論 文

Technical Papers

Support Vector Machineを用いた文書の重要文節抽出 — 要約文生成に向けて —

A Method for Extracting Important Segments from Documents Using Support Vector Machines: Toward Automatic Text Summarization

鈴木 大介
Daisuke Suzuki
電気通信大学大学院電気通信学研究科システム工学専攻
Department of Systems Engineering, The University of Electro-Communications
dajie@utm.se.uec.ac.jp

内海 彰
Akira Utsumi
電気通信大学電気通信学部システム工学科
Department of Systems Engineering, The University of Electro-Communications
utsumi@se.uec.ac.jp

keywords: segment extraction, dependency structure, automatic summarization, Support Vector Machines

Summary

In this paper we propose an extraction-based method for automatic summarization. The proposed method consists of two processes: important segment extraction and sentence compaction. The process of important segment extraction classifies each segment in a document as important or not by Support Vector Machines (SVMs). The process of sentence compaction then determines grammatically appropriate portions of a sentence for a summary according to its dependency structure and the classification result by SVMs. To test the performance of our method, we conducted an evaluation experiment using the Text Summarization Challenge (TSC-1) corpus of human-prepared summaries. The result was that our method achieved better performance than a segment-extraction-only method and the Lead method, especially for sentences only a part of which was included in human summaries. Further analysis of the experimental results suggests that a hybrid method that integrates sentence extraction with segment extraction may generate better summaries.

1. はじめに

計算機を用いて文書を自動要約する研究の多くは，文章から重要な文を抽出して要約とする抜粋（extract）による手法に基づいている [Mani 01, 奥村 99, Radev 02]．一方，人間が作成する要約であるアブストラクト（abstract）には，複数の原文を再構成して生成される新たな文が多く含まれる．よって抜粋による手法だけでは，アブストラクトに匹敵する質の要約を生成できない．現時点では，アブストラクトの自動生成は技術的に困難であり，研究も概念実証の段階であるが [Radev 02]，文中から重要個所を抽出したり非重要個所を除去したりするだけでも要約の質の大幅な向上が期待できる [Knight 02, Radev 02]．

一方，重要文や重要個所の抽出には，文書中のキーワードの出現頻度や位置などの複数の表層的な情報の利用が効果的であることが示されている [Edmundson 69, Luhn 58]．さらに近年では，大量の情報の中のどれをどの程度考慮すれば重要文をより適切に抽出できるかを調べるために，機械学習を用いた重要文抽出の研究が行われている [Kupiec 95, Lin 99]．その中で，平尾ら [平尾 03, Hirao 03] は，多くの分類問題において優れた汎化性能を持つ機械学習手法である Support Vector Machine[Vapnik 95]（以下，SVM）を重要文抽出に適用して，優れた結果を報告している．

そこで本研究では，アブストラクト生成への第一歩として，文より小さい抽出単位として文節を考え，以下の 2 ステップからなる重要個所抽出型の要約手法を提案する．

（1）SVM を用いて文書中の文の重要文節を決定する．
（2）上記の結果と原文の係り受け構造から，文法的に妥当な文節のまとまりを抽出して要約とする．

ステップ (1) では，文節や文に関する複数の情報に基づく重要文節の分類を SVM を用いて行い，重要文節を決定する．しかし，これらの重要文節を集めただけでは文法的に適切な文や文節群になるとは限らない．そこでステップ (2) では，SVM による重要文節抽出の結果をできるだけ反映させながら，原文の文節間の係り受け関係を考慮して文法的に適切な文節群を抽出して要約文とする．

本稿の以下では，まず 2 章で重要個所抽出による既存の要約研究と本研究の違いを明らかにした上で，3 章と 4 章で，提案手法のステップ (1) と (2) の各詳細について説

明する．次に 5 章で，人手で抽出された重要個所データに基づく評価実験の方法とその結果について述べる．そして 6 章で，評価結果に対する詳細な分析を行い，重要個所抽出による要約手法や係り受け情報の有効性などについて考察する．

2. 重要個所抽出による要約

重要個所抽出による要約に関する既存研究の多くは，まず重要文抽出を行ってから，抽出された文の非重要個所を削除するという手法を採用している [Hirao 03, Jing 00a, Knight 02, Takeuchi 01]．非重要個所の削除手法としては，構文情報（構文木 [Jing 00a]，二重修飾 [大竹 02]，動詞連体修飾節 [酒井 04]）に基づいて削除個所を決定するものや，要約コーパスを用いて機械学習（SVM[Hirao 03, Takeuchi 01]，決定木 [Knight 02]，確率モデル [Knight 02, Jing 00a]）により削除知識を獲得するものが挙げられる．これらの手法に共通するのは，要約に含まれるべき情報を選択するのは文抽出処理に任せておき，非重要個所の削除はあくまでも要約をよりコンパクトにするために行うという考え方である．よって，非重要個所の削除による手法は，多くの知見が蓄積されている重要文抽出法がそのまま利用できる，文全体の構成に影響のないような個所を削除することによって文法的な正しさを保存しやすいなどの利点がある．

しかし，この手法には以下のような問題点もある．

- 文抽出処理で重要文と認識されなかった文中に重要な情報を持つ個所がある場合，それらの情報は要約から欠落する可能性が高い [和田 02]．抽出する重要文の量を目的の要約率に対して多く設定することでこの問題に対処することも考えられるが，そうすると非重要個所の削除だけで目的の要約率を満たすのが困難になる．よって，いずれにしろ重要個所を積極的に選択する必要がある．
- 人間の行うアブストラクトとしての要約を生成するためには，複数の文に含まれる重要個所を取り出してそれらを単文にまとめるという文結合の処理が必要である [Jing 00b][*1]．しかし非重要個所の削除による手法だけでは，このような要約を生成するための要素技術を提供できない [大石 03]．

以上の問題点に対処するために，本研究では，重要文を抽出せずに，原文から重要個所を直接抽出する手法を提案する．重要個所を直接抽出する研究としては，ニュース文に特有の情報を用いた重要文節の同定 [和田 02] や遺伝的アルゴリズムによる重要文節の同定 [大石 03] が挙げられる．しかし，前者は重要文節の同定に用いる情報の規模が小さい上に一般的ではなく，後者は他の手法との比較が行われていないなど，まだまだ重要個所抽出法の有効性は明らかになっていない．さらにこの手法は，抽出した重要個所を列挙するだけでは文法的な正しさを保証するのが困難であるという欠点を持つ．既存研究 [和田 02] では重要個所の抽出時に係り受け情報を考慮することでこの問題に対処する（大石ら [大石 03] はこの問題を扱っていない）のに対し，本研究では重要個所抽出とは独立に係り受け情報を用いた文法性の保証を試みる．

3. SVM を用いた重要文節抽出

3·1 SVM

SVM は二値分類のための教師あり学習アルゴリズムである [Vapnik 95]．学習データとして m 個の事例に対応する特徴ベクトル $\boldsymbol{x_1}, \cdots, \boldsymbol{x_m} \in R^n$ とそれらの事例に対する正解のクラスラベル（正例，負例）$y_1, \cdots, y_m \in \{+1, -1\}$ が与えられたとき，SVM は正例と負例間の距離が最大となるような分離超平面 $f(\boldsymbol{x}) = 0$ を決定する．

$$f(\boldsymbol{x}) = \boldsymbol{w} \cdot \boldsymbol{x} + b = \sum_{j=1}^{n} w_i x_i + b \tag{1}$$

