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

2005－ＦＩ－ 80 (13)
2005－ＮＬ－169 (13)
2005／9／30

新聞記事からの交通事故事例および事故原因表現の抽出

酒井 浩之† 梅村 祥之† 増山 繁†,††

† 豊橋技術科学大学 知識情報工学系
†† 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
E-mail: †{sakai,umemura}@smlab.tutkie.tut.ac.jp, ††masuyama@tutkie.tut.ac.jp

あらまし 新聞記事から交通事故事例の記事を抽出し，さらに，その中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．本手法では，まず，Support Vector Machine(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する．そして，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，交通事故事例抽出は精度 82.0%，再現率 84.3% 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 77.7%，再現率が 39.8% であった．
キーワード 情報抽出，原因表現抽出

Extraction of Articles concerning Traffic Accident and Expressions concerning Accident Cause

Hiroyuki SAKAI†, Yoshiyuki UMEMURA†, and Shigeru MASUYAMA†,††

† Department of Knowledge-based Information Engineering, Toyohashi University of Technology,
†† Intelligent Sensing System Research Center, Toyohashi University of Technology,
E-mail: †{sakai,umemura}@smlab.tutkie.tut.ac.jp, ††masuyama@tutkie.tut.ac.jp

Abstract We propose a method for extracting articles concerning traffic accident and expressions concerning accident cause (e.g., "mishandling of the steering wheel control") from a newspaper corpus. Our method extracts articles concerning traffic accident from a newspaper corpus by using SVMs, and extracts expressions concerning accident cause from the extracted articles. 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 of extraction of articles concerning traffic accident from a newspaper corpus attained 82.0% precison and 84.3% recall. Here, we define a sentence containing both an expression concerning accident cause and a seed expression as a cause sentence and the precision and the recall of extraction of cause sentences attained 77.7% and 39.8%, respectively.
Key words 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[9] [7].

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[7]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[8]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，当事者の聞取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．本稿では，新聞記事の電子テキストデータから交通事故を扱った記事を抽出し，さらに，その中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．

2. 事故事例記事の抽出

本手法は，まず，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．)を抽出し，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 Machine(SVM) [6]を用いる．

2.1 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人手を必要とするため大きな労力が必要になるが，本タスクでは「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で行える．

すなわち，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する．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．

2.2 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，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，重みが上位から半分以上である語を抽出する．重み付けには次の式1を用いる．

$$W_p(t_i, S_p) = \frac{Tf(t_i, S_p)}{\sum_{t_i \in Ts_{S_p}} Tf(t_i, S_p)} H(t_i, S_p) \qquad (1)$$

ただし，

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，

式1の第一項は正例の文書集合における語 t_i の出現確率を表す．$H(t_i, S_p)$ 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．この指標の導入の理由は，正例の文書集合の中でも少数の文書に集中して出現している語より多くの文書に分散して出現している語の方が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2で定義される．

$$H(t_i, S_p) = -\sum_{d \in S_p} P(t_i, d) \log_2 P(t_i, d) \qquad (2)$$

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
対抗車線	はみ出す	右カーブ	右折
中央分離帯	即死	死亡	重傷

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 3 を用いて重み付けを行い，重みが上位半分の語を抽出する．

$$W_n(t_i, S_n) = \frac{Tf(t_i, S_n)}{\sum_{t_i \in Ts_{S_n}} Tf(t_i, S_n)} H(t_i, S_n) \qquad (3)$$

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．

$$W_p(t_i, S_p) > 2W_n(t_i, S_n) \qquad (4)$$

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あまり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 における正例における重み $W_p(t_i, S_p)$ と負例における重み $W_n(t_i, S_p)$ を比較し，$W_p(t_i, S_p)$ の方が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いくつか示す．

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，本稿では線形カーネルを利用した(注1)．

3. 事故原因表現の自動獲得

抽出された交通事故事例記事から，交通事故の原因を表す表現 (以降，事故原因表現とする．) を自動的に獲得する手法について述べる．本手法では，例えば，「前方不注意」，「スピードの出し過ぎ」といった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，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) 表現と定義する．例えば，「スピードの出し過ぎが原因」という文では，「スピードの出し過ぎ」が事故原因表現であるので，種表現は「が原因」となる．また，「スピードの出し過ぎとみて」という文では，「とみて」が種表現となる．次に，種表現に直接係っている文節の文末から助詞と名詞「こと」を削除したものを「核文節」と定義する．例えば，「スピードの出し過ぎが原因」という文では，「出し過ぎ」が核文節となる．図 1 に例を示す．

