新聞記事群からのイベント抽出と時間情報に基づく編成

Inoue, kota / 井上, 晃太

(出版者 / Publisher)
法政大学大学院情報科学研究科
(雑誌名 / Journal or Publication Title)
法政大学大学院紀要. 情報科学研究科編 / 法政大学大学院紀要. 情報科学研究科編
(巻 / Volume)
9
(開始ページ / Start Page)
77
(終了ページ / End Page)
82
(発行年 / Year)
2014-03
(URL)
https://doi.org/10.15002/00010525

新聞記事群からのイベント抽出と時間情報に基づく編成
Organizing Events Extracted from Multiple Document by Time

井上　晃太
Kota Inoue
法政大学大学院情報科学研究科情報科学専攻
E-mail: kota.inoue.6h@stu.hosei.ac.jp

Abstract

When the news such as a large-scale or long-term case, accident or disaster is reported, they take a long time to arrange events which happened in these cases by human because they have huge data. This paper suggests a method for quickly picking up information about events including when and what accidents happened and arranging them when the case occurred. To pick up information from news articles, we use named entity extraction and an original rule which helps picking up particular extraction. Such a rule can extract more information from news articles than using only named entity extraction. In addition, as well as comparison of the time when events happened, events are united which report same or similar events by sentence from news articles for users of this system to check easily when the flow of events is visualized. To confirm the precision of the method, this paper checked how results reproduced real events and what informations are changed by addition of new events from collected news.

1. 序論

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．例として，新聞記事やニュースサイトで掲載される報道記事を収集し，記述内容を要約することで「いつ」「何が起きたか」ということを知ることが出来るが，規模の大きいニュースほど全体の把握のために大量の記事を収集する必要がある．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収集し，時間情報を記事中から見つけ，それを基に情報を整理する必要がある．しかし人の手によって大量のニュース記事から情報をまとめる作業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る．

本研究ではインターネット上のニュースサイトからニュース記事を収集し，文中から必要な情報を取得し，取得した情報から事件・事故・災害内で発生した詳細な出来事を比較することでそれらの発生順序を整理することをコンピュータ上で行う．本研究では事件・事故・災害のニュースから，「いつ」に着目し，その時「何が起きたか」という形で出来事をまとめたものをイベントと呼び，そのイベントが発生した日付・時刻などの要素をイベント情報とする．

ニュース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[1]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を抽出する研究が行われている．文書解析器 CaboCha[2]には IREX 定義に基づく固有表現抽出機能が実装されており，それを利用したニュース記事からの情報抽出の研究[3]も存在する．しかし，この方法では文中に記された固有表現を抽出できるが，「同日」や「夜」といった相対的な表現や広い時間の範囲を示すもの等，具体的な時間を与えていないため，システム上で具体的な時間を与える必要がある．そこで本研究では取得できない表現を補う形でイベント情報の抽出方法を考えた．

また，複数の報道機関から一定の期間内に出されたニュース記事を収集することから，同イベントについてのニュースや関連性の高いニュースを結びつける必要がある．本研究では収集し整理した情報をシステムの利用者が理解しやすくするために，情報を簡潔にまとめる必要性がある．そこで，ニュース記事内容をイベント別に分割し，同イベントに関する記事文章をイベントに関する単語からクラスタリングすることを試みた．

本論文では，収集したニュース記事を整理し，イベントの流れを可視化するためのシステム形成のために，イベント情報の抽出方法とイベントの整理方法を提案し，人力による文章整理と比較することで提案手法の性能について調査した結果について述べる．

2. 概要

図 1 は本研究におけるイベントの整理における流れを表したものである．本研究では収集したある事件・事故もしくは災害に関する複数のニュース記事を入力し，記事ごとの内容をイベント別にまとめた後，同じイベントの重複を避けるため，すべての記事から収集したイベントの中から類似する物をクラスタリングした後，ニュース内におけるイベントの流れを視覚的に把握しやすくするためにガントチャートによって整理した結果を示すこととする．記事からの情報抽出の際，記事の内容を時間情報の有無を基にイベントごとに分割するために一度記事を文単位に分割してからイベント情報を抽出し，イベントごとにまとめ直す．イベントの抽出方法は固有表現抽出を利用した抽出ルールを考案し，イベントの整理をしやすくしている．イベントの整理では，イベントに含まれる単語の種類から同じイベントを判別するために形

