1C4-2

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

The Research about Dealing Time Relation of Events from Plural Documents

井上 晃太[*1] 佐藤 真[*2] 赤石 美奈[*2]
Kota Inoue Makoto Sato Mina Akaishi

[*1] 法政大学大学院情報科学研究科
Graduate School of Computer and Information Sciences, Hosei University

[*2] 法政大学情報科学部
Faculty of Computer and Information Sciences, Hosei University

This paper proposes a framework to clarify a context of events reported in newspaper articles. A huge number of newspaper articles are published in real time. It is difficult to grasp the context of events in a large amount of articles. First, we extract event information from newspapers articles. Then, the event information is arranged in chronological order to show its context. In this paper, we evaluate a method for event information extraction.

1. はじめに

大規模な事件・事故・災害は大量の情報を包括しており，それらの最新情報は報道によって随時公開され，私たちはそれらの報道を収集し，整理することによって事件・事故・災害の出来事の流れを把握することが出来る．特に新聞記事やニュースサイトで掲載される報道記事は文章として残されるため，それらを追って情報を収集・整理しやすいが，1つの記事に記述される情報は全体のごく一部のみである．このため，全体の流れを時間順に把握するには関連するニュースを全て収集し，時間情報を記事中から見つけ出し，それを基に情報を整理する必要がある．人手により，大量のニュース記事から情報をまとめるには時間がかかるため，効率よく情報を集約する仕組みが必要とされている．

新聞記事から情報を取り出すための研究は既に存在し，日本では IREX[関根 98]における日付，時間の表現抽出の課題があり，固有表現抽出を用いた新聞記事から必要な情報の抽出をはじめとする研究[山田 02][斉藤 13]が行われている．また，IREX の固有表現抽出の定義を基に作成された CaboCha[工藤 02]等の固有表現抽出が可能な文書解析器も公開されている．

本研究ではニュース記事で報道されるイベントの時間順序に着目し，これを容易に把握できるような情報提示手法を提供することを目的とする．このため，まず固有表現抽出によってニュースサイトの記事からイベント情報を取り出し，それらのイベント情報を，時間的に整理するための手法について提案する．

2. 概要

図 1 は本論文におけるイベント情報の抽出の流れである．本節では，記事からイベント情報を抽出する手法について述べる．

本論文では大規模ニュース内に含まれる事象をイベントとして扱い，その中に含まれる時間情報や主体，発生場所などをイベント情報として定義する．また記事文の特徴から，固有表現抽出とそれを補う抽出ルールを提案することにより，イベント情報を抽出している．特に時間情報の抽出について相対的な時間や，曖昧な時間情報について，取得済みのイベント情報やニュース記事の公開日時，時間情報の範囲からイベント情報を推測できるよう抽出ルールを定義している．

連絡先：井上 晃太，kota.inoue.6h@hosei.stu.ac.jp

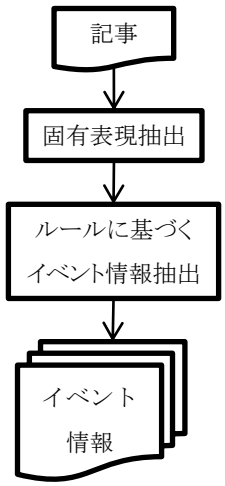

図 1： 記事からのイベント情報抽出の流れ

3. 記事からのイベント情報の抽出

3.1 イベントとイベント情報

本研究では大規模な事件・事故・災害といった大きな出来事を大規模ニュースとし，大規模ニュース内で発生した出来事をイベントとして扱う．ここで言うイベントとは，例えば「大地震の影響で発生した火災」や「家屋の倒壊」といった詳細な情報のことを示す．イベントには発生した場所や時間などのイベントの状況に関する情報が含まれている．これをイベント情報として大規模ニュース内のイベントの時間関係の整理やイベントの関連性を調べるために用いる．表 1 にイベント情報のデータ構造を示す．イベントの発生した日時の比較に発生日と発生時刻．他のイベントとの関連性を調べるために主体と発生場所の情報を用いる．また，一部の日時情報との比較や同一記事内で書かれたイベントの参照のためにニュース記事の公開日時を使用する．イベントの内容は解析にかける前の文章そのものとし，同イベントに関する文章は新しいイベントを取得しない限り記事内の記述順にまとめる．

表 1: イベント情報のデータ構造

情報	内容
発生日	イベントの発生した日
発生時刻	イベントの発生した時刻
主体	イベントの主体となる語
発生場所	イベントの発生した場所
ニュース記事の公開日時	ニュース記事の公開日時
内容	元の記事文. 同記事内で同イベントに関する文章は全てまとめる.

