社団法人　情報処理学会　研究報告
IPSJ SIG Technical Report

2005－NL－167（19）
2005－SLP－56（19）
2005／5／27

# 新聞記事中の事故・事件名の自動抽出

野畑　　周†　　佐田いち子†　　井佐原　均††

† シャープ株式会社
†† 独立行政法人 情報通信研究機構

**あらまし**　ある特定の出来事について知ろうとして新聞記事などを読むとき、その出来事を示す表現は何らかの形でその記事の中に現われている。しかし、その表現の文字列は一意でないことが多い。文章中の人名や組織名などの表現は、現われる文章に依らず固定していることが多く、それらの表現を自動的に取り出す固有表現抽出システムの精度は近年の研究によって高まっている。それを利用して自由度のより高い出来事を示す表現を汎用的な手法で自動的に抽出することは、情報抽出のための固有表現抽出としては拡張の方向性の一つであり、また自動要約や機械翻訳などの分野においても、文書間の話題のつながりを捉えたり、二言語間で対応する表現の範囲を広げたりする点で有用である。本論文では、特定の出来事を指す表現のうち、「事件・事故名」を対象として、その抽出方法の提案と評価を行う。

**キーワード**　情報抽出、固有表現、編集距離

# Automatic Extraction of Incident Expressions on Newspaper Articles

Chikashi NOBATA†, Ichiko SATA†, and Hitoshi ISAHARA††

† Sharp Corporation
†† National Institute of Information and Communications Technology

**Abstract**　When we read newspaper articles to obtain knowledge about a specific event, some expressions that denote the event appear in each article, but these expressions are more flexible and elusive than named entities like person names, organization names. Since the performance of a named entity recognizer has recently become better, it is one of the next steps to use recognized named entities for recognizing event expressions. The recognition of event expressions is also useful in detection of the same topic between multiple documents for automatic summarization, and between different languages for machine translation. In this paper, we present a method and evaluation results of extraction of specific incident names as a part of event expressions.

**Key words**　Information Extraction, Named Entity, Edit Distance

## 1. はじめに

報道記事で書かれる対象は主に出来事であり、出来事を示す表現は、各記事が伝える内容の主題となる表現である。新聞記事などの文章データから必要な情報を取り出す情報抽出の分野においても、最終的な目標はある種類の出来事を抽出することにおかれていた [2]。その発展として、特定の出来事に限定するのでなく、汎用的に情報抽出を行おうという研究も行われている。

情報抽出の部分タスクとして定義された固有表現抽出 (Named Entity Recognition) は、出来事を示す構成要素である固有名詞や数値表現等を認識するタスクである [1]。固有表現抽出は独立したタスクとして発展し、また抽出の対象を拡張する研究も行われてきている [6]。

自動要約の分野では、いくつかの新聞記事を一つの要約にまとめる複数文書要約が盛んに研究されている。複数文書要約においても、特定の個人や組織に関する要約と同様に、特定の出来事に関する要約が課題として設定されている [3], [4]。個々の新聞記事を要約する場合と比べて、複数の記事を要約する場合には記事の間で重複する情報が多くなり、これらの情報を適切にまとめる必要がある。同じ出来事を示す表現を適切に認識できれば、ある出来事に関する記事が集められたときに、個々の記事に現れるこのような、同じ出来事を示す表現を認識して一つにまとめることで要約文の冗長性を減少させることができる。

ある出来事を示す表現を自動的に認識する際に問題となるのは、同じ出来事を示す個々の表現の違いが大きいことである。新聞記事では、同じ出来事に対し報道機関や報道過程によってそれを示す表現が異なる場合がある。例えば、ある記事では「○○で何日に■■が△△した」と表現されるが、それ以降の記事では「○○で何日に起きた△△事件」や「○○における事件」と表現されたりするなどの自由度がある。重大な事件として扱われると、さらに「○○事件」「■■事件」といった表現になり、固有名詞化していくものもある。このように、出来事を示す表現を認識することは、固有名詞から句や節の範囲までの広がりをもつ表現を認識することになる。我々はこれまで、見出しに出現する出来事に関連する表現を分類した [10]。対象は見出しに限定しているが、出来事表現全般について、同じ出来事を示す異なる表現間での差異について分析している。

雨宮 [7] は、出来事を示す表現のうち、特に新聞社会面記事における特定の事件を示す表現について、主に構文構造に注目し、

- 連体修飾部分の種類 (節、句、語のいずれか、あるいは存在しない)
- 「事件」の前に複合する要素の有無 (末尾が「汚職事件」などとなっているか単に「事件」か)

