2008－NL－187 (5)
2008／9／24

推移律を考慮した機械学習手法による時間的順序関係推定

吉川 克正 † 浅原 正幸 † 松本 裕治 †

† 奈良先端科学技術大学院大学 情報科学研究科

†{katsumasa-y, masayu-a, matsu}@is.naist.jp

あらまし 本稿では，英文の新聞記事における時間表現と事象表現を対象として，その時間的な順序関係の推定を機械学習手法を利用して行う手法について述べる．従来より，時間順序関係の推定は多値分類問題の一つとして定義付けられてきたが，学習に必要な素性が明確とは言えず，学習データとして利用できる言語資源も限られているため，十分な精度が得られていないのが現状である．本研究の主旨は，TempEvalで行われた3種類タスクについて再考し，全体の精度を向上させることにある．そのための工夫として，3つのタスク間の関連性に着目し，その関連性を推移律として捉えて新たな学習素性を作り出す手法を考案した．この手法によって得られた結果をTempEvalの最終結果と比較・考察し，素性の有効性を示す．

Machine Learning of Temporal Relation Identification

Katsumasa Yoshikawa† Masayuki Asahara† Yuji Matsumoto†

†Graduate School of Infomation Science, Nara Institute of Science and Technology

Abstract

This paper describes methods for the task: "temporal relation identification with machine learning." The task aims at identifying the relation between two expressions about events or time in English newspaper texts. Although recent works have defined the task as a classification task, any work has not yet achieved enough accuracy since features required to train are not clear and available training data size is small. Our goal is to reconsider the three tasks in TempEval and to improve the accuracy of them by original feature engineering. In this work, we focus on the relations between the three tasks, devise some new features, and also come up with a method to apply them to the tasks effectivly. We compare our results with the final results of TempEval and show the effectiveness of our method for the tasks.

1 はじめに

近年，自然言語処理において盛んになりつつある意味解析研究の一分野として，時間を対象とした意味を解析する時間解析が存在する．機械学習によってこの解析を可能とするためには，時間表現及び事象表現について，その表現ごとに適切な属性を付与したタグ付きコーパスが必要になる．このコーパスに関しても英語をはじめとする各言語について徐々に整備されつつある．このような背景の中で，時間解析の第一歩として提案されているのが，時間または事象の言語表現間における時間的な順序関係を推定する問題である．この問題は各表現間について，Allenの時区間論理 [1] を基にした時間関係ラベルを割り当てる多値分類問題として定義されている．これまでにもその順序関係推定を行う試みがなされてきているが，未だに十分な精度が得られず，決定的な手法も確立されていない．この順序関係推定の先には，時間表現や事象表現が生起した時間をを全て実時間軸上に写像するという，時間正規化の試みも考えられている．時間正規化には様々な応用が期待できるが，それを実現するためにも高い精度で表現間の順序関係を推定する手法が求められている．

本稿に関連する研究として，Boguraevらによって行われた事象-時間表現間の関係推定 [2] 及び，Maniらによって行われた事象-事象表現間における関係推定 [3] がある．彼らの研究ではいずれもTimeBank [4] を利用しており，その中の時間や事象の表現に対して関係推定を行っている．

また，同じく時間関係推定の代表的な試みとして，

SemEval 2007 の Shared Task の一つである TempEval [5] で行われた 3 種類のタスクがある．本研究では，この TempEval のデータを対象に，その 3 種類のタスクそれぞれについて，推定精度の向上を目指すことにする．この目的のために，3 種類のタスクの間に強い依存関係があることに着目し，精度の高いタスクの結果を他のタスクに利用することを考える．この着想に加え，推移律という本来ルールベースで利用する手法を，機械学習の素性として利用する工夫を行っている．この工夫が全体の精度の向上に寄与していることを実験を通して示す．尚，本研究の結果は，Shared Task の最終結果と比較し，素性の有効性を評価する．

以下，2 章では時区間論理とそれをフォーマットとして組み込んだ TimeML の概要について，3 章ではデータとして利用する TimeBank と TempEval のデータに関して，4 章からでは提案手法の詳細を．そして 5 章では実際の実験設定から結果，考察までを述べることにする．6 章は結論と今後の指針である．

