Vol.2010-DBS-151 No.34
2010/11/13

整合性の分析のための Wikipedia からの事象データベースの構築

島田裕司[†1]　浅野泰仁[†1]　吉川正俊[†1]

Wikipedia から事象知識データベースを構築する手法を提案する．扱う事象は記事内に記述されている現実に起こったとされるすべての事象である．これらの事象のモデル化を行い，事象に関する情報を獲得，データベースに格納する．事象モデルでは 5W1H に関する情報のうち，When,Who,Where,What の要素に着目する．異なる記述から抽出された各要素の情報を元に，機械学習を用いて同一の事象に関する記述であるかを判定する．

Constructing event knowledge database from Wikipedia for consistency analysis

HIROSHI SHIMADA,[†1] YASUHITO ASANO[†1] and MASATOSHI YOSHIKAWA[†1]

In this reserch, we propose the method constructing event knowledge database from Wikipedia. The targets are all events that described in the article. Modeling these events, next we collect informations of events by using the model, and store them into a database. The event model use When,Who,Where,What elements in 5W1H information. Each events sometimes are described in different articles or different parts of same article. We identify them using machine learning.

†1 京都大学
Kyoto University

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おり，Web 上に存在するの情報の量は膨大な量である．これらの情報を有効に活用するため，Web からの知識発見に関する研究は盛んに行われている．特に事象に関する情報の取得を目的としている研究がある．事象に関する情報とはいつ誰がどこで何をしたといったような実際に発生した出来事に関する情報を指す．このような情報を知識として獲得することができると，ある人物に関連する事象や，ある場所である時期に発生した事象を容易に知ることができる．また，このような知識を分類，蓄積し，データベースを作成しておくことで，他のアプリケーションから利用することが可能となる．このような分野では，5W1H 情報に着目して事象に関する情報を取得している研究[2)]や，日付情報に着目し，事象に関する情報を集めて年表を作る研究[1)]などがある．これらには多義性の解決などの難しさにより異なった記述のなされた同一の事象への対処などの問題がある．

Web コンテンツのひとつである Wikipedia は，ユーザの誰もが閲覧，編集することのできる大規模な百科事典である．今回使用する日本語版の Wikipedia には約 60 万の記事がある．Wikipedia の記事は様々なタグによって構造化された半構造化データである．また，密なリンク構造や，URL により語彙の意味が一意に特定できる，200 以上の言語をサポートしているなどといった特徴がある．これらの特徴を利用して Web を利用するだけでは困難であった知識の獲得を行う研究がなされている．URL により語彙の意味が一意に特定できることやリンクのアンカーテキスト，曖昧さ回避の記事を用いて語義の曖昧性解消をはかる研究[7)]や，記事に付与されたカテゴリや適用されているテンプレートの情報から語の間の関係を抽出する研究[8)]などが行われている．加えて，半構造化されていること，リンク構造が利用できること，記事数が多く情報の網羅性が高いことは事象に関する情報を扱う上で Wikipedia を利用する利点といえる．

本研究では，Wikipedia から事象知識データベースを構築する手法を提案する．扱う事象は記事内に記述されている現実に起こったとされるすべての事象である．事象モデルを用いて事象に関する情報を獲得，データベースに格納する．事象モデルでは 5W1H に関する情報のうち，When,Who,Where,What の要素に着目する．また，その他の重要であると思われる名詞をキーワード集合として用いる．これらの情報の取得には，固有表現抽出ツール NExT[9)]，構文解析システム KNP[10)] を用いる．重要な概念を表す名詞には，Wikipedia 内

にそれをタイトルとする記事が存在するという性質を利用している．さらにデータベースに格納された各要素について，同一である，包含関係を持っている，共通部分がある，類似度が高い，などの基準によりそれぞれの記述が同一の事象に関する記述であるかを判定する．事象には様々な粒度があり，それぞれの事象の間には同一であるという関係の他，一方が他方を包含しているという関係がある．今回は，これらの記述を比較することで，記事間，記述間での情報の不整合を発見することが可能となる．情報の不整合の発見は，誤った情報の発見にもつながる．誤りを発見することは，包含関係にある事象についての記述も含めて，同一の事象に関する記述として扱う．多数のユーザが編集を行うことによって誤った情報を含むことが多い Wikipedia において，その情報の有効利用のための重要な処理である．

