格文節ボックスに基づく新聞記事要約システムの開発

山森　一人[a)]・原口　隼翔[b)]・相川　勝[c)]

Development of Newspaper Article Summarization System Based on “Clause Box with a Case-marking Particle”

Kunihito YAMAMORI, Hayato HARAGUCHI, Masaru AIKAWA

Abstract

We develop a newspaper article summarization system. Our system summarizes newspaper articles based on grammatical rules. Our system firstly separates articles into some words by morphological analyzer. Secondly, the system combines these words into clauses. Thirdly, evaluation value is given for each clause from the grammatical points of view. Here we propose “Clause Box with a Case-marking Particle” called “Kaku-Bunsetsu Box” as one of the grammatical rules. “Kaku-Bunsetsu Box” is a pair of clauses that includes a case-marking particle and the next appeared clauses. Finally, system selects clauses to maximize the sum of evaluation values within the given number of characters, then a summary is assembled from these clauses. We solve the final process as 0-1 knapsack problem. We evaluate our system by “ROUGE-2” that is one of the evaluation index. Our system shows good performance as the same as the previous works using templates.

Keywords: Text Summarization,Sentence reduction,Case-Marking

1. はじめに

高度情報化が急激に進む近年、インターネットや携帯情報端末の急速な普及により、個人でも簡単に情報を送受信できるようになった。それに伴い、個人が受け取るテキスト情報の量は急激に増加した。寸暇を惜しんで活動する現代人にとって、受信した大量のテキスト情報をすべてじっくりと読むことは効率が悪い。必要な情報を取捨選択し、そこから速やかに主旨を理解するためテキスト情報の自動要約への需要が高まっている。

本研究では、次々と変化する社会情勢の動きを素早く、かつ広く配信している新聞記事の要約を扱う。

テキスト自動要約の既存手法として、山本ら[1)2)]は人間が作成した要約文をテンプレートとして用いる方法や、文章を文法的に解析し、様々な日本語の特徴を利用して要約文を作成する方法を提案している。

既存の要約文からテンプレートを作成する手法では、膨大な要約サンプルやコーパスが必要であり、それらを収集、管理するコストがかかる。また、人間が文章を要約する際にはテンプレートを使うことはない。つまり、テンプレートを使用して機械的に要約するよりも、文法的に文章を解釈し要約したほうがより人間らしい要約文が作成できると考えられる。

本研究の目的は、文法的な視点、特に文中で意味関係を表す「格助詞」を含む文節を基に新聞記事を要約する手法を開発することである。格助詞を含む文節に対し、「格文節ボックス」の考えを提案し、特定の格助詞を含む文節、および格文節ボックスに対して得点付けを行うことで、要約文を作成する。

2. 形態素解析

言語学では、意味の通じる最小の言語的な要素を形態素と呼ぶ。文章を文法的に正しく要約するためには、文章に含まれている形態素の品詞や活用形を知る必要がある。日本語の文章には明確な形態素の境界がなく、そのままではコンピュータが形態素の区切りを理解できない。そのため、あらかじめ文章を形態素に分ける処理、形態素解析が必要である。本研究では単語の素性が解析できる形態素解析器 MeCab[3)]を単語の素性を求めるために使用する。

形態素解析で得られる結果は文法的に「単語」と呼ばれ、文章をなす要素の中では最小の単位である。単語は文法的な処理を行うにあたり、短く扱いづらいため、提案するシステムでは処理単位を文節とする。

