文章の構造化による修辞情報を利用した自動抄録と文章要約

比留間 正樹[†] 山下 卓規^{††}
奈良 雅雄^{†††} 田村 直良^{††††}

本論文では、重回帰分析にもとづいた文章構造解析を利用した自動抄録手法とその評価、および文章要約への展開について述べる。文章構造の解析は、文章中の様々な特徴をパラメタとした判定式や局所的な言語知識により、文章セグメントの分割統合を進めて構造木を作るものである。得られた文章構造上の各種特徴をもとに、さらに文章抄録の観点から選択されたパラメタを加えて、文抽出のための判定式を作る。本研究では被験者5人にのべ350編の新聞社説の抄録調査を実施し、これを基準に、重回帰分析によりパラメタの重みを決定し判定式を得、また、本方式を評価する。また、自動生成された抄録文に対して、照応情報の欠落による文章の首尾一貫性の低下を避けるための補完や、論旨を損なわない冗長な表現の削除を行なうことで要約文章を生成する手法を紹介する。

キーワード: 文章抄録, 文章要約, パラメタ, 学習, 文章構造, 重回帰分析, 修辞構造

Extract Generation using Rhetorical Information by Text Structuring and its Extension to Summary Generation

MASAKI HIRUMA [†], TAKUMI YAMASHITA ^{††}, MASAO NARA ^{†††}
and NAOYOSHI TAMURA^{††††}

This paper presents a method of automatic extract generation from text by using the rhetorical structure of the text which is built by discriminants trained by the multiple regression analysis. We also show an extension to the summary generation. The way of our method of text structuring is to divide or to combine text segments one after another according to the discriminant with various parameters for superficial features of sentences. Another discriminant to extract important sentences has parameters selected from the point of view of text extraction along with several features on the text structure (rhetorical structure). In the experiment, five examinees select important sentences of newspaper editorials from total 350 articles. The results are used to the calculation of weights of parameters by the multiple regression analysis, and also to the evaluation of accuracy of our method. We also show an extension to summary generation in which sentences or phrases are restored not to decrease the coherency of text caused by deletion of referred sentences, and in which redundant expressions independent to the context are deleted.

KeyWords: *extract, summary, parameter, learning, text structure, multiple regression analysis, rhetorical structure*

1 はじめに

本論文では、文、文章上の特徴、および文章の解析により得られた構造上の特徴をパラメタとして用いた判定式による文章の自動抄録手法を示す。さらに、抽出された文の整形や照応を考慮した文章要約手法について述べる。

近年のインターネットなどの発展により、大量の電子化された文書が我々の周りに溢れている。これら大量の文書から必要とする情報を効率良く高速に処理するために、キーワード抽出や文章要約、抄録といった研究が行なわれている。それらのためには、計算機を用い、必ずしも深い意味解析を行わずに文章の表層的特徴から解析を行なう方法が有効である。

文章抄録とは文章から何らかの方法で重要である文を選び出し、抽出することである。山本ら(山本和英, 増山繁, 内藤昭三 1995)は照応, 省略, 語彙による結束性など多くの談話要素から重要文を選択していく論説文要約システム (GREEN) を発表している。このシステムは談話要素を利用したものではあるが、文章の局所的な特徴を基に文を抽出するもので、本研究の立場からすれば文章全体の構造に基づく抽出と、電子化された大量のコーパス利用を考慮した抽出手法や手法の評価が必要と考える。

また、亀田(亀田雅之 1997)は重要文の抽出の際に文章の中で小さなまとまりを示す段落や、一種の要約情報である文の見出しに着目する手法を提案、実現しているが、重要度計算の調整は人手により、系統的でないところが感じられる。

さて、重要文の抽出に用いられるテキスト中の表層的特徴については、(奥村学 難波英嗣 1998)にサーベイがある。これによると、Paice(Paice. 1990)の分類として、(1) キーワードの出現頻度によるもの、(2) テキスト、段落中の位置情報によるもの、(3) タイトル等の情報によるもの、(4) 文章の構造によるもの、(5) 手がかり語によるもの、(6) 文や単語間のつながりによるもの、(7) 文間の類似性によるものがあげられている。本研究での手法は、上記のかなりの要素を組み合わせるパラメタとして利用している。

いくつかの観点からのパラメタを組み合わせるといふ同様な手法として、(Watanabe 1996)、(野本忠司 松本裕治 1997)がある。それぞれ、重回帰分析、決定木学習により訓練データから自動学習するものである。われわれの手法は、構造木に関する情報を特に重視している。

人間は、目的の意見、主張を読み手に伝えるために、意識下／無意識下に文章構成の約束に基づいて文章生成を行なっているが、それらの文章に論証性を持たせるためのものが文章構造である。また逆に、文章を理解し論旨を捉える際に文章構造を活用していると考えられる。したがって、文章の抄録にあたり、論旨を捉え、文章構造を理解した上で重要文を抽出していく手

† 日本 IBM 株式会社, IBM Japan, Ltd.

