# 交通事故事例に含まれる事故原因表現の新聞記事からの抽出

酒井　浩之[†]　梅村　祥之[†]　増山　繁[†,††]

新聞記事に含まれる交通事故事例の記事から事故原因を表す表現(例えば,「ハンドル操作を誤った」)を自動的に抽出する手法を提案する．抽出結果に基づき交通事故事例の原因を分析することで，例えば交通事故防止装置の開発に役立てることができる．本手法では，まず，前処理としてSupport Vector Machines(SVM)を用いて新聞記事コーパスから交通事故事例の記事を抽出し，抽出された交通事故事例の記事から事故原因を表す表現を，新聞記事コーパスから得られる統計的な情報を使用して抽出する．具体的には，事故原因を表す表現がいくつか係る表現を種表現と定義して人手で1つ与え，種表現に係っている事故原因表現を自動的に取得する．そして，取得したいくつかの事故原因表現から自動的に種表現を取得し，さらに，取得した種表現から再び事故原因表現を取得する．このプロセスを繰り返すことで，事故原因表現，および，その種表現を取得していく．本手法を評価したところ，事故原因表現抽出の精度は77.2%であり，再現率は38.6%であった．また，事故原因表現，および，種表現を共に含んでいる文，もしくは，事故原因表現に「らしい」が追加された表現を含む文を原因文と定義し，その抽出精度，再現率を求めたところ，精度が87.2%，再現率が40.8%であった．

**キーワード**:　情報抽出，原因表現抽出

# Extraction of Expressions concerning Accident Cause contained in Articles on Traffic Accidents

Hiroyuki Sakai[†] , Shouji Umemura[†] and Shigeru Masuyama[†,††]

We propose a method for extracting expressions concerning accident cause (e.g., “mishandling of the steering wheel control”) contained in articles of traffic accidents from a newspaper corpus. It is effective to develop traffic accident prevention devices by analyzing cause of the traffic accident cases obtained by our method. Our method extracts expressions concerning accident cause from articles of traffic accidents extracted as a preprocessing from a newspaper corpus by using SVMs. Here, we define an expression modified by expressions concerning accident cause as “a seed expression”. Our method acquires expressions concerning accident cause from an initial seed expression provided manually. Moreover, our method acquires seed expressions from the expressions concerning accident cause and acquires new expressions concerning accident cause from the acquired seed expressions. By iterating these processes, expressions concerning accident cause and seed expressions are acquired. Experimental results showed that our method attained 77.2% precison and 38.6% recall. Here, we define a sentence containing both an expression concerning accident cause and a seed expression or a sentence containing an expression that adds ”*rashii*(らしい: seem to)” to an expression concerning accident cause as a cause sentence and the precision and the recall of extraction of cause sentences attained 87.2% and 40.8%, respectively.

**KeyWords:** *Information extraction, Cause expression extraction*

# 1 はじめに

自動車の保有台数は，年々増加の一途をたどり，平成13年には8,972万台と，国民1.4人に1台の割合となった．それに伴い，交通事故が大きな社会問題となっている．事故統計によれば，道路交通事故発生件数は昭和45年の720,880件を最高に，一旦は減少した．しかし，昭和50年代半ばから再び増加傾向を示し，平成5年には昭和45年の件数を上回る．その後，増加の一途をたどり，平成16年には，952,191件に達している (内閣府 2005)(交通事故総合分析センター 2002)．

交通事故低減に向けた効果的な対策のために，事故原因の分析が重要であることに疑問の余地はない．そのためのデータとして代表的なものは交通事故統計年報 (交通事故総合分析センター 2002) である．同事故統計は，発生時間帯別や当事者の年齢別など事故を様々な角度から集計しており，マクロ分析と呼ばれている．その中で，上記目的にあった集計には，道路形状別・昼夜別・事故類型別全事故件数，法令違反別・当事者別全事故件数，当事者別・行動類型別死亡事故件数などがある．例えば，事故類型別の集計では，事故を「出会い頭」,「追越・追抜時」等，36種類の類型に分類して集計している．しかし，なぜ，出会い頭あるいは追越・追抜で事故になったかという原因は，このデータだけでは分からない．

マクロ分析に対してミクロ分析がある．資料 (交通事故総合分析センター 2005) は，出会い頭事故をミクロ事故調査により，詳細に分析したものである．事故原因をドライバの「認知エラー」,「判断・予測エラー」で大別し，それぞれの原因を数項目に分類して分析している．このようなミクロ分析は，当事者の聞き取り調査によって初めて可能となった分析であり価値の高いものである．しかし，交通事故防止に関する研究のため，自分の研究テーマに沿った観点で分析したくても，同資料に開示された以上のことは不明である．しかし，同資料は公的な組織による交通事故現場での詳細な調査に基づくものであり，このようなデータを追加収集することは容易ではない．

一方，近年のインターネットの普及などにより大量の電子テキストが利用可能となった．その中にあるニュース記事やWebページなどに，膨大な交通事故のテキスト情報が含まれている．そこで，情報検索・情報抽出などの言語処理技術を利用して，交通事故に関する大量のテキスト情報を抽出し，さらに，事故原因に関する情報を抽出できれば，従来のマクロ分析，ミクロ分析を補完する有用な情報になり得ると期待できる．本論文では，新聞記事の電子テキス

† 豊橋技術科学大学知識情報工学系, Department of Knowledge-based Information Engineering, Toyohashi University of Technology
†† 豊橋技術科学大学インテリジェントセンシングシステムリサーチセンター, Intelligent Sensing System Research Center, Toyohashi University of Technology

トデータに含まれる交通事故を扱った記事[1]から，事故原因に関する情報として事故原因を表す表現（例えば,「ハンドル操作を誤った」）を自動的に抽出する手法を提案する．

第2章では，まず，前処理として新聞記事コーパスから交通事故事例記事を抽出し，その中から事故原因を表す表現を自動的に抽出する手法を提案する．第3章では，手法の実装と，本手法によって新聞記事コーパスから実際に抽出された事故原因を表す表現を示す．第4章では，抽出した事故原因表現の応用について述べる．第5章では本手法の評価について述べ，第6章では評価結果を考察する．第7章では関連研究について述べ，関連研究と本研究の違いや本手法の特徴について述べる．

# 2　提案手法

## 2.1　交通事故事例記事の抽出 (前処理)

本論文で提案する手法は，交通事故事例に含まれる事故原因を表す表現を抽出する手法であるが，その前処理として，新聞記事コーパスから交通事故を扱った記事 (以降，交通事故事例記事とする) を抽出する．そして，抽出された交通事故事例から交通事故の原因を表す表現 (以降，事故原因表現とする) を抽出する．新聞記事コーパスからの交通事故事例記事の抽出には Support Vector Machines(SVM)(Vapnik 1995)(Vapnik 1999) を用いる．

### 訓練データの作成

SVM の学習に用いる訓練データの作成について述べる．SVM を用いる場合，訓練データの作成には人手を必要とするため大きな労力が必要になるが，本手法では「衝突」と「乗用車」という語が含まれている表題をもつ記事は交通事故事例記事である可能性が高いという特性を利用することで，訓練データの作成も半自動的に行なっている．訓練データは1998年の読売新聞記事から，表題に「衝突」と「乗用車」が含まれている記事を正例とし，その結果，90記事が取得された．(ただし，90記事のうち1記事のみ「衝突安全ボディ」に関する記事であったので，その記事を除外したが，後は全て交通事故事例であった．そのため，89記事が正例となった．) そして，正例と同数の記事を無作為に選び，負例とした．その結果，178記事が訓練データとなる．

### 素性選択

SVM における素性選択について述べる．本タスクにおける素性は，正例にのみ多く含まれている内容語 (名詞，動詞，形容詞) とした．つまり，交通事故と関連がある内容語 (例えば,「正面衝突」,「スピード」,「軽傷」など) を訓練データから抽出する必要がある．そのために，まず，