の 2 つの観点から分類を行っている。また、表現中の構成要素として「犯罪行為」「舞台」「被害者」「加害者」「道具」「対象物」「日時」「修飾的要素」「関連事項」の 9 つを設定し、分類ごとのこれらの現れ方の違いを分析している。設定された構成要素のうち「舞台」「被害者」「加害者」「道具」「対象物」「日時」は固有表現抽出の対象クラスと重なっている。

記事の検索を行う際には、出来事の表現を捉えなくても、その表現の構成要素を列挙することで記事の特徴を示し、それらの要素に検索手法を適用することで検索の目的は達成される。しかし、その次の過程として情報抽出や自動要約を行う際には、精度を向上させるために、集められた記事の主題となる出来事を示す表現をより限定された形で認識することが有用であると考えられる。

出来事表現を認識する研究の最終的な目標としては、汎用的に出来事の表現を抽出し、それらを関係づけることである。その過程として本研究では、出来事を示す表現のうち「事件」と「事故」を主辞(注1)に取る表現に限定して、それらの表現の自動抽出を試みた。また、抽出できた表現の間で、同一の出来事と思われる表現をまとめるタスクについても実験を行った。実際に文章に現われる表現としては、事故や事件を示す表現でも主辞にそれらの語が現われない場合や、完全な文によって出来事が表現される場合があるが、これらについては今回対象としていない。

以下、パタンによる抽出手法・編集距離によるグループ化手法それぞれについて説明し、手法を適用

(注1)：ここでいう主辞は、名詞句の中で最も末尾に近い内容語を意味している。

した実験結果とそれに対する考察を述べる。

## 2. 事故・事件名の認識手法

### 2.1 事故・事件名の抽出

抽出手法としては、単語数に制限を設けずに任意の長さの表現を得るために、パタンベースで行っている。パタンは人手で作成したものである。パタンの記述は形態素ベースで、パタンの個々の要素は字面、文字種、品詞、辞書情報、固有表現クラスを用いて記述している。形態素解析には juman4.0 [9] を用いた。

特定の出来事を示す表現を抽出するパタンの作成方針として、ここでは固有表現をその中に含むことを前提とした。具体的には、「事件」「事故」の主辞と組織名、地名、人名などの固有表現とをパタンの主要な構成要素にしている。

個々の形態素に対するタグ付けは、BIO 記法を用いて行っている。即ち、固有表現の開始要素を B、中間・終端要素を I で表している。例えば、「茨城県東海村で起きた臨界事故」という表現は、

```
(NE:B-LOCATION) (NE:I-LOCATION)* (で)
(起きた) (POS:接頭辞|形容詞|名詞)* (事故|事件)
→ 1=B-INCIDENT 2〜6=I-INCIDENT
```

といったパタンでタグ付けされる。

パタン作成時には、毎日新聞 96〜99 年版の新聞記事から取り出した事故・事件名のうち、一定の頻度以上のものを参照した。具体的には、事件名については頻度 19 以上の表現 94 個、事故名については頻度 9 以上の表現 49 個を認識できるパタンを作成した。

今回、事故・事件名の抽出に必要な固有表現クラスは同様にパタンベースの固有表現抽出システムの出力を利用した。事故・事件名のパタン記述に用いた固有表現クラスは、人名 (PERSON)、地名 (LOCATION)、組織名 (ORGANIZATION)、施設名 (FACILITY)、製品名 (PRODUCT)、日付表現 (DATE) の 6 種類。

### 2.2 事故・事件名の関連付け

事故・事件名が与えられた後、同じ出来事を表すかどうかを判定する手法について述べる。ここでは、二つの表現が与えられたときに、それらがどの程度類似しているかを指標で示し、その指標の値が一定値以上になるものを同一とみなす手法でどこまでできるかを調べた。表現によっては、同じ出来事を表す表現が一つしかないものがあるが、それらも実験時には対象として含めている。従って、そのような表現については他のどの表現とも関連付けられてはならないことになる。

