見出し一致度と文の相対位置を利用した記事重要文抽出方式

Sentence extraction using the similarity between title and sentence and the relative position of sentence

松村　繁男[*1]　　山田　剛一[*2]　　絹川　博之[*2]　　中川　裕志[*3]

[*1] 東京電機大学大学院工学研究科　[*2] 東京電機大学工学部　[*3] 東京大学情報基盤センター

１．はじめに

近年，パーソナルコンピュータ，携帯電話などの情報機器と，それらを相互に接続するインターネットやイントラネットの急激な発達・普及により，電子メール，Web，ネットニュース，電子書籍など電子化されたテキスト情報が氾濫する状況となっている．また同時に，情報検索システムもその適用分野やデータベースサイズを急速に拡大しつつある．これに伴い，膨大な情報の概要や内容を短時間で把握することが必要となっており，これを補助する技術として，テキスト自動要約技術が注目されている[1].

従来の研究では，要約手法として，文書中の表層的な情報を基に重要と判断される文を抽出する重要文抽出型の手法が多く用いられている．これらの情報についてPaiceは，①文書中のキーワードの出現頻度，②文書中あるいは段落中での位置情報，③文書の見出しの情報，④文書中の文間関係を解析したテキスト構造情報，⑤文書中の手がかり表現，⑥文書中の文あるいは単語間のつながり情報，⑦文書中の文間の類似性情報，の 7 つに分類している[2]．そして，①とその他の表層情報による重み補正を組み合わせ，重要文を抽出する手法が多く提案されている．また，①を軸としない手法としては，③のみを用いた手法[3]や，②⑦を組み合わせた手法が提案されている[4]．しかし，単語重みを考慮しないこれらの 2 つの手法はテストコレクションによる評価を行っていないため，定量的な評価を行うことができていない．そして，②については，経験的に重み補正に有効であるとされているが，統計的な実証はされていない.

本研究では，見出し内容語の文書内における出現分布の傾向の調査を行い，その結果を踏まえた上で，(1)見出し情報と，(2)段落または文(以下，段落／文と略す)の位置情報，(3)文間関連度を組み合わせた，見出し一致度と文の相対位置を利用した重要文抽出方式を提案する．提案方式は，文書内の単語の出現頻度を考慮しないという点で，従来からの提案手法と異なる．本論文では，見出し内容語の文書内における出現分布の傾向がNTCIR Workshop2 TSC1の重要文抽出型要約タスク(Task A1)用テストコレクション(以下，テストコレクションと略す)における重要文の出現分布の傾向と類似していることに着目した，提案方式の有効性を同テストコレクションを用いて評価し，考察する.

２．見出し内容語の出現傾向

2.1　見出し内容語の本文内出現傾向

新聞記事の見出しに含まれる内容語の本文内の出現分布を集計し，テストコレクションの重要文の出現分布傾向と比較する．見出し内容語出現分布の集計処理手順は，以下のとおりである.

(1) SGML 化した記事データから 1 記事分のデータを抽出する.
(2) 抽出した記事の見出しから“［社説］”のように複数の見出しに共通する定形表現を除去した後，形態素解析を行い，見出し内容語を抽出する．なお，本研究では形態素解析器として茶筌[5]を用い，内容語として名詞，動詞，形容詞，未知語を選出している.
(3) 抽出した記事の本文を，段落／文に分割する.
(4) 段落／文ごとに形態素解析し，内容語を抽出する.
(5) 見出しに含まれる内容語が，記事本文の第何段落目に出現しているのか，第何文目に出現しているのかを集計する．しかし，段落／文の数は記事によって違いがあるため，正規化が必要となってくる．そこで，段落／文の平均値を最大長として，正規化を行う．それぞれの正規化の式は，(式 1)に示す.

$$Norm(i) = \frac{Real(i)}{Len} \times Ave \quad \text{(式 1)}$$

Real(*i*)は実際の段落／文の位置($1 \leqq i \leqq Len$)，*Len* はそれぞれの全体の長さ，*Ave* はそれぞれのジャンルの平均値である．それぞれのジャンルの平均値は表 1 に示す.