1 なお，本論文における交通事故事例は道路交通事故に限る．

正例に含まれている内容語(以降，語とする)に対して重み付けを行い，重みが上位半分となる語を抽出する．重み付けには次の式1を用いる．

$$W_p(t_i, S_p) = P(t_i, S_p)H(t_i, S_p) \tag{1}$$

$$P(t_i, S_p) = \frac{Tf(t_i, S_p)}{\sum_{t \in Ts_{S_p}} Tf(t, S_p)} \tag{2}$$

ただし，

$P(t_i, S_p)$:　正例の文書集合$S_p$における語$t_i$の出現確率

$S_p$:　訓練データにおいて正例に属する文書集合

$Tf(t_i, S_p)$:　正例の文書集合$S_p$に含まれる語$t_i$の数

$Ts_{S_p}$:　正例の文書集合$S_p$に含まれる語の集合

$H(t_i, S_p)$:　正例の文書集合$S_p$に含まれる各文書における語$t_i$の出現確率に基づくエントロピー

$H(t_i, S_p)$は，正例の文書集合$S_p$に含まれる各文書における語$t_i$の出現確率に基づくエントロピーを表し，エントロピーが高い語ほど，正例の文書集合に均一に分布している語であることが分かる．この指標を導入した理由は，正例の文書集合中でも多くの文書に分散して出現している語の方が，少数の文書に集中して出現している語と比較して，よりその文書集合の特徴を表し，素性としても有効であるという仮定に基づく．$H(t_i, S_p)$は次の式3で定義される．

$$H(t_i, S_p) = -\sum_{d \in S_p} P(t_i, d) \log_2 P(t_i, d) \tag{3}$$

$$P(t_i, d) = \frac{tf(t_i, d)}{\sum_{d \in S_p} tf(t_i, d)} \tag{4}$$

ここで，$P(t_i, d)$は文書$d$における語$t_i$の出現確率を表し，$tf(t_i, d)$は文書$d$に含まれる語$t_i$の数を表す．

次に，正例の場合と同様に，負例に含まれる語に対して次の式5を用いて重み付けを行い，重みが上位半分となる語を抽出する．

$$W_n(t_i, S_n) = P(t_i, S_n)H(t_i, S_n) \tag{5}$$

$$P(t_i, S_n) = \frac{Tf(t_i, S_n)}{\sum_{t \in Ts_{S_n}} Tf(t, S_n)} \tag{6}$$

ただし，$S_n$は訓練データにおいて負例に属する文書集合である．そして，ある語$t_i$の正例における重み$W_p(t_i, S_p)$が負例における重み$W_n(t_i, S_p)$の2倍より大きければ，その語$t_i$を素性として選択する．すなわち，以下の条件が成り立つ語$t_i$を素性として選択する．

$$W_p(t_i, S_p) > 2W_n(t_i, S_n) \tag{7}$$

**表 1**　選択された素性の例

| | | | |
|---|---|---|---|
| 大破 | 正面衝突 | センターライン | 横転 |
| 対向車線 | はみ出す | 右カーブ | 右折 |
| 中央分離帯 | 即死 | 死亡 | 重傷 |

式1で表した重みでは，一般的な語であれば交通事故事例とは関係のない語でも高い重みが付与される．しかし，そのような語は負例においても高い重みが与えられる可能性が高い．そこで，ある語$t_i$に対する正例における重み$W_p(t_i, S_p)$と負例における重み$W_n(t_i, S_n)$を比較し，$W_p(t_i, S_p)$の方が$W_n(t_i, S_n)$の2倍よりも大きい語を選択することで，一般的な語が素性として選択されることを防ぐ．178記事の訓練データから素性を選択したところ，104個の素性が選択された．表1に選択された素性を例示する．

SVMによる学習に用いる素性ベクトルの各要素は，訓練データの各文書における素性として選択された語の出現確率とした．また，本研究では線形カーネルを利用した[2]．

## 2.2　事故原因表現の獲得手法の概要

前処理として抽出された交通事故事例記事から，事故原因表現を自動的に獲得する．本手法では，例えば,「前方不注意」,「スピードの出し過ぎ」などの表現を自動的に獲得することができる．

手法の説明にあたり，まず，**核文節**，**種 (たね) 表現**という用語を以下のように定義する．

**核文節:**　事故原因表現を構成する文節の最後尾の文節から，助詞や形式名詞「の」を削除したもの

**種表現:**　事故原因表現が係る文節に，その事故原因表現を構成する最後尾の文節に含まれる助詞や形式名詞「の」を追加した表現

例えば,「スピードの出し過ぎが原因」という文において，これを文節に区切ると「スピードの」「出し過ぎが」「原因」となる．この文に含まれている事故原因表現は「スピードの出し過ぎ」であるので，事故原因表現を構成する文節は「スピードの」「出し過ぎが」の2文節となり，その最後尾の文節は「出し過ぎが」となる．この文節から助詞「が」を削除した「出し過ぎ」が核文節となる．また，削除された助詞「が」を「原因」に追加した「が原因」が種表現となる．また,「スピードを出し過ぎたのが原因」という文において，これを文節に区切ると「スピードを」「出し過ぎたのが」「原因」となる．この文に含まれている事故原因表現は「スピードを出し過ぎた」であるので，その最後尾の文節は「出し過ぎたのが」となる．この文節から助詞「が」，および，形式名詞「の」を削除した「出し過ぎた」が核文節となる．そして，削除された形式名詞「の」，および，助詞「が」を「原因」に追加した「のが原因」が種表現となる．

2 予備実験として，2次の多項式カーネルを用いて比較評価実験を行ったが結果に差はみられなかった

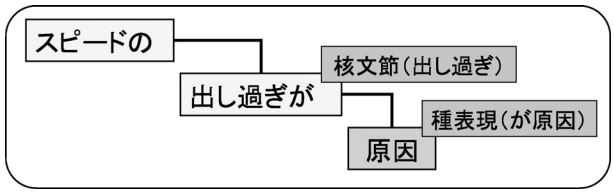


**図 1**　核文節と種表現

**表 2**　事故原因表現を含んだ文における核文節と種表現

| 文 | 事故原因表現 | 核文節 | 種表現 |
|---|---|---|---|
| スピードの出し過ぎが原因 | スピードの出し過ぎ | 出し過ぎ | が原因 |
| スピードの出し過ぎとみて | スピードの出し過ぎ | 出し過ぎ | とみて |
| スピードを出し過ぎたのが原因 | スピードを出し過ぎた | 出し過ぎた | のが原因 |
| スピードを出し過ぎたとみて | スピードを出し過ぎた | 出し過ぎた | とみて |

事故原因表現を構成する文節の最後尾の文節から，助詞や形式名詞「の」を削除した理由は，同じ事故原因表現を含む文でありながら，種表現が異るため事故原因表現が変化することを防ぐためである．例えば,「スピードの出し過ぎが原因」の場合，事故原因表現を構成する最後尾の文節は「出し過ぎが」となるが，そこから助詞「が」を削除することで核文節は「出し過ぎ」となり，種表現は「が原因」となる．それに対し,「スピードの出し過ぎとみて」の場合，事故原因表現を構成する最後尾の文節は「出し過ぎと」となるが，そこから助詞「と」を削除することで核文節は「出し過ぎ」となり，種表現は「とみて」となる．このように，同じ事故原因表現「スピードの出し過ぎ」を含む文でありながら，種表現が異るため事故原因表現を構成する最後尾の文節が異るが，核文節を定義することで種表現が異る場合でも同じ事故原因表現を獲得することが可能となる．また，核文節の取得の際に削除された助詞や形式名詞「の」を事故原因表現が係っている文節に追加することで，事故原因表現が前に出現している確率が高い種表現となる．

図1に例を示す．また，表2に事故原因表現を含んだ文における核文節と種表現をいくつか示す．

本手法の概要を以下に示す．

**Step 1:**　1つの初期種表現を人手で与え，それに係る事故原因表現を獲得する．

**Step 2:**　獲得した事故原因表現から，新たな種表現を獲得する．

**Step 3**　獲得した種表現から，新たな事故原因表現を獲得する．

**Step 4:**　Step 2，3を予め定めた回数まで繰り返す．　□

本手法では，初期種表現として「が原因」を人手で与えた．この初期種表現により，Step 1に

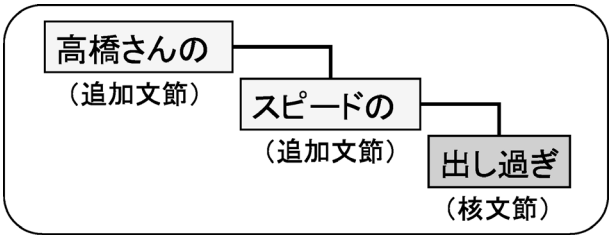


**図 2**　核文節「出し過ぎ」の拡張

おいて，例えば「前方不注意が原因」という文から「前方不注意」という事故原因表現を獲得する．Step 2において，獲得した事故原因表現を使用して，例えば「前方不注意とみて」という文から「とみて」という種表現を獲得する．Step 3において，初期種表現と獲得した種表現を使用して新たな事故原因表現を獲得し，Step 2にもどる．次節で，種表現からの事故原因表現を獲得する手法について述べる．

