重要語句抽出を利用した要旨作成システム

金子 満生　恵谷 淳一郎　松澤 由梨枝　韓 東力

日本大学文理学部　情報システム解析学科

1. はじめに

学術論文では、本文の前に要旨・要約が存在することが多い。これにより読み手は論文の概要を短い時間である程度把握することができる。特に社会科学系の論文には非常に長いものが多く、読むのに時間がかかるため、要旨・要約の必要性が高い。しかし、論文によっては著者の意見を反映していない要約や、要旨・要約が存在しない論文もある。そこで本研究では、社会科学系の論文を対象にした要旨の作成を行う。ここで、要旨とは「述べられたことの、最も重要なこと。もしくは肝要な事柄。」であり、要約とは「長い話や文章を短くまとめて要点を明らかにすること。また、まとめたもの。」[1]である。

先行研究についての調査では、要旨作成の研究がほとんどなかったため、自動要約を中心にサーベイを行った。自動要約に関する研究では、文章中の重要な部分を用いて要約を作成するものもあれば[2,3]、文生成に伴う要約もある[4,5]。「重要文抽出中心の方が文生成より情報・論理展開ともにシンプルである」[6]という考えに指摘されているように、重要文抽出に基づく要約システムが比較的に多く存在する。しかし、これらの重要文（語句）抽出を中心にした研究では文の抽出もしくは文簡約で終わっているものが多く、文章の形になっていないために読み手には読みにくい。一方、文生成を伴う要約では、対象文章の体裁や書き方に依存するなど、重要度計算を中心にした研究は少ない。

そこで我々は、次の3点に重点を置き重要文抽出を中心にした要旨作成を試みる。

- 頻出度や位置情報だけでない重要語句・重要文抽出を行う
- 単に文を羅列するのではなく、読み手が読みやすいようにする
- 出来るだけ論文中の重要な点・著者の意見を簡潔に述べる

実際に作成したシステムは大きく分けて前処理、文内整理、重要単語抽出、重要文抽出、要旨作成の5つの段階からなる。具体的な処理の流れについては2～6章で述べ、7章以降では評価と結論を述べる。

2. 前処理

論文を構成している文のみを格納したCSVファイルと各文の位置情報を格納したCSVファイルに対し、図1のようにTTM[7]を用いて以下の処理を行う。ここで使われているキーワード辞書は、「現代日本政治小辞典」[8]と「現代思想を読む事典」[9]から作成したものであり、社会科学の専門家から推薦されたものである。

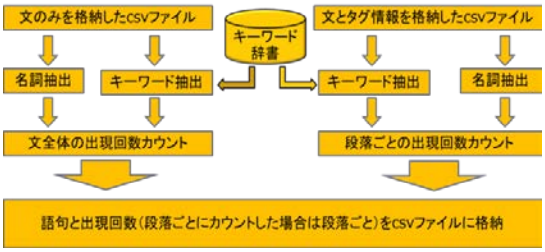

図1. keywords , Nkeywords 抽出

ここで抽出されたキーワードを keywords と呼び、キーワード辞書内に含まれない頻出名詞を Nkeywords と呼ぶ。この Nkeywords に対し形態素解析エンジン mecab[10]で数詞・副詞可能・代名詞・ナイ形容詞語幹・形容動詞語幹と判断されたものを除外し、Nkeywords の出現段落数を求めておく。

最後にあらかじめ用意された次の辞書ファイ

ルを読み込む。

- 副詞辞書　：「副詞的表現の諸相」[11]から程度を表す副詞を抽出したもの
- 末尾表現辞書　：「日本語表現文型」[12]から助動詞と同様の働きをする表現すべてを抽出したもの
- 変換規則対応表　：接続助詞や並立助詞とそれに対応する接続詞を格納したもの
- 必須格情報　：「EDR 日本語共起辞書」[13]を用い、各動詞に対しともに使われやすい格の統計をとり、上位 3 つの格情報を格納したもの

3. 文内の整理

3.1.形態素解析

mecab を用いて論文中の各語句に対して表層形、品詞、品詞細分類 1、品詞細分類 2、原型の 5 つの情報を抽出する。

3.2.丸括弧の削除

論文を読む際、“()”で挿入されている文は、削除しても意味が通じることが殆どであるので削除する。

3.3.括弧内の処理

括弧（丸括弧を除く）内のもので 15 文字以下の長さで、読点を含まないものをその論文独自のキーワードとして抽出し、Tkeywords と呼ぶ。括弧は以後の解析の際には不必要であるため削除する。

3.4.三人称文削除

我々が作成したい要旨では、著者の意見を抽出したいため、三人称が主語の文は要旨に不適切であるため、文内に係助詞の“は”か、格助詞の“が”を含み、かつ直前が固有名詞で人名のものか、直前が接尾語で 2 つ前が人名のもの、もしくは“彼”、“彼ら”など明らかに三人称と判断されるものの 3 通りは、三人称が主語の文として削除する。