本研究では、文節を以下のように定義する。また形態素解析で得られた単語すべてを自立語か付属語のどちらかに分類する。

a) 工学教育研究部准教授

b) 情報システム工学科

c) 宮崎大学工学部教育研究支援技術センター技術職員

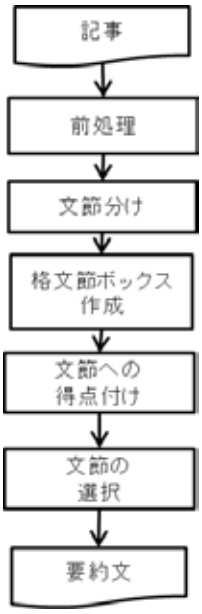


図 1 本システムの処理手順

- 文節：自立語が 1 つ、直後に付属語が 0 個以上連続している最長のまとまり
- 自立語：単体で文節をなす単語
- 付属語：単体で文節をなすことができず、自立語の直後に出現することで意味をなすことができる単語

3. 提案手法

本研究で提案する自動要約システムの処理の流れを図 1 に示す。入力された新聞記事に対して、文章を整形する前処理を行ったあと形態素解析を行い、その結果をもとに 2 章で示した定義に基づき文節分けを行う。次に、格文節ボックスを作成し、各文節への得点付けを行ったあと、得点の高い文節を選び出して要約文とする。

3.1. 前処理

入力された新聞記事について前処理を行う。
前処理とは、形態素解析を行う前に
文章を整形することである。具体的には、

- 全角英数字を半角英数字に変換、
- スペースの削除、
- カギ括弧内の単語を一つの固有名詞に変換、
- 丸括弧内の単語を削除、

の処理を行う。

政治経済について記述された記事が入力された場合を考慮し、2012 年度において現存する政党名を MeCab の辞書に登録する。これは政党名が助詞を含む場合、固有名詞として処理されるべきであるにもかかわらず、政党名を分割して文節が作成されるのを防ぐためである。

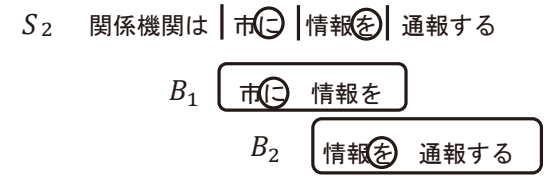


図 2 格文節ボックスの作成例

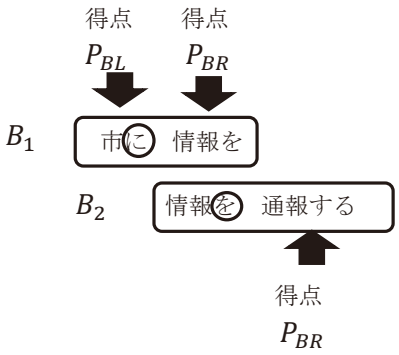


図 3 格文節ボックスへの得点付け

MeCab の処理の都合上、全角英数字は半角英数字に変換する前処理を行う。

3.2. 格文節ボックスへの点数付け

本研究では要約にあたり、格文節ボックスの導入を提案する。格文節ボックスとは、格助詞を含む文節と、直後に出現する文節を 1 つのまとまりとしたものである。

格文節ボックスで作成例を図 2 に示す。図 2 中で格助詞は丸で囲んでおり、文節の区切りを縦棒で示している。例文S_1に含まれる文節のうち、格助詞を含む文節は「市に」「情報を」の 2 つがある。したがって、格文節ボックスはB_1の「市に情報を」と、B_2の「情報を通報する」の 2 つが作成できる。

1 つの格文節ボックスとなった 2 つの文節に対して、前の文節に得点P_{BL}、後の文節にP_{BR}を与える。得点はユーザが任意に指定できる。また、図 3 の「情報を」の文節のように、ある文節が 2 つの格文節ボックスに属している場合は、重複して得点を与えない。

3.3. 文法格への点数付け

文法格とは、名詞を含む文節がどのような意味的な関係にあるかを示すものである。井上 [4]によると、
日本語の助詞「が」「に」「を」は文法格を表示する格助詞であり、名詞句と助詞の関係的役割を示す意味的な関係が指定されていなくても、ある述語が選んだ名詞句の数に応じて、図 4 のように左から順に格助詞「が」「を」「に」が付加されると定義した。