事象に関する情報のうち特に発生日時に関する情報については，実際に日付に関する情報が明示的記述されているものだけではなく，記事内に潜在的に持っていると思われる情報の推定を行う．

本研究では句点で区切られる 1 文をひとつの記述として扱う．ひとつの記述は，ひとつの事象を表し，逆にひとつの事象を表す記述は複数存在するものとする．また，その有効性の検証のため，そのデータベースを用いて，同一の事象の判定を行い，結果を分析した．

本稿の構成は以下の通りである．まず 2 章にて，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

事象に関する情報を扱った研究として，奥村ら[2], 木村ら[1] の研究がある．

奥村ら[2] は，新聞記事を対象に 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．この研究では各記事の要約であるヘッドラインからその記事に関する情報を抽出している．すなわちこの研究では各記事をひとつの出来事として扱い，5W1H の各要素を抽出している．これに対して，我々は，Wikipedia の記事を扱うが，記事内にあるより細かい事象の記述について扱う．すなわち，ひとつの記事は多くに事象の記述によって成り立っていると考える．

ここで用いられた固有表現に関して，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．

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，それを人物に関する事象として扱い，時系列順に提示しすることで年表を作成している．日付情報については「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関し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．その他，同姓同名の人物の存在への対処は行っているが，同一の事象に関する複数の記述が得られることに関しての対処は行っていない．

その他，年表を作成する研究として，金田[3]，Kim ら[4]，Schiffman ら[5] の研究がある．

3. 提 案 手 法

3.1 事象モデル

同一の事象に関する記述であるかどうかを判別するため，事象モデルを構築する．Wikipedia の記事内の事象に関する記述について，それぞれの事象を表現する重要な要素について調査を行った．事象とは，句点で区切られたひとつの文章で表現されているものを指し，ある出来事の発生を記述したもの，およびある対象の状態を表現しているものを含める．

まず注目した要素は，事象の発生日時である．どのような事象にも発生した日時，期間は存在するため，これによってある程度他の事象と判別が可能である．また，例えば選挙や，スポーツの試合などのように，同じ出来事が繰り返し行われるような場合，発生日時の情報を用いれば容易にそれらを区別可能であるが，発生日時の情報を用いずに区別することは困難な場合がある．

次に，事象が発生した場所である．これも発生日時と同様，すべての事象に対して，存在する．ただし，場所の範囲はまちまちであり，世界的に発生した事象や，ある国，ある都市で発生した事象，ある点で発生した事象が考えられる．また，後に述べる事象の包含関係にも関連するが，複数の場所で同時発生的に起こる事象も存在する．複数の場所で同時発生的に起こる事象に関して，それらの場所を含む地域，例えば，様々な国で起こっている場合

には，世界で発生した，日本国内の各地で発生した場合には，日本で発生したと考えると，異なる場所で発生した事象は，別の事象であると判断することができる．

それから事象に関わった人物，組織に着目した．この要素は，すべての事象に対して存在するわけではないが，自然現象以外の事象に関しては，何らかの組織，人物が関わっている．また，自然現象に関しても，被害を受けたもの，対策を行ったものなど，人物，組織が関わっていることがある．また，出来事であればその主体，すなわち「だれが」もしくは「なにが」行ったか，状態であれば,「だれが」もしくは「なにが」の状態であるのかというのは事象を判別する重要な要素である．

発生日時，発生場所，関わった人物，組織，動詞，これらの要素は，事象を表現する上で重要であることが分かった．しかし，これだけで事象を特定するのに十分であるとは言えない．例えば,「2003 年 9 月安倍晋三は自由民主党幹事長に就任した.」「2006 年 9 月安倍晋三は第 90 代内閣総理大臣に就任した.」といった事象では，前者は日時:2003 年 9 月, 人物：安倍晋三, 場所:なし, 動詞:就任，後者は日時:2006 年 9 月, 人物：安倍晋三, 場所:なし, 動詞:就任，となり発生日時は異なっているものの同じ事象が発生したと判断されてしまう．これはそれぞれの事象を表す特徴的な語である「自由民主党幹事長」「第 90 代内閣総理大臣」などが反映されていないことが原因である．そのため，上記の各要素には該当しないが，事象を特定するための重要となる語をキーワード集合として利用する．

