Evaluation of Importance of Sentences based on Connectivity to Title

Takehiko Yoshimi and **Toshiyuki Okunishi**
Takahiro Yamaji and **Yoji Fukumochi**
Software Business Development Center, SHARP Corporation
492 Minosho-cho Yamatokoriyama Nara, Japan

Abstract

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 recall of 78.2% and precision of 57.7%, 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.

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，あるい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 and 栃内香次, 1988; 間瀬久雄 et al., 1989; Salton et al., 1994; Brandow et al., 1995; 松尾比呂志 and 木本晴夫, 1995; 佐藤円 et al., 1995; 山本和英 et al., 1995; Watanabe, 1996; Zechner, 1996; 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，あるい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，本稿では，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 et al., 1989; Ono et al., 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 et al., 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) が，接続表現の使用頻度はあまり高くない．このため，前者の手法には，接続表現だけでは文間のつな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 文の重要度の評価

2.1 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．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

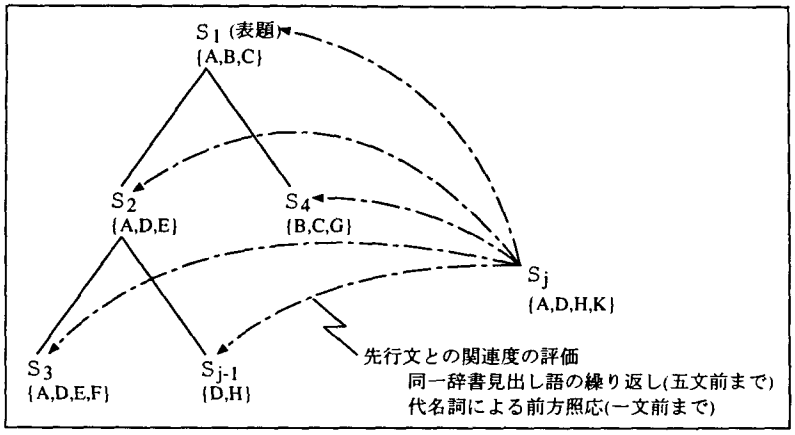

図 1: 文の重要度の評価

で，図1に示すように，テキスト構造は冒頭文 S_1 を根節点とする木で表される．

1節で述べた基本的な考え方は次のように具体化できる．

1. テキストの冒頭文 S_1 は，多くの場合，そのテキストの表題であるので，S_1 にはテキスト全体で最大の重要度を与える．
2. 冒頭文 S_1 以外の文 S_j の重要度は S_j から先行文 S_i へのつながりの強さ (関連度) と S_i の重要度によって決まると考え，文 S_j の重要度を求める式を次のように定める．

$$S_j\text{の重要度} = \max_{i<j}\{S_i\text{の重要度} \times S_i\text{と}S_j\text{の関連度}\} \quad (1)$$

文の重要度を (1) 式で求めることにすると，テキストの冒頭から順に処理を行なっていけば，テキストを構成する文すべての重要度が決定できるが，そのためには，まず，二つの文の関連度をどのようにして求めるかを定めなければならない．

2.2 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．

$$S_i\text{と}S_j\text{の関連度} = \frac{M_{i,j}}{N_i} \quad (2)$$

ここで，$M_{i,j}$ は文 S_j 中の重要語のうち先行文 S_i の題述 (rheme) 中の重要語につながるものの重みの和を表し，N_i は先行文 S_i の題述中の重要語の数を表す．

(2) 式の意味は 2.2.1 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 (2.2.1 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 (2.2.2 節) によって行なう．重要語への重み付けについては 2.2.3 節で述べ，本稿でいう文の題述の定義は 2.2.4 節で与える．

2.2.1 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．

2.2.2 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．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，二つの重要語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．

2.2.3 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 (Edmundson, 1969; 間瀬久雄 et al., 1989; Watanabe, 1996) のが適切である．本稿では，表題語に対する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．

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

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．

2.2.4 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，2.1 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 実験と考察

提案手法の評価には，訓練テキストとは異なる英文テキスト 80 編を用いた．評価テキストの総文数は，最も短かいもので 12 文，最も長いもので 64 文，一テキスト当たりの平均では 29.0 文であった．各テキストについて，第三者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．正解文の数は，平均で元テキストの総文数の 17.9% であった．

まず，各テキストについて，正解文と同じ数だけ文を選択するように設定して重要文選択実験を行なった．この場合の精度 (再現率と適合率は同じ値となる) は，72.3% であった．各テキストごとの精度分布を図 2 に示す．

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 3 に示す．図 3 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 and 栃内香次, 1988) が，ここでは次式で文 S の重要度を評価した．

$$\text{文}S\text{の重要度} = \frac{F}{N}$$

ここで，F は文 S 中の各重要語のテキスト全体での出現頻度の和を表し，N は文 S 中の重要語の数を表す．図 3 によれば，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，市販されている三つのシステム 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 2 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 2 によれば，一般ユーザに利用されている実働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．