この定義をもとに、文末から述語と助詞を探索し、図 4 のK_1、K_2、K_3のいずれかのパターンに当てはまる文節

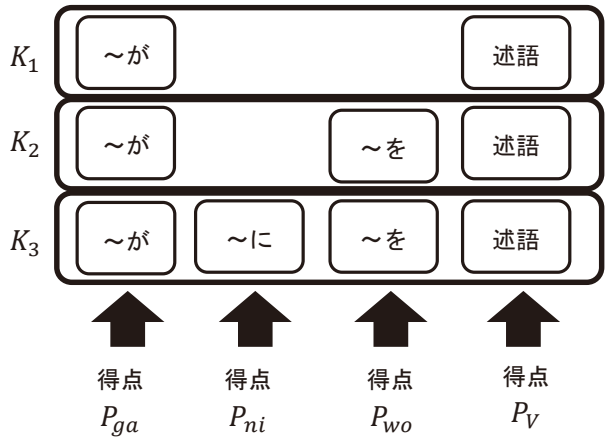


図 4 井上による文法格への得点付け

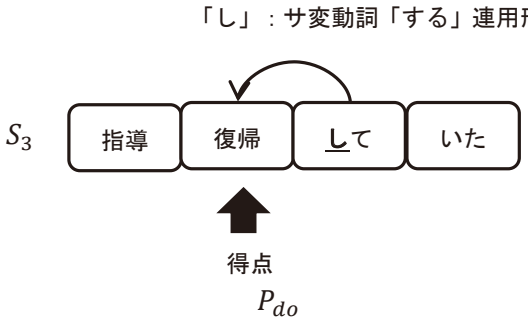


図 5 体言止めへの得点付け

を抜粋することで要約文が作成できると考えた。そこで、述語にあたる「動詞を含む文節」と、助詞「が」「に」「を」を含む文節に対しても、それぞれ得点P_V、P_{ga}、P_{ni}、P_{wo}を与える。

3.4. 体言止めの点数付け

人間が文章の不要な部分を削除して要約文を作成する際、作成された要約文の文末は、新聞記事の見出しのように名詞や動詞の終止形で終わることが多く、「～した」となることは少ない。また、池田ら [5]は、新幹線の電光掲示板やニュースの字幕は、格助詞や名詞で終わる文が多く見られると述べた。これらを参考にして、動詞「した」の活用形を含む文節の、直前の文節に対して得点P_{do}を与える。

3.5. 文節の決定

格文節ボックスを構成する文節などに対して、ユーザが指定した最大文字数以下で、かつ得点の総和が最大になる文節の組み合わせを求め、それらを結合することで要約文とする。

これはナップサック問題と考えることができる。ナップサックの容量、つまり指定された文字数を超えない条件の下で、各文節の価値である文節の得点の総和が最大になるように選択する。本研究ではナップサック問題のソルバとして、Algorithm::Knap01DP のバージョン 0.25 を使用する。

4. 実験と評価

4.1. 実験条件

提案したシステムは CentOS6.3 上に Perl を使用して実装し、形態素解析器には MeCab を使用する。各文節のコスト値は文節の文字数、価値は 3 章で述べた文節の得点の合計値とする。3.2 節、3.3 節および 3.4 節で述べた、各得点付けで与える点数を表 1 に示す。

山本ら [2]によると、新聞記事の冒頭文は記事の主題を示すことが多く、さらに中間文は要約文を作成する際に削除されやすいと考察している。この考察から、提案する要約システムへの入力は新聞記事の冒頭 1 段落とする。さらに、形態素解析を行う際に名詞が連続して出現した場合、それらを一つの名詞とする処理を行う。これは、名詞の連続で企業名になっているなどの場合に、途中で文節として区切られてしまう問題を防ぐためである。

対象の記事として、読売新聞社、毎日新聞社、朝日新聞社の三社が配信しているニュースサイトのトップページから無作為に記事を選び、そのニュース記事の最初の 1 段落をシステムへの入力とする。要約文として出力してよい文字数は、元の文章の 50%までとする。

