時間表現抽出と文間の関係による事件記事の時間構造解析

Temporal Structure Analysis for Newspaper Article
by Temporal Expression Extraction and Sentence Relation Analysis

大山 朋敬†　　田村 直良‡
横浜国立大学大学院環境情報学府†
横浜国立大学大学院環境情報研究院‡
tomonori@tamlab.ynu.ac.jp†
tam@ynu.ac.jp‡

１　はじめに

本研究は、新聞記事、特に事故や窃盗などの事件記事に関して、文章内にある動作記述に注目し、発生した動作間の時間的関係を解析することを目的とする。

近年、インターネットの普及や、計算機技術の発達などによって電子化ドキュメントの数は膨大となっている。このような状況において、それぞれのドキュメントの内容を効率的に理解するため、内容要約やキーワード抽出などの研究が活発に行われている。

ここで、特に動作記述に対する内容理解に関しては、動作間の時間関係を把握することが重要であると考えられる。新聞記事、特に事件に関する記事に注目してみると、記事の内容は、「何が起こったか」という記述が中心となっている。たとえば次のような文章がある。

横浜市内の中学一年の男子生徒（１３）が十二月二十六日、家を出たまま行方が分からなくなったが、県警が三十一日、同県市内の旅館にいるところを保護し、一緒にいた容疑者（５８）を未成年者誘拐の疑いで逮捕。[1]

このような文章では、各動作の発生時間が文章内の出現順とはなっておらず、記述内容によって、各動作の時間関係が付けられていることが分かる。これらの前後関係を正確に把握することで文章の理解を手助けすることになる。

文章の情報抽出に関しては、固有表現抽出の研究がある。固有表現は、時間に限らず、人名、場所などを含んでいるが、時間情報を抽出できる点では同じである。NeXT[1]やNameLister[2]など固有表現抽出の商用製品も開発されている状況であり、固有表現抽出に関しては研究が進んでいる。しかし、表現抽出のみでは動作記述間の時間関係は解析できず、もっと別の情報を解析していく必要がある。また、ブログ記事に限定を行い、動作記述の生起時間帯判定の研究もおこなわれている[3]。これは、ブログ記事内では時間幅が限定されるので、各動作記述の生起時間帯を朝、昼、夕、夜に分類分けを試みている。しかし、対象がブログ記事であるため時間幅の限定が有効であるが、もっと一般的な文章においては、その限定を利用することができない。

そこで本研究では、事件記事は動作記述が主となる内容であることに注目し、事件記事の内容理解には、記述された動作間の時間関係の把握が重要であり、それには総合的な情報解析の必要があることを述べる。また、本研究では、時間に関して、その表現の分類、文章のセグメント化、構造解析によって動作間の時間関係の解析を行う。

以下２章で時間表現抽出、３章で文章の時間セグメント作成、４章で文章構造解析、５章で評価実験・考察に関して述べる。

２　時間表現抽出

本研究では、動詞句によって記述される動作をイベントとよび、イベントの発生時間を表現している文節を時間表現と呼ぶことにする。ここでは特に時間表現の抽出に関して記述する。

２．１　時間表現の分類

本研究における時間表現は、１文節を１単位としている。時間表現の分類に関しては、大きく２種類に分類される。また、間接表現はさらに３種類に細分類される。

直接表現　時間を直接表現しているもの。

（1）*１９９０年、５月、２時、など*

間接表現　時間を直接表現しているわけではないが、基準からの時間など、相対的に時間を表現しているもの。

照応表現　指示代名詞により表現される時間表現で、先行詞の時間を基準とするもの。

（2）*それから、その時、など*

相対表現　基準の時点からの相対的な時間で表現されているもの。

[1] 実際の記事を参考に創作

（3） *明日、三日後、など*

従属複文 従属文の示すイベントが、その動作時間を示すように表現されているもの。

（4） *焼いた後、投げてから、など*

２．２　時間表現の構文的分類

前述の時間表現はそれぞれ構文的には表１のように分類される。

意味分類		構文的分類
直接表現		名詞句
間接表現	照応表現	指示代名詞＋名詞後置詞
		指示代名詞＋助詞
		手掛かり語
	相対表現	名詞＋名詞後置詞
		名詞＋助詞
		手掛かり語
	従属複文	文＋名詞後置詞
		文＋助詞
		文＋手掛かり語

