表題へのつながりに基づく文の重要度評価

吉見　毅彦[†]　奥西　稔幸[††]
山路　孝浩[††]　福持　陽士[††]

本稿では，表層的な情報を手がかりとして文と文のつながりの強さを評価し，その強さに基づいて重要な文を選び出す手法を提案する．文の重要度の評価に際して，表題はテキスト中で最も重要な文であり，重要な文へのつながりが強い文ほど重要な文であるという仮定を置き，文から表題へのつながりの強さをその文の重要度とする．二つの文のつながりの強さは，人称代名詞による前方照応と，同一辞書見出し語による語彙的なつながりに着目して評価する．平均で 29.0 文から成る英文テキスト 80 編を対象とした実験では，文選択率を 25%に設定したとき，従来手法による精度を上回る再現率 78.2%，適合率 57.7%の精度を得，比較的短いテキストに対して提案手法が有効であることを確認した．

キーワード:　抄録，重要文選択，照応，語彙的なつながり

Evaluation of Importance of Sentences based on Connectivity to the Title

Takehiko Yoshimi[†] , Toshiyuki Okunishi[††] , Takahiro Yamaji[††]
and Yoji Fukumochi[††]

This paper proposes a method of selecting important sentences from a text based on the evaluation of the connectivity between sentences by using surface information. We assume that the title of a text is the most concise statement which expresses the most essential information of the text, and that the closer a sentence relates to an important sentence, the more important this sentence is. The importance of a sentence is defined as the connectivity between the sentence and the title. The connectivity between two sentences is measured based on correference between a pronoun and a preceding (pro)noun, and on lexical cohesion of lexical items. In an experiment with 80 English texts, which consist of an average of 29.0 sentences, the proposed method has marked 78.2% recall and 57.7% precision, with the selection ratio being 25%. The recall and precision values surpass those achieved by conventional methods, which means that our method is more effective in abridging relatively short texts.

KeyWords:　*Abridgement, Selection of Important Sentences, Correference, Lexical Cohesion*

† シャープ (株) ソフト事業推進センター / 神戸大学大学院自然科学研究科, Software Business Development Center, SHARP Corp. / Graduate School of Science and Technology, Kobe University
†† シャープ (株) ソフト事業推進センター, Software Business Development Center, SHARP Corp.

1　はじめに

電子化テキストの急増などに伴い，近年，テキストから要点を抜き出す重要文選択技術の必要性が高まってきている．このような要請に現状の技術レベルで応えるためには，表層的な情報を有効に利用することが必要である．これまでに提案されている表層情報に基づく手法では，文の重要度の評価が主に，1) 文に占める重要語の割合，2) 段落の冒頭，末尾などのテキスト中での文の出現位置，3) 事実を述べた文，書き手の見解を述べた文などの文種，4) あらかじめ用意したテンプレートとの類似性などの評価基準のいずれか，またはこれらを組み合わせた基準に基づいて行なわれる (Luhn 1958; Edmundson 1969; 喜多壮太郎 1987; 鈴木康広 栃内香次 1988; 間瀬久雄, 大西昇, 杉江昇 1989; Salton, Allan, Buckley, and Singhal 1994; Brandow, Mitze, and Rau 1995; 松尾比呂志 木本晴夫 1995; 佐藤円, 佐藤理史, 篠田陽一 1995; 山本和英, 増山繁, 内藤昭三 1995; Watanabe 1996; Zechner 1996; 福本文代, 福本淳一, 鈴木良弥 1997; 仲尾由雄 1997)．

本稿では，表層的な情報を手がかりとして文と文のつながりの強さを評価し，その強さに基づいて文の重要度を決定する手法を提案する．提案する手法では文の重要度に関して次の仮定を置く．

(1)　表題はテキスト中で最も重要な文である．

(2)　重要な文とのつながりが強ければ強いほど，その文は重要である．

表題はテキストの最も重要な情報を伝える表現であるため，それだけで最も簡潔な抄録になりえるが，多くの場合それだけでは情報量が十分でない．従って，不足情報を補う文を選び出すことが必要となるが，そのような文は，表題への直接的なつながりまたは他の文を介しての間接的なつながりが強い文であると考えられる．このような考え方に基づいて，文から表題へのつながりの強さをその文の重要度とする．文と文のつながりの強さを評価するために次の二つの現象に着目する．