3.2 情報の収集

各要素の取得方法について述べる．When 要素に関しては 3.3 において述べる．Who 要素として，記述における主語，および記述に含まれる人名，組織名を取得する．Where 要素として記述に含まれる地名を取得し，What 要素として，記述における述語を取得する．

主語，述語，人名，組織名，地名の情報の取得については NExT(Named Entity Extraction Tool)[9]，KNP[10] を利用する．NExT は日本語文から固有表現を抽出するシステムである．テキスト文書情報に含まれる人名，組織名，地名，数量表現を自動的に判別し，様々な形式で抽出・タグ付けする．形態素解析システムの解析結果を入力とする．KNP は日本語文の構文解析を行うシステムである．形態素解析システムの解析結果を入力とし，それらを文節単位にまとめ，文節間の係り受け関係を決定する．また，文節の格に関する情報も付与する．これらのシステムへの入力データを作成する形態素解析システムとして，JUMAN[11] を利用する．

NExT を用いて，Who 要素，Where 要素を取得する．また，KNP を用いて主語要素，述語要素を取得する.「安倍は小泉前首相の靖国神社参拝問題のために途絶えていた中国，韓国への訪問を表明」という例では，人名として「安倍」,「小泉」，地名として「中国」,「韓国」がそれぞれ得られ，主語として「安倍」，動詞として「表明」が得られる．また，人名をタイトルとする記事のではタイトルとなっている人物が主語の場合，主語が省略されていることが多くみられる．そのため，主語がない文に関しては，記事のタイトルを主語として扱う．抽出に利用したツールによるものであるが，季節の「夏」が人名をして抽出されたり，人名「安倍晋太郎」の一部「安倍晋」が人名として抽出されたりといった場合ある．これらを排除するため，Wikipedia のリンクを用いる方法を考える．Wikipedia の各記事はリンクによってつながっている．すなわち記事内には記事のタイトルに関係のある記事へのリンクが含まれている．そのためこれらのリンク先の記事のタイトルは元の記事のタイトルに関係のある事柄であると見なすことができる．そこでリンク先の記事のタイトルを各要素の候補とし，上記の手法で抽出された要素のうちリンク先のタイトルとして出現するものを各要素とする．Wikipedia の記事は 1 個の実体につき 1 個以下の記事になっているため，このマッチングにより例えば記事中で「安倍晋三」,「安倍首相」,「安倍元幹事長」などのように様々に表現されるものにも「安倍晋三」という記事のタイトルに統一することができ，表記の揺れを吸収することができる．このマッチングには，文字列マッチングのほか，リンクのアンカーテキストやリダイレクトリンクの情報を用いる．これらはともにそれぞれの文字列をひとつの記事に対応させていることから，ひとつの実体を表す複数の表現を得ることができるためである．

記事内に出現する重要な名詞を，キーワード集合として扱う．Wikipedia には重要な概念に関しては，それをタイトルとする記事が存在する．その性質を利用し，事象に関する記述に含まれる名詞をすべて抽出し，そのうち，Wikipedia のタイトルとして出現しているものを，キーワード集合とする．

Wikipedia の記事のタイトルのうち，キーワード集合の候補として，対象の記事からリンクのある記事のタイトルを候補とする．このとき，複合名詞がタイトルとなっているものもあるため，名詞を結合したマッチング，及び，部分マッチングを行う．候補となる名詞は形態素解析の名詞とするため，複合名詞は最小単位の名詞に分割される．

連続する名詞をセットとして以下のルールを用いてマッチングを行う．ここではセット T=abc(a,b,c は名詞)

(1) 記事内からアンカーテキストとリンク先のタイトル組を収集するこの集合を L とする．アンカーテキストとリンク先のタイトルが一致しているものも含む．
(2) N=a(T の先頭の候補) とする．