表 1　それぞれのジャンルにおける平均値

ジャンル	対象記事数 (テストコレクション)	平　均 段落数	平均 文数	平　均 内容語数
解説	6891　(47)	7	19	10
社説	11096　(28)	7	17	9
社会	99901　(15)	3	9	10

これらを用いて，それぞれの正規化した位置 *Norm*(*i*) を求める．ただし，この式による正規化は，最後尾を基準としたもので，これのみでは正確に集計したとはいえない．そこで，先頭を基準とした正規化も行うことにした．これにより，正確に出現分布の傾向を集計できるようになる．先頭基準の正規化の式を(式 2)に示す.

$$RevNorm(i) = Ave - \left(\frac{Len - Real(i) + 1}{Len} \times Ave \right) + 1 \quad \text{(式 2)}$$

したがって，実段落が 1，全体の長さが 5，正規化最大長が 10 だった場合，正規化段落として 2 と 1 が与えられることになる.

(6) 形態素解析した見出しの内容語が，段落／文のどの位置に出現しているかを集計する.

2.2　テストコレクション内の重要文出現傾向

次にテストコレクション内の重要文出現分布の集計処理手順について述べる.

(1) テストコレクションより，記事 ID，重要文を得る.
(2) 取得した記事 ID を持つ記事データを抽出し，記事本文を段落／文に分割する.

(3) テストコレクションより取得した重要文が，本文中のどの段落に属する文なのか，どの位置の文なのかを集計する．段落／文の大きさは(式 1)，(式 2)を用いて正規化する．

2.3 見出し内容語の出現傾向の比較予備実験

テストコレクション以外の記事データにおける見出しの内容語出現分布傾向と，テストコレクション内の重要文出現分布傾向についての比較予備実験を行う．

2.3.1 比較予備実験対象データ

(1) 見出し内容語出現分布傾向集計データ：CD 毎日新聞-1994 年,1995 年,1998 年,2000 年版
  (a) 解説：6844 件
  (b) 社説：11068 件
  (c) 社会：99886 件

(2) テストコレクション重要文出現分布傾向集計データ：NTCIR Workshop2 TSC1 の重要文抽出型要約タスク(Task A1)で用いられたテストコレクション
  (a) 解説：47 件
  (b) 社説：28 件
  (c) 社会：15 件

(1)の集計には，1 つ以上の見出し内容語および段落で構成されている記事を対象としている．形態素解析器として ChaSen Version2.3.3 を用い，名詞，動詞，形容詞，未知語を内容語とした．辞書等の編集は行っていない．

2.3.2 比較予備実験結果

上記に述べたデータを用いてそれぞれの出現分布傾向の比較評価を行った．比較評価の結果を表 2 に示す．なお，出現分布値の高い上位 3 段落は網掛けで表している．社会面については，最上位のみを網掛けしている．

表 2　段落単位における集計結果

	解説		社説		社会	
	(1)	(2)	(1)	(2)	(1)	(2)
第 1 段落	**65050**	273	**76820**	**273**	**2740578**	**309**
第 2 段落	**52564**	**309**	**81722**	**277**	1400550	208
第 3 段落	41160	**285**	65306	256	1400116	177
第 4 段落	41342	284	62236	232		
第 5 段落	41198	281	62502	221		
第 6 段落	50090	**298**	**76770**	232		
第 7 段落	**52768**	258	75504	**349**		

※網掛け：出現分布値の高い上位 3 件(社会は最上位のみ)

2.3.3 比較予備実験結果の考察

見出し内容語の出現傾向とテストコレクション内の重要文出現傾向の比較評価結果より，上位段落の出現位置が近接していることがわかる．このことより，重要文抽出の際に，見出しと本文の段落／文単位での一致度を用いるということは，有効なのではないかといえる。

3．見出し一致度と文の相対位置を利用した記事重要文抽出方式

前章での比較予備実験結果をふまえ，見出し一致度と文の相対位置を利用した重要文抽出方式を提案する．

これまで我々は，社説に関し，見出しとの一致度を基にした重要文抽出方式[6]を提案し，テストコレクションを用いての評価を行った．この結果，同テストコレクションを用いた最良結果を上回ることができ，提案手法の有効性を示した．しかしこの手法は，社説のみに限定されているということや抽出できる重要文の数が極端に少ないという欠点があった．

今回は，これらの欠点の解消を目指して，見出しとの一致度に加えて，段落／文の相対位置情報，文の関連度，段落／文の絶対位置情報を組み合わせた重み付け方式を提案している．この拡張により，どのジャンルにも適用可能となり，抽出可能な文の数が限定されるという問題点は解消することができた．以下にそれぞれの重み付けについて説明する．

