自然言語処理 121－17
情報学基礎 47－9
（1997. 9. 12）

段落間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

亀田 雅之
(株) リコー・研究開発本部 情報通信研究所
kameda@ic.rdc.ricoh.co.jp

要旨 文書中の重要文を抽出するために、高順位の段落中の文を優先して文の重要レベルを付与する段落シフト法を主体にした手法を述べる。段落や文内のキーワード候補群間で構成単語レベルでの重複を計数して得た段落間及び文間関連度を、正規化分母の相違に応じて参照関連度と被参照関連度とに区別し、見出しは参照関連度平均値、段落は被参照関連度平均値、文は両関連度平均値の線形和を基本スコアとして順位付けする。この上で、見出しは見出し群内での順位値により、本文中の文は段落内での順位値を段落の順位でシフトして、各々重要レベルを付与し、さらに、見出しとの関連度等で補正を行う。本手法による重要文抽出の小実験(平均12.1文, 42新聞記事)を行い、最高レベルの文で再現率約64%/適合率約71%の結果を得た。

Key-sentences Extraction based on paragraph-shift method

Masayuki KAMEDA
Information and Communication R & D Center, RICOH COMPANY, LTD.
3-2-3, Shin-Yokohama, Kohoku-ku, Yokohama, 222

ABSTRACT This paper describes a method to extract key-sentences from texts based on preference-ranking sentences within preference-ranked paragraph. Two types, a referring type and a referred-to type, of degree of relevance between two elements, determined by counting component words common to the two groupes of keyword-candidates, are used for ranking the elements, such as paragraphs or sentences. The key-sentence level assigned to a sentence is calculated by shifting the rank-value of the sentence within the paragraph by the rank-value of the paragraph including the sentence. The result of the highest level key-sentence extractions from 42 newspaper articles shows 64%-recall and 71%-precision compared to sentences selected by human readers.

1 はじめに

近年、インターネットを始めとした情報流通メディアの進展に伴い、文書情報が急激に増大しており、文書検索とともに、文書を縮約して提示する要約あるいは抄録のニーズが高まっている[1]。

[狭義の] 要約は、文書内容を理解した上で、重要情報を抽出・再構成し、要約文を生成することが必要であり、多くの難しい課題を含んでいるため、現状では、文書中の重要文を判定・抽出する抄録[広義の要約]のアプローチが取られることが多い。

重要文抽出あるいは抄録の手法としては、

(a) 文書構造の解析をベースとする手法
(b) 統計的手法
(c) 両者をハイブリッドした手法

等がある。(a) としては、文種別、照応、省略等の日本語の様々な表層的な特徴や段落内の文位置(冒頭文/最終文)を利用する手法[1]、接続詞等の手掛かり語により文間の論理的な構造を解析し、重要文を判定する手法[2]、(b) としては、単語の文書内頻度tfに基づき代表文等を抽出する手法[3]、さらに、一般語の影響を文書頻度の逆数idfで排除したtf*idf法により単語の重要度を計算し、重要度の高い単語を含む文を抄録する手法[4]、(c) としては、キーワード頻度、文体、位置、長さ等の様々なパラメータによる文重要度を学習する手法[5]などが提案されている。

我々は、簡易日本語解析系QJP[6, 7]の上に実装した日本語文書読解支援系QJR[8]の一部機能の拡張として、擬似キーワード相関法に基づく重要キーワードと重要文[2]の抽出手法を提案した[9, 以下「前報告」と呼ぶ]。擬似キーワード相関法は、2つの言語要素内の「擬似キーワード」と呼ぶ複合語レベルのキーワード候補群同士での構成単語レベルでの重複の計数により要素間の関連度を得る手法であり、これにより文書内の文間関連度マトリクスを求め、さらに、他文との関連度の平均値及び他文とどれだけ広く関係しているかというカバレージにより重要度を与え、文をランキングした。

同様の文間の関連度に基づく手法としては、文書検索等で一般的なベクトル空間モデル[10]に基づき、文を単語ベクトルで表現し、文の単語ベクトル同士の内積値を正規化して文間の結合度を求め、これをもとにユーザの注目語を含む重要文を抽出する手法[11]が報告されている。

ここでは、さらに、重要文抽出に際し、段落と見出しに着目する。

段落については、既に、段落を単位にした重要文抽出の手法[1][3]や重要段落の抽出の手法[12]等が提案されているが、重要文抽出手法の多くは、文を個々に独立して扱い、段落を考慮していない。しかし、段落(パラグラフ)は、「読み」の技法から見て、文書構成上、注目するべき重要なまとまりである[13]。

見出しについては、前報告で文間関連度を利用した関連文抽出機能の例として、簡潔な見出し文を本文中の関連文により詳細化展開する例を示した。さらに進めて、一種の要約情報でもある見出しに着目し、見出しに関連する文を重要文抽出に利用する方法が提案されている[14, 15, 16]。

