情 報 学 基 礎 61−12
自 然 言 語 処 理 142−12
（2001. 3. 5）

# 単語の中心性に基づくテキスト自動要約システム

石井弘志　林日華　古郡廷治

電気通信大学　情報工学科
{ishii-h,lin}@phaeton.cs.uec.ac.jp, furugori@cs.uec.ac.jp

概要

本稿では、単語の中心性に関する仮説を利用した文抽出による自動要約手法を提案する。本手法では、テキストの要点がテキスト上で中心的な単語が使われている文にあると仮定して要約文を抽出する。しかし、単純に重要文を並べた要約文には文間に結束性を欠くという難点がある。ここでは、要約に使う文を選択する際に、ある種の制限を設けることによって読みやすさの向上を図った。提案手法に基づく要約システムを実装し、生成された要約文を「読みやすさ」と「内容の適切さ」に関し、5点満点で評価したところ、その平均で前者は3.50点、後者は3.36点の値を得た。

キーワード: 要約、文抽出、単語の中心性、選択制限

# An automatic text summarization system based on the centrality of word roles in sentences

Hiroshi Ishii, Rihua Lin, and Teiji Furugori

Department of Computer Science
The University of Electro-Communications

Abstract

This paper presents a system for automatic text summarization based on the roles various noun words play in sentences. We calculate the importance score of each sentence by the values attached to the role of each noun in it and extract sentences with higher scores assigned. A summary by sentence extraction sometimes lacks coherence. We use an algorithm so that the sentences selected are less independent each other syntactically and semantically. An experimental result shows that the summaries produced by our method are reasonable ones both in their readability and suitability.

Keyword: summarization, sentence extraction, word roles, selectional restriction

## 1．はじめに

インターネットなどの普及によって、広範にわたり電子化された文書がつくられるようになった。それにつれ、氾濫する文書情報の中から、的確な情報を手短に得る手段が必要となっている。最近脚光を浴びている自動要約研究は、その要求に応える有効な手段の一つである[1][2][3]。

自動要約の研究では、実現の容易さから、主に表層的な情報に基づいて重要文の抽出を行う方法がとられてきた。文の重要度計算の際に用いる情報には、これまで単語の出現頻度、テキスト中の位置情報、タイトルの情報、文間の関係を解析したテキスト構造、手がかり表現、文あるいは単語間のつながりの情報、文間の類似性の情報などがある[1]。

本稿では、文中およびテキスト中での単語の役割の軽重（中心性）に基づき、文中の単語に重み付けを行い、文の重要度を計ることによって、文を抽出する要約文作成手法を提案する。また、それを実装したシステムによる要約文生成の実験結果を報告する。

## 2．要約手法とシステムの概要

単語の重み付け手法に、単語の出現頻度に基づく tf*idf 法がある[4]。tf*idf 法では、重要度を計算する指標として単語の出現回数を使っている。しかし、テキストのそれぞれの出現箇所において、単語は主格や目的格などとなって異なる役割を担っている。したがって、その役割の如何により、テキスト内での重要性が異なると考えられる。本研究では、単語の中心性に着目し、次の手順で要約を生成する。

（1）テキストを入力する。
（2）文の形態素解析および構文解析を行う。
（3）文中に出現する単語に中心性に基づく重み付けをする。
（4）単語の重要度から文の重要度の計算をする。
（5）重要文をもとに、選択制限を加味し、要約文を出力する。

本手法では、テキスト上で中心をなす単語をより多く含む文を重要文とする。ただし、単純に重要文を並べた要約文は文間の結束性を欠く場合が多い。そこで、要約文の読みやすさの向上を図るため、文選択上の制限を行ったうえで、最終的な要約文を生成する。

## 3．単語の中心性

単語の役割の軽重、ここでいう中心性とは、主に照応解析で用いられるセンター理論の概念である。センター理論では、談話の中心となる要素が格情報によって判別されると仮定している。

亀山[5]は、 文内の指示対象（文の要素）の中心性に関し次の仮説を設けている。

> 発話文は、以下の順序で指示対象の中心性を決める：話題格＞主格＞目的格＞その他

本研究では、この仮説をもとに、単語（名詞）と結合している表層的な助詞の情報によって、次の順序をもった単語の中心性を仮定する。

> （ハ、モ等の副助詞）＞ガ＞ヲ＞（ガ、ヲ以外の格助詞）

さらに、本研究では複文を処理する場合、主節に現れる語は従属節に現れる語よりも中心的であると仮定する。単語の中心性の順序例を次に示す。

例１：猫がネズミを捕まえた。
猫 ＞ ネズミ

例２：ネズミが猫に捕まえられた。
　　ネズミ ＞ 猫
例３：ネズミを捕まえた猫を知っている。
　　猫 ＞ ネズミ

## ４．単語の重要度の計算

本システムでは、テキスト中で中心性が高い語は重要であるという考えに基づき、単語の重要度の計算を行う。この計算には、３節で述べた単語のもつ格や、単語の属する節の情報を利用する（なお、テキストの内容を表す単語は主に名詞であることから、名詞のみに重要度を与える）。