Vol.2010-DBS-151 No.34
2010/11/13

(3) N を含む L のアンカーテキスト M のうち |N|/|M| が最小となるものを候補 (a,M) とする．

(4) N=b とし，N を含む L のアンカーテキスト M のうち複合名詞 ab を含むものがあれば，そのうち |ab|/|M| が最小となるものを候補 (ab,M) とし，候補 (a,M) を破棄する．

(5) N=c とし，以上の操作を繰り返す．

例えば「安倍晋三」の記事内の「与謝野官房長官」は形態素解析によって「与謝」「野」「官房」「長官」と分割される．まず「与謝」が「与謝野馨」にマッチングされ，次にそれが破棄されて「与謝野」が「与謝野馨」とマッチングされる．その後，同様に「官房長官」が「内閣官房長官」にマッチングされることとなる．ここで,「与謝野」を含むタイトルの Wikipedia の記事はほかに「与謝野町」,「与謝野晶子」などがあるが候補を「安倍晋三」からリンクがある記事に限定することでこれらを候補から外し，正しいマッチングを行うことができる．

3.3 日付情報の推定

3.3.1 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．

3.3.2 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．

3.3.3 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

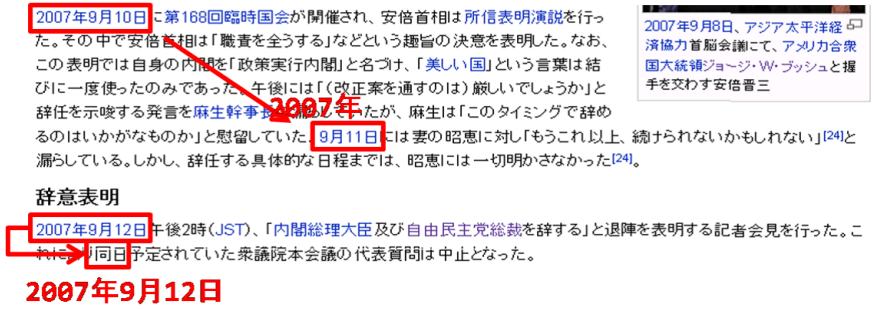

図 1 日付表現の補完
Fig. 1 Complementation of date explession.

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．例えば，2005 年 10 月 31 日という記述のみを持つ事象では「20051031」～「20051031」と推定し，2006 年 9 月 20 日，2006 年 9 月 26 日の記述を有する事象では「20060920」～「20060926」と推定する．

3.3.4 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．

3.3.5 文章構造の利用

3.3.2，3.3.4 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，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．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 に変換して利用する．

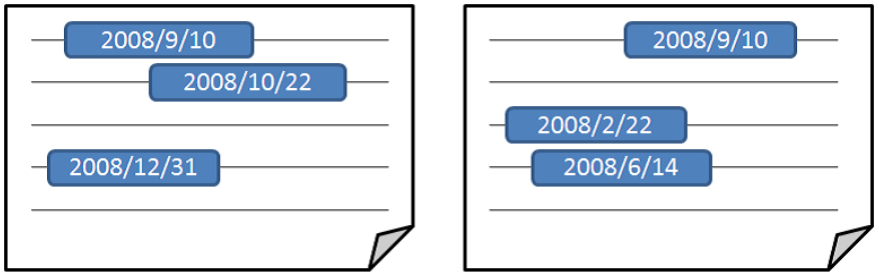

図 2 手法の適用範囲の決定の例
Fig. 2 Example coverage of the method.

