっているものである．もう 1 つは「*年」、「*月」、「*日」など表層的な特徴がないものの「改造内閣はわずか 31 日の短命に終わった」「義務教育 9 年の規定や男女共学の項を削除した」の「31 日」「9 年」のように文脈で期間や期限を示すもの、あるいは「安倍は父、晋太郎の秘書官時代に光永の設立した「光カメラ販売」(1991 年倒産)では取締役に、」のように「1991 年」組織などの属性の値を表しているものであった．

日付表現の補完の結果を表 2 に示す．誤抽出原因の多くは、補完の基準とする直前の表現が誤抽出によって得られたものになってしまったことであった．もう 1 つの原因は直前に過去の事象が話題の対象として出現したため、過去の日付表現が出現し、それを基準に補完を行ったため発生したものであった．例を図 5 に示す「同年 11 月」の「同年」は「2003 年 9 月」を基準に補完すべきであるが、直前に出現した「1979 年」を基準としてしまうことによって誤りが生じている．

日付表現を持たない事象について 3.4 の手法を用いて事象の発生日時の推定を行った結果を表 3 に示す．誤推定数には、わずかであるが日付情報の誤抽出、誤補完による日付表現記述を有する事象の発生日時の推定における誤りも含まれている．一方で誤推定数の 9 割以上を占めるのは、日付情報の誤抽出、誤補完による日付表現記述を持たない事象の発生日時の推定における誤りと、文章構造間での記述に時系列的なつながりがない部分で 3.4 の手法を用いたことによる誤りであった．例を図 6 に示す「60 年」と「2006 年 11 月 14 日」が抽出されるため、その間に記述された事象は「00600000」～「20061114」と推定される．60 年は期間が誤抽出されたものであり、推定に誤りを生じさせる．これは同様に前方へも波及する．また、ここでは

```

- <Section level="2">
  <Title>来歴</Title>
- <Section level="3">
  <Title>出生</Title>
  - <Paragraph>
    <Sentence start="19540921" end="19540921">1954年9月21日、当時毎日新聞の記者だった安倍晋太郎とその妻、洋子の次男として東京で生まれる(本籍地は山口県大津郡油谷町(現・長門市))</Sentence>
    <Sentence start="19540921" end="99999999">父方の祖父は衆議院議員の安倍寛(この時代に故人)、母方の祖父は後の首相・岸信介で、大叔父にはやはり後の首相・佐藤栄作がいる、政治家一族であった</Sentence>
    <Sentence start="19540921" end="99999999">「幼い頃から私には身近に政治がありました」と安倍は振り返っている</Sentence>
  </Paragraph>
  - <Paragraph>
    <Sentence start="00000000" end="99999999">幼い頃は、野球選手や刑事になることに憧れていた</Sentence>
  </Paragraph>
</Section>

```

図 3 XML 化され日付情報が付与された Wikipedia の記事

Section

来歴 Subsection

出生 Paragraph

1954年9月21日、当時毎日新聞の記者だった安倍晋太郎とその妻、洋子の次男として東京で生まれる(本籍地は山口県大津郡油谷町(現・長門市)。父方の祖父は衆議院議員の安倍寛(この時代に故人)、母方の祖父は後の首相・岸信介で、大叔父にはやはり後の首相・佐藤栄作がいる、政治家一族であった。「幼い頃から私には身近に政治がありました」と安倍は振り返っている^[1]。幼い頃は、野球選手や刑事になることに憧れていた^[2]。

学生時代

成蹊学園(小学校・中学校・高等学校・大学法学部政治学科)を卒業。小学4年生から5年生にかけての1964年から2年間平沢勝栄(後の衆議院議員、日本人拉致問題で助言したと言われる)が家庭教師にっていた^[3]。高校でクラブは地理研究部に所属^[4]。

略歴 ListItem

- 1954年9月21日・東京都に生まれる。本籍地は山口県大津郡油谷町(現・長門市)。
- 1977年3月・成蹊大学法学部政治学科卒業
- 1977年4月・米カリフォルニア州ヘイワードの英語学校に入学。その後、ロングビーチの語学学校に転校した。