## 2.3　種表現に係る事故原因表現の獲得

種表現に係る事故原因表現は，次に示す「核文節取得」「拡張」「縮約」の3つの処理を順番に行うことで獲得される．

**核文節取得:**　種表現に係っている核文節を取得

**拡張:**　核文節に文節を追加して表現を拡張

**縮約:**　拡張された表現から不要な文節を除去して事故原因表現を生成

核文節取得では，前節で定義した核文節 (例えば,「出し過ぎ」) を取得する．しかし，核文節だけでは事故原因表現としては不十分である．そこで，核文節に係っている文節を追加して核文節の拡張を行う．例えば,「前方不注意」や「居眠り運転」は拡張の必要はないが,「出し過ぎ」や「誤った」といった核文節には拡張が必要になる．拡張では，核文節に係っている文節を核文節に追加する．そして，追加した文節に係っている文節をさらに追加し，表現を拡張していく．図2に，核文節「出し過ぎ」の拡張の例を示す．しかし，単純に文節を追加して表現を拡張していくだけでは，事故原因表現には不必要な文節も追加されてしまう．図2における「高橋さんの」という文節は明らかに事故原因表現には不必要な文節であり，このような文節は除去する必要がある．次節では，拡張された表現から不必要な文節を除去する処理である「縮約」について述べる．

## 2.4　縮約

縮約の手法を以下に示す．

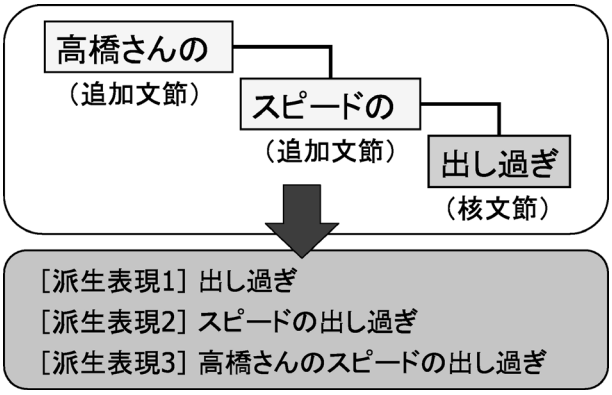


**図 3**　核文節「出し過ぎ」からの派生表現

**Step 1:**　核文節に文節を追加することで派生する表現を全て取得(図3を参照).

**Step 2:**　各表現のスコアを計算.

**Step 3:**　核文節から2回以上派生し，かつ，スコア最大の表現を事故原因表現として獲得．□

Step 1では，例えば，文書$A$に「高橋さんのスピードの出し過ぎ」という文が存在していたとすれば，この文から「出し過ぎ」「スピードの出し過ぎ」「高橋さんのスピードの出し過ぎ」という3つの表現が派生する．また，文書$B$に「大庭さんのスピードの出し過ぎ」という文が存在していたとすれば，この文から「出し過ぎ」「スピードの出し過ぎ」「大庭さんのスピードの出し過ぎ」という3つの表現が派生する．そして，文書$A$と文書$B$からは「出し過ぎ」が2回,「スピードの出し過ぎ」が2回,「高橋さんのスピードの出し過ぎ」が1回,「大庭さんのスピードの出し過ぎ」が1回，派生したことになる．

Step 2では，核文節$c$から派生した各表現$e$に対して，式8で表されるスコアを計算する．

$$Score(e,c) \quad = \quad -pf(e)ef(e,c)\log_2 P(e,c) \tag{8}$$

$$P(e,c) \quad = \quad \frac{ef(e,c)}{Ne(c)} \tag{9}$$

ただし，

$P(e,c)$**:**　核文節$c$から派生する表現$e$の派生確率

$ef(e,c)$**:**　核文節$c$から派生する表現$e$の派生回数

$Ne(c)$**:**　核文節$c$から派生する表現の総数

$pf(e)$**:**　表現$e$に含まれる文節の数

例えば，前述の文書$A$と文書$B$の例では,「スピードの出し過ぎ」の$ef(e,c)$の値は2であり$Ne(c)$の値は6であるため，$P(e,c)$の値は2/6となる．また$pf(e)$の値は,「スピードの」と「出し過ぎ」で2となる.「出し過ぎ」の場合は，$P(e,c)$の値は「スピードの出し過ぎ」と同一であ

**表 3**　「誤った」から派生した表現の例

| 表現$e$ | $ef(e,c)$ | $P(e,c)$ | $Score(e,c)$ |
|---|---|---|---|
| 誤った | 114 | 0.235 | 238.5 |
| ハンドル操作を誤った | 93 | 0.19 | 443.7 |
| 何らかの原因でハンドル操作を誤った | 4 | 0.008 | 110.8 |
| 運転を誤った | 11 | 0.023 | 120.2 |
| 出し過ぎてハンドル操作を誤った | 7 | 0.014 | 128.5 |
| 運転操作を誤った | 9 | 0.019 | 103.6 |
| 原因でハンドル操作を誤った | 4 | 0.008 | 83.1 |
| 出し過ぎ、ハンドル操作を誤った | 4 | 0.008 | 83.1 |
| スピードを出し過ぎてハンドル操作を誤った | 7 | 0.014 | 171.3 |

**表 4**　核文節から取得された事故原因表現

| 核文節 | 事故原因表現 |
|---|---|
| 誤った | ハンドル操作を誤った |
| 出し過ぎ | スピードの出し過ぎ |
| 見ていなかった | 前をよく見ていなかった |
| 前方不注意 | 前方不注意 |
| 踏み間違えた | ブレーキとアクセルを踏み間違えた |

るが，$pf(e)$の値が1であるため，スコアは「スピードの出し過ぎ」より低くなる．

Step 3では，$ef(e,c)$の値が2以上である表現のうち，スコアが最大の表現を事故原因表現として獲得する．そのため，文書$A$と文書$B$の例では「高橋さんのスピードの出し過ぎ」,「大庭さんのスピードの出し過ぎ」は事故原因表現として獲得されず,「スピードの出し過ぎ」が事故原因表現として獲得される．表3に，1999年，2000年，2001年の3年分の読売新聞905,373記事において，核文節「誤った」から派生した表現の一部と，そのスコアを示す．表3の例では,「ハンドル操作を誤った」がスコア最大であるため，それが事故原因表現として獲得される．ただし，縮約された表現と核文節が同一の場合(つまり，1つの文節も核文節に追加されていない場合)，かつ，その核文節を構成している最初の語が動詞なら，それだけでは事故原因表現として不十分である場合が多い．例えば，核文節が「誤った」であり，その縮約の結果が「誤った」であった場合は，必要な格要素が除去されたため事故原因として何を誤ったのか分からない．そのため，そのような場合は事故原因表現として認定しない．

表4に，いくつかの核文節から取得された事故原因表現を示す．このように,「前方不注意」のような1文節で構成される事故原因表現や「ブレーキとアクセルを踏み間違えた」のような3文節で構成される事故原因表現も，縮約によって適切に取得される．

## 2.5　事故原因表現の選別

ある種表現から，核文節取得，拡張，縮約を行って事故原因表現が獲得されても，事故原因表現として不適切な表現も獲得される．そこで，種表現から獲得された事故原因表現の中から適切な事故原因表現を選別する．具体的には，様々な種表現に係っている事故原因表現は適切であるという仮定に基づき，事故原因表現が種表現に係る確率に基づくエントロピーを求め，その値がある閾値以上の事故原因表現を選別する．例えば，事故原因表現「前方不注意」は，「が原因」「とみて」「と見ている」など様々な種表現に係るため，高いエントロピーをもつ．そして，高いエントロピーをもつ事故原因表現のみを選別することで，様々な種表現に係っている事故原因表現が適切な事故原因表現として選別される．事故原因表現が種表現に係る確率に基づくエントロピーは式10で求める．

$$H(e) \quad = \quad -\sum_{s \in Se} P(e,s) \log_2 P(e,s) \tag{10}$$

$$P(e,s) \quad = \quad \frac{f(e,s)}{N(e)} \tag{11}$$

ただし，事故原因表現を抽出する交通事故事例記事集合において，

$P(e,s)$:　事故原因表現 $e$ が種表現 $s$ に係る確率

$f(e,s)$:　種表現 $s$ に係る事故原因表現 $e$ の数

$N(e)$:　事故原因表現 $e$ の総数

$Se$:　事故原因表現 $e$ が係る種表現の集合

閾値 $T_e$ は，以下の式12によって設定する．

$$T_e = \alpha \log_2 Ns \tag{12}$$

ただし，

$Ns$:　事故原因表現を獲得するのに使用した種表現の数

$\alpha$:　定数 $(0 < \alpha < 1)$

$\log_2 Ns$ は，事故原因表現が種表現に係る確率に基づくエントロピーの最大値を表し，その値と定数 $\alpha$ との積が閾値として設定される．ただし，初回は種表現の数が初期種表現「が原因」の1つなので，事故原因表現のエントロピー，および，閾値が0になる．そのため，初回のみ全ての事故原因表現が選別される．