3.3.6 記事中の日付表現記述を利用した手法の適用範囲の決定

ここでは 3.3.4 の手法を 3.3.5 で述べた文章構造のうち手法をどこに適用すべきであるかを述べる．

まず，3.3.2，3.3.3 の手法を用いて日付表現記述がある事象に関して発生日時の推定を行う．ここで発生日時が推定された事象に着目し，各文章構造内で事象の記述順が発生順と合致しているかどうかを判別する．そして合致しているとみなされた文章構造の範囲に 3.3.4 の推定手法を適用する．

記述順と発生順が合致しているかどうかを判別の手法について述べる．対象の文章構造内で，隣接する日付表現記述がある事象を比較し，その記述順と発生順とが合致しているものの数の割合を評価値とし，その評価値が閾値を超えている場合，その文章構造は記述順と発生順が合致していると判別する．例を図 2 に示す．図の四角を文章構造とし，青い四角は日付表現を持つ事象，中の数字は発生日時を示しているとする．このとき左の例では，1 番目の事象と 2 番目の事象，2 番目の事象と 3 番目の事象がそれぞれ記述順と発生順が合致しているため評価値は 2/2，すなわち 1 となる．右の例では，1 番目の事象と 2 番目の事象は記述順と発生順が合致しておらず，2 番目の事象と 3 番目の事象は合致しているため，評価値は 1/2，すなわち 0.5 となる．

3.4 同一の事象に関する記述の発見

各事象に関する情報は，複数の記事において記述がなされている．また，同一の記事内において複数の記述がなされている場合もある．各事象の発生日時に関する不整合を発見するためには，推定された発生日時を同一の事象に関する複数の記述間で比較を行う必要がある．ここでは同一の事象に関する記述を発見する手法について述べる．

ここで事象には様々な粒度があり,「ニューヨーク事務所，加古川製鉄所，東京本社で勤務した．(安倍晋三)」,「安倍晋三がサラリーマン時代勤務していた場所でもある．(神戸製鋼所加古川製鉄所)」のように同一であるという関係の他，一方が他方を包含しているという関係がある．例えば，事象 A：同年 9 月から 11 月にかけ，… 郵政造反組復党問題が政治問題化する（安倍晋三），事象 B：2006 年 9 月 26 日，… 造反無所属議員 12 人は … 自民党に賛同する行動をとり続ける (郵政造反組復党問題)，事象 C:見解を出す日である 2006 年 11 月 27 日，造反議員 12 人は復党届を執行部に提出した (郵政造反組復党問題) という 3 個の事象について考える．() 内はそれぞれの記述のある記事のタイトルである．事象 B，および事象 C は郵政造反組復党問題という事象すなわち事象 A に含まれる出来事である．また，これと同様に記述の仕方によって発生する包含関係もある．例えば，事象 A：12 月には，懸案だった教育基本法改正と防衛庁の省昇格を実現した (安倍晋三)，事象 B：現行法は，2006 年 12 月 15 日午後には参議院の本会議で成立した．(教育基本法)，事象 C：12 月 15 日 - 同法案 (防衛庁設置法等改正法案) が参議院で可決．(防衛省) に関して，事象 A における教育基本法改正は事象 B のことであり，防衛庁の省昇格は事象 C のことである．ここでは，このような包含関係にある事象の記述も含めて同一の事象に関する記述として扱う．

Wikipedia 記事内の同一の事象に関する記述を調査し，その結果から同一の事象に関する記述が満たすべき条件を設定した．ここで用いた条件は 3 項目である．1 項目は，発生日時に重なりがあること，2 項目は，関連人物に少なくとも 1 人は共通して出現する人物がいること，3 項目は発生場所の少なくとも 1 カ所は共通していることである．しかし，省略などの可能性があるため，片方の事象に発生日時，関連人物，発生場所の情報がない場合にはそれぞれ対応する条件を満たす必要はないものとする．

収集されたすべての記述の組に対して上記の 3 項目でフィルタリングを行い，通過した組を候補としてさらに同一の事象に関する記述であるか否かの判別を行う．記述の組に対し，それぞれの要素間で語の出現数により，特徴ベクトルを作成し，コサイン類似度を算出する．このコサイン類似度をそれぞれの属性における評価値とし，各記述の組に対し評価値の列を与える．正例と負例それぞれの記述の組の評価値列に対し SVM による学習を行い，そこから得られた基準で未知の記述の組に対して判別を行う．正例は，人手で発見した同一の事象に関する記述とし，負例は，上記の 3 項目を満たしたものの内，同一の事象に関する記述ではないものからランダムに選択し利用する．

ここで人物，組織名，キーワードに関しては，すべて wikipedia の記事のタイトルとなっているため，Wikipedia のひとつの実体にひとつの記事が対応するという性質から，表記