本システムは、単語の重要度の計算に、ローカルな重要度とグローバルな重要度を使う。ローカルな重要度では、テキストの各出現箇所において、その単語がどれほどの重要さを持っているかを測る。グローバルな重要度では、テキスト全体を見たときに、その単語がどれほどの重要さを持っているかを測る。

次に単語のローカルな重要度とグローバルな重要度の計算方法を述べる。

**単語のローカルな重要度**　まず、構文解析器KNP[6]によって文を構文解析する。その結果から、単語の各出現箇所ごとに、次の式によって単語 w のローカルな重要度を計算する。

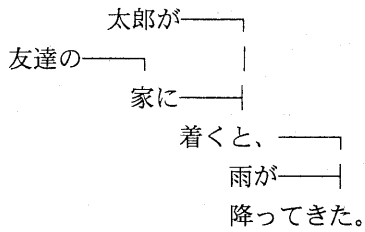
単語$w$のローカルな重要度
　＝$w$に後続する助詞による重要度
　＋$w$の文節の深さによる重要度

「w に後続する助詞による重要度」は、表層格による中心性の順序に基づき、中心性が高い格要素に、高い重要度を与えるようにする。なお、複合名詞および助詞「の」で接続された名詞句は１つの処理単位とする。また、「には」「よりも」のように、名詞に複数の助詞が接続する場合の重要度は最後尾の助詞によって決める。

「w の文節の深さによる重要度」は、KNPで文を解析した際の、文末の述語からの距離に基づいて計算する。主節に含まれる文節の深さは浅い位置となり、従属節に含まれる文節は比較的深い位置となるが、浅い位置にある格要素ほど文内での中心性が高いと仮定し、それに準じた重要度を与える。

図１は KNP での構文解析例である。

太郎が友達の家に着くと、雨が降ってきた。




重要度を与える単語：　太郎、友達の家、雨
重要度の順序：　雨 ＞ 太郎 ＞ 友達の家

図１　KNP での構文解析例

**単語のグローバルな重要度**　単語 w のグローバルな重要度は、w と類似性のある単語のローカルな重要度を結合して計測する。

単語$w$のグローバルな重要度
$$= \sum_{w'} (w'\text{のローカルな重要度} \times \text{類似度}(w, w'))$$

w'：テキスト中の w と類似性を持つ単語

グローバルな重要度を計算するため、複合語や、「の」で接続された名詞句を含めた単語間で、類似度を計算する必要がある。その際、同一指示性を持つ単語間の類似度は１とする。

以下に、単語間の類似性と同一指示性の判定方法を述べる。

**類似性の判定**　単語間の類似性は、

2単語間で同一の形態素が含まれる場合、単語間に類似性がある。

という仮説によって判定する。単語間の類似度は次式で計測する。

$$\text{類似度} = \frac{\text{2単語間での同一形態素数}}{\text{2単語の形態素数の合計}}$$

ただし、この計算をするうえで、形態素数に助詞の「の」は含めないこととする。以下に類似度の計算例を示す。

例1：政治+改革, 政治
　　→　2/3
例2：日本+映画, アメリカ+映画
　　→　2/4

**人物に関する同一性の判定**　人物（品詞に「人名」を含む単語とする）を指示する同一表現には、次のような場合がある。（以下、形態素の区切りを+で表す）。

例1：鈴木+太郎（姓+名），鈴木（姓），太郎（名）
例2：細川+護熙+首相（姓+名+役職），細川+首相（姓+役職），首相（役職）

このような表現では、形態素単位で次の（1）から（3）のいずれかのマッチングで同一指示を判定する。

（1）　前方部分一致
　　鈴木+太郎 ＝ 鈴木
（2）　後方部分一致
　　鈴木+太郎 ＝ 太郎
　　細川+護熙+首相 ＝ 首相
（3）　前方＋後方部分一致
　　細川+護熙+首相 ＝ 細川+首相

