同一文内の表現対を対象とした日本語における時間関係認識

稲田 和明　　松林 優一郎　　乾 健太郎

東北大学

{kazuaki.inada, y-matsu, inui}@ecei.tohoku.ac.jp

1　はじめに

本稿では、日本語における時間表現やイベント表現間の時間関係を自動解析することを目的とし、日本語での時間関係認識の課題を調査する。

文章中に出現する時間表現やイベント表現の時間関係を正しく認識することは、イベントの時制判断などの時間情報に関する解析だけでなく、イベント発生の有無や順序関係の認識が必要となる含意関係認識や談話関係認識などのより応用的な処理に有用である。

英語では、時間関係が付与された TimeBank コーパス [6] が存在することもあり、時間関係認識に関する研究が盛んに行われてきた。時間関係認識のワークショップである TempEval では英語以外の言語でも時間関係認識が研究された [10] が、日本語については、学習や評価に用いるコーパスが存在しなかったため、これまであまり行われてこなかった。そのような背景の中、近年、日本語でも時間関係の情報が付与された BCCWJ-TimeBank が作成された [17]。

そこで本稿では、英語での時間関係認識のタスクに習い、BCCWJ-TimeBank を用いた日本語初の時間関係認識を試みる。まず、英語での研究を参考に、時間関係認識の手がかりとなる素性の抽出や整備、作成を行った。その後、作成した素性を日本語へ適用した時、どの程度の解析精度になるかを確かめ、その効果を検証した。

2　関連研究

英語での時間関係認識の研究は、時間関係の情報が付与された TimeBank コーパス [6] が作成されたことにより活発になった。TimeBank は、TimeML[5] の基準に従い、時間表現とイベント表現、及びそれらの表現間に 13 種類の時間関係ラベルが付与されている。

この TimeBank を用いた時間関係認識の取り組みとして、TempEval というワークショップが開催されている [9]。最新の TempEval-3 では、「同一文内の時間表現とイベント間、文書作成日時とイベント間、同一文内のイベント間、隣接文の主要イベント間」の 4 種類の表現対間の時間関係を認識するタスクが用意された。教師データとして使用できる、6 千語規模の質の高いコーパスや、自動解析を利用してタグ付けした 60 万語規模のコーパスが作成されたこともあり、13 種類の細かい時間関係を捉えなければならないにも関わらず、F1 値で約 56 %の精度を達成している。

英語では、主に機械学習を用いた手法が一般的だが、その一つとして、D'Souza らは、類義語や対義語、上位・下位語などの多数の語彙知識と、述語項構造や談話関係などの解析器の出力を利用したモデルを提案している [7]。統語構造などの素性と語彙知識を利用した素性を、ルールによって組み合わせることで、語彙知識の効果を引き出している。

一方、日本語における時間関係認識に関連する研究として、拡張固有表現抽出により得られた時間表現を用いて、イベントの発生時間を推定する取り組みが存在する [11]。この研究では、事件・事故・災害などの大規模なニュース内で発生したイベントの発生時刻の推定を行うが、ニュース記事の特徴を利用したルールベースの手法であるため、一般の文章に用いることは難しいと考えられる。

このように、日本語における時間関係認識に関連する研究が乏しい中、日本語に時間関係の情報を付与した BCCWJ-TimeBank コーパスが作成された [17]。BCCWJ-TimeBank の時間関係に関する詳細は次節で述べる。

3　使用データとタスク設定

3.1　BCCWJ-TimeBank

BCCWJ-TimeBank は、現代日本語書き言葉均衡コーパス (BCCWJ)[12] に、時間表現を示す「TIMEX3」とイベント表現を示す「EVENT」と、それらの間の時間関係「TLINK」が付与されたコーパスである。以下に BCCWJ-TimeBank の例を示す。

例「十月末に食品部門を分離し、・・・」

```
<TIMEX3 type="DATE" tid="t15" value="2001-10">
十月末</TIMEX3>に食品部門を
<EVENT class="OCCURRENCE" eid="e39">分離し
</EVENT>、・・・
 <TLINK eid="e39" tid="t14" task="T2E"
relA="before" relB="before" relC="before"/>
```