(1)　人称代名詞と先行 (代) 名詞の前方照応

(2)　同一辞書見出し語による語彙的なつながり

重要文を選択するために文間のつながりを解析する従来の手法としては，1) 接続表現を手がかりとして修辞構造を解析し，その結果に基づいて文の重要度を評価する手法 (間瀬久雄他 1989; Ono, Sumita, and Miike 1994) や，2) 本稿と同じく，語彙的なつながりに着目した手法 (Hoey 1991; Collier 1994; 福本淳一 1997; 佐々木一朗, 増山繁, 内藤昭三 1993) がある．文と文をつなぐ言語的手段には，照応，代用，省略，接続表現の使用，語彙的なつながりがある (Halliday and Hasan 1976; Jelinek, Yoshimi, Nishida, Tamura, and Murakami 1995) が，接続表現の使用頻度はあまり高くない[1]．このため，前者の手法には，接続表現だけでは文間のつな

1 文献 (Halliday and Hasan 1976) で調査された七編のテキストでは，照応，代用，省略，接続表現の使用，語彙的なつながりの割合は，それぞれ，32%，4%，10%，12%，42%である (Hoey 1991)．

がりを解析するための手がかりとしては十分でないという問題点がある．後者の手法では，使用頻度が比較的高い照応を手がかりとして利用していない．

2　テキストの結束を維持する手段

適格なテキストでは通常，文と文の間につながりがある．二つの文をつなぐ言語的手段のうち照応と語彙的なつながりは，他の結束維持手段よりも頻繁に見られる．

照応は，二つのテキスト構成要素が一つの事象に言及することによってテキストの結束を生む手段である．前方照応では，ある要素の解釈が，テキスト中でその要素より前方に現れる先行要素に依存して決まる．ある要素 Y とその先行要素 X の間で照応が成り立つためには，1) Y は X を縮約した言語形式であり，2) Y の意味と X の意味は矛盾してはならない (Jelinek et al. 1995)．例えば，代名詞は名詞句を，名詞句は分詞節を，分詞節は文をそれぞれ縮約した言語形式である．次のテキスト 1 では斜体の表現の意味は互いに矛盾しないので，それらはいずれも同一事象を指しているとみなせる．

テキスト 1　*The Soviet National Emergency Committee dismissed President Gorbachov from office.* As well as *dismissing the President*, the Committee embarked upon choosing his replacement. *Gorbachov's dismissal* is bound to put Western policies vis-a-vis the Soviet Union into great turmoil. *It* will have grave repercussions on the exchange rates.

語彙的なつながりでは，照応と異なり，二つのテキスト構成要素が同一事象に言及しているとは限らない (Halliday and Hasan 1976)．次のテキスト 2 では第二文の “boy” は先行文の “boy” と同じ少年に言及しているが，テキスト 3 では別の少年に言及している．

テキスト 2　There's a *boy* climbing that tree. The *boy*'s going to fall if he doesn't take care.

テキスト 3　There's a *boy* climbing that tree. And there's another *boy* standing underneath.

テキスト 2 と 3 では同一辞書見出し語が繰り返されているが，類義語や上位概念語などの使用によって語彙的なつながりが生じることもある．次のテキスト 4 では “boy” の類義語 “lad” が用いられている．

テキスト 4　There's a *boy* climbing that tree. The *lad*'s going to fall if he doesn't take care.

3　文の重要度の評価

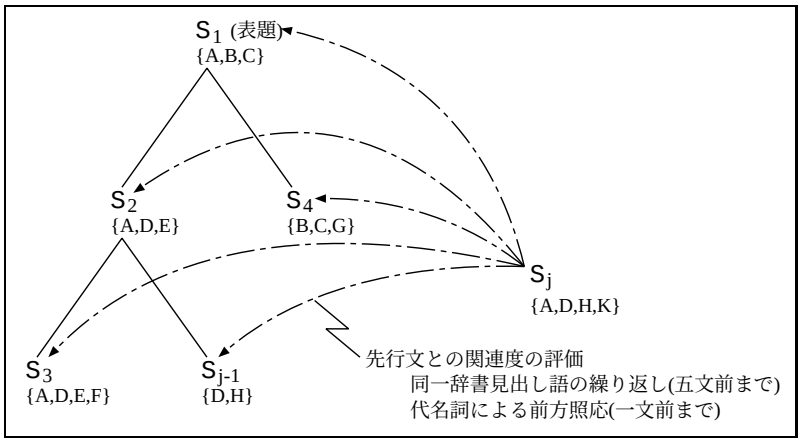
3.1　テキスト構造と文の重要度に関する仮定

本稿では，テキストを構成する文 $S_1, S_2, \cdots, S_n$ の間で次の条件が成り立つと仮定する．

(1)　冒頭文 S_1 はどの文にもつながらない．

(2)　S_1 以外の各文 S_j について，S_j が直接つながる先行文 $S_i(i<j)$ が唯一つ存在する．

この仮定は，二つの文 (の構成要素) のつながりに，後続文 (の構成要素) から先行文 (の構成要素) への方向性があることを意味する．この方向性に対する反例として後方照応 (Hirst 1981) があるが，後方照応が用いられることは希である[2]．また，この仮定に従えば，文が同時に複数の先行文に直接つながることはないので，テキスト構造は，図 1 に示すように，冒頭文 S_1 を根節点とする木で表される．