†† 株式会社 東芝, TOSHIBA Corporation

††† 日立ソフトウェアエンジニアリング株式会社 ネットワーク本部 ネットワーク第1設計部, Network Systems Dept. Hitachi Software Engineering Co., Ltd.

†††† 横浜国立大学 教育人間科学部 情報認知システム講座, Department of Information and Cognition Systems, Faculty of Education and Human Science

法は人間の文章抄録の流れに沿っており、ごく自然であると考えられる。実際、(Marcu 1997)では、人間の手による生成ではあるが文間の関係を解析した修辭構造生成後の文抽出の再現率、適合率は良好と報告されている。われわれの手法でも、修辭構造を含めた文章構造解析による情報を利用する。

文章構造解析には田村ら(田村直良 和田啓二 1998)の分割と統合による構造解析手法を利用する。文章抄録には、構造解析で用いたパラメタに加えて、得られた文章構造上の情報についてのパラメタにより文抽出のための判定式を作り、それを基にして抄録を作成する。判定式とパラメタの重みの決定は重回帰分析に基づき、その訓練のため、およびシステムの評価のための基準データは、被験者に対するのべ350編の抄録調査による。なお、実験の対象とした文章は、均一な文章が容易に入手可能であるとの理由から、新聞の社説を用いる。

一方、原文から単に文を選ぶだけの文章抄録では、選択された文間の隣接関係が不自然になる場合がある。また、たとえ選択された一文でも文内には冗長な表現が残っている場合がある。そこで、自動要約に向けては、抄録後になんらかの文章整形過程が必要である。本研究では、抄録の整形過程としての照応処理と、一文の圧縮処理を行なう。

以下、第2章では文章抄録、要約のための文章構造解析について述べ、第3章では文章の自動抄録の手法について説明する。第4章では、提案の手法について再現率、適合率により評価検討を行う。最後に付録として、抄録の整形過程について述べ、実際に要約した文章例を示す。

2 分割と統合による文章構造解析

文章とはある首尾一貫性を持った文の集まりである。それゆえ文章を理解し、抄録を作成するためには、その文章がどのような構成(構造)になっているかを知る必要がある。よってここでは文章の構造解析について述べる。

文章の構造化としてはMann(Mann 1987)の修辭構造理論を基にして論旨の展開を木構造として表現する手法(田村直良・和田啓二 1998)を用いる。これは大量の文章を高速に処理するために深い意味解析に立ち入らず、主に表層的な処理のみでセグメントの分割(トップダウン的アプローチ)、およびセグメントの統合(ボトムアップ的アプローチ)を行なっている。こうすることで、文章の木構造を根から葉へ、かつ葉から根へと交互に生成してゆき、一方の欠点を他方の利点で補う効果的な文章解析を行なうことができる。この解析処理の例を模式的に示したのが図1である。

トップダウン的解析では、望月ら(望月源, 本田岳夫, 奥村学 1996)の文章の表層表現を情報とした、テキストセグメンテーションの手法を利用して文章の分割を行なう。すべての文間について判定式による評価値を求め、評価値の高い順にセグメントの分割を行なう。また、ボトムアップ的解析でも判定式を用い、すべての文間について評価値を求め、評価値の高い順にセ

図 1 分割と統合による構造解析

グメントの統合を行なう。それぞれの判定式のパラメタの重みは、訓練データ¹のテキスト中のすべての文と文の境界についてパラメタを評価し、重回帰分析により求める。

パラメタは以下の観点から選択する(望月源他 1996)(詳細は 3.2 の 1(単独の文に関するもの)参照)。

- 助詞「は」と「が」の出現
- 接続語の有無
- 指示詞の有無
- 時制の情報
- 文末表現の情報
- 語彙連鎖情報

本手法によって新聞社説²を解析してできた修辞構造木の例を図 2 に示す。

この修辞構造木において(m,n) は 1 文に対応し、m が原文章の形式段落の番号、n が形式段落内の文の番号を表す。

また、「順接」、「強調」、「説明」等は隣接するセグメント間の修辞関係³を表している(Mann 1987)。修辞関係は、cue word により文頭の接続表現を 18 種、文末表現を 14 種に分類し、接続表現があればそれにより同定し、なければ、隣接する前文、後文の文末表現その他の組み合わせをもとに決定している。それぞれの関係に応じて、どちらが核(nucleus)、衛星(satellite)で