3.2 固有表現抽出によるメタデータ付与

イベント情報を得るには記事文から1つのイベントに関する記述の範囲を特定し, イベントの状況に関する情報を見つけ出す必要がある. 固有表現抽出は, テキスト中から人名や日時などの表現を抽出することが出来るため, イベントの日時, 主体, 発生場所を特定することに適している.

日本語係り受け解析器 CaboCha には固有表現抽出を行う機能が実装され, IREX 定義に基づくメタデータタグが付与される. IREX 定義には日付を示す DATE タグ, 時間を示す TIME が存在し, 記事文から日時情報を抽出することが可能である. またイベント関連性も調べることを考慮し, 主体として利用するタグに ORGANIZATION, 発生場所に LOCATION を対応させる.

本研究では固有表現抽出によって抽出されるメタデータから, イベントの発生時間, イベントの主体, イベントの発生場所を推測する. 抽出する情報と対応する IREX タグを表 2 に示す.

表 2: 抽出する情報と対応する IREX タグ

イベント情報	対応する IREX タグ
イベント発生日	DATE
イベント発生時間	TIME
イベントの主体	ORGANIZATION
イベントの発生場所	LOCATION

3.3 時間情報の抽出

イベントの発生時間の前後関係を調べるにはイベント発生日時の抽出が重要となる. 固有表現抽出によって「○月○日」「○時○分」といった表現を抽出することが出来る. しかし, 「同日」という表現が DATE に対応していない, TIME に「朝」「午前」といった具体的な時間が不明な表現が抽出されるなど, 固有表現抽出だけではイベントの発生時間を特定することは困難であると考えられる. 表 3 は, CaboCha で解析した記事文章から発見して時間を表現する単語のメタデータタグ DATE, TIME を分類した一例をまとめたものである.

記事文中からイベントの発生時間を特定するために日付・時間表現の抽出ルールを決める必要がある. また時間の表現には「～から～まで」のような一定の期間を表す表現, 「～ごろ」「～前」のようなある時間付近ではあるが曖昧性を持つ表現が存在し, イベントの発生時間の比較を行う場合, これらの不明な時間の範囲や曖昧性を考慮したイベント発生時間の設定を行う必要がある. よって本研究では抽出した時間情報に不明な時間範囲や曖昧性が含まれる場合を考えた抽出ルールを提案する.

表 3: ニュース記事で表現される日時情報の例

	例
DATE もしくは TIME タグが付与	午前, 午後, 朝, 昼, 夕, 夜, 未明, 前日
メタデータタグの付与なし	同日, 昨日, ～後, ～前, ～過ぎ, ～から, ～まで

3.4 ニュース記事の特徴

ニュース記事におけるイベントの記述はイベントの発生日, 発生時間, 主体, 発生場所すべてを含むとき, 「主体」「発生日」「発生時間」「発生場所」という順番でイベント情報が提供され, 発生場所の後にイベントの内容に関する記述がされているという特徴を持つ. イベントの内容にあたる部分は複数の文で記述されていることもあり, このような場合, 文中に日時情報を含まず, 前の文を補足する形で記述されている. また, イベント内容に関する記述の中に別のイベントが含まれていることもあり, 「～について」「記者会見で～と発表」といった表現から, イベント中のイベントの存在を確認することができる.

4. イベント情報抽出ルールの提案

本研究では, ニュース記事の解析によって得たイベント情報から, 時間情報を比較しイベントの発生順に整理し直すことを目的としている. そこで, ニュース記事の解析から判明したニュース記事の特徴, 特に抽出された時間情報の特徴を考慮し, イベント情報の抽出ルールを提案する.