Supervisor: Prof. Mina Akaishi

態素解析を用いる．イベント情報の抽出およびイベントの整理の詳細は後述する．

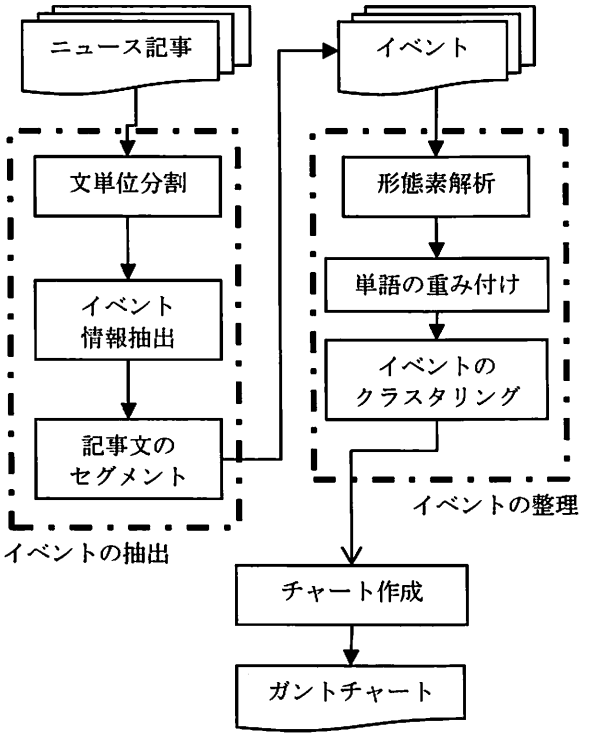

図 1　ニュース記事からのイベント情報整理の流れ

3.イベント情報の抽出

3.1. イベントとイベント情報

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
ニュース記事の公開日時	ニュース記事の公開日時
内容	ニュース記事本文

本研究では事件・事故・災害などのニュース内で発生した「いつ」「どのようなこと起きたか」という情報をイベントとして扱う．イベントには発生した日付・時間等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記事本文を用いる．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する前の文そのものを使用し，複数のイベントが記事内に存在していた場合は，該当するイベントに合わせて文章が分けられている．

3.2. イベントのセグメント

イベント情報を得るには記事文からイベントに関する記述の範囲を特定し，複数のイベントが記事内に含まれていることを想定して，イベントの状況に関する情報を探す必要がある．そのために，ニュース記事をイベント別にセグメントさせる．本研究では大量のニュース記事を素早く整理することと，「いつ」「何が起きたか」という情報をイベントとして取得することを目標としているため，関連語を用いたセグメントではなく，文単位分割後に抽出したイベント情報，特に文中における日付や時間の有無によってセグメントを行う．

ニュース記事文の構成を調べたところ，ある出来事について報道する際，最初に出来事の趣旨について述べ，そこに日時情報が含まれている．そして次の文以降でその出来事についての詳細な説明が記述されている傾向が見られた．このことから，本研究ではニュース記事文を文単位に分割してイベント情報を抽出した後，日時情報の有無によって文をセグメントすることによってイベントごとに文章を分けることとする．イベントのセグメントのフローを図 2 に示す．図 2 では記事内で文単位に分割された文を $S_i(i=1,2,\ldots)$，イベントを $E_j(j=0,1,2,\ldots)$として扱う．イベントは文内に日時情報を発見した場合に新規に作られ，日時情報が文 S 内になければその文は既存のイベントに追加されることになる．これを記事ごとに文単位分割された文がなくなるまで繰り返す．