¹ 日本経済新聞の社説 100 編

² 日本経済新聞 1993 年 1 月 28 日(付録に示す)

³ それ以外に「逆接」、「換言」、「添加」、「条件」、「結論」、「一般化」、「相反」、「提起」、「根拠」、「因果」、「並列」、「選択」、「対比」、「転換」、「例示」がある。

```

[ ]
|- (1,1)
|- 順接
|- [ ]
    |- [ ]
        | |- [ ]
        | | |- [ ]
        | | | |- [ ]
        | | | | |- [(2,1), 順接, (2,2)]
        | | | | |- 逆接
        | | | | |- [[ (3,1), 順接, (3,2)], 順接, (3,3)]
        | | | |
        | | | |- 順接
        | | | |- (4,1)
        | | |
        | | |- 順接
        | | |- [ ]
        | | | |- [ ]
        | | | | |- [(5,1), 順接, (5,2)], 順接, (5,3)]
        | | | | |- 順接
        | | | | |- (5,4)
        | | | |
        | | | |- 順接
        | | | |- (5,5)
        | | |
        | | |- 順接
        | | |- [ ]
        | | | |- [(6,1), 順接, (6,2)], 順接, (6,3)]
        | | | |- 順接
        | | | |- [(7,1), 順接, (7,2)]
        | | |
        | | |- 順接
        | | |- [[ (8,1), 結論, (8,2)], 転換, (8,3)]

```

図 2 修辞構造解析結果の例

あるかも決定する。また、セグメント間の修辞関係の同定は、核をたどることによってセグメントを代表する 1 文を求め、これらの文間の関係により同定する。なお、「順接」を default としている。(詳細については、(田村直良・和田啓二 1998) を参照されたい。)

3 文章抄録

本論文で提案する抄録手法は、まず文章の構造化を行ない、得られた修辞構造木の情報やその他抄録の観点から選択されたパラメタに基づく判定式により各文の重要度を算出し、重要度の高い文から抽出していく。判定式中のパラメタの重みは被験者が作成した抄録について重回帰分析により訓練する。

また、本研究では、文だけでなく構造木の部分木の削除も行なう手法についても検討する。部分木とは、根に近い枝から順に構造木を分割したもののリスト構造で、文章の段落に相当するものである。部分木の数は、社説の全文数を形式段落内の文の数の平均 2.69⁴ で除することにより求める。この処理を模式的に示したのが図 3 であり、構造木の根に近い節点から (1,2,3) 順に分割していったものが部分木である。実際に図 2 の構造木から再構成される部分木の例を図 4 に示す。この例では文の数が 20 個なので、それを 2.69 で割り、部分木の数が 7 となる。

図 3 構造木の部分木

3.1 抄録手法

本研究で提案する抄録手法は、基本的には 2 節で述べた文章の構造化を行ない、その構造上の情報と 3.2 節で述べるパラメタより重要文を選択、抽出するものである。ここでは、さらに文の抽出の前に部分木を削除する過程を加えた手法も検討対象とする。また、構造化は行なわずに判定する手法も加え、以下の 3 種の手法を検討する。

手法 1 (文章構造解析 + 部分木削除 + 重要文抽出による抄録)

この手法では、(i) 文章構造解析により修辞構造木を作成し、作成された修辞構造木をさらに部分木に分割する。(ii) 各部分木の重要度を後出の式 (1) を用いて算出し、重要度の低い部分木から削除していく。(iii) 残った各文の重要度を算出し、重要度の高い文から抽出していく。という 3 つの段階を踏む。(ii)、(iii) での削除、抽出の量は実験時に設定

⁴ 1993, 1994 年の日本経済新聞の社説 1227 編から求めた。

```

[ ]
|- (1,1)
|
|-[ ]
|  |- [ ]
|  |  |- [(2,1), 順接, (2,2)]
|  |  |- 逆接
|  |  |- [[ (3,1), 順接, (3,2)], 順接, (3,3)]
|  |
|  |  |- 順接
|  |  |- (4,1)
|  |
|  |-[ ]
|  |  |- [[ (5,1), 順接, (5,2)], 順接, (5,3)], 順接, (5,4)]
|  |  |- 順接
|  |  |- (5,5)
|  |
|  |  |- [(6,1), 順接, (6,2)]
|  |  |- (6,3)
|  |  |- [(7,1), 順接, (7,2)]
|  |  |- [[ (8,1), 結論, (8,2)], 転換, (8,3)]

```

図 4 構造木から再構成される部分木の例 (図 2 の構造木より)

する。

手法 2 (i) + (iii)