4.1 イベントの分割

1つの記事内に複数のイベントに関する記述が存在する場合, 記事の記述内容から個々のイベントを判断する必要がある. 記事内で別のイベントに関する記述が始まる場合, 日時, 主体, 場所を示すイベント情報がそれまでの記述と異なるため, これらが変更されたときにその文から前とは別のイベントの記述がされていると考えることができる. ニュース記事の文書解析は文ごとに行い, 新規にイベントを発見したかどうかを確認するために, イベント情報を最後に取得したイベントのものと比較し, 値が異なれば新たなイベントが取得されたとみなす.

発言内容などのイベントの中に含まれるイベントについても「発言した」というイベントと発言内容のイベントに分割させることを考える. 発言内容ならば「～と発言した」というような表現を取得した際に通常のイベントの取得と同様に発言された内容をイベントとして取得すればよい. しかし, 発言内容やイベント内のイベントと取れる表現が取得できない場合, ニュース記事の特徴である, 「主体」「発生日」「発生時間」「発生場所」という順番で記述された表現のうちいずれかが 1 文内で重複していないか調べ, 重複していたならば 2 つ目の「主体」「発生日」「発生時刻」「発生場所」の並び以降を新しいイベントとして扱うこととする.

4.2 イベント発生日時の決定

メタデータタグ DATE, TIME が付与された日時情報から, イベントの発生時間を推測する. 記事文の解析によって判明した時間情報の特徴から, イベント発生時間は発生日時と終了日時の 2 つに分け, イベントが発生したと考えられる範囲の時間を取る. 時間情報が期間を表すものでなく, 曖昧な表現でもなければ, 発生時間・終了時間は同じ値を入れる.

(1) 基本的な日時情報の格納

日時の値を取得する際，日付は「年」「月」「日」，時刻は「時」「分」「秒」に分割し，それぞれに値が入っているか確認する．値が入っていない場合は記事公開日に記されている値を代入させる．たとえば 2013 年 5 月 15 日公開の記事文から得た情報として「14 日」とある場合，14 を「日」の値に入れ，記事文中から抽出できなかった年，月の値は記事公開日の「2013 年 5 月」を割り当てる．ただし，「年」，「月」，「時」，「分」の値を代入させた場合はそれ以下の値は代入させないこととする．なお，時間情報について，時間の前に「午前」「午後」とついていた場合や他国との時差がある場合は 24 時間表記で日本時間に修正しておく．

(2) 相対的な日時情報

「同日」や「前日」など相対的な時間表現は，記事公開日時を比較したものと，記事内に書かれたひとつ前のイベントの発生時間と比較したものがある．あるイベント内で「同日」または「前日」というイベントとの相対的な時間表現を取得した場合，そのひとつ前のイベントの日付情報を取得し，相対時間を参考にイベント発生日時を調整する．また「明日」「昨日」といった記事の著者の視点から見た相対時間表現については，記事公開日を取得し，日時を調整する．表 4 は，相対的な時間表現取得時のイベント日時の割り当て方の対応の一例である．

表 4: 相対時間取得時のイベント日時の割り当て例

相対時間	対象	割り当て方
同日	ひとつ前に取得したイベント日時	対象のイベントと同じ日付を割り当て
前日	ひとつ前に取得したイベント日時	対象のイベントの 1 日前の日付を割り当て
翌日	ひとつ前に取得したイベント日時	対象のイベントの1日後の日付を割り当て
昨日	記事が公開された日付	記事が公開された日の1日前の日付を割り当て
明日	記事が公開された日付	記事が公開された日の1日後の日付を割り当て

(3) 期間を表す日時情報の範囲

時間を表す表現には，時刻だけではなく，「午前」や「朝」のような期間を表す表現がある．そのような期間発生時間，終了時間にそれぞれ別々の値を入力させる．表 5 は，期間を表す表現における，イベント発生時間と終了時間の対応表である．時間情報として格納する場合，開始時間は発生時間として扱う．

表 5: 期間を表す表現の開始，終了時間の範囲

表現	開始時間	終了時間
午前	0:00:00	12:00:00
午後	12:00:00	24:00:00
朝	6:00:00	12:00:00
昼	12:00:00	16:00:00
夕	16:00:00	18:00:00
夜	18:00:00	24:00:00
未明	0:00:00	6:00:00