しかし多くの実問題では線形分離不可能であり，このような場合への対処法として，ソフトマージン法がある．この方法では完全な線形識別をあきらめ，事例の誤識別の度合いをスラック変数 $\xi_i (\geq 0)$ を用いて表し，できるだけ誤識別の度合いを小さくするように最適化を行う．

以上の分類問題は，以下に示す最適化問題（２次計画問題）として定式化できる．

$$\begin{aligned} \text{Minimise}_{\boldsymbol{w},b,\xi} \quad & \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{m}\xi_i \\ \text{subject to} \quad & y_i(\boldsymbol{w}\cdot\boldsymbol{x_i} + b) \geq 1 - \xi_i \end{aligned} \tag{2}$$

式(2) におけるパラメータ C は，誤識別の度合いをどれだけ考慮するかを決める定数である．式(2) の最適化問題を Lagrange の未定乗数法を用いて解くと，以下の判別関数 $f(\boldsymbol{x})$ が得られる．

$$f(\boldsymbol{x}) = \sum_{i=1}^{m} \alpha_i y_i (\boldsymbol{x_i} \cdot \boldsymbol{x}) + b \tag{3}$$

ここで α_i は Lagrange 乗数の最適解である．式(3) の符号によって，正例か負例かを分類することができる．

さらに，式(3) の内積をカーネル関数 $K(\boldsymbol{x_i}, \boldsymbol{x})$ で置き換えることによって，線形 SVM を非線形に容易に拡張することができる．

$$f(\boldsymbol{x}) = \sum_{i=1}^{m} \alpha_i y_i K(\boldsymbol{x_i}, \boldsymbol{x}) + b \tag{4}$$

本研究では，学習結果の解釈が比較的容易であるため，カーネル関数として式(5) に示す多項式関数を用いる．

$$K(\boldsymbol{x_i}, \boldsymbol{x}) = (\boldsymbol{x_i} \cdot \boldsymbol{x} + 1)^d \tag{5}$$

*1 Jing & McKeown[Jing 00b] は，人手による要約の分析を通じて，文結合の他に，文縮約，構文変形，語彙換言，一般化・特殊化，並べ換えの操作によって原文から要約が生成できることを示している．文縮約は非重要個所の削除手法で対処可能であり，他の操作は重要個所抽出とは異なる処理が必要である．

表 1 特徴ベクトルを構成する素性の一覧

素性	次元数	内容
文節 P を含む文 S に関する素性		
文のタイトル類似度	10	式(6) で定義される文 S とタイトル T の類似度
文の TF・IDF 値	10	式(7) で定義される文 S の TF・IDF 値
文書内位置	10	文 S を含む文書 D の総文字数に対する，D の先頭から文 S までの文字数の割合
段落内位置	10	文 S を含む段落 Q の総文字数に対する，Q の先頭から文 S までの文字数の割合（段落は原文における字下げに基づいて決定する.）
文の長さ	10	0 から 1 までの実数に正規化した，文 S の文字数
文節 P に関する素性		
文節のタイトル類似度	1	タイトル T に含まれる名詞が文節 P に出現するかどうかを示す二値
文節の TF・IDF 値	10	式(9) で定義される文節 P の TF・IDF 値
文内位置	10	文 S の総文字数に対する，S の先頭から文節 P までの文字数の割合
形態素（大分類）	14	JUMAN の解析結果から得られる「感動詞，形容詞，指示詞，助詞，助動詞，接続詞，接頭辞，接尾辞，動詞，特殊，判定詞，副詞，名詞，連体詞」の 14 種類の大分類
形態素（小分類）	62	JUMAN の解析結果から得られる普通名詞，格助詞，読点等の計 62 種類の小分類
係り方（大分類）	5	KNP の解析結果から得られる「一般，並列，同格，独立，文末」の 5 種類の大分類
係り方（小分類）	24	KNP の解析結果から得られる，連体，連用，文末，ヲ格等の計 24 種類の小分類
係る文節数	6	文節 P に直接係る文節の数（0 から 5 までの 6 種類）
固有表現	7	NExT の固有表現抽出結果に基づいて得られる「日付表現，時間表現，金額表現，割合表現，人名，場所名，組織名」の 7 種類の分類
用言の意味属性	36	日本語語彙大系 [池原 99] で定められた，物理的移動，精神的移動，思考動作，身体動作等の用言意味属性 36 種類

3·2 特徴ベクトルの構成

SVM による重要文節抽出を行うために，対象とする文書中の各文に対して JUMAN[*2]による形態素解析，KNP[*3]による係り受け解析，NExT[*4]による固有表現抽出の処理を行い，表 1 に示す素性に基づき各文節 P_i の特徴ベクトル $\boldsymbol{x_i}$ を構成する．なお，文節の単位は係り受け解析器 KNP の出力に基づいている．

表 1 に示す素性は，文節 P を含む文 S の性質を表現したものと，文節 P の性質を表現したものに大別される．文の性質に関する素性は既存研究 [平尾 03] で用いられた情報を参考にして選択し，文節に関する素性は，他の文節との関係，文と文節の関係，文書と文節の関係などに注目して適切と思われる素性を選択した．

特徴ベクトルの各要素（次元）は 0 か 1 の二値を取るように定義する．そこで実数の値を取る素性については，0 から 1 までの値を取るように文書内での最大値で正規化して，その値が $[0.0, 0.1), [0.1, 0.2), \cdots, [0.9, 1.0]$ の 10 区間のいずれに含まれるかを表す 10 次元の二値ベクトルで表現する[*5]．例えば，正規化した値が 0.55 ならば，それに対応する 10 次元の値は 0000010000 となる．また，種類を表す素性については，それぞれのカテゴリに次元を割り当て，該当するカテゴリに対応する次元を 1，それ以外の次元を 0 とする．以上により，各文節は 225 次元（$n = 225$）の二値ベクトルとして表現される．以下に，表 1 の中で詳細な説明が必要な素性について述べる．

§ 1 文のタイトル類似度

文 S における名詞の出現頻度ベクトル $\boldsymbol{v}(S)$ と，S を含む文書 D のタイトル T における名詞の出現頻度ベクトル $\boldsymbol{v}(T)$ の余弦として，文 S のタイトル類似度 $\mathrm{Sim}(S)$ を定義する．

$$\mathrm{Sim}(S) = \frac{\boldsymbol{v}(S) \cdot \boldsymbol{v}(T)}{\|\boldsymbol{v}(S)\| \cdot \|\boldsymbol{v}(T)\|} \tag{6}$$

§ 2 文の TF・IDF 値 [平尾 03]

文 S の TF・IDF 値 $\mathrm{TI}(S)$ を，次式で定義する．

$$\mathrm{TI}(S) = \sum_{t \in S} tf(t, S) \cdot w_s(t, D) \tag{7}$$

式(7) において，$tf(t, S)$ は文 S における名詞 t の出現頻度，$w_s(t, D)$ は文書 D における単語 t の TF・IDF 値であり，次式で定義する．

$$w_s(t, D) = 0.5 \left(1 + \frac{tf(t, D)}{tf_{\max}(D)}\right) \cdot \log_2 \left(\frac{N}{df(t)}\right) \tag{8}$$

ここで，$tf(t, D)$ は文書 D における名詞 t の出現頻度，$tf_{\max}(D)$ は文書 D における $tf(t, D)$ の最大値，N は対象とする文書集合に含まれる文書の数，$df(t)$ は名詞 t を含む文書数である．

*2 http://www.kc.t.u-tokyo.ac.jp/nl-resource/juman.html
*3 http://www.kc.t.u-tokyo.ac.jp/nl-resource/knp.html
*4 http://www.ai.info.mie-u.ac.jp/~next/next.html
*5 10 次元の二値ベクトルを用いて実数値を表現した理由は，6·3 節で述べるように，すべての素性を二値で表現することによってどの素性（の組み合わせ）が重要／非重要の判断に有効かを分析しやすくするためである [平尾 03].