3.5.文分割

論文の中には 1 文で複数行に及ぶような長文が存在する。このような長文の解析を行う際、重要な個所が文の中で偏ることが予想される。そこで、表 1 のように文分割を行い不要箇所が抽出されないようにする。また、接続助詞、並立助詞の場合は変換規則対応表を用いて分割を行う。

分割前		分割後
動詞（+接尾）「、」	＋	原型+「。」+そして+「、」
接続助詞	＋「、」	言い切り（原型）+「。」+「接続詞」+「、」
並立助詞	＋「、」	言い切り（原型）+「。」+「接続詞」+「、」

表 1. 文分割規則

3.6.位置情報の付加

ここでは文分割により増えた文や、三人称文削除などで減った文など、論文を読み込んだ時と文の総数が異なるため、各文に対して文の出現する章、段落、出現番号を付与する。また、1 文内に終助詞“か”を含む文を疑問文、代名詞を含む文を指示詞がある文と定義し疑問文と直後の文 [1]と指示詞を含む文と直前の文 [2]をそれぞれペア文として抽出する。

4. 重要単語抽出

今までに抽出した 3 種類のキーワードをそれぞれ文字数の多い順に並び替える。これは「階級政治」のような文字数が多いキーワードに、「階級」、「政治」のような文字数が少ないキーワードが含まれることがあるので重複を防ぐためである。

次に式 1 を基本式とし、単語の重み付けを行う。ここでは単語の出現頻度を wc、単語が出現した段落数を wp、全段落数を dp とする。

$$\mathrm{Score} = \mathrm{wc} \cdot \left(\frac{\mathrm{wp}}{\mathrm{dp}} + 1\right)$$

式 1.重み付けの基本式

まず、頻出度による重み付けを行う。keywords、Tkeywords、Nkeywords の重みを Socre_K、Score_Tk、Score_Nk とする。Nkeywords は単なる頻出名詞のため、他の二つに比べると重要度が

[1] 疑問文が最後の文なら抽出を行わない
[2] 指示詞を含む文が第 1 文目、もしくは指示詞の前に読点があり、かつ読点より前に名詞がある文は抽出を行わない

低いため係数 i(0.9)を掛けて差別化を図る。この数字は我々が幾度か実験を繰り返した結果、適当と思われた数字である。次に位置情報による重み付けを行う。　社会系論文を 30 編以上調査した結果、章の最初と最後の段落、また、最後の章に書かれる文章はより重要度が高いことが分かった。重要な位置にキーワードが出現した場合、$Score_{\_K}$、$Score_{\_Nk}$ に関しては重要な位置に出現した回数だけ加算した。また $Score_{\_Tk}$ に関しては出現位置に関係なく重要なので差別化を図るためキーワードの文字数を加算する方法をとった。

5. 重要文抽出

論文では要旨の下にキーワードを表示しているものが多くみられるので、総合上位 5 つのキーワードを抽出するため Keywords ・ Nkeywords ・ Tkeywords の重複を確認し Score の高い順に上位 5 つのキーワードを抽出した後文字数の多い順に並べ替える。

次に文単位で重要度を考える。文中に程度を表す副詞、助動詞と同様の働きをする文末表現があれば、その文は筆者が強調したいところであると考えられる。そこで強調度チェックに必要な副詞辞書と末尾チェックに必要な末尾表現辞書に載っている語句と表層的に一致する語句を含んでいれば、文の重み付け時に考慮する。また末尾表現の表層一致を確認する際に日本語係り受け解析器 cabocha [14]を用いて係り受け解析を行う。これは、末尾表現では文末を利用するため、文末表現が文末以外から検出されることを防ぐためである。

最後に文の重み付けを行う。文の重みは基本的には文中に含まれるキーワードの重みの総和である。これに加えて文内に強調チェックに一致する語句があれば文の重みを α 倍にする。末尾チェックに一致した場合も同様に α 倍、両方一致した場合は $(\alpha)^2$ 倍にする。α は経験的に定めた数値であり、評価実験では 1.1 を用いている。

以上の手順により重み付けをした文を重みの高い順に格納する。

6.要旨作成

要旨では文内の最も重要なことが述べられるのが望ましいたいため文簡約を行い、不要個所を削除する。はじめに、一文ずつ cabocha で係り受け解析を行い、文を受け文節と係り文節に分けた後、必須格情報を用いて係り文節から受け文節の動詞に対する必須格を含む文節を抽出する。次に、文簡約をするにあたり、以下の要素を文の必要な要素とし、それ以外の文節を削除する。

- 必須格
- 文節内に Tkeywards、keywords が含まれているもの
- 上の 2 つを満たす文節に直接的若しくは間接的に係られているもの
- 受け文節（述語）
- 受け文に直接係っている文節