この手法では、パラメタを採集するために文章の構造化、さらに部分木の再構成を行い、その後で構造上の情報その他を用いて各文の重要度を算出し、抽出する。

手法 3 (iii)

この手法では、文章の構造化は行わずに、単独の文個々についてのパラメタにより各文の重要度を算出し、抽出する。

3.2 判定式を用いた自動抄録

観測点の各種パラメタを基にそれが属するクラスを判定する手法はいくつかあるが、本研究では重回帰分析を用いる。

判定のための式は以下である。ただし、採用、不採用の判定は、対象文章の各文を式 (1) により評価し、評価値が大きいものから、順に所定の要約率となるまで採用することとする。

$$\hat{z}_{sum} = \delta_0 + \delta_1 y_1 + \delta_2 y_2 + \cdots + \delta_q y_q \quad (1)$$

(y_i : パラメタ i の点数, δ_i : パラメタ i の重み)

文の重要度算出の判定式

各文のパラメタとして以下を用いる。

- (1) 単独の文に関するもの

- 文末表現 (3 種)
表層表現から文を「意見」,「断定」,「陳述」の 3 タイプに分類し, 該当するパラメタを 1 とする.
 - 時制 (1 種)
文を「現在」,「過去」に分類し,「現在」ならパラメタを 1 とする.
 - 参照表現 (1 種)
文に「これ」,「その」,「あれ」などの指示語による参照表現が存在すれば 1 とする.
 - 助詞「は」と「が」の出現 (2 種)
文に助詞「は」,「が」が存在すれば, 該当するパラメタを 1 とする.
 - 接続詞 (2 種)
「補足」的,「展開」的な接続詞が存在すれば, 該当するパラメタを 1 とする.
 - 文に含まれる重要語の数 (1 種)
キーワード抽出で用いられる tf*idf 法⁵により抽出された重要語の出現数を値とする.
 - 文を通過する重要語の語彙連鎖の数 (1 種)
語彙連鎖を形成する重要語の出現数を値とする. ただし, 同義語はまとめて 1 語とし, 連鎖中の短いギャップに対しては重要語が存在しなくても数に加える.
 - 文の位置 (2 種)
文章の先頭からの距離, 終わりからの距離を値とする.
- (2) 構造木上の情報に関するもの
- 修辞関係の核 (nucleus) または衛星 (satellite) であるかどうか (34 種⁶)
各パラメタは「順接」を除く修辞関係 17 種それぞれについての核, 衛星に対応する. 文が文章構造解析で得られた修辞関係の核, 衛星であれば対応するパラメタを 1 とする.
- (3) 部分木 (段落) の情報に関するもの
- 文の位置 (6 種)
文が含まれる部分木の位置, 部分木中の位置, など.
- (4) 部分木の削除についても, 以下のパラメタを持つ判定式を用い, 重回帰分析により訓練する.
- 部分木内の文の文末タイプの情報 (3 種)
部分木内で,「意見」,「断定」,「陳述」についてそれぞれのタイプの文が出現する割合を値とする.

⁵ 実験では, idf 値は, 1993,1994 年の日本経済新聞の社説 1227 編より求めた.

⁶ 訓練の結果, 使用されなかったものもある.

- 部分木内の文に含まれる重要語の数 (1 種)
前述の重要語の数.
- 部分木内の文を通過する重要語の語彙連鎖 (1 種)
同様に, 重要語の語彙連鎖の数.
- 文章の開始または終りからの距離 (2 種)
この部分木 (段落) が, 初めまたは終わりから幾つめの部分木であるか.
- 部分木内の文の数 (1 種)
部分木 (段落) 内に含まれる文の数.

4 実験と検討

4.1 パラメタの訓練

実験にあたって, まず訓練と評価のために人間の被験者の抄録結果による正解を定める. 抄録は, 学生 5 人⁷を被験者として, 日本経済新聞 CD-ROM'93~'94 版の中の社説のべ 350 編について重要であると思う文を全文数の 3 割程度選択させる.

調査用社説は表 1 のように A~L のセットに分け, A~F セットを訓練用セット, G~L セットを評価用セットとする. 被験者と調査させたセットの対応を表 2 に示す.

訓練セット	記事数	評価セット	記事数
A	20	G	10
B	30	H	10
C	30	I	10
D	30	J	10
E	30	K	10
F	30	L	10

表 1 調査セットと記事数

被験者	調査セット
1	A,B,G,L
2	A,C,H,L
3	A,D,I,L
4	A,E,J,L
5	A,F,K,L

表 2 調査セットの被験者への配分

⁷ 本学学部 1 年生. 自然言語処理についての教育は特に受けていない.

