情 報 学 基 礎 63−16
(2001. 7. 25)

Support Vector Machine による重要文抽出

平尾 努† 前田 英作† 松本 裕治‡
†NTT コミュニケーション科学基礎研究所
〒619-0237 京都府相楽郡精華町光台 2-4
{hirao,maeda}@cslab.kecl.ntt.co.jp
‡奈良先端科学技術大学院大学 情報科学研究科
〒630-0101 奈良県生駒市高山町 8916-5
matsu@is.aist-nara.ac.jp

概要

重要文抽出では，文に関する複数の情報を統合的に扱うことにより精度の向上が期待できる．近年，機械学習手法を用いて文に関する複数の情報を統合的に扱う研究が盛んになってきている．本稿では，機械学習手法の一つである，Support Vector Machine (SVM) を用いた重要文抽出手法を提案する．TSC (Text Summarization Challenge) のコーパスを用いて，提案手法と決定木学習を用いた手法，Lead 手法の従来手法との比較評価を行なった結果，提案手法が統計的に有意な差で優れていることを確認した．

キーワード: 重要文抽出，文書要約，機械学習，Support Vector Machine

Sentence Extraction Based on Support Vector Machines

Tsutomu HIRAO† and Eisaku MAEDA† and Yuji MATSUMOTO‡
†NTT Communication Science Laboratories
2-4, Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0237 Japan
{hirao,maeda}@cslab.kecl.ntt.co.jp
‡Graduate School of Information Science, Nara Institute Science and Technology
8916-5, Takayama, Ikoma, Nara 630-0101, Japan
matsu@is.aist-nara.ac.jp

Abstract

It is known that integrating multiple information is effective for sentence extraction. In this paper, we describe a method of sentence extraction based on Support Vector Machines (SVM). To confirm the performance of our method, we have conducted experiments wiht two other comparison methods, one is a Lead-baed method, the other is a Decision Tree-based method. Results on TSC corpus show that our system has the highest accuracy.

Keywords: Sentence Extraction, Text Summarization, Machine Learning, Support Vector Machine

1 はじめに

重要文抽出とは，文書を構成する文集合から重要な文のみを抽出する技術のことである．抽出された文集合は，単なる文の抽出であるが故に可読性が損なわれることもあるが，それ自体を要約と考えることも可能である．したがって，重要文抽出は，より自然な要約を目指すための基礎的な技術の一つとして位置づけることがで

きる．

重要文抽出に関する研究は従来から数多く行なわれている．それら従来研究の多くは，単独あるいは複数のてがかりに基づき文の重要度を決定している．てがかりとしては，文を構成する単語の重要度，文の出現位置，文書構造，てがかり表現，などの情報が用いられる．こうした複数のてがかりを統合的に扱う手法として，Edmundson [2]，野畑ら [6] は，各てがかりに対してスコアを与え，人手によって決定した重みを考慮したスコアの線形和を文の重要度とする手法を提案している．しかし，てがかりの数が多くなると，人手により重みの最適値を決定することが困難になるという問題があった．

一方，大量の訓練データが与えられた場合には機械学習手法が有効であることが知られており，近年，自然言語処理の様々な研究分野において注目されている．Aone ら [1]，Mani ら [5]，野本ら [14]，奥村ら [10] は複数のてがかりを利用した重要文抽出を機械学習手法によって実現している．ただし，これらは決定木学習を基本としたものであった．機械学習手法の一つである Suport Vector Machine (以下，SVM) [9] は，文書分類 [3, 15]，チャンキング [4]，係り受け解析 [13] などの自然言語処理に応用されその有効性が報告されている．

そこで本稿では，SVM を用いた重要文抽出手法を提案する．評価型ワークショップである NTCIR (NII-NACSIS Test Collection for IR Systems) Workshop-2 のサブタスクとして開催された TSC (Text Summarization Challenge) の重要文抽出タスクのデータを用いて評価を行ない，その有効性を示す．

以下，2 章では SVM の概要と SVM を利用した重要文抽出手法について述べ，3 章で評価実験の概要と結果を示し，考察を行なう．

2 Support Vector Machine に基づく重要文抽出手法

2.1 Support Vector Machine (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．概念図を図 1 に示す．

訓練データとして以下のベクトル集合を考える．

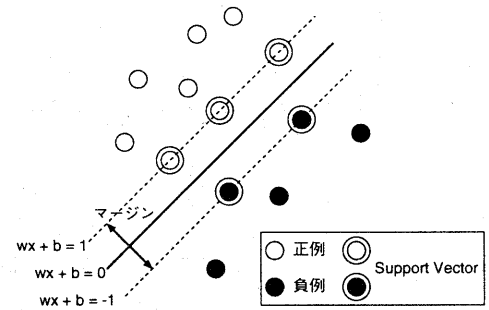

図 1: SVM の概念図

$$(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_l, y_l), \quad \mathbf{x}_i \in \mathbf{R}^n,\ y_i \in \{+1, -1\}$$

