

事象間の接続関係に基づく時間的順序関係推定

高木宏伸[†]嶋田和孝[‡][†]九州工業大学大学院 情報工学府[‡]九州工業大学大学院 情報工学研究院

{h_takagi, shimada}@pluto.ai.kyutech.ac.jp

1 はじめに

近年、文章中に出現する時間表現や事象表現間の時間的順序関係を正しく認識する時間情報解析が盛んに行われている。事象の発生した時間やそれらの時間的順序関係を同定することは、因果関係知識の獲得や含意関係認識への応用など、深い言語理解の実現に必要である。

英語を対象とした時間的順序関係推定は、時間情報が付与された TimeBank[3] が公開されたことによって、機械学習による推定手法が提案されてきた。さらに、時間的順序関係推定のワークショップである TempEval[4] を中心に、英語、スペイン語、中国語といった主要な言語については、さまざまな研究が行われている [1, 2]。しかしながら、日本語を対象とした研究はまだ少ない。

そのような背景の中、近年日本語での時間情報が付与された BCCWJ-TimeBank[6, 9] が構築され、機械学習による統計的解析や、定量的な性能評価が可能になった。これまで、機械学習を用いた統計的手法が提案されてきたが、事象間の表層的特徴についての詳しい分析は行われていない。事象間の時間的順序関係と表層構造の相関が明らかになれば、表層上の特徴に着目するだけでも、高精度の時間的順序関係推定が実現できると考えられる。

そこで本稿では、BCCWJ-TimeBank の文章を事象間の接続関係に基づき分類し、それぞれについて時間的順序関係推定を試み、その性質を考察する。

2 関連研究

BCCWJ-TimeBank を用いた時間的順序関係推定手法として、吉川ら [10] は、時間表現と事象表現の品詞や活用形といった語彙的な素性、構文木上での距離といった統語的な素性を使用した時間的順序関係推定モデルを提案している。英語での時間的順序関係推定に一般的に使われる素性を日本語で使用できるように整備し、これらの素性が日本語での時間的順序関係推定に

においても有用であることを示した。また、稲田ら [5] は、同一文内の事象間に焦点を当てたモデルを提案し、類義語関係や大規模データの頻度情報を用いた素性なども使用することで、高精度の結果を報告している。

一方で、言語学の分野では、複文における事象間の時間的順序関係について、接続関係の種類やそれに含まれる語に基づく考察が多くされている。その一つとして、丹波 [7] は、複文における連体修飾節についての時間関係について考察している。

BCCWJ-TimeBank は、2013 年に公開されたばかりで、現在これを用いた時間的順序関係推定は、英語を対象とした手法の日本語における有用性が検証されている段階である。そういった中で本研究では、日本語独特の表層構造の特徴に着目し、複文の節間の接続表現が時間的順序関係推定に有用であるかを BCCWJ-TimeBank を用いて検証していく。

3 提案手法

同一文内の隣接する事象対に対して、事象間の表現から接続関係を抽出する。抽出された接続関係ごとに時間的順序関係を出力する分類モデルを作成し、評価を行う。加えて、結果について分析を行う。

3.1 使用データ

時間的順序関係分類モデルを作成するにあたり、事象間の時間的順序関係が付与された BCCWJ-TimeBank を使用する。BCCWJ-TimeBank は、「現代日本語書き言葉均衡コーパス¹」のコアデータ 54 文書に対して、時間表現を示す「TIMEX3」、事象表現を示す「EVENT」、それらの間の時間的順序関係を示す「TLINK」が付与されている。「TLINK」には、「before, after, overlaps」など 17 種類のラベルが 3 人の作業員によって付与されている。