本稿では、前報告の重要文抽出手法を、段落や見出しの観点を組み入れて改訂し、文に重要レベルを付与する方法について述べる。具体的には、文書は見出し群と本文段落群の組からなる1以上のブロック群からなる、という簡単な文書構造モデルの上で、段落間関連度を導入して段落のランキングを行い、上位ランクの段落中の文に優先的に高いレベルを付与する段落シフト法を用いる。さらに、重要な見出し文と強く関連する文や、重要性を示唆する機能語句を含む文を重視する等の補正を加える。

以降、第2節では、前報告で示した擬似キーワード相関法と同相関法による文間関連度を利用した重要文抽出手法を簡単にレビューし、第3節では、前報告の改訂として、段落シフト法を主体にした重要文抽出手法方法と例を示す。さらに、第4節で評価実験の結果を報告し、第5節で考察、第6節でまとめを行う。

[1] 実際、抜粋あるいは要約機能として、重要文をマークしたり、抽出する機能が一部の市販の文書管理システムやワープロソフトに装備されるようになった

[2] QJRでの重要文抽出は、飛ばし読み(Skip Reading)支援を主要な目的とした

2 文間関連度を利用した重要文抽出

擬似キーワードと擬似キーワード相関法

前報告及び本稿の手法は、文書の内容や分野を特徴付けるキーワードの候補に相当する擬似キーワードを用いた擬似キーワード相関法をベースとする。

擬似キーワードには、形態素解析系[3]により文書から抽出した名詞のうち、形式名詞や副詞名詞等の機能性名詞、数名詞、1字(和語)名詞等を除いて用いる。

擬似キーワード相関法では、擬似キーワードW_iのW_jに対する関連度$RW_{W_i}(W_j)$を次のように与える[4]。

$$RW_{W_i}(W_j) = CW(W_i, W_j)/NW(W_i)$$
$CW(W_i, W_j)$：W_iとW_jの重複構成単語数
$NW(W_i)$：W_iの構成単語数

さらに、この関連度を、文や段落、文書全体といった言語単位に含まれる擬似キーワード群に拡張し、擬似キーワード群G_iのG_jに対する関連度$RG_{G_i}(G_j)$を次のように与える。

$$RG_{G_i}(G_j) = CG(G_i, G_j)/NG(G_i)$$
$$CG(G_i, G_j) = \sum_{W_n \in G_i, W_m \in G_j} CW(W_n, W_m) \quad [5]$$
$$NG(G_i) = \sum_{W_n \in G_i} NW(W_n) \quad [6]$$

文間関連度と文重要度の指標

文内の擬似キーワード群に擬似キーワード相関法を適用する。即ち、上記のG_iを文(内の擬似キーワード群)S_iで読み替える等、'G'を'S'に置き換え、文S_iのS_jに対する関連度$RS_{S_i}(S_j)$を考える。

前報告では、この文間関連度$RS_{S_i}(S_j)$[7]を用い、次の2つの指標を導入した。

$$\text{平均関連度}：ARS_{S_i} = \sum_{S_j \in D, i \neq j} RS_{S_i}(S_j)/n'$$
$$\text{カバレージ}：CRS_{S_i} = \sum_{S_j \in D, i \neq j} \delta(RS_{S_i}(S_j))/n'$$
n'：文書内文数-1　$\delta(x) = 0(x = 0), 1(x \neq 0)$

平均関連度は他の文との関連度の平均値であり、カバレージは他の文とどの程度広く関連しているかを示す。各々は、その観点で文の重要度に関わると考え、この2つの指標の積あるいは和を重要度スコアとして、文のランキングに用い、重要文を抽出した。

[3]形態素解析にはQJPを用いる。擬似キーワードは通常は複合名詞であるが、QJPの機能の制約から構成単語に分割せずに扱う

[4]擬似キーワードは構成単語レベルで分割されていないので、実装は、擬似キーワードの{重複}文字列長を文字種ごとに定めた基本単語長で正規化し、{重複}構成単語数を模擬している。前報告では、$CW/CG/CS$を「重複文字列」、$NW/NG/NS$の代わりに$LW/LG/LS$を「文字列長」とした実装レベルの定義を示した

[5]G_i,G_jに含まれる擬似キーワード同士の重複構成単語数の総和

[6]G_iに含まれる擬似キーワードの構成単語数の総和

[7]QJRでは、擬似キーワードの重み付けに用いる修正単語頻度のためにCWの計数を行うが、この過程で副次的に文間関連度マトリクスが得られる

3 段落シフト法に基づく重要文抽出

本節では、前節の重要文抽出法に段落や見出しの観点を入れて改訂した手法を示す。

3.1 文書構造モデル

文書内の文をフラットに扱う重要文抽出の一般的な手法に対し、見出しや段落を扱うために、簡単な文書構造モデル（図1）を想定する。

このモデルでは、見出し部分と本文部分からなるブロックという単位を導入する。ブロックは大きなトピックのまとまりであり、文書は1以上のブロックからなる。また、見出し部分は見出し文群[8]、本文部分は段落群からなり、段落は1以上の文からなる[9]。