提案法の評価には、要約システムの自動評価法として最も広く使用されている ROUGE[6]の 1 つである、ROUGE-N[6]を使用する。ROUGE-N は、人間が作成した要約文と構築したシステムの要約文との一致するn-gram の割合を式(1)で評価する。n-gram とは、自然言語処理において言語モデルを構築する際、文章中の句や単語の、隣接するn文字の連続である。特に$n = 2$のときをバイグラム（bi-gram）と呼ぶ。

$$ROUGE\text{-}N = \frac{\sum_{S\in\{Sum\}}\sum_{G_n\in S} C_m(G_n)}{\sum_{S\in\{Sum\}}\sum_{G_n\in S} C(G_n)} \quad (1)$$

ここで、

- Sum ： 人間が作成した要約文の集合、
- S ： 各要約文、
- G_n ： 言語モデル n-gram、
- $C(G_n)$ ： 人間の要約文における、G_nの出現回数、
- $C_m(G_n)$ ： 人間の要約文と、自動要約でのG_nの同時出現回数、

である。式(1)の分子は本システムと人間の生成した要約文で一致したn-gram の数、分母は人間の作成した要約文に含まれるn-gram の総数である。評価値は 1 に近いほど良い結果となる。

Lin[6]によると、単一ドキュメントの要約の評価には$n = 2$、すなわちバイグラムの一致数を基に評価を行うと最も高い相関が得られる。したがって、本システムが生成する要約文の評価にも、ROUGE-2 を使用する。

表 1　各得点付け方法で与える得点

対象の文節	得点
P_{BL}	30
P_{BR}	40
P_{V}	25
P_{ga}	50
P_{ni}	30
P_{wo}	30
P_{do}	30

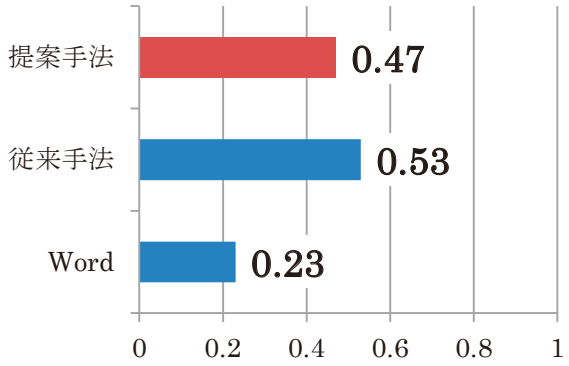
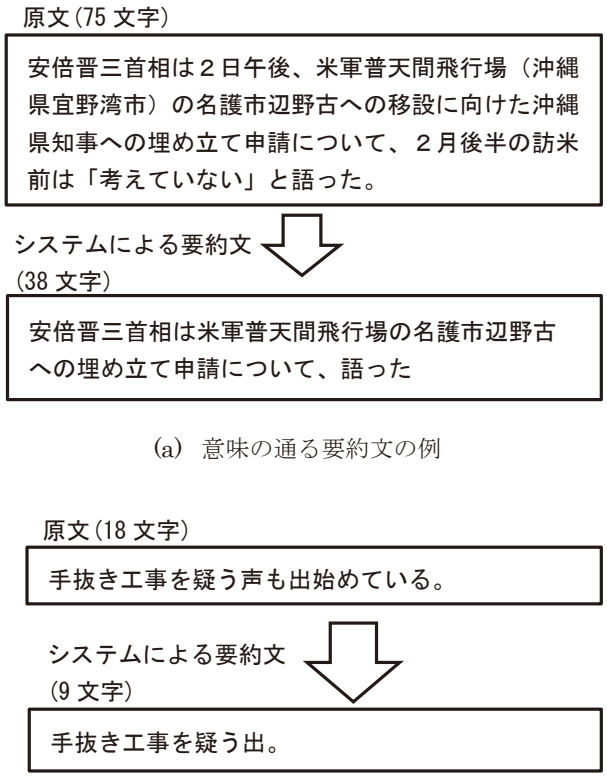