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

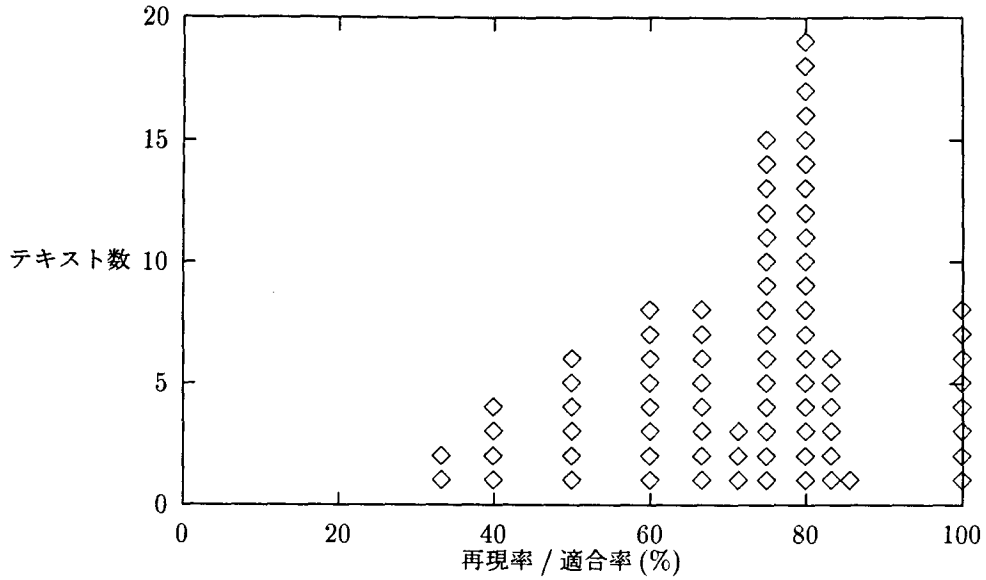

図 2: 提案手法による精度分布

表 2: 提案手法と他の実働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%

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働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．実働システムでは，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．

4 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，あるい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．

References

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

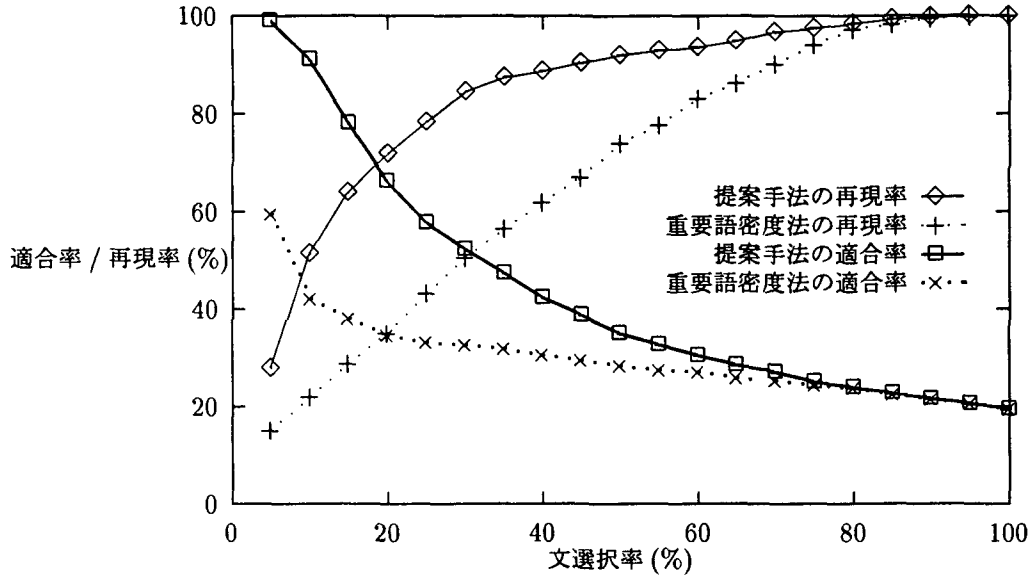

図 3: 提案手法と重要語密度法の精度比較

A. Collier. 1994. A System for Automating Concordance Line Selection. In *Proceedings of NeMLaP*, pages 95–100.

H. P. Edmundson. 1969. New Methods in Automatic Extracting. *Journal of the ACM*, 16(2):264–285.

T. Givon. 1979. From Discourse to Syntax: Grammar as a Processing Strategy. In T. Givon, editor, *Discourse and Syntax*, pages 81–112. Academic Press.

M. A. K. Halliday and R. Hasan. 1976. *Cohesion in English*. Longman.

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

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

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

K. Ono, K. Sumita, and S. Miike. 1994. Abstract Generation based on Rhetorical Structure Extraction. In *Proceedings of COLING*, pages 344–348.

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

H. Watanabe. 1996. A Method for Abstracting Newspaper Articles by Using Surface Clues. In *Proceedings of COLING*, pages 974–979.

K. Zechner. 1996. Fast Generation of Abstracts from General Domain Text Corpora by Extracting Relevant Sentences. In *Proceedings of COLING*, pages 986–989.

間瀬久雄, 大西昇, and 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.

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.

佐々木一朗, 増山繁, and 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, and 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.

山本和英, 増山繁, and 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.

松尾比呂志 and 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.

福地肇. 1985. 談話の構造. 大修館書店.

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

鈴木康広 and 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.