3.2 二文間の関連度の評価

提案手法への入力はテキストの形態素解析結果である．形態素解析によってテキスト中の各語の辞書見出し語と品詞が得られる．今回利用した形態素解析系からの出力では品詞は一意に決定されている．以降，品詞が名詞，人称代名詞，動詞，形容詞，副詞のいずれかである辞書見出し語を重要語と呼ぶ．

文 S_j の先行文 S_i へのつながりの強さ (関連度) を求める式を次のように定める．

$$S_i \text{と} S_j \text{の関連度} = \frac{S_j \text{中の重要語のうち} S_i \text{の題述中の重要語につながるものの重みの和}}{S_i \text{の題述中の重要語の数}} \quad (2)$$

(2) 式の意味は 3.2.1 節以降で説明する．二つの重要語の間につながりがあるかどうかの判定は，人称代名詞と先行 (代) 名詞の前方照応を検出すること (3.2.1 節) と，同一辞書見出し語による語彙的なつながりを検出すること (3.2.2 節) によって行なう．重要語への重み付けについては 3.2.3 節で述べ，本稿でいう文の題述 (rheme) の定義は 3.2.4 節で与える．

3.2.1 人称代名詞と先行 (代) 名詞の照応の検出

人称代名詞と先行名詞または先行代名詞との照応を検出するためには，両者の人称，性，数，意味素性をそれぞれ照合する必要がある．しかし，今回は，名詞の性と意味素性が記述されていない辞書を用いたので，照応の検出は両者の人称，数をそれぞれ照合することによって行なった．

しばしば指摘されるように，代名詞との間で照応が成り立つ先行 (代) 名詞は，その代名詞を含む文 S_j あるいは S_j の直前の文 S_{j-1} に現れることが多い[3]ので，先行 (代) 名詞の検索対象文を S_j と S_{j-1} に限定する．検索は S_j，S_{j-1} の順で行ない，S_j 中の (代) 名詞との照合が成功した場合は，S_{j-1} に対する処理は行なわない．

3.2.2 重要語の語彙的なつながりの検出

二つの文に現れる重要語が文字列として一致するとき，両者の間に語彙的なつながりがあるとみなす．ここでは，2 節で述べたような，二つの語が同一事象に言及しているかどうかの区別は行なわない．文字列照合において，照合対象が両方とも単語である場合は，二つの重要語が完全に文字列一致したときに限り照合成功とみなすが，照合対象の両方またはいずれか一方が辞書に登録されている連語である場合は，両者が前方一致または後方一致したときも照合成功とみなす．例えば，“put pressure on” と “put” は前方一致で，“cabinet meeting” と “meeting” は後方一致で照合が成功する．

二つの文がある一定の距離以上離れていると，それらに含まれる重要語の文字列照合が成功しても二つの文の間に直接的なつながりはないと考えられる．このため，二文間の距離に関し

3 訓練テキスト 20 編では，人称代名詞による前方照応のうち 96%がこのような事例であった．

て制限を設ける．提案手法を開発する際に訓練用として用いた英文テキスト 20 編において，文字列照合が成功する重要語 (人称代名詞は除く) を含む二つの文の間の距離と，その重要語が二つの文を直接つなぐ役割を実際に果たしているかどうかとの関連を調べた結果に基づいて，処理対象範囲を文 S_j から五文前までの先行文 $S_i(j-5 \leq i < j)$ とする．

直観的には，単に処理対象範囲を制限するだけでなく，文字列照合が成功する重要語を含む二文間の距離に応じて照合結果に重み付けを行なう方が自然かもしれない．このため，訓練テキストを対象とした実験において，文 S_j から五文前までの先行文 S_i の範囲で，二つの文の距離が離れるにつれてつながりの強さが弱まるように重み付けを試みた．しかし，重み付けを行なわない場合の再現率と適合率を上回る結果は得られなかった．このため，本稿では処理範囲を制限するに留める．