**人物以外の同一性の判定**　人物以外（品詞に「人名」を含まない単語とする）を同一指示する表現には、次のような場合がある。

例1：政治+改革　改革
例2：日本+の+映画+界　日本+映画+界
例3：フェニックス+計画　フ+計画

人物以外の表現の場合、前方部分一致および、前方＋後方部分一致では、同一指示とならないことが多い（例：「政治＋改革」と「政治」、「感染+防止+力」と「感染+力」）。そのため、人物以外を指示する表現は、後方部分一致によって同一指示性の判定をする。これによって例1が同一指示とみなされる。ただし、同一テキストに「日本+映画」、「アメリカ+映画」、「映画」が出現し、「映画」で双方を含む一般的な「映画」を指すことがある。このことを考慮し、同一指示と判定するのは、後方部分一致が一つのときに限る。

例2のように「の」で形態素が接続された表現は、「の」を削除してできる複合語と同一とする。例3には文字の省略がある。このような場合に対処するため、「形態素は、その形態素の始めの1文字に置き換えることができる」ものとする。

## 5．文の重要度の計算

文の重要度は、単語のグローバルな重要度とローカルな重要度に基づいて計算する。これは、テキスト中で重要となる単語が、中心的に使用されている文を重要な文とするためである。

文の重要度を計算する際には、文に現れる各単語の重要度を足し合わせたものを文の重要度とする方法が考えられる。しかし、この方法では、多くの述語を含む複文は、格要素を多く

持つため、文の重要度が必然的に高くなる傾向がある。これを回避するため、ここでは、文に含まれる述語ごと、つまり節単位で重要度を決定する。

節の重要度の計算は次式によって行う。

$$\text{節の重要度} = \sum_{w} \sqrt{w\text{のローカルな重要度}} \times w\text{のグローバルな重要度}$$

w :節中の単語

ただし、名詞句に係る節は係り先名詞への従属性が高いため、係り先の節と切り離して重要度を計算するのは問題である。そこで、名詞句に係る節の重要度は、係り先の節へ結合する（加算）。名詞句に係らない節の重要度は独立に計算する。この修正を経た後、文中の節の重要度のなかで、最も高い値をとって文全体の重要度とする。

以下に文の重要度を計算するアルゴリズムを示す。

1. 単語のローカルな重要度を計算する（単語の格、文節の深さにより、単語の各出現箇所での重要度を計算する）。
2. 単語のグローバルな重要度を計算する（単語のローカルな重要度を結合し、テキスト全体での単語の重要度を計算する）。
3. （名詞句に係らない）節ごとに、節中の単語の、ローカルな重要度とグローバルな重要度をもとに節の重要度を計算する。
4. 文中の節の重要度のなかで最も高い値を文の重要度とする。

## 6．要約文に採用する文の選択

本手法では、重要度の高い文をテキストの要点と見なし、要約文として抽出する。ただし、重要度が高い文から順に採用を行うと、採用された文間の結束性が悪い場合が多く、要約文が読みづらくなる恐れがある。この問題を解消するため、文の選択に関し２つの処理を行う。

テキスト中の文には、古い情報と新しい情報がある[7]。文抽出による要約文が読みづらくなる一因は、抽出された文が、新しい情報（未知情報）のみからなり、古い情報（既知情報）を含まないため、唐突な文となってしまうことにある。しかし、この問題は、既知情報を含む文を採用すれば、ある程度、解決できると思われる。そこで、本手法では、抽出文中の未知情報が初出する文も採用することにする。

図２はこの操作の一例である。ここで、「残念ながら「あずさ」の振り子は・・」という文が採用されたとき、「あずさ」という単語が初出する、「昨年十二月デビューした・・」という文も採用する。この処理により、要約文内で「あずさ」という単語の解釈が明白になる。

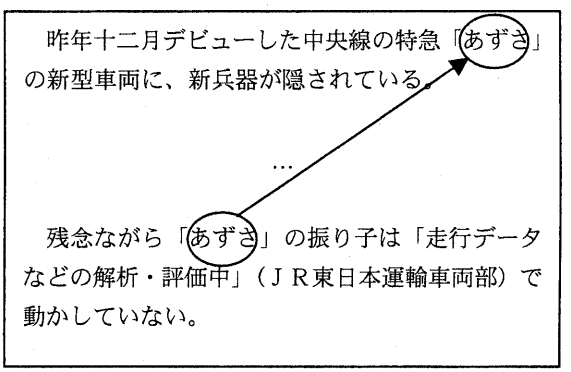



図２　未知情報が初出する文の採用

文抽出の要約文が読みづらくなる別の一因に、文頭に指示詞や接続詞（手がかり語）がくる場合がある。山本らの要約システムGREEN[8]は、このような場合や、主語が省略されている場合、テキスト中でその直前にある文も採用している。本手法では、指示詞や接続詞の種類により、その文を要約文とするか否かにつき、次の選択制限を設ける。

