自 然 言 語 処 理 140−16
（2000. 11. 22）

# 文間の時間制約モデルと事象の時系列化への応用に関する研究

小倉 牧人† 田村 直良††

†横浜国立大学 工学研究科 電子情報工学専攻
††横浜国立大学 教育人間科学部 情報認知システム講座
{makito,tam}@tamlab.dnj.ynu.ac.jp

関連のある一つの事例について調べていこうとすれば、必然的に複数の文章を対象としなければならない。そこで、本研究では、それらの文書をひとつにまとめることを目指し、記事の収集、時系列化、重要事項の抽出という要約モデルを提唱し、特に時系列化について論じる。時区間関係を捉える枠組みとして Allen の時区間推論が有名である。このことから時系列化に Allen のモデルを適用させる。そのためには、時間情報のない文の defualt 処理が必要である。我々はこの default 処理について提案する。また、提案したモデルを用いて実際に実装をめざす。本報告では第一段階として一文章中の出来事の時系列化を目標とし、新聞文書からのデータセットを用いて時系列化し、その評価を行った。

# A Model of Temporal Constraints between sentences and Its Application to Serialization of Actions

Makito Ogura† Naoyoshi Tamura††

†Department of Electrical and Computer Engineering,
Yokohama National University

††Department of Information and Cognition Systems
Faculty of Education and Human Science
Yokohama National University

{makito,tam}@tamlab.dnj.ynu.ac.jp

In this paper we describe a model of temporal constraints between sentences and its application to serialization of actions. When we make a report of an event, we have to retrieve information from multiple texts about the event and summarize them. We present a summarization model, which proceeds as follows: gathering articles from newspaper, serialization of actions in the events and extraction of characteristic actions. Especially in serialization process, we make an extension of Allen's time interval model to deal with temporal default relations between actions in the event. We also present a prototype implementation and evaluation of our method.

## 1 はじめに

インターネットに限らず、昨今は大量の電子化されたドキュメントが日々増え続けている。そのような中で、それらのドキュメントを人が逐一読んで処理するのは困難となってきている。そこで、必要な情報をすばやく効率よく手にいれるために、それらのドキュメントの自動要約や抄録などといった要求が高まってきている。

特に、新聞記事のデータベースやオンライン記事などのように、出来事について書かれたドキュメントは、別の視点で書かれたものや、その出来事の続きなど複数のドキュメントが書かれている場合も多い [1]。近年、そのような multi-document を要約してひとつにまとめるという、複数文書の要約の要求が高まってきている。

本研究では、このような複数文書をひとつにまとめることを目指す。

複数文書の要約の問題のひとつに、どのような観点で文書をひとつにまとめるかがある。本研究では、特に時間に着目する。つまり、複数文書をひとつにまとめる方針として、出来事を記述している文が表すイベントに関して時系列化することは有効な方針だと考えている。

時間推論を扱うモデルのひとつに、James F.Allen によって提唱された時間モデル [2] がある。これは、時間をある大きさを持つ時区間 (time interval) として扱い、その順序関係のみで表したものである。本研究では、Allen のモデルを基礎に文の時系列化のためのモデルの拡張を提案する。

## 2 理論的枠組み

### 2.1 出来事の定義

時間に関するモデルを述べるために、まず基礎とすべき「出来事」を以下のように定義する。

**[定義] 出来事**

**出来事** $A$ は、イベント $e_1, e_2, ..., e_n$ の集合と時間軸上の時間 $t_1, t_2, ..., t_m$ と、それらの間の時間関係 $r_1, r_2, ..., r_l$ の集合の三つ組である。

$$A = (\{e_1, ..., e_n\}, \{t_1, ..., t_m\}, \{r_1, ..., r_l\})$$

ここで、イベント e とはフレーム構造（格フレーム）である。また、時間関係 $r$ とは時区間の間に結ばれるリンク $l$ であるとする。

$$r = l(C, e_a, x)$$

ここで、$C$ は Allen の定義する時区間どうしの関係の集合である。この関係を図 1に示す。また、$x$ は、イベント $e_b$ もしくは時間軸上の時区間 $t$ のどちらかをとる。