2 時区間論理と TimeML

この章では，本研究における主題である時間順序関係の推定について述べるにあたり，前提となる時区間論理及び，TimeML のアノテーションについて述べることにする．

時区間論理は Allen [1] によって提案されたもので，2 つの時区間の間に起こり得る時間的関係を 13 種類の関係ラベルによって定義している (図 1)．さらに，

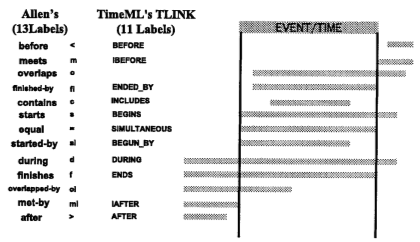


図 1: Allen's time interval logic and TimeML's TLINK Labels

この時区間論理をテキストデータ中の表現に利用する研究も進められているが，その中の一つに時間・事象の表現に対し，時間情報を付与するために開発された XML フォーマット，TimeML [6] がある．TimeML が持つアノテーションには，“the third quater” のような時間表現に対する $\langle TIMEX3 \rangle$ タグ，“have crashed” など，事象表現に対する $\langle EVENT \rangle$ タグと $\langle MAKEINSTANCE \rangle$ タグ，そして “while”，“since” など，手がかり句に対する $\langle SIGNAL \rangle$ などがある．さらにもう一つ，重要なタグとして，時間・事象の表現間における関係情報を付与する $\langle TLINK \rangle$ タグが存在している．このタグが，Allen の時区間論理を基にした，表現間の関係ラベルを属性値として持っている．以下には $\langle TLINK \rangle$ が持つ関係ラベルの例として，“DURING_INV” と “IBEFORE” の例を示した (図 2)．尚，TimeML は英語を対象とし

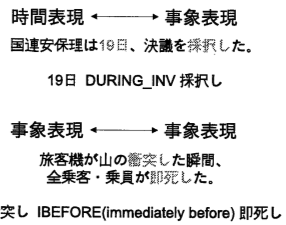


図 2: TLINK Labels Examples

て提案されたものであるため，英語では表現されない “overlaps”，“overlapped-by” の 2 つが除外され，ラベルの数は 11 種類に減じている．次章では，この TimeML を実際に用いてタグ付けした TimeBank コーパスと TempEval データについて述べることにする．

3 TimeBank と TempEval

本章では，時間情報を扱った 2 つのコーパス，TimeBank corpus [4] と，TimeBank を元にして，SemEval 2007 Shared Task [1]のために，いくつかの簡略化と最適化を行い，再構成した TempEval [5] について述べることにする．TimeBank と TempEval の相違点など，実験に使う上で必要になる情報を中心として概説する．

3.1 TimeBank

TimeML を現実のテキストデータに対して利用し，時間情報のタグ付けを行ったのが TimeBank である．TimeBank は Wall Street Journal をはじめとする英語の新聞記事を対象としている．TimeBank 1.1 [2]には 186 の新聞記事が含まれており，TLINK の総数は 5982 である．TimeBank で扱われている TLINK の関係ラベルは TimeML に準拠しており，明確に 11 種類 (BEFORE, IBEFORE, ENDED_BY, INCLUDES, BEGINS, SIMULTANEOUS, BE-

[1] http://nlp.cs.swarthmore.edu/semeval/
[2] http://www.timeml.org/site/timebank/timebank.html

GUN_BY, ENDS, DURING, IAFTER, AFTER) である．

TimeBank について現在知られている主要な問題点として，その関係ラベルの分布に大きな偏りがあることが挙げられる．TimeBank のデータの内，ここでは同一文内の時間表現－事象表現間の TLINK について，そのラベルの分布を以下に示す．

TLINK type	# occurrences
BEFORE	30
IBEFORE	1
ENDED_BY	48
INCLUDES	45
BEGINS	31
SIMULTANEOUS	70
BEGUN_BY	24
ENDS	100
DURING	951
IAFTER	5
AFTER	31
TOTAL	1336