3.1 見出しとの一致度に基づく重み付け

記事の見出しとは，本文の極端な要約である．したがって，見出しと一致度の高い文は重要文である可能性が高いといえる．そこで，見出しと文の間で共通する内容語の異なり数を一致度として，文に重み付けを行う．

3.2 段落／文の相対位置に基づく重み付け

見出し内容語の出現分布の集計結果より，段落／文の位置によって見出し内容語の出現分布の傾向は変わってくる．また，この結果は，テストコレクションの重要文出現分布傾向と類似している．この特徴を利用し，集計結果より第 1 段落／文を基準とした見出し内容語出現分布の相対値を文の重みとして用いる．相対値の算出式は(式 3)のとおりである．また，それぞれの相対値を表 3 に示す．*NormFreq*(*i*)は *i* 段落／文の出現分布値，*Rel*(*i*)は *i* 段落／文の相対値である．

$$Rel(i) = \frac{NormFreq(i)}{NormFreq(1)} \qquad \text{(式 3)}$$

表 3　段落・文の相対値

段落相対値				文相対値			
ジャンル	解説	社説	社会	ジャンル	解説	社説	社会
第 1 段落	1.000	1.000	1.000	第 1 文	1.000	1.000	1.000
第 2 段落	0.808	1.064	0.511	第 2 文	0.694	0.842	0.484
第 3 段落	0.633	0.850	0.511	第 3 文	0.662	0.774	0.472
第 4 段落	0.636	0.810		第 4 文	0.540	0.701	0.271
第 5 段落	0.633	0.814		第 5 文	0.502	0.698	0.318
第 6 段落	0.770	0.999		第 6 文	0.469	0.639	0.251
第 7 段落	0.811	0.983		第 7 文	0.468	0.631	0.240
				第 8 文	0.468	0.642	0.210
				第 9 文	0.457	0.605	0.240
				第 10 文	0.482	0.617	
				第 11 文	0.460	0.604	
				第 12 文	0.462	0.604	
				第 13 文	0.465	0.621	
				第 14 文	0.461	0.633	
				第 15 文	0.490	0.708	
				第 16 文	0.491	0.762	
				第 17 文	0.586	0.861	
				第 18 文	0.552		
				第 19 文	0.640		

文への重み付けは，以下の手順によって行われる．

(1) 文の属する段落／文の位置を取得する．
(2) 取得した段落／文の位置による正規化値を算出する．正規化値は先頭基準と最後尾基準の値を算出する．
(3) それぞれの正規化値における相対値の合計を文の重みとして付与する．

3.3 文の関連度に基づく重み付け

関連度とは，亀田の提案した重み付け方式で，参照関連度と被参照関連度という 2 つの指標から成り立っている。参照関連度の算出式を(式 4)に，被参照関連度の算出式を

(式 5)に示す.

$$RS_{S_i}(S_j) = \frac{CW(S_i, S_j)}{NW(S_i)} \qquad \text{(式 4)}$$

$$RS_{S_j}(S_i) = \frac{CW(S_j, S_i)}{NW(S_j)} \qquad \text{(式 5)}$$

$CW(S_i,S_j)$は, 文S_iと文S_jの重複構成単語数を表し, $NW(S_i)$は文 S_i の構成単語数を表している．参照関連度は，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, を示す指標である.

構成単語数が少なく, そのほとんどがほかの文の構成単語と重複するような文は，参照関連度が高く，一方，多くの構成単語をもち, 少しずつでもほかの文の構成単語と重複するような文は，被参照関連度が高くなる．前者は，簡潔で重要な文, 後者は総合的な長い文がその例として考えられる．ここでは，これらの指標の線形和平均をとり，その値を関連度としている．関連度 *ARS* は(式 6)のように定義される．*D* は文書全体を表している.

$$ARS_{s_i} = \sum_{S_j \in D, i \neq j} \frac{\left(RS_{S_i}(S_j) + RS_{S_j}(S_i)\right)}{2} \qquad \text{(式 6)}$$

関連度は, その文が文書全体のどの程度の情報を網羅した文なのかを知る尺度となる.

3.4　段落／文の絶対位置に基づく重み付け

文書の重要なことは，初めの方に書くことが多いと経験的にされている．そこで，正規化しない段落／文の位置，すなわち絶対位置における文の重み補正を行った．補正方法は以下のとおりである.

(1)　文書中の第 1 段落に含まれる文の文重みを 2 倍する.
(2)　各段落内の第 1 文の文重みを 2 倍する.

したがって，第 1 段落の第 1 文は，文の重みが 4 倍されることになる.