政見・政策

社会保障 Definition

中国残留孤児

中国残留孤児問題における訴訟では請求を取り下げられた原告団に面会し、新たな支援を検討していくことを確認した。慈善病院の「こうのとりのゆりかご」

赤ちゃんポストも参照

2007年2月23日に、熊本市の慈善病院が赤ちゃんポストの設置を計画していることについて、「ポスト」という名前や匿名で子供を置いていけるものだということに大変抵抗を感じると反対の意向を示した^{[6][8]}。

年金問題

年金記録問題および宙に浮いた年金記録も参照

年金記録問題では民主党の小沢一郎との党首討論で「消えた年金はどうするのか」という野党からの追及に対し「年金は消えたわけではない」として年金時効撤廃特例法案など具体的な解消案を提示した。該当者不明の年金記録5000万件の照合作業

図 4 文章構造の例

2003年9月、小泉首相によって自民党幹事長に抜擢された。自民党は総幹分離の原則が長く続いており、総裁派閣幹事長は1979年の大平正芳総裁時代の斎藤邦吉幹事長以来24年ぶりであった。大臣経験もない若手議員が第一与党幹事長に就任するのは極めて異例のことであるが、苦戦が予想される同年11月の総選挙のために安倍の「人気」が必要とされたためと見られた。11月の総選挙で与党は絶対安定多数の確保に成功した。

図 5 日付表現の誤補完の例

事象数			評価値
日付表現あり	日付表現なし	合計	
183	323	506	0.1800

表 3 文章構造を用いない推定の結果

「憲法」「安全保障」の Section にまたがって推定がなされているが「憲法」の Section に記述されている国民投票法案が制定されたのは2007年5月14日であり、「安全保障」の Section に出現する「2006年11月14日」以前に発生したとの推定は誤ったものである。

4.2 文章構造を利用した場合

次に文章構造を利用し、提案手法を適用した「安倍晋三」の

記事を XML に変換して利用し、読点で区切られる各文章を1つの事象として扱うため各文章に Sentence タグを付与した。Section, Paragraph, ListItem, Definition の4つの文章構造に関して、Paragraph, ListItem, Definition は Section の内部で利用されていることから、日付表現の補完、事象の発生日時の推定において該当する Section の内部の情報を全て用いる場合と、Paragraph, ListItem, Definition のそれぞれの内部の

憲法

総裁選では施行60周年を迎えた日本国憲法を改正すると宣言し、総理就任後の国会で、「現行の憲法は、日本が占領されている時代に制定され、60年近くを経て現実にそぐわないものとなっているので、21世紀にふさわしい日本の未来の姿あるいは理想を憲法として書き上げていくことが必要と考えている」と述べた^[94]。また「私は、国会議員になった当初から改憲論者だが、3つの点で憲法を改正すべきだと主張してきた。第一の理由だが、現行憲法は占領軍の手によって、憲法の専門家ではない人たちによって2週間そこそこで書き上げられた、と言われており、やはり国の基本法である限り、制定過程にもこだわらざるを得ない」と述べた^[95]。

改憲のための「国民投票法案(日本国憲法の改正手続に関する法律)」の整備をおこなうが、選挙に向けて野党の反対を押し切り成立させたためメディアでは法案の不備について批判が出た。

安全保障

日本版「国家安全保障会議」(NSC)構想を推進した。総理就任以前は憲法改正に関しては集団的自衛権の容認を打ち出して来た。総理就任後は、防衛庁を防衛省へ格上げした。これは「戦後レジーム(体制)から脱却し、新たな国造りを行うための第一歩」と意義付けられた。

2005年11月14日 安倍内閣は閣議で、核保有についての鈴木宗男の質問主意書^[96]に対して、「政府としては、非核三原則の見直しを議論することは考えていない」と強調しながらも、「核兵器であっても、自衛のための必要最小限度にとどまれば、保有は必ずしも憲法の禁止するところではない」との答弁書^[97]を出した。