## 2.6　種表現の獲得

抽出した事故原因表現から，新たな種表現を獲得するための手法について述べる．まず，抽出した事故原因表現を含む文を抽出し，その中で事故原因表現が係っている文節を獲得する．そ

して，その文節に対して，係っている事故原因表現を構成する最後尾の文節に含まれる助詞や形式名詞「の」を追加し，それを種表現候補とする．そして，種表現候補が事故原因表現によって係られる確率に基づくエントロピーを求め，ある閾値以上の種表現候補を種表現として抽出する．これは，適切な種表現には様々な事故原因表現が係っているという仮定に基づく．例えば，種表現「が原因」には,「前方不注意」「スピードの出し過ぎ」など様々な事故原因表現が係っている．種表現候補が事故原因表現によって係られる確率に基づくエントロピーは式13で求める．

$$H(s) = -\sum_{e \in Es} P(s,e) \log_2 P(s,e) \tag{13}$$

$$P(s,e) = \frac{f(s,e)}{N(s)} \tag{14}$$

ただし，事故原因表現を抽出する交通事故事例記事集合において，

$P(s,e)$**:**　種表現$s$が事故原因表現$e$によって係られる確率

$f(s,e)$**:**　事故原因表現$e$によって係られる種表現$s$の数

$N(s)$**:**　種表現$s$の総数

$Es$**:**　種表現$s$に係る事故原因表現の集合

閾値$T_s$は，以下の式15によって設定する．

$$T_s = \alpha \log_2 Ne \tag{15}$$

$Ne$は種表現を獲得するのに使用した事故原因表現の数である．また，定数$\alpha$は，事故原因表現獲得の閾値を求めるときの定数と同じである．

## 3　新聞記事からの事故原因表現の抽出

本手法を実装し，新聞記事コーパスから事故原因表現を抽出した．まず，前処理として1998年の読売新聞記事から訓練データを取得し，1999年，2000年，2001年の3年分の読売新聞905,373記事に対して交通事故事例記事の抽出を行った．その結果，21,097記事が交通事故事例記事として抽出された．そして，抽出された交通事故事例記事から事故原因表現を抽出した．なお，初期種表現は「が原因」とし，事故原因表現と種表現の獲得の繰り返し回数は5回とした．その結果，例えば，閾値を設定するための定数$\alpha$を0.6に設定した場合，57個の事故原因表現，6個の種表現を抽出した．表5に，定数$\alpha$を0.6に設定した場合に抽出した57個の事故原因表現を，表6に同じ条件で抽出した種表現を全て示す．(表に示した各事故原因表現，および，種表現は，人手による選別を行っていない．)

なお，実装にあたり，$SVM^{light}$ [3]を使用した．また，形態素解析器としてChaSen[4]，係り受

3 http://svmlight.joachims.org
4 http://chasen.aist-nara.ac.jp/hiki/ChaSen/

**表 5**　抽出した事故原因表現

| |
|---|
| 気付くのが遅れた，後退した， 交差点に進入した， どちらかの信号無視， |
| 無視して交差点に入った， 故障， 反対車線にはみだした， 右折した， |
| 前方不注意， 確認せずにドアを開けた，居眠り運転，前をよく見ていなかった， |
| センターラインを超えた，少年が一時停止しなかった，スリップ，侵入した， |
| 無理な追い越し，居眠り運転をしていた，対向車線にはみ出した， |
| ブレーキとアクセルを踏み間違えた，一時停止をしなかった，信号無視した， |
| スリップした，不十分だった，スピードを出し過ぎた，脱輪した，安全不確認， |
| いずれかが信号を無視した，一時停止を怠った，無理に追い越そうとした， |
| 急ハンドルを切った，居眠り，安全をよく確かめていなかった，勘違いした， |
| 曲がりきれず、対向車線に飛び出した，カーブを曲がり切れなかった， |
| 赤信号を見落とした，カーブを曲がりきれなかった，安全確認が不十分だった， |
| 右折しようとした，車を追い越そうとして対向車線に出た，スピードの出し過ぎ， |
| 信号無視をした，センターラインを越えた，トラブル，スピードの出しすぎ， |
| 運転していた，安全をよく確認しなかった，交差点に入った，前方不注視， |
| ことなど，前をよく見てなかった，ハンドル操作を誤った，わき見運転， |
| トラックがスリップした，ハンドル操作の誤り，対向車線に飛び出した |

**表 6**　抽出した種表現

| | | |
|---|---|---|
| のが原因らしい， | 可能性が， | のが原因と |
| が原因， | とみている， | とみて |

け解析器としてCaboCha[5]を使用した．

# 4　抽出した事故原因表現の応用例

## 4.1　交通事故原因の傾向分析

本手法によって抽出された事故原因表現を使用することで，交通事故原因の傾向を見る．具体的には，1999年，2000年，2001年の3年分の読売新聞記事から前処理で抽出された交通事故事例記事に含まれる事故原因表現の総数を調べ，交通事故原因の傾向を見る．図4に結果を示す．

## 4.2　交通事故事例記事の事故原因を含む文の抽出による要約

本手法によって抽出された事故原因表現を使用することで，交通事故事例記事において事故原因が記述してある文の抽出による要約を行う．そのことにより，特に多量の交通事故事例記事を対象として分析するときなどに，それぞれの記事の事故原因を素早く把握することができ

5 http://chasen.org/~taku/software/cabocha/

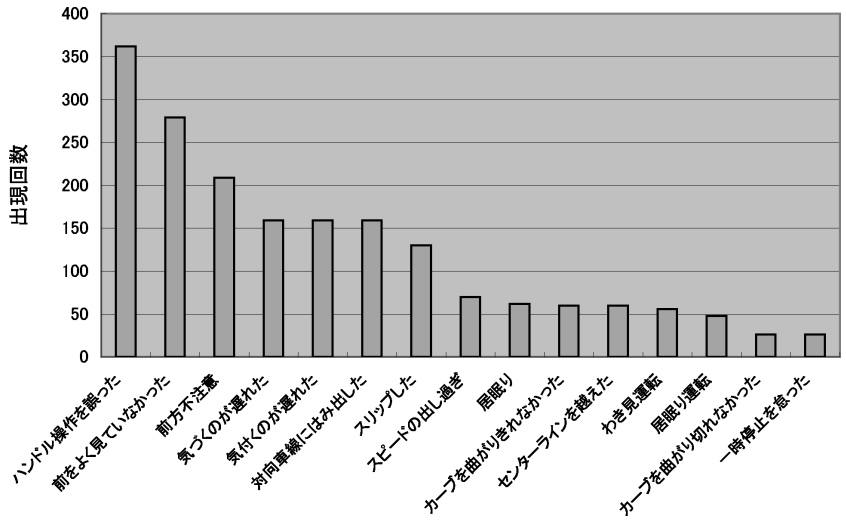


**図 4**　交通事故原因の傾向分析

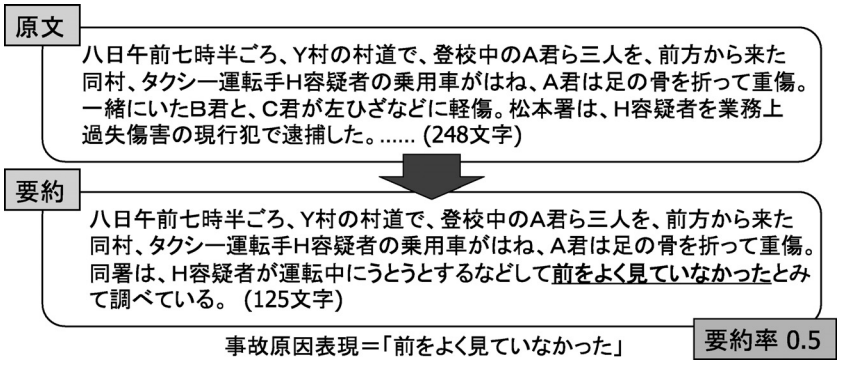


**図 5**　事故原因表現を使用した要約例

る．具体的には，交通事故事例記事において，事故原因表現，および，種表現が共に1つ以上含まれている，もしくは，事故原因表現に「らしい」が追加された表現 (例えば,「前方不注意らしい」) を含む文を事故原因が記述してある文として抽出する．それに，その交通事故事例記事の第一文を加えて，交通故事例記事を2つの文で構成する．(まれに，1つの記事から複数の文が事故原因の記述してある文として抽出されることもあるが，その場合は，抽出された全ての文を要約に含める)．図5に，元の交通事故事例記事と要約された記事の例を示す．なお，定数 $\alpha$ を 0.55 とした場合，抽出した 21,097 の交通事故事例記事のうち 2,211 記事に事故原因表現が含まれた文が含まれていた．そして，記事の第一文と事故原因表現が含まれた文とで構成された要約の要約率の平均は 0.63 であり，半数近くの文字が削減された．