の揺れが排除されており，単純にテキスト比較することが可能である．一方，場所の表記に関してはその粒度は様々であることが考えられる．例えば同じ場所を指すのに,「ATC」「大阪」「日本」など複数の表記が考えられるためである．そのため，ここでは GeoNames を用いて，それぞれの地名を包含する地名，例えば区を包含する市町村，またそれを包含する都道府県，国名を獲得し，それらを含めて比較を行う．GeoNames とは，各国の地図製作，統計，郵便の各当局ならびにアメリカ陸軍，Wikipedia などさまざまなソースから集めた無料データを統合して作成された地理情報データベースである．GeoNames のデータは英語であるため，地名の英語表記を取得する必要がある．ここで，出現する地名は Wikipedia の記事のタイトルとなっている．そのため，Wikipedia の言語間リンクを用いることで地名の英語表記を獲得することが可能であり，それを用いて GeoNames のデータとの照合を行う．

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，それらの記事と，それらの記事へのリンクを持つ記事から事象データベースを作成した．

4.1 事象情報の収集

記事内のテキストを句点で区切られる文章に分割し，各記述ごとに発生日時，発生場所，関連人物，組織，キーワード集合の各要素を抽出する．事象の発生日時の取得及び推定に関しては別途述べる．

安倍晋三の記事から情報を取得し，100 個の事象をランダムに選択して評価を行った．この結果の評価は，事象を表す適切な情報が取得できているかどうかではなく，文章内の人物名，地名，組織名が正しくとれているかどうかを適合率，再現率を用いて評価した．また，キーワード集合に関しても，キーワードとして適しているかどうかではなくマッチングされた Wikipedia のタイトルと文章中に出現している名詞が同一のものを指しているかどうかの評価を行った．また，固有表現であるかどうかの判断は IREX の固有表現抽出課題の定義に基づいて行った．結果を表 1 に示す．改善手法適用前とは NExT によって抽出されたものをそのまま用いたものである．表 1 における A は本来の固有表現の一部のみ抽出されたもの（略称，名字のみ，名前のみなど）や固有表現に不要な部分もともに抽出してしまっているもの（役職名やその一部など）も正解と判断したものであり，B はそれらを不正解としたものである．人物においては A では高い数値が得られているものの，B ではかなり低い数値となっていることから，人物を一意に特定出来るだけの情報が得られていないことが分かる．これに対し，改善手法では一定の数値を保ったまま Wikipedia のタイトルを取得しているため人物を一意に特定することができている．これは事象の識別において重要な情報である．一方で組織については B に対しても低い数値になっていることから手法の改善が必要であることが分かる．低い数値になっている原因の多くは，本文中の「自民党幹事長」から NExT によって「自民党」が抽出されているがそれを「自民党幹事長」にマッチングするというような状況によって発生している．これに対しマッチングの手法の改善とともにマッチングされた記事の情報からそれが組織であるかどうかを判定することを検討している．場所に関しても同様の問題を抱えている．

表 1　情報の取得
Table 1　Collecting information

	人物		組織		場所		キーワード
	適合率	再現率	適合率	再現率	適合率	再現率	適合率
改善手法不適用 A	.74	.81	1.0	.45	.64	.50	
改善手法不適用 B	.29	.31	.81	.37	.44	.34	
改善手法適用	.74	.78	.20	.26	.59	.26	.44

4.2 日付情報の推定

提案手法を適用し，記事のうち「安倍晋三」の記事を使用して結果の評価を行った．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．各事象の正解の時区間は正解は人手で作成した．この正解は，事象が実際に発生した日時であり，必ずしも記事内の情報のみから得られるものではない．

$$e = \frac{|T_1 \cap T_2|}{|T_1 \cup T_2|} \quad \cdots (1)$$

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．

3.3.4 で提案した手法を，文章構造を利用せず記事全体に適用した場合，特定の文章構造の内部で適用した場合，3.3.6 で提案した手法で適用範囲を決定した場合で実験を行い結果を比較した．特定の文章構造の内部で適用する際には，Section 内の情報に適用する場合 (S) と，Paragraph 内，ListItem 内，Definition 内のそれぞれの情報のみを適用する場合 (P) で行った．これは，Paragraph，ListItem，Definition は Section の内部で，すなわち子要素として利用されているためである．また，手法の適用範囲を決定する際に用いる閾値は

