# 概念ベースと関連度計算を用いた 事件記事要約のための重要キーワード抽出

鈴木 修一　　渡部 広一　　河岡 司

同志社大学工学部知識工学科

## 1　はじめに

近年のパソコンや携帯電話の急激な普及により，人間が得る情報は膨大なものとなっている．それらの情報から必要なものだけを得るためには，コンピュータが情報の重要な部分だけを厳選し，提供する必要がある．コンピュータを利用して情報量の圧縮を行うことができれば，短時間でのユーザの長文理解支援につながる．

本研究では，事件記事を対象とし，概念ベース[1]と関連度計算[2]を用いて，記事から重要キーワードを抽出する手法を提案する．本研究において抽出される重要キーワードにより，単語レベルでの意味を考慮した情報量の圧縮が可能になる．

## 2　概念ベースと関連度計算

概念ベースとは，電子辞書などから自動構築された知識ベースである．ある一つの概念 $A$ を属性 $a_i$ と重み $w_i$ によって，次のように定義する．

$$A=\{(a_1,w_1),\cdots,(a_N,w_N)\} \tag{1}$$

一つの語（概念 $A$）は概念 $A$ の意味特徴を表す単語（属性と呼ぶ）とその属性の重要度（重み）の対の集合で表される．概念数は，約 9 万語で一つの概念につき平均 30 個の属性が存在する（図 1）．


| 概念 | 属性／重み |
|---|---|
| 医者 | 医師／0.3, ･･･, 治療／0.1 |
| 太陽 | 日光／0.2, ･･･, 月／0.05 |
| 美術 | 芸術／0.4, ･･･, 鑑賞／0.07 |
| : | : |

概念：約 9 万語

**図 1　概念ベース**

関連度とは，二つの概念 $A$ と $B$ の関連の強さを定量化した相対的な値である．

関連度は 0 から 1 までの連続値をとり，関連の強い概念同士では高い値となり，関連の弱い概念同士では低い値となる．例えば，概念「医者」と「病院」の関連度は 0.72，概念「医者」と「太陽」の関連度は 0.04 となる．このように概念同士の関連の強さを定量化すれば，数値の大小比較によって，曖昧である概念間の関連性の強弱をコンピュータに判断させることができるようになる．この概念ベースを利用して概念と概念の関連の強さを定量化する手法が関連度計算である．本研究では，重み比率付き関連度計算[2]を利用する．

### 2.1 重み比率付き一致度

2 つの概念 $A$, $B$ でその一次属性を $a_i$, $b_j$, 重みを $u_i$, $v_j$ とし，属性がそれぞれ $L$ 個，$M$ 個 ($L \leqq M$) とすると

$$\begin{aligned} A&=\{(a_i,u_i)\,|\,i=1\sim L\} \\ B&=\{(b_j,v_j)\,|\,j=1\sim M\} \end{aligned} \tag{2}$$

と表現でき，概念 $A$, $B$ の一致度 MatchWR($A$,$B$) は以下のようになる．

$$\begin{aligned} \mathrm{MatchWR}(A,B)&=\sum_{a_i=b_j}\min(u_i,v_j) \\ \min(\alpha,\beta)&=\begin{cases}\alpha(\beta\geq\alpha)\\ \beta(\alpha>\beta)\end{cases} \end{aligned} \tag{3}$$

重み比率付き一致度は一致する属性のうち小さい方の重みの和となるが，これは両方の属性に共通して存在する重み分は有効だと考えるためである．

### 2.2 重み比率付き関連度

概念 $A$, $B$ のうち属性数の少ない概念を $A$ ($L \leqq M$) とし，概念 $A$ の一次属性の並びを固定する．

$$A=\{(a_1,u_1),\ (a_2,u_2),\ \cdots,\ (a_L,u_L)\} \tag{4}$$

概念 $B$ の各一次属性を対応する概念 $A$ の各一次属性との一致度(MatchWR)の合計が最大になるように並べ替える．

$$B_x=\{(b_{x1},v_{x1}),\ (b_{x2},v_{x2}),\ \cdots,\ (b_{xL},v_{xL})\} \tag{5}$$

概念 $A$ と概念 $B_x$ との関連度 ChainWR($A$,$B$) は，

$$\mathrm{ChainWR}(A,B_x)=\sum_{i=1}^{L}\mathrm{MatchWR}(a_i,b_{xi})\times\frac{(u_i+v_{xi})}{2}\times\frac{\min(u_i,v_{xi})}{\max(u_i,v_{xi})}$$