(注1)：予備実験として，2 次の多項式カーネルを用いて評価実験を行ったが結果に差はみられなかった

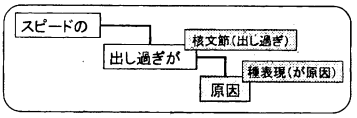

図 1　核文節と種表現

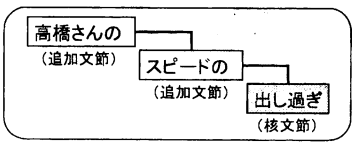

図 2　核文節「出し過ぎ」の拡張

3.1 交通事故表現の獲得手法の概要

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 を指定回数まで繰り返す．　□

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，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ことができる．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．

3.2 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「核文節獲得」「拡張」「縮約」の 3 つの処理を順番に行うことで獲得される。

核文節獲得:　種表現に直接係る文節を核文節として獲得

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

核文節獲得では，前節で定義した核文節 (例えば，「出し過ぎ」) を獲得する．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」は拡張の必要がないが，「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」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

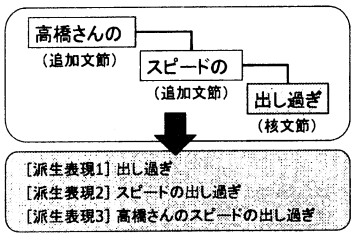

図 3　核文節「出し過ぎ」からの派生表現

去する処理である「縮約」について述べる.

3.3 縮　　約

縮約の手法を以下に示す.

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 5 で表されるスコアを計算する.

$$Score(e,c) = -pf(e)f(e,c)\log_2 P(e,c), \tag{5}$$

$$P(e,c) = \frac{f(e,c)}{N(c)} \tag{6}$$

ただし,

$P(e,c)$:　核文節 c から派生する表現 e の派生確率.

$f(e,c)$:　核文節 c から派生する表現 e の派生回数.

$N(c)$:　核文節 c から派生する表現の総数.

$pf(e)$:　文節 e に含まれる文節の数.

例えば，さきほどの文書 A と文書 B の例では,「スピードの出し過ぎ」の $f(e,c)$ の値は 2 であり $N(c)$ 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，$f(e,c)$ 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 2 に，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 2 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，事故原因表現として獲得される．また，表 3 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.

表 2　「誤った」から派生した表現の例

表現 e	$f(e,c)$	$P(e,c)$	$Score(e,c)$
誤った	39	0.23	82.2
ハンドル操作を誤った	37	0.22	161.5
運転操作を誤った	2	0.01	25.6
出し過ぎてハンドル操作を誤った	2	0.01	38.4

表 3　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
見落とした	どちらかが信号を見落とした

3.4 事故原因表現の選別

ある種表現から，核文節獲得，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当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抽出する．具体的には，事故原因表現の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抽出する．これは，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く．事故原因表現のエントロピーは式 7 で求める.

$$H(e) = -\sum_{s \in Se} P(e,s)\log_2 P(e,s) \tag{7}$$

$$P(e,s) = \frac{f(e,s)}{N(e)} \tag{8}$$

ただし,

$P(e,s)$:　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

エントロピーが大きい事故原因表現は様々な種表現に係っていることが分かる．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，本手法によって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抽出される.

閾値 T_e は以下の式 9 によって設定される.

$$T_e = \alpha \log_2 Ns \quad (9)$$

ただし，

Ns: 事故原因表現を抽出するのに使用した種表現の総数．

α: 定数 $(0 < \alpha < 1)$.

$\log_2 Ns$ は事故原因表現のエントロピーの最大値を表し，その値と 1 以下の定数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抽出される．

3.5 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が係っている文節を獲得し，種表現候補とする．そして，種表現候補の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」といった様々な事故原因表現が係っている．種表現候補のエントロピーは式 10 で求める．

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

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

ただし，

$P(s,e)$: 種表現 s が事故原因表現 e によって係られる確率．

$f(s,e)$: 種表現 s が事故原因表現 e によって係られる数．

$N(s)$: 事故原因表現によって係られる種表現 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 12 によって設定される．

$$T_s = \alpha \log_2 Ne \quad (12)$$

Ne は種表現を抽出するのに使用した事故原因表現の総数である．また，定数 α は，事故原因表現の抽出の閾値を求めるときの定数と同じである．ただし，初回のみ，事故原因表現の総数 Ne の値が小さく，閾値が低く設定されることを防ぐため，閾値を 1 とする．

4. 評　　価

4.1 手法の実装

評価を行うために本手法を実装した．本評価実験では，1998 年の読売新聞記事から訓練データを取得し，1999 年の読売新聞 242,985 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，4,524 記事が交通事故事例記事として抽出された．さらに，抽出された交通事故事例記事から事故原因表現を獲得する．なお，初期種表現は「が原因」とした．そして，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．例えば，閾値を設定するための定数 α を 0.4 に設定した場合，46 個の事故原因表現，15 個の種表現を獲得した．また，表 4 に，定数 α を 0.6 に設定した場合に獲得した事故事例表現を，また，表 5 に同じ条件で獲得した種表現を全て示す(注2)．