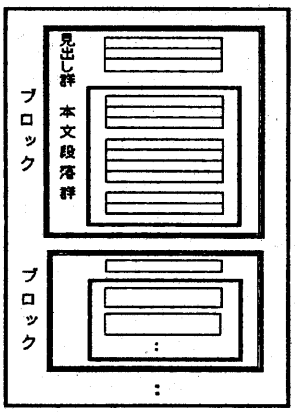


図1．文書構造モデル

この文書構造モデルを用いるために、見出し文を判定し、見出し部分と本文部分を分け、さらに見出し部分と本文部分の組をブロックとする方法を取る。

尚、段落は、改行による形式段落を用いる。

見出し文の判定

見出し文の判定は、次のような見出し文のいくつかの特徴に対応したヒューリスティックスコアの総和が閾値[10]を越えた文を見出し文とする。

- 1段落1文：10
- 文書の先頭：6
- 短い文　　：10(30字まで) ～ 0(60字まで)
- 文頭記号　：7
- {字下げ/文末語尾/句点}の省略：6/7/7

[8]多くの場合、見出し文群内の見出し文は1つである

[9]より詳細には、ブロック以上の階層もありうるが、ここではブロックまでの簡単なモデルに止めた

[10]実装では40を用いた。また、現状、1段落1文を必須条件としている

3.2 参照関連度と被参照関連度

以下の考察から、関連度を2つのタイプー参照関連度と被参照関連度ーに区別し、これらを使い分ける。

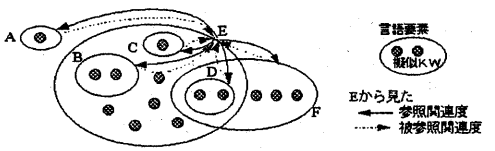
擬似キーワード群G_iとG_jについて、G_iから見たG_jとの関連度$RG_{G_i}(G_j)$と、G_jから見たG_iとの関連度$RG_{G_j}(G_i)$は、次のように与えられる。

$$RG_{G_i}(G_j) = CG(G_i, G_j)/NG(G_i)$$
$$RG_{G_j}(G_i) = CG(G_j, G_i)/NG(G_j)$$

ここで、分子の重複構成単語数CGは、i,jに対して対称だが、分母の正規化項の構成単語数NGは異なるため、関連度RGは、i,jに対して対称とはならない。そこで、G_iとG_jの関連度について、正規化項が自身のNGか、他方のNGかにより、各々G_iから見たG_jとの参照関連度、被参照関連度と呼び、区別し、参照関連度$RG_{G_i}(G_j)$は、自身のどの程度が相手を参照しているか、被参照関連度$RG_{G_j}(G_i)$は、自身のどの程度が相手から参照されているか、を示す指標と見る。

図2の観察から、擬似キーワードが少なく、そのほとんどが他の言語要素と重複するような言語要素(B,C,D)は、参照関連度(平均値)が高く、一方、多くの擬似キーワードを含み、少しづつでも他の言語要素と重複するような言語要素(E,F)は、被参照関連度(平均値)が高くなる。前者は、簡潔な見出し文、後者は、総合的な長い文がその例として考えられる。

前報告の平均関連度は、ここで参照関連度としたものの平均値である。一方、カバレージは、他の文とどの程度広く関連しているかという観点で導入したが、関連度が0でなければ、その大小にかかわらず1で扱うため、識別力が弱い。総合的な文が高くなる被参照関連度は、カバレージと同じ傾向をもつとともに、図2にみるように識別力もあることから、カバレージの代替として被参照関連度を考える。



言語要素	A	B	C	D	E	F	参照関連度平均値	カバレージ	被参照関連度平均値
A	*	0/1	0/1	0/1	0/1	0/1	.00	.00	(.00)
B	0/2	*	0/2	0/2	2/2	0/2	.20*	.20	(.04)
C	0/1	0/1	*	0/1	1/1	0/1	.20*	.20	(.02)
D	0/2	0/2	0/2	*	2/2	2/2	.40*	.40*	(.12)
E	0/10	2/10	1/10	2/10	*	2/10	.14	.80*	(.68*)
F	0/5	0/5	0/5	2/5	2/5	*	.18	.40*	(.24*)
被参照関連度平均値	.00	.04	.02	.12	.68*	.24*			