(4) 曖昧な表現の範囲

時間情報には数字によって具体的に表現されたものに「～ごろ」「～前」のような数字に曖昧性を持たせてしまう表現を持つものが存在する．曖昧性を持った時間情報は，他のイベント情報の時間情報との比較を行う際，その情報がはっきりとしないために時系列の整理において正確な結果を出せない可能性があるため，その曖昧さの範囲をあらかじめ指定し，時間情報の比較を行いやすくする．図 2 は時間の曖昧部分の範囲を示したチャート図である．曖昧表現の前に記述された具体的な日時情報を時間 n とし，曖昧部分が n 以外の時間を示さないようにするため，「ごろ」「前」「過ぎ」といった表現はその時間±1(n と同単位)の時間範囲を取る．

図 2 ：時間の曖昧部分の範囲

4.3 イベントの主体・発生場所の決定

ニュース記事の特徴から，イベントの主体，発生場所は記事文中においてイベントの内容より先に表記されていると考えられる．このことから，一つのイベントの主体・発生場所はイベント情報を記述した文において初めに出現した主体・場所を示す言葉が該当すると考えられる．よって，解析されたイベントの記事文の最初に出現した ORGANIZATION タグの内容をイベントの主体に，LOCATION タグをイベントの発生場所とする．

5. 実験

提案した大規模ニュースにおける情報整理のためのイベント情報の抽出，特に時間情報の取得について，抽出精度を調べるために実験を行った．

5.1 実験用データの作成

本提案手法を評価するために、ニュース記事からイベント情報を抽出し、正解データと比較する．まず，読売新聞社のオンラインニュース提供サイトから，大規模な災害・事故についての報道として福島第一原子力発電所の事故に関する 10 記事を選び，それらを句点で分割した 83 文に分割した．これらすべてに CaboCha によって IREX 定義に基づくメタデータタグを付与し，記事が公開された日付と時間の情報としてメタデータ RELEASE を追加し，xml 形式のデータとして作成した．これを 3 章で提案したイベント情報の抽出ルールに基づき，メタデータタグが付与された情報を抽出し，表 6 の項目に分けて各文の情報をまとめた．1 文ごとに項目に情報を割り当てることとするが，値を抽出できなかった項目に関しては空のままにしておき，値の代入は行わないものとする．

表 6: 実験用データの各項目

項目名	内容
ID	文の ID
Start Date	イベントの発生日
Start Time	イベントの発生時刻
End Date	イベントの終了日
End Time	イベントの終了時刻
Release Date	記事が公開された日
Release Time	記事が公開された時刻
Main	文の主体
Location	イベント発生場所
Description	固有表現抽出前の文

また抽出した情報の適合率，再現率を調べるために，人手による正解データも作成する．

5.2 評価方法

抽出データ，正解データのイベント情報から，Start Date，Start Time，End Date，End Time，Main，Location の 6 項目を比較し，抽出データが正解データの値に一致すれば正解とし，各分から抽出されたイベント情報の各項目における適合率，再現率を算出する．正解データに値が存在し，抽出データに検出された値と一致した数を tp，正解データに値が存在するが，抽出データで検出された値と異なった数を fp，正解データの値が存在しないが抽出データで値が検出された数を fn とすると，適合率，再現率は次式(1)，(2)で表される

$$適合率: p = tp / (tp + fp) \qquad (1)$$

$$再現率: f = tp / (tp + fn) \qquad (2)$$

(i)提案手法を用いた場合と(ii)CaboCha の抽出のみでイベント情報を抽出した結果を比較する．なお(ii)CaboCha のみの抽出においても，4.2 で述べた相対時間・範囲時間の調整は行っている．表 7 は各項目の適合率，再現率を示したものである．また，各文の項目の正解数の分布を表 8 に示す．ただしこちらの正解には正解データ，抽出データ共に値が入ってない物も含まれている．