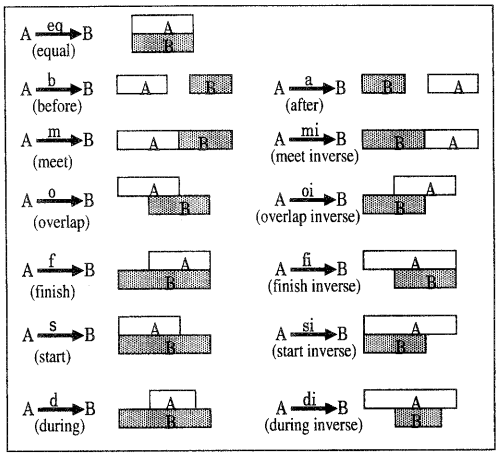


図 1: Allen の 13 の関係 [2]

それぞれのイベントをあらわすフレームには、おのおの表 1に示すようなスロットを持つとする。

| | |
|---|---|
| agent(動作主) | 動作主または状態の主格。 |
| item(道具格) | 動作の原因となること。 |
| place(場所格) | 動作の起きている場所。 |
| source(源泉格) | 変化や移動の起点。 |
| goal(目標格) | 変化や移動の終点。 |
| object(対象格) | 変化や移動の対象。 |
| action(動作) | 主体の動作。 |
| status(状態格) | 主体の状態。 |

表 1: フレームのスロット

主となるのは、動作主を表す agent とその動作をあらわす action である。他に主格の状態を表すフレームでは、status(状態) スロットを持つとする。つまり、ひとつのイベントは文書中のひとつの動作もしくは状態に対応している。ここで注意するのは、時間に関する情報はこのスロットには含まれていないということである。イベントの時間を表すのは、あくまで時間軸上の時間 $t$ と、その間に張られたリンクか、もしくはイベント間に張られたリンクで表される。時間軸上の時刻 $t$ もある大きさを持つ時区間として取り扱う。

リンク $l$ には、以下の 4 種類が存在する。

1. 直接リンク
   文中に時間情報がある場合、それを用いて張られるリンクのことである。
2. 推論リンク
   直接リンク、もしくは他の推論リンクから時間推論して張られるリンクのことである。時間推論については次節で述べる。
3. default リンク
   文中に時間情報が無い場合に、文中の時間的な要素以外から張られるリンクである。

4. default 推論リンク
   default のリンク、もしくは他の default 推論リンクをもちいて default 時間推論されたリンクである。default 時間推論についても次節で述べる。

このようにして、定義した出来事の例を図 2に示す。なお、推論に関するリンクについては触れず、純粋に文書中の情報のみを表してる。ここで図中の $e_3$ は、文中に時間情報がなかったイベントであり、それは前のイベント $e_2$ と default のリンクを張っている。このようなイベントを default のイベントと呼ぶことにする。

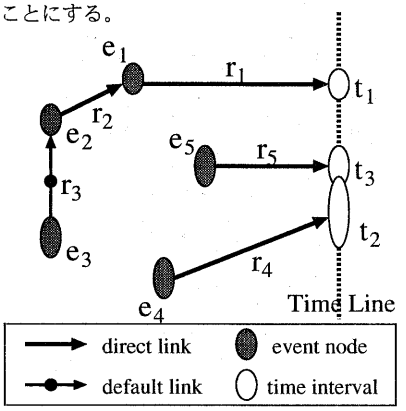


図 2: 出来事の概観

## 2.2 時間推論の定義

時間推論とは、リンクの張られていないイベントの間の時間関係を遷移的に求めることである。いま、イベント $e_1$ と $e_2$ の時間関係を $r_1$ とし、$e_2$ と $e_3$ の時間関係を $r_2$ とする。$e_1$ と $e_3$ の関係 $r_3$ は、$r_1$,$r_2$ の関係から Allen の提唱する遷移表 (Transitivity Table) を用いて図 3のように推論することができる [2]。