# 5　評価

## 5.1　評価方法

本手法の評価を行った．そのために，まず，本手法を評価するための評価用データを作成した．本論文では，まず，前処理として1998年の読売新聞記事から訓練データを取得し，1999年，2000年，2001年の3年分の読売新聞905,373記事に対して交通事故事例記事を抽出する．そして，その中から，事故原因表現の抽出を行った．しかしながら，3年分の記事数は膨大であり，それに含まれる全ての事故原因表現を対象として，精度，再現率を求めるのは困難である．そこで，以下のようにして評価用データを作成した．まず，1999年，2000年，2001年の読売新聞905,373記事のうち，9人の工学系の学生に一人約3,000記事の新聞記事を読んでもらい，交通事故を扱った記事を選別してもらった[6]．その結果，27,722記事から699の交通事故事例記事を評価用データとして取得した．次に，選択した交通事故事例を読んでもらい，事故原因が記述してある文を選別してもらった(以降，事故原因が記述してある文のことを原因文と定義する)．その結果，201の原因文を評価用データとして取得した．そして，取得した原因文の中から，事故原因表現を人手で抽出した．その結果，延べ215個，異なり数132の事故原因表現を評価用データとして取得した．(一部の原因文には，2つの事故原因表現が記述されていた．)

評価では，これらの評価用データを使用して精度，再現率を測定する．具体的には，1999年，2000年，2001年の読売新聞905,373記事に対して，交通事故事例記事，および，事故原因表現の抽出を行う．そして，評価用データである27,722記事に含まれる699個の交通事故事例記事における交通事故事例記事抽出の精度，再現率，および，事故原因表現抽出の精度，再現率を測定する．

ここで，交通事故事例記事抽出の精度，再現率を以下のように定義した．

$$\text{精度}\,(\text{交通事故事例抽出}) = \frac{|Sd \cap Ad|}{|Sd \cap Nd|},\quad \text{再現率}\,(\text{交通事故事例抽出}) = \frac{|Sd \cap Ad|}{|Ad|}$$

ただし，

$Sd$:　3年分の新聞記事コーパスから抽出した交通事故事例記事を要素とする集合

$Ad$:　27,722記事に含まれる，人手で抽出した699個の交通事故事例記事を要素とする集合

$Nd$:　27,722記事を要素とする集合

事故原因表現抽出の評価は，精度に関しては，1999年，2000年，2001年の読売新聞905,373記事から本手法によって抽出した事故原因表現を人手で判定して測定した．しかし，再現率を測定するためには3年分の新聞記事コーパスに含まれる事故原因表現を全て人手で抽出する必要があり，全ての事故原因表現を対象として再現率を求めるのは困難である．そのため，評価用データである延べ215個の事故原因表現における再現率を測定した．具体的には，以下のよ

6 なお，正解とした交通事故事例は道路交通事故に限る．例えば，列車と人との衝突は含めない．ただし，列車と自動車の衝突は含める．

うに再現率を定義し測定した．

$$\text{再現率 (事故原因表現抽出)} = \frac{|Sc \cap Ac|}{|Ac|}$$

ただし，

$Sc$:　3年分の新聞記事コーパスから抽出した事故原因表現を要素とする集合

$Ac$:　699個の交通事故事例記事に含まれる，人手で抽出した延べ215個の事故原因表現を要素とする集合

また，抽出した事故原因表現，および，種表現を使用して，評価用データの対象となっている27,722記事に含まれる交通事故事例記事から原因文を認定し，その精度，再現率を測定した．具体的には，抽出した事故原因表現，および，種表現を共に1つ以上含む，もしくは，事故原因表現に「らしい」が追加された表現(例えば,「前方不注意らしい」)を含む文を原因文として抽出し，評価用データである201個の原因文と比較して，原因文抽出の精度，再現率を測定する．ここで，原因文抽出の精度，再現率を以下のように定義した．

$$\text{精度 (原因文抽出)} = \frac{|Ss \cap As|}{|Ss \cap Ns|},\ \text{再現率 (原因文抽出)} = \frac{|Ss \cap As|}{|As|}$$

ただし，

$Ss$:　3年分の新聞記事コーパスから抽出した原因文を要素とする集合

$As$:　699個の交通事故事例記事に含まれる201個の原因文を要素とする集合

$Ns$:　699個の交通事故事例記事に含まれる文を要素とする集合

## 5.2　評価結果

前処理である交通事故事例記事抽出の評価結果は，精度82.0%，再現率84.1%であった．事故原因表現抽出の精度，再現率は，2.5節で定義した定数$\alpha$を0.6に設定した場合にF値が最大となり，精度77.2%，再現率38.6%であった．ただし，抽出される事故原因表現，および，種表現の数は定数$\alpha$によって大きく変化する．そのため，$\alpha$を0.9から0.1まで変化させた場合の精度，再現率を測定した．結果を図6に示す．そして，表7に，定数$\alpha$を変化させた場合の事故原因表現の抽出数，種表現の抽出数を示す．なお，参考のため，その場合の精度，再現率，F値も並記する．

また，本手法は事故原因表現，および，種表現の獲得を繰り返すことで抽出される事故原因表現を増やしていくため，繰り返す回数によって精度，再現率が変化する．そのため，繰り返し回数を変動させた場合の精度，再現率を測定した．その際，定数$\alpha$の値は0.6と0.3に設定した．結果を表8，表9に示す．

原因文抽出に関しては，定数$\alpha$を0.55に設定した場合にF値が最大となり，精度87.2%，再現率40.8%であった．定数$\alpha$を0.9から0.1まで変化させた場合の精度，再現率を図7に示す．ま

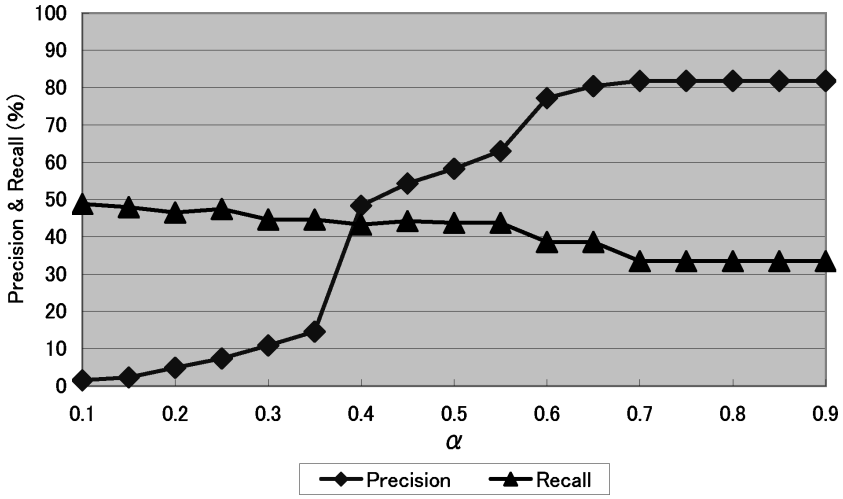


**図 6**　事故原因表現抽出の精度，再現率

**表 7**　定数$\alpha$を変化させた場合の事故原因表現，種表現の抽出数，および，精度，再現率

| $\alpha$ | 事故原因表現抽出数 | 種表現抽出数 | 精度 (%) | 再現率 (%) | F 値 |
|---|---|---|---|---|---|
| 0.9 | 33 | 1 | 81.8 | 33.5 | 47.5 |
| 0.7 | 33 | 1 | 81.8 | 33.5 | 47.5 |
| 0.6 | 57 | 6 | 77.2 | 38.6 | 51.5 |
| 0.5 | 91 | 16 | 58.2 | 43.7 | 49.9 |
| 0.4 | 120 | 29 | 48.3 | 43.3 | 45.7 |
| 0.3 | 738 | 304 | 10.8 | 44.7 | 17.4 |
| 0.2 | 1930 | 1906 | 4.9 | 46.5 | 8.9 |
| 0.1 | 6149 | 11788 | 1.5 | 48.8 | 3.0 |