図 6 発生日時の誤推定の例

	両側推定		片側のみ推定	情報なし	合計
	補完済み	未補完を含む			
S-S	300	51	136	19	506
S-P	175	40	141	150	506
P-P	156	59	141	150	506

表 4 文章構造を用いた補完、推定の結果

情報のみを用いる場合の 2 通り行った。また、ある事象の日付表現の補完はその事象の記述と別の記述との間に時系列的なつながりがなくとも各事象が同じ年もしくは月に発生しているなど発生日時が近い場合には記述の出現順序が事象の発生順序と食い違っていても可能であるため、事象の発生日時の推定のほうがより制約が必要であると考えた。例えば「2008 年 4 月 21 日に事象 A が発生した。15 日の事象 B が原因である。」といった場合、後に記述された「事象 B」は「事象 A」より前に発生しているが、「2008 年 4 月」という情報を補完することが可能である。したがって補完と推定をともに Paragraph, ListItem, Definition の内部 (P-P)、ともに Section の内部 (S-S)、補完を Section の内部、推定を Paragraph, ListItem, Definition の内部 (S-P)、と 3 通りで実行した。記事内の各構造の数は、Section: 50, Paragraph: 93, ListItem: 56, Definition: 50, Sentence: 506 であった。またこれらの構造に着目したため、画像およびキャプションが格納されている構造である Image に関しては対象から除外した。したがって構造を用いなかった場合の事象数に比べ、事象数すなわち Sentence の数が減少している。結果を表 4 に示す。ここで「両側推定」とは推定した結果、事象の発生日時の両端が得られたものを指す。「片側のみ推定」とはある日時以降に発生した、もしくはある日時以前に発生したという情報のうち一方のみが得られたものをさし、「情報なし」とはそのどちらも得られなかったものを指す。「未補完」は得られた日付表現が年月日の情報がそろっていないものであり、「補完済み」は年月日の情報がそろって得られたものである。また、各項目の括弧内の数字は推定された発生日時が誤っていたものの数である。精度に関しては発生日時の両端の情報が得られ、かつそれらの年月日の情報がそろっていたもののうち正しく推定できているものの割合とした。再現率は、全事象のうち発生日時の両端の年月日の情報がそろって正しく推定できているものの割合とした。文章構造を用いなかった場合の結果と文章構造を用いなかった場合との比較を表 5 に示す。文章構造を用い、Section から Paragraph, ListItem, DefinitionList へ補

	全体	両側+片側	両側
構造を用いない	0.1800	-	-
S-S	0.1534	0.1594	0.2184
S-P	0.1358	0.1930	0.3127
P-P	0.1224	0.1740	0.2820

表 5 評価値の比較

完・推定の範囲を狭めることで誤った推定を行わなくすることを目的としている。そのため、出現する事象全体の評価値、推定しなかった事象を省いた評価値 (両側+片側)、開始日時、終了日時の一方でも推定できなかったものを省いた評価値を計算し比較を行った。それぞれの制限ごとに評価値の最大値を見ることで文章構造の利用による誤り回避の効果を見て取ることができる。一方で全体での評価値が補完・推定の範囲を狭めることで、正しく推定されていた部分を制限してしまっていることが分かる。

5. おわりに

本稿では、Wikipedia の特性、特徴を用いることで手掛かりとなるような日付表現記述のない事象について発生日時を推定する手法を提案した。具体的には、文章の時系列的なつながりを利用した、直前に出現する日付表現記述を有する事象の終了日時から直後に出現する日付表現記述を有する事象の開始日時の範囲が事象の発生日時であると推定する手法である。さらに文章構造を利用することで事象の発生日時の推定の向上を図った。また、実験により手法の効果を検証した。

実験の結果から、事象の発生日時の推定の精度の向上のためにはまず日付情報の抽出精度、および補完精度の向上、および Section, Subsection などの文章構造間の関係を利用することが必要であることがわかった。

日付情報の補完における誤りは日付情報の抽出の精度を向上することによりある程度減らすことができる。日付情報の抽出に関しては、「*日間」、「*年生」などのように言葉の 1 部を日付