表 1: TLINK type Distributions in TimeBank

表 1 から，DURING と ENDS の関係ラベルだけで，全体の 80% 近くを占めていることが分かる．逆に IAFTER や IBEFORE などは出現回数が少ない．このため，評価の際にはその偏りを十分に考慮する必要がある．

3.2 TempEval

次に TempEval についての概略に移る．TempEval のデータが Shared Task に向けて TimeBank を最適化したデータであることは既に述べたが，データの説明に移る前に，この Shared Task で与えられたタスクについて述べる必要がある．

与えられたタスクは 3 種類あり，全て TLINK の type，即ち，関係ラベルを推定する問題であるが，その TLINK の種類によってタスクが分類されている (図 3)．

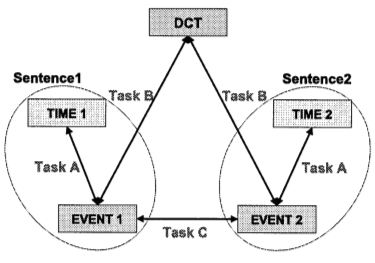


図 3: TempEval 3 Tasks

TASK A 同一文内にある時間表現ー事象表現間の関係推定

TASK B 記事が作成された時間 (Document Creation Time: DCT) と事象表現間の関係推定

TASK C 隣接する文における主述語間の関係推定

TempEval データはこの 3 つのタスクのために，TimeBank を元にして，7 人のタグ付け作業者により人手で関係ラベルを付与し直している．それに際して，関係ラベルは元の TimeBank オリジナルの 11 種類から，6 種類 (BEFORE, OVERLAP, AFTER, BEFORE-OR-OVERLAP(B-O), OVERLAP-OR-AFTER(O-A), VAGUE) へ簡略化されている．しかしながら，タグ付け一致率は Task A,B では 72%, Task C では 65% 程度であることが報告されており，人にとっても難しいタスクであることが窺える．

また TaskA,B で扱われる事象表現については，その語幹 (stem) が 20 回以上出現していることを保証している．もう 1 点記述フォーマットの点で異なるのは，TempEval が，MAKEINSTANCE タグを省略し，EVENT タグに時制やアスペクトの情報までも含めている点である．TimeBank と同じく，TempEval で与えられた訓練データ中の関係ラベルの分布をタスクごとに表 2 に示す．

TLINK type	Task A	Task B	Task C
BEFORE	276	1588	434
OVELAP	742	487	732
AFTER	369	360	306
B-O	32	47	66
O-A	35	35	54
VAGUE	36	39	152
TOTAL	1490	2556	1744

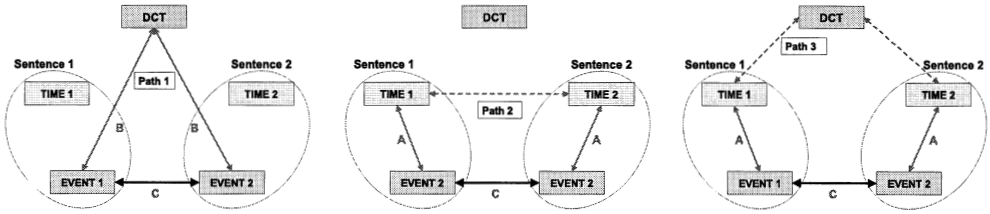
表 2: TLINK Type Distributions in TempEval

関係ラベルの偏りは TimeBank と比較すれば緩和されていると言えるが，Task A ならば OVERLAP, Task B ならば BEFORE といったタスクの性質上避けることのできない偏りもある．

次章から提案手法とその実験について述べるが，以後，本研究では TimeBank の内包する極端な関係ラベルの偏りを避けるとともに，Shared Task との結果の比較を行いやすいという利点から，TempEval のデータを利用していることをここで追記しておく．

4 提案手法

本研究では，時間順序関係推定のタスクにおいて，時間推移律を考慮した学習素性を，効果的に活用する手法を考案した．ここで言う時間推移律とは，例えば “EVENT1-BEFORE-DCT”, “EVENT2-AFTER-DCT” が共に既知である場合に，