●行成分：参照関連度　：見出し的な簡潔な重要文が高くなる
●列成分：被参照関連度：総合的な長い重要文が高くなる

図2. 参照関連度と被参照関連度

この上で、本文(段落)中の文のランキングには、前報告の手法の延長上で、参照関連度平均値と被参照関連度平均値を指標として用いる一方、上記に示した参照関連度と被参照関連度の特性から、簡潔な見出し文は参照関連度平均値、擬似キーワード数が多くなる本文段落は被参照関連度平均値を用いる方針を取る。

ここで、新たに必要となる段落のランキングには、段落間関連度を導入する。

3.3 重要レベルと段落シフト法

3.1の文書構造モデルの重要文抽出において、大きなトピックのまとまりである各ブロックを独立して扱い、また、各ブロック内では、見出し部分は、本文部分の要約的な機能を担う独立した部分であると考え、見出し部分と本文部分を別々に扱う。

このように、各ブロックや見出し部分と本文部分を独立して重要文抽出を行うことから、単なる順位付けではなく、同一レベルに複数の文を割り当てられるような、文に重要レベルを付与する方法を取る。

また、本文部分は、各段落が小さなまとまりをなし、段落にも重要性の違いがあることから、重要文を選ぶ際に、その文の属する段落の重要性を考慮する。即ち、重要性の高い「キー段落」の中から優先的に「キー文」を選んでいくことを指針として、段落の重要度の順に重要文を抽出していく方法を考える。

具体的には、段落内の文の順位値に対し、その文の属する段落の順位値をベース値(ペナルティ)として加算し、重要レベルとする。段落の順位値によりレベルのベース値をシフトすることから、これを「段落シフト法」と呼ぶ（図3の棒グラフ状部分参照)。

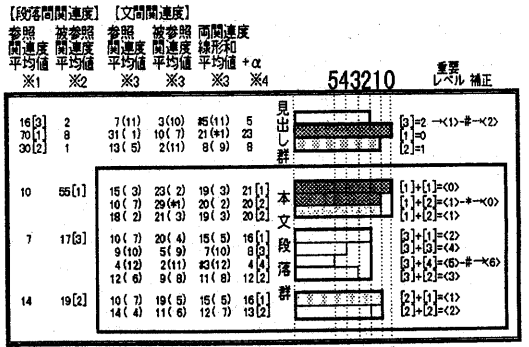


図3. 段落シフト法

3.4 重要レベルの付与

上記の方針に基づき、文に対し、0を最高レベルとする重要レベル付けを行う。

見出し文の重要レベル

見出し文は、「ブロック内での見出し文群内での順位値-1」[11]を重要レベルとし、さらに、見出し文のもつ本来的な重要性を考慮し、レベルが2以上の場合は、さらに1だけレベルアップする。

ここで、見出し文群内でのランキングのスコアには、3.2の方針から参照関連度平均値を用いる。

本文段落中の文の重要レベル

本文中の各文は、段落シフト法に基づき、「その文の属する段落のブロック内での順位値-1」と「段落内での文の順位値-1」の和を重要レベルとする。

上記と同様に、3.2の方針から、ランキングのスコアとしては、段落は、段落間関連度マトリクスによる被参照関連度平均値を用いる。文は、文間関連度マトリクスによる参照関連度と被参照関連度の両平均値[12]の線形和平均[13]を基本スコアとした上で、さらに、次に示す重要文を示唆する語句等の重要文手掛かりスコアを加算する[14]。

重要文手掛かりスコア

上記の重要文手掛かりのために、次のような語句や文体に対するスコアを用いる。

- **直接手掛かり語句**：重要[5:重要だ,大切だ,..],要約[7:要するに,..],結果[5:従って,結局,..],原因/理由[5:何故なら,..;(前文に2)],..
- **間接手掛かり語句**：希望[3:たい,ほしい,..],義務[3:べきだ,..],..
- **手掛かり文体**：非用言文[1:体言止め,..],主題文[2:～は],新情報文[4:(主節中に)～が]

重要レベルの補正

また、見出しとの関連が強い文の重要性の観点他を取り入れるために、重要レベルに、次のようなレベルのアップ(－)あるいはダウン(＋)の補正を行う。

- －1：レベル0の見出し文と強い関連度をもつ文
- －1：基本スコアあるいは被参照関連度平均値による順位が上位の文
- ＋1：基本スコアが最高値と比べ充分小さい文

[11]順位値から1を減ずるのは、0を最高レベルとするため
[12]各々の関連度は100で正規化
[13]実装では、1:1の平均
[14]実装では、重要文手掛かりスコアの1/2を加算している

【段落間関連度マトリクス】

i＼j	123456	<A>	<B>	<C>	段落頭テキスト
1	*00800	16[3]	2	9(6)	!通常兵器関連の工業製品
2	0*0#69	70[1]	8	39(2)	!輸出規制が始動
3	00*#50	30[2]	1	15(5)	!４ヶ国対象
4	121*33	10	55[1]	58(*1)	-通常兵器の部品や加工機械に転用できる工
5	0104*2	7	17[3]	24(3)	-規制対象となる国は、イラン、イラク、リ
6	02063*	14	19[2]	24(3)	-輸出貿易管理令などに基づいて定められた