表１．時間表現の意味分類と構文的分類

２．３　時間表現の解析

時間表現の抽出は、基本的にパターンマッチと構文情報を利用して行う。しかし、直接表現や相対表現の抽出は名詞句に関して行うが、構文情報だけでは対象パターンが限りなく多くなってしまう。そこで、対象となる名詞句の意味を考慮し、パターンマッチと併せて機械学習による判定器を用いる。以下に時間表現抽出のアルゴリズムを示す。

また、形態素解析には茶筌[4]を、構文解析には南瓜[5]を利用して抽出・解析実験を行う。

ステップ１　形態素・構文解析を行い、各文節について、動詞句とその他の句に分類分けを行う。

ステップ２　それぞれの文節に対して、以下の判定を行う。

ステップ２．１　形態素情報、パターンマッチング、係り受け関係から、従属複文と照応表現の判定を行う。判定された場合は次の文節へ。

ステップ２．２　機械学習によって、時間表現であるかどうかの判定を行う。

２．４　機械学習による時間表現判定

前述の分類である相対表現と直接表現に関しては機械学習によって判定を行う。学習器にはC5.0[6]を利用する。以下に利用した素性を示す。判定は文節ごとに行う。

- 助詞を除いた最後の形態素の品詞
- 名詞句の最終形態素の意味分類
- 南瓜の解析結果にDATAかTIMEタグがあるか
- 助詞の種類

２．５　判定器の作成

前述の判定器の作成に関して、以下の予備実験を行った。テストデータは、毎日新聞１９９５年の事件記事約１０００件から、無作為に抽出した約７０００文節（副詞句、動詞句はのぞく）に対して、それぞれ時間表現であるかどうかを被験者２名によって判断を行った。判断が異なる文節に対しては筆者の一名が判断した。表２に、テストデータを用いて作成した判定器の判定結果に対するクローズテストと５分割交差検定の結果を示す。

		判定結果			
		クローズテスト		５分割交差検定	
		TRUE	FALSE	TRUE	FALSE
正解	TRUE	3401	303	3374	331
	FALSE	291	2872	424	2739
精度		91.3%		89.0%	

表２　時間表現判定の学習結果

３　時間セグメント作成

通常、文章とは、ある程度のまとまりで意味分けがされており、特に動作記述が中心の文章では、時間的にセグメント分けすることができる。

そのため、動作間の時間関係を特定するためには、時間表現の抽出をするとともに、文章を時間的なまとまりでセグメント分けすることが重要である。そこで、このような時間的なまとまりを時間セグメントということにする。ここでは文章の時間セグメント分割について述べる。

３．１　時間セグメントの設定

セグメントはイベントの集合で構成され、基本的には前述した時間表現か、あるいはセグメントを作成する手掛かり語(以下セグメントワード)が出現した直前のイベントで区切りをつけるが、それ以外にも動詞のテンスや、段落情報からも判断を行う。以下にセグメント設定の詳細を記述する。

ステップ１　改行などの、文章作成者が故意に作成したと思われる部分をセグメントの境界とする。

ステップ２　セグメントワードを抽出し、ステップ１で作成した各セグメントに対し、セグメントワードあるいは時間表現の直前のイベント記述でセグメント分割を行う。

ステップ３　ステップ２までで作成された各セグメントに対し、現在から過去、過去から現在というように、テンスの整合がとれていない部分を分割する。

３．２　時間セグメント間の時間順序関係解析

時間セグメント内のイベントの時間順序関係は原則出現順になっているため、改めて解析を行う必要はない。そこで、時間セグメント間の時間関係の解析を行う必要があるが、時間セグメント間の時間関係は以下の順序で解析を行う。

ステップ１ 時間表現によりセグメント分けされた時間セグメント同士の時間関係を、セグメントワードあるいは時間表現間の関係を解析することで特定する。

ステップ２ 時間表現のない時間セグメントについては、テンスの情報や段落情報から、前後に出現したセグメントとの時間関係を特定する。

ステップ３ ステップ１とステップ２の情報から、すべてのセグメント間の時間関係に関して推移律を適用し、例えば、AはBより後で、BはCより後なら、AはCより後であるというような推論を行う。