(a) Path 1 (b) Path 2 (c) Path 3

図 4: Temporal Relation Paths

"EVENT1-BEFORE-EVENT2" は常に成立することを導いたり，"EVENT1-BEFORE-DCT"，且つ "EVENT2-INCLUDE-DCT" のとき、少なくとも "EVENT1-(AFTER or IAFTER)-DCT" は成り立たないことを導くルールのことである．本手法では、このようなルールを暗に素性として機械学習器に与えることを想定している．

本手法では 3 つの関係推定タスクは $B < A < C$ の順で難度が高いことを前提とする．この事実は TempEval の最終結果から既に明らかであり，また Mani らの研究 [3] によっても事象表現間 (EVENT - EVENT) の関係推定が困難であることは確認されている．そこで最も困難とされている Task C に対して，比較的精度の高い Task A, Task B の結果を，そして Task A に対しては Task B の結果をそれぞれ利用することを考える．

具体的な方法は，Task A,B の結果のラベル（BEFORE，OVERLAP, AFTER など）を組み合わせて，時間関係パスという形で素性にするというものである．つまりは，対象としている事象関係の間に，基軸となる時間を仲介させることで直接的には獲得できない関係を導き出すということになる．これは即ち時間推移律に準ずるものであると言える．

Task C に対して我々が考案した関係パスは図 4 に示す 3 種類である．

Path 1 は Task B の結果を最も直接的に利用する手法で，Task B で得られた DCT と事象表現間の関係ラベルをそのまま組み合わせて素性にするものである (図 4a)．Path 2 は Task A の結果を利用したパスで，時間表現間の関係は TIMEX3 が持つ value 値を比較することで，BEFORE，AFTER，OVERLAP のいずれかを決定し，パスを繋いでいる (図 4b)．Path 3 は Task A の結果を介して，さらに DCT との関係も含めるパスである (図 4c)．

また，Task A に関しても，Task B の結果を利用するために，図 5 のような関係パスを用意している．

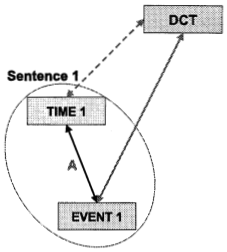


図 5: The Path with DCT (for Task A)

これらの関係パスを素性として用いて関係推定を行うのが提案手法になる．

ここで，我々の手法が，時間推移律をあくまでも素性として利用しており，完全なルールとしては利用していない理由を 2 つ上げる．

まずは，TempEval における関係ラベルが Allen の時区間論理から簡略化された 6 種類であり，その中には "VAGUE" も含む．その結果，明確な意味での時間推移律を構築するのは容易ではないことが一つ目の理由である．

次に，時間推移律とは前提条件の基に厳密なルールとして導かれるものであるから，条件によってはルールを導けない場合もある．特に Path 3 のように考慮すべきパス上の関係数が多くなれば矛盾が生じ易くなる．そして，本研究のタスクのように偏りの大きいデータを扱う際には，厳密なルールを導けない場合が多くなることも有り得る．しかし，我々の手法は，そのような場合であっても，パスという形の素性で，一定の手掛かりを与えることが可能になる．この点において厳密な推移律よりも有効であり，これが 2 つ目の理由になる．

次章では実験環境など，具体的な設定について述べることにする．

5 実験

本研究で行った実験概要を，実験設定，実験素性，そして実験結果に分けて述べていく．TempEval で

用意された3つのタスクそれぞれについて結果を出力し，個別に評価を行っている．さらに，提案手法の核である時間関係パスの利用による結果の違いについては，Task A，及び Task C を対象として別途に評価を行った．

5.1 実験設定

まず，実験に利用した学習ツールを網羅しておく．初めに TempEval データに行った前処理は2つあり，1つ目はデータ中の単語全てに対する品詞タグ付けである．これには TnT ver2.2 [3]を利用した．

2つ目として，その品詞タグ付きのデータに対し，MaltParser 1.0.0 [4] により係り受け解析を行っている．いずれの処理も，次節で示す拡張素性の生成に必須のものである．

さらに順序関係推定のタスクに対する多値分類学習器として，本研究では SVM-Struct [5]を利用している．尚，本実験のタスク全てに渡り，利用したのは線形カーネルのみである．