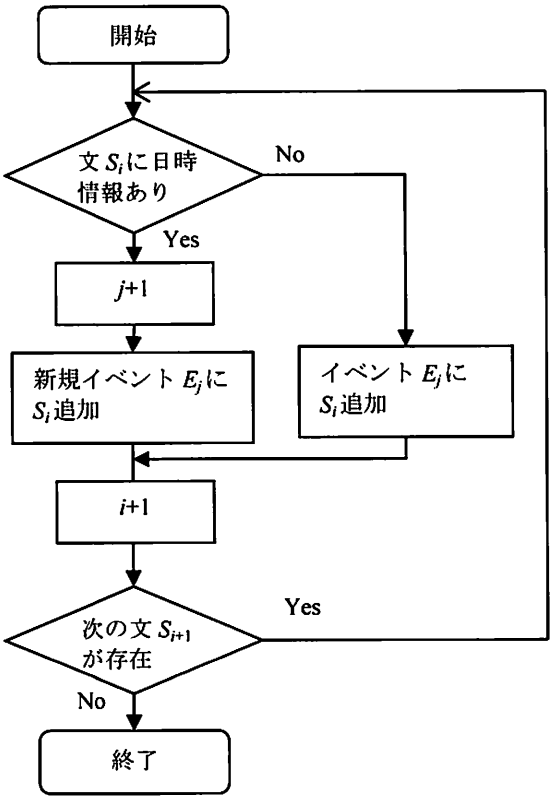

図 2　記事ごとのイベントのセグメントのフロー

3.3. 固有表現抽出を用いたイベント情報抽出

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できるため，イベントの日時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ており，IREX 定義に基づくメタデータタグが該当する表現に付与される．IREX 定義には日付を示す DATE タグ，時刻または時間を示す TIME タグが存在し，記事文から日時情報を抽出することが可能である．

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，イベントの日時を推測する．CaboCha では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できるが，「同日」という表現が DATE に対応していない，TIME に「朝」「午前」といった時間帯は分かるが具体的な時刻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がいつ起きたかを特定することは困難であると考えられる．表 2 は CaboCha で解析した記事文章から発見した日付・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日時を特定するためには日付・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，「～後」「～前」のようなある時間付近を示すが曖昧性を持つ表現が存在し，イベントの日時の比較を行う場合，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 2　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

抽出する日時情報と対応する IREX タグを表 3 に示す．メタデータタグ DATE，TIME が付与された日時情報から，イベントの日時を推測する際，表 2 における日時情報の特徴から，取得できるイベントの中にはある時点で発生し何日，あるいは同日内のある時刻まで続いていたものが出ることが考えられる．そのようなイベントの取得を想定し，イベントの日時情報には発生日時と終了日時を用意し，イベントが発生したと考えられる範囲の時間を取る．日時情報が期間を表すものでなく，あいまいな表現でもなければ，発生日時・終了日時は同じ値を入れる．

表 3　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
イベント終了日	
イベント発生時刻	TIME
イベント終了時刻	

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．時間について，「午前」「午後」の表記があれば 24 時間表記に修正しておく．

「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日時とを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	割り当てる日付
同日	1 つ前に取得したイベント日時	対象のイベントと同じ
前日	1 つ前に取得したイベント日時	対象のイベントの 1 日前
翌日	1 つ前に取得したイベント日時	対象のイベントの 1 日後
昨日	記事公開日	記事公開日の 1 日前
明日	記事公開日	記事公開日の 1 日後

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．そのような表現には期間の発生時刻，終了時刻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，イベント発生日時と終了日時の対応表である．時間情報として格納する場合，開始時刻は発生時刻として扱う．

表 5　期間を表す表現の開始，終了時刻の範囲

表現	開始時刻	終了時刻
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

3.4. イベント情報の抽出実験

提案手法におけるイベント情報の抽出，特に時間情報の取得精度を評価するために実際の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．この実験では読売新聞社のオンラインニュース提供サイト[5]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 文を用意した．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．これを提案手法における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とそれ以外に必要な情報を抽出し，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．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．

表6　実験用データの各項目

項目名	内容
ID	文またはイベントの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
Article	イベントに含まれる記事文

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の4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，各文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．

$$\text{適合率} = \frac{tp}{tp + fp} \quad (1)$$

$$\text{再現率} = \frac{tp}{tp + fn} \quad (2)$$

3.5. イベント情報の抽出結果

(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)においても，相対時間と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．

表7　イベント情報の抽出精度評価の結果

	(i)		(ii)	
項目名	適合率	再現率	適合率	再現率
Start Date	0.85	0.94	0.53	0.86
Start Time	0.93	0.96	0.86	0.96
End Date	0.82	0.90	0.41	0.82
End Time	0.93	0.96	0.83	0.96