3.2.3　表題語への重み付け

テキストの表題中に現れる重要語 (以降，表題語と呼ぶ) は，そのテキストにおいて重要な情報を伝えると考えられる．従って，表題語を含む文の重要度を大きくするために，他の重要語に与える重みの値よりも大きな値を与えること (Edmundson 1969; 間瀬久雄他 1989; Watanabe 1996) が適切である．本稿では，表題語への重み付けを行なう際にテキスト中での表題語の出現頻度を考慮する．すなわち次のような仮定を置く．

> 表題語を含む文の重要性は，表題語がテキスト中に頻繁に現れる場合は，表題語を含まない文の重要性に比べて特に高いわけではないが，表題語がテキスト中に希にしか現れない場合には，表題語を含まない文に比べて特に高くなる．

訓練テキスト 20 編を分析した結果に基づいて，表題語を含む文の数がテキストの総文数の 1/4 以下である場合に限り，表題語の重みを $w(>1)$ とする．表題語以外の重要語の重みは常に 1 とする．

$$\text{重要語 } kw \text{ の重み} = \begin{cases} w(>1) & kw \text{ が表題語であり，かつ } kw \text{ を含む文の数が総文数の} \\ & 1/4 \text{ 以下の場合} \\ 1 & \text{その他} \end{cases}$$

重み w の具体的な値は，訓練テキストを対象とした実験で再現率と適合率ができるだけ高くなるように調整し，最終的に $w = 5$ とした．

3.2.4　先行文の題述へのつながり

テキストは，通常，先行文 S_i における題述 (rheme) が文 S_j においてその主題 (theme) として受け継がれ，それに新たな情報が付け加わるという形で展開する (Givon 1979)．従って，文 S_j の先行文 S_i へのつながりの強さの評価を，S_j が S_i の題述をどれだけ多く主題として受け継いでいるかに基づいて行なう．

主題と題述は，文の前半部分が主題，後半部分が題述というように文中の位置で区別されることが多い (福地肇 1985) が，本稿では，文中の位置ではなく，関連文とのつながりに基づいて区別する．ここで，S_j の関連文とは，3.1 節の (1) 式において，{S_i の重要度 $\times S_i$ と S_j の関連度 } の値が最大となるときの先行文 S_i を意味する．この値を最大にする先行文が複数存在する場合は，S_j との距離が最も近いものを関連文と呼ぶ．関連文とのつながりに基づいて主題と題述を次のように定める．

> 文 S_j の主題は，S_j 中の重要語のうち S_j の関連文中の重要語につながるものから構成され，文 S_j の題述は，つながらない重要語から構成される．ただし，関連文を持たない冒頭文 S_1 では，それに含まれる重要語すべてが題述を構成する．

例えば，図 1 において，括弧 { と } で括った英大文字を各文に現れる重要語とすると，各文の主題と題述は表 1 のように分けられる．

表 1　図 1 の各文の主題と題述

文	関連文	主題	題述
S_1	—	—	A,B,C
S_2	S_1	A	D,E
S_3	S_2	A,D,E	F
S_4	S_1	B,C	G
:	:	:	:
S_{j-1}	S_2	D	H

3.3　重要文選択手順と処理例

3.1 節と 3.2 節で述べた考え方に従って重要文を選ぶ処理は図 2 のようにまとめられる． 例として，図 3 のテキストを処理して得られる結果を表 2 に示す．このテキストでは，表題語 “amorphous”，“Si”，“TFT” を含む文はそれぞれ三文，三文，五文存在し[4]，いずれもテキスト総文数 10 文の 1/4 を越えるので，表題語への重み付けは行なわれない．表 2 の「つながり語」欄に現れる記号 ϕ は，先行文の題述中の重要語につながる重要語が存在しなかったことを意味する．このテキストからは，文選択率が 25%に設定されているとき，表題 S_1，S_1 につながる文 S_4，S_4 につながる文 S_6 の三文が重要文として選び出される．

4　実験と考察

重要文選択実験には英文報道記事 100 編を用いた．100 編のテキストを訓練用の 20 編と試験用の 80 編に分けた．まず，訓練テキスト 20 編を対象として実験を繰り返し，再現率と適合

4 表題も含めて数えている．

ステップ 1　入力を形態素解析する.
ステップ 2　表題語への重み付け処理を行なう.
ステップ 3　冒頭文 S_1 の重要度を次式で求める.

$$S_1\text{の重要度} = \frac{S_1\text{中の重要語の重みの和}}{S_1\text{の重要語の数}}$$

ステップ 4　各文 $S_j (j = 2, 3, \cdots, n)$ について，S_j から五文前までの先行文 S_i の範囲 $(j - 5 \leq i < j)$ で，3.1 節の (1) 式と 3.2 節の (2) 式に従って重要度を求める.
ステップ 5　あらかじめ定められた数だけ文を重要度の順に選択し，それらをテキストでの出現順に出力する.