**表 8**　繰り返し回数を変化させた場合の事故原因表現，種表現の抽出数，および，精度，再現率 ($\alpha = 0.6$)

| 繰り返し回数 | 事故原因表現抽出数 | 種表現抽出数 | 精度 (%) | 再現率 (%) |
|---|---|---|---|---|
| 1 | 33 | 5 | 81.8 | 33.5 |
| 2 | 52 | 6 | 76.9 | 38.6 |
| 3 | 57 | 6 | 77.2 | 38.6 |
| 4 | 57 | 6 | 77.2 | 38.6 |
| 5 | 57 | 6 | 77.2 | 38.6 |

た，表10に，定数$\alpha$を変化させた場合の，1999年，2000年，2001年の読売新聞905,373記事からの原因文の抽出数を示す．なお，参考のため，その場合の原因文抽出の精度，再現率，F値も並記する．

**表 9**　繰り返し回数を変化させた場合の事故原因表現，種表現の抽出数，および，精度，再現率 ($\alpha = 0.3$)

| 繰り返し回数 | 事故原因表現抽出数 | 種表現抽出数 | 精度 (%) | 再現率 (%) |
|---|---|---|---|---|
| 1 | 33 | 29 | 81.8 | 33.5 |
| 2 | 129 | 64 | 46.5 | 43.7 |
| 3 | 335 | 133 | 22.1 | 44.7 |
| 4 | 665 | 194 | 12.2 | 44.7 |
| 5 | 738 | 304 | 10.8 | 44.7 |

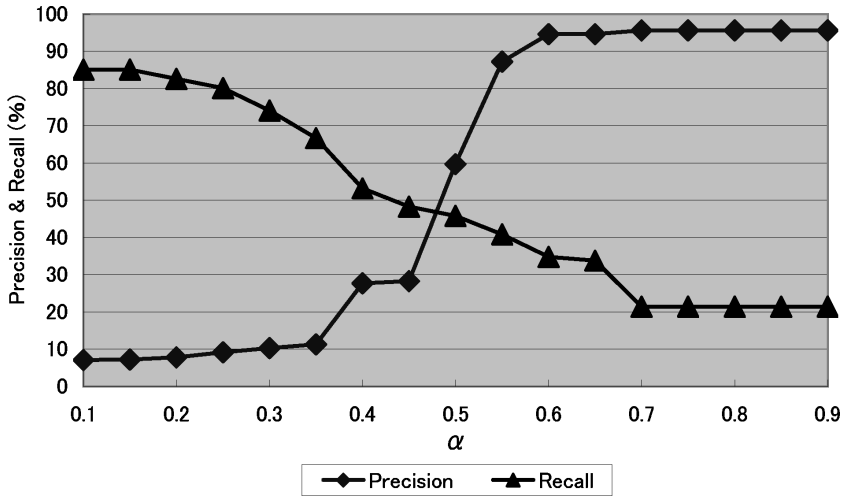


**図 7**　原因文抽出の精度，再現率

## 6　考察

まず，前処理である新聞記事からの交通事故事例記事抽出における誤り分析を行う．誤って交通事故事例として抽出された記事には，車の不審火や火事に関する記事，列車に人がはねられた記事，人が転落して死亡した記事が多かった．特に列車に人がはねられた記事が最も多く，誤認定の 10% を占めていた．表 11に，誤って抽出された記事に多く含まれていた素性をいくつか示す．これらの語は列車の事故，ないし，人の転落事故に関する記事にも頻出する語であるため，それらの記事が交通事故事例として誤認定されたと考える．そこで，1998 年の読売新聞記事から誤認定されそうな交通事故以外の死亡記事を 89 記事，人手で選択し，それを負例として交通事故事例記事を抽出し，精度，再現率を測定した．その結果，精度 95.7%，再現率 63.2% であり，精度は大きく向上したが再現率は大幅に低下した．これは,「調べ」「死亡」「事故」といった交通事故事例記事に頻出し，交通事故以外の死亡記事にも頻出する語が有効な素性ではなくなったからであると考える．そのため，再現率を落とさず精度をより向上させるためには，

**表 10**　定数$\alpha$を変化させた場合の原因文の抽出数，および，精度，再現率

| $\alpha$ | 原因文抽出数 | 精度 (%) | 再現率 (%) | F 値 |
|---|---|---|---|---|
| 0.9 | 1094 | 95.6 | 21.4 | 35.0 |
| 0.7 | 1094 | 95.6 | 21.4 | 35.0 |
| 0.6 | 1887 | 94.6 | 34.8 | 50.9 |
| 0.55 | 2743 | 87.2 | 40.8 | 55.6 |
| 0.5 | 4918 | 59.7 | 45.8 | 51.8 |
| 0.4 | 11067 | 27.7 | 53.2 | 36.4 |
| 0.3 | 43779 | 10.3 | 74.1 | 18.1 |
| 0.2 | 63425 | 7.8 | 82.6 | 14.3 |
| 0.1 | 70825 | 7.1 | 85.0 | 13.1 |

**表 11**　誤って抽出された記事に多く含まれる素性の例

| | | | |
|---|---|---|---|
| 調べ | 乗用車 | 死亡 | 事故 |
| 午前 | 現場 | 午後 | けが |

交通事故事例記事の性質を利用したヒューリスティクスに基づく規則を導入する必要があると考える．

次に，本手法である事故原因表現の抽出について考察する．表5より，事故原因を表す表現が精度よく抽出されている．また，表8，および，表9から，繰り返し回数を増やすことで抽出される事故原因表現の数が増えていくことが分かる．ここで，定数$\alpha$の値を0.6と設定した場合，繰り返し回数が3回目と4回目で抽出される事故原因表現の数が変化していない．これは，これ以上，事故原因表現，および，種表現が抽出されないことを示し，これ以上繰り返し回数を増やしても抽出される事故原因表現がないことが分かる．よって，繰り返し回数を増やしていき，事故原因表現，および，種表現の抽出数が，ともに変化しなくなれば処理を終了することができる．しかしながら，定数$\alpha$の値を0.3と設定した場合では，5回の繰り返しが終わっても事故原因表現，および，種表現の抽出数が増えている．そして，5回目の事故原因表現抽出の精度が10.8%であることから，多くの不適切な表現が事故原因表現として抽出されたことが分かる．不適切な種表現，事故原因表現が多く獲得されるようになると，不適切な種表現から多くの不適切な事故原因表現が獲得され，その不適切な事故原因表現から不適切な種表現が獲得されるという悪循環となり，いつまでたっても処理が終了しなくなる．そのため，繰り返し回数の上限を定める必要があり，本評価実験ではそれを5とした．

定数$\alpha$を0.6と設定した場合，事故原因表現が精度よく抽出された．しかし，定数$\alpha$を0.3と低く設定した場合，例えば「雨で地盤が緩んでいた」のような明らかに交通事故事例記事ではない記事の原因表現も抽出されていた．このように，交通事故事例記事の抽出精度が事故原因表現の抽出精度に影響することもあり，交通事故事例記事抽出の評価結果と事故原因表現抽出

**表 12**　必要な文節が除去された表現も正解とした場合の事故原因表現抽出の精度，再現率 (参考)

| $\alpha$ | 精度 (%) | 再現率 (%) | F 値 |
|---|---|---|---|
| 0.9 | 87.9 | 40.5 | 55.4 |
| 0.7 | 87.9 | 40.5 | 55.4 |
| 0.6 | 86.0 | 44.2 | 58.4 |
| 0.5 | 63.7 | 50.7 | 56.5 |
| 0.4 | 56.7 | 54.4 | 55.5 |
| 0.3 | 13.6 | 66.5 | 22.5 |
| 0.2 | 7.0 | 82.8 | 12.9 |
| 0.1 | 3.3 | 99.5 | 6.3 |

