文の時間表現の分類と時間表現抽出システム

Classification and Extraction System of Temporal Representation

榎村 麻里 †　　田村 直良 ††

† 横浜国立大学教育人間科学部
†† 横浜国立大学大学院環境情報研究院
{emura, tam}@tamlab.ynu.ac.jp

1　はじめに

本論文では、日本語文章における時間表現の出現パターンを分類し、これを抽出するためのモデル化および、文章中から時間表現を自動抽出するシステムについて述べる。

インターネット上の新聞記事や日記等、様々な出来事について記述された電子文書が膨大な量で入手できる現在、これらを自動的に解析する必要性が高まっている。出来事の発生時間や順序を自動的に解析することができれば、膨大な文書の中から特定の時間や時期に発生した出来事を検索したり、複数の出来事を前後関係によって並べ替えたりといったことができる。

このような文章中での文間の時間的な参照関係と、文が記述する出来事間の時間的な関係を、文章の**時間構造**と呼ぶことにする。文章の時間構造を解析するためには、まず各文から時間に関する表現を抽出することが必要である。

現在行われている時間表現の抽出に関する研究は、渡辺ら [1] の固有表現抽出の研究における時間表現の抽出など、直接的な時間表現のみを抽出するものが主である。他に、土屋ら [2] の研究は、概念の類似度を用いて暗示的に時間を示す語から時間を判断する。また、野呂ら [3] の研究では機械学習を利用して時間を連想させる表現を判別している。いずれも語から直接特定の時間を示す表現に関する研究である。

だが、文章中に記載されている時間に関する情報は、このような直接特定の時間を表す表現だけではなく、ある時点への参照を用いて、「〜を行った際」「その後」などのような、間接的な表現でもしばしば記載される。このため、文章から時間に関する情報をすべて抽出するには、間接的な時間表現も抽出することが不可欠である。

そこで本研究では、以下の手順によって、間接的な表現を含めた時間表現の抽出を試みる。

まず、時間表現の出現パターンをいくつかに分類する。そして、文の係り受け構造により時間表現をモデル化し、これを利用して時間表現を抽出するシステムを作成する。システムは、抽出エンジンと抽出ルールの集合から構成される。抽出結果は XML の形式で出力される。

2　時間表現の分類とモデル化

2.1　対象とする時間表現

出来事の開始や終了、状態の変化を**事象**と定義する。本研究が対象とする時間表現は、文の述語が記述する事象の発生時間あるいは状態の継続期間を限定する表現である。

例１　今日の空は青い。

例２　今日は空が青い。

例３の「今日の」は「空」という名詞に対する修飾であるため、直接事象の発生時間を記述してはいないので、時間表現としない。例４の「今日は」は「青い」という述語に対して時間を限定しているので、時間表現である。

なお、特定の表現が暗にある時期や時間を連想させるものは含まない。例えば、

例３　日が昇り、あたりが明るくなった。

例４　日が昇ったとき、あたりが明るくなった。

このような場合、例５は時間表現とはせず、例６の「日が昇ったとき」を「明るくなった」に対する時間表現（時間格）とする。

2.2　時間表現の分類

時間表現を、**内部参照を用いた間接的時間表現**、**外部参照を用いた間接的時間表現**、**直接的時間表現**の３つに分類する。内部参照を用いた間接的時間表現、外部参照を用いた間接的時間表現についてはそれぞれを表層表現の形によって、更に２つずつに分類する。

1. 内部参照を用いた間接的時間表現

(a) 指示代名詞

(b) 連体詞の指示詞＋名詞

2. 外部参照を用いた間接的時間表現

(a) 文＋名詞

(b) 名詞＋名詞

3. 直接的時間表現

2.2.1 内部参照を用いた間接的時間表現

内部参照を用いた間接的時間表現は、照応詞を用いた時間表現であり、文中にその参照先（先行詞）が存在する。参照先が示す時間によって文が記述する事象の時間を限定する。

(a) 指示代名詞

指示代名詞は、「ここ」、「そこ」、「これ」、「それ」などである。この指示代名詞が時間を表す名詞を指す照応詞である場合、時間表現となる。しかし、時間を表す名詞以外を指している場合もあり、これだけでは判断できない。

例）「それから」

(b) 連体詞の指示詞＋名詞

時間を表す名詞に指示詞が係っている場合、その指示詞が指す参照先自体の時間もしくは参照先の事象の発生時間、状態の継続期間を表す時間表現となる。

例）「その時」、「この際」

2.2.2 外部参照を用いた間接的時間表現

外部参照を用いた間接的時間表現は、間接的な時間表現ではあるが、その参照先が文内、文章内に存在しなくてもよい。文が記述する事象の時間的制約を示すために、別の事象について記述し、別の事象の時間との関係によって時間を限定する。