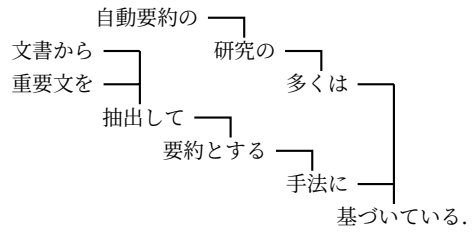

図 1 係り受け木の例

§3 文節の TF・IDF 値

文節 P の TF・IDF 値を次式で定義する.

$$\mathrm{TI}(P) = \sum_{t \in P} tf(t,P) \cdot w_p(t,D) \qquad (9)$$

式(9) において，$tf(t,P)$ は文節 P における名詞 t の出現頻度，$w_p(t,D)$ は文書 D における単語 t の TF・IDF 値であり，次式で定義する.

$$w_p(t,D) = tf(t,D) \cdot \log_2 \left(\frac{N}{df(t)} + 1 \right) \qquad (10)$$

3·3 SVM を用いた重要文節抽出による要約

SVM による重要文節抽出は，テキスト中の各文節（の特徴ベクトル）に対して重要か非重要かの判別を行う二値分類問題と考えられる．したがって，学習で得られた判別関数 $f(\boldsymbol{x})$ の値が正となる特徴ベクトル $\boldsymbol{x}$ を持つ文節を重要文節として抽出すればよい.

しかし，判別関数の値が正となる文節を抽出するだけでは，抽出文節数の割合が指定された要約率に一致するとは限らない．そこで本研究では，学習された分離平面との距離に相当する $f(\boldsymbol{x})$ の値の大きい順に要約率を満たすまで文節を抽出するという手法を採用する[*6].

4. 係り受け情報を用いた重要文節群の抽出

4·1 係 り 受 け 木

日本語文は，文節をノード，文節どうしの係り受け関係をノード間のリンクとした木構造（係り受け木）で表すことができる．図 1 に係り受け木の例を示す.

係り受け木の根ノードはどの文節にも係らない文末尾の文節であり，葉ノードはどの文節からも係らない文節である．根ノードを含む部分木は，それぞれの文節の係り先の文節を含むため，日本語として文法的に正しい文となることが一般的に期待できる [伊藤 03]．また，根ノードを含まない部分木を原文に出現する順に並べた場合，文末の一部が文法上不適切な表現になる可能性はあるものの，文よりも小さい意味的なまとまりを持つ単位になると考えられる.

*6 なお 4 章で述べるように，係り受け情報を用いて重要文節群を抽出する際に要約率を考慮すれば，必ずしもこの時点で要約率を考慮する必要はない．しかし係り受け情報を用いる方法と用いない（つまり SVM による抽出のみを行う）方法の比較検討を行うために，本研究では SVM による文節抽出時にも要約率を考慮した.

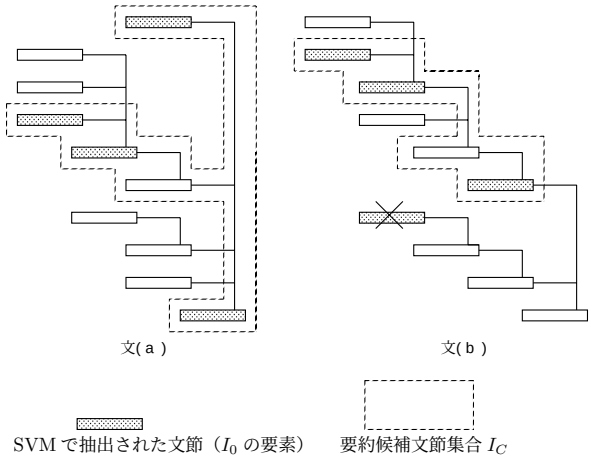

図 2 係り受け木を利用した重要文節群の抽出法による抽出例

4·2 係り受け木を利用した重要文節群の抽出法

以下に示すアルゴリズムを用いて，SVM で抽出された重要文節をもとに文法的に適切な文節のまとまりを要約文として抽出する.

(1) 係り受け解析器 KNP により各文が木構造表現された文書 D と，SVM によって重要と判断された文節集合 I_0 を入力として受け取る.

(2) 要約候補文節集合 $I_C = \phi$，要約文節集合 $I_S = \phi$ とする.

(3) 文書 D に含まれるすべての文節 P に対し，以下のいずれかの処理を行う.

- a $P \in I_0$ の場合，P の先祖または P の子孫のいずれかが I_0 の要素であるならば，文節 P を I_C に加える.
- b $P \notin I_0$ の場合，P の先祖に I_0 の要素が存在し，かつ，P の子孫に I_0 の要素が存在するならば，文節 P を I_C に加える.

(4) I_C の要素からなるすべての部分木 T（文節集合）に対し，次式の $R(T)$ の値を計算する.

$$R(T) = \frac{T \text{ に含まれる } I_0 \text{の要素数}}{T \text{ に含まれる } I_C \text{の要素数}}$$

そして与えられた要約率を満たすまで，$R(T)$ の大きい順に部分木 T を選択して，その部分木を構成するすべての文節を I_S に加える.

(5) I_S に含まれるすべての文節を原文で出現する順に並べて要約文とする.

このアルゴリズムによる要約文抽出の概要を図 2 に示す．この図において，網かけされたノードは I_0 の要素を表し，破線で囲まれた部分は I_C の要素を表す．図 2 の文 (a)，(b) ともに，上記のステップ (3b) において，I_0 の要素でない（つまり SVM で重要文節として抽出されなかった）文節を含む文法的に適切なまとまりが要約として抽出されることになる．逆に文 (b) の例のように，SVM によって重要と判断された文節であっても他の重要文節か

ら孤立している場合には，ステップ (3a) によって要約文には含まれないことになる．

5. 評 価 実 験

5·1 重要文節の正解データ

本実験で用いる重要文節の正解データは，第 2 回 NTCIR ワークショップ（NTCIR-2）における自動要約タスク TSC-1[Fukusima 01] で作成されたテストコレクション（NTCIR-2 SUMM）のうちの重要個所抽出データをもとに作成した．TSC-1 の重要個所抽出データは，CD-毎日新聞 94，95，98 年版に含まれる 180 個の記事を対象として，20%と 40%の要約率で人手により要約されたものである．本実験では，これらのうちの報道記事（全 69 記事）と社説（全 43 記事）を用い，それ以外の記事は必ずしもジャンルが明確でないため用いなかった．なお，3·2 節で述べた文や文節の TF・IDF 値を求める際に必要な文書集合には 180 記事すべてを用いた．

具体的には，以下に示す手順にしたがって，TSC の重要個所抽出データから文節単位の正解データを構成した．

(1) TSC の重要個所抽出データに含まれる各文 S_i に対して，各文字を要素とするベクトル表現での余弦類似度が最も高くなる原文書中の文 $T_{f(S_i)}$ を選択する．（この対応付けの方法は文献 [Takeuchi 01] を参考にした．）