指標としては、Bag-of-Words と編集距離を用いた。編集距離は、さらに文字単位と単語単位の 2 種類で比較した。類似度を求める際には、主辞の「事故」「事件」は除いてある。それぞれの値は、値域が [0,1] で値が大きいほど類似していることを示すように、長い方の表現の文字数または単語数で正規化している。Bag-of-Words(BOW)、文字単位の編集距離 (EDC)、単語単位の編集距離 (EDW) による類似度を式で表わすとそれぞれ以下の通り。

$$S_{\mathrm{BOW}}(E_1, E_2) = \frac{\mathrm{BOW}}{\max(|E_1|_w, |E_2|_w)}$$

$$S_{\mathrm{EDC}}(E_1, E_2) = \frac{\max(|E_1|_c, |E_2|_c) - \mathrm{EDC}}{\max(|E_1|_c, |E_2|_c)}$$

$$S_{\mathrm{EDW}}(E_1, E_2) = \frac{\max(|E_1|_w, |E_2|_w) - \mathrm{EDW}}{\max(|E_1|_w, |E_2|_w)}$$

例えば、$E_1$:「茨城/県/東海/村/で/起きた/臨界/事故」と $E_2$:「ＪＣＯ/東海/事業所/の/臨界/事故」のペアについて各々の指標を計算すると、「事故」を除いたそれぞれの単語長が $|E_1|_w = 7$, $|E_2|_w = 5$、文字長が $|E_1|_c = 12$、$|E_2|_c = 11$ であり、両者に共通する語が「東海」「臨界」の 2 語、文字単位の編集距離が 8、単語単位の編集距離が 5 なので、両者の類似度はそれぞれ

$$S_{\mathrm{BOW}}(E_1, E_2) = \frac{2}{\max(7,5)} = 0.286$$

$$S_{\mathrm{EDC}}(E_1, E_2) = \frac{\max(12,11) - 8}{\max(12,11)} = 0.333$$

$$S_{\mathrm{EDW}}(E_1, E_2) = \frac{\max(7,5) - 5}{\max(7,5)} = 0.286$$

と計算される。

## 3. 実　　験

毎日新聞 95 年版の新聞記事を対象として作成された、拡張固有表現タグ付きデータ [5] において事故・事件名としてタグ付けされた表現をテストデータとし、評価実験を行った。

表 1　事故・事件名の抽出精度

| データ | トレーニング | テスト |
|---|---|---|
| 総表現数 | 23058 | 833 |
| 異なり数 | 143 | 274 |
| 再現率 | 89.36 | 76.83 |
| 適合率 | 80.68 | 81.63 |
| F 値 | 84.08 | 79.16 |

### 3.1　抽出の精度

トレーニングデータ (96〜99 年)、テストデータ (95 年) それぞれに対して、事件・事故名抽出パタンを適用した場合の精度を調べた。トレーニングデータはパタン作成時に参照した表現が現われている文を取り出して用いている。

各々のデータでの抽出精度を表 1 に示す。評価の方法は IREX ワークショップ [8] のものに沿っており、表現の範囲が全く同一でない場合は誤りと判定している。従って、正解である表現がシステムが抽出した表現に完全に含まれている場合、あるいはその逆の場合も不正解となり、それぞれ正解の抽出ミス 1 回と冗長な抽出 1 回としてカウントされる。

### 3.2　関連付けの精度

関連付けを判定する際には、まず任意の二つの表現に対して類似度を求め、得られた類似度を予め与えたしきい値と比較し、そのしきい値より大きい値をもつ表現のペアは同じ出来事を示すと判定している。

ここでは、関連付け単体の精度を示すため、システムが抽出した表現ではなくトレーニングデータ・テストデータ双方の事故・事件名を対象として評価を行っている。類似度に対するしきい値を変化させたときの各類似度による関連付け結果の再現率-適合率曲線を、トレーニングデータ (図 1) とテストデータ (図 2) それぞれについて示す。

トレーニングデータでは、文字単位の編集距離に基く類似度が他の類似度よりも良い。テストデータでは、トレーニングデータと比較して類似度の差は小さいが、編集距離に基く類似度の方が、Bag-of-Words に基づく類似度よりも上回っている。

次に、双方のデータにおけるしきい値の変化による F 値の変化をトレーニングデータ (図 3) とテストデータ (図 4) それぞれについて示す。

表 2 に示すように、トレーニングデータとテストデータでのしきい値の変化を比較するとトレーニングデータの方が、最も高い F 値を与えるしきい値が低い。この結果は、トレーニングデータ中の事故・事件名の方がより表現同士の差が大きいことを示していると考えられる。