本研究では、事象間の時間的順序関係が示された「TLINK」のみを学習データとして使用する。また、

¹http://www.ninjal.ac.jp/corpus_center/bccwj/

表 1: 時間的順序関係ラベルの簡略化

元ラベル	変換後
after	AFTER
met-by	AFTER
before	BEFORE
meets	BEFORE
contains	OVERLAP
during	OVERLAP
equal	OVERLAP
finished-by	OVERLAP
finishes	OVERLAP
identity	OVERLAP
includes	OVERLAP
is_included	OVERLAP
overlaped-by	OVERLAP
overlaps	OVERLAP
started-by	OVERLAP
starts	OVERLAP
vague	VAGUE

表 2: 時間的順序関係が付与された事象対数

ラベル	事象対数
AFTER	212
BEFORE	601
OVERLAP	387
VAGUE	41
合計	1241

時間的順序関係を示すラベルの一部はわずかしか出現せず、疎データ問題を引き起こす。そのため、17種類のラベルを表1に示す4種類に簡略化する。実験では、作業員3人の付与したラベルが一致したもののみを用いる。現時点で同一文内の事象対への時間的順序関係が付与されているのは1241事例である。表2に事例の内訳を示す。

3.2 接続関係抽出

一般に複文における接続節は、「補足節、副詞節、連体節、並列節」の4種類であると知られている[8]。本研究では、事象の活用形と手がかり表現、またそれに繋がる助詞の組み合わせを手で整理し、パターンマッチを用いた抽出モデルを作成する。表3に接続節ごとの代表的な手がかり表現を示す。この抽出モデルは、事象対とその間の表現から第一事象が含まれる節の種類を同定し、それに繋がる節との接続関係を抽出

表 3: 接続節ごとの代表的な手がかり表現

補足節	副詞節
こと、の、ところ のは、と、よう	から、結果、あいだ のに、述語テ形+も 述語連用形+ながら
連体節	並列節
述語+名詞、という との	が、たり、述語テ形 述語連用形

表 4: モデルごとの使用した素性

素性	補足	副詞	連体	並列
第一事象の語尾	○	○	○	○
第二事象の語尾	○	○	○	○
手がかり表現	○	○		○
手がかり表現に続く助詞	○		○	

する。以降、これらを「補足接続、副詞接続、連体接続、並列接続」と呼ぶ。全てのパターンにマッチしなかった事象対は、抽出漏れとして、3.3節で説明する時間的順序関係推定の対象とはしない。

3.3 時間的順序関係推定

3.2節で述べた「補足接続、副詞接続、連体接続、並列接続」の4種類の接続関係について、それぞれ時間的順序関係推定モデルを構築する。本稿では、4種類の接続関係と事象間の時間的順序関係との相関を分析するため、極めて単純な素性のみを用いたC4.5決定木による時間的順序関係推定モデルを構築する。表4に使用する素性一覧を示す。

4 実験と結果

ここでは、3節で説明した、事象間の接続関係抽出とその時間的順序関係推定についての実験結果を述べる。

4.1 接続関係抽出

表2に示す事象対1241事例について、接続関係抽出モデルを用いて「補足接続、副詞接続、連体接続、並列接続」の抽出を行った結果を表5に示す。4種類の接続関係について合計1148事例が抽出でき、抽出漏れは93事例であった。抽出漏れの主な原因は、抽出モデルの知識不足に加えて、「運転手殴り料金踏み倒す」といった新聞の見出しのように、助詞や形容詞が省略され、接続関係が抽出できない文が存在したことが挙げられる。なお、結果には誤抽出も含まれるが、

表 5: 接続関係抽出結果

ラベル	補足	副詞	連体	並列	合計
AFTER	46	20	123	8	197
BEFORE	64	108	206	173	551
OVERLAP	53	97	151	62	363
VAGUE	0	0	26	11	37
合計	163	225	506	254	1148