次に Task A, B, C それぞれについて，訓練データ中でパラメータの設定まで行うために，ディベロップメントセット (DEV) を用意した．そのデータの分割を含め，その中に含まれている TLINK の数を以下の表3に示す．

	TRAIN	DEV	TEST	TOTAL
Task A	1191	299	169	1659
Task B	2106	450	331	2888
Task C	1394	350	258	2002

表 3: TLINKs for the 3 tasks

5.2 実験素性

本実験で問題とする素性には，コーパスから直接得られる素性 (基本素性) と，何らかの方法でデータを加工，解析して作り出す間接的な素性 (拡張素性) があるが，その両方について，有効な素性の概要を分類して順に述べる．

5.2.1 基本素性

基本素性として利用できるのは，事象・時間表現のローカルな情報が主となる．各タスクでどの素性を利用するのかも合わせて示していく．

事象表現情報 (EVENT) まず，事象表現に対しては words, aspect, polality, POS, stem, string, class, そして tense の情報がある．Task A, C についてはこれらの情報を全て利用する．それに対して Task B ではローカルの情報が活用しにくいため，class, tense, aspect のみを利用する．

時間表現情報 (TIMEX3) 時間表現が持つ情報は，value, words, POS, type, TemporalFunction, FunctionInDocument, anchorTimeID がある．これについても Task A, C では全ての情報を利用しているが，Task B に関しては直接時間表現が関わっていないので，対象となる事象表現と同じ文内の時間表現に対して，DCT と比較した時の順序関係のみを利用する．

その他 Task A は同一文内に共起する表現間の関係推定なので，その文内での順序も有力な素性となる．その他，実際の素性としては，対象の時間・事象だけでなく，隣接する文内の表現なども対象としている．

基本素性に関しては素性ごとにその有効性を細かく吟味しているわけではなく，利用できる情報を全て含める形に止めている．ただし，Task B では全ての情報を利用すると明らかにパフォーマンスが低下することも確認しているので，その点だけは考慮した．

5.2.2 拡張素性

拡張素性としては，品詞情報と係り受け情報をまず利用することを考え，さらに提案手法に従い，3つのタスクの情報を相互に関係パスという形で素性として利用する工夫を行う．

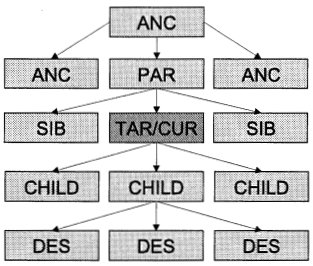
品詞情報と係り受け情報 特に Task A においては，対象となる文の係り受け情報が有力な素性となる．これは対象の時間と事象が，係り受け木においてどのような位置関係にあるかを素性に含めるのが狙いである．大まかには，対象となる文の係り受け木において，対象の表現よりも上にある単語には “ANC”，下にあるものには “DES”，兄弟関係のものには “SIB” という係り受けラベルを付与する．直上や，直下の単語には特別に “PARENT”，“CHILD”，をつけるが，“ANC” と “DES” の延長と捉えてもよい．(図6)

そして，この係り受けのラベルと品詞情報を合わせて素性とする．また，Task A の場合は個々の単語だけではなく，係り受け木における対象表現間のパスもそのまま素性にしている．

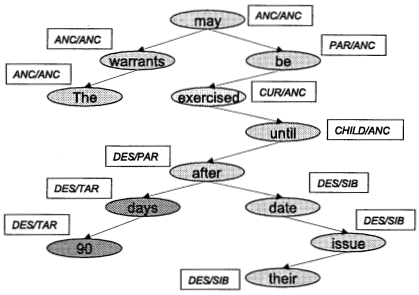
[3] http://www.coli.uni-saarland.de/~thorsten/tnt/

[4] http://w3.msi.vxu.se/~nivre/research/MaltParser.html

[5] http://svmlight.joachims.org/



(a) DepTree Position Labels



(b) DepTree Position Labels (expample)