$$\min(\alpha,\beta)=\begin{cases}\alpha(\beta\geq\alpha)\\ \beta(\alpha>\beta)\end{cases}\ \max(\alpha,\beta)=\begin{cases}\beta(\beta\geq\alpha)\\ \alpha(\alpha>\beta)\end{cases} \tag{6}$$

となる．すなわち，重み比率付き関連度は対応する一次属性の一致度と，それらの属性の重みの平均および重みの比に比例する．

## 3　キーワード抽出

提案したシステムでは，記事を入力し，記事関連度計算を用いた記事分類方式[3]を利用してカテゴリに分類する．そして事件カテゴリに分類された記事に対して，時間，場所，加害者，被害者項目に加え，それ以外の重要な語を格納した自立語項目からなる重要キーワードを出力するというものである（図 2）．なお本研究では，事件カテゴリのみを対象にキーワードの抽出を行った．このシステムの出力により，文章全体に目を通さなくても，その文章の内容を把握することができる．

具体的な処理の流れは，まず記事分類方式を利用し記事を事件カテゴリに分類する．次に事件カテゴリに分類された記事に対し，形態素・構文解析を行う．なお時間・場所項目には形態素解析結果を，加害者・被害者項目には構文解析結果を利用して該当する語を抽出し，各項目に格納する．さらに，事件カテゴリを交通，火災，水難，強盗からなる 4 つの小カテゴリに分類する．そして各小カテゴリ毎に用意した知識と記事内に存在する自立語の関連度を求め，関連度の高いものを自立語項目に格納する．知識はそれぞれ概念ベースにある概念 10 語からなる．

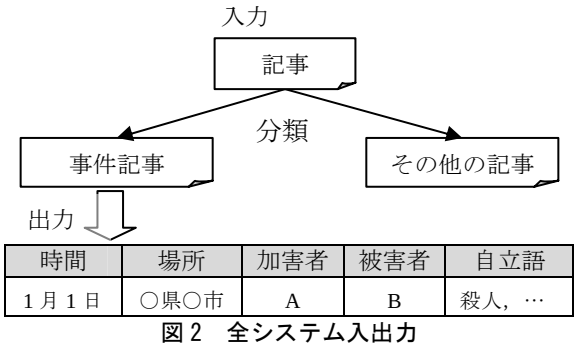


**図 2　全システム入出力**

### 3.1 記事分類方式

使用している記事分類方式は，概念ベースと関連度計算を用い，記事を最適なカテゴリに自動で分類するものである．新聞記事は経済面や社会面というようにカテゴリに分類することができる．このカテゴリに記事を分類するためには，分類の基準が示されなくてはいけない．分類の基準とは，カテゴリを包括的に表す「経済」という語や，各カテゴリのサンプルとなる記事のように，それらの特徴を表す情報である．そこでカテゴリの情報を表す語を代表概念，各カテゴリのサンプルとなる記事を代表記事と呼ぶ．これをふまえ，記事分類の手がかりとしてコンピュータに代表概念を与え，その代表概念を手がかりに分類対象の記事群の中から代表記事を自動で抽出し，その代表記事をもとに記事を分類する．

### 3.2 時間項目格納方式

時間項目は，記事内の日付と時刻に関する情報のなかで，記事の更新日に最も近い語を格納するものである．日付を形態素解析すると，「○日」という表記に対し，品詞は「名詞-数＋名詞-接尾-助数詞」となる．この結果と一致する，記事内の表記を時間項目の候補として抽出する．記事内に複数の日付が記述されている場合には，記事の更新日に注目し，最も更新日に近い日付を時間項目に格納する．一つに絞った日付に対し，その日付での時刻情報が記事内にある場合には，日付に加えて時刻も時間項目に格納する．これにより事件が起きた日付を獲得する．

