

F-056

常識判断を用いた文書要約のための情報整理手法
Method for Organizing Information for Document Summaries
Using Common Sense Judgment

洞井 知彦†
Tomohiko Horai

吉村 枝里子†
Eriko Yoshimura

土屋 誠司†
Seiji Tsuchiya

渡部 広一†
Hirokazu Watabe

1. はじめに

近年、コンピュータやネットワークの発達に伴って、個人が扱える情報量は膨大なものとなってきている。よって、一個人が膨大な情報の中から、必要となる情報だけを抽出することは大変な作業となる。その手助けとなる手法のひとつとして、情報の重要な部分だけを厳選し提供する文書要約技術の研究開発が活発に行われている。

現在、文書要約技術に着目すると、自動要約ソフトといった複数の製品^{[1][2]}が既に存在する。これらの技術の多くは、文書に含まれる語の表記を手がかりに処理を行っている。しかし、語の表記が同じでも異なる意味を有したり(多義性)、同じ意味でも語の表記が異なったりする(表記ゆれ、同義語、類義語)問題がある。そのため、文書要約技術は広く一般的に使用される技術レベルには達していないのが現状である。

そこで本研究では、事件記事を対象に、文章の内容ならびにその重要度合いを考慮して情報を階層的に整理、分類する情報整理手法を提案する。文章の内容を理解するためには、語が有する意味特徴を語と重みで表現する概念ベース^[3]と、語と語の間にある関連の強さを同義性・類義性にとらわれず、数値として算出する関連度計算方式^[3]を用いる。

2. 提案する情報整理手法

本研究で提案する情報整理手法を説明する。記事文章を入力として、記事の内容や重要度に応じて要約レベル(6章参照)により記事本文を3段階に整理する。そして各分野に重要視されるキーワードの抽出を要約フレーム(7章参照)により実現している。これらの目的は、文章の内容とそれに含まれる単語の重要性という観点から、情報を分かりやすく整理することにある。また、本研究では記事を事件の内容に応じて、交通、火災、水難、強盗、暴力団、性犯罪、麻薬、鉄道、殺人の9つの分野に分類する。分類の種類は、読売新聞社、朝日新聞社が使用している分類コードおよびYahoo!ニュース^[4]の分類において共通する項目を基に決定した。

3. 関連技術

語(概念)が有する意味特徴を語(属性)と重みで表現する概念ベースと、語と語の間にある意味的な関連性を数値として算出する関連度計算方式、一般名詞の意味的用法を表す2710個の意味属性(ノード)の上位-下位関係、全体一部分関係を木構造で示したシソーラス^[5]を用いる。なお関連度計算には動的関連度計算方式^[6]を用いる。

3.1. 概念ベース

概念ベースとは、複数の国語辞書や新聞等から機械的に構築した語(概念)とその意味特徴を表す単語(属性)の集合からなる知識ベースであり、約12万語の概念が収録されている。概念は、ある語 A を属性 a_i と重み $w_i(>0)$ の対の集合として式2.1によって定義する。

$$A = \{(a_1, w_1), (a_2, w_2), \dots, (a_n, w_n)\} \quad (3.1)$$

ここで、属性 a_i を概念 A の一次属性と呼ぶ。また、属性 a_i も概念ベースに登録されている1つの概念である。従って、 a_i からも同様に属性を導くことができる。 a_i の属性 a_{ij} を概念 A の二次属性と呼ぶ。

3.2. シソーラス

シソーラスとは、一般名詞の意味的用法を表す約2700語の意味属性の上位-下位関係、全体一部分の関係を木構造で示したものであり、約13万語が登録されている。例えば「ビール」の上位は「酒」となる(図1)。

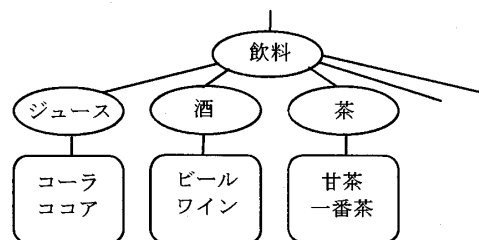


図1 シソーラスの構成(一部)

