5Q-04

日本語文内で表現されるイベント間の時間的な順序関係の認識*

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

文章中のイベント間の時間的な順序関係の認識 (**時間関係認識**) は，自然言語処理分野における挑戦的なタスクの 1 つとして，近年盛んに研究が行われている．英語では，時間関係認識のための学習・評価用コーパス TimeBank[1]が早期に開発されたこともあり，評価型ワークショップ **TempEval** [1] が開催されるなど，多数の研究成果が報告されている．一方，日本語では TimeBank に相当するようなコーパスが長らく存在しなかったこともあり，時間関係認識に関する研究は乏しかった．このような中，近年日本語の文章に時間関係をはじめとする時間情報を付与した **BCCWJ-TimeBank** [2] が開発されたことを受けて，吉川らが日本語文を対象に時間関係認識を実施した [3] が，その性能・分析は十分ではなかった．そこで我々は，時間表現の係り先のイベント表現の文脈情報と，大規模データから獲得したイベント表現間の頻度情報の 2 つの情報を利用した独自の素性を導入し，さらに分類手法に多項式カーネルを用いたサポートベクターマシンを採用することで，既存研究を上回る性能を有する時間関係の分類器を実現した．また訓練データ量と精度の関係から，学習に用いる事例数を増加させることで，更なる性能向上が期待できることが分かった．

2 BCCWJ-TimeBank

BCCWJ-TimeBank [2] は，日本語での時間関係認識のために開発された学習・評価用コーパスである．図 1 にその概観を示す．BCCWJ-TimeBank では，文章上の時間表現やイベント表現の位置を同定した上で，「E2E: 隣接イベント表現間」,「MATRIX（以降，MAT と略記する）: 隣接文の末尾のイベント表現間」,「T2E: 同一文内の時間表現とイベント表現間」,「DCT: 文書作成日時とイベント表現」の 4 種類の表現間に時間関係を示すラベル (時間関係ラベル) が付与されている．そして，それらの表現間の時間関係として「after(X → Y)（X の前に Y が発生する）」などの計 17 種類のラベル[2]が採用されている．また BCCWJ-TimeBank 上で同定された時間表現やイベント表現には，時間表現を正規化した値 (value)，時間表現が示す期間などによる分類 (type)，イベント表現の語彙的意味を考慮した時間的性質に関わる分類 (class) の 3 つが時間情報として付与されている．既存研究では，これらの時間情報を時間関係ラベルの分類に取り入れているため，本研究でも同様にこれらの値を時間関係認識に利用する．

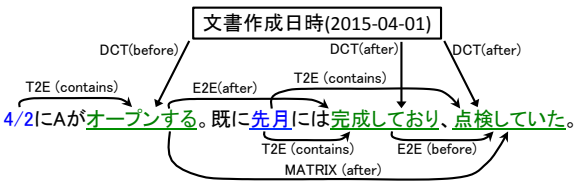

図 1: BCCWJ-TimeBank の概観

表 1: 本研究で使用した素性

ID	素性	E2E	MAT	T2E	DCT
1	対象表現とその前後 2 形態素の基本形	✓	✓	✓	✓
2	BCCWJ-TimeBank に付与された value			✓	✓
3	BCCWJ-TimeBank に付与された type			✓	✓
3	BCCWJ-TimeBank に付与された class	✓	✓	✓	✓
4	対象表現間の係り受け距離	✓		✓	
5	対象表現が同一文内に存在するか？	✓			
6	対象表現のどちらが先に出現するか？			✓	
7	対象表現間に時間表現が存在するか？	✓	✓	✓	
8	対象表現が文頭もしくは文末の語か？	✓	✓	✓	✓
9	対象表現に付随する助詞・助動詞（機能表現）	✓	✓	✓	✓
10	イベント表現に特定の機能表現が付随するか？	✓	✓	✓	✓
11	拡張モダリティタグ [5] (時制，真偽判断，仮想)	✓	✓	✓	✓
A	時間表現の係り先のイベント表現の文脈情報			✓	
B	大規模データ上のイベント表現間の頻度情報	✓	✓		