簡約後の文に、簡約前の重みと 3.6.で付与した位置情報を継承させ、指定された割合に基づいて式 2 により要旨に用いる文を取得する。

$$取得文字数 = 指定された割合 \times 全文章の文字数$$

式 2.要旨文取得の基準式

文の取得過程において、必要な文字数を超えてしまった場合、抽出途中の文を破棄する。これは後のペア文追加で文が増える可能性があるためである。取得した文章をより自然な形にするために位置情報を元に文を出現順に格納する。この際に、3.6.で関連づけした文が存在しかつ既に文中に同文が含まれていない場合そのペア文を指示詞の文ならば指示詞の前に、疑問詞の文ならば疑問詞の後に追加する。最後に接続詞の調整など一連の整形作業を行う。

7. システム評価

アンケート評価では、論文 A を 10%、論文 B を 5%で要旨を生成し、文字数を基準に割合指定を行う文字数選択型と文数を基準に割合指定を行う文数選択型それぞれで要旨を作成し、「文章が文法的に自然かどうか」、「意味の通る日本語かどうか」、「文の繋がりが自然かどうか」の 3 点においてこの研究に関わっていない 9 人に 4 段階で評価実験を行った。表 2 は文字数選択型、文数選択型のそれぞれの項目の平均である。この表からは「文章が自然かどうか」、「意味の通る日本語かどうか」の 2 つの項目では 7～8 割の評価を得ることができた。しかし「文の繋がりが自然かどうか」の項目についての評価は 6 割程度に留まった。

	文章が文法的に自然かどうか	意味の通る日本語かどうか	文の繋がりが自然かどうか
文字数選択型	3.2 (73%)	3.4 (80%)	2.9(65%)
文数選択型	3.3(78%)	3.2 (73%)	2.7 (56%)

表 2：論文選択型ごとの平均

「文の繋がりの自然さ」の評価が平均で 6 割に止まった点に関しては、指示詞、疑問詞、接続詞のみの考慮では不十分であったためであろう。

8. 結論

本研究では、重要文抽出に基づく社会科学系論文の要旨作成を試みた。重要文抽出では、ジャンル独自の辞書や末尾・副詞表現などを採用することにより、頻出度と位置情報によるものに比べて対象論文の特徴的な部分の抽出が可能となった。要旨作成部分では、ペア文と共に表示し、文章化することにより重要文を羅列するのに比べ読み手に読みやすい文章を作成することができた。

しかし、評価実験の結果を通してみると、まだまだ満足のいく結果とは言い難く、段階ごとに実験を細かく重ねることにより手法全体のアルゴリズムを大幅に改善する必要があると思われる。たとえば、単語重み付けの際に強調表現を考慮しているが、否定的なのか肯定的なのかまでは検討していないので、新たな手法の導入を行っていきたい。また、文の繋がりについて、指示詞と疑問詞、接続詞についてしか考慮していないので判断基準の拡張が必要である。さらに、文の主語の判断基準が 3 通りのみなので、同様に判断基準の拡張を行っていこうと思う。

参考文献

[1]梅棹忠夫, 金田一春彦, 阪倉篤義, 日野原重明, 「日本語大辞典講談社カラー版第二版」, 講談社. (1995).

[2]諸岡祐平, 江嵜誠, 高木一幸, 尾関和彦, 「重要文抽出と文簡約を併用した新聞記事の自動要約」, 言語処理学会年次大会発表論文集, A10 P4-04, (2004).

[3]畑山満美子, 松尾義博, 白井諭, 「重要語句抽出による新聞記事自動要約」, 情報処理学会研究報告, NL−141, pp.95-101. (2001).

[4]和田裕二, 奥村明俊, 浦谷則好, 白井克彦, 「属性を用いた文節重要度に基づくニュース文要約」, 言語処理学会年次大会発表論文集, pp.543-546. (2002).

[5]川端正法, 山本和英, 「話題の継続に着目した国会会議録要約」, 言語処理学会年次大会発表論文集, pp.696-699. (2007).

[6]望主雅子, 萩野紫穂, 太田公子, 井佐原均, 「重要文と要約の差異に基づく要約手法の調査」, 情報処理学会研究報告, NL−135, pp.95-102. (2000).

[7] http://www.mtmr.jp/ttm/

[8]内田満, 「現代日本政治小辞典」, ブレーン出版. (2005).

[9]今村仁司, 「現代思想を読む事典」, 講談社現代新書. (1988).

[10] http://mecab.sourceforge.net/

[11]仁田義雄, 「副詞的表現の諸相」, くろしお出版. (2002).

[12]森田良行, 松木正恵, 「日本語表現文型—用例中心・複合辞の意味と用法」, アルク. (1989).

[13] http://www2.nict.go.jp/r/r312/EDR/J_index.html

[14] http://code.google.com/p/cabocha/