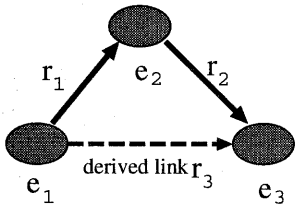


図 3: 時間推論 [2]

この遷移表は、縦軸に $r_1$、横軸に $r_2$ をとったもので、値として考えられる得る $e_1$ と $e_3$ の関係 $r_3$ を持っている。表 2に Allen の遷移表の一部を示す。

表中の「no info」は、その $r_1$ と $r_2$ の関係からは、$r_3$ の関係を決定する情報が何もないということを表している。

| $r_2$ / $r_1$ | b | a | o | oi | m | mi | .... |
|---|---|---|---|---|---|---|---|
| b | b | no info | b | b o m d s | b | b o m d s | .... |
| a | no info | a | a oi d mi f | a | a oi mi d f | a | .... |
| o | b | a oi di mi si | a o m | o oi d di s si f fi eq | a | si oi di | .... |
| oi | b o m d fi | a | o oi d di s si f fi eq | o di fi | o di fi | a | .... |
| m | b | a oi mi di fi | a | o d s | a | f fi eq | .... |
| mi | b o m di f | a | oi d f | a | s si eq | a | .... |
| : | : | : | : | : | : | : | |

表 2: Allen の遷移表の一部 [2]

ここまでが Allen によって提唱された時間推論の機構であるが、この機構では、時間に関するリンクが無いノードは他のノードへ遷移できないので推論することができないことになる。しかし、文書を対象に考えると時間情報のないイベントも多く出現する。本研究では、そのようなイベントに関しては default のリンクを張ることにし、それを以下のように定義する。

**[定義] default リンク**

**default リンク**とは、文中に時間に関する情報がない default のイベントと、文書中でその直前に記述されたイベントとの間に張られるリンクである。

これは、時間に関する情報がないイベントは、通常、もとの文書中で出現した順番にイベントが進行していると仮定したためである。このリンクもやはりその関係は Allen の提唱する 13 種類の関係をとるとした。またその決定には、イベントの文書中での位置などの構造的な要素と、イベントの種類などの要素を用いて決定するとし、通常は b(before) の関係を採用した。

これにより時間情報のない default のイベントにも文書中の特徴からリンクが張られることになり、時間推論が可能となる。default 時間推論を以下のように定義する。

**[定義] default 時間推論**

**default 時間推論**とは、$e_1$ と $e_2$ の間の関係を $r_1$、$e_2$ と $e_3$ の間の関係を $r_2$ とするとき、$r_1$ もしくは $r_2$ の少なくともどちらかが default リ

ンクもしくは、default 推論リンクであり、かつ、$e_1$ と $e_3$ の間に直接リンクもしくは推論リンクが張られていない時に、遷移表により関係 $r_3$ を張ることである。

これは、default リンクの時間関係は直接リンクの時間情報に比べて正確さが落ちるためである。

推論機構においては、まず最初に直接リンク、推論リンクの時間推論を行い、次に残りのイベント間の関係について default 時間推論を行うとする。（図5参照）

### 2.3 ソーティングの方法

このようにして、イベント間に張られた時間関係のリンクを用いてソーティングを行う。時間推論によってリンクが張られたとはいえ、文中に出てくる時間情報ではすべてのイベント間の時間関係が一意に決まるとは限らない。例えば、表2の遷移表においても「no info」となっていればそのリンクに関しては時間関係が分からないことになる。つまり、イベント間の順序は半順序関係となっている。したがって、そのような仕組みを取り扱うトポロジカルソートを用いてソーティングを行うことにする。