(2) 文 S_i の各文節 s_{ij} に対して，先頭 s_{i1} から順に，以下の基準により $T_{f(S_i)}$ 中の対応文節を決定する．文節 s_{ij} の対応文節は，直前の文節 $s_{i,j-1}$ の対応文節 $t_{f(S_i),g(s_{i,j-1})}$ よりも後ろにあるすべての文節 $t_{f(S_i),k}$ $(k > g(s_{i,j-1}))$ の中で，文字ベースの余弦類似度が最も高い文節とする．このようにして選択された文 $T_{f(S_i)}$ 中の対応文節 $t_{f(S_i),g(s_{i1})}, t_{f(S_i),g(s_{i2})}, \cdots$ を重要文節の正解データとする．

5·2 方 法

5·1 節の正解データを用いて，以下の手順で実験を行った．まず，各ジャンル・要約率ごとに，正解データをできるだけ同じ個数になるように 5 個のグループに分割した．（例えば，報道記事では，14 記事のグループが 4 個，13 記事のグループが 1 個となった．）次に，それらのうちの 4 個のグループを学習データとして SVM による学習を行い，残りの 1 グループをテストデータとして学習結果による重要文節の抽出を行うという学習・テストのサイクルを，すべてのグループがテストデータとなるように 5 回繰り返した．

SVM による学習において，カーネル関数として用いた多項式関数の次数 d は，1 と 2 の場合でそれぞれ実験を行った．平尾らの研究 [平尾 03] で得られている知見（SVM による重要文抽出において次数が 2 のときに最も良い性能を示す）に従い，3 次以上の多項式関数は用いなかった．また，式(2) の目的関数のパラメータ C については，予備実験で良好な結果を得た値 $C = 0.1$（$d = 1$），$C = 0.0005$（$d = 2$）を用いた．

表 2 実験結果（F 値）

	$d=1$		$d=2$		
	SVM-D	SVM	SVM-D	SVM	LEAD
報道 20%	0.397	0.392	0.433	0.430	<u>0.457</u>
報道 40%	0.493	0.488	<u>0.501</u>	0.489	0.483
社説 20%	0.214	0.229	0.279	<u>0.284</u>	0.251
社説 40%	<u>0.491</u>	0.482	0.490	0.486	0.387

5·3 結 果

実験結果の評価基準には，再現率と適合率の調和平均である F 値を用いる．

$$\text{F 値} = \frac{2PR}{P+R}$$

$$\text{再現率 (R)} = \frac{\text{システムが出力した正解の重要文節数}}{\text{正解の重要文節数}}$$

$$\text{適合率 (P)} = \frac{\text{システムが出力した正解の重要文節数}}{\text{システムが重要文節として出力した文節数}}$$

表 2 に実験結果（F 値）を示す．SVM と係り受け情報を用いた提案方法（SVM-D）による結果の他に，比較対象として SVM による重要文節抽出のみを行う方法（SVM）および重要文抽出法である LEAD 法（要約率を満たすまで文書の先頭から文を選択する手法）による結果も示す．また，ジャンル・要約率ごとの F 値の最大値を下線で示す．

表 2 の実験結果は，以下のようにまとめられる．

- 社説 40%（SVM-D）を除くすべてのジャンル・要約率において，多項式関数の次数 $d = 2$ のほうが，$d = 1$ よりも F 値が高い．この結果は素性の組み合わせを考慮した学習の有効性を示しており，平尾らの結果 [平尾 03] とも一致する．よって，本稿の以下では次数が 2 の場合の結果のみを議論の対象とする．
- 係り受けを考慮する手法（SVM-D）と SVM による抽出のみを行う手法（SVM）を比較すると，それほど大きな差ではないが，社説 20%要約以外は提案手法（SVM-D）のほうが F 値が高い．これは係り受け情報を考慮する提案手法の有効性を示すものと考えられるが，さらに詳細な考察は 6·2 節で行う．
- 新聞記事の要約に有効であることが知られている LEAD 法 [Brandow 95, Mani 01] との比較では，報道 20%要約以外は，提案手法（SVM-D）のほうが F 値が高い．特に社説の要約でその差が大きい．この結果は，社説の要約には先頭から文を抽出するだけでは不十分であるのに対し，報道記事の要約では（特に要約の長さが短い場合には）LEAD 法が効果を発揮するためであると考えられる．さらに，6·1 節や 6·3 節では，実験結果を詳細に分析することによって，LEAD 法のような重要文抽出に基づく要約

表 3 重要文節の割合による分類別の正解データの統計

ジャンル＋要約率	分類	文数	文節数	重要文節数	重要文節数の割合
報道 20%	C1	66	440	440	1.000
	C2	205	2609	1681	0.644
	C3	107	1789	612	0.342
	C4	904	8923	0	0.000
	計	1282	13761	2733	0.199
報道 40%	C1	188	1385	1385	1.000
	C2	370	4948	3503	0.708
	C3	102	1625	617	0.380
	C4	622	5803	0	0.000
	計	1282	13761	5505	0.400
社説 20%	C1	68	409	409	1.000
	C2	263	2636	1731	0.657
	C3	123	1660	598	0.360
	C4	1006	8925	0	0.000
	計	1460	13630	2738	0.201
社説 40%	C1	225	1537	1537	1.000
	C2	483	4759	3325	0.699
	C3	113	1460	572	0.392
	C4	684	5874	0	0.000
	計	1460	13630	5434	0.399

手法に比べて，重要文節抽出による要約手法がどのような点で優れているか／劣っているかを考察する．

6. 考　　察

6·1 重要文節の割合の違いによる性能の比較

1 章や 2 章で述べたように，文抽出を中心とする要約手法は，人間が行うような複数文の分割・再構成による要約のための要素技術を提供できない．したがって，原文の一部のみが重要文節として正解データに含まれる文に対する本手法の性能を調べることで，重要文節抽出に基づく本手法がアブストラクト生成のための要素技術として有効であるかどうかを検討することができる．

そこで本節では，評価実験に用いた文書に含まれるそれぞれの文を，原文に対する正解重要文節の文字数の割合によって以下の 4 つのクラスに分類し，各クラスにおける提案手法と LEAD 手法の性能を比較する．

C1　文全体が重要文節である文（100%）
C2　文全体の 50%以上 100%未満の重要文節を含む文
C3　文全体の 50%未満の重要文節を含む文
C4　重要文節を全く含まない文（0%）

これらの中で C2 と C3 に分類される文に対して本手法の性能が LEAD 手法よりも優れていれば，それは本手法の重要文節抽出の有効性を示すことになる．逆に，これらのクラスにおける本手法の性能が良くなければ，本手法による重要文節抽出が有効に働いていないと言える．

表 3 に 4 つのクラス別の正解データの統計情報を示す．この表を見ると，例えば，実験に用いた報道 69 記事に含まれる文の総数は 1282 文，それらの総文節数は 13761 個であり，20%の要約率ではそのうちの 19.9% にあたる 2733 個の文節が重要文節であることがわかる．さらに全体が重要文節である文（C1）は 66 文（総文節数は 440 個），重要文節が半分以上である文（C2）は 205 文（総文節数 2609 個，重要文節数 1681 個）で，平均してクラス C2 に含まれる文の 64.4%が重要文節であること等が読み取れる．