Vol.2010-DBS-151 No.34
2010/11/13

表 2 推定を行った件数
Table 2 Count of date estimation

	両側推定		片側のみ推定	情報なし	合計
	日付情報あり	なし			
構造を用いない	183	323	0	0	506
S	183	168	136	19	506
P	183	32	141	150	506
0.4	183	230	27	66	506
0.5	183	191	38	94	506
0.7	183	132	21	170	506

表 3 結果の評価値
Table 3 Evaluation value of date estimation

	全事象	両側+片側	両側
全体	0.2576	-	-
S	0.2371	0.2463	0.3007
P	0.2605	0.3703	0.4668
0.4	0.2775	0.3191	0.3150
0.5	0.2796	0.3434	0.3478
0.7	0.2831	0.4263	0.4101

表 4 結果の評価値 (日付表現記述を持たない事象)
Table 4 Evaluation value of date estimation(events without date explession)

	全事象	両側+片側	両側
全体	0.0601	-	-
S	0.0732	0.0778	0.0550
P	0.1099	0.2053	0.1264
0.4	0.0913	0.1148	0.0834
0.5	0.0947	0.1335	0.1003
0.7	0.1000	0.2112	0.1382

0.4,0.5,0.7 として行った．結果を表 2,3,4 に示す．

表 2 は，手法を適用した結果，各記述が指す事象の発生日時を推定できたものの件数である．両側推定とは，事象の開始日時と終了日時の双方が推定できたものである．片側推定とは開始日時と終了日時のどちらか一方が推定できたもので，情報なしとはどちらも推定できなかったものである．また，日付表現ありとは記述に日付表現が含まれていたものである．表 3,4 は，推定した日付情報の評価値である．今回の推定では，特に記述に日付表現が含まれていないものに着目していたため，表 4 では記述に日付表現を含まないものだけを評価した結果を示している．また，評価値における「全事象」,「両側+片側」,「両側」に関しては，出現する全事象の評価値の平均値 (全事象)，推定しなかった事象を省いた評価値の平均値 (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の平均値 (両側) である．補完・推定の範囲を狭める本手法は，誤った推定を行わなくすることを目的としているため，正しい推定が増加することだけではなく，誤った推定が減少していることを確認するためにこれらの評価値を求めた．

表 5 正解時区間別の評価値
Table 5 Evaluation value of each time section

正解期間	両側+片側	両側
年	0.2835	0.3071
月	0.0827	0.0970
日	0.2128	0.1013

表 6 正解時区間の一部を含む件数 (両側のみ)
Table 6 Evaluation value of each time section(estimationg both sides)

正解期間	該当件数	全件数
年	24	24
月	10	13
日	68	95

両側+片側，両側に関して，より手法の適用範囲が広い場合の評価値より狭い場合の評価値が優れているということは，誤った推定を減らすことができたことを意味している．誤った推定を行われていた事象は評価値が低く，それを推定せずに片側や，情報なしにすることによって全体の評価値計算に含まなくするためである．しかし，日付表現記述を持たない事象に注目してみると P では 317 件のうち両側が推定できているものは 32 件となっており，推定の範囲を減らしすぎていると考えられる．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．

記事内の日付表現から推定手法を適用する範囲を決定する手法では，閾値を高く設定することで手法の適用範囲が狭まるため，推定を行えた事象数が増加し，行えなかった事象数は減少する．一方で推定の評価値は減少するというように，評価値と推定ができた事象数はトレードオフの関係にある．しかし，日付表現記述を持たない事象に関して，こちらは P と比較し，近い評価値を保ちながら，両側推定ができた件数を大幅に伸ばすことができている．このことからこの手法がより効果的に事象の発生日時の推定を行えていることがわかる．

