

日本語テキストの畳み込み型要約のための 単語・文間の関連付け手法の提案

及川 中[†] 伊藤 久祥[†]

[†] 岩手県立大学大学院 ソフトウェア情報学研究科

〒020-0193 岩手県滝沢村滝沢字巣子 152-52

E-mail: †g231b004@edu.soft.iwate-pu.ac.jp, †hito@soft.iwate-pu.ac.jp

あらまし オンラインヘルプ等の読みにくさを解消し、ユーザに応じた動的なコンテンツの再構成を実現するため、日本語のニュース文を対象とし、ある文に含まれる名詞に対し、その名詞と他の文との相関を見出し、関連付けを生成することにより、日本語テキストの畳み込み型要約を行う手法を提案する。本稿では、被験者を使った実験を行った結果と、それを踏まえた関連付け生成規則を用いたシステムの試作について報告する。

キーワード 文章要約, 畳み込み型要約, 単語・文の関連付け

A proposal of detection method of relationship between words and sentences in fold up summarization for Japanese text

Ataru OIKAWA[†] and Hisayoshi ITO[†]

[†] Graduate School of Software and Informaton Science, Iwate Prefectural University

Takizawa Aza Sugo 152-52, Takizawa-mura, Iwate, 020-0193 Japan

E-mail: †g231b004@edu.soft.iwate-pu.ac.jp, †hito@soft.iwate-pu.ac.jp

Abstract The authors propose a detection method of relationship between words and sentences, using in fold up summarization for Japanese text. The authors report about the result of experiment and prototype system of automatic fold up summarization.

Key words summarization, fold up summarization, relationship between words and sentences

1. はじめに

近年のソフトウェアにおいて、テキストを使用した説明は未だに欠かせないものである。具体的には、オンラインヘルプ・オンラインマニュアル等である。しかし、現在のオンラインヘルプ等は、単一の静的コンテンツによって構成され、その内容の専門性の度合い等を考慮すると、必ずしも「誰にとっても読みやすい」とは言えず、ソフトウェアを使い始めた段階の初心者における壁になっている。

このような問題を解決する方法として、習熟度別に複数のコンテンツを用意する方法が考えられるが、これはコンテンツ作成者への負担が大きくなる。また、できるだけ詳しいコンテンツにするという方法もあるが、これは熟練者にとっては既知の事項が大半を占めてしまうため、本当に探している情報が埋もれてしまい、非常に読みづらくなる。これらのように、単一または複数の静的コンテンツを用意する方法では、万人にとって読みやすいとは言えないのが現状である。

そこで筆者らは、どのような習熟度のユーザに対しても適切なコンテンツの表示を行えるようにするため、テキストコンテンツをユーザの要求に応じて動的に再構成し表示するためのフレームワークの一部として、ある名詞とある文のを数値化し、それらの間に関連付けを生成することにより、文章を自動的に構造化する『畳み込み型要約』を提案する。

本稿では、畳み込み型要約の実現にあたって、被験者を使った名詞・文相互の関連付けを行う実験を行った結果と、それを踏まえた名詞と文の関連付けの自動生成規則を用いたシステムの試作について報告する。

2. 畳み込み型要約

文章要約の基本的な考え方として、現在利用されているものは以下の2種類に大別できる。[1]

- 抜粋
- アブストラクト

抜粋は不要と思われる部分を削除することによって実現され、

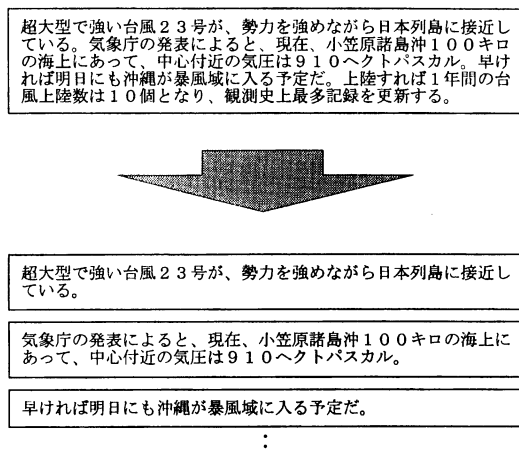


図1 文章→文への分解

Fig. 1 Decompose an article into sentences

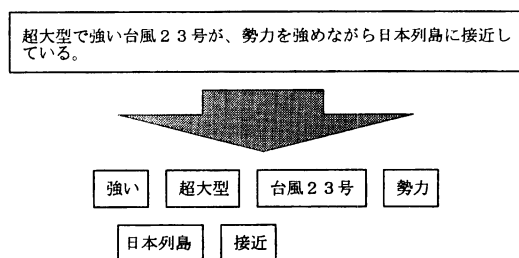


図2 文→名詞節への分解

Fig. 2 Decompose a sentence into noun clauses

アブストラクトは元の文章の意味を損なわないように、より短い一般的な表現に換言することによって実現される。

本稿において提案する『畳み込み型要約』は、これらの要約とは「一度短くした文章の任意の部分に元に戻すことができる」という点において異なる。

本稿における『畳み込み型要約』とは、以下に示す手順をもって文章を自動的に構造化する手法のことを指すものとする。

2.1 概要

(1) 文章の分解

入力された文章を、まず初めに句点を区切りとする文の単位に分解する。この際、釣括弧(「」『』など)で囲まれた部分に含まれる句点は分解の対象としない。(図1)