クラス C2 または C3 に分類される文から抽出された重要文節は，全重要文節の約 70%から 85%を占める．(特に全体の 60 数%の重要文節が C2 に含まれる.）このことは，C2 や C3 での要約手法の性能が全体の性能に大きく影響を与えることを示している．一方で，重要文節を全く含まない文（C4）も，要約率 20%で全体の約 70%，要約率 40%でも全体の 50%弱存在している．よってこれらの文をいかに抽出しないかも全体の性能に影響する．

次に表 4 に，4 つのクラスごとの 3 手法の F 値および再現率，適合率の値を示す．なお，クラス C4 は正解となる重要文節が存在しないので，再現率と F 値は計算できない.（適合率は必ず 0 となる.）したがって，これらの再現率の欄には非重要文節における再現率（非重要文節の正解数に対して各手法が非重要と判断した文節数の割合）を括弧書きで示す．本節の以下では，文書のジャンルごとに SVM を用いる 2 手法と LEAD 手法の比較を行い，SVM-D と SVM の両手法の比較は 6·2 節で行う．

まず報道記事の要約について考察する．クラスごとに両手法の F 値を比較すると，20%要約ではいずれのクラスでも LEAD 法のほうが高いのに対して，40%要約のクラス C2 と C3 では提案手法（SVM-D もしくは SVM）のほうが高い．また，各手法におけるクラス間の F 値を比較すると，LEAD 法ではクラス C1 で最も F 値が高いのに対して，SVM-D 法および SVM 法ではクラス C2 で最も高い F 値が得られている．これらの結果は，SVM による重要文節抽出に基づく本要約手法が文の一部のみが要約に用いられる場合には有効であることを示している．一方，全体が要約に含まれるような文には LEAD 法に代表される文抽出法に利点があると言える．

クラス C2 や C3 における両手法の F 値の差の主原因は，LEAD 法の適合率の低さにある．この結果は当然と言えば当然であるが，LEAD 法の適合率はランダムに文節を選択したときの期待値（表 3 における重要文節数の割合に等しい）とおおよそ同じ値であるのに対して，SVM による手法ではその値を大きく上回っている．一方，再現率については，20%要約では LEAD 法のほうが高い値であるが，40%要約では SVM による手法のほうが高い．このことは，重要文節を直接抽出する本研究の手法が，特に 40%要約において，非抽出文に含まれる重要情報の欠落という 2 章で述べた重要文の非重要個所を削除する手法の問題点に対処できていることを示している．さらに，C4 における非重要文節の再現率でも，SVM に基づく手法が LEAD 法より高い値を示している．

一方，社説においては，要約率やクラスに関係なく，

表 4 重要文節の割合に基づく分類別の要約手法の比較

ジャンル＋要約率	分類	F 値			再現率			適合率		
		SVM-D	SVM	LEAD	SVM-D	SVM	LEAD	SVM-D	SVM	LEAD
報道 20%	C1	0.511	0.533	0.656	0.343	0.364	0.489	1.000	1.000	1.000
	C2	0.543	0.543	0.580	0.448	0.444	0.518	0.690	0.699	0.660
	C3	0.358	0.364	0.376	0.335	0.332	0.397	0.384	0.404	0.356
	C4	—	—	—	(0.931)	(0.921)	(0.904)	0.000	0.000	0.000
	全体	0.433	0.430	0.457	0.406	0.406	0.486	0.464	0.456	0.432
報道 40%	C1	0.544	0.541	0.611	0.374	0.371	0.440	1.000	1.000	1.000
	C2	0.629	0.599	0.606	0.553	0.504	0.529	0.730	0.739	0.711
	C3	0.471	0.480	0.424	0.501	0.475	0.457	0.444	0.486	0.395
	C4	—	—	—	(0.713)	(0.740)	(0.670)	0.000	0.000	0.000
	全体	0.501	0.489	0.483	0.502	0.467	0.498	0.499	0.513	0.470
社説 20%	C1	0.482	0.488	0.427	0.318	0.323	0.271	1.000	1.000	1.000
	C2	0.382	0.384	0.370	0.260	0.258	0.254	0.723	0.747	0.682
	C3	0.298	0.328	0.293	0.232	0.256	0.259	0.415	0.457	0.336
	C4	—	—	—	(0.851)	(0.848)	(0.814)	0.000	0.000	0.000
	全体	0.279	0.284	0.251	0.263	0.267	0.257	0.298	0.302	0.245
社説 40%	C1	0.639	0.637	0.454	0.470	0.468	0.294	1.000	1.000	1.000
	C2	0.584	0.574	0.533	0.478	0.461	0.427	0.751	0.760	0.708
	C3	0.501	0.522	0.419	0.491	0.493	0.470	0.511	0.555	0.378
	C4	—	—	—	(0.700)	(0.703)	(0.581)	0.000	0.000	0.000
	全体	0.490	0.486	0.387	0.477	0.467	0.394	0.503	0.508	0.380

SVM に基づく手法の F 値が LEAD 法の F 値を上回っている．この傾向は再現率や適合率でも同じであり，特に適合率の差が大きい．さらに 40%要約においては，LEAD 法の非重要文節の再現率が特に低い値となっている．以上の結果は，文書の前半に重要文が必ずしも含まれない社説記事では LEAD 法が不利であることを考えると必ずしも適切な比較とは言えないが，それでもクラス C2 や C3 における重要文節抽出に基づく本手法の有効性を示すものと考えられる．

6·2 係り受け情報の利用の有効性

本節では，係り受け情報を考慮する提案手法（SVM-D）と SVM による重要文節抽出のみによる手法（SVM）の比較検討を行う．5 章の実験結果に基づく比較とともに，文の読みやすさに関して行った簡単な評価実験の結果から，係り受け情報の利用の有効性を検証する．

§1 F 値による評価

5 章の表 2 で示したように，全体の F 値による評価では，社説 20%要約を除いては係り受けを考慮する SVM-D 法のほうが良好な結果である．さらに表 4 で示すように，適合率は係り受けを考慮しない SVM 法のほうが高いが，再現率は係り受けを考慮した SVM-D 法のほうが高いという傾向が見られる．この傾向はクラス別に見た場合でも同様である．特に F 値の差が最も大きい報道 40%要約では，クラス C2 や C3 において，適合率の減少が低く抑えられた上で，再現率が大きく向上している．

以上の結果を総合すると，提案手法である SVM-D 法は，SVM 法に比べて多くの非重要文節を重要文節として抽出してしまうという欠点があるが，重要文節をより多く抽出できるという利点がそれに勝っていると言える．つまり，係り受け情報を考慮することによって，特徴ベクトルでは表現するのが難しい「人間が重要個所と判断する文法的に妥当な文節のまとまり」をより適切に選択でき，これが性能の向上に貢献していると考えられる．

さらに SVM-D 法の性能が SVM による学習結果を適切に反映した結果である（つまり，係り受け情報によって SVM の学習結果と大きく異なった判断をしていない）ことを確かめるために，SVM-D 法と SVM 法で判断の異なる文節の割合を求めたところ，報道 20%で 0.040，報道 40%で 0.074，社説 20%で 0.073，社説 40%で 0.066 となった．両手法で判断の異なる文節は全体の 1 割もなく，重要文節の判断に大きな違いがないことがわかる．よって，SVM-D 法が SVM による学習結果を適切に反映しながら性能を向上させる妥当な手法であると言える．