*一部のタグ表現を省略・改変している

TIMEX3 は、文章中の「日付表現、時刻表現、時間表現、頻度集合表現」に付与され、この 4 種類の内、どの表現であるかを示す type と、表現を文脈情報から正規化した value が記述されている。

EVENT は、時間情報を持つイベントに付与されており、項に事象を取るかどうか、また項を取る場合はその性質を分類した class の情報が与えられている。ただし、TimeBank に付与されている時制や相などの情報は付加されていない。

TLINK は、「隣接するイベント間 (E2E)、同一文内の時間表現とイベント間 (T2E)、文書作成時とイベント間 (DCT)、隣接する文の末尾のイベント表現間 (MATRIX)」の 4 種類の表現間に、TimeBank と同義の 1 3 種類のラベルに加え、部分事象の関係が全く同一の事象の関係にある場合に与える 3 種類のラベルと

表 1: 縮退後の時間的順序関係ラベル

縮退後	縮退前	意味
after	after, met-by	A が B より後に起こる
before	before, meet	A が B より前に起こる
overlaps	上述と vague 以外	A と B が時間に重なる
vague	vague	A と B の時間関係が不明

表 2: 時間関係が付与された表現対数 (文内限定)

	E2E	T2E
after	177	323
before	582	303
overlaps	378	974
vague	27	100
合計	1164	1700

時間関係無しを示す vague を加えた、計 17 種類の時間関係が付与されている。ただし、付与対象は文書作成日時が存在する新聞サブコーパスのみに限定している。なお、時間関係の付与は 3 人の作業者によって行われたが、1 つの時間関係に対し、作業者それぞれが判断した 3 人分の時間関係ラベルが記述されている。

3.2 タスク設定

離れた場所に存在する表現対の時間関係認識では、同一文内の表現対に比べ、談話構造の認識や語彙的な知識などが必要となる。このように、文外の離れた表現対は、文内と取り扱うべき問題が異なるため、同一のモデルで解析すると、分析が複雑となる可能性がある。このような背景から、本研究では BCCWJ-TimeBank を用いた最初の分析として、T2E と E2E の中で表現対が同一文内に存在するものに絞り、時間関係認識を行う。

また、現時点で作成された BCCWJ-TimeBank の規模は E2E が 2972 事例、T2E が 2188 事例と少量であり、さらに一部の時間表現ラベルは、コーパス中にわずかしか出現しない。そこで、疎データ問題を防ぐために 3 人の作業者が付与した 17 種類の時間関係のラベルを表 1 に示す 4 種類に縮退させる。実験では、作業者 3 人のラベルが一致したものを用いた。表 2 は、上述の処理を行った結果得られた E2E と T2E の事例数である。

時間関係認識の際は、入力として、時間関係を認識する対象となる表現対、それらに付加されている情報、表現対を含む文を与える。コーパスの仕様に準じ、入力として与えられる表現対 A、B は、E2E では A、B の順で出現する隣接イベントとし、T2E では A が時間表現、B をイベント表現とする。また、出力は、与えた表現対間の時間関係ラベルとする。

4 提案手法

入力された表現対の時間関係を識別するため、出力となる 4 種類のラベルの多値分類問題とみなし、教師ありの機械学習、L2 正則化ロジスティック回帰（最大エントロピー法）を用いる。表 3 は本稿で使用した素性の一覧である。

4.1 基本的な素性

表 3 の 1〜13 の素性は、英語における時間関係認識の代表的なシステム [2, 7] が使用している基本的な素性の中で、日本語でも効果が期待できそうな素性を、既存の解析器を用いて作成したものである。

Algorithm 1 synsetID のグループ化

W : $WordNet$ 内の語の集合
G_w : 語 w に割り振られた $synsetID$ の集合
$G \Leftarrow \{G_w | w \in W\}$

$return\ \ merge(G)$

$def\ merge(G)$
  $G' \Leftarrow \emptyset$
  $foreach\ g\ in\ G$
    $foreach\ g'\ in\ G'$
      $if((|g \cap g'|/|g \cup g'|) \geq 0.1)$
        $G' \Leftarrow G' \cup \{g \cup g'\}$
        $break$
    $if(G'\ is\ not\ updated)$
      $G' \Leftarrow G' \cup \{g\}$
  $if(G == G')$
    $return\ \ G$
  $else$
    $merge(G')$