表 4　獲得した事故原因表現

信号無視	交差点に進入した
安全確認を怠った	右折しようとした
前方不注意	スピードの出し過ぎ
居眠り運転	前をよく見ていなかった
気づくのが遅れた	対向車線にはみ出した
一時停止をしなかった	安全をよく確認しなかった
スリップした	前方不注視
ハンドル操作を誤った	わき見運転

表 5　獲得した種表現

として、逮捕した。	とみて、	が原因と
のが原因らしい。	のが原因と	が原因
ことが	が原因らしい。	と見て
とみている。	ものと	とみて

なお，実装にあたり，SVM^{light}(注3) を使用した．また，形態素解析器として Chasen(注4)，係り受け解析器として Cabocha(注5) を使用した．

4.2 正解データの作成

本手法を評価するために正解データを作成した．正解データは，テストデータとして使用した 1999 年の読売新聞記事のうち，9 人の工学部の大学生に一人約 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(注6)．その結果，27,722 記事から 699 の交通事故事例記事を正解データとして取得した．また，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 の原因文を取得した．評価では，これらの正解データを使用して再現率，精度を測定する．事故原因表現の評価では，事故原因表現，および，種表現を使用することで原因文を抽出し，正解データと比較する．具体的には，1999 年の読売新聞 242,985 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，正解データの対象となっている 27,722 記事の中の 699 の交通事故事例記事における精度，再現率を測定する．また，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，原因文の精度，

(注2)：表に示した各表現は，人手による選別を行っていない．

(注3)：http://svmlight.joachims.org

(注4)：http://chasen.aist-nara.ac.jp/hiki/ChaSen/

(注5)：http://chasen.org/~taku/software/cabocha/

(注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．

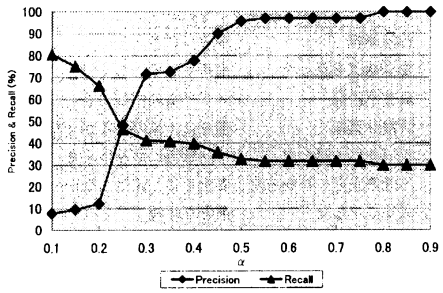

図 4　原因文の精度，再現率

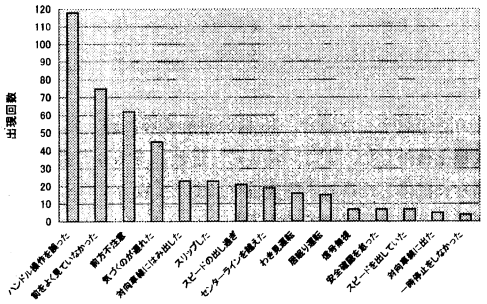

図 5　交通事故原因の傾向分析

再現率を測定する．

4.3　評価結果

交通事故事例抽出の評価結果は，精度 82.0%，再現率 84.1% であった．事故原因表現の評価は原因文の精度，再現率を測定することで行い，定数 α を 0.4 に設定した場合，精度 77.7%，再現率 39.8% 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 α によって大きく変化する．そのため，α を 0.9 から 0.1 まで変化した場合の精度，再現率を測定した．結果を図 4 に示す．また，表 6 に，定数 α 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．また，参考のため，その場合の原因文の精度，再現率，F 値も並記する．

4.4　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，1 年分の新聞記事から本手法によって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 5 に結果を示す．

4.5　交通事故事例記事の原因文抽出による要約

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，交通事故事例記事の原因文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る．具体的には，交通事故事例記事の第一文と原因文を重要文として抽出し，交通事故事例記事

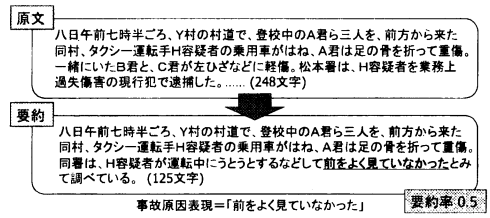

図 6　事故原因表現を使用した要約例

表 7　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

を 2 つの文で構成する (まれに複数の文が原因文として抽出されることもあるが，その場合は，全ての原因文を要約に含める)．なお，原因文は，事故原因表現，および，種表現が共に 1 つ以上，含まれている文である．図 6 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，α を 0.4 とした場合，抽出した 4524 の交通事故事例記事のうち 567 記事に本手法で抽出した原因文が含まれていた．そして，記事の第一文と原因文で構成された要約の要約率の平均は 0.64 であり，半分近くの文字が削減された．