3.5　重要文の重み付け

文の重みは，見出しとの一致度，段落／文の相対位置重み，文の関連度，段落／文の絶対位置補正を組み合わせることによって決定される. 重みの組み合わせを図 1 に示す. 合計 12 通りの文重み付け方式における性能評価を行う.

4．提案した重要文抽出方式の評価実験

テストコレクションを用いて, 提案した重要文抽出方式の精度を評価する.

4.1　実験対象データ

本実験では，NTCIR Workshop2 TSC1 の重要文抽出型要約タスク (Task A1) で用いられたテストコレクションを使用している. これは, 毎日新聞記事データベースの社説，解説面からの 60 記事を対象とした Dry run データと，社説，社会面からの 30 記事を対象とした Formal run データからなる．データとして，要約する文書番号と，3 種の要約率(10%，30%，50%)に対する抽出文数と，人手により抽出された正解要約が与えられている. 形態素解析器として ChaSen Version2.3.3 を用い，名詞，動詞，形容詞，未知語を内容語とした．辞書等の編集は行っていない．また，今回はジャンル毎の文相対位置重みの効果を調べるため，テストコレクションを解説，社説，社会に分類して評価を行った.

4.2　評価方法

評価対象システムによる要約対象文書の要約結果と正解要約を文単位で比較し，F 値を算出する．要約率ごとに 30 文書の平均 F 値を算出し評価を行う．F 値は(式 7)により定義される.

$$F = \frac{2PR}{P+R}$$
$$P = \frac{\text{作成された要約中の正解文数}}{\text{生成された要約の総文数}} \qquad \text{(式 7)}$$
$$R = \frac{\text{作成された要約中の正解文数}}{\text{正解要約の総文数}}$$

ただし本評価では, 常にテストコレクション中で指定された数の重要文を抽出するため，常に *P*=*R*=*F* となる.

4.3　実験結果

それぞれの組み合わせパターンにおける性能評価結果を表 4 に示す.　表では，ジャンル毎の組み合わせパターンにおける上位 5 方式と Formal run データ全体での最良結果, NTCIR-2 TSC1 参加システムにおける最良結果を示す.

4.4　実験結果の考察

性能評価実験の結果より，NTCIR-2 TSC1 参加システムの最良結果に比べ，解説は，要約率 50%で 4.0%，Average で 0.3%上回り，社説は，要約率 10%で 2.4%，社会は，要約率 10%で 4.2%, 要約率 30%で 8.7%, 要約率 50%で 3.6%, Average で 7.1%上回った. 全体では, 要約率 10%で 2.9%, 要約率 50%で 0.4%, Average で 2.3%上回ることができた. このことより，単語の重み付けを用いず，見出しとの一致度と文の相対位置情報を用いた提案方式の有効性を示すことができた. 以下にそれぞれの重みにおける効果について述べる.

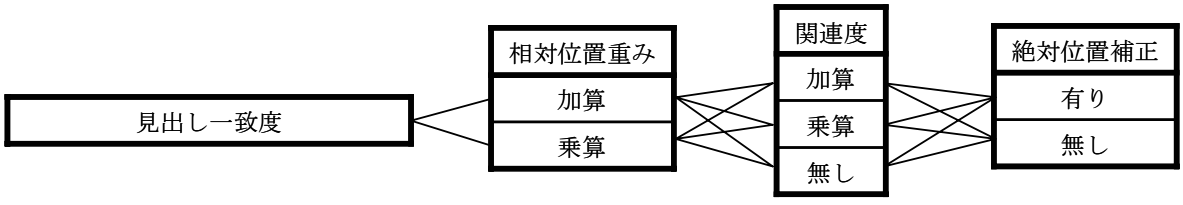

図 1　文の重み付けの組み合わせ

(1) 見出し一致度と段落／文の相対位置重みについて

見出し一致度と段落／文の相対位置重みについては，関連度などを組み合わせた重み付け方式に劣っている．したがって，この二つの重みは他の重みと組み合わせていかないと効果を発揮できないということがわかる．また，段落／文の相対位置重みについては，加算するより，乗算するほうが精度が向上していることがわかった．

(2) 文の関連度について

関連度を適用した場合と適用しない場合においては，適用した場合のほうが精度が上がることがわかった．また，関連度を加算した場合と乗算した場合で比較すると，加算したほうが精度が向上することがわかった．