1 は対象表現対に関する形態素レベルの基本的な素性であり、2〜3 は BCCWJ-TimeBank の付加情報を用いた素性である。また、日本語でも、表現対の構文的な支配関係が時間関係に関連すると考えられるため、4〜9 の文構造に関する素性を採用した。これらの素性を作成する為に、CaboCha[8] の解析結果を使用した。

10〜11 は同一文節内の機能表現や拡張モダリティ情報を用いた素性である [15]。日本語では、イベント表現と同じ文節内の機能表現が、時間関係に影響を与えるため、機能表現を素性として取り入れた。TempEval では、TimeBank に付与されている時制や相などを素性として使用している。そこで、これらに近い情報として、拡張モダリティの「時制、仮想、真偽判断」を素性として用いた。BCCWJ には、拡張モダリティ情報が人手で付与されているが、BCCWJ-TimeBank のイベント表現と一致する部分がわずかだったため、BCCWJ に付与された拡張モダリティ情報は使用せず、水野らの自動解析器 [16] の出力を素性として用いた。

12〜14 は係り受け木を辿ったとき、最も近いイベントに関する情報を捉える素性であるが、対象表現対が係り先のイベントからも、時間関係の手がかりを得ることを期待している。

4.2 表現を一般化した素性

表 2 に示した通り、時間関係認識に使用できるデータ数が乏しい。そこで、疎データ問題に対応するため、同義語や類義語を 1 つの表現にまとめ、一般化した表現を用いた素性 (a) を作成した。

まず、動詞・形容詞・サ変名詞に属する語の一般化に、日本語 WordNet[1] の synsetID を利用した。直接 synsetID を一般化した表現として用いると、一つの語が複数の synsetID を持つ為、一般化の際に曖昧性が生じる。また、約 24000 語に 10000 種類以上の synsetID を割り振っているため、疎データ問題の対策にはなりにくい。

これらの対応策として、動詞・形容詞に属する synsetID と、複数の synsetID が付与された語に着目したアルゴリズム 1 を用いて、synsetID を 464 のグループにまとめ上げた。このアルゴリズムでは、1 つの語に付与された synsetID を集合として捉え、類似した synsetID の集合同士を合わせていくことで、グループ化をしている。なお、 一般化の際には、単語が持つ synsetID 集合が、グループ化された synsetID 集合のうち、部分集合の関係にある集合に割り振った ID を用いた。

表 3: 素性一覧

id	素性	E2E	T2E
1	対象表現対と前後２形態素の基本形と品詞	✓	✓
2	対象表現対の BCCWJ-TimeBank に付与されているイベントの class	○	○
3	対象表現対の BCCWJ-TimeBank に付与されている時間表現の type、value		○
4	文上で、A が B より先に出現するか		○
5	文上で、対象表現対の間に、DATE か TIME を持つ固有表現中が存在するか	○	○
6	A と B が同じ文節内に存在するか	○	○
7	A から B へ係り受け木を辿った時の距離	○	○
8	係り受け上で、A が B の親か、またはその逆	○	○
9	係り受け上で、対象表現対の間の固有表現中に、DATE か TIME を持つものが存在するか		○
10	対象表現対が含まれている文節に存在する機能表現の表層	○	○
11	対象表現対の内、イベントの拡張モダリティタグの時制、仮想、真偽判断	○	○
12	対象表現対が係る最も近い動詞・形容詞までの係り受け距離	○	○
13	対象表現対が係る最も近い動詞・形容詞自身とその周辺２形態素の基本型、品詞	○	○
14	対象表現対が係る最も近い動詞・形容詞を含む文節内に存在する機能表現の表層	○	○
a	対象表現対と前後２形態素の基本形を一般化した素性	○	
b	対象表現対を一般化した時、それらが同じ値か	○	
c	大規模データから取得したイベント対の頻度を利用した素性	○	

表 4: 時間関係判断の手がかり表現

before	後, 結果, 後々, すぐ, 将来, 未来, この後, 直後, 今後, 以降
after	前, 前もって, 先, 先立つ, 予め, 直前, この前, 以前