【文間関連度マトリクス】

i＼j	123456789012	<A>	<B>	<C> /+α	レベル	文頭テキスト
1	*00800000000	7(11)	3(10)	5(11)/ 5[3]	<2>	!通常兵器関連の工業製品
2	0*0686300363	31(*1)	10(7)	21(*1)/23[1]	<0>	!輸出規制が始動
3	00*055500000	13(5)	2(11)	8(9)/ 8[2]	<1>	!４ヶ国対象
4	420*32100141	15(3)	23(2)	19(3)/21[1]	<0>	-通常兵器の部品や加工機
5	0212*3210111	10(7)	29(*1)	20(2)/20[2]	<0>	英米などの主要先進七カ
6	03135*3001#1	18(2)	21(3)	19(3)/20[2]	<1>	Ｇ７は既に対象となる品
7	011132*11111	10(7)	20(4)	15(5)/18[1]	<2>	-規制対象となる国は、イ
8	0000303*0004	9(10)	5(9)	7(10)/ 8[3]	<4>	北朝鮮は、既に共産圏と
9	00002020*200	4(12)	2(11)	3(12)/ 4[4]	<6>	また、イラクは経済制裁
10	020222302*20	12(6)	9(8)	11(8)/12[2]	<3>	今回の措置で輸出に大き
11	0203221001*2	10(7)	19(5)	15(5)/16[1]	<1>	-輸出貿易管理令などに基
12	0202222304*	14(4)	11(6)	12(7)/13[2]	<2>	その際、特定地域に指定

※ <A>/<B>/<C>:参照関連度平均値/被参照関連度平均値/両平均値の線形和平均
α:重要文手掛かりスコア　00/[N]:全体での順位/群内順位
!/-:見出し文/本文段落(先頭)

図4.段落間及び文間関連度マトリクスとランキング情報

重要レベル付与の例

図6(付録)の文書に対する段落間関連度マトリクス、文間関連度マトリクス[15]とそれらのランキング情報（含 重要レベル）を図4に示す。

段落間関連度のランキング情報としては、<A>に、参照関連度平均値と[ ]内に同値による見出し3文の順位値、<B>に、被参照関連度平均値と[ ]内に同値による3段落の順位値が示されている。

文間関連度のランキング情報としては、<A>,<B>,<C>に、参照関連度平均値、被参照関連度平均値、両平均値の線形和平均と( )内に各々の値による12文中での順位値、<C>の右(＋α欄)には、<C>の値に重要文手掛かりスコアを加えたスコア値と[ ]内に同値による段落内での文の順位値、さらに、< >内に重要レベル値が示されている。

重要レベルは、見出し3文は，見出し文群内での順位値(上段<A>欄)から1減じた値、本文段落内の各文は、その文の段落内順位値(文間マトリクスによる<C>＋α欄)から1減じた値に、その文の属する段落の順位値(段落間マトリクスによる<B>欄)から1減じた値を加えた値となる。ただし、図3右の補正欄にあるように、第5文は<B>の第1位文のためにレベルアップ、第1文(見出し)と第9文は<C>の値が最大値に比べ小さいためにレベルダウンされている。

重要レベル欄から、最高レベル0の文は、第2文(見出し文)と第4、5文となる。さらに、準最高レベル1では、第3文(見出し文)と第6、11文が加わる。

[15]図中のマトリクスの各要素(関連度)は、最大値を10で正規化し、1桁の数値で示してある。10の場合は'*'で示した

4 評価実験

42新聞記事(平均12.1文)に対する重要文抽出実験を行ない、2人の被験者により抽出された重要文[平均抽出文数:ともに3.6文]に対する再現率と適合率を求めた。

重要文抽出としては、本手法(Ver.2)による最高レベル0のみ及び準最高レベル1まで(0及び1)の2通りの抽出と、比較のために、ほぼ同じ抽出文数となる、前報告の手法(Ver.1)による上位33%及び50%の文数の抽出、さらに、市販の2ソフトA,Bの同等機能により各々上位3文及び5文の抽出を行った。

図5及び表1a、表1bに、抽出文数が3文及び5文程度になる各々の重要文抽出について、2つの正解セットに対する再現率(R:Recall)と適合率(P:Precision)を示す。ただし、表では、2つの正解セットに対する再現率と適合率の平均値を示す。また、指標は、