(a) 重要度の順位によって文を採用するとき

段落の途中に出現し、接続詞、指示詞を文頭に持つ文　→　同段落内で前に出現する文が採用されていなければ、採用順を１つ下げる

「例えば」のような例示の接続表現で始まる文　→　採用を行わず、重要度を０にする

(b) 未知情報が初出する文を採用するとき

段落の途中に出現し、接続詞、指示詞を文頭に持つ文　→　同段落内で前に出現する文が採用されていなければ、採用しない

「例えば」のような例示を示す接続表現で始まる文　→　採用しない

次に、要約文に採用する文の選択アルゴリズムを示す。

1. 重要度の高い方から１文を選ぶ。
2. 選択した文に手がかり語が含まれている場合、手がかり語処理(a)を行い、ステップ１に戻る。
3. 要約率を超えないならその文を採用する（超えるなら終了）。
4. 次の手順で未知情報が初出する文を採用する。
   - 4.1 ステップ３で採用した文に含まれている単語と、類似性のある単語がテキストで初めて出現する文を選ぶ（グローバルな重要度が高い単語順）。
   - 4.2 選択した文に手がかり語が含まれず、また、要約率を超えないならその文を採用する（超えるなら終了）。
   - 4.3 次の単語があればステップ4.1に戻る。
5. ステップ１に戻る。

なお、要約率は以下の式で計算する。システムはこの要約率を超えない範囲で文の選択をする。

$$\text{要約率} = \frac{\text{要約文の総文字数}}{\text{原文の総文字数}} \times 100[\%]$$

## ７．実験および評価

提案した要約手法に基づき、システムを実装して得た実験結果を以下に示す。ここでは、単語のローカルな重要度を計算する際の助詞による重要度および文節の深さによる重要度として、表１、表２に示す値を使用した。

表１　助詞による重要度

| 助詞 | 重要度 |
|---|---|
| は　も　等の副助詞 | 4 |
| が | 3 |
| を | 2 |
| が　を　以外の格助詞 | 1 |

表２　文節の深さによる重要度

| KNPで解析した際の文節の深さ | 重要度 |
|---|---|
| 1 | 3 |
| 2 | 2 |
| 3以上 | 1 |

表３は、システムによって作られた要約文を「読みやすさ」、「内容の適切さ」において評価した結果である。ここでの評価点は、10名の被験者（大学院生および学部学生）に原文と本システムで要約した要約文（要約率40%）を提示し、それぞれに１点から５点（「非常に悪い」「悪い」「普通」「良い」「非常に良い」の順）を与えてもらったものである。なお、要約に用いたテキストは、CD-毎日新聞94年版[9]から抽出した科学コラム５記事である（付録に実際の要約の一例を、その原文とともにあげておく）。

表3　要約の評価

| | 読みやすさ | 内容の適切さ |
|---|---|---|
| 記事1 | 3.90 | 3.80 |
| 記事2 | 3.30 | 3.60 |
| 記事3 | 4.20 | 3.40 |
| 記事4 | 2.70 | 3.20 |
| 記事5 | 3.40 | 2.80 |
| 平均 | 3.50 | 3.36 |

要約の評価結果では、「読みやすさ」が平均3.50点、「内容の適切さ」が平均3.36点である。このことは、本手法による要約が概して良好なものであることを示している。

しかし、評価点の悪い要約を精査してみると、問題点もある。その一つは、「は」などの副助詞が付く単語を重要性の高い単語としているが、これが必ずしもテキストで重要な概念を表さない場合もあるということである。今回、要約の対象としたテキストでは発言者・行為者としての人物や機関に「は」が使われていたが、科学コラムではこれらの語の実際の重要性は高くないはずである。

また、単語の前提説明は必ずしも必要ではないため、未知情報が初出する文の選択が適切ではない場合もみられる。これに関しては、文の選択に関するアルゴリズムのさらなる検討が必要である。

## 8．おわりに

本稿では、単語の中心性を仮定し、それに基づき文の重要度を計算して行う自動要約の手法を提案した。一般に、テキストの自動要約を行うには、要点箇所の特定と、要約文の結束性を向上させる処理が必要である。今回提案した要約手法では、「要点」を中心性の高い単語が使われている文として近似し、さらに、文の採用アルゴリズムで要約文の結束性の向上を図っている。