図 6: Dependency Tree Position Labels

時間関係パス 4 章で示した提案手法を元に生成する時間関係パスである．Task C に関しては 3 種類の Task A に関しては 1 種類の関係パスを用意した．

5.3 実験結果

TempEval はその Shared Task 中において，実験結果の評価にも異なる 2 種類のスコアリングスキームを提案している．よって本研究での結果の評価にもそのスキームを利用する．

提案されたスキームは strict と relaxed の 2 種類である．strict は一つの TLINK に対して，正解であれば 1 を，不正解であれば 0 を評価得点とするスキームで，曖昧さを許さない厳密な評価となる．一方，relaxed は正解であれば 1 を得点とするのは変わらないが，不正解であっても，システムの出力した正解ラベルがゴールドラベルに近い場合には，その近さを考慮して得点を与える手法である．relaxed の得点対応表を以下の表 4 に示す．

	B	O	A	B-O	O-A	V
B	1	0	0	0.5	0	0.33
O	0	1	0	0.5	0.5	0.33
A	0	0	1	0	0.5	0.33
B-O	0.5	0.5	0	1	0.5	0.67
O-A	0	0.5	0.5	0.5	1	0.67
V	0.33	0.33	0.33	0.67	0.67	1

B: BEFORE　　O: OVERLAP
A: AFTER　　B-O: BEFORE-OR-OVERLAP
O-A: OVERLAP-OR-AFTER　　V: VAGUE

表 4: Relaxed Scoring Weights

表 4 の行が表すのはゴールドラベルであり，列が表すのがシステムの出力ラベルである．あるゴールドラベルに対し，システムがどのラベルを出力したかで，得点が一意に決定される．例えば，ゴールドラベルが BEFORE であるものに対し，システムが AFTER を出力した場合の得点は 0 であるが，B-O(BEFORE-OR-OVERLAP) を出力した場合には半分正解という解釈で，与えられる得点は 0.5 となる．

次節以降は，全ての実験結果を strict と relaxed，その両方のスキームにおいて評価する．

5.3.1 Task A (EVENT-TIME)

表 5 の左表は Task A の結果を Precision, Reacall, 及び F 値で示したものある．表には TempEval Shared Task の最終結果を共に示してあり，本研究の結果は最終段となっている．

Shared Task における各チームの詳細については SemEval 2007 [5] に譲るが，CU-TMP, NAIST-japan, USFD の 3 チームは機械学習のみによる手法，LCC-TE 及び WVALI の 2 チームが機械学習にヒューリスティックなルールを追加したハイブリッドのシステム，そして XRCE-T がルールベースでのシステムを採用していることだけを述べておく．

表 5 の右表はゴールドラベルとシステムの出力した正解ラベルの対応表であり，行がゴールドラベル，列がシステムが出力したラベルである．尚，この表で利用した関係ラベルの略称は，全て表 4 と同一である．

Task A の関係推定にも，係り受け解析など，主に対象の文から得られるローカルの情報を使うと同時に，DCT を基軸にし，Task B の結果を含めた関係パスを素性にしていることは既に述べた通りである．この素性が relaxed スコアにおいて大きな精度の改善に貢献していることを，表 5 から読み取ることができる．Task A は同一文内の関係推定という性質上 OVERLAP と AFTER が正解となることが多いが，システムが出力したラベルはこの性質を強く受け過ぎている傾向にあることも分かる．

5.3.2 Task B (EVENT-DCT)

次に Task B の結果を同様にして表 6 に示した．

最終結果を得る際の実験の順序としては，この

team	strict			relaxed		
	P	R	F	P	R	F
CU-TMP	0.61	0.61	0.61	0.63	0.63	0.63
LCC-TE	0.59	0.57	0.58	0.61	0.60	0.60
NAIST.Japan	0.61	0.61	0.61	0.63	0.63	0.63
USFD	0.59	0.59	0.59	0.60	0.60	0.60
WVALI	0.62	0.62	0.62	0.64	0.64	0.64
XRCE-T	0.53	0.25	0.34	0.63	0.30	0.41
Our Method	**0.63**	**0.63**	**0.63**	**0.68**	**0.68**	**0.68**