図 2　重要文選択手順

S_1 Amorphous Si TFT
S_2 Active matrix LCDs which are typically used in products such as LCD color TVs are controlled by a switching element known as a thin-film transistor or thin-film diode placed at each pixel.
S_3 The fundamental concept was revealed in 1961 by RCA of America, a U.S. company, but basic research only began in the 1970's.
S_4 Amorphous Si TFT LCDs introduced in 1979 and 1980 have become the mainstream for today's active matrix displays.
S_5 These units place an active element at each pixel, and taking advantage of the non-linearity of the active element, are able to apply sufficient drive-voltage margin to the liquid crystal itself, even with the increase in the number of scan lines.
S_6 As shown in Figure 1, TFT LCDs that use amorphous Si thin-film transistors (TFTs) as the active elements are becoming the mainstream today, and full-color displays achieving contrast ratios of 100:1 and which compare favorably to CRTs are being developed.
S_7 The driver electronics for TFT LCDs consist of data-line drive circuitry that applies display signals to the data lines (source drivers) and scanning line drive circuitry that applies scanning signals to the gate lines (gate drivers).
S_8 A signal control circuit to control these operations and a power supply circuit complete the system.
S_9 Liquid crystal materials used in TFT LCDs are TN (twisted nematic) liquid crystals, but despite the fact that pixel counts have increased and a drive element is placed at each pixel, we have still been able to rapidly increase the contrast, viewing angle, and image quality of these displays.
S_{10} However, manufacturing technologies to fabricate several hundred thousand such elements onto the surface of a large screen are extremely problematic, and the fundamental approach developed in 1987 is still being used today.

図 3　テキスト例

表 2　図 3 のテキストに対する処理結果

文	関連文	つながり語	関連度	重要度	選択順位
S_1	—	—	—	1	1
S_2	—	ϕ	0	0	—
S_3	—	ϕ	0	0	—
S_4	S_1	amorphous, Si, TFT	3/3	1	2
S_5	S_4	active	1/9	1/9	7
S_6	S_4	LCD, active, become, mainstream, today, display	6/9	2/3	3
S_7	S_4	LCD, display	2/9	2/9	4
S_8	S_7	signal	1/13	2/117	8
S_9	S_4	LCD, display	2/9	2/9	5
S_{10}	S_6	element, develop, use	3/16	1/8	6

率ができるだけ高くなるように，3.2.3 節で述べた表題語の重みを調整した．次に，訓練テキストを対象とした実験で最も高い再現率と適合率が得られた設定で，試験テキスト 80 編を対象として実験を行なった．

テキストの総文数は，訓練テキストの場合，最も短いもので 15 文，最も長いもので 36 文，一テキスト当たりの平均では 26.2 文であり，試験テキストの場合，それぞれ 12 文，64 文，29.0 文であった．

各テキストについて，第三者 (一名) によって重要と判断された文を，選択すべき正解文とした．人手による正解文の選択では，システムが行なっているような各文についての選択順位付けは行なわず，テキスト中の各文についてそれが重要な文であるかそうでないかを判断するに留めた．正解文の数は，訓練テキストの場合，平均で元テキストの総文数の 20.8%であり，試験テキストの場合 17.9%であった．

4.1　訓練テキストでの実験結果

3.2.3 節で述べた表題語への重み付けに関して次のような三種類の設定で，各訓練テキストについて正解文と同じ数だけ文を選択した場合の平均精度 (再現率と適合率は同じ値となる) を表 3 に示す．

設定 1　表題語を含む文の数がテキスト総文数の 1/4 以下である場合に限り，表題語の重みを 5 とする．表題語以外の重要語の重みは 1 とする．

設定 2　表題語の重みをその出現頻度に関係なく常に 5 とする．表題語以外の重要語の重みは 1 とする．

設定 3　表題語の重みを他の重要語の重みと同じ 1 とする．

表 3 によれば，設定 1 での精度が最も高くなっており，3.2.3 節で示した，出現頻度を考慮した表題語への重み付けが有効であることがわかる．

表 3　訓練テキスト 20 編での実験結果

設定	1	2	3
精度	71.0%	70.0%	62.5%

4.2　試験テキストでの実験結果

訓練テキストを対象とした実験で最も高い再現率と適合率が得られた設定で，80 編の各試験テキストについて正解文と同じ数だけ文を選択した場合の精度は，平均で 72.3%であった．各テキストごとの精度分布を図 4 に示す．