調査の結果、全文数 8164、選択された文数 2172、平均要約率 26.6%となった。

まず、A セット 20 編 (以下、訓練セット 1)、B～F セット合計 150 編 (以下、まとめて訓練セット 2) の 2 セットについてそれぞれ重回帰分析により、式 (1) におけるパラメタの重みを決定する。それぞれは、次のような方針に基づく。

訓練セット 1 訓練セット 1 の基準値としては各文についてその文が重要であるとした被験者の人数を与える。つまり、各文に 0～5 の値を与える。この訓練セットは同一の文章を多人数で調査することによって判定の質を重視した訓練セットである。

訓練セット 2 訓練セット 2 の基準値としては各文について担当した被験者が重要であるとした文に 1、そうでない文に 0 の値を与える。この訓練セットは標本集合を大きくすることによって訓練の規模を重視する訓練セットである。

手法 1 では、部分木の選択においても重回帰分析による手法を用いるが、被験者には文単位での重要文選択しか行なわせていないため、あらかじめ前述の手法により各社説の構造化を行ない、部分木を作成する。その上で、被験者が選択した文が部分木に含まれていればその部分木の観測値を 10、含まれていなければ 0 とする。これを正解として重回帰分析によりパラメタの重みを決定する。

4.2 生成された抄録文の評価

抄録は 3.1 節に示した 3 種類の抄録手法で作成し、さらに手法 1 については部分木の削除の段階で、部分木 (段落) を 50%, 25%, 10% ずつ削除した場合についてそれぞれ抄録を作成する。抽出する文の数は、人間の重要文選択調査における平均要約率 26.6% を目安とする。

評価セット G～K、および評価セット L に対する実験結果から、それぞれについて以下の式により、再現率、適合率を求める。

$$\text{再現率} = \frac{\text{自動抄録により得られた抄録文に含まれる正解文数}}{\text{正解文数 (人間によって選択された文数)}}$$

$$\text{適合率} = \frac{\text{自動抄録により得られた抄録文に含まれる正解文数}}{\text{自動抄録により得られた抄録文数}}$$

ここで、正解とは以下のように定義する。

評価セット G～K については、個別に評価するのではなくひとまとまりとして扱い、担当した被験者が選択したかどうかによるものとする。

評価セット L については、被験者 5 人中 2 人以上が重要であると判断された文を抄録の正解とする。

評価セット G～K に対する再現率、適合率を表 3 に、評価セット L に対する再現率、適合率を表 4 に示す。

	部分木の削除率	訓練セット 1 で訓練		訓練セット 2 で訓練	
		再現率	適合率	再現率	適合率
手法 1	50%	26.9%	36.9%	25.3%	36.7%
	25%	31.9%	37.4%	30.9%	36.4%
	10%	33.1%	36.4%	35.3%	38.8%
手法 2	—	33.4%	34.5%	37.5%	38.7%
手法 3	—	32.8%	33.9%	34.1%	35.2%

表 3 評価セット G~K に対する再現率・適合率

	部分木の削除率	訓練セット 1 で訓練		訓練セット 2 で訓練	
		再現率	適合率	再現率	適合率
手法 1	50%	33.3%	40.7%	34.7%	42.4%
	25%	31.9%	39.0%	34.7%	42.4%
	10%	33.3%	40.7%	35.4%	43.2%
手法 2	—	33.3%	40.7%	34.7%	42.4%
手法 3	—	40.3%	49.2%	33.3%	40.7%

表 4 評価セット L に対する再現率・適合率

4.3 抄録文に対する検討

- 評価セット G~K に対する検討

再現率に関しては手法 1 の場合が低い値となった。これは部分木の削除の際に被験者が選択した文が含まれていても、その文を含む部分木の重要度評価値が低かったために部分木が削除されてしまい、抄録文として抽出されなかったためと思われる。

訓練セット別に見ると訓練セット 1 より訓練セット 2 で訓練した方が、高い値となっている。これは訓練セット 1 の文章の数が社説 20 編と少なかったからと考えられる。

適合率に関してはすべての手法でほぼ同じ値となった。訓練セット間の比較では訓練セット 1 の方が低い値となったが、再現率のときと同様の理由と思われる。

抄録手法別に見ると再現率・適合率ともに手法 1(10%) と手法 2 が高い値となった。これは本研究で提案する文章構造の情報をパラメタに含めた判定式による抄録手法が有効であることを示している。また、過度に部分木を削除することの不適切性を示している。

- 評価セット L に対する検討

再現率・適合率ともに評価セット G~K に対する結果に比べ高い値となった。これは特異的に優しいセットだったためだと思われる。あるいは、被験者 5 人の判定を重ね合わせているので判定の質がアップしていることも寄与していると考えられる。