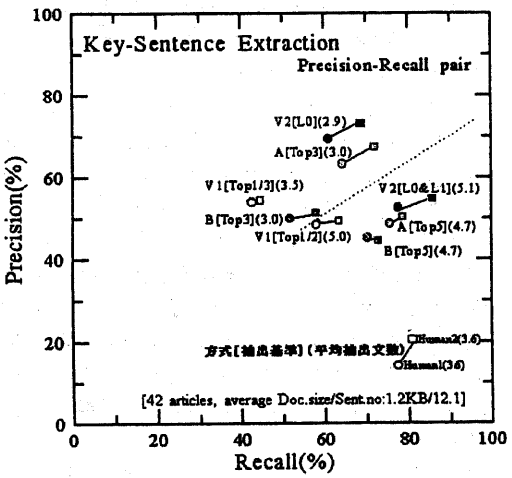


図5.重要文抽出の再現率と適合率

方式	抽出基準	(抽出文数)	再現率/適合率	指標
Ver.2	Level 0	(2.9)	64.8%/71.2%	.678
Ver.1	Top33%	(3.5)	43.3%/54.3%	.482
A	Top3	(3.0)	68.5%/65.0%	.667
B	Top3	(3.0)	54.7%/50.5%	.525

表1 a. 重要文抽出[3文程度]の再現率と適合率

方式	抽出基準	(抽出文数)	再現率/適合率	指標
Ver.2	Level 0,1	(5.1)	81.4%/53.6%	.646
Ver.1	Top50%	(5.0)	60.6%/49.0%	.542
A	Top5	(4.7)	77.1%/49.1%	.596
B	Top5	(4.7)	71.7%/45.1%	.552

表1 b. 重要文抽出[5文程度]の再現率と適合率

$(\beta^2+1)P\cdot R/(\beta^2\cdot P+R)$による評価値である[16]。

本手法の結果は、前報告の手法より明らかに、精度の向上を得た。また、評価指標によれば、2つのソフトA,Bのうち、Bに対しては十分に良く、Aに対しては若干良い(あるいは同等程度の)結果となった[17]。

5 考察

文書構造モデルと段落シフト法

簡単ではあるが、文書構造モデルに基づき、大きなトピックをもつブロックごとに重要文抽出を行うとともに、段落シフト法により「キー段落」の重要文を優先し、また、見出しに着目し、見出し部分を本文部分から独立して扱った。これらにより、全体の文をフラットに扱うボトムアップ的手法に比べ、トップダウン的で重要文の取りこぼしが少ない手法となった。

関連度マトリクスの利用

前報告では、QJRのキーワードの重み付け処理の過程で副次的に生成される関連度マトリクスを利用して、重要文抽出と関連文抽出を示したが、重要文抽出で重要度スコアとして用いた平均関連度とカバレージは、関連度マトリクスの特性を十分に生かしておらず、基本的には単語頻度法[18]と大きな相違はないと考えられる。

本手法では、関連度を参照関連度と被参照関連度に区別し、見出し、段落、本文の違いに応じてこれらを使い分けるとともに、見出し文との関連文抽出も重要文抽出に利用する等、関連度マトリクスの特性を生かした利用を行った。

重要レベル

重要文抽出では、抽出量をサイズ(文数)や文書サイズに対する割合で指定するのが一般的である。サイズ指定では、ほぼ一定量の抽出が得られ、割合指定では、文書サイズに応じた抽出が得られる。

本手法では、重要レベルにより、レベル0/1を最重要/準重要とするような重要基準に応じた抽出を与える。抽出文数は、レベル0で2文×ブロック数、レベル1までで4文×ブロック数が目安となる。

しかし、基本的に重要レベルの補正で加点主義(レベルアップ)を取っているので、これより多めになる問題がある。また、サイズ/割合指定の形式的要求も

[16] $\beta=1$を用いた

[17] 両ソフトの手法は不明であるが、Aは、文書のほぼ前半から抽出する傾向がある。Aの手法が文の位置情報を利用し、前方を優先しているとすると、最初の段落が重要なリード段落であることが多い新聞記事は、Aにとって有利な対象だった可能性がある

[18] 前報告も本手法も、idfは効果は考慮されていない

強い。こうしたことから、同一レベル内で改めてランキングすることで、これらの問題や要求に対応することが現実的である。

大規模文書での問題

現状、新聞記事程度のサイズの文書を対象にしたが、大規模な文書では、関連度マトリクスの操作/作成、文書構造の観点から次のような問題が出てくる。

スコアとしている関連度の平均は、段落間及び文間関連度マトリクスを文書全体で1つづつ作成し、全体での関連度の平均として得た上で、ブロック内の見出し群内、本文段落群内、段落内の文群内での比較に利用している。しかし、大規模な文書になると、0部分が多いスパースなマトリクスとなり、平均値が小さくなる問題がある。これに対しては、部分マトリクスを取り出し、平均を取る等の方法の検討が必要である。

一方、擬似キーワード相関法では、文書サイズの二乗の処理時間がかかる。この点でも、大きな文書に対しては、全体の関連度マトリクスを作成するのは不利である[19]。従って、大規模な文書については、事前に適切な文書分割を行って、それぞれの部分の関連度マトリクスを作成することが望ましい。