3.3. 関連度計算方式

関連度計算方式とは、概念ベースに定義された語と語の関連の強さを、同義性、類似性に関わらず定量化する手法である。

関連度は、0以上1以下の連続的な実数で表され、概念同士の関連が大きいほど関連度は高くなる。関連度は、それぞれの概念を二次属性まで展開し、その重みを利用した計算によって最適な一次属性の組み合わせを求め、それらが一致する属性の重みを評価することで算出する。表1に、概念「自動車」における関連度計算の例を示す。

表1 関連度計算の例

基準概念	対象概念	関連度
自動車	車	0.919
	電車	0.781
	鉛筆	0.025

†同志社大学理工学部
Faculty of Science and Technology, Doshisha University

‡同志社大学大学院工学研究科
Graduate School of Engineering, Doshisha University

3.4. TF・IDF

TF・IDF法^[7]とは、語の頻度と網羅性に基づいた重み付け手法である。TFはある文書に出現する索引語の頻度を表す尺度である。IDFはある索引語が全文書中のどれくらいの文書に出現するかという特定性を表す尺度である。なお、 N を検索対象となる文書集合中の全文書数、 $df(t)$ を索引語 t が出現する文書数とする。このとき、IDFは式2.2で定義される。

$$idf(t) = \log \frac{N}{df(t)} + 1 \quad (3.2)$$

3.5. 常識判断システム

常識判断システムとは、人間が普段、意識的または無意識のうちに用いている「常識的判断力」をコンピュータに持たせるためのシステムである。人間が日常的に用いている常識には様々なものがあり、例えば、「語と語の間にある関係」に関する常識、季節や時期などの「時間」に関する常識、学校や病院など「場所」に関する常識、大きさや重さなどの「量」に関する常識、暑い・騒がしい・美味しい・美しいといった「感覚や知覚」に関する常識、嬉しい・悲しいといった「感情」に関する常識などを挙げることができ、これらが円滑なコミュニケーションにおいて重要な役割を果たしている。

本研究では、常識判断システムの中でも場所判断システムと感覚・知覚判断システムを利用する。

場所判断システム^[8]では、ある名詞が場所に関する語であるか否かを判断し、場所である場合にはその場所に存在する人や物とその場所で行われる事象を想起することができる。

感覚・知覚判断システム^[9]では、ある語に対して人間が常識的に抱く印象を形容詞・形容動詞の形で出力させることができる。

4. 要約レベル

要約レベルとは、記事文章の内容や重要度に応じ、表記情報を元にして「発生」、「対処」、「説明」の3段階に情報を整理する手法である。

項目「発生」とは、いつ、どこで、どのような事件が起きたかを示した文章である。事件記事中では、「発生」、「起き」、「出来」という単語が使用されると共に記事の書き出しに主として記載されていることが多く、この3語のうちのいずれかを文中に含んでいる文を項目「発生」と分類する。

項目「対処」とは、発生した事件に対して警察や目撃者がどのような対応を取ったかを示した文章である。事件記事中では、「捜査」、「逮捕」、「通報」、「指名」、「手配」などの単語が使用されることが多く、この5語のうちのいずれかを文中に含んでいる文を項目「対処」と分類する。

項目「説明」とは、発生した事件の時系列や「発生」、「対処」をより詳細に記載している文章である。事件記事中では、「調べ」などの単語が使用されることが多く、「調べ」という単語を文中に含んでいる文や、項目「発

生」や項目「対処」のどちらにも分類されなかった文を項目「調べ」と分類する。

本研究では、上述した処理により、各文章を各段階に分類している。このように記事文章を3段階に情報整理することにより、記事中の語を抽出した際、同じ語であっても、「発生」、「対処」、「説明」のどのレベルの文から抽出された語であるかということ考慮することにより、その語が持つ微妙な意味の違いを理解することが可能になる。

5. 要約フレーム

要約フレームとは、文章の内容に着目して情報を階層的に整理、分類する体系化手法である。本研究では、要約フレームを用いて記事中で重要となるキーワードの抽出を行う。事件記事における要約フレームのキーワード項目は、「時間」、「場所」、「加害者」、「被害者」、「被害」、「自立語」の6項目を使用する。表2に、「4月2日、公園で、AがBをナイフで刺し殺害した。」という事件記事の例文に対してキーワード抽出を行った結果を示す。