(a) 文＋名詞

時間を表す名詞に、事象を記述する文が係っているとき、この事象の時間との関係によって表される時間表現となる。

例）「〜した時」、「〜が起こった瞬間」、「〜する前」

(b) 名詞＋名詞

時間を表す名詞に、事象を記述する名詞が係っているとき、この事象の時間との関係によって表される時間表現となる。

例）「発生前」、「事件前日」

2.2.3 直接的時間表現

直接的時間表現は、時間を表す固有表現を用いることによって直接に事象の時間を限定する。

例）「２００７年」、「2時５分前」、「元旦」

2.3 時間表現モデル

2.3.1 文の構造

前節で述べた分類に基づき、文の係り受け構造に着目した抽出を行う。そのために、文を係り受け構造に基づく木構造として表現する。

文構造木を以下のように定義する。

- 木のノードは文中の節である。
- 木において、子ノードはその親ノードが係る節である。

2.3.2 ルールとしての時間表現抽出

本研究では、前述の時間表現の分類により文の係り受け構造を基にしたルールにより時間表現の抽出をモデル化する。

ひとつの抽出ルールは、**ルール ID**、**パターン**、**適用条件**、**動作記述**から成る。

ルール ID ルールを識別し、どのルールによって抽出されたのかを示すために用いられる。

パターン ルール適用について係り受け構造を文構造木上のパターンとして表現したものである。例を図 1に示す。

適用条件 パターンに合致した各節のについて、意味的な条件を記述するものである。

動作記述 適用条件が成立した場合の動作について記述したものである。本システムは、結果を XML により記述するため、抽出した節にタグをつける際どのような属性をつけるかを指定するものである。

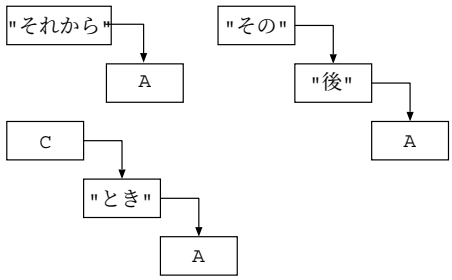

図 1: パターンの例

2.3.3 抽出ルールの記述

抽出ルールの記述で用いられる記法を示す。

まず、例を示す。このルール（shiji01）は、直列的（[A,[B,[C]]]）に、指示詞（変数C で与えられる）が時間を表す名詞（変数B で与えられる）に係り、それが述語（変数A で与えられる）に係っているという（構文的）パターンで規定され

る。さらに、適用条件部では、それら（A∼C）の文法カテゴリを各判定関数によりに規定する。

動作記述では、B が時間を表す名詞で、C が時間についての参照であることを示す XML のタグの属性を付加するための関数呼び出しである。

なお、jyoshi_1(B) は、特定の助詞を持つ名詞節であるかをみる判定関数である。

```
name: shiji01
pattern: [A, [B, [C]]]
condition: is_jyutsugo(A) &&
 is_jikan_meishi(B) && is_shijishi(C) &&
 jyoshi_1(B)
action: tag_time(B), tag_timeref(C)
```

3 時間表現抽出システム

時間表現抽出システムは、前章で述べた時間表現モデルに基づいて、文中から時間表現を抽出しタグを付与するものである。本章ではその実現方法について述べる。

3.1 抽出システムの構成

システムの動作は、以下の通りである（図 2）。

1. あらかじめ、抽出ルールを Perl のプログラムに変換する。
2. 解析対象となる文章を、茶筌、南瓜により形態素解析、構文（係り受け）解析し、文構造木（の列）を生成する。
3. プログラムに変換されたルールごとに文構造木上でパターンマッチを試みる。
4. マッチしたルールに基づき動作部の動作を実行する。