また、章節などの本格的な階層文書構造をもった文書では、ここで用いた簡易な文書構造モデルは不十分である。このためにも、階層的な文書の分割や構造の認識が求められる。

6 まとめ

重要文抽出に際し、小さなまとまりを示す段落や一種の要約情報である見出しに着目し、簡単な文書構造モデルの上での重要文抽出手法を提案した。段落については、キー段落から優先して重要文を抽出する指針に基づき、段落の順位により文の重要レベルに段階的にペナルティを与える段落シフト法を示し、また、見出し文については、その重要性から見出し部分からの重要文抽出を独立させるとともに、見出しと関連が強い本文のレベルを上げる補正を示した。

本手法では、擬似キーワード相関法に基づく段落間及び文間関連度を基本情報とし、さらに、関連度を正規化の仕方により参照関連度と被参照関連度を区別し、見出しと段落及び文のランキングでは、それらの特徴に応じて使い分け、また、関連度を重要見出しの関連文の判定にも用いた。

本手法に基づく新聞記事を対象にした小規模な重要文抽出実験により、前報告の手法や既存システムでの類似機能に対し、比較的良い抽出精度を得た。

今後は、考察に示した部分マトリクス、文書分割、階層的抽出といった大きな文書への対応を中心に展開を図る予定である。

謝辞 QJP及びQJRの高速化等の改良により重要文抽出等の機能の実用性能を大幅に高めてくれたシステムソリューション事業本部の後藤淳之氏、重要文抽出の正解作成に協力してくれた情報通信研究所の望主雅子氏に感謝する。

[19] ただし、高速化改良によりほぼ問題のない処理時間になっている。Pentium 133MHzのPCで、100KBテキストの処理時間は、形態素解+構文解析のQJP処理時間が約10秒、関連度マトリクス作成を含むQJR処理が約5秒である

参考文献

[1] 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, pp.39-55, 1995.

[2] 住田一男, 伊藤悦雄, 三池誠司, 武田公人：対話的抄録生成機能をもつ文書検索システム, 情報処理学会 ヒューマンインターフェース 52-3, 1995.

[3] 小部正人： 文章抄録装置, 特開昭61-117658, 1984.

[4] Klaus Zechner : Fast generation of abstracts from general domain text corpora by extracting relevant sentences, COLING'96, pp.986-989, 1996.

[5] Hideo Watanabe : A Method for Abstracting Newspaper Articles by Using Surface Clues, COLING'96, pp.974-979, 1996.

[6] 亀田雅之： 軽量・高速な日本語解析ツール『簡易日本語解析系Q_JP』, 言語処理学会 第1回年次大会, 1995.

[7] Masayuki KAMEDA : A Portable & Quick Japanse Parser : QJP, COLING'96, pp.616-621, 1996.

[8] 亀田雅之： 日本語文書読解支援系QJRの検討, 情報処理学会 自然言語処理研究会報告 110-9, 1995.

[9] 亀田雅之：擬似キーワード相関法による重要キーワードと重要文の抽出, 言語処理学会 第2回年次大会, 1996.

[10] Gerard Salton : Introduction to Modern Information Retrieval, McGraw-Hill, 1983.

[11] 福本淳一： 文間関連度に基づく内容抽出手法, 言語処理学会 第3回年次大会, 1997.

[12] 福本文代, 福本淳一, 鈴木良弥：文脈依存の度合を考慮した重要パラグラフの抽出, 自然言語処理 Vol.4 No.2, pp.89-109, 1997.

[13] 小河原誠： 読み書きの技法, ちくま新書, 1996.

[14] 亀田雅之： 文書表示装置及び文書要約装置及びディジタル複写装置, 特開平06-259434, 1993.

[15] 仲尾由雄： 見出しを利用した新聞・レポートからのダイジェスト情報の抽出, 情報処理学会 自然言語処理研究会報告 117-17, 1997.

[16] Ryo Ochitani, Yoshio Nakao, Fumihito Nishino : Goal-Directed Approach for Text Summarization, ISTS'97(Intelligent Scalable Text Summarization), pp.47-50, 1997.