今、それぞれのイベントの時間関係は Allen の提唱する 13 種類の関係になっているので、これを大小関係のみに変換し比較可能にする。基本方針としては、始点が前に出てきたものを前のイベントとして扱う。始点が同じときは時間の大きさが大きい場合を前にする。また、イベントの時間が全く同じ場合は、文中に記述された順番を尊重することにする。表 3にその分類を示す。

| イベント $e_1$ と $e_2$ の関係 | その順序 |
|---|---|
| b , m , o, si ,fi ,di | $e_1 < e_2$ |
| a, mi, oi, s, f, d | $e_1 > e_2$ |
| eq | 文中での順番 |

表 3: 時間関係の変換

このようにして並び替えた場合、一意に順序が求まるとは限らないが、条件をみたす一つの解が求まれば良いと考える。

## 3 実装

### 3.1 システムの概略

この節では、前述のモデルを用いて、複数の文書を時系列順に並び替えるシステムを実現するためにはどうしたらよいかについて論じ、実際の実装方法について述べる。

図 4に提唱するシステムの概略図を示す。

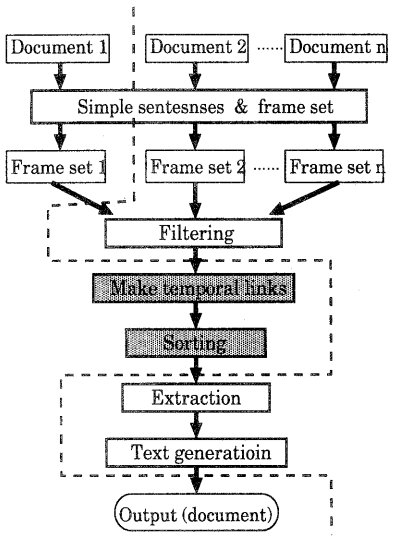


図 4: システムの概要

以下の順で処理を行っていく。

1. 複数文書をそれぞれ単文化モジュール、格フレーム化モジュールに順次入力し、それらのイベントをひとつにまとめて出力する。
2. 同じイベントを表す部分をフィルタモジュールでまとめる。
3. 作成されたフレームの集合を時間リンク作成モジュールに入力し、時間推論してでき得るすべてのリンクを張る。
4. 張り終わったリンクを条件としてソーティングを行う。
5. 重要部分の抽出を行い出力する。
6. 格フレームをもちいて文章を作成し出力する。

実際の研究のプロトタイプとして図中の破線の左側の機構で、一文書についての時系列化を目指す。今回実装したのは、時間リンク作成モジュールとソー

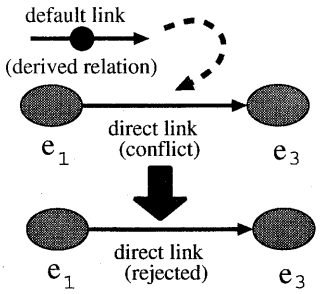


図 5: default 時間推論の機構

ティングであった。以下、用いたデータセットとそれぞれのモジュールについて述べる。

### 3.2 データセット

今回の実装では、文章から格フレームを構造に変換する部分は含まれない。複文は単文化され、文ひとつひとつのイベントと考え、格フレーム化し、これらをすべて人手により記述した。特に agent については必ず補うようにした。文中の時間情報もそれぞれ照応を考えた上でデータセット中に埋め込んだ。