表2 事件記事における要約フレームの例

時間	場所	加害者	被害者	被害	自立語
4月3日	公園	A	B	殺害	ナイフ

5.1. 共通項目

事件記事に限らず、全分野の記事に共通して重要であるキーワード項目として「時間」、「場所」、「自立語」の3項目を使用する。

5.1.1. 項目「時間」

項目「時間」は、記事に記載されている内容が発生した日付、時間である。記事本文に対して形態素解析を行うと時間表現特有の品詞列が出力される。本研究で使った形態素解析ソフト茶筌^[10]では、「名詞一数+名詞一接尾一助数詞」という品詞列が出力される。この特有の品詞列を手がかりに時間表現を抽出する。

5.1.2. 項目「場所」

項目「場所」は、記事に記載されている内容が発生した場所である。項目「時間」の処理と同様に、記事本文に対して形態素解析を行うと場所表現特有の品詞列が出力される。本研究で使った茶筌では、「名詞一固有有名詞一地域一一般」または「名詞一接尾一地域」が出力される。これらの前後にある単語の品詞が人名に関するものでない場合に項目「場所」として格納する。さらに、3.5節で説明した場所判断システムによる処理を行い、場所表現に関連する物事なども項目「場所」として格納する。また、「現場」や「県警」など、場所を示す語ではあるが事件が発生した場所の説明にはならないような語は、どの分野の事件記事中にも比較的多く出現するため、索引語の特定性を表すIDFを用い、実験により求めた閾値以上となる単語のみを抽出することにより、上記のような不適切な語を取り除くという処理も行っている。

5.1.3. 項目「自立語」

項目「自立語」は、上記で説明した各項目と次の7.2節で説明する分野別項目に格納される語以外の単語（名詞）で記事内容を表現する際に重要となる単語の集合である。主な例として、事件発生の原因、被害状況、事後の対処などが挙げられる。表3に示した分野ごとに付与された各分野を特徴付ける10語の単語と項目「自立語」の候補になる単語との関連度を関連度計算方式により算出し、関連度の最大値が実験により求めた閾値以上である単語を項目「自立語」として格納する。

表3 事件記事の各分野を特徴付ける単語の例

分野分類	交通	火災	水難	強盗
各分野に 付与された 単語 (10語)	交通	火災	水難	強盗
	道	火傷	転覆	盗み
	事故	放火	海	強盗
	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

5.2. 分野別項目

事件記事に特有のキーワード項目として「加害者」、「被害者」、「被害」の3項目を使用する。

5.2.1. 項目「加害者」・項目「被害者」

項目「加害者」、「被害者」は、記事に記載された加害・被害の対象となる人物や組織である。

加害者・被害者の対象が人名の場合、記事本文に対して形態素解析を行い、人名を特定し、その後存在する助詞から主格、目的格を判断する。人名を形態素解析すると、品詞について「名詞-固有名詞-人名-姓」、「名詞-固有名詞-人名-名」という結果が得られる。記事を形態素解析したとき、これらの品詞が存在した場合、その語を人名として抽出する。特に「名詞-固有名詞-人名-姓+名詞-固有名詞-人名-名」となった場合は、姓名と繋げて人名として抽出する。「名詞-固有名詞-人名-名」の前の語の品詞が「名詞-固有名詞-地域-一般」となっている場合にも、場所とするのではなく、姓名と繋げて人名とする。

次に、抽出した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」という表記がある場合には、その人名を加害者項目に格納する。人名の後ろに「さん」表記がある場合には、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から人名の係り受け関係を考慮して、加害者・被害者の判定を行う。

係り受け解析システム南瓜^[11]による構文解析では、文章を文節ごとに区切り、語の係り受けを知ることができる。そこで記事内の人名に対して、文節と係り受け関係より助詞、動詞、受身について判断し、加害者・被害者の判定を行う。助詞では、人名が書かれた文節に対し主格（が、は、も）、目的格（を、に）を判定する。また