表 2　各類似度の最適値を与えるしきい値

| トレーニングデータ | | |
|---|---|---|
| 類似度 | しきい値 | F 値 |
| BOW | 0.28 | 0.616 |
| EDC | 0.24 | 0.690 |
| EDW | 0.27 | 0.619 |
| テストデータ | | |
| 類似度 | しきい値 | F 値 |
| BOW | 0.35 | 0.659 |
| EDC | 0.39 | 0.677 |
| EDW | 0.27 | 0.704 |

## 4.　考　　察

### 4.1　抽出手法に関する考察

パタン作成時において、一意に定まる出来事名であるための必要条件として固有表現の存在を前提としたが、

> “一口サイズのこんにゃく入りゼリーを子供がのどに詰まらせる事故”

といったように、特定の出来事であるが固有表現を含まない事故名も存在した。抽出の対象とする表現の定義を、固有表現によらずに定める必要がある。

また今回は、人手で作成したパタンによる抽出であったが、このようなパタンの自動生成も今後の課題である。今回用いたような手書きのパタンを種として、bootstrapping や co-training によってパタンを学習していくことも考えられる。

### 4.2　関連付け手法に関する考察

Bag-of-Words、編集距離ともに、2 つの表現の字面上での比較を行っているので、字面上に共通する語がない場合には、しきい値をどのように設定してもそれらが同じ出来事であるとは判定されない。

しかし、例えば

> “ペルー人質事件”
> “リマの日本大使公邸占拠事件”

といったように、字面上は共通点がなくても同じ出来事を示す表現が存在する。このような場合に対処するためには、

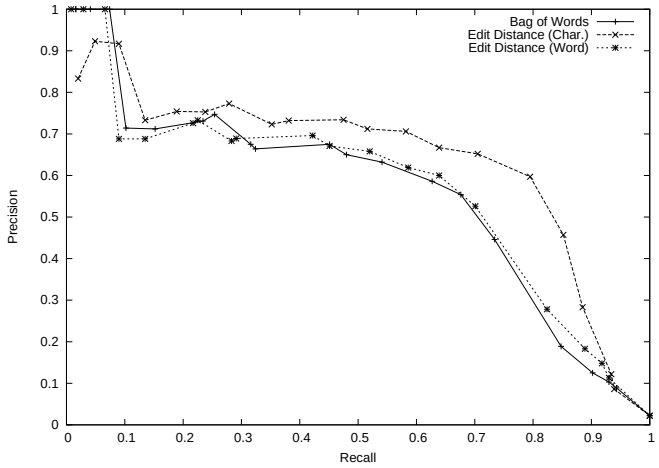


図 1　事故・事件名の関連付け結果 (トレーニングデータ)

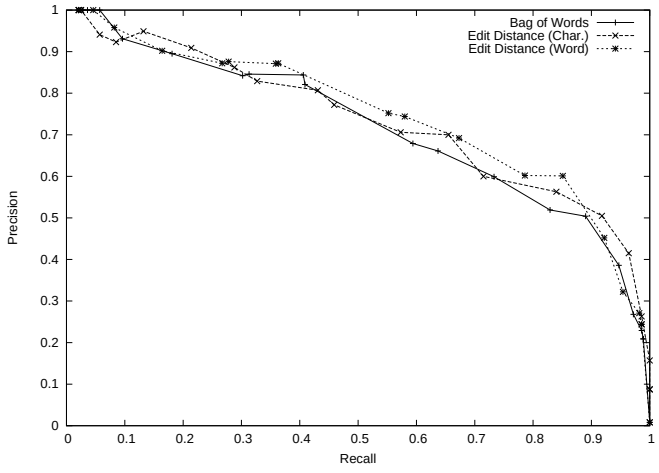


図 2　事故・事件名の関連付け結果 (テストデータ)

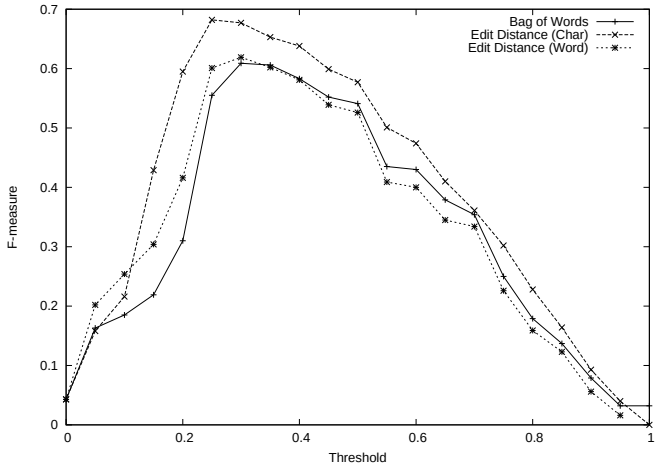


図 3　しきい値を変化させたときの事故・事件名の関連付け結果の変化 (トレーニングデータ)