４　文章構造の解析

前述した時間表現と時間セグメントを利用することによって、主文の記述に対しての時間関係は解析が可能である。しかし、それだけでは情報欠如が引き起こされてしまうため、構造的な情報も解析する必要がある。ここでは考慮する構造と、時間セグメントとの関係について述べる。

４．１　時間ユニット

次のような例を考えてみる。

（5）　*三十一日、昨日釈放されたA氏が再逮捕された。*

この場合、主文の記述のみに注目してしまうと「A氏が昨日釈放された」という情報が欠落してしまう。そのため、このような入れ子構造も考慮する必要がある。ここで、入れ子構造により作成される時間的まとまりを考慮したものを時間ユニットと定義する。すると、この例では、「昨日釈放された」と「三十一日・・・再逮捕された」がそれぞれ時間ユニットにあたる。後者は前者を含んだ構造になっており、時間関係の比較を行う場合には、包含関係を考慮しなくてはならない。この情報欠如に対応するため、以下の２点を考慮する。

- 従属節中のイベント記述

（6）　*お昼を食べた時に、*

- 名詞句への修飾節

（7）　*昨日作ったケーキを今日食べた。*

４．２　時間ユニットによる時間関係の比較

先ほどの時間ユニットは、時間セグメントの情報も併せて持っているため、時間ユニットを利用して時間関係の比較を行う。しかし、通常の比較では時間表現のない時間セグメントまたは時間ユニットとの関係は解析できないが、構造的な情報を利用すれば、時間関係を解析できる場合がある。

名詞句への修飾節に関して、以下に例を挙げる。

（8）　*二十日に、持ち帰ったケーキを食べた。*

この場合、時間ユニットは「持ち帰った」と「食べた。」が作成される。時間ユニット「持ち帰った」に時間表現がないため、通常は両者の関係を解析できない。しかし、時間ユニット「持ち帰った」は時間ユニット「食べた。」内で入れ子構造になっているため、「入れ子構造になっているイベントの時間関係は原則出現順になっている」という規則を適用することができる。この規則の適用により、イベント「持ち帰った」はイベント「食べた。」よりも前に起こったと考えることができる。

５　評価実験・考察

これまでに述べた解析手法を用いて、時間ユニット間の時間関係を把握するプログラムを作成した。以下に結果と考察を示す。

５．１　判別例

プログラムは、各ユニット間の時間関係を判別した結果を出力とした。基準をFrom、比較対象をToとし、時間的にFromがToより過去の場合は１、FromがToより未来の場合は－１、FromとToが同時の場合は０、判断がつかない場合は？を付与し、縦をFrom、横をToにとったマトリクスで出力される。以下の文章に対する出力例を表３に示す。()内の数字はユニット番号を示す。

Aちゃんが殺された(1)事件で、三十一日夕未成年者誘拐容疑で前日逮捕した(2)書籍販売業、B氏を殺人、死体遺棄容疑で再逮捕した(3)。

調べによると、B氏は二十八日午後五時十分ごろ、近所の友人宅から戻って来た(4)Aちゃんを殺害した(5)疑い。[2]

５．２　評価実験

毎日新聞１９９５年[7]の事件記事の中から無作為に選んだ事件記事５件に対して実験を行った。正解データとして、各記事に対して前述のマトリクスを人手で作成し、プログラムの出力結果と比較を行い、精度と再現率を求めた。精度、再現率は次を用いる。結果を表４に示す。

精度　＝　出力結果の中で正解データと一致した数／プログラムが出力したイベントの組み合わせ総数

再現率　＝　正解データの中で出力と一致した数／正解データのイベントの組み合わせ総数

また、時間セグメント間の時間関係の解析も行った。結果を表５に示す。

---

[2] 実際の記事を参考に創作

		TO				
	ユニット番号	1	2	3	4	5
FROM	1	－	？	？	？	？
	2	？	－	1	−1	−1
	3	？	−1	－	−1	−1
	4	？	1	1	－	1
	5	？	1	1	−1	－