抽出結果を XML 形式で出力するように構成されており、抽出ルールの動作部は、 XML のタグの属性を記述している。

全体として、 XSLT により結果の表示形式を定義し、ブラウザで表示させる。

3.2 ルールトランスレータ

ルールトランスレータは人手で記述された抽出ルールを、 Perl プログラムの一部としてそのまま実行できるような形式に変換するプログラムである。

例えば、 2.3.3節の抽出ルールの記述例は、以下のような Perl プログラムの実行文に変換される。

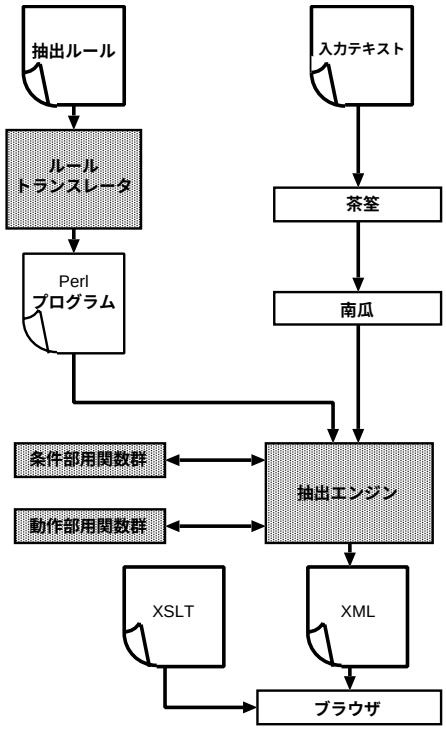

図 2: 時間表現抽出システムの処理の流れ

```
$name = "shiji01";
$pattern = [A, [B, [C]]];
$condition =
 "is_jyutsugo(\$variables{'A'}) &&
 is_jikan_meishi(\$variables{'B'}) &&
 is_shijishi(\$variables{'C'}) &&
 jyoshi_1(\$variables{'B'})
$action = "&tag_time(\$variables{'B'});
 &tag_timeref(\$variables{'C'});";
```

4 システムを利用した実験と結果

4.1 実験

前章で述べた時間表現抽出システムの評価実験を行った。実験に使用する文章は、日本経済新聞 1998 年の新聞記事である。

4.1.1 正解コーパスの作成

時間表現抽出システムを評価するため、以下のような手順で人手によって時間表現を抽出した正解コーパス 7 記事を作成した。

まず、実験に使用する新聞記事を茶筌、南瓜を利用して形態素解析、係り受け解析を行う。そしてその結果に基づいて、 XML 形式で出力する。

次に、出力された XML ファイルに対して、人手（大学生 2 名）で時間表現に関する情報を付加し

ていく。すなわち、 phrase タグの type の属性に対して、 phrase タグ内の節が時間を表している場合は ”time”、節が時間の参照先を表している場合は ”timeref”なる属性値を与える。

4.1.2 評価方法

正解コーパスおよびシステムが出力した XML テキストについて、 type 属性が ”time”もしくは ”timeref”となっている節（phrase タグ）を時間表現の節とする。システムが出力した XML の各節（phrase タグ）における type 属性がコーパス内の同一の節における type 属性と一致している場合、これを正解とする。

再現率 R、適合率 P を以下のように定義する。

$$R = \frac{\text{システムとコーパスで一致した節の数}}{\text{正解コーパスに含まれる時間表現の節の数}}$$

$$P = \frac{\text{システムとコーパスで一致した節の数}}{\text{システムが抽出した時間表現の節の数}}$$

4.2 結果

前述した日本経済新聞 1998 年の新聞記事から 7 記事について実験を行った。その結果は以下の通りである。

表 1: 実験の結果

再現率	0.384
適合率	0.625

4.2.1 検討

抽出に失敗した例としては以下のようなものがある。

- 今年は “ 夢持つ ” 年に

「今年は」と「年に」に対してコーパス作成者が時間の属性を付与してしまった。コーパス作成時の指示のあいまいさからくる認識の差が原因である。

また、コーパスでは時間の属性が付与されていないのに、システムが誤って抽出した例には以下のようなものがある。

- ビジネスウィーク一月十二日号は「九七年のトップ経営者二十五人」を特集

「一月十二日号」の節に対して時間の属性が付与されてしまった。

4.3 分類結果

また、日本経済新聞 1998 年の新聞記事から 200 記事についてシステムによる解析を行い、各分類ごとの出力結果をまとめた。

表 2: 分類

内部参照を用いた間接的時間表現	13
外部参照を用いた間接的時間表現	107
直接的時間表現	436

上記のように、直接的時間表現が多数を占めた。これは実験に利用した文章が新聞記事であるため、その特性上明確に時間を表現する場合が多いためであると考えられる。今後は新聞記事以外の日記や小説のような文章についても同様の実験を行うことで、それぞれの傾向を分析することができるとともに、より多様な文章に対応した時間表現が抽出できると思われる。

5 おわりに

本研究ではさまざまな種類の時間表現を分類し、時間表現抽出をルールとしてモデル化し、時間表現の自動抽出システムを試作した。

システムの評価としては、現在のところ十分な性能が得られていないが、今後、ルールを充実させていくことで性能の向上が期待される。

抽出ルールは、単に文内の時間表現抽出のためだけではなく、文間の時間関係の解析に用いることを前提に設計された。このような文章中の時間的な構造解析により、例えば事件記事などに対しては、事件の推移などが抽出される。文間の関係解析への展開が今後の課題である。

参考文献

[1] 渡辺一郎, 桝井文人, 福本淳一. 固有表現抽出ツール next の精緻化とユーザビリティの向上. 言語処理学会, Vol. 10, pp. 413–415, 3 2004.

[2] 土屋誠司, 渡辺広一, 河岡司. 連想メカニズムを用いた時間判断手法の有効性の検証. Vol. 168, pp. 113–118, 7 2005.

[3] 野呂太一, 乾孝司, 高村大也, 奥村学. イベントの生起時間帯判定. 情報処理学会 自然言語処理研究会, Vol. 170, pp. 7–14, 11 2005.