§2 読みやすさによる評価

大学生・大学院生 10 名に対して，要約の読みやすさの評定実験を行った．実験では，5 章の評価実験に用いた報道記事と社説から各 5 文書，計 10 文書を無作為抽出し，各文書に対して SVM-D 法と SVM 法で作成した 20%要約および 40%要約の読みやすさを 5 段階（1:読みにくい，2:やや読みにくい，3:どちらともいえない，4:やや読みやすい，5:読みやすい）で評定してもらった．

表 5 に評定結果（10 名の評定値の平均と標準偏差）を示す．報道 20%以外では，SVM-D 法のほうが文章の読

表 5 抽出文節群の読みやすさに関する要約手法の比較

	SVM-D		SVM	
	平均	標準偏差	平均	標準偏差
報道 20%	2.84	0.72	2.94	0.59
報道 40%	3.72	0.86	3.50	0.70
社説 20%	2.96	0.86	2.20	0.90
社説 40%	3.56	0.82	2.00	0.54

みやすさの平均評定値が高く，特に社説ではこの傾向が顕著であった．また，報道 20%要約においても，5 文書のうちの 3 文書の平均評定値は SVM-D 法のほうが高かった．これらの結果は，係り受け構造を考慮することによって文としての読みやすさがおおむね向上することを示しており，文の読みやすさの観点からも提案手法が有効であると言える．

6·3 重要文節の抽出に有効な素性の分析

本節では，SVM による重要文節抽出において，どのような素性が重要／非重要の判断に有効かを分析する．特に文節固有の素性がこの判断にどのくらい有効かを調べることによって，重要文節抽出法の有効性を検討する．

どのような素性（の組み合わせ）が重要／非重要の判断に大きな影響を及ぼすかは，SVM の学習で得られる式(4) の $f(\boldsymbol{x})$ から知ることができる [平尾 03]．カーネル関数として用いた多項式関数(5) の次数 $d=2$ の場合，式(5) の内積を展開して，さらに特徴ベクトルの各要素が 0 か 1 の二値であることを考えると，$f(\boldsymbol{x})$ は以下の式で表される．

$$f(\boldsymbol{x}) = A + \sum_{j=1}^{n} B_j x_j + \sum_{j=1}^{n-1} \sum_{k=j+1}^{n} C_{jk} x_j x_k \quad (11)$$

$$\text{where} \quad A = b + \textstyle\sum_{i=1}^{m} \alpha_i y_i \quad B_j = 3 \sum_{i=1}^{m} \alpha_i y_i x_{ij}$$
$$C_{jk} = 2 \textstyle\sum_{i=1}^{m} \alpha_i y_i x_{ij} x_{ik}$$

上式における x_{ij} は学習事例の特徴ベクトル $\boldsymbol{x_i}\,(1 \leq i \leq m)$ の次元 j の要素を表す．式(11) を見ると，係数 B_j は素性 x_j が成立するときのスコア，係数 C_{jk} は 2 つの素性 x_j, x_k が同時に成立するときのスコアと解釈することができる．よってこれらの係数の絶対値が大きいほど，対応する素性（の組み合わせ）が重要（係数が正の場合）または非重要（係数が負の場合）の分類に有効であると考えられる．

各ジャンル・要約率について，以上の方法で求めた係数 B_j, C_{jk} の絶対値の大きい順に 10 位までを正負ごとに示したのが表 6 である．この中で，太字で示されている素性が文節に固有の素性である．表 6 を概観すると，まず一つの素性よりも素性の組み合わせが多く出現していることがわかる．(正・負いずれの場合も，85%が素性の組み合わせである.) この傾向は，多項式関数の次数 $d=2$ のほうが判別性能が高いという 5 章の結果と整合する．

文節固有の素性に注目すると，社説 20%の場合を除いて，重要文節の判断に有効な素性の中に文節固有のものが多く含まれている．これらのほとんどは文に関する素性との組み合わせで出現している．この結果は，文の重要さと文内における文節の重要さの両方が重要文節の判断に寄与していると解釈できる．一方，非重要文節の判断に有効な素性の中に文節固有のものはほとんど見られず，文に関する素性（の組み合わせ）が多く含まれている．このことは，非重要文節の判断（主にクラス C4 に含まれるような非重要文の文節を非重要と判断する場合であると思われる）には主に文に基づく素性で十分であると解釈できる．

なお社説 20%要約には文節固有の素性が全く出現していないが，このことは報道 20%要約に比べて著しく性能が悪いという評価結果（大石らの研究 [大石 03] でも同様である）と関係していると思われる．文に関する素性についても，非常に類似した（組み合わせ）素性が重要・非重要の判断の両方に有効であるなど，首尾一貫した特徴が見られない．社説を短く要約するためには，本研究で用いた文や文節などの局所的な情報だけでなく，より文書の構造を反映した大域的な情報 [綾 05] が本質的に重要になってくるのであろう．

次に，どのような素性が重要／非重要の判断に有効かを具体的に見ていく．重要文節に有効な素性を見ると，文に関する素性である「文書内位置」を含むものが圧倒的に多く，続いて「文のタイトル類似度」が多く出現している．報道記事では，文書内位置が文書の先頭であるという素性（$[0.0, 0.1)$）と文節固有の素性（文節のタイトル類似度や名詞・助詞などの形態素，係り方の種類など）の組み合わせに高いスコアが与えられている．この結果は，報道記事には LEAD 法が有効であることを裏付けており，さらに文書先頭の文中の名詞句（特にタイトルの名詞を含む）や（サ変）動詞句が重要文節として選ばれやすいことを示している．一方，社説（特に 40%要約）では文書の末尾を示す素性（$[0.9, 1.0]$）に高いスコアが与えられており，社説では結論や主張を述べるのが文書の最後であるという一般的な特徴が学習されたと考えられる．また，文のタイトル類似度が低いまたは文が長いという特徴も有効である．これらの結果はおおむね既存研究 [平尾 03, 大石 03] とも一致している．

一方，非重要文節の判断に有効な素性に関しては,「文のタイトル類似度」が圧倒的に多く関与しており，続いて「文書内位置」と「文の長さ」が多く出現している．報道・社説いずれにおいても，文のタイトル類似度が低く，文の長さが相対的に長い文に大きな負のスコアが与えられている．これらはおそらく非重要文の特徴であり，このような文に含まれる文節が非重要と判断される．さらに，報道記事では，文書の先頭の文であっても，長い文の文節や他のどの文節からも係らない文節が非重要と判断されやすくなっている．このような文節固有の素性は，重要文中の非重要文節の同定に役立っていると言える．