	B	O	A	B-O	O-A	V	TOTAL
B	6	7	7	0	0	1	21
O	2	84	10	1	0	0	97
A	5	7	16	1	0	1	30
B-O	0	2	0	0	0	0	2
O-A	0	3	2	0	0	0	5
V	5	2	7	0	0	0	14
TOTAL	18	105	42	2	0	2	169

表 5: Results for Task A

team	strict			relaxed		
	P	R	F	P	R	F
CU-TMP	0.75	0.75	0.75	0.76	0.76	0.76
LCC-TE	0.75	0.71	0.73	0.76	0.72	0.74
NAIST.Japan	0.75	0.75	0.75	0.76	0.76	0.74
USFD	0.73	0.73	0.73	0.74	0.74	0.74
WVALI	**0.80**	**0.80**	**0.80**	0.81	**0.81**	**0.81**
XRCE-T	0.78	0.57	0.66	**0.84**	0.62	0.71
Our Method	0.76	0.76	0.76	0.78	0.78	0.78

	B	O	A	B-O	O-A	V	TOTAL
B	174	6	6	0	0	0	186
O	20	47	12	0	0	2	81
A	9	10	30	0	0	0	49
B-O	5	3	0	0	0	0	8
O-A	2	0	0	0	0	0	2
V	4	1	0	0	0	0	5
TOTAL	214	67	48	0	0	2	331

表 6: Results for Task B

team	strict			relaxed		
	P	R	F	P	R	F
CU-TMP	0.54	0.54	0.54	0.58	0.58	0.58
LCC-TE	0.55	0.55	0.55	0.58	0.58	0.58
NAIST.Japan	0.49	0.49	0.49	0.53	0.53	0.53
USFD	0.54	0.54	0.54	0.57	0.57	0.57
WVALI	0.54	0.54	0.54	**0.64**	**0.64**	**0.64**
XRCE-T	0.42	0.42	0.42	0.58	0.58	0.58
Our Method	**0.56**	**0.56**	**0.56**	0.62	0.62	0.62

	B	O	A	B-O	O-A	V	TOTAL
B	31	23	2	1	1	1	59
O	25	85	9	1	0	2	122
A	5	9	27	0	1	0	42
B-O	4	4	1	2	0	1	12
O-A	1	3	3	0	0	0	7
V	3	9	4	0	0	0	16
TOTAL	69	133	46	4	2	4	258

表 7: Result for Task C

Task B の関係推定を最初に行うことになる．この Task B については，特に新しい素性を利用しておらず，その精度も他のチームの結果と比較して大きく変わるものではない．それでも機械学習のみを利用して推定を行ったチームの結果としてはよい数字となっている．

しかしながら，Task B はローカルな情報が利用しにくいこともあり，差がつきにくいタスクであると言える．新聞記事が対象である以上，既に起こった事柄について書くことが多いのは BEFORE のラベルが半分以上になることから容易に分かる．さらには他の 2 つのタスクより比較的曖昧さが少ない．即ち，BEFORE，OVERLAP，AFTER，の 3 ラベルにうまく分かれている TLINK が多い．結果としては，この Task B が 3 つのタスクの中では各チームとも，最も高い数値になっている．

5.3.3 Task C (EVENT-EVENT)

同様にして Task C の結果を表 7 に示した．Task C は Task B とは対称的に，3 つの中で最も精度が低いタスクである．事象表現間の関係推定は，対象が動詞に限らず，名詞化した動詞なども対象になっているため，時制やアスペクトなど，基本素性だけでの推定は困難であると言える．

本研究の Task C の結果は，strict スコアにおいて最高の値を出している．出力ラベルの分布から推察できることは，Task A と同じくデータの分布が多い BEFORE, OVERLAP, AFTER, のラベルに偏ってしまい，例えば，VAGUE などは全く正解していないということである．