サ変名詞以外の名詞の一般化には、ALAGIN 文脈類義語データベース [3] に存在する、約 100 万語の名詞が 500 個のクラスに分類されたデータを使用した。このクラスに割り振られた ID を、名詞の一般化した表現として使用した。

機能表現の同定には、日本語機能表現辞書 [14] の「L3:文法的機能」レベルの ID を用いて、9041 の機能表現を 418 種類にまとめ上げた。一つの機能表現が複数の ID を持つ場合には、その機能表現が持つ ID の中で、辞書中に最も出現する ID を割り当てた。

また、入力された表現対が時間表現の場合は、BCCWJ-TimeBank に付与された type の値を使用した。固有表現抽出の結果、固有表現を持つ形態素は、固有表現タグを一般化した表現として用いた。記号は「句点、読点、括弧開、括弧閉、アルファベット、数字、それ以外」の７つの表現にまとめた。なお、表現の一般化は、時間表現、動詞・形容詞・サ変名詞、名詞、固有表現、機能語、記号の順で行い、一般化できなかったものはその語の基本形を一般形として用いた。

4.3 類義語関係を利用した素性

D'Souza ら [7] は、入力された表現対が同義語や類義語の関係ならば、同じ時間帯に発生している可能性が高いという予測に基づき、WordNet などの語彙知識を用いて、同義語・類義語判断を行う素性を採用している。本稿でも、4.2 節の表現の一般化を用いて、入力された表現対を汎化したときに、それらの値が同じ表現になるかを判断する素性 (b) を導入した。

4.4 大規模データの頻度情報を用いた素性

因果関係のように、特定のイベント間には、時間的順序関係が存在するのではないかと仮定し、特定の文字列表現を手がかり表現として用いた因果関係抽出の手法 [13] をイベントの時間順序抽出に応用し、事象の成立順序に一定の傾向があると思われるイベント対を抽出した。

まず、表 4 に示す before と after の時間関係判断の手がかりとなる表現を作成した。そして、Web 上の約１０億文に対して、表 4 の語を含む文節を基点とし、その文節へ係っている文節に含まれるイベント表現 A と、その文節が係っている文節に含まれるイベント表現 B を抽出した。この時、CaboCha による形態素解析の結果、動詞・形容詞・サ変名詞を持つ形態素をイベント表現として扱った。このようにして得られた AB 間には、手がかり表現が属する時間関係が存在すると見なし、時間関係と (A,B) の頻度を求めた。最終的に、抽出した (A,B) の頻度が、before もしくは after の片方に偏っている場合、A と B はその時間関係になりやすいと判断し、以下の式を用いて求めた確信度 R を求め、対数を用いて R を圧縮したスコア S を素性 (c) として使用した。なお、before(A,B) は、before の手がかり表現で抽出したイベント対 (A,B) の頻度を表す。

$$R = before(A,B) + after(B,A) - before(B,A) - after(A,B)$$

$$S = \begin{cases} 5 & (R > 0 \wedge \log_5 R \geq 5) \\ \lfloor \log_5 R \rfloor & (R > 0 \wedge 0 < \log_5 R < 5) \\ -\lfloor \log_5(-R) \rfloor & (R < 0 \wedge 0 < log_5(-R) < 5) \\ -5 & (R < 0 \wedge \log_5(-R) \geq 5) \\ 0 & (else) \end{cases}$$

5 実験・結果

英語に対して行われた機械学習の手法と類似の戦略で、どの程度の解析精度が実現できるか、また提案した素性が有効かを調べるために実験を行った。実験には表 2 のデータを使用し、E2E と T2E のそれぞれに対して５分割交差検定を行い、各ラベルの F1 値とその Micro 平均を評価尺度として、性能を比較した。なお，モデルの学習及び評価には、岡崎の Classias[4] を用いた。

表 5 は表 3 の素性で最も基本的な素性である 1 のみを用いたモデルに、2〜3、4〜9、10〜11、12〜13 の素性を順に加えた時の結果である。さらに E2E では、基本的な素性の中で最も性能が良かった 1〜11 の素性を用いたモデルに、提案した素性 abc を加えたモデルの評価も行った。