実際は、1993 年日本経済新聞のコーパスから事件記事 50 件を対象にデータセットの作成を人手で行った。格フレームは prolog 形式で書いた。図 6に作成した格フレームの例を示す。

```
%-------------------------------------------------------
%イベント
event(e(1),[
        agent: 会社員らを狙った路上強盗,
        place: 大阪,
        action: 発生する,
]).
event(e(2),[
        agent: 大阪府警,
        goal: 強盗傷害事件,
        action: みて調べる,
]).
event(e(3),[
        agent: 会社員、Aさん,
        place: 大阪府警大正署,
        object: 「路上強盗の被害に遭った」,
        action: 届け出る,
]).

% 時間軸上の時間
time(t(1),[1992,12,31]).      % 1992 年 12 月 31 日
time(t(2),[1992,12,31,2,35]). % 1992 年 12 月 31 日 2 時 35 分
% リンク
link(l(1),e(1),t(1),dr,[eq]). % 時間軸へのリンク
link(l(2),e(1),e(2),df,[b]).  % default リンク
link(l(3),e(3),t(2),dr,[eq]).
%-------------------------------------------------------
```

図 6: 作成したデータセットの例

作成したデータセットにおいて時間情報について調査した。調査は、

1. 各イベントの時間情報の有無とその数
2. default のイベントとその直前のイベントの関係

について行った。調査 1 の結果を表 4に、調査 2 の結果を表 5に示す。

| イベントの種類 | 出現イベント数 | |
|---|---|---|
| 時間軸とのリンクがある | 343 | |
| イベント間の順序情報がる | 79 | 422 |
| default のイベント | 547 | |
| 総調査イベント数 | 969 | |

表 4: イベントの種類の調査結果

| 要素 | b 系 | a 系 | その他 | 計 |
|---|---|---|---|---|
| 段落先頭 | 83/90% | 8/9% | 1/1% | 92 |
| その他 | 89/72% | 32/26% | 2/2% | 123 |
| 動作 | 108/81% | 25/19% | 0/0% | 133 |
| 状態 | 64/78% | 15/18% | 3/4% | 82 |
| 時間軸 | 62/87% | 9/13% | 0/0% | 71 |
| 順序 | 25/93% | 2/7% | 0/0% | 27 |
| なし | 85/72% | 29/25% | 3/3% | 117 |
| 同一文 | 54/90% | 5/9% | 1/1% | 60 |
| 非同一文 | 118/76% | 35/23% | 2/1% | 155 |
| 全体 | 172/80% | 40/19% | 3/1% | 215 |

表 5: default のイベントの調査結果

調査 1 の結果から、default のイベントが全体の半数以上をしめていることが分かる。したがって、default の処理が重要であることが分かる。また、順序を表す時間情報よりも時間軸上の時間の方が多いことが分かった。これは、事件記事では事件の発生した時間を正確に伝える必要があるからだと思われる。

調査 2 では、default のイベントとその直前に記述されたイベントとの関係に表 3の変換を適用してみた。表中の、b(before) 系は、直前に記述されたイベントの後に default のイベントが来るという関係であり、a(after) 系は直前のイベントの前に default のイベントが来るという関係を表している。全体を通して default のイベントは b 系の関係にな

る場合が8割であった。このことから、defaultのイベントは通常はbの関係をもつとして良いことが分かる。特に、直前のイベントが段落先頭である時や、時間情報を持つ場合やまた単文化前はdefaultのイベントと同一文だった場合は9割ほどであった。a系の場合になるものが多かったのは、前のイベントが段落の先頭ではなく、かつそれ自身もdefaultのイベントであった。この場合に関してはさらなる調査が必要である。また、b系、a系のどちらかに絞れない場合はほとんどなかったが、それらは直前の文が状態の文である時のみに発生していた。