図 6 既存手法および Microsoft Word と
提案システムとの評価値の比較



(a) 意味の通る要約文の例

(b) 非文となった例

図 7 要約文の例

4.2.　評価と考察

4.1 節の条件のもと、提案システムを使用して作成した要約文について、ROUGE-2 を用いて要約結果を比較したグラフを図 6 に示す。既存手法との比較対象として、山本 [2]らによるテンプレート方式による要約手法と、Microsoft Word 2007 の文章要約機能をとりあげる。

テンプレートを用いて要約を行う方法に対して、提案手法の ROUGE-2 の値は 0.06 低い結果となった。この理由として、テンプレートの文章の構成が人間の作成する要約文に類似しており、テンプレートに対し元の文章の適切な単語をあてはめることができたためと考えられる。しかし、テンプレート方式の弱点である膨大な要約サンプルやコーパスの収集、管理によるコストがかからない点で、提案手法は優れているといえる。

また、実際に要約を行った例を図 7(a)と図 7(b)に示す。図 7(a)の要約例では、元の文章からある程度正しい要約文が作成されている。一方、図 7(b)に示す要約結果では、要約した文章が正しい文章になっていない、非文となった部分が存在する。太字で示した部分「疑う出」は、日本語として意味が通じない。

非文となった理由として、得点が高い文節にもかかわらず、文節に含まれる文字が長いために選ばれなかったものがあることと、また出力してよい文字数が元の文章のその 50%と制限されており、元の文章が短いのにもかかわらず、さらに短くしようとしたことが原因であると考えられる。

ある文節が文字数が長いために選ばれない問題は、4.1 節で実験条件に挙げた、名詞の連続を一つに結合する処理によるものや、3.1 節で述べた前処理のカギ括弧内の文字列を一つの固有名詞とみなす処理が大きな原因となっていることがわかった。この問題は、名詞が連続している場合にはそれらから不要と思われる名詞を削除したり、カギ括弧内の文章に対してあらかじめ要約処理を行ったりすることで解決できると考えられる。また、文章が短い場合には出力してよい文字数を増加させることでも解決できると考えられる。

5.　おわりに

情報端末やインターネットの進化、様々なウェブサービス登場により、世界中で今現在何が起こっているのか、身近な友人は今何をしているのかリアルタイムで知ることができるようになった。それに伴い個人が扱う情報の量は膨大になり、短く要点がまとめられた文章を求める風潮が強まっている。

本研究は、新聞記事を文法的な視点、特に文章の格助詞を中心に要約する手法の開発を目的とした。ROUGE-2 の結果から、人間が作成する要約文に近い文を作成することができた。文節に対して、格文節ボックスを使用し

た得点付けと、述語と格助詞の出現順序について得点付けを行うことで、有効に文章の要約を行えることを示した。

今後の課題として、要約結果が非文となってしまう場合は得点付けの重みを動的に変更する方法を追加することが挙げられる。

参考文献

1) 山本 和英, 牧野 恵: 要約事例を用例として模倣利用したニュース記事要約, 自然言語処理, Vol.15, pp.115–158, 2008.
2) 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, 情報処理学会研究報告 自然言語処理, NL-099, pp.17–24, 1994.
3) 工藤 拓, 山本 薫, 松本 裕治: Mecab: Yet another part-of-speech and morphological analyzer, http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html.
4) 井上 和子: 変形文法と日本語, 大修館書店, 1976.
5) 池田 諭史, 大橋 一輝, 山本 和英: 「新幹線要約」のための文末の整形, 情報処理学会研究報告 自然言語処理, FI-076, pp. 161–168, 2004.
6) C. yew Lin, Rouge: a package for automatic evaluation of summaries, Proc. ACL Workshop on Text Summarization Branches, pp. 25–26 2004.