### 3.3 場所項目格納方式

場所項目は，記事内に記述されている地名を格納するものである．地名を形態素解析すると，品詞について「名詞-固有名詞-地域-一般」「名詞-接尾-地域」という結果が得られる．記事を形態素解析したとき，連続してこれらの品詞が続いた場合，もしくは「名詞-固有名詞-地域-一般」の品詞の前後が人名に関する品詞でない場合に，その語を場所項目に格納する．

### 3.4 加害者・被害者項目格納方式

加害者・被害者項目は，記事内にある加害者・被害者となる人名を格納するものである．記事内の人名に対して，加害者・被害者の判定を行う．判定の例外処理として，人名の後ろに「容疑者」「被告」表記がある場合，その人名を加害者項目に，人名の後ろに「さん」という表記がある場合，その人名を被害者項目に格納する．これ以外の人名は構文解析結果から人名の修飾する助詞，動詞，受身を判定し，加害者・被害者の判定を行う（図 3）．

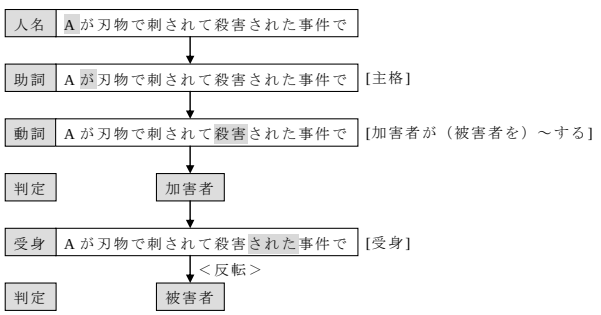


**図 3　加害者・被害者抽出の流れ**

助詞では，人名が書かれた文節に対し主格，目的格を判定する．

動詞では，まず「加害者が（被害者を）～する」，「被害者が（加害者を）～する」という 2 通りの形に当てはまる動詞，名詞-サ変接続を用意する．今回は前者に「殺害」，後者は「死亡」といった語である．これを利用し記事を構文解析することで，人名のある文の述語となる動詞がどちらになるかを判定する．以上の助詞と動詞の判定から，一度，加害者・被害者の判定を行う．例えば，助詞が主格で動詞が「加害者が（被害者を）～する」の形を取るものなら，その人名を加害者と判定し，助詞が目的格なら被害者と判定する．さらにその動詞が受身形の場合には，判定結果を加害者なら被害者，被害者なら加害者と反転させる．最後に加害者もしくは被害者と判定された人名に対し，それぞれ重複と包含関係を考慮し，人名を加害者・被害者項目に格納する．

### 3.5 自立語項目格納方式

自立語項目は，各小カテゴリ毎に用意した知識（表 1）と記事内に存在する自立語との関連度を求め，関連度の高い自立語を格納するものである．自立語と小カテゴリにおける知識 10 語で関連度計算を行うので，関連度も 10 個取得される（図 4）．そこで自立語と知識の関連度として，10 個の関連度の最大値を用いることにする．そして関連度に閾値を設定し，自立語を絞り込む．この関連度の閾値には，実験で関連度の閾値を 0～1 の間を 0.01 刻みで検証したところ，0.05 でF値[4]が 0.69 となる最適な自立語が得られたため，この値に設定する（図 5）．結果の図には閾値 0.1 までしか示していないが，F値が閾値 0.1 以上で上がることがなかったため省略した．F値とは，再現率と精度の両方を考慮した尺度である．