(ii)と比較して，(i)の結果のほうが適合率，再現率が高く，特にStart Date，End Dateにおいては適合率が大きく上昇し，再現率も上昇した．一方でStart Time，End Timeは適合率の上昇も小さく，再現率にも変化は見られなかった．

4. イベントの整理

抽出したイベント情報を基にイベントを比較し，イベントの流れを特定する．整理したイベントをガントチャートによって可視化した際にユーザーが分かりやすくするために，同じ事柄を示すイベントはクラスタリングし，抽出したイベントの日時からイベントの前後関係を整理する．

4.1. イベントのクラスタリング

イベントの時間関係の整理を行う前に，記事ごとにイベント別に分割したものから，別々の記事に含まれる同じイベント同士をまとめるためのクラスタリングを行う．同じイベントに関する記述は文中に含まれる単語の種類も類似することから，類似性を利用して同じイベントに関する記述であるかを判断する．

イベント情報にはイベント内容にニュース記事の文章が含まれているため，これを形態素解析することによって単語を抜き出し，他のイベントとの単語の種類や出現頻度の類似性を調べる．本研究では形態素解析には形態素解析器MeCab[4]を使用しMeCabに含まれる辞書の品詞分類から名詞を抽出した．このとき，イベントの総数を m，全イベントに k 種類の単語が含まれているとする．イベント E_m の単語ベクトルを e_m，e_m に対する単語 w_k の重みを $n(w_k, E_m)$ とすると，

$$e_m = \left(n(w_1, E_m), n(w_2, E_m), \ldots, n(w_k, E_m)\right) \quad (3)$$

となる．単語の重み n はtf-idf法を用いて以下の式で計算する．

$$\mathrm{tf}(w_k, E_m) = \frac{E_m \text{における} w_k \text{の出現数}}{E_m \text{における総単語数}} \quad (4)$$

$$\mathrm{idf}(w_k) = \frac{m}{w_k \text{が出現するイベント数}} \quad (5)$$

$$n(w_k, E_m) = \mathrm{tf} \cdot \mathrm{idf}(w_k, E_m) = \mathrm{tf}(w_k, E_m) \times \mathrm{idf}(w_k) \quad (6)$$

これらイベントの単語ベクトルに含まれる各単語の重みを用いて，イベント同士の類似度をコサイン類似度で計算する．類似度計算を行う単語ベクトルを e_x，$e_y (x, y \in m)$ とすると，次式(5)で表される．

$$\mathrm{sim}(e_x, e_y) = \cos\theta = \frac{e_x \cdot e_y}{|e_x||e_y|} \quad (7)$$

この類似度計算の結果から，一定の数値を超えたイベント同士は類似イベントとして分類する．なお，クラスタリングが行われた後，そのイベントのクラスと他の未分類のイベントとの類似性を調べる場合には，クラスタリングされたイベントの中から記事公開日時が最新のものを参照する．

4.2. イベントの整理と可視化

ニュース記事ごとに記事文から日時情報を含むイベントに分け，他の記事の類似イベントとクラスタリングす

ることによってイベントをまとめたデータを整理し，どの時間でどのようなイベントが起きたか可視化させる．

イベントの流れの可視化は図 3 のようにガントチャートで示すこととする．イベント情報抽出の際，イベントの発生日と発生時間は開始日，開始時間，終了日，終了時間という形で取得しているため，これらをガントチャート上でのイベントの期間をグラフで表す際の始点，終点として扱う．またイベント発生時間が期間ではなく，ある時点で発生したものについてはマイルストーンで示す．イベントは記事公開順で並べてあるが，図 3 のように イベントの発生順は記事公開順と異なる場合があるため，イベントの流れを追うには記事公開順とは別に追う必要がある．これを用いることによって，イベントの発生順序でイベントの内容を読むことが出来る．また，最新のイベントから，過去にさかのぼって記事を読むことができ，ユーザーは事件の流れを理解しやすくなる．