表現として抽出しているものは、抽出ルールを細かく設定することで回避することが可能であると考えられる。

文章構造の利用について、今回は主要な文章構造の要素の種類に対して、内部で時系列的なつながりがあるかどうかを判断する形で利用した。これによって誤り推定を回避することができたが、正解推定を排除してしまっている部分もある。これは文章構造の要素の種類ののみを利用したためであると考えられる。また、同じ文章構造を持つ記事であっても時系列的なつながりの有無は異なる。例えば Wikipedia の「安倍晋三」の記事において、「来歴」の Section と「政見・政策」の Section とでは時系列的なつながりの有無が異なる。前者の内容は来歴についてであるので時系列に沿って記述されており、「来歴」内部の各 Subsection の間には時系列的なつながりがある。一方後者では、より重要なものが先に記述されていたり、時系列的に並列に進行している事象が異なる部分に記述されているなど時系列に沿って記述されていないため、「政見・政策」内部の各 Subsection の間に時系列的なつながりはない。今後、各構造に含まれるテキストの特徴や、記事全体の文章構造の特徴を利用することで、このような例についても有効に推定を行う手法を開発する予定である。

また、現在の評価基準では片側推定の結果をうまく評価できるとは言えない。そのため、評価基準の検討を行う予定である。

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します。

文 献

- [1] 木村 壘, 小山 聡, 田中 克己, 人物情報辞典の自動生成のための時系列情報自動収集 (ウェブ属性抽出, 夏のデータベースワークショップ DBWS 2006) 情報処理学会研究報告, データベース・システム研究会報告, 社団法人情報処理学会, 2006
- [2] 奥村 明俊, 池田 崇博, 村木 一至, 5W1H 分類・ナビゲーションによる情報活用プラットフォーム, 情報処理学会研究報告, DD, [デジタル・ドキュメント], 社団法人情報処理学会, 1997
- [3] 金田泰, “ 百科事典から動的に年表を生成するテキスト検索法のための年代情報の抽出法と表現法, ” 情報処理学会 情報学基礎研究会報告 Vol.1999, No.57, pp.81-88, 1999.
- [4] S. Kim, H. Alani, W. Hall, P. Lewis, D. Millard, N. Shadbolt, and M. Weal, "Artequakt: Generating Tailored Biographies from Automatically Annotated Fragments from the Web," In Proceedings of Workshop on Semantic Authoring, Annotation and Knowledge Markup (SAAKM'02), the 15th European Conference on Artificial Intelligence, (ECAI'02), pp. 1-6, 2002.
- [5] B. Schiffman, I. Mani, K. J. Concepcion, " Producing Biographical Summaries: Combining Linguistic Knowledge with Corpus Statistics, " In Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL-EACL 2001), July. 2001.
- [6] 渡邊 陽太郎, 浅原 正幸, 松本 裕治: “ グラフ構造を持つ条件付確率場による Wikipedia 文書中の固有表現分類 ”, 人工知能学会論文誌, Vol. 23, No. 4, pp.245-254 2008.
- [7] 山田 寛康, 工藤 拓, 松本 裕治, Support Vector Machines を用いた日本語固有表現抽出, 情報処理学会研究報告. 情報学基礎研究会報告, 社団法人情報処理学会, pp.121-128, 2001.

- [8] 内元清貴, 馬青, 村田真樹, 小作浩美, 内山将夫, 井佐原均, “ 最大エントロピーモデルと書き換え規則に基づく固有表現抽出. ” 自然言語処理, pp.63-90, 2000.
- [9] Sekine S. Grishman R. and Shinnou, H., A Decision Tree Method for Finding and Classifying Names in Japanese Texts, In Proceedings of the Sixth Workshop on Very Large Corpora, pp.171-178, Montreal, Canada, August, 1998.
- [10] Manabu Sassano and Takehito Utsuro. 2000. Named Entity Chunking Techniques in Supervised Learning for Japanese Named Entity Recognition. In Proceedings of COLING pp.705-711, 2000.