実験結果は、おおむね手法の適切性を指示していると思われるが、問題点、改良の余地も残されている。要約文に文間の結束性を保たせるためには、照応問題の解決や、文間の因果関係などにも注目する必要がある。一方、手短な要約にするためには、前処理段階か後処理段階、あるいはその双方で不要箇所の削除を行うようなことも考慮に値するものと思われる。

## 付録（要約例：記事1）

### 原文（717文字）

直径約八メートルの巨大な丸いガラス板が、米国のコーニング社に出現した＝写真。丸窓にはめこもうというのではない。日本がハワイ島マウナケア山に建設中の大型光学赤外線望遠鏡「すばる」の心臓部となる主鏡材料ができあがったのだ。

完成の暁には東京から富士山の頂上のテニスボールが見分けられるほどで、「太陽系外の惑星探しにも威力を発揮する」と大きな期待がかけられるが、その高性能を実現するための秘密がいくつかある。

まずは、八メートルと世界最大級の口径。実は十メートル級がすでに存在するが、これは三十六枚の鏡を組み合わせた分割鏡で、一枚鏡の「すばる」とはちょっと違う。「押したり引いたりして合わせ込む分割鏡と違って、一枚鏡は分子の力で結びついているので精度を上げやすい」と国立天文台の唐牛宏助教授は説明する。

さらに、鏡が自重でたわまない程度の厚さが必要という常識を覆し、厚さ約二〇センチの薄型にして裏から支えることにした。

主鏡材に蜂（はち）の巣のような模様が見えるのは、六角形の特殊なガラス板四十四枚を、熱変形が最小になるように計算して配置し、融合したからだ。このあと板を球面にし、三年かけて研磨。鏡面の凸凹を〇・一ミクロン以下に仕上げる。

「東京ディズニーランドから高尾山まで入る平地を鏡にたとえると、世田谷区と杉並区を一ミリの誤差で地ならしするようなもの。まさに職人芸です」と唐牛さん。これを二百四十数点で支え、重力、熱などの影響に対応してコンピューター制御し、鏡面の精度を保つ。

コンピューター制御は日本の技術だが、鏡の製作は米国まかせ。基礎科学用の大物単品には日本

の企業は手を出さない、そんな現状も巨大鏡材の裏に隠されているようだ。

**要約文（254 文字、要約率 35.4%）**

直径約八メートルの巨大な丸いガラス板が、米国のコーニング社に出現した＝写真。
日本がハワイ島マウナケア山に建設中の大型光学赤外線望遠鏡「すばる」の心臓部となる主鏡材料ができあがったのだ。
実は十メートル級がすでに存在するが、これは三十六枚の鏡を組み合わせた分割鏡で、一枚鏡の「すばる」とはちょっと違う。
さらに、鏡が自重でたわまない程度の厚さが必要という常識を覆し、厚さ約二〇センチの薄型にして裏から支えることにした。
基礎科学用の大物単品には日本の企業は手を出さない、そんな現状も巨大鏡材の裏に隠されているようだ。

## 参考文献


[1] 奥村学,難波英嗣：テキスト自動要約に関する研究動向, 自然言語処理, Vol.6, No.6, pp.1-26, 1999

[2] I. Mani, M. Maybury : Intelligent scalable text summarization, Universidad Nacional de Educacion a Distancia, Madrid, 1997

[3] I. Mani, M. Maybury (eds.) : Advances in Automatic Text summarization, MIT Press, London, 1999

[4] K. Zechner : Fast Generation of Abstracts from General Domain Text Corpora by Extracting Relevant Sentences, In Proc. of the 16th International Conference on Computational Linguistics, pp.986-989, 1996

[5] 田窪行則,西山佑司,三藤博,亀山恵,片桐恭弘：岩波講座 言語の科学 7 談話と文脈, 岩波書店, pp.101-107, 1999

[6] 黒橋禎夫：日本語構文解析システム KNP 2.0b6 使用説明書, 京都大学大学院 情報学研究科, http://pine.kuee.kyoto-u.ac.jp/nl-resource/knp.html, 1998

[7] H.H. Clark : Inferences in Comprehension, In LaBerge, D. and Samuels, S. J.(eds.): Basic Process in Reading: Perception and Comprehension, Erlbaum:NJ, pp.243-263, 1975

[8] 山本和英,増山繁,内藤昭三：文章内構造を複合的に利用した論説文要約システム GREEN, 自然言語処理, Vol.2, No.1, 1994

[9] CD-毎日新聞 94 年版, 毎日新聞社