助詞、「の」「と」に対しては、文における次の文節の助詞を用いてこの判定を行う。それ以外の助詞は判定を行わない。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」「逃走」といった語を24語、後者に「死亡」「逮捕」といった語を7語ずつ用意した。これを利用して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。

加害者・被害者の対象が物または組織の場合、その名称を特定することができないため、構文解析（係り受け解析）を行い、述語から主語を導くことで加害者・被害者の対象を判断する。

5.2.2. 項目「被害」

項目「被害」は、記事に記載されている被害の種類や大きさを表す語である。例えば「死亡」、「全焼」など、項目「被害」のキーワードになり得る語は感覚・知覚判断システムにかけると「つらい」、「危ない」、「悪質な」、「強引な」という4語のうちの最低1語を知覚語として出力するという傾向が見られる。よって記事本文に対して形態素解析を行い、名詞である語を3.5節で説明した感覚・知覚判断システムにかけ、項目「被害」のキーワードがとる知覚語として特徴的な4語のうちいずれかが出力されればその語を項目「被害」に格納するという処理を行う。また、例えば「自動車」には「危ない」、「懲役」にも「つらい」という知覚語が出力されるが、シソーラスで親ノードが「具体」、「犯罪」、「刑」である語を取り除くという処理を追加することにより、これらの語は項目「被害」のキーワードとして抽出しないようにしている。

6. 情報整理手法の評価

本研究で提案した情報整理手法の評価実験を行った。

6.1. 要約レベルについての評価

記事文章の内容の重要度合いから「発生」、「対処」、「説明」の3種類に整理する精度を評価した。無作為に選んだ20件の事件記事を評価データとして使用し、被験者3名のうち全員の意見が一致した場合を「正解（○）」、2名の意見が一致した場合を「誤りではない（△）」として評価した。結果を図3に示す。

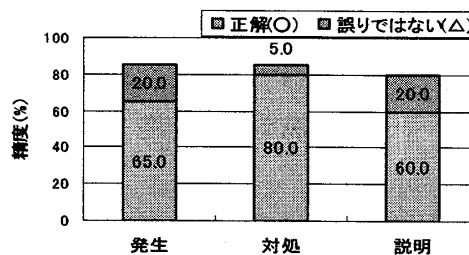


図3 要約レベルの評価結果

「誤りではない」までを正解とした場合、全項目の精度が 80%を超えており、本研究で提案した手法を用いることにより、事件記事における情報の階層化が十分に可能であることがわかる。

6.2. 要約フレームについての評価

事件記事の内容を表現する際に重要となるキーワードを抽出し、各項目に分類する精度と再現率を評価した。評価実験では、各分野につき 11 件ずつ計 99 件の事件記事の評価データとして使用した。内容を表現するキーワードとして重要とみなす判断は人によって揺らぎが生じるため、本実験では 3 名の被験者のうち 2 名以上の意見が一致したものを正解として評価した。

「完全一致」とはキーワードが完全に一致したもの、「部分一致」とはキーワードの表記が部分的に一致しており、意味的にもほぼ同様のものを表すと判断できるものを指す。要約フレームについての評価結果を「完全一致」、「部分一致」の場合に分けて、それぞれ図 4、図 5 に示す。

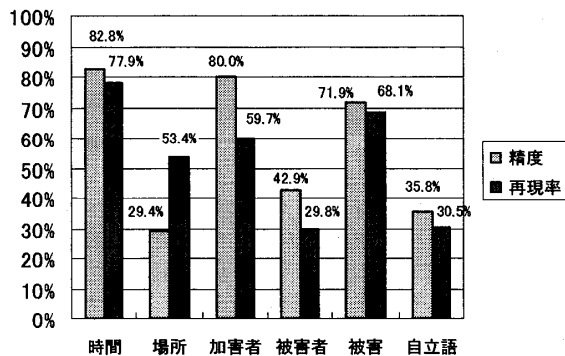


図4 キーワード抽出の評価結果 (完全一致)