これらのパラメータは、defaultのリンクの関係を決定するのに用いることができる。

### 3.3 実装したモジュールについて

今回、実装したのは、時間リンク作成モジュールとソーティングモジュールである。使用したプログラム言語はprologである。以下それらのモジュールについて述べる。

図7に時間リンクの作成モジュールの概略を示す。

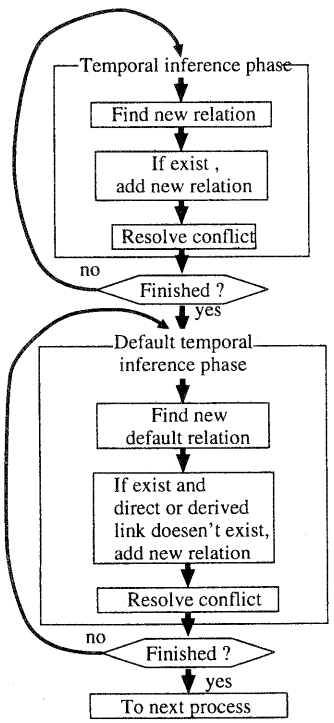


図 7: 時間リンク作成モジュールの概略

以下のように時間リンク作成を行う。

1. 入力されたリンクに対して推論リンクを張る。
   (a) 時間推論することができる関係を探し、あれば推論リンクを張る。
   (b) 同じイベント間に複数のリンクが張られている場合は、それぞれの関係の和集合を取り、絞り込みを行う。
2. すべての推論リンクを張り終えたら、default推論リンクを張る。
   (a) default時間推論できる関係を探す。
   (b) 張られたリンクの中から直接リンクまたは推論リンクが既にあったら、defaultリンクもしくはdefault推論リンクは張らない。そうでなければ、default推論リンクを張る。
   (c) 同じイベント間に複数のdefaultリンクもしくはdefault推論リンクが張られている場合は、それぞれの関係の和集合を取り、絞り込みを行う。

このように、時間リンクの作成は二段階の処理を行う。

ソーティングは、トポロジカルソートを行い、作成されたリンクの条件を満たす解をひとつ出力する。その際にそれぞれのリンクの関係は表3を用いて変換する。実際には、以下の二つの処理を繰り返し行う。

1. リンクの張られていないイベントがあれば、残っているイベントの中で、その文中での位置が前でかつ一番近いイベントと同グループにする。さらに、そのグループ内で文中の出現順通りソーティングする。
2. 張られているリンクの関係すべてが、「b」のイベントを探し、それを出力し、そのイベントに張られていたすべてのリンクを切る。

### 3.4 抽出機構について

今回の実装では範囲外であったが、抽出機構についても考察する。

文書が多くなるほど同一の事象を表すイベントを判別し、一つにまとめる必要が出てくる。そのような処理を行うモジュールをフィルターモジュールと

呼ぶことにする。フィルターモジュールでは、以下の二つのルーチンが存在するだろう。

1. 同じイベントの判別

2. ひとつにまとめる機構。

1に関しては、格フレームの各スロットを比較した類似検索になると思われる。2に関しては、二つの場合が考えられる。すなわち、内容が同じ場合と異なっている場合である。内容が同じ場合は、それらのイベントの情報の和集合をとり、一つのイベントのスロットに統合する機構が必要である。また、異なっている内容の場合は、信頼性の高い文を選択する機構が必要となる。いづれにせよ、そのようなイベントの統廃合を行った後のリンクの修正も必要であると考えられる。このフィルターモジュールの詳細なアルゴリズムは今後の課題である。

また、ソティーング後に重要事項の抽出を行うが、これは、文書を抄録する上でより重要度の高いイベントを選択する機構である。状態を表すイベントやdefaultのイベントの中には、時系列化するうえで取り除くべきイベントも存在していると思われる。そのようなイベントを取捨選択するためのパラメータを模索し、選択の機構を考えるのも、今後の課題としてあげられる。例えば、イベントに張られたリンクの数などもそのパラメータの一つの候補として挙げられるのではないかと考える。

## 4 結果と評価

実際に実装した機構を用いて、時系列化を行った例を示す。まず、実験に使用した例文を以下に示す。

---

＜単文化された事件記事＞

1．愛知県で二十九日深夜、タクシー運転手が客の男に顔を殴られた。
2．売上金など十五万円が入ったカバンを奪われた。
3．愛知県警は三十一日、B容疑者を強盗傷害の疑いで緊急逮捕した。
4．B容疑者は二十九日午後十一時十分ごろ建設工事現場で、Cさんのタクシーを停車させた。
5．Cさんの顔に軽いけがを負わせた。
6．売上金など約十五万円入りのカバンを奪った。