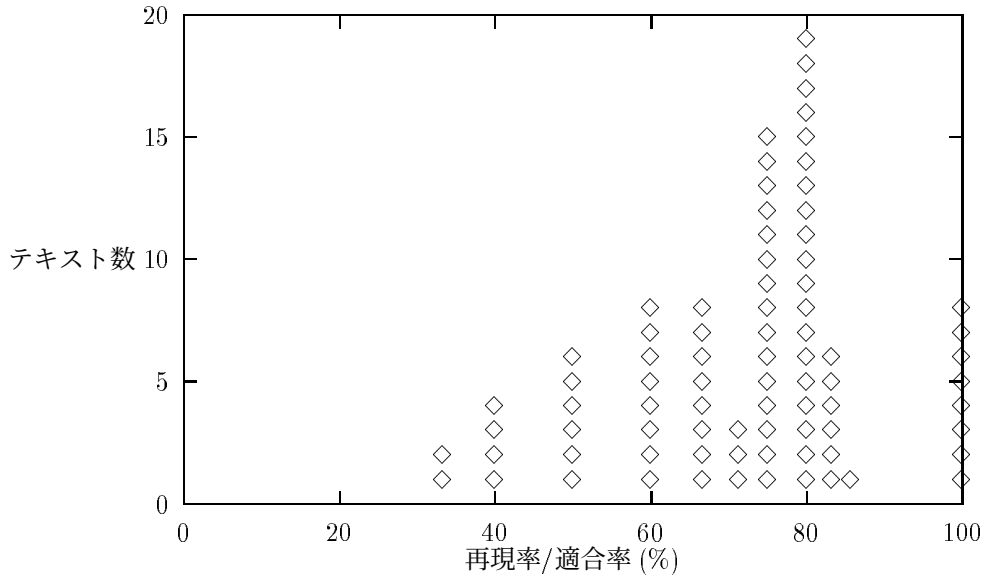


図 4　提案手法による精度分布

文選択率を 5%から 100%まで五刻みで変化させたときの平均再現率と平均適合率の変化の様子を図 5 に示す．図 5 には，精度比較のために実装した重要語密度法による実験結果を併せて示す．重要語密度法に関して改良手法が提案されている (鈴木康広・栃内香次 1988) が，ここでは次式で文 S の重要度を評価した．

$$\text{文 } S \text{ の重要度} = \frac{\text{文 } S \text{ 中の各重要語のテキスト全体での出現頻度の和}}{\text{文 } S \text{ 中の重要語の数}}$$

図 5 によれば，一般的な抄録において適切な文選択率であるとされる 20%から 30%までの付近で，特に，提案手法の精度が重要語密度法の精度を大きく上回っている．