ここまで，手法間での優劣を調査した，次に，推定がどの程度の効果を示しているか，閾値 0.7 の時の結果を元に調査した． 表 5 は正解時区間が, 日単位 (30 日以内)，月単位 (365 日以内)，年単位 (366 日以上) に分けて評価値を求めたものである．この値から，日単位のものに関しては 1/10 以上であり，平均して前後 10 日程度の範囲で推定が行われていることが分かる．また同様に，月単位のものに関しては 1/12 以上であることから 1 年程度の範囲で，年単位のものに関しては 1/3 程度であることから前後 1 年程度の範囲で推定ができ

ていることが分かる．また，表 6 は，両側推定されたもののうち，正解時区間と共通範囲を持つものの件数である．ここから高い精度で正解時区間を含んだ推定ができていることが分かる．この件数は推定範囲を広くとることで大きくすることができるが，表 5 の評価値と複合的にみることで，どの程度の範囲で，どの程度の正確さで推定が行われているかを確認できる．

4.3 同一の事象に関する記述の判別，記述間の不整合の発見

学習データを小泉純一郎の記事に含まれる記述とこの記事からリンクの張られている記事に含まれる記述との比較により作成し，それを用いて作成したモデルデータを元に，安倍晋三の記事に含まれる記述とその記事からリンクの張られている記事に含まれる記述との組に適用し，判別を行った．ここで用いたデータは，情報の抽出において改善手法を適用したものであり，日付情報の推定においては閾値を 0.7 として行ったものである．学習及び判別には SVM light[12] を用いた．安倍晋三の記事に含まれる記述の数は 442 であり，判別対象となった記述の組は 20373752 組であった．このうち 3 項目の条件を満たしたものが 9079246 組であり，165802 組が同一の事象に関する記述であると判別された．今回正解としたのは 47 組である．内 24 組が同一の事象のものであると判別されいたことが確認された．不正解を多く正解としてしまっているだけでなく，正解とするものも約半数落としてしまっている．原因として学習に用いたデータに正例が少なすぎたこと，負例に評価値列が 0 並びであるものが多すぎたことなどが考えられる．また，各事象の人物場所組織について情報の取得についても改善が考えられる．また今後，各属性間の評価値や学習手法などを見直し，結果の改善を図る．

5. おわりに

本稿では，Wikipedia の記事から事象データベースを構築する手法を提案した．作成した事象モデルを元に情報を収集し，同一の事象に関する記述を収集することで，様々に記述される事象に関する情報を統合したり，記述間に存在する不整合を発見することが可能となる．事象モデルでは Where,What,Who,When に関する情報を用いた．事象の発生日時に関しては記事内に明示的に記述されている情報だけでなく潜在的に有している情報を推定する手法を提案し，実験を行い，効果を検証した．

引き続き，同一の事象に関する記述の判別，記述間の不整合の発見の実験を行い，結果の分析，手法の改善を行う．さらに同一の事象に関する記述から，情報の不整合を発見する手法についても検討する．

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．

参考文献

1) Rui Kimura,Satoshi Oyama,Hiroyuki Toda,Katsumi Tanaka Creating Personal Histories from the Web Using Namesake Disambiguation and Event Extraction Proc. of the 7th International Conference on Web Engineering (ICWE2007), LNCS, Vol.4607, July 2007, pp.400414
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) 渡邉 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) 道下 智之，中山 浩太郎，原 隆浩，西尾 章治郎, ウィキペディアを用いた語義曖昧性解消手法と 情報検索への応用, 人工知能学会全国大会（第 23 回）, 2009
8) 桜井慎弥，手島拓也，森田 武史，和泉 憲明，山口 高平，Wikipedia オントロジーに基づくドメインオントロジー構築支援環境の実現と評価, 人工知能学会全国大会（第 23 回）, 2009
9) 固有表現抽出ツール NExT,http://www.ai.info.mie-u.ac.jp/next/
10) 日本語構文解析システム KNP,http://www-lab25.kuee.kyoto-u.ac.jp/nl-resource/knp.html
11) 日本語形態素解析システム JUMAN,http://www-lab25.kuee.kyoto-u.ac.jp/nl-resource/juman.html
12) サポートベクターマシン SVM light,http://svmlight.joachims.org/