---

＜データセットに埋め込んだリンク＞

```
1 -- dr[eq] --> 時間軸（29日深夜）
1 -- df [b] --> 2
3 -- dr[eq] --> 時間軸（31日）
4 -- dr[eq] --> 時間軸（29日23時10分）
4 -- df [b] --> 5
5 -- df [b] --> 6
```

---

drは直接リンク、dfはdefaultリンクを表している。これを用いて張られたリンクと並び替えられたイベントを以下に示す。

---

＜最終的に張られたリンク＞

```
1 -- dr[eq]   --> 時間軸（29日深夜）
1 -- df [b]   --> 2
1 -- inf[b]   --> 3
1 -- inf[di] --> 4
1 -- dfi[b,o,m] --> 5
1 -- dfi[b,o,m]  --> 6
2 -- dfi[b,o,m,di,fi] --> 3
2 -- dfi[a]   --> 4
2 -- dfi[a,oi,mi,d,f] --> 5
2 -- dfi[a,oi,mi,d,f] --> 6
3 -- dr[eq]   --> 時間軸（31日）
3 -- inf[a]   --> 4
3 -- dfi[a]   --> 5
4 -- dr[eq]   --> 時間軸（29日23時10分）
4 -- df [b]   --> 5
4 -- dfi[b]   --> 6
5 -- df [b]   --> 6
```

---

＜時系列化されたもの＞

1．愛知県で二十九日深夜、タクシー運転手が客の男に顔を殴られた。
4．B容疑者は二十九日午後十一時十分ごろ建設工事現場で、Cさんのタクシーを停車させた。
5．Cさんの顔に軽いけがを負わせた。
6．売上金など約十五万円入りのカバンを奪った。
2．売上金など十五万円が入ったカバンを奪われた。
3．愛知県警は三十一日、B容疑者を強盗傷害の疑いで緊急逮捕した。

---

infは推論リンクをdfiはdefault推論リンクを表している。

最終的なリンクをみると、3と6の間のリンク以外はすべて推論により張られた。3と6の間のリン

クが張られなかったのは、推論の結果、時間関係が「no info」であったためである。このように、イベント間の時間関係は半順序関係になることが分かる。この例において、張られたリンクを用いて実際にソーティングを行うと、おおむね時系列化は正しくなされていると思われる。

また、6と2は同じactionを表すイベントであった。このようなイベントはフィルターでひとつにまとめる必要性があると分かる。

## 5 まとめと今後の展望

複数文書をひとつにまとめるために、時系列化という方針を採用し、そのための出来事の枠組みを定め、Allenの時間モデルを基礎に文の時系列化のためのモデルの拡張を提案した。また、そのモデルを用いて、プロトタイプとして時間リンク作成モジュールとソーティングモジュールを実装し、実際に新聞文書の事件記事1文書の時系列化が正しく行われていることを確認した。

今後の展望としては以下のものが挙げられる。

1. defaultのイベントと直前のイベントのさらなる関係の調査
2. フィルターモジュールの機構の考案と実装
3. 重要事項抽出のアルゴリズム
4. 複数文書適用時のモデルの考案

1に関しては、defaultのイベントの直前のイベントがdefauitのイベントである時などに関して、文末表現、接続関係などの調査パラメータを追加し調べていく。2と3に関しては3.4節で述べた通りである。

以上のことがらを踏まえて、4の複数文書の時系列化をめざす。

## 参考文献

[1] Kathleen Mckeown and Dragomir R.Radev, "Generating Summaries of Multiple News Articles," SIGIR 1995,pp.74-81.

[2] James F.Allen, "Maintainig Knowledge about Temporal Intervals," Representation of Commonsense Knowledge 1983, pp.510-521.