表 7: 評価実験の結果

	(i)		(ii)	
項目名	適合率	再現率	適合率	再現率
Start Date	0.8529	0.9354	0.5294	0.8636
Start Time	0.9259	0.9600	0.8620	0.9600
End Date	0.8235	0.9032	0.4075	0.8181
End Time	0.9259	0.9600	0.8333	0.9600
Main	0.2925	0.3939	0.2925	0.3939
Location	0.0888	0.1111	0.0888	0.1111

表 8: 各文の項目の正解数

	文数	
正解数	(i)	(ii)
0	1	2
1	11	1
2	11	8
3	21	4
4	26	58
5	11	8
6	2	0

6. 考察

実験の結果から，提案手法における適合率，再現率を CaboCha のみの抽出と比較する．提案手法では CaboCha で取得できなかったイベント発生日も抽出しているため，Start Date，End Date の適合率・再現率が CaboCha のみの場合よりも向上し，正解数の分布も偏りが少なくなっている．その他の項目ついては CaboCha と同じ値が抽出されたために適合率・再現率ともに変化が見られなかった．

また，今回の比較では抽出データ，正解データともに曖昧性のある表現を取得することができなかったため，曖昧性の表現を持つニュース記事データを使用して評価実験を行った場合の適合率，再現率を確認する必要がある．さらに今回の実験は文ごとの解析であったため，イベント分割により他の文から日時情報を得る必要があるイベントの日時情報で不正解となる項目が出現した．提案手法によるイベント分割ルールを適用し記事文全体を解析にかけた場合，他の文から情報を得ることができるため，適合率・再現率はさらに上がると考えられる．

一方，提案手法における各文の項目の正解数を見ると，Main，Location の抽出において，適合率，再現率が低かったことから 5 つ以上の項目を正解した文は少ないことがわかり，さらに項目が 6 つとも一致した文は 2 つしかなかった．これについては「福島第一原子力発電所」などのそのニュースでのみ頻出している語が CaboCha の辞書に登録されていないためにメタデータタグが付かなかったことと，Location にあるべき情報が，ORGANIZATION タグが付いていたため，Main の情報として扱われていることが原因であることがわかった．抽出できなかった語に関しては，抽出キーワードの設定を行う必要があることが考えられるが，Main，Location においては特に情報抽出ルールの定義が不足しているため，今後さらに細かな抽出ルールを設定していくことが必要である．

7. まとめ

本論文では，ニュース記事からのイベントの時間情報の取得について，固有表現抽出を利用した日時情報の抽出ルールを提案し，高い適合率，再現率を計測することができた．記事文中に時間に関係する表現が存在する場合，今回の提案手法における日時の判別によってイベントの発生日時を特定しやすいことがこの事から分かる．今回の評価実験では1イベントではなく1文ごとに記事文を解析にかけたため，日時情報を特定できない文も存在したが，記事ごとのイベント情報の解析においてもこの可能性は考えられるため，解析記事内で発見できない日時情報特定する必要がある．

イベントの主体とイベント発生場所の抽出においても今回の実験では定義が不十分であることに加え，イベント情報を調査するニュース内でのみ頻繁に使用される単語を抽出キーワードとして定義する必要が考えられる．そのためには，イベントごとの分割や抽出を行う前に記事文章を解析し，頻出語を特定することで適合率，再現率の向上を目指す．

イベント抽出の課題解決の後，イベント情報ごとの時間関係を比較・整理する手法を考案し，発生したイベントが時間順に正しく整理されているかを評価する．

参考文献

[関根 98] 関根 聡, 伊佐原 均:「IREX:情報検索, 情報抽出コンテスト」, 情報処理学会研究報告 第 127 回自然言語処理研究会報告, pp109-116, (1998).

[山田 02] 山田 寛康, 工藤 拓, 松本裕治:「Support Vector Machine を用いた日本語固有表現抽出」, 情報処理学会論文誌, vol43 , no1, pp44-53, (2002).

[斉藤 13] 斉藤 悠, 佐藤 真, 赤石 美奈:「文要素解析と固有表現抽出によるメタデータ抽出」, 情報処理学会第 75 回全国大会, no.2, pp125-126,(2013).

[工藤 02 年] 工藤 拓, 松本裕治:「チャンキングの段階適用による日本語係り受け解析」, 情報処理学会論文誌, vol43, no.6 , pp1834-1842, (2002).