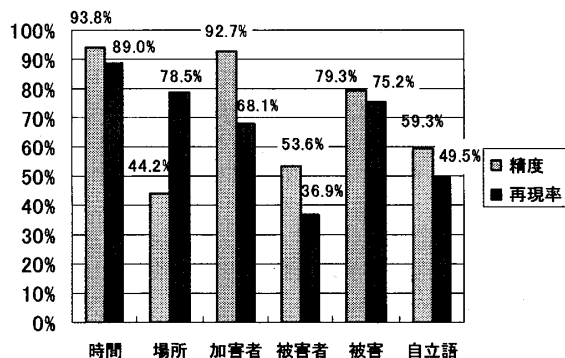


図5 キーワード抽出の評価結果 (部分一致)

項目「時間」、項目「被害」に関しては非常に高い精度、再現率でキーワードの抽出・分類ができており、本研究で提案した手法が有効であるといえる。項目「場所」や項目「加害者」、項目「被害者」などの精度や再

現率が低い項目に関しては、人名辞典などの固有表現を適切に抽出できる形態素解析手法を導入することによって、完全一致による評価値を部分一致による評価値へ近づけることが可能になると考えられる。

7. おわりに

本研究では、文章の内容と重要度合いに着目して情報を階層的に整理、分類する体系である情報整理手法を提案した。また、文章の内容を理解するために、その語が持っている意味特徴を、語と重みで表現する概念ベースと、語と語の間にある意味的な関連性を数値として算出する関連度計算方式、語間の上位-下位関係、全体-部分関係を木構造で表したシソーラス、場所判断システム、感覚・知覚判断システムを利用する手法を提案した。要約レベルによる分類は、「誤りではない(△)」まで含めた精度が 83.3% (「正解(○)」の精度は 68.3%)、要約フレームを用いたキーワード抽出においては、全体として、「部分一致」まで含めた精度が 60.8%、再現率が 71.4%という結果が得られた(「完全一致」では精度が 47.2%、再現率が 56.0%)。

階層的に情報を整理することにより文書要約では、字数制限など必要とされる情報のみを容易に抽出することができ、また、情報検索では検索対象を絞り込むことで検索の高速化を質問応答システムでは TPO に合わせた受け答えを実現することが期待できる。

謝辞

本研究の一部は、科学研究費補助金(若手研究(B) 21700241)の補助を受けて行った。

参考文献

- [1] 渡部広一, 河岡司, “常識判断のための概念間の関連度評価モデル”, 自然言語処理, Vol.8, No.2, pp.39-54, 2001.
- [2] ソースネクスト・ドットコム/文書作成ソフト/ズバリ要約, <http://www.sourcenext.com/products/youyaku/>
- [3] 製品情報[CB Summarizer], <http://www.justsystems.com/jp/km/product/cb104.html>
- [4] Yahoo!ニュース, http://dailynews.yahoo.co.jp/fc/domestic/toppage/crime_1.html
- [5] NTT コミュニケーション科学研究所, “日本語彙関係”, 岩波書店, 1997.
- [6] 荒木孝允, 奥村紀之, 渡部広一, 河岡司, “比較対象概念の共通属性を重視する動的関連度計算方式”, 同志社大理工学研究報告, Vol.48-3, pp.14-24, 2007.
- [7] 徳永健伸, “言語処理と計算 5 情報検索と言語処理”, 東京大学出版会, 1999.
- [8] 杉本二郎, 渡部広一, 河岡司, “概念ベースを用いた常識場所判断システムの構築”, 情報処理学会自然言語処理研究会資料, 2003-NL-153, pp.81-88, 2003.
- [9] A. Horiguchi, S. Tsuchiya, K. Kojima, H. Watabe, T. Kawaoka, “Constructing a Sensuous Judgment System Based on Conceptual Processing”, Computational Linguistics and Intelligent Text Processing (Proc. of CICLing-2002), Springer, pp.86-95, 2002.
- [10] 松本裕治, 北内啓, 山下達雄, 平野善隆, 今一修, 今村友明, “日本語形態素解析システム『茶筌』version1.0 使用説明書”, NAIST Technical Report, NAIST-IS-TR97007, 1997.
- [11] CaboCha/南瓜: Yet Another Japanese Dependency Structure Analyzer, <http://chasen.org/~taku/software/cabocha/>