表 6 重要文節／非重要文節の分類に有効な素性（各ジャンル・要約率ごとに上位 10 個）

	重要文節の判断に有効な（正の係数を持つ）素性	非重要文節の判断に有効な（負の係数を持つ）素性
報道	文書内位置 [0.0, 0.1)，**文節のタイトル類似度 1**	文のタイトル類似度 [0.0, 0.1)，文書内位置 [0.0, 0.1)
20%	文書内位置 [0.0, 0.1)，**形態素（小分類）：サ変名詞**	文書内位置 [0.0, 0.1)，文の長さ [0.5, 0.6)
	文書内位置 [0.0, 0.1)	文のタイトル類似度 [0.5, 0.6)，文の長さ [0.9, 1.0]
	文のタイトル類似度 [0.7, 0.8)	文書内位置 [0.0, 0.1)，**係る文節数 0**
	文書内位置 [0.0, 0.1)，文のタイトル類似度 [0.4, 0.5)	文の長さ [0.7, 0.8)，**文節の TF・IDF 値 [0.9, 1.0]**
	文書内位置 [0.0, 0.1)，段落内位置 [0.2, 0.3)	文のタイトル類似度 [0.0, 0.1)，文書内位置 [0.1, 0.2)
	文書内位置 [0.0, 0.1)，文の長さ [0.8, 0.9)	文のタイトル類似度 [0.1, 0.2)，**文節のタイトル類似度 1**
	文書内位置 [0.0, 0.1)，**係り方（大分類）：一般**	文書内位置 [0.0, 0.1)，段落内位置 [0.5, 0.6)
	文書内位置 [0.1, 0.2)，文のタイトル類似度 [0.2, 0.3)	文のタイトル類似度 [0.1, 0.2)
	文書内位置 [0.0, 0.1)，**係る文節数** 1	文のタイトル類似度 [0.3, 0.4)，文の長さ [0.7, 0.8)
報道	文書内位置 [0.0, 0.1)	文のタイトル類似度 [0.0, 0.1)
40%	文のタイトル類似度 [0.4, 0.5)	文のタイトル類似度 [0.0, 0.1)，文書内位置 [0.0, 0.1)
	文書内位置 [0.0, 0.1)，**係り方（大分類）：一般**	文のタイトル類似度 [0.1, 0.2)
	文書内位置 [0.0, 0.1)，**形態素（大分類）：名詞**	文のタイトル類似度 [0.3, 0.4)，文の長さ [0.6, 0.7)
	文書内位置 [0.0, 0.1)，**形態素（大分類）：助詞**	文のタイトル類似度 [0.1, 0.2)，文の長さ [0.4, 0.5)
	文のタイトル類似度 [0.4, 0.5)，**係り方（大分類）：一般**	文のタイトル類似度 [0.2, 0.3)
	文書内位置 [0.0, 0.1)，**文節のタイトル類似度 1**	文の長さ [0.5, 0.6)，文の TF・IDF 値 [0.3, 0.4)
	文節のタイトル類似度 1	文の長さ [0.6, 0.7)，文書内位置 [0.5, 0.6)
	文のタイトル類似度 [0.4, 0.5)，**形態素（大分類）：名詞**	文のタイトル類似度 [0.1, 0.2)，文書内位置 [0.8, 0.9)
	文書内位置 [0.0, 0.1)，**係る文節数 1**	文のタイトル類似度 [0.1, 0.2)，**形態素（小分類）：サ変名詞**
社説	文書内位置 [0.0, 0.1)，文の TF・IDF 値 [0.2, 0.3)	文のタイトル類似度 [0.1, 0.2)，文書内位置 [0.9, 1.0]
20%	文書内位置 [0.0, 0.1)，文の長さ [0.7, 0.8)	文のタイトル類似度 [0.0, 0.1)，文書内位置 [0.9, 1.0]
	文書内位置 [0.9, 1.0]，文の TF・IDF 値 [0.3, 0.4)	文のタイトル類似度 [0.2, 0.3)，文の長さ [0.4, 0.5)
	文書内位置 [0.9, 1.0]，文の長さ [0.3, 0.4)	文の TF・IDF 値 [0.6, 0.7)，文書内位置 [0.0, 0.1)
	文の TF・IDF 値 [0.2, 0.3)，文の長さ [0.7, 0.8)	文のタイトル類似度 [0.1, 0.2)，文の長さ [0.7, 0.8)
	文書内位置 [0.9, 1.0]，文のタイトル類似度 [0.2, 0.3)	文のタイトル類似度 [0.3, 0.4)，文の長さ [0.5, 0.6)
	文書内位置 [0.0, 0.1)，文のタイトル類似度 [0.3, 0.4)	文のタイトル類似度 [0.3, 0.4)，文の TF・IDF 値 [0.6, 0.7)
	文書内位置 [0.9, 1.0]，段落内位置 [0.9, 1.0]	文のタイトル類似度 [0.0, 0.1)，段落内位置 [0.8, 0.9)
	文のタイトル類似度 [0.1, 0.2)，文の TF・IDF 値 [0.5, 0.6)	文の TF・IDF 値 [0.7, 0.8)，文書内位置 [0.0, 0.1)
	文書内位置 [0.5, 0.6)，文の TF・IDF 値 [0.8, 0.9)	文のタイトル類似度 [0.4, 0.5)，文の TF・IDF 値 [0.2, 0.3)
社説	文書内位置 [0.9, 1.0]	文の長さ [0.5, 0.6)，文のタイトル類似度 [0.0, 0.1)
40%	文書内位置 [0.9, 1.0]，**形態素（大分類）：助詞**	文のタイトル類似度 [0.0, 0.1)
	文書内位置 [0.9, 1.0]，**係り方（大分類）：一般**	文の長さ [0.3, 0.4)，文のタイトル類似度 [0.0, 0.1)
	文書内位置 [0.9, 1.0]，**形態素（大分類）：名詞**	文の長さ [0.4, 0.5)，文の TF・IDF 値 [0.5, 0.6)
	文書内位置 [0.9, 1.0]，**文節の TF・IDF 値 [0.0, 0.1)**	文書内位置 [0.6, 0.7)
	文書内位置 [0.9, 1.0]，文の TF・IDF 値 [0.3, 0.4)	文の長さ [0.7, 0.8)，文書内位置 [0.5, 0.6)
	文の TF・IDF 値 [0.5, 0.6)，文の長さ [0.5, 0.6)	文書内位置 [0.3, 0.4)，文のタイトル類似度 [0.0, 0.1)
	文書内位置 [0.9, 1.0]，**形態素（小分類）：格助詞**	文の長さ [0.5, 0.6)，文書内位置 [0.3, 0.4)
	文書内位置 [0.9, 1.0]，文のタイトル類似度 [0.0, 0.1)	文の長さ [0.8, 0.9)，文の TF・IDF 値 [0.9, 1.0]
	文のタイトル類似度 [0.0, 0.1)，文の長さ [0.7, 0.8)	文書内位置 [0.8, 0.9)，文のタイトル類似度 [0.0, 0.1)

6･4 重要文節抽出と重要文抽出の併用法の可能性

ここまで実験結果の分析を通じて，重要文節抽出による手法の有効性を実証・考察してきたが，一方で，得られた結果は重要文抽出と重要文節抽出の併用による要約手法がより有効な手法になり得る可能性も示している．

例えば，まず重要文抽出法によって要約の下地となる文を選択（つまりクラス C4 に属するような非重要文を除去）し，選択された文の重要度に応じて，重要文節の選択もしくは非重要文節の削除を行うという手法が考えられる．このような併用法によって，クラス C1 に属する文をできるだけ選択しつつ，クラス C2 や C3 に属する文を適切に処理することが期待できる*7．また 6･3 節の分析（非重要文節の判断には文に基づく素性を考慮するだけで十分である）からも，重要文抽出法を用いて非重要文を除去する方法は妥当である．さらに，このような併用法によって，クラス C2 や C3 に含まれる文節のみを学習データとして用いて，正例と負例をバランス良く学習させることが可能になる．その結果，重要文中の非重要文節の判断に有効な文節固有の素性が適切に学習されることが期待できる．

*7 なお，ここで言う併用法は，「非重要文削除＋（重要文節抽出／非重要文節削除）」という重要文節の判断を生かす手法であり，その点で既存手法の「重要文抽出＋非重要文節削除」とは異なる手法である．