表３　結果出力の例

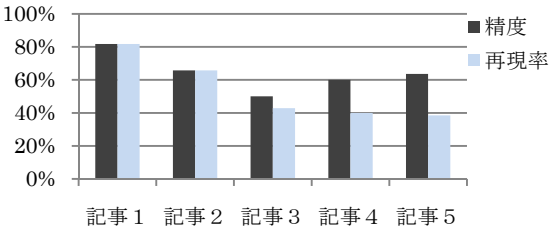

表４　ユニット間の時間関係比較結果

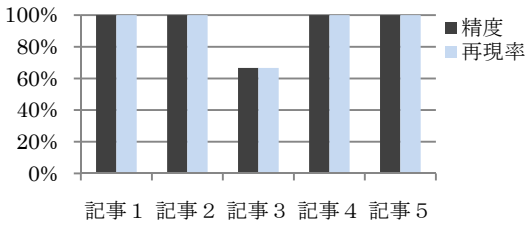

表５　セグメント間の時間関係比較結果

５．３　考察

セグメント間の比較結果をみると、今回提案した手法により、時間関係は十分に把握することができると判断できる。しかし、ユニット間の比較結果を見てみると、記事１、２と比較して、記事３，４，５の結果が良くなかった。これは、形態素・構文解析の解析誤りが大きな要因であった。また、結果には表れていないが、文章に表現上の問題があった。以下に詳しく述べる。

形態素・構文解析の解析誤り

特に文節区切りに関する誤りが多かった。もともと事件記事は長文ではないため、絶対的な文節数が少ない。そのため、１つの文節の解析誤りが結果に大きな影響を及ぼすことが多かった。１つの文節が解析できないだけで５%～１０%と大きく結果が変わってしまったようである。記事３，４，５は精度に対して再現率が特に悪かった。これは、動詞句がうまく抽出しきれていなかったことが原因であった。例えば「書き換え所有者に」が１文節と解釈されていた。事件記事は基本的に、動作記述のあとには読点が入るように記述されているが、記事３，４，５はそれが徹底されておらず、このような解析誤りが多かった。

また、すべての記事に対して、係り受け関係の解析誤りや形態素の分割誤りなどもあり、結果を下げる要因となっていた。

文章の表現上の問題

記事の中には時間表現がほとんど含まれていないような記事も見受けられた。そのような記事は、ほとんど時間関係を特定することができていなかった。しかし、人間が文章を呼んだ場合に、時間表現がないが、時間関係を特定できる状況として以下の３点が挙げられた。

・表現が違うが、同一の内容を指している
・文脈により明らかに時間関係が特定できる
・各動詞の意味的な因果関係により特定できる

今回提案している手法ではこの問題については取り扱わなかったが、この問題の解決が実際の時間関係把握には重要であることがわかった。正確に時間関係を把握するためには文脈やそれぞれの語彙について細かく解析を行っていく必要があると思われる。

６　おわりに

時間表現の抽出と、文章の構文情報を利用して、イベント間の時間関係を解析する方法を述べた。そして、プログラムを作成し、実際の記事に適用できるかどうかを調べた。セグメント間の時間関係の把握については、時間表現や構文情報により時間関係が特定される場合も多く、時間関係の把握には有効であることが分かった。

しかし、これらの情報だけではイベント間の時間関係を正確に把握することができない場合もあることが分かった。文脈やそれぞれの語彙に関する細かい情報を駆使して解析を行うことでより正確な時間関係の把握ができるようになると思われる。

謝辞

本研究では、「毎日新聞全文記事（１９９５年度版）」を利用した。

参考文献

［1］三重大学工学部 Named Entity Extraction Tool(NExT)
http://www.ai.info.mie-u.ac.jp/~next/next.html
［2］NTT 日本語固有表現抽出「NameLister」
http://www.ntt-tec.jp/technology/A122.html
［3］野呂太一ら：イベントの生起時間帯判定、情報処理学会論文誌 Vol.48 No.10 pp.3405-3414.
［4］奈良先端科学技術大学院：形態素解析機 Chasen（茶筌）http://chasen-legacy.sourceforge.jp/
［5］工藤拓：日本語係り受け解析器 CaboCha（南瓜）
http://chasen.org/~taku/software/cabocha/
［6］RULEQUEST RESEARCH: Data Minig Tools See5 and C5.0
http://www.rulequest.com/
［7］毎日新聞「毎日新聞全文記事（１９９５年度版）」