の評価結果には依存性がある．しかし，定数$\alpha$を0.6と設定した場合では，交通事故事例以外の記事から誤って抽出された原因表現はない．本手法では，様々な種表現に係っている事故原因表現は適切であるという仮定に基づき，事故原因表現が種表現に係る確率に基づくエントロピーを求め，その値がある閾値以上の事故原因表現を選別する．そのため，たとえ，交通事故事例以外の記事が多少含まれていても，定数$\alpha$を高く設定すれば，そこから適切な事故原因表現を抽出することが可能であると考える．

しかし，定数$\alpha$を0.6と設定した場合でも「交差点に進入した」や「右折しようとした」という，それだけでは事故原因表現として不十分な表現もある.「交差点に進入した」という表現は,「赤信号を無視して」,「周囲を良く見ずに」といった表現が前方から係っていたが，縮約の段階でこれらは除去された．また,「右折しようとした」という表現は,「安全を十分に確かめず」,「左右の安全をよく確認しないで」などの表現が除去された．これらは，本手法における縮約において，必要な文節が除去されてしまったため不適切な表現となってしまった．参考として,「交差点に進入した」や「右折しようとした」などの，必要な文節が除去されてしまったために不適切な事故原因表現とされた表現も正解とした場合の精度，再現率を測定した．結果を表12に示す．本手法における縮約では，例えば「交差点に進入した」に複数の表現が係っている場合,「交差点に進入した」に最も高いスコアを割り当て，係っている表現を除去する手法となっている．これは，核文節「進入した」から派生する表現のうち，最もスコアが高い表現を1つだけ事故原因表現として抽出しているからであり，閾値を設定し，核文節から複数の事故原因表現を抽出できるようにすれば解決できると考える．しかし，その閾値をどのように決定すればよいかといった問題が生じると考える．

本評価実験では，3年分の新聞記事コーパスから事故原因表現を抽出し，その精度は人手で測定したが，再現率は27,722記事から人手で抽出した延べ215の事故原因表現を評価データとして測定した．そのため，参考として，27,722記事から事故原因表現を抽出し，その精度，再現率を測定した．結果は，定数$\alpha$の設定値にかかわらず，事故原因表現の抽出数は「前方不注

**表 13**　同じ事故原因を表す表現

| | |
|---|---|
| 前方不注意 | 前をよく見ていなかった，<br>前方不注視，前をよく見てなかった，わき見運転 |
| 安全不確認 | 安全をよく確かめていなかった，<br>安全確認が不十分だった，<br>安全をよく確認しなかった |
| 居眠り運転 | 居眠り運転をしていた，居眠り， |
| 対向車線に飛び出した | 反対車線にはみだした，センターラインを超えた，<br>車を追い越そうとして対向車線に出た，<br>センターラインを越えた |
| スピードの出し過ぎ | スピードの出しすぎ，スピードを出し過ぎた |

視」,「前方不注意」,「ハンドル操作を誤った」の3つだけであり，種表現は初期種表現の「が原因」のみあった．そして，事故原因表現抽出の精度，再現率は，精度が100%，再現率が16.7%であった．また，原因文抽出の精度，再現率は，精度が100%，再現率が11.4%であった．事故原因表現が3つしか獲得できなかった理由は，本手法では，事故原因表現を新聞記事コーパスから得られる統計的な情報を使用して抽出しており，27,722記事では有効な統計情報を取得することができなかったからである．事故原因表現が本手法によって獲得されるためには，新聞記事コーパスにおいて同一の事故原因表現が同一の種表現に少なくとも2回以上係る必要がある．また，事故原因表現から獲得される新たな種表現は，少なくとも2種類以上の事故原因表現が係っていなければならない．そのため，本手法によって適切に事故原因表現を抽出するためには，ある程度のコーパス量が必要になる．

表5より，同一の意味をもつ異った事故原因表現があることが分かる．例えば,「前方不注意」「前をよく見ていなかった」「わき見運転」は同一の原因を表す事故原因表現である．表13に，表5に示した事故原因表現の中から同じ事故原因を表している表現を示す．これらの事故原因表現を1つの事故原因表現にまとめることができれば，より正確な事故原因の傾向分析が可能であると考える．また，交通事故原因に基づく交通事故事例記事のクラスタリングも可能になる．しかし,「前方不注意」「前をよく見ていなかった」「わき見運転」といった表現を同一の表現と認定することは，人間でも事故原因表現の意味を理解する必要があるため機械処理は容易ではないと考えられ，これは今後の課題としたい．

本論文では，新聞記事コーパスに含まれる交通事故事例記事における事故原因表現の抽出手法について述べたが，新聞記事では死亡事故のような大きな交通事故しか掲載されないため，そのような事例の原因しか抽出できないという問題がある．そして，交通事故事例の原因分析には，事故になりかけた事例やヒヤッとした瞬間の事例の原因も必要である．しかし，そのような事例は新聞記事には掲載されない．事故になりかけた事例を取得するためには，新聞記事ではなくWeb上のブログを対象にして本手法を適用してみる必要があると考える．しかし，新聞

記事と異り，ブログでは表現が定型的ではない．本手法では同一の表現がコーパス中に何回か出現しないと事故原因表現を取得できないため，本手法をそのままブログに適用することは困難である．そこで，本手法によって新聞記事から取得された事故原因表現をキーワードとしてブログから交通事故事例を取得し，さらに，取得された交通事故事例からブログ特有の事故原因表現を獲得するといったアプローチが有効であると考え，今後，取り組んでいく予定である．

## 7　関連研究

本研究は，抽出すべき情報を交通事故事例における事故原因表現と設定し，新聞記事コーパスから抽出した交通事故事例記事から事故原因表現を取得する．新聞記事やWebといった知識源から，抽出すべき情報を表す表現を取得する研究はいくつか行われており，それらをいくつか挙げ，本研究との違いについて述べる．

那須川らは，好評文脈，不評文脈を分析し，好不評表現の性質を利用することでネット上の掲示版から好評表現，不評表現を取得する手法を提案している (那須川哲哉, 金山博, 坪井祐太, 渡辺日出雄 2005)．那須川らの手法では，種表現として少数の好評表現，不評表現を人手で与え，その種表現から好不評表現の性質を利用して文書中の好不評文脈を推測し，その中からさらに好評表現，不評表現を取得することを繰り返して，ブートストラップ的に多くの好評表現，不評表現を自動的に抽出している．それに対して，本手法は種表現として事故原因表現自身を与えるのではなく，事故原因表現が係っている文節を種表現として定義し，それを1つ与え，ブートストラップ的に事故原因表現と種表現を取得する．ただし，与える情報は1つの種表現のみであり，事故原因表現の性質を利用して事故原因表現を抽出するのではなく，新聞記事コーパスから得られる統計情報を使用して抽出を行う．そのため，抽出すべき情報の表現に適した種表現を与えれば原因表現以外の表現抽出への本手法の適用も可能である．野畑らは，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，新聞記事から事故・事件名を人手で作成したパタンによって自動抽出する手法を提案している (野畑周, 佐田いち子, 井佐原均 2005)．また，2つの事故・事件名が与えられた場合，編集距離を用いることでそれらが同一の事故・事件を表しているかどうかを判定している．小林らは，特定の商品やサービスに対する意見を，意見を＜対象，属性，評価値＞という3つ組で表し，それぞれに該当する表現を，対象名辞書，属性表現辞書，評価値表現辞書や，人手で作成した共起パタンを使用して半自動的に収集する手法を提案している (小林のぞみ, 乾健太郎, 松本裕治, 立石健二, 福島俊一 2005)．しかし，1つの文節を越える属性表現や評価値表現を収集できないことを問題点として挙げている．また，Morinagaらは，Webページから，ある製品に関する意見情報を自動的に収集し，分析する手法を提案している (Morinaga, Yamanishi, Tateishi, and Fukushima 2002)．意見情報抽出は，ある製品に関する評価表現を含む文を評価表現辞書を使用して抽出し，評価表現を含む文が適切な意見情報かどうかは，人手で設定されたルールを適用することで判定する．それらに対して，

本手法で人手で与える情報は1つの種表現のみであり，人手で作成したパタンや辞書は不要である．また，複数の文節で構成される事故原因表現も抽出可能である．Riloff らは，意見を示す手がかり表現を人手で与え，それを使って抽出された意見文から意見を抽出するためのパタンを学習する手法を提案している (Riloff and Wiebe 2003)．それに対して，本手法は事故原因を表す表現を抽出する手法であり，パタンを抽出するわけではない．