7. お わ り に

本研究では，重要個所抽出による新たな要約手法として，SVM によって抽出された重要文節をもとに，原文の係り受け情報を用いて文法的に適切な文節群を抽出する手法を提案した．そして，TSC-1 の正解データを用いた評価実験や読みやすさの評定実験を通じて，提案手法の

有効性を示した．特に文の一部だけが要約として重要であるような文に対して，重要文抽出型手法である LEAD 法よりも良好な結果が得られることを明らかにした．この結果は，重要文中の非重要文節を削除する手法の欠点に本手法が対処できることを示している．一方で，文全体が重要であるかまったく重要でない文では重要文抽出に基づく手法も優れており，重要文抽出と重要文節抽出の併用法がより性能の高い要約手法になる可能性も示唆した．

今後の課題としては，本研究の結果をもとにした併用法による要約手法の開発が挙げられる．さらに，1 章や 2 章で述べたように，アブストラクトとしての要約を生成するためには，文の再構成が不可欠である．提案手法による抽出結果をもとに，複数の文をまとめる等の再構成の手法の開発にも取り組んでいきたい．

謝　　辞

本研究で使用した NTCIR-SUMM テストコレクションは，国立情報学研究所の許諾を得て使用させて頂きました．また，本論文の執筆にあたり，数多くの有益なコメントを頂きました査読者の方々に感謝の意を表します．

◇ 参　考　文　献 ◇

[綾 05] 綾 聡平, 松尾 豊, 岡崎 直観, 橋田 浩一, 石塚 満: 修辞構造のアノテーションに基づく要約生成, 人工知能学会論文誌, Vol.20, No.3, pp.149–158 (2005).

[Brandow 95] Brandow, R., Mitze, K. and Rau, L.: Automatic condensation of electronic publications by sentence selection, *Information Processing and Management*, Vol. 31, No. 5, pp. 675–685 (1995).

[Edmundson 69] Edmundson, H.P.: New methods in automatic abstracting, *Journal of the Association for Computing Machinery*, Vol. 16, No. 2, pp.246–285 (1969).

[Fukusima 01] Fukusima, T. and Okumura, M.: Text summarization challenge: Text summarization evaluation at NTCIR Workshop2, in *Proc. of the Second NTCIR Workshop on Research in Chinese and Japanese Text Retrieval and Text Summarization* (2001).

[平尾 03] 平尾 努, 磯崎 秀樹, 前田 英作, 松本 裕治: Support Vector Machine を用いた重要文抽出法, 情報処理学会論文誌, Vol. 44, No. 8, pp. 2230–2243 (2003).

[Hirao 03] Hirao, T., Takeuchi, K., Isozaki, H., Sasaki, Y. and Maeda, E.: SVM-Based multi-document summarization integrating sentence extraction with *Bunsetsu* elimination, *IEICE Transactions on Information and System*, Vol. E86–D, No. 9, pp.1702–1709 (2003).

[池原 99] 池原 悟, 宮崎 正弘, 横尾 昭男, 中岩 浩巳, 小倉 健太郎, 大山 芳史, 林 良彦: 日本語語彙大系 CD － ROM 版, 岩波書店 (1999).

[伊藤 03] 伊藤 潤, 酒井 哲也, 平澤 茂一: 係り受け木を用いた日本語文書の重要部分抽出, 情報処理学会研究報告 NL–158–4, pp.19–24 (2003).

[Jing 00a] Jing, H.: Sentence reduction for automatic text summarization, in *Proc. of the 6th Applied Natural Language Processing Conference (ANLP'2000)*, pp.310–315 (2000).

[Jing 00b] Jing, H. and McKeown, K.R.: Cut and paste based text summarization, in *Proc. of the First Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL'2000)*, pp.178–185 (2000).

[Knight 02] Knight, K. and Marcu, D.: Summarization beyond sentence extraction: A probabilistic approach to sentence compression, *Artificial Intelligence*, Vol. 139, No. 1, pp. 91–107 (2002).

[Kupiec 95] Kupiec, J., Pedersen, J. and Chen, F.: A trainable document summarizer, in *Proc. of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'95)*, pp.68–73 (1995).

[Lin 99] Lin, C-Y.: Training a selection function for extraction, in *Proc. of the 18th International Conference on Information and Knowledge Management (CIKM'99)*, pp.55–62 (1999).

[Luhn 58] Luhn, H.P.: The automatic creation of literature abstracts, *IBM Journal of Research and Development*, Vol. 2, No. 2, pp.159–165 (1958).

[Mani 01] Mani, I.: *Automatic Summarization*, John Benjamins Publishing Company (2001), 奥村 学, 難場 英嗣, 植田 禎子（訳）: 自動要約, 共立出版 (2003).

[大石 03] 大石 亨, 西尾 修一郎, 藤田 純, 遠藤 雅人, 奥村 学, 難波 英嗣: 遺伝的アルゴリズムによる重要文節概念の獲得, 言語処理学会第 9 回年次大会発表論文集, pp.489–492 (2003).

[大竹 02] 大竹 清敬, 岡本 大吾, 児玉 充, 増山 繁: 重要文抽出, 自由作成要約に対応した新聞記事要約システム YELLOW, 情報処理学会論文誌：データベース, Vol. 43, No. SIG02(TOD13), pp.37–47 (2002).

[奥村 99] 奥村 学, 難波 英嗣: テキスト自動要約に関する研究動向, 自然言語処理, Vol.6, No.6, pp.1–26 (1999).

[Radev 02] Radev, D.R., Hovy, E. and McKeown, K.: Introduction to the special issue on summarization, *Computational Linguistics*, Vol. 28, No. 4, pp. 399–408 (2002).

[酒井 04] 酒井 浩之, 増山 繁: 動詞連体修飾節の省略可能性に関するコーパスからの知識獲得, 電子情報通信学会論文誌 D-II, Vol. J87, No. 8, pp.1641–1652 (2004).

[Takeuchi 01] Takeuchi, K. and Matsumoto, Y.: Acquisition of sentence reduction rules for improving quality of text summaries, in *Proc. of the 6th Natural Language Processing Pacific Rim Symposium (NLPRS2001)*, pp.447–452 (2001).

[Vapnik 95] Vapnik, V.: *The Nature of Statistical Learning Theory*, Springer-Verlag (1995).

[和田 02] 和田 裕二, 奥村 明俊, 浦谷 則好, 白井 克彦: 属性を用いた文節重要度に基づくニュース文要約, 言語処理学会第 8 回年次大会発表論文集, pp.543–546 (2002).

〔担当委員：奥村　学〕

2005 年 4 月 12 日　受理

著　者　紹　介

鈴木　大介

2004 年電気通信大学電気通信学部システム工学科卒業．現在，同大学院電気通信学研究科システム工学専攻博士前期課程に在学中．自然言語処理の研究に従事．

内海　　彰(正会員)

1988 年東京大学工学部反応化学科卒業．1993 年東京大学大学院工学系研究科情報工学専攻博士課程修了．博士（工学）．東京工業大学大学院総合理工学研究科助手，講師を経て，2000 年から電気通信大学電気通信学部システム工学科助教授．言語やその周辺を対象とした認知科学や言語情報処理の研究に従事．日本認知科学会，情報処理学会，言語処理学会，日本語用論学会，Cognitive Science Society 等各会員．