本研究の目的は接続関係ごとの性質を分析することなので、接続表現抽出の精度について今回は言及しない。

4.2 時間的順序関係推定

4.1 節で抽出された 4 種類の接続関係をもつ事象対を対象に、それぞれの時間的順序関係推定モデルを 10 分割交差検定で評価した。

4.2.1 補足接続

表 5 のうち、補足節接続であると抽出された事象対 163 事例について分類評価を行った結果を表 6 に示す。加重平均で適合率 0.723、再現率 0.699、F 値 0.693 という結果になった。現在、BCCWJ-TimeBank を対象にした同一文における時間的順序関係推定の最高精度である稲田ら [5] の精度が F 値 0.672 であることから、補足接続のみであればこの結果を上回っている。また、事象対 163 事例中、AFTER が 46 事例、BEFORE が 64 事例、OVERLAP が 53 事例とほぼ偏りなく存在している。このことから、補足接続である事象間の時間的順序関係推定では、単純な素性のみでも高精度の分類を行うことができたといえる。

次に、各ラベルについて考察する。AFTER、BEFORE、OVERLAP のうち BEFORE が F 値 0.873 と一番良い結果であり、AFTER や OVERLAP とは特徴の違いが顕著であった。構築された決定木では、根節点に第一事象の語尾がきており、語尾が過去を表す“た”のときに BEFORE に分類され、それ以外の場合で手がかり語から AFTER と OVERLAP に分類される。AFTER と OVERLAP の F 値は、それぞれ 0.643 と 0.518 で、手がかり表現という素性のみでは、両者間の分類が正しく行えていなかった。

4.2.2 副詞接続

副詞接続であると抽出された事象対は、BEFORE と OVERLAP が共に 100 事例程度で、AFTER のみ 20 事例という少ない事例数であった。表 2 に示す抽出前のラベル比率を考慮すると、OVERLAP の関係をとりにやすく、AFTER の関係をとりにづらいことが推測できる。

表 6: 時間的順序関係推定 (補足接続)

ラベル	適合率	再現率	F 値
AFTER	0.536	0.804	0.643
BEFORE	0.887	0.859	0.873
OVERLAP	0.688	0.415	0.518
VAGUE	—	—	—
加重平均	0.723	0.699	0.693

表 7: 時間的順序関係推定 (副詞接続)

ラベル	適合率	再現率	F 値
AFTER	0.909	0.500	0.645
BEFORE	0.648	0.769	0.703
OVERLAP	0.698	0.619	0.656
VAGUE	—	—	—
加重平均	0.693	0.680	0.678

表 7 は、この事象対 225 事例について分類評価を行った結果である。加重平均で適合率 0.693、再現率 0.680、F 値 0.678 という結果になった。F 値は 3 つのラベルで 0.645 から 0.703 の値となり、ある程度分類は成功している。構築された決定木は、根節点の手がかり表現であった。このことから、副詞接続においては、手がかり表現と時間的順序関係には相関があると考えられる。しかしながら、今回の実験では、1 事例や 2 事例でしか出現しなかった手がかり表現が多く存在し、学習がうまくできなかった部分もあった。このことから、手がかり語のグループ化や他の素性を用いることでモデルの改善を行う必要がある。

4.2.3 連体接続

連体接続であると抽出された事象対は、合計 506 事例と、他の接続関係に比べ最多であった。そのうち大多数が「述語+名詞」のパターンで抽出されたため、手がかり表現が素性として使用できなかった。結果、表 8 に示すように、AFTER、VAGUE は F 値が 0 で、加重平均の F 値が 0.479 と分類に失敗している。今後、名詞に続く表現に対して、時間的順序関係を表す表現の有無を分析する必要がある。

4.2.4 並列接続

並列接続であると抽出された事象対は合計 254 事例あるが、その内 AFTER の関係として抽出されたのはわずか 8 事例のみであった。並列節には、順接的並列と逆接的並列の 2 つがあるが、順接的並列の出現頻度が圧倒的に高かった。さらに、順接的並列では、第一事象が第二事象より時間的に後に起こることはなく、

表 8: 時間的順序関係推定 (連体接続)