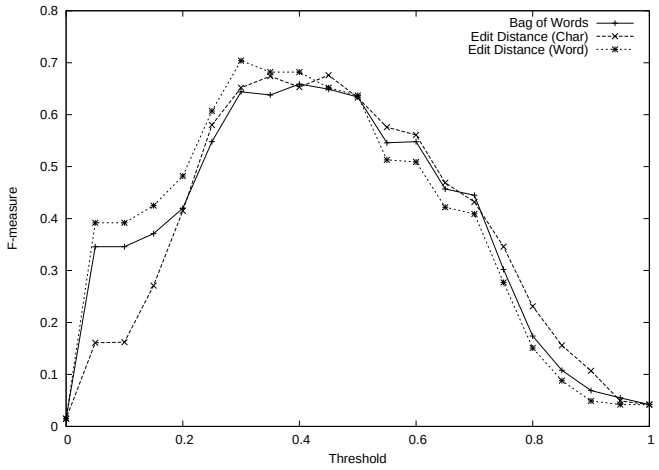


図 4　しきい値を変化させたときの事故・事件名の関連付け結果の変化 (テストデータ)

- 表現周辺の文脈情報の導入

(周囲の表現が類似しているかどうかを考慮できるようにする)

- 語彙知識の導入 (シソーラス等)

(「人質 (をとる)」行為と「占拠 (する)」行為の意味的な近さを考慮できるようにする)

- 背景知識の導入

(「リマはペルーの首都」といった背景知識を導入し、それに基づく推論が行えるようにする)

などの手法が考えられる。

また、ここでは同一の出来事を示す表現を関連付けることを目的としたが、事件や事故が拡大していった場合など、実際にはある出来事が別の出来事の一部分になっている場合がある。これについては、関連付けを行う際に、関連付けられたものは全て同一の出来事とみなすのではなく、関連付けの種類をラベル付けできるようにするなど、タスクを拡張することが考えられる。

## 5. 結　　論

本研究では、新聞記事中の出来事を表す表現の認識の部分タスクとして、事故・事件名の抽出とそれらの間での関連付けを行った。表現の抽出は人手で作成したパタンによって行い、テストデータに対して再現率 76.8%、適合率 81.6%の結果を得た。また

関連付けの実験では、Bag-of-Words と編集距離を用いて表現が異なる事故・事件名が同じ出来事を示すかどうかを判定し、文字単位の編集距離を類似度として用いた関連付けが、他の類似度よりもデータの違いに対して安定していることを示した。

## 文　　献


[1] DARPA. *Proceedings of the Sixth Message Understanding Conference (MUC-6)*, Columbia, MD, USA, November 1995. Morgan Kaufmann.

[2] DARPA. *Proceedings of the Seventh Message Understanding Conference (MUC-7)*, Fairfax, VA, USA, May 1998.

[3] DUC. http://duc.nist.gov, 2001–. Document Understanding Conference.

[4] NII, editor. *Proceedings of the Fourth NTCIR Workshop on Research in Information Retrieval, Automatic Text Summarization and Question Answering*, Tokyo, Japan, June 2004. National Institute of Informatics.

[5] Satoshi Sekine and Chikashi Nobata. Definition, dictionary and tagger for extended named entities. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation 2004*, pp. 1977–1980, Lisbon, Portugal, 2004.

[6] Satoshi Sekine, Kiyoshi Sudo, and Chikashi Nobata. Extended Named Entity Hierarchy. In *Proceedings of the LREC-2002 Conference*, pp. 1818–1824, 2002.

[7] 雨宮雄一. 新聞社会面記事における「事件」の表現. 計量国語学, Vol. 24, No. 1, pp. 19–39, 2003.

[8] IREX 実行委員会（編）. IREX ワークショップ予稿集. IREX 実行委員会, 9 月 1999.

[9] 黒橋禎夫, 河原大輔. 日本語形態素解析システム JUMAN version 4.0. 東京大学大学院情報理工学系研究科, 2003.

[10] 野畑周, 関根聡, 内元清貴, 井佐原均. 新聞記事における出来事を示す表現の分類と分析. 言語処理学会 第 10 回年次大会併設ワークショップ, 3 月 2004.