訓練セット別に関しても抄録手法別に関しても、前述の評価セット G~K に対する結果と同様の傾向を示した。

また、訓練セット 1 と評価セット L の組合せでは手法 3 の再現率・適合率が他より高いが、これは訓練セット、評価セット共に少ないため信頼できる評価ができないためで、より多くの評価セットで評価を行なう必要があると思われる。

- パラメタ学習に関する検討

セット A で訓練し、それを評価セットとして抄録を作成したときの再現率、適合率を表 5 に示す。ただし、正解は被験者 5 人中 2 人以上が重要であると判断された文とする。

	部分木の削除率	訓練セット 1 で訓練	
		再現率	適合率
手法 1	50%	41.1%	54.3%
	25%	42.3%	55.9%
	10%	42.9%	56.7%
手法 2	—	42.9%	56.7%
手法 3	—	39.3%	52.0%

表 5 セット A で訓練、評価したときの再現率・適合率

本来、訓練セットそのものを評価した場合、再現率、適合率とも 100% になってもよさそうなものである。表 5 の値は、訓練の限界を示す目安になっており、先の評価結果は、再現率ではこの値にかなり近付いている。

- 人間の判断のばらつきに関する検討

被験者 5 人に対して、重要文選択調査した杜説のうち 4 人の結果を正解として、残り 1 人の結果と比較する。以上の実験を順に 5 人すべてに対して評価を行なった結果を表 6 に示す。なお、正解は 4 人中 2 人以上が選択した文とする。

人間同士の比較でも再現率、適合率は表 6 で表される程度にばらつきがある。ただし、再現率、適合率間でそれほど大きな差はない。

自動抄録についての研究の一つの目標を示しているものと考えられる。

被験者	1	2	3	4	5	平均
再現率 (%)	61.1	61.8	53.5	48.6	55.9	57.2
適合率 (%)	56.4	67.2	67.9	58.4	68.6	63.7
要約率 (%)	30.8	26.3	24.2	24.9	24.8	26.2

表 6 被験者の抄録についての評価

以上の結果より、本研究で提案する文章構造情報を利用した抄録手法が、利用しない抄録手法より有効であると言える。また、文章構造木を部分木に分割し、重要文抽出の前段階で部分木を削除する手法も提案したが、小規模な部分木削除 (10%) でもっとも良好な結果を示している。

部分木の削除を考慮する場合、周囲との関連がない重要な文の存在が結果の精度に大きく影響する。すなわち、ノイズ的にパラメータの評価値が高くなってしまふ文では、部分木の削除

を行う場合(手法1), これが削除され, 精度は向上する.

5 おわりに

本論文では, 文章中のさまざまな特徴をパラメタとした判定式から文章の構造化を行ない, さらに文章抄録の観点から選択したパラメタを加えた判定式による自動抄録手法を提案, 実現し, 評価, 検討した.

従来の抄録手法は文章中の情報を用いて各文について重要度を算出し, 重要度の高い文から順に抽出していくものが多かった. 本論文における手法も同様のものであるが, 判定式に用いるパラメタは抄録作成の観点から選択したものに加えて, 文章構造解析による構造木の修辞関係や部分木の位置などの構造的なものも選択した. パラメタの重みは, 人間による重要文抽出調査の結果をもとに重回帰分析により求めた. また, 文章構造の情報を考慮した場合としない場合での比較を行なうために, 3種類の手法で抄録を作成し, 検討を行なった. 結果からパラメタとして文章構造の情報を使用する方が使用しない場合よりも良い結果が得られることが分かった. また, 個人差があっても大規模な訓練セット(150編)を用いた方が, 同一の記事を複数人で調査した訓練セット(20編)を用いるよりも結果は良好であった.

なお, 開発途中であるが, 照応解析および一文内の圧縮をすることによる抄録から要約への展開および自動要約の実現を付録に示した.

今後の課題として, 他の学習方法についての検討, より大きい訓練セットでの実験, 抄録文章の整形過程の検討, また, システムとしてのユーザインタフェースの整備があげられる.

謝辞

本実験で使用したコーパスは, 日本経済新聞 CD-ROM'93~94 版から得ている. 同社, および使用に関して尽力された方々に深く感謝します.