ラベル	適合率	再現率	F 値
AFTER	0	0	0
BEFORE	0.687	0.864	0.766
OVERLAP	0.451	0.735	0.559
VAGUE	0	0	0
加重平均	0.414	0.571	0.479

表 9: 時間的順序関係推定 (並列接続)

ラベル	適合率	再現率	F 値
AFTER	0	0	0
BEFORE	0.712	1.000	0.832
OVERLAP	0	0	0
VAGUE	0.286	0.182	0.222
加重平均	0.497	0.689	0.576

逆接的並列の一部でのみ第一事象が時間的に後に起こる場合がある。このことから、並列接続である事象対が AFTER の関係をもつ場合は稀であるため、抽出数が 8 事例と少なかったことも妥当な結果だと考えられる。

時間的順序関係の分類結果は表 9 に示す通りで、AFTER と OVERLAP の F 値が 0、F 値の加重平均は 0.576 と分類に失敗している。これは、254 事例中 199 事例の手がかり表現が「述語連用形」であったため、この素性が分類精度に貢献しなかったことが原因である。分類精度を上げるために、順接的並列では AFTER の関係をとらないという特徴を活かせる素性を作成をする必要がある。

4.2.5 共通の課題

全ての接続関係における時間的順序関係推定に共通した考察として、第二事象の語尾という素性が全く分類精度に貢献しなかったことが挙げられる。今回、接続関係を抽出する時点では、第二事象が含まれる節の種類については考慮していなかった。第二事象の語尾は、第二事象が含まれる節の種類によって変化するが、その変化を素性として反映できなかったことが、分類精度に貢献しなかった原因であると考えられる。また、全てにおいて VAGUE の精度が非常に低いが、これは事例数が合計で 41 事例と少ないことから学習ができなかったためと考えられる。

5 おわりに

本稿では、時間的順序関係推定において、事象間の表現から接続関係ごとにそれぞれのモデルを作成することで、接続関係ごとの特徴を分析した。

今後は、エラー分析をより詳細に行い、新たな素性を作成するとともに、共通の課題で挙げられた第二事象が含まれる節の種類についても考慮したい。また、今回用いた BCCWJ-TimeBank は事例数が少なかったため、別の文書に対しての評価実験や精度調査なども今後の課題として挙げられる。

参考文献

- [1] Steven Bethard. Cleartk-timeml: A minimalist approach to tempeval 2013. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pp. 10–14, 2013.
- [2] Kenton Lee, Yoav Artzi, Jesse Dodge and Luke Zettlemoyer. Context-dependent Semantic Parsing for Time Expressions. In *Proceedings of the Conference of the Association for Computational Linguistics*, pp. 1437–1447, 2014.
- [3] James Pustejovsky, Jose Castano, Robert Ingria, Reser Sauri, Robert Gaizauskas, Andrea Setzer and Graham Katz. The timebank corpus. In *Proceedings of Corpus Linguistics 2003*, pp. 647–656, 2003.
- [4] 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 *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pp. 1–9, 2013.
- [5] 稲田和明, 松林優一郎, 乾健太郎. 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. 言語処理学会第 20 回年次大会発表論文集, pp. 348–351, 2014.
- [6] 小西光, 浅原正幸, 前川喜久雄. 『現代日本語書き言葉均衡コーパス』に対する時間情報アノテーション. 自然言語処理, Vol. 20, No. 2, pp. 201–222, 2013.
- [7] 丹波哲也. 連体修飾節のテンスとアスペクト. 『言語』30, pp. 56–62, 大修館書店, 2001.
- [8] 益岡隆志, 田窪行則. 基礎日本語文法. くろしお出版, 1992.
- [9] 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. 自然言語処理, Vol. 20, No. 5, pp. 657–682, 2013.
- [10] 吉川克正, 浅原正幸, 飯田龍. BCCWJ-TimeBank を対象とした時間的順序関係の推定. 言語処理学会第 20 回年次大会発表論文集, pp. 1103–1106, 2014.