E2E と T2E の両方で、BCCWJ-TimeBank の付加情報 2〜3 と、文構造に関する素性 4〜9 を加えることで精度が上昇した。特に T2E では、文構造の素性により、入力されたイベント表現 B が時間表現 A に支配されているかどうかを判断することができたため、精度が大きく上昇したと考えられる。

機能表現と拡張モダリティの素性 10〜11 は、E2E と T2E どちらにも有効だった。対象表現対のイベント表現が持つ機能表現や拡張モダリティ情報は、時間関係認識において有用であることがわかった。本稿では、自動解析により拡張モダリティの情報を得ていたため、より正確な拡張モダリティ情報を用いることで、さらなる精度上昇が期待できる。

表 5: 提案した素性の効果に関する評価結果

	E2E						T2E				
	1	1-3	1-9	1-11	1-14	1-11 +abc	1	1-3	1-9	1-11	1-14
after	42.2	45.3	48.6	49.9	47.5	49.8	55.3	57.2	64.5	65.7	71.9
before	75.2	75.1	75.0	77.1	76.3	78.1	41.8	47.0	49.7	51.2	60.9
overlaps	52.6	53.3	52.3	56.3	55.4	57.3	75.9	76.4	79.3	79.6	82.9
vague	0	0	0	5.7	0	5.71	34.8	50.2	44.8	46.0	61.2
Micro	63.4	63.8	63.9	66.4	65.1	67.2	66.2	67.6	70.8	71.4	76.5

表 6: E2E での追加実験における適合率と再現率

	適合率				再現率			
	1-11	1-11 +a	1-11 +b	1-11 +c	1-11	1-11 +a	1-11 +b	1-11 +c
after	58.1	59.3	59.8	60.2	44.8	42.0	40.9	39.8
before	70.9	71.4	70.1	70.1	84.6	84.9	87.1	86.9
overlaps	60.2	58.9	61.3	60.7	53.1	54.1	52.0	52.1
vague	20.0	20.0	0	0	3.3	3.3	0	0

係り受け先のイベント情報を見る素性 12〜14 では、T2E では精度が上昇したが、E2E では低下した。イベント表現の場合は、係り先のイベントの情報を見てもあまり有用ではないが、時間表現の場合は、係り先イベントの周辺の語や機能表現が時間関係の判断に活用できていると推測される。

また、E2E では vague の精度が非常に低いが、これは事例数が 27 件と少なすぎて、学習できなかったためと考えられる。

E2E では、基本的な素性の中で最も性能が良かった 1〜11 の素性を用いたモデルに、表 3 の素性 abc を加えると、さらに精度が上昇した。そこで、新たに加えた素性が、精度にどのような影響を与えているのかを調べた。表 6 は、3 の素性 1〜11 を使用したモデルに、素性 abc を別々に加えて実験したときの適合率と再現率である。

表現の一般化の素性 a は、before の再現率を低下させずに適合率が上昇している点や、overlaps の再現率が上昇している点から、一般化の素性が有効に働いていると推測できる。しかし、after の再現率が大幅に低下しているため、表現を一般化しすぎている面もあると予想できる。

対象表現の一般形が同一の表現になるかを判断する素性 b は、仮定通り overlaps の適合率が上昇したが、after の適合率と before の再現率以外は低下している。これは、素性 b を二値の素性として導入した為、必要以上に発火してしまった可能性が原因の一つとして考えられる。

大規模データから抽出したイベント対の頻度を利用した素性 c は、after の適合率が上昇していることから、ある程度機能していると判断できる。しかし、頻度を求めている before の適合率が上昇していないため、大規模データからの抽出方法や before の手がかり表現を検討する必要があると言える。

実験により、英語で使用されている素性が、日本語でも概ね有効であることを確認した。また、本稿で提案した素性も、組み合わせによっては時間関係認識に有効であることがわかった。なお、BCCWJ-TimeBank の時間関係ラベルを 4 種類に縮退した時の作業者間一致率は、文内のみの E2E が 66.0%、T2E が 75.1%であるが、本稿のシステムはこれに近い性能を達成している。

6 おわりに

本稿では、BCCWJ-TimeBank を用いて時間関係を行い、日本語において時間関係認識が可能であることを示した。英語での時間関係認識に用いられている基本的な素性を、日本語で使用できる素性として整備し、それらがどの程度の効果があるのかを確認した。また、表現の一般化を用いた素性や大規模データの頻度情報を利用した素性を提案し、それらの効果を検証した。