提案手法の精度と，インターネット上で試用可能なシステム A と，市販されている三つの

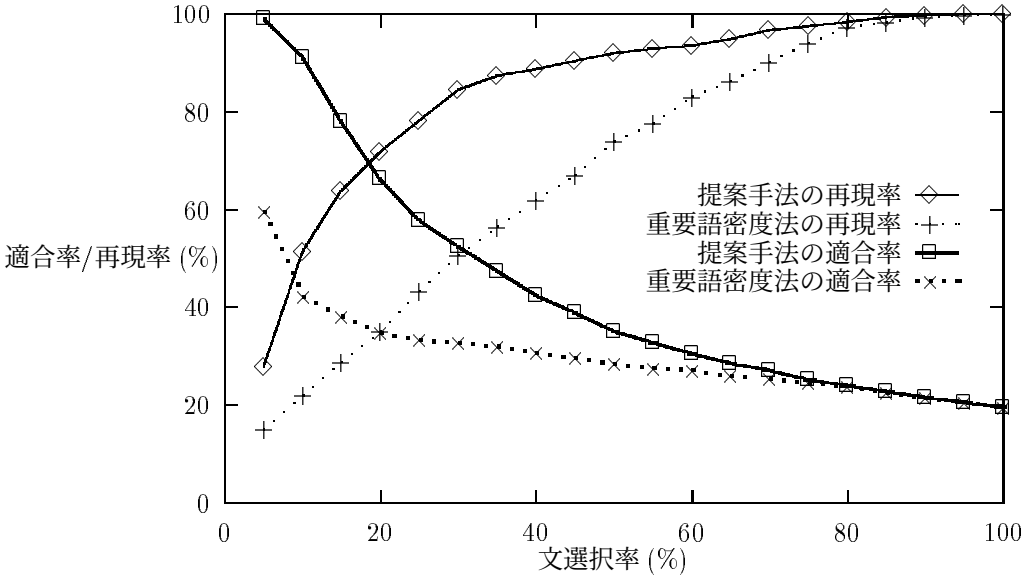


図 5　提案手法と重要語密度法の精度比較

システム B，C，D の精度を比較した．それぞれの平均再現率と平均適合率を表 4 に示す．システム A，B，C，D の文選択率は，各システムの既定状態で選ばれた文の数とテキストの総文数から逆算したものである．提案手法の文選択率は，四システムの文選択率とほぼ同じである 25%とした．表 4 によれば，一般ユーザに利用されている実動システムの精度を提案手法の精度が上回っており，提案手法の実用的な抄録システムとしての有効性が示されている．

表 4　提案手法と他の実動システムの精度比較

	再現率	適合率	文選択率
提案手法	78.2%	57.7%	25%
システム A	72.3%	52.6%	26%
システム B	61.7%	39.5%	29%
システム C	61.4%	40.9%	29%
システム D	57.5%	42.2%	27%

4.3　考察

提案手法によって正解文に与えられた重要度が小さく，正解文が選択されなかった原因を分析した．ここでは代表的な原因を二つ挙げる．一つは，辞書見出し語の文字列照合では，語彙的なつながりが捉えられなかったことである．あるテキストでは，“shooting” と “gunfire” の類

義関係が把握できないため，“gunfire” を含む正解文はどの先行文にもつながらないとみなされ，重要文として選択できなかった．このような語彙的なつながりを捉えるためにはシソーラスが必要となるが，他のテキストでは，辞書見出し語の文字列照合の代わりに語基 (base) の文字列照合を行なえば，つながりが捉えられる可能性もあった．例えば，“announce” と “announcement” は，辞書見出し語としては異なるが語基は同一であるので，文字列照合が成功するだろう．

本研究では，一般ユーザに利用される実動システムへの組み込みを前提として，高速な処理を実現することを目標の一つとした．実動システムでは，プロトタイプシステムと異なり，重要文選択の精度と共に処理速度も重要視される．シソーラスの検索に比べて，文字列照合は処理効率の点で有利である．

正解文に十分大きい重要度が与えられなかったもう一つの原因は，テキストが複数のサブトピックから構成されていることであった．一般に，トピックが切り替わると，それまでとは異なった語彙が用いられるようになる．このため，提案手法のように同一辞書見出し語による語彙的なつながり (と人称代名詞による前方照応) に基づいて文と文のつながりを評価する手法では，トピックが切り替わる文から先行文へのつながりが弱いと判定され，トピック切り替わり文に対して与えられる重要度は小さくなる．従って，トピック切り替わり文が正解文であるようなテキストでは，高い精度を得ることが難しくなる．

5　おわりに

本稿では，人称代名詞による前方照応と，同一辞書見出し語による語彙的なつながりを検出することによって，テキストを構成する各文と表題との直接的なつながりまたは他の文を介しての間接的なつながりの強さを評価し，その強さに基づいて各文の重要度を決定する手法を提案した．平均で 29.0 文から成る英文テキスト 80 編を対象とした実験では，文選択率を 25%に設定したとき，再現率 78.2%，適合率 57.7%の精度を得，提案手法が比較的短いテキストに対して有効であることを確認した．

複数のサブトピックから成るような比較的長いテキストの扱いは今後の課題である．同一辞書見出し語の出現頻度と出現分布を利用してトピックの切り替わりを検出し (Hearst 1997)，各サブトピックごとに提案手法を適用すると，長いテキストに対してどの程度の精度が得られるかを今後検証したい．

参考文献

Brandow, R., Mitze, K., and Rau, L. F. (1995). “Automatic Condensation of Electric Publications by Sentence Selection.” *Information Processing & Management*, **31** (5), 675–685.

Collier, A. (1994). "A System for Automating Concordance Line Selection." In *Proceedings of International Conference on New Methods in Language Processing (NeMLaP94)*, pp. 95–100.

Edmundson, H. P. (1969). "New Methods in Automatic Extracting." *Journal of the Association for Computing Machinery*, **16** (2), 264–285.

福地肇 (1985). 談話の構造, 新英文法選書, 10 巻. 大修館書店.

福本文代, 福本淳一, 鈴木良弥 (1997). "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, **4** (2), 89–109.