(3) 段落／文の絶対位置補正について

絶対位置補正を用いた場合と用いない場合で比較すると，絶対位置補正を適用したほうが精度が上がることがわかった．このことより，段落／文の絶対位置補正の適用は有効であるといえる．

これらの考察より，見出し共通度と段落／文の相対位置重み，関連度，段落／文の絶対位置補正すべてを用いる方式で最も精度がよい結果が得られると考えられる．

５．おわりに

5.1 得られた成果

本研究における成果は以下のとおりである．

(1) 見出し内容語の本文内における出現分布の傾向
見出し内容語の出現傾向とテストコレクション内の重要文出現傾向は類似していることがわかった．

(2) 重要文抽出方式の提案
(1)の結果に着目して，見出しとの一致度，段落／文の相対位置情報，文の関連度，段落／文の絶対位置情報を組み合わせた重要文抽出方式を提案した．

(3) 提案方式における性能評価
(a) (2)で挙げた 4 つの重みにおける，12 通りの組み合わせの性能評価を行い，Average で解説は 0.3%，社説はほぼ同等，社会は 7.1%，全体では 2.3%，NTCIR-2 TSC1 参加システムの最良結果を上回った．
(b) 最適な重要文重み付けの組み合わせは，見出し一致度と段落／文の相対位置重みの積に関連度を加算し，最後に段落／文の絶対位置補正を適用する組み合わせであった．

5.2 今後の課題

今後の課題として，以下のものが挙げられる．

(1) 他のジャンルにおける本手法の性能評価
(2) 他の新聞社の記事における本手法の性能評価

謝辞

CD 毎日新聞-94,95,98,2000 版の使用許諾を頂いた毎日新聞社，およびテストコレクションをご提供頂いた国立情報学研究所の方々，形態素解析器茶筌の開発にあたられた多くの方々に深謝いたします．

参考文献

[1] 奥村 学, 難波 英嗣, "テキスト自動要約に関する研究動向(巻頭言に代えて)," 自然言語処理, Vol.6, No.6, pp.1-26, 1999.

[2] C. Paice, "Constructing Literature Abstracts by Computer: Techniques and Prospects," Information Processing and Management, Vol.26, No.1, pp.171-186, 1990.

[3] 中尾 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," 情報処理学会研究報告, NL-117-17, pp.121-128, 1997.

[4] 亀田 雅之，"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方式," 情報処理学会研究報告, NL-121-17, pp.119-126, 1997.

[5] 奈良先端科学技術大学院大学,「形態素解析器茶筌」, http://chasen.aist-nara.ac.jp/.

[6] 松村 繁男，山田 剛一，絹川 博之，中川 裕志, "新聞記事コーパスの共通文を利用した重要文抽出方式" FIT2004 情報科学技術フォーラム, 分冊 2，pp.171-172，2004.9.

表４ 提案方式の性能評価結果

ジャンル	相対位置重み	関連度	絶対位置補正	10%	30%	50%	Average
解説	加算	加算	無し	0.295	0.464	**0.652**	**0.470**
	乗算	加算	有り	0.304	0.455	**0.643**	**0.467**
	加算	加算	有り	0.304	0.458	**0.634**	0.465
	乗算	加算	無し	0.286	0.464	**0.632**	0.461
	加算	無し	有り	0.321	0.427	**0.622**	0.457
社説	加算	加算	有り	**0.387**	0.405	0.607	0.466
	乗算	加算	有り	**0.377**	0.415	0.603	0.465
	加算	無し	有り	**0.368**	0.425	0.593	0.462
	乗算	無し	有り	**0.368**	0.405	0.599	0.457
	乗算	加算	無し	0.358	0.412	0.581	0.450
社会	乗算	加算	有り	**0.405**	**0.570**	**0.638**	**0.538**
	乗算	加算	無し	0.351	**0.535**	**0.648**	**0.511**
	加算	加算	有り	0.351	**0.535**	**0.643**	**0.510**
	乗算	無し	有り	**0.378**	**0.526**	**0.622**	**0.509**
	加算	加算	無し	0.324	**0.561**	**0.638**	**0.508**
Formal Run	乗算	加算	有り	**0.392**	0.462	**0.616**	**0.490**
NTCIR-2 TSC1 TaskA1 Formal run 最良結果				0.363	0.483	0.612	0.467

※網掛け：ジャンル毎の組み合わせパターンの中で最も性能の高かった組み合わせ

※**太字**：NTCIR-2 TSC1 TaskA1 参加システム最良結果を上回った組み合わせ