付録 (文章要約)

抄録文章は、原文章から文を抜き書きしたものであり、参照表現、接続詞等も原文章のままであるので、そのままでは人間が読むときに不自然さを感じる。要約とは、より少ない文字数で原文章と同じ内容を表現することであるが、1つの文章として文を越えた何らかの意味的なつながり、つまり論旨が成り立っている必要がある。

そこで本研究では、照応関係を解析し、その情報を利用することで照応情報の欠落による抄録文の首尾一貫性の低下を避ける。また、文内で比較的重要度が低いと考えられる表現を削除することで要約文章の圧縮を行なう。

本方法では、対象とする照応表現は指示詞および指示連体詞とし、以下のような照応関係の判定を行なう。

- 指示対象が直前文である。
- 指示対象が前段落 (部分木) である。
- 指示対象が前出の名詞句である。
- 指示対象が文章内に存在しない。

判定の結果、指示対象が文章内に存在する場合、指示対象を抄録文に採用する。ただし、指示対象が名詞句の場合はその名詞句を含む文を指示対象と判断する。

抄録処理で文を抜き出すだけでは、要約文章として冗長な部分があったり、文間の接続が不自然な場合がある (結束性の悪化)。そこで、省略することによって結束構造を発生することができる主題や文の論旨に影響を及ぼさない表現を削除することによって、結束性を高めるとともに、一文の量的な圧縮を行なう。

本方法では、以下のような表現を削除の対象とする。

削除対象となる表現

- 接続詞 (文頭にある場合のみ)
- 修飾句
 - － 固有名詞への修飾句
 - － 例示句
 - － 話括弧 “「」” の前語句
 - － 時間格
- 副詞
 - － 量副詞 (「ほとんど」, 「ほぼ」等)
 - － 程度副詞 (「きわめて」, 「かなり」等)
 - － 時制相副詞 (「まだ」, 「もう」等)
- 丸括弧 “()” および丸括弧内語句

- 主題 (維持されている場合のみ)

以上の要約手法によって生成された要約文章の例を, 原文章, 抄録文章 (下線部分) とともに以下に示す.

原文章と抄録文章 (下線部分)

日本経済新聞 93 年 1 月 28 日の社説

改憲論の前に日本の“自画像”を描け (社説)

(1,1): 通常国会の代表質問が二十七日終わったが, 今回の特色は与野党の党首や主要役員級の質問者が異なった発想に基づく憲法改正論を展開, 慎重な姿勢をとった宮沢首相との間で踏み込んだ論議を交わした点にある.

(2,1): もとより憲法には改正条項があり, その長所と短所を点検し, 論議を深めること自体は望ましいことである.

(2,2): この意味で「神棚に上げず日常的に議論することは, 憲法を自分のものとするためにも有意義だ」(首相) といってよい.

(3,1): しかし, まず改憲ありき, といった前提で衆参両院に憲法に関する協議会を設置し, 直ちに憲法九条を中心とする改正の具体化に進むという提案 (三塚自民党政調会長) はいささか性急にすぎるといふほかない.

(3,2): まず, 広く国民各層, 各界の間で, 国際貢献, 人権, 環境権, 地方分権など幅広く総点検を進め, じっくりと時間をかけて日本の国づくりの方向はどうあるべきかを徹底的に議論するのが大事である.

(3,3): その過程から, 法律改正で足りるものと, 憲法改正なしには改善されないものが自然に区別されてくるはずである.

(4,1): 私たちが, 拙速を避け, 慎重な対処をすべきだと主張する理由を, 別の角度から整理し直してみると次の二点が挙げられよう.

(5,1): 第一に, 憲法改正の実現のためには, 高いハードルを越えねばならない.

(5,2): 憲法九六条では, 改正は衆参各議院の総議員の三分の二以上の賛成で国会が発議し, 特別の国民投票または国政選挙の際に行われる投票で過半数の賛成を必要とする.

(5,3): この厳しい条件を満たすには, 何を実現するための憲法改正で, その結果, 得られるプラス点と甘受すべきマイナス点は何か, という肝心な点をきちんと国民に説明し, 納得を得なければならない.

(5,4): 政党側に都合のいいことばかり宣伝して, 国民にとって苦痛となりうる点を隠しているようでは政治不信を招くだけである.

(5,5): この困難な過程を, 目をつぶって急ぎ足で通り抜けてはならない.

(6,1): 第二に, 改憲により自衛隊が正規の軍隊と同じ扱いとなった場合, しばしば海外に派遣されよう.

(6,2): ガリ国連事務総長は報告書で, 紛争防止のための平和維持活動 (P K O) だけでなく, 停戦合意が守られない事態に対処する平和執行部隊の創設を含む平和創造活動を提唱している.

(6,3): これは武力行使を覚悟した派遣であり, カンボジアに派遣中の P K O とは質的に異なる.

(7,1): 将来, 仮に日本が国連安保理の常任理事国になった場合, 平和執行部隊派遣に賛成しながら自国の部隊派遣を拒否することは利己的な態度として非難されよう.

(7,2): 日本は国際貢献と犠牲の調和をどこに求めるのか.