(2) 文の分解

次に、分解された文を形態素に分解し、品詞情報の分析により、名詞節のみを抽出する。(図2)

(3) 名詞節と他の文の関連付けの生成

抽出された各々の名詞節に対し、他の文との関連付けを以下の手順で行う。

- (a) 名詞節に含まれる全ての名詞を抽出する
- (b) ある他の文に含まれる全ての名詞を抽出する

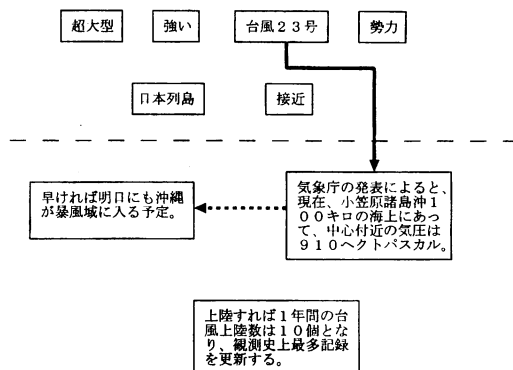


図3 名詞節・文の関連付け

Fig. 3 Detection of relationship between words and sentences

(c) 名詞節と他の文についてスコアを取得し、スコアが最も高い名詞節と文に対して関連付けを生成する(図3, 実線矢印)

(d) (3)において関連付けが行われなかった文に対して、(3)で関連付けが行われた文に対し、文の分解・名詞節と他の文の関連付けの生成を繰り返す(図3, 破線矢印)

名詞節と文の関連を調べるために、同一文章中での名詞の共起関係が使用できるのではないかと考え、共起関係を格納するデータベースを作成することにした。次節では、文章データの取得・共起関係データベース作成について述べる。

3. データの取得

畳み込み型要約の自動化を実現するため、本稿では、以下に挙げる理由から、日本語のニュース文を対象として単語・文の関連付けを行うこととした。

- 種類・量が豊富である
- 文法が整っている
- 多くの記事で、第一文が全体の大まかな内容を説明している

3.1 ニュース記事の取得

ニュース記事の取得はPCにVine Linux 3.0をインストールしたサーバ上で、ニュースを配信しているWebサイトからcronを利用して1時間毎に記事一覧のページからURLを抽出し、新たに投稿された記事のHTMLページをダウンロードした。さらに、Perlスクリプトで記事からHTMLタグを取り除くことによりテキスト形式に変換した。

3.2 共起関係データベースの作成

3.1節で示したサーバ上で、形態素解析システムに茶筌[2]、DBMSにPostgreSQL 7.4.5[3]を使用し、下記の条件で共起関係データベースを作成した。

あるニュース記事を形態素解析した結果において、

- (1) 以下の(a)~(d)に示す品詞分類に該当する単語のリストを作成する(同一単語の重複は認めず、1つと数える)

- (a) 名詞-一般
- (b) 名詞-固有名詞
- (c) 名詞-サ変接続
- (d) 名詞-動詞非自立的

(2) (1) のリストからの全ての 2 単語の組み合わせについて、データベース上のカウントを 1 増やす

(3) (1)~(2) の作業を記事の件数分繰り返す

作業の結果、ニュース記事 14,529 件に対する共起関係のデータベースが作成された。このデータベースに基づいた単語・文間の関連のスコアを算出する方法として、以下の方法を考案した。

(1) ある文の中に含まれる名詞と、他の一文に含まれる名詞全てについて、データベースよりカウントを取得する

(2) (1) の中で最大の値を、その名詞と文のスコアとする

4. 実験

共起関係データベースの作成と並行して、人間が自然に感じることのできる名詞節と文の関連付けの方法を明らかにするため、被験者を使った実験を行った。実験は、無作為に抽出した 20 件のニュース記事に対して、

(1) 第一文中のある名詞節と関連があると思われる文

(2) ある文と関連があると思われる第一文中の名詞節を記入するという内容で行った。「どの候補とも関連がない」という選択肢も設けた。以上のことにより、

- 他の文/名詞節との関連が存在するかどうか
- 回答がどの程度同一のものに集中するか
- 関連付けられた名詞節と文双方に含まれる名詞には共起関係において特徴があるか