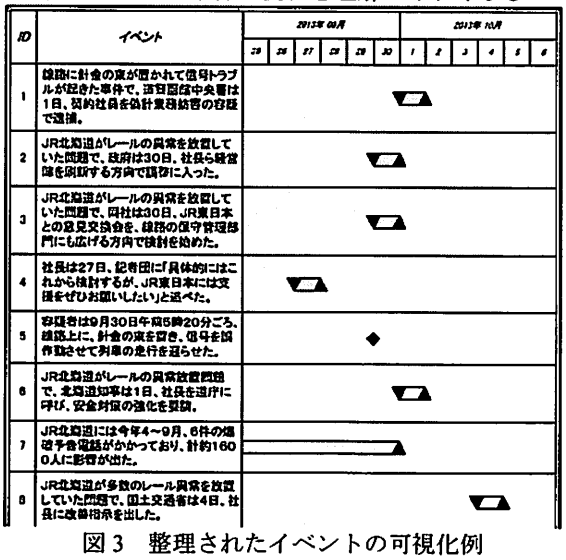

図 3　整理されたイベントの可視化例

4.3. イベントの整理実験

イベント情報の抽出後，そのイベントを提案手法による情報整理の精度についても評価実験を行った．この実験では，複数の報道機関から同じ大規模ニュースに関する情報を収集することで，同じイベントに対する情報の違い等を考察することを考え，yahoo ニュース[6]から JR 北海道の不祥事に関するニュースを収集し，イベント情報の抽出実験と同様にして抽出データを作成後，イベント整理実験を行った．また，提案手法によるイベントの発生時間の抽出，イベントのセグメントとクラスタリングの精度を評価するために，提案手法によって整理された結果と，人手によるイベント整理の結果とを比較して行った．この評価実験では 10 件のニュース記事を使用し，5 人の実験協力者に以下の方法でイベントの整理を行ってもらった．

① 1 つの記事ごとに，異なる出来事について述べていると思われる箇所で記事文を分割し，それぞれをイベントとする．

② ①で分割したイベントから表 6 に当てはまる情報を抜き出し，該当項目に記入する．

③ 全ての記事で①②の作業を終えた後，整理されたイベントのうち，同じことについて述べていると思われるイベントをクラスタリングする．

これによって抽出されたイベントと提案手法を比較し，記事別のイベント抽出及び類似イベントのクラスタリング後の結果を比較した．

4.4. 記事別のイベント取得の結果

表 8 は提案手法が記事ごとにイベントを取得した際のイベント分割箇所の実験協力者の結果との一致度である．イベントの分割箇所が提案手法のものと一致した箇所の傾向としては，イベント内容の最初の文に必ず日時情報が含まれているということである．

表 8　提案手法のイベント分割箇所の協力者との一致度

記事	分割箇所	実験協力者の一致度(%)
I	I -1	80
	I -2	60
II	II -1	20
III	III -1	75
	III -2	80
IV	IV -1	60
	IV -2	60
V	V -1	40
VI		
VII	VII -1	80
	VII -2	80
VIII		
IX	IX -1	80
	IX -2	80
X	X -1	100
	X -2	60

表 9　協力者のイベント分割箇所の提案手法との一致度

単位(%)	実験協力者				
記事	A	B	C	D	E
I	100	67	60	25	100
II	67	33	20	20	20
III	100	100	75	75	75
IV	50	100	60	40	50
V	100	100	67	40	20
VI	100	100	20	33	33
VII	100	100	100	50	75
VIII	100	100	20	50	33
IX	100	100	100	100	100
X	100	67	75	60	75

表 9 は実験協力者のイベント分割箇所の提案手法との一致度をまとめたものである．一方では 1 つのイベントとして扱われるものがもう一方の結果では複数に分かれていた場合でも，日時情報が含まれているイベントでは，その日時情報が含まれる文がイベント内容の最初に来ているため，提案手法と分割箇所の一致が多かった．このことから，記事ごとからのイベントのセグメントでは，

文中に日時を表す表現が含まれているイベントの分割は人手による作業に近い結果を出せていると考えられる．

4.5. イベントのクラスタリングの結果

次に，全ての記事から取得したイベントから類似イベントをクラスタリングした結果について比較した．この実験での提案手法における他記事内の類似記事とのクラスタリングは，文書類似率が30%を超えた場合とした．表10は最終的に整理されたイベント数である．提案手法が最終的にまとめたイベント数は14であり，協力者A，C，Dの結果に近い数値となった．