3 分類手法

与えられた表現間に付与された時間関係ラベルの分類に，2 次多項式カーネルによるサポートベクターマシン (SVM) を用いる．SVM は分類学習器として頻繁に利用されるモデルであり，再現が容易なことで知られる．本研究では，SVM に多項式カーネルを採用することで，複数の素性の組み合わせを考慮する非線形モデルを扱う．使用した素性の一覧を表 1 に示す．なお，素性作成の際の構文解析に CaboCha[3]を利用した．表 1 の素性は主に英語での既存研究 [1, 4] に基づくものであるが，本研究独自の素性 (A, B) も加えてあり，以下ではそれらに関して説明する．

A. 時間表現の係り先のイベント表現の文脈情報

例えば「12 日 に大会の日程を**発表した**が，取り消すこととなった」という文で,「12 日」と「取り消す」の間の時間関係を考える際,「12 日」の係り先のイベント表現に当たる「発表した」の周辺文脈を素性として取り入れることで，時間表現に不足しがちな手がかりを補う．時間表現の係り先のイベント表現とその周辺の 2 形態素の基本形，依存構造木上の距離，付随する機能的な表現を素性として採用した．

B. 大規模データ上のイベント表現間の頻度情報

各イベントが成立する時間的な順序関係の偏りを大規模データから取得することを試みた素性である．具体的には,「寝る」と「消す」の間には,「（電気など）を消した後に寝る」のように「消す」から「寝る」方向に時間順序の偏りが存在しているといった考えである．

まず，時間的な順序関係を判断する手がかりとなるフレーズを表 2 のように定めた．その上で，Web 文書

*Analyzing Temporal Relations between Events in Japanese Sentenses
Kazuaki Inada†, Yuichiroh Matsubayashi†, and Kentaro Inui†
† Tohoku University

[1]http://www.timeml.org/site/timebank/timebank.html

[2]各時間関係ラベルの詳細は，参考文献 [2] を参照されたい

[3]http://code.google.com/p/cabocha/

表 2: 大規模データからの頻度抽出に用いた表現

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

表 3: 時間関係ラベルのグループ化

グループ化後	グループ化前
AFTER	after, met-by
BEFORE	before, meet
B-or-O	starts, overlaps
O-or-A	overlapped-by, finishes
OVERLAP	during, started-by, equal, is-included, identity, contains, finished-by,includes
VAGUE	vague

約 10 億文に対し，イベント表現 X と Y の間の統語係り受けパスを取る．その中で，X と Y がそれぞれ祖先と子孫の関係にあり，パスの経路中に表 2 の Before および After のフレーズを含むパタンの頻度をカウントし，以下の式で時間関係の偏り (TB) を求めた．ただし，TB は取り得る値の範囲が疎になることが予想されたため，TB を対数で抑えた上で，-5〜5 の整数値に制限した値を素性として用いた．

$$TB(X,Y) = Before(X,Y) + After(Y,X) - \{Before(Y,X) + After(X,Y)\}$$

4 実験と結果

実験では BCCWJ-TimeBank を用いるが，既存研究と問題設定を合わせるため，17 種類の時間関係ラベルを表 3 に従って，6 種類のラベルにグループ化した表 4 のデータを使用した．さらに既存研究に従い，解析対象の種類 (E2E，MAT，T2E，DCT) 毎で問題を 4 つのタスクに分割し，それぞれ個別に分類器を作成した．また，評価指標には正答率 (Accuracy) を用い，5 分割交叉検定よる評価を実施した．なお，学習・分類には LIBSVM[4]を利用した．