表 1　小カテゴリの知識

| カテゴリ | 事件 | | | |
|---|---|---|---|---|
| 小カテゴリ | 交通 | 火災 | 水難 | 強盗 |
| 知識 | 交通 | 火災 | 水難 | 強盗 |
| | 道 | 火傷 | 転覆 | 盗み |
| | 事故 | 放火 | 海 | 強奪 |
| | 運転 | 建物 | 水死 | 脅し |
| | 車 | 消防 | 船 | 金 |
| | ひき逃げ | 爆発 | 衝突 | 侵入 |
| | 酒 | 出火 | 転落 | 刃物 |
| | 信号 | 身元 | 川 | 銃 |
| | 衝突 | 火事 | 溺死 | 財布 |
| | 交差点 | 焼死 | 釣り | 逃走 |

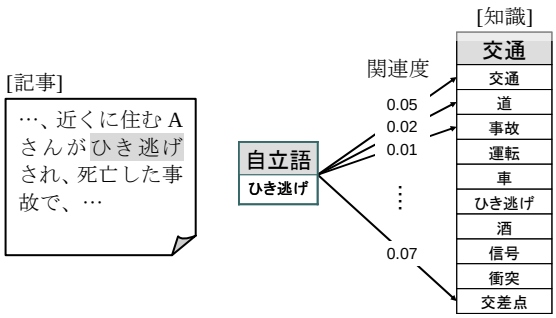


図 4　自立語と知識の関連度計算

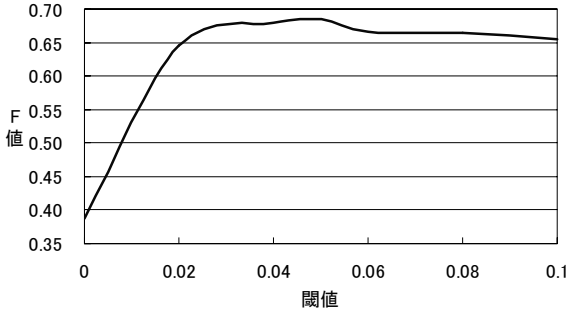


図 5　F 値による閾値の検証結果

## 4　分類の検証

3.1 節に示した記事分類方式を用い，事件記事を小カテゴリ（交通，火災，水難，強盗）に分類する再現率を検証した．再現率とは，システムの有効性を評価する尺度であり，ここでは用意した記事が正しく元のカテゴリに分類されたかを評価した．過去の研究[3]より代表記事件数は 10 件とする．

事件記事を交通，火災，水難，強盗からなる小カテゴリ分類を検証するため，テストデータにYahoo! ニュース[5]より小カテゴリ（交通，火災，水難，強盗）ごとに 100 件ずつ合計 400 件用意した．このデータを用い，代表概念として「交通」，「火災」，「水難」，「強盗」を入力し，事件カテゴリに分類された記事を小カテゴリ（-交通，火災，水難，強盗）に分類する再現率を検証した．全体の平均で 85%という分類の再現率が得られた（図 6）．

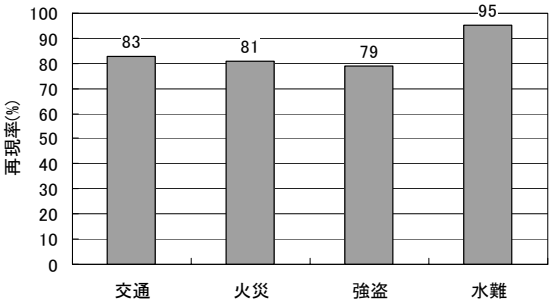


図 6　事件記事の分類

## 5　時間，場所，加害者，被害者項目の検証

提案手法における 4 項目格納方式の評価のため，加害者と被害者が特定されている記事 100 件からなるデータベースを作成した. このデータベースには，更新日と記事がセットで格納されている.

時間項目は，事件が起きた日が取れているか，場所項目は，事件の起きた場所（地名）が正確に取れているか，加害者・被害者項目は，正しい人名が加害者・被害者に正しく格納できているかという基準で評価した.

記事 100 件での事件フレームの評価結果は，時間項目で 98%, 場所項目で 93%, 加害者項目で 86%, 被害者項目で 78%，4 項目全体で平均 89%の正答率が得られた(図 7). 正答率とは記事 100 件に対する，項目ごとで正解と評価された記事の割合である.