福本淳一 (1997). "文の結合度に基づく内容抽出法." 言語処理学会第 3 回年次大会発表論文集, pp. 321–324.

Givon, T. (1979). "From Discourse to Syntax: Grammar as a Processing Strategy." In Givon, T. (Ed.), *Discourse and Syntax*, Vol. 12 of *Syntax and Semantics*, pp. 81–112. Academic Press.

Halliday, M. A. K. and Hasan, R. (1976). *Cohesion in English*. English Language Series 9. Longman.

Hearst, M. A. (1997). "TextTiling: Segmenting Text into Multi-paragraph Subtopic Passages." *Computational Linguistics*, **23** (1), 33–64.

Hirst, G. (1981). *Anaphora in Natural Language Understanding: A Survey*. Lecture Notes in Computer Science 119. Springer-Verlag.

Hoey, M. (1991). *Patterns of Lexis in Text*. Describing English Language. Oxford University Press.

Jelinek, J., Yoshimi, T., Nishida, O., Tamura, N., and Murakami, H. (1995). "Text-Wide MT Grammar." In *Proceedings of the 3rd Natural Language Processing Pacific Rim Symposium (NLPRS95)*, pp. 449–454.

喜多壮太郎 (1987). "説明文を要約するシステム." NL63-6, 情報処理学会.

Luhn, H. P. (1958). "The Automatic Creation of Literature Abstracts." *IBM Journal for Research and Development*, **2** (2), 159–165.

間瀬久雄, 大西昇, 杉江昇 (1989). "説明文の抄録作成について." NLC89-40, 電子情報通信学会.

松尾比呂志 木本晴夫 (1995). "抽出パターンの階層的照合に基づく日本語テキストからの内容抽出法." 情報処理学会論文誌, **36** (8), 1838–1844.

仲尾由雄 (1997). "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." NL117-17, 情報処理学会.

Ono, K., Sumita, K., and Miike, S. (1994). "Abstract Generation based on Rhetorical Structure Extraction." In *Proceedings of the 15th International Conference on Computational*

Linguistics (COLING94), pp. 344–348. Also published as http://xxx.lanl.gov/abs/cmp-lg/9411023.

Salton, G., Allan, J., Buckley, C., and Singhal, A. (1994). "Automatic Analysis, Theme Generation, and Summarization of Machine-Readable Texts." *Science*, **264** (3), 1421–1426.

佐々木一朗, 増山繁, 内藤昭三 (1993). "語彙的結束性に着目した文章抄録法の提案." NL98-9, 情報処理学会.

佐藤円, 佐藤理史, 篠田陽一 (1995). "電子ニュースのダイジェスト自動生成." 情報処理学会論文誌, **36** (10), 2371–2379.

鈴木康広 栃内香次 (1988). "キーワード密度方式自動抄録法の改良—高頻度隣接語による改善—." 情報処理学会論文誌, **29** (3), 325–328.

Watanabe, H. (1996). "A Method for Abstracting Newspaper Articles by Using Surface Clues." In *Proceedings of the 16th International Conference on Computational Linguistics (COLING96)*, pp. 974–979.

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.

Zechner, K. (1996). "Fast Generation of Abstracts from General Domain Text Corpora by Extracting Relevant Sentences." In *Proceedings of the 16th International Conference on Computational Linguistics (COLING96)*, pp. 986–989.

略歴

吉見 毅彦: 1987 年電気通信大学大学院計算機科学専攻修士課程修了．現在，シャープ (株) ソフト事業推進センターにて機械翻訳システムの研究開発に従事．在職のまま，1996 年より神戸大学大学院自然科学研究科博士課程在学中．

奥西 稔幸: 1984 年大阪大学基礎工学部情報工学科卒業．同年シャープ (株) に入社．1985〜89 年 (財) 新世代コンピュータ技術開発機構に出向．現在，同社情報システム事業本部ソフト事業推進センターに勤務．機械翻訳システムの研究開発に従事．

山路 孝浩: 1990 年大阪市立大学理学部数学科修士課程修了．同年シャープ (株) に入社．1993〜95 年 (財) 新世代コンピュータ技術開発機構に出向．現在，同社 OA システム事業部においてワープロの開発に携わる．

福持 陽士: 1982 年インディアナ大学言語学部応用言語学科修士課程修了．翌年，シャープ (株) に入社．現在，情報システム事業本部ソフト事業推進センター副参事．機械翻訳システムの研究開発に従事．

(1998 年 5 月 18 日 受付)
(1998 年 8 月 24 日 再受付)

(1998 年 10 月 2 日 採録)