提案手法と協力者によって整理された結果を比較してみると，最終的なイベント数が近いA，C，Dの結果との比較では，複数の記事に記述されていた「JR北海道の経営陣の刷新」「契約社員による妨害行為」「国土交通省からのJR北海道社長の呼び出し」といったイベントがまとめられており，提案手法のクラスタリング結果と似た複数のイベントのまとめ方をしていた．しかし経営陣の刷新に関する話題に関して，提案手法と実験協力者の結果では経営陣刷新の前に行われた国土交通省による特別保安監査という扱いが異なっていた．これに関しては記事別のイベントのセグメントで特別保安監査のイベントとして取得された文の中に，「経営陣の刷新を検討する」という記述が含まれていたことが原因であると考えられる．これにより，イベントの類似度を計算する際に同じイベントとして提案手法ではクラスタリングしてしまったと考えられる．また，提案手法では他のイベントに含むことができると考えられる短い文が個別のイベントとして認識され，クラスタリングされない例もあった．これらのことから，イベント内容の単語を用いたイベントの類似度計算だけでは，全ての記事からのイベントのクラスタリングに問題があると考えられる．

提案手法と比べてイベント数が少ないBはレールの異常を放置していた事に関する「経営陣の刷新」「国土交通省による社長の呼び出し」といった別々のイベントとして扱えるものを全て同じイベントとしてクラスタリングしてしまっていため，他の結果と比べて極端にイベント数が少なくなってしまい，提案手法との比較が困難になってしまった．逆にイベント数が提案手法より多いEの結果からは「いつ」「何をした」という形に加えて「誰が（または何が）」という要素でイベントがまとめられていたため，提案手法よりイベント数が増加していたが，経営陣の刷新や契約社員の業務妨害など，一部の類似イベントのクラスタリングに提案手法と似た結果が見られた．

表10　実験で得た最終的なイベント数

	提案手法	協力者				
		A	B	C	D	E
イベント数	14	12	5	11	13	27

5. 考察

本研究の実験結果について，イベント情報の抽出，記事ごとのイベントのセグメント，イベント整理の3点について考察する．

本研究におけるニュース記事からのイベント情報抽出方法をCaboChaのみによるイベント情報抽出と比較したところ，表7よりStart Date，End Dateの適合率・再現率がCaboChaのみの場合よりも向上したことから，提案手法では記事文からのイベント情報，特にイベント発生日が特に取得しやすくなったと考えられる．

ニュース記事ごとのイベントのセグメントに関しては，日時を表す表現が含まれている文によってイベントを分割する方法を行ったところ，人の手による記事ごとのイベントのセグメントでも同様のイベントの区切り方が見られた．しかし人手の作業では日時情報が含まれない文でもイベントの分割が行われており，人によってその区切りも異なっていた．日時情報に頼らないイベントのセグメントのためには，人間のニュース記事を読んだ際の主題の認識の変化についても調べる必要があると考えられる．

イベントの整理については今回の実験の結果から，単語の頻度を利用した類似度計算によるイベントのクラスタリングだけでは問題があり，より精密なイベントの整理を行うには整理するための方法を改善する必要がある．人手によるイベントの整理と比較して今回の結果から考えられることは，イベントに含まれる日時情報の違いによって，イベント内容の文書類似度の高いイベント同士が異なるイベントとして扱うことが出来ることである．また，実験協力者Eの結果から見られた，イベント内容の主語やイベントの主体によるイベントの類似性の判定も，より精密なイベントの整理が可能と考えられる．

6. まとめ

本研究では，事件・事故・災害に関する情報を，固有表現抽出を利用した手法で抽出し，時間情報を軸にニュースをイベントとして整理する方法を提案した．実際に手法に従った整理を行った結果から収集したニュースから出来事を把握することを試み，人の手によるイベントの整理結果を比較した．時間情報の有無によるイベントの分割によって，ある程度は人の手による作業結果に近い形となったが，イベント内容の類似性の判断には，他の要約手法を利用することで更なるイベント整理の精度向上を見込める可能性も出てきたと考えられる．

文　献

[1] 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.

[2] 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，no.1，pp44-53，2002.

[3] 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.

[4] Kudo Taku, Yamamoto Kaoru, Matsumoto Yuji, “Applying Conditional Random Fields to Japanese Morphological Analysis”, *EMNLP*, vol4, pp. 230-237, 2004.

[5] 読売オンラインニュース, http://www.yomiuri.co.jp/

[6] Yahooニュース, http://news.yahoo.co.jp/