(8,1): その答えを出すには, まず日本はどのようなコストを負担して世界に寄与するのか, という国家としての自画像について国民的合意を固めねばならない.

(8,2): この意味で首相と自民党政調会長が二十七日, 党の憲法調査会で論議を深めることで合意したのは, その第一歩として歓迎したい.

(8,3): 先入観を捨て、謙虚な姿勢で論議を尽くしてほしい。

要約文章

日本経済新聞 93 年 1 月 28 日の社説

改憲論の前に日本の“自画像”を描け(社説)

(1,1): 通常国会の代表質問が二十七日終わったが、今回の特色は与野党の党首や主要役員級の質問者が異なった発想に基づく憲法改正論を展開、宮沢首相との間で踏み込んだ論議を交わした点にある。

(2,1): 憲法には改正条項があり、その長所と短所を点検し、論議を深めること自体は望ましいことである。

(7,1): 将来、日本が国連安保理の常任理事国になった場合、平和執行部隊派遣に賛成しながら自国の部隊派遣を拒否することは利己的な態度として非難されよう。

(7,2): 日本は国際貢献と犠牲の調和をどこに求めるのか。

(8,1): その答えを出すには、どのようなコストを負担して世界に寄与するのか、という国家としての自画像について国民的合意を固めねばならない。

(8,2): この意味で首相と自民党政調会長が二十七日、党の憲法調査会で論議を深めることで合意したのは、その第一歩として歓迎したい。

参考文献

- Mann, W. C. (1987). "Rhetorical Structure Theory : Description and Construction of Text Structure." In Kempen, G. (Ed.), *Natural Language Generation*, pp. 279-300. Martinus Nijhoff Publishers.
- Marcu, D. (1997). "From discourse structures to text summaries." In *Proc. of the ACL Workshop on Intelligent Scalable Text Summarization*, pp. 82-88.
- 望月源, 本田岳夫, 奥村学 (1996). "重回帰分析とクラスタ分析を用いたテキストセグメンテーション." 言語処理学会 第 2 回年次大会 発表論文集, 325-328.
- 野本忠司 松本裕治 (1997). "人間の重要度判定に基づいた自動要約の試み." 情報処理学会自然言語処理研究会報告, 120, 71-76.
- 奥村学 難波英嗣 (1998). "テキスト自動要約技術の現状と課題." 言語処理学会 第 4 回 年次大会 ワークショップ論文集, pp. 80-87.
- Paice., C. (1990). "Constructing literature abstracts by computer: Techniques and prospects." *Information Processing and Management*, 26 (1), 171-186.
- 田村直良 和田啓二 (1998). "セグメントの分割と統合による文章構造解析." 自然言語処理, 5 (1), 59-78.
- Watanabe, H. (1996). "A method for abstracting newspaper articles by using surface clues." In *Proc. of the 16th International Conference on Computational Linguistics*, pp. 974-979.
- 亀田雅之 (1997). "段落及び文間関連度を利用した段落シフト法に基づく重要文抽出." 情報処理学会自然言語処理研究会報告, 121, 119-126.

山本和英, 増山繁, 内藤昭三 (1995). “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, 2 (1), 39-55.

略歴

比留間 正樹: 1997 年横浜国立大学工学部電子情報工学科卒業. 1999 年横浜国立大学大学院工学研究科電子情報工学専攻前期課程修了, その間, 自然言語処理の研究に従事. 同年, 日本アイ・ビー・エム入社, 現在に至る.

山下 卓規: 1997 年横浜国立大学工学部電子情報工学科卒業. 1999 年横浜国立大学大学院工学研究科電子情報工学専攻前期課程修了, その間, 文章要約などの自然言語処理の研究に従事. 同年, 株式会社東芝入社, 現在に至る.

奈良 雅雄: 1996 年横浜国立大学工学部電子情報工学科卒業. 1998 年同大学大学院工学研究科電子情報工学専攻前期課程修了, 自然言語処理, 文章要約の研究に従事, 同年, 日立ソフトウェアエンジニアリング (株) 入社, 現在に至る.

田村 直良: 1985 年東京工業大学大学院博士課程情報工学専攻修了, 工学博士. 同年東京工業大学工学部助手. 1987 年横浜国立大学工学部講師, 同助教授を経て, 1995 年米国オレゴン州立大学客員教授, 1997 年横浜国立大学教育人間科学部教授, 現在に至る. 構文解析, 文章解析, 文章要約などの自然言語処理の研究に従事. 情報処理学会, 人工知能学会, 言語処理学会各会員.

(1998 年 9 月 30 日 受付)

(1998 年 12 月 21 日 再受付)

(1999 年 1 月 21 日 採録)