Task C においては特に，他の 2 つのタスクの結果を利用することが重要である．その根拠となるのが，表 8 に示した関係パスの利用による精度の違いである．

表 8 は，Task A と C についての結果だが，まず Task C は，左端が関係パスを利用しない場合であり，strict スコアで 54%，relaxed スコアで 60% 程度であった．これに 3 つのパスを加えた結果をそれぞれ示している．Path 2 については利用することでかえって精度を引き下げてしまう結果になったが，Path 1 では strict スコアで凡そ 2%，Path 3 では 0.5% 程度の精度の向上が見られた．また “All” の列が示すのは 3 つのパスを全て利用した場合だが，この結果は Path 1 の精度を越えられなかった．

さらに，この表 8 で各 Path の右隣にある “(G)” の列が意味するのは，Task A,B の実験結果の変わ

	Task C									Task A		
Path	None	Path 1	Path 1 (G)	Path 2	Path 2 (G)	Path 3	Path 3 (G)	All	All (G)	None	Path	Path (G)
strict	0.539	0.562	0.562	0.531	0.539	0.547	0.558	0.558	0.562	0.598	0.627	0.609
relaxed	0.601	0.622	0.622	0.591	0.600	0.604	0.615	0.616	0.620	0.654	0.682	0.668

表 8: Results with Various Paths

りに，ゴールドラベルを利用した場合の結果である．このことを踏まえると，もう一つ興味深い点を見つけられる．Path 3については，ゴールドデータを利用した方が明確に精度が改善されるのに対し，Path 1を利用した結果は，Task Bの関係ラベルをゴールドデータで置き換えた場合でも数値に変化が見られないという点である．この理由として考えられるのは，Task A, Bのタグ付け一致率が72%程度であることから，Task Bの結果を手がかりにして引き上げられる精度としては，この数値が限界であるという可能性である．

この仮説を支持するのが表8の右側にあるTask Aの結果である．Task AではPathを利用することにより精度は向上しているが，ゴールドデータを利用した場合の方がシステムの出力ラベルを利用した場合よりも精度が下がっている．これはタグ付けの不確かさを顕著に示していると考えられる．

以上でTask A, B, C全ての結果を示したことになり，Task A, Cについては，確かに精度の向上を確認することができたが，各チームのデータを見ても，概してTask Aで65%，Task Bで80%，Task Cで60%程度というのが現在のアプローチの限界であることが分かる．さらに高い精度を求めるのであれば，TimeBank，或いはTempEvalのデータサイズが小さいことを補う意味でも，外部知識の活用が必要であることを示唆する結果となった．

6 結論

本研究では，SemEval2007 Shared Taskを基にして，3種類の関係推定タスクをそれぞれ行い，精度の向上を試みた．精度向上の方針は，3つの関係タスクを相互に関連付けることにより，新たな素性を作り出すことだったが，その工夫によって確かに精度を全体の精度を向上させることに成功した．

しかしながら，本稿の結論として，TimeBank，或いはTempEvalという，規模の小さなコーパスから得られる情報だけでは，既に限界であることが実験結果の点で明らかなため，今後，動詞オントロジーをはじめとした，外部知識を積極的に活用する手法を模索するとともに，英語に限らず，他の言語においても利用可能な素性を同定する試みを予定している．

参考文献

[1] James Allen. Maintaining knowledge about temporal intervals. In *Communications of the ACM*, pp. 832–843, 1983.

[2] Branimir Boguraev and Rie Kubota Ando. Timeml-compliant text analysis for temporal reasoning. In *Proceedings of IJCAI-05*, pp. 997–1003, 2005.

[3] Inderjeet Mani, Marc Verhagen, Ben Wellner, Chong Min Lee, and James Pustejovsky. Machine learning of temporal relations. In *Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics, Sydney Austoralia.*, pp. 753–760, 2006.

[4] 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.

[5] Marc Verhagen, Robert Gaizaukas, Frank Schilder, Mark Hepple, Graham Katz, and James Pustejovsky. Semeval-2007 task 15: Tempeval temporal relation identification. In *Proceedings of the 4th International Workshop on SemEval-2007.*, pp. 75–80, 2007.

[6] James Pustejovsky, Jose Castano, Robert Ingria, Reser Sauri, Robert Gaizauskas, Andrea Setzer, and Graham Katz. Timeml: Robust specification of event and temporal expressions in text. In *Proceedings of the Fifth International Workshop on Computational Semantics(IWCS-5)*, 2003.