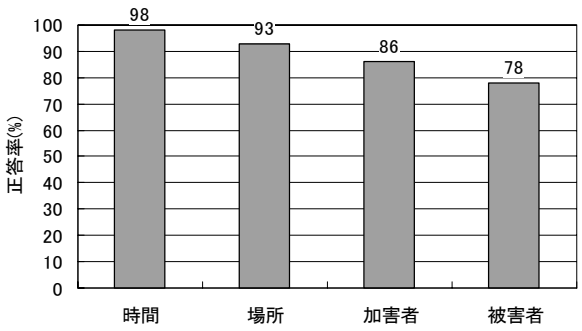


図 7　4 項目評価結果

## 6　自立語項目の検証

提案手法における自立語項目評価のため，小カテゴリ（交通，火災，水難，強盗）ごとに 100 件ずつ合計 400 件の記事からなるデータベースを作成した. この記事は，4 章で用いた事件記事と一致する.

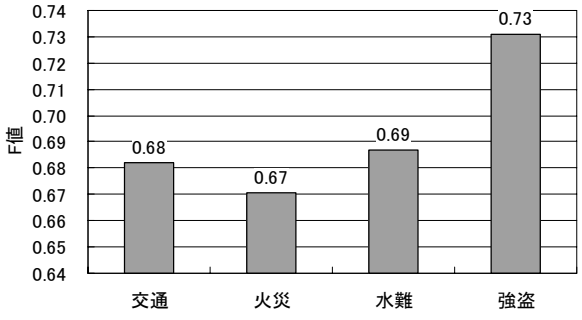


図 8　自立語項目の評価結果

評価は要約に必要と考えられる自立語（正解語）を記事 1 件ごとに用意し，関連度計算により絞り込んだ自立語に対して再現率と精度から F 値での評価を行った（図 8）. なお再現率とは，正解語数に対する絞り込んだ自立語中の正解語数の割合, 精度とは, 絞り込んだ自立語数に対する絞り込んだ自立語中の正解語数の割合を表す. 関連度の閾値は 0.05 である.

## 7　考察

記事分類に関しては，小カテゴリへの分類が平均 85%という再現率より，この分類システムの利用が有効であることがわかる. 加害者･被害者項目では，形態素解析による人名取得失敗が合わせて 23 件であったので，人名抽出が正しく行われれば正答率が上がると思われる. 自立語項目での自立語と知識の関連度計算に関しては，関連度の取り方で最大値を用いたが，平均値で評価したところ F 値は 0.69 とほぼ等しい評価結果が得られた.

## 8　おわりに

提案した方式により，時間，場所，加害者，被害者項目で正答率の平均 89%，自立語項目で F 値から 0.70 付近という評価が得られた. 今後の検討課題として，加害者・被害者項目において人名から「容疑者」などの表記を見つける操作を逆にし，人名抽出の精度を向上させる. 自立語項目では，小カテゴリの属性（例えば，交通という概念の属性）を小カテゴリの知識に追加し，不要語を閾値で切り捨て，新たな知識とすることを考える. システムの出力である重要キーワードの精度向上により，長文理解支援のための要約自動生成が期待される.


## 謝辞

本研究は文部科学省からの補助を受けた同志社大学の学術フロンティア研究プロジェクトにおける研究の一環として行った.


## 参考文献


[1]小島一秀，渡部広一，河岡司，“連想システムのための概念ベース構成法－属性信頼度の考え方に基づく属性重みの決定”，自然言語処理，Vol.9，No.5，pp.93-110，2002.

[2]渡部広一，奥村紀之，河岡司，“概念の意味属性と共起情報を用いた関連度計算方式”，自然言語処理，vol.13，No.1，pp.54-74, 2005.

[3]若月紀之，松田全弘，渡部広一，河岡司，“概念ベースと関連度計算を用いた新聞記事の分類”，情報処理学会研究報告，2005-NL-165，pp.67-72，2005.

[4]徳永健伸, 情報検索と言語処理, 東京大学出版会, 1999.

[5]Yahoo!ニュース，http://dailynews.yahoo.co.jp/fc/