という点について調査を行った。

被験者は 18~35 歳の学生 34 名、教員 1 名 (男性 23 名、女性 12 名) を対象とした。なお、被験者が普段新聞や Web サイト等で文字のニュースを閲覧する頻度を調査したところ、表 1 のような回答が得られた。

実験の結果、

(1) 第一文中のある名詞節と関連があると思われる文

に関しては、被験者のニュース閲覧頻度に関わらず同一の回答に集中する名詞節が見られた一方、どの文とも関係がないと判断された名詞節も多かった。

(2) ある文と関連があると思われる第一文中の名詞節に関しては、どの名詞節とも関係がないと判断された文は少なかったものの、回答がばらつく傾向が見られた。

4.1 実験から得られた知見

実験の結果について、

(1) 第一文中のある名詞節と関連があると思われる文

に注目し、同一の文に回答が集中した名詞節について、共起関係のスコアから実験と同じ解が導き出せるか、詳細に内容を検証した。

実験において被験者の回答が集中した名詞節について、

- ほとんどの被験者が関連が深いと回答した文

表 1 被験者のニュース閲覧頻度の調査結果

Table 1 Frequency of reading news in text

回答	人数
よく読む	4
たまに読む	17
ほとんど読まない	9
まったく読まない	5

図 4 実験用紙

Fig. 4 Experiment sheet

- それ以外の文

に分け、それぞれの全ての組み合わせに対して共起関係のスコアを抽出し、前者と後者で特徴的差異があるかどうかを確認した。さらに、3 節で示したスコアを算出し、実験結果と同じ文への関連付けが行われるかを検証した。

その結果、共起関係のスコアは、一般的によく使われる名詞に対して高いスコアを示し、実験結果と異なる文への関連付けが行われることが多く、双方の関連付けの間に共通性は認められなかった。共起関係において高いスコアを示した文は文脈とは無関係になりがちであり、名詞節と文の関連付けにおいて共起関係を使用することは難しいということが分かった。

また、ニュースの閲覧頻度に関する質問 (表 1) の回答と実験の回答には相関が見られなかった。

5. 単語・文の関連付け手法

実験結果を踏まえ、共起関係とは別の観点で名詞節と文の関連付けを行う手法を検討した。被験者の回答を個々に読んだ結果、第一文における主語に相当する名詞節において、回答が集中する傾向が見られた。したがって、主語からのつながりを見出すことで関連性を導出することができるのではないかと考え、

超大型で強い台風23号は、勢力を強めながら日本列島に接近している。二

図 5 試験システム (1)

Fig. 5 Experimental system

超大型で強い台風23号は、勢力を強めながら日本列島に接近している。
気象庁の発表によると、現在、小笠原諸島沖100キロの海上にあって、
中心付近の気圧は910ヘクトパスカル。二

図 6 試験システム (2)

Fig. 6 Experimental system(2)

以下の手順によって単語・文の関連性を導く手法を考案した。

- (1) 第一文の名詞節を抜き出し、記憶しておく
- (2) 第二文以降のそれぞれの文に対して、
 - ・ 主語が存在すれば抜き出す。存在しない場合は直前の文の主語と同一であると見なす
 - ・ 該当する文の主語と第一文の名詞節を比較し、一致すれば関連付けを生成する
- (3) どの名詞節とも関連づけられなかった文に対しては、補足的な情報であると見なし、名詞節→文という形での関連付けは行わない。

この手法を使用した折り畳み型要約の提示システムを、3.1 節で示したサーバ上で、Apache 1.3.31 を使用して実装した。Perl による CGI スクリプトで折り畳み型要約を行い、処理結果を JavaScript を使用した HTML に整形し、Web ブラウザから閲覧できるようにした (図 5, 図 6)。

6. ま と め

本稿において、日本語ニュース文に対する折り畳み型要約のための単語・文を自動的に関連付けする手法を提案した。主語のつながりを利用することで、人間にとって自然な関連付けを生成することができた。

しかし、現状では複雑な構文を持った記事に対しての提案手法の有効性の検証がなされていないため、今後はより多くのニュース文に対して提案手法の有効性を検証していく必要がある。また、ニュース文以外の日本語テキスト(オンラインヘルプ・オンラインマニュアル等)においても、有効な単語・文の関連付け手法を構築し、畳み込み型自動要約システムを完成させたい。

文 献

- [1] Inderjeet Mani, “自動要約,” 奥村学 他訳, 共立出版, 2003
- [2] “ChaSen’s Wiki”, <http://chasen.naist.jp/wiki/ChaSen/>
- [3] “PostgreSQL”, <http://www.postgresql.org/>
- [4] “Apache Software Foundation”, <http://www.apache.org/>