原因と結果の関連を表す因果関係の知識を新聞記事から獲得する研究がいくつか行われている．乾らは，接続標識「ため」に着目し，例えば「タイでマングローブを破壊したため，大水害が発生した」という文から「タイでマングローブを破壊した」ことが原因で,「大水害が発生した」という結果を得る因果関係知識を獲得する手法を提案している (乾孝司, 乾健太郎, 松本裕治 2004)．Khoo らは，接続標識を利用した人手で作成したパターンを適用することで，新聞記事から因果関係情報を取得している (Khoo, Kornfilt, Oddy, and Myaeng 1998)．また，構文構造を考慮した人手で作成したパターンを適用することで，医療関係のデータベースから因果関係知識を取得している (Khoo, Chan, and Niu 2000)．これらの研究では，接続標識やパターンを適用する必要上，1 文の中に原因と結果が含まれていなければならない．それに対して本研究は，結果が交通事故でありその原因を表す表現を抽出しているが，結果である「交通事故」は既知であり，通常は原因を表す表現と同一文に出現しないことを前提としている．例えば，本手法では「県警は A さんの前方不注意が原因とみて調べている」という文から「前方不注意」という事故原因表現を取得するが，この文には結果となる「交通事故」に関する記述がない．そのため，本手法では，1 文に事故原因表現のみが出現していても，事故原因表現を識別し取得することが可能である．さらに，本手法では，統計的な情報を使用して不必要な文節を除去することで，汎用的な事故原因表現の知識を取得できる．例えば，前述の例文の「県警は A さんの前方不注意が原因とみて調べている」という文から,「A さんの」という不要な文節を除去し,「前方不注意」という事故原因表現を取得する．それによって,「B さんの前方不注意が原因とみて調べている」という文が含まれている交通事故事例記事の原因も特定することができる．

以上の関連研究に対して，本論文で提案する手法は交通事故事例記事に含まれる事故原因表現を抽出する手法であり，抽出すべき情報が異る．そのため，関連研究におけるパタンをそのまま適用することはできない．また，原因表現は複数の文節で構成されることも多いが (例えば,「ハンドル操作を誤った」は「ハンドル操作を」と「誤った」の 2 文節で構成されている．), 本手法では，原因表現に文節を追加 (拡張) し，統計的な情報を使用して追加された文節の中で不要な文節の除去 (縮約) を行う手法であると考えれば，交通事故事例に特化せずに複数の文節で構成される適切な原因表現を抽出することも可能であると考える．

## 8　まとめ

本論文では，新聞記事に含まれる交通事故事例記事から事故原因表現を自動的に抽出する手法を提案した．本手法では，例えば,「前方不注意」,「スピードの出し過ぎ」などの事故原因表現を，交通事故事例記事から自動的に抽出することができる．本手法では，前処理として1999年，2000年，2001年の読売新聞記事コーパスから，SVMを用いて交通事故事例の記事を抽出し，そして，抽出された交通事故事例の記事から統計的な情報を用いて事故原因表現を抽出した．具体的には，まず，事故原因表現が係る文節に助詞を追加した表現を種(たね)表現と定義し，また，種表現に直接係っている文節の文末から助詞と名詞「の」を削除したものを「核文節」と定義した．そして，初期種表現として「が原因」を人手で与え,「が原因」に係っている事故原因表現を自動的に取得する．その際,「核文節取得」「拡張」「縮約」の3つの処理を順番に行うことで,「前方不注意」のような1文節で構成される事故原因表現や「ブレーキとアクセルを踏み間違えた」のような3文節で構成される事故原因表現も適切に取得された．次に，取得したいくつかの事故原因表現から別の種表現を自動的に取得し，さらに，取得した種表現から再び事故原因表現を取得する．このプロセスを繰り返すことで，事故原因表現，および，種表現を自動的に取得した．本手法を評価したところ，事故原因表現抽出の精度は77.2%であり，再現率は38.6%であった．また，事故原因表現，および，種表現を共に含んでいる文，もしくは，事故原因表現に「らしい」が追加された表現を含む文を原因文と定義し，その精度，再現率を求めたところ，精度が87.2%，再現率が40.8%であった．今後の課題として，同一の意味をもつ異った事故原因表現(例えば,「前方不注意」「前をよく見ていなかった」「わき見運転」など)のまとめ上げ，および，ブログからの事故原因表現抽出のための手法拡張を挙げる．

**謝辞** 本研究の一部は，文部科学省科学研究費特定領域研究(B)(2)16092213，及び，２１世紀COEプログラム「インテリジェントヒューマンセンシング」(豊橋技術科学大学)の援助により行われた．また，言語データとして，読売新聞CD-ROMの使用を許可して頂いた読売新聞社に深謝する．

## 参考文献

乾孝司, 乾健太郎, 松本裕治 (2004). “接続標識「ため」に基づく文書集合からの因果関係知識の自動獲得.” 情報処理学会論文誌, **45** (3), pp. 919–933.

Khoo, C. S., Kornfilt, J., Oddy, R. N., and Myaeng, S. H. (1998). “Automatic Extraction of Cause-Effect Information from Newspaper Text Without Knowledge-based Inferencing.” *Literary and Linguistic Computing*, **13** (4), pp. 177–186.

Khoo, C. S., Chan, S., and Niu, Y. (2000). “Extracting Causal Knowledge from a Medical

Database Using Graphical Patterns." In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 336–343.

小林のぞみ, 乾健太郎, 松本裕治, 立石健二, 福島俊一 (2005). "意見抽出のための評価表現の収集." 自然言語処理, **12** (3), pp. 203–222.

Morinaga, S., Yamanishi, K., Tateishi, K., and Fukushima, T. (2002). "Mining product reputations on the Web." In *Proceedings of Eighth ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD2002)*, pp. 341–349.

那須川哲哉, 金山博, 坪井祐太, 渡辺日出雄 (2005). "好不評文脈を応用した自然言語処理." 言語処理学会第11回年次大会発表論文集, pp. 153–156.

野畑周, 佐田いち子, 井佐原均 (2005). "新聞記事中の事故・事件名の自動抽出." 情報処理学会研究報告 2005-NL-167, pp. 125–130.

Riloff, E. and Wiebe, J. (2003). "Learning Extraction Patterns for Subjective Expressions." In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 105–112.

Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. Springer.

Vapnik, V. (1999). *Statistical Learning Theory*. Wiley.

交通事故総合分析センター (2002). 交通事故統計年報 平成13年度版. 財団法人交通事故総合分析センター.

交通事故総合分析センター (2005). イタルダ・インフォメーション No.56 出会い頭事故における人的要因の分析. 財団法人交通事故総合分析センター (http://www.itarda.or.jp/).

内閣府 (2005). 交通安全白書 平成17年度版. 国立印刷局.

## 略歴

**酒井 浩之:** 2002年 豊橋技術科学大学大学院工学研究科修士課程 知識情報工学専攻 修了．2005年 豊橋技術科学大学大学院工学研究科博士後期課程 電子・情報工学専攻 修了．2005年 豊橋技術科学大学知識情報工学系助手．博士 (工学)．自然言語処理，特に，テキスト自動要約の研究に従事．言語処理学会，人工知能学会各会員．e-mail: sakai@smlab.tutkie.tut.ac.jp

**梅村 祥之:** 1981年名古屋大学大学院修士課程工学研究科修了．同年，東京芝浦電気株式会社入社．1988年株式会社豊田中央研究所入社．現在に至る．同社人間特性研究室に所属し，主に，自動車の予防安全のための人間特性の研究に従事．2003年豊橋技術科学大学博士後期課程社会人コース入学．言語処理学会，ヒューマンインタフェース学会，情報処理学会，日本音響学会，自動車技術会会員．e-mail: umemura@smlab.tutkie.tut.ac.jp

**増山 繁:** 1977年 京都大学工学部数理工学科卒業．1982年 同大学院博士後期

課程単位取得退学．1983年 同修了 (工学博士)．1982年 日本学術振興会奨励研究員．1984年 京都大学工学部数理工学科助手．1989年 豊橋技術科学大学知識情報工学系講師．1990年 同助教授．1997年 同教授．2005年 豊橋技術科学大学インテリジェントセンシングシステムリサーチセンター教授 併任．アルゴリズム工学，特に，並列アルゴリズム等，及び，自然言語処理，特に，テキスト自動要約等の研究に従事．言語処理学会，電子情報通信学会，情報処理学会等会員．e-mail: masuyama@tutkie.tut.ac.jp

(2005 年 9 月 23 日 受付)
(2005 年 11 月 30 日 再受付)
(2005 年 12 月 27 日 採録)