付録：重要文抽出例

図6,図7,図8に、最重要文(最高レベル0)は太字＋下線、準重要文(準最高レベル1)は下線でマークした重要文抽出例を示す。

通常兵器関連の工業製品
輸出規制が強化
4ヶ国対象

通常兵器の開発や加工機械に転用できる工業製品の輸出規制が二十日、日本でも始まった。基準などの主導先進七カ国（G7）の合意に基づいた規制であり、イラクなど懸念国を対象にして、対共産圏輸出統制委員会（ココム）のリストを見直する形。G7は既に対象となる品目、国を広げるための話し合いを始めており、冷戦終結で変わる新たな輸出規制に発展しそうだ。

規制対象となる国は、イラン、イラク、リビア、朝鮮民主主義人民共和国（北朝鮮）の四カ国である。北朝鮮は、既に共産圏として特定地域に指定されているため、新たに追加されるのは三カ国である。また、イラクは経済制裁で業輸措置が取られている。今回の措置で輸出に大きな変化が出るのはイランとリビアの二国になりそうだ。

輸出貿易管理令などに基づいて定められたコンピュータや工作機械などの規制品目を輸出する時には、通産省に許可申請を行う。その際、特定地域に指定されている国に対しては、明らかに民生用とわかる場合でなければ許可が下りず、事実上、規制品は輸出できない。

出典：朝日新聞 1993年1月21日

図6.重要文抽出例1

イメージコミュニケーション――このリコーの新しい、コーポレートスローガンは、ビジネスコミュニケーションの未来を、そしてお客様に新たなベネフィットを提供していくために、私たちが取り組んでいく重点領域を表しています。

デジタル技術やネットワーク技術の進歩は、インターネットをはじめとする世界規模の新しいコミュニケーションを生み出し、ビジネスの進め方も大きく変化しました。そこで待たれているのが「人にとって、よりわかりやすく、柔軟性のある」コミュニケーション手段の登場です。

リコーがいま進めているのは「絵や写真・文字・数字などのイメージ情報を、誰もが簡単に加工・処理したり、ネットワークで思いのままに」コミュニケーションできる環境づくり。そこでは時間や空間を超えて、仕事をする人々のパワーをひとつにできます。

リコーは、コンピューターの情報と、人間が直感的に理解しやすいイメージ情報をつなぐデジタル画像処理を駆使して、コミュニケーションを豊かにする製品、サービス、総合的な環境をお届けしたいと考えています。私たちの未来は、もう、はじまっています。

出典：リコー ホームページ

【段落間関連度マトリクス】

i＼j	1234	<0>
1	4811	36 [1]
2	5413	24 [3]
3	1143	10 [4]
4	533*	25 [2]

【文間関連度マトリクス】

i＼j	1234567	<0>+α	
1	*3231041	32[1]	<0> -イメージコミュニケーション
2	4421020	17[1]	<2> -デジタル技術やネットワーク
3	84*0040	17[1]	<2> そこで待たれているのが「人
4	210*050	11[1]	<3> -リコーがいま進めているのは
5	0000*00	1[2]	<4> そこでは時間や空間を超えて
6	42230*0	22[1]	<1> -リコーは、コンピューターの
7	000000*	6[2]	<2> 私たちの未来は、もう、はじ

図7.重要文抽出例2

最新鋭ＭＰＵ搭載パソコン、ＩＢＭ、来夏発売
低価格・高速処理を実現

【ニューヨーク１０日＝松本元裕】

米ＩＢＭは十日、最新鋭マイクロプロセッサー（ＭＰＵ、微小回路高集積装置）「パワーＰＣ」を搭載したパソコンを来年夏に発売することを明らかにした。まずノート型を発売、続いてデスクトップ型二機種を投入する。低価格で処理速度の高いパワーＰＣは、ＩＢＭ再生のカギを握ると言われるコンピューターの心臓部品。同社がパワーＰＣ搭載パソコンの商品計画を明らかにしたことで、他の日米欧のパソコン各社も対抗策を迫られることになりそうだ。

発売する三機種はＣＤ－ＲＯＭ、マイク、ステレオオーディオ、音声認識機能を標準装備して、マルチメディア機能を高める計画。ＯＳ（基本ソフト）はＩＢＭの「ＯＳ／２」の他、米マイクロソフトの「ウィンドウズＮＴ」、サンマイクロシステムズの「ソラリス」などにも対応できるようにする。

パワーＰＣはＩＢＭ、アップルコンピュータ、モトローラの三社が共同開発したＲＩＳＣ（縮小命令セットコンピュータ）型ＭＰＵ、パソコン用ＭＰＵ市場で事実上の標準機種になっているインテル製ＭＰＵに対応するための商品で、低価格・高速処理が特徴だ。

パソコン業界二位のアップルがパワーＰＣ搭載パソコンを来年発売する計画を発表。同一位のＩＢＭはパワーＰＣ内蔵のワークステーションをすでに発売しているが、パソコンについては製品計画を明らかにしていなかった。

ＩＢＭはパワーＰＣを外販するだけでなく、搭載パソコンの技術仕様を外部に有償で公開、パワーＰＣ搭載パソコンのファミリー作りを進める計画。アップルなどと合わせたパワーＰＣ搭載パソコン全体で、世界市場に占めるシェアを最低二〇％程度までもっていきたい考え。

出典：日本経済新聞 1993年11月1日

図8.重要文抽出例3