今後の課題としては、時間関係のラベル数を増加させた時、本稿で用いた素性が有効かを検証することや、時間関係認識に有用な素性の作成やその活用方法の吟味、複数の素性を組み合わせたときの精度変化の調査などが挙げられる。

謝辞 本研究の一部は JST, CREST の助成を受けたものです。

参考文献

[1] Francis Bond, Hitoshi Isahara, Sanae Fujita, Kiyotaka Uchimoto, Takayuki Kuribayashi, and Kyoko Kanzaki. Enhancing the Japanese WordNet. In *Proceeding of ACL-IJCNLP*, pp. 1–8, 2009.

[2] Nathanael Chambers, Shan Wang, and Dan Jurafsky. Classifying temporal relations between events. In *Proceedings of ACL*, pp. 173–176, 2007.

[3] Jun'ichi Kazama and Kentaro Torisawa. Inducing gazetteers for named entity recognition by large-scale clustering of dependency relations. In *Proceedings of ACL-08:HLT*, pp. 407–415, 2008.

[4] Naoaki Okazaki. Classias: a collection of machine-learning algorithms for classification, 2009.

[5] James Pustejovsky, Jose Casta, Robert Ingria, Roser Sauri, Robert Gaizauskas, Andrea Setzer, and Graham Katz. TimeML: Robust specification of event and temporal expressions in text. In *Proceedings of IWCS-5*, 2003.

[6] James Pustejovsky, David Day, Lisa Ferro, Robert Gaizauskas, Patrick Hanks, Marcia Lazo, Roser Sauri, Andrew See, Andrea Setzer, and Beth Sundheim. The TIMEBANK corpus. In *Proceedings of Corpus Linguistics*, pp. 647–656, 2003.

[7] Lucian Silcox and Emmett Tomai. Identification of temporal event relationships in biographical accounts. In *Proceedings of NAACL-HLT*, pp. 529–533, 2013.

[8] Yuji Matsumoto Taku Kudo. Japanese dependency analysis using cascaded chunking. In *Proceedings of CoNLL 2002*, pp. 63–69, 2002.

[9] Naushad UzZaman, Hector Llorens, Leon Derczynski, James Allen, Marc Verhagen, and James Pustejovsky. Semeval-2013 task 1: Tempeval-3: Evaluating time expressions, events, and temporal relations. In *Proceedings of SemEval 2013*, pp. 1–9, 2013.

[10] Marc Verhagen, Roser Sauri, Tommaso Caselli, and James Pustejovsky. Semeval-2010 task 13: Tempeval-2. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pp. 57–62, 2010.

[11] 井上晃太, 佐藤真, 赤石美奈. 複数文書から抽出したイベントの時間関係処理に関する研究. 人工知能学会論文誌, 2013.

[12] 橋本泰一, 中村俊一. 拡張固有表現タグ付きコーパスの構築 - 白書, 書籍, yahoo!知恵袋コアデータ -. 言語処理学会第 16 回年次大会発表論文集, pp. 916–919, 2010.

[13] 坂地泰紀, 増山繁, 酒井浩之. 新聞記事中の文が因果関係を含むか否かの判定. 電子情報通信学会技術研究報告, 2010.

[14] 松吉俊, 佐藤理史, 宇津呂武仁. 日本語機能表現辞書の編纂. 自然言語処理, 第 14 巻, pp. 123–146, 2007.

[15] 松吉俊, 佐尾ちとせ, 乾健太郎, 松本裕治. 拡張モダリティタグ付与コーパスの設計と構築. 言語処理学会第 17 回年次大会発表論文集, pp. 147–150, 2011.

[16] 水野淳太, 成田和弥, 乾健太郎, 大竹清敬, 鳥澤健太郎. 拡張モダリティ解析器の試作と課題分析. ALAGIN NLP 若手の会 合同シンポジウム, 2013.

[17] 保田祥, 小西光, 浅原正幸, 今田水穂, 前川喜久雄. 『現代日本語書き言葉均衡コーパス』に対する時間表現・事象表現間の時間的順序関係アノテーション. 第 3 回コーパス日本語学ワークショップ, 2013.