5.　考　　察

まず，本手法による新聞記事からの交通事故事例記事の抽出の誤り分析を行う．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 7 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，人の転落事故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，素性から表 7 に示した語を除去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 80.8%，再現率 72.5% であり，再現率が大きく低下した．これは，「調べ」「死亡」「事故」といった語は交通事故事例記事に頻出する語であるため，再現率が低下したと考える．さらに，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，交通事故事例記事の性質を利用したヒューリスティックに基づく規則を導入する必要がある

表 6　定数 α を変化した場合の事故原因表現，種表現の抽出数，および，原因文の精度，再現率

α	事故原因表現の抽出数	種表現の抽出数	精度 (%)	再現率 (%)	F 値
0.9	14	12	100	29.9	46.0
0.8	14	12	100	29.9	46.0
0.7	15	12	97.0	31.8	47.8
0.6	16	12	97.0	31.8	47.8
0.5	23	12	95.7	32.8	48.8
0.4	46	15	77.7	39.8	52.6
0.3	55	19	72.6	40.8	52.2
0.2	153	148	12.2	66.2	20.6
0.1	819	2418	7.8	80.5	14.1

と考える．

次に，事故原因表現の抽出について考察する．表 4 より，事故原因を表す表現が精度よく抽出されている．しかし，「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外された．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」といった表現が除外された．本手法における縮約では「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．

6. 関連研究

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 [4]．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原因表現自身を与えるのではなく，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 つ与える．与える情報は 1 つの種表現のみであり，原因表現の性質を利用して原因表現を抽出するのではなく，統計情報を使用して抽出を行う．そのため，抽出する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 [5]．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で収集する手法を提案している [2]．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 [3]．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，本手法では与える情報は 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る [1]．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，パターンを抽出するわけではない．

以上の関連研究に対して，本研究は交通事故事例記事から事故原因を表す表現を抽出する手法が主であり，抽出すべき情報が異る．本手法は新聞記事からの交通事故事例記事抽出，および，交通事故事例記事からの事故原因表現の抽出の 2 つに分けられる．新聞記事からの交通事故事例記事を抽出する手法は，交通事故事例記事の特徴 (例えば表題には「衝突」「乗用車」といった語が出現することが多い．) を利用しており，本タスクに特化した手法となっている．それに対し，事故原因表現の抽出手法は，交通事故事例記事からの抽出に特化した手法ではなく，適切な種表現を 1 つ与えれば原因表現を自動的に抽出する．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)を行うことで，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できることが特徴である．

7. むすび

本稿では，新聞記事から交通事故事例の記事を抽出し，さらに，その中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，まず，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，そして，抽出された交通事故事例の記事から事故原因表現を抽出した．具体的には，事故原因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ていく．本手法を評価したところ，交通事故事例抽出が精度82.0%，再現率84.3%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77.7%，再現率が39.8%であった．

今後の課題として，以下の点が挙げられる．

- 同一の意味をもつ異った事故原因表現をまとめる必要がある．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．そのため，これらの事故原因表現を1つの事故原因表現(例えば，この場合では「前方不注意」)にまとめることができれば，より正確な傾向分析が可能であると考える．また，交通事故原因による交通事故事例記事のクラスタリングも可能になる．
- 本稿では，新聞記事コーパスからの交通事故事例記事，および，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，および，そ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できないと考える．そこで，本手法によって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や事故原因を取得し，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る．
- 新聞記事コーパスからの交通事故事例記事，および，事故原因表現の抽出という目的で手法を考案したが，本手法はこの目的に特化した手法ではなく，対象が交通事故事例でなくとも，その原因を表す表現が抽出可能であると考える．また，適切な初期の種表現を設定すれば，原因を表す表現以外にも根拠を表す表現なども抽出可能であると考える．ただし，適切な初期の種表現は，抽出する表現，および，対象となる文書の内容に依存する．そのため，初期の種表現を自動獲得できる手法を開発することで，より汎用的な手法になると考える．

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．

文献

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

[2] 小林のぞみ，乾健太郎，松本裕治，立石健二，福島俊一：意見抽出のための評価表現の収集，自然言語処理，Vol. 12, No. 3, pp. 203–222 (2005).

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

[4] 那須川哲哉，金山博，坪井祐太，渡辺日出雄：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 (2005).

[5] 野畑周，佐田いち子，井佐原均：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 (2005).

[6] Vapnik, V.: *Statistical Learning Theory*, Wiley (1999).

[7] 交通事故総合分析センター：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 (2002).

[8] 交通事故総合分析センター：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/) (2005).

[9] 内閣府：交通安全白書 平成17年度版，国立印刷局 (2005).