表 5 は我々のモデルを，最頻出の時間関係ラベルを全て選択した場合 (Baseline)，吉川らのモデル [3]，TempEval-2 [1] での最高性能のシステムと比較した結果である．我々のモデルは，全てのタスクで吉川らのモデルより高い Accuracy を達成した．特に T2E では，10 %以上と大きく上回っていることが分かる．その主な理由として，我々のモデルでは，解析対象の表現対の周辺に存在する基本形や時間表現の係り先のイベント表現の文脈情報など，吉川らが採用していない素性を取り入れており，さらに分類手法として非線形モデルである多項式カーネルを用いた SVM を採用したためと考えられる．また TempEval-2 と比較すると，各タスクとも遜色無い結果となった．しかし MAT は，他のタスクと比較すると相対的に日本語での性能の低さが目立つ．これは，TempEval-2 では MAT の学習に約 1600 の事例を使用しているが，日本語では 776 事例しか存在せず，使用可能な訓練データ量に倍以上の差が存在するためと考えられる．

次に，現状のコーパス規模で訓練データの量が十分かを調査した．図 2 は，表 4 のデータ量から対数的に使用データ量を増加させた際の学習曲線である．図 2 より，全てのタスクでデータ量増加に伴う Accuracy の飽和が発生していないことを確認できる．従って，学習に用いる事例数を増加させることによって，更なる性能向上が期待できる．また MAT と DCT では，データ量増加に伴う Accuracy の上昇度合いが低いことから，現状の素性のみでは時間関係認識のための素性が不足していると考えられる．実際に表 1 を見ると，これらのタスクで使用した素性の種類数が少ないことを確認できる．よって，訓練データを増加させるのみではなく，特に MAT と DCT では，時間関係認識に有用な素性を確保することも重要と言える．

[4]http://www.csie.ntu.edu.tw/ cjlin/libsvm/

表 4: 実験で使用した BCCWJ-TimeBank のデータ量

	E2E	MAT	T2E	DCT
AFTER	441	201	323	1961
BEFORE	804	305	303	581
B-or-O	0	0	43	1
O-or-A	1	0	22	2
OVERLAP	488	222	814	290
VAGUE	128	48	100	38
合計	1862	776	1605	2873

表 5: 既存研究との Accuracy の比較

	E2E	MAT	T2E	DCT
Baseline	43.18	39.30	50.72	68.26
本研究	61.55	51.03	69.84	80.30
吉川ら [3]	59.9	50.0	55.7	75.6
TempEval-2	60	58	65	82

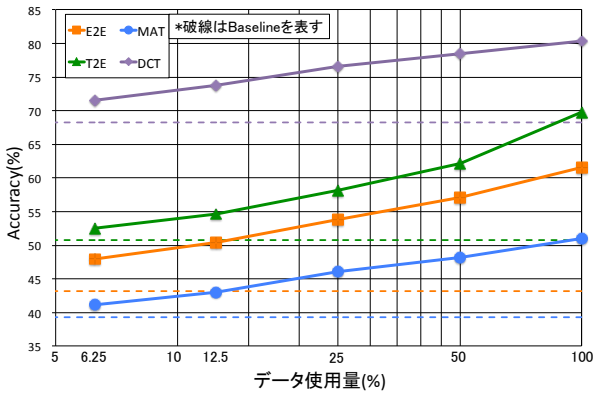

図 2: 訓練データ量増加に伴う Accuracy の変化

5 まとめ

本稿では，日本語でイベント間の時間的な順序関係の認識を実施した．本研究独自の素性と，複数の素性の組み合わせを考慮する非線形モデルを用いたことで，既存研究を上回る性能を達成した．また学習曲線の分析から，現状のコーパス量では性能の飽和が発生していないことが分かり，更なるデータ量の増加に価値があることが分かった．今後の課題として，更なる性能向上のため新規素性の模索や訓練データ量の増加，定性的な観点からの分析などが挙げられる．

参考文献

[1] 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.

[2] Masayuki Asahara, Sachi Yasuda, Hikari Konishi, Mizuho Imada, and Kikuo Maekawa. Bccwj-timebank: Temporal and event information annotation on japanese text. In *Proceedings of 27th PACLIC*, pp. 206–214, 2013.

[3] 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.

[4] Paramita Mirza and Sara Tonelli. Classifying temporal relations with simple features. In *Proceedings of EACL*, pp. 308–317, 2014.

[5] 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.