$\mathbf{x}_i$ は事例 i を表わす n 次元の特徴ベクトルである．y_i は，事例 i が正例であるときに 1，負例であるときに-1 をとる．SVM は，$\mathbf{x}_i$ を以下の分離平面で正例，負例に分類する．

$$\mathbf{w} \cdot \mathbf{x} + b = 0,\ \mathbf{w} \in \mathbf{R}^n, b \in \mathbf{R} \qquad (1)$$

ただし，このような分離平面は一般的に多数存在するが，SVM ではマージン[1] が最大になるように $\mathbf{w}$ と b を決定する．

訓練事例を線形分離できない場合も考慮に入れてマージンを最大化するためには，非負の変数である ξ を導入し，

$$\frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^{l}\xi_i \qquad (2)$$

を最小化すればよい．第一項はマージンの大きさに関する項，大二項は分離できなかった訓練事例がそれぞれの平面 $\mathbf{w}\cdot\mathbf{x}+b=\pm 1$ からどれだけ離れているかを示す項である．C はこれら二つの項の度合いを決めるパラメータである．

ここで，Lagrange 乗数 $\boldsymbol{\lambda}$，$\boldsymbol{\mu}$ を用いて Lagurange 関数 $L(\mathbf{w}, b, \boldsymbol{\lambda}, \boldsymbol{\mu})$ は以下の式で表わされる．

$$\frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^{l}\xi_i$$

[1] 最も負例よりの正例側の境界面と最も正例よりの負例側の境界面の間の距離をマージンと呼ぶ．

$$+\sum_{i=1}^{l}\lambda_i(1-\xi_i-y_i(\mathbf{w}\cdot\mathbf{x}_i+b))$$
$$-\sum_{i}^{l}\mu_i\xi_i \qquad (3)$$

したがって，以下の制約のもとに Lagurange 関数 ((3) 式) を最大にすればよい．

$$\begin{aligned}\frac{\partial L(\mathbf{w},b,\boldsymbol{\lambda},\boldsymbol{\mu})}{\partial \mathbf{w}} &= \mathbf{w}-\sum_{i=1}^{l}\lambda_i y_i\mathbf{x}_i=0\\ \frac{\partial L(\mathbf{w},b,\boldsymbol{\lambda},\boldsymbol{\mu})}{\partial b} &= -\sum_{i=1}^{l}\lambda_i y_i=0\\ \frac{\partial L(\mathbf{w},b,\boldsymbol{\lambda},\boldsymbol{\mu})}{\partial \xi_i} &= C-\lambda_i-\mu_i=0 \qquad (4)\end{aligned}$$

これらの関係を用いて，テスト事例 $\mathbf{x}$ を判別する判別関数 $f(\mathbf{x})$ として次の式が得られる．

$$\begin{aligned}f(\mathbf{x}) &= \mathrm{sgn}\left(\sum_i \lambda_i y_i(\mathbf{x}_i\cdot\mathbf{x})+b\right) \qquad (5)\\ &= \mathrm{sgn}\,(\mathbf{w}\cdot\mathbf{x}+b)\end{aligned}$$

判別関数は Support Vector によってのみ決定され，Support Vector 以外の事例は判別関数の決定に影響を与えない．

さらに，SVM の特徴の一つは判別関数を非線型に容易に拡張できる点にあり，非線型の分離平面を実現できる．これは式 (5) の内積を Kernel 関数で置き換えることで実現される．Kernel 関数を用いた場合の判別関数は，以下の式となる．

$$f(\mathbf{x})=\mathrm{sgn}\left(\sum_i \lambda_i y_i K(\mathbf{x}_i,\mathbf{x})+b\right) \qquad (6)$$

本稿では次式で定義される d 次の Polynomial 関数を Kernel 関数として用いた．

$$K(\mathbf{x},\mathbf{y})=(\mathbf{x}\cdot\mathbf{y}+1)^d \qquad (7)$$

2.2 SVM に基づく重要文抽出手法

重要文抽出とは，ある文書から重要な情報を持つ文を抽出することであり，文書中の各文に対して，重要 (正例)，非重要 (負例) のラベルを付与する 2 値分類問題であるといえる．すなわち，訓練データが与えられれば，前節で説明した SVM を用いて重要文，非重要文の特徴を学習し，未知データとして入力された文書中の各文を重要文，非重要文に分類することができる．ただし，文書中の何割の文が重要文として分類されるかは分からない．

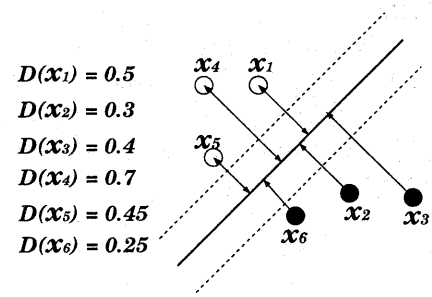

図 2: 分離平面からの距離に基づく評価

一方，重要文抽出タスクではあらかじめ要約率の形で重要文の全文数に対する比率が与えられるのが一般的である．そこで本稿では，分離平面から事例までの距離に基づき評価値を計算し，評価値の高い事例から順に重要文として採用する手法をとった (図 2)．分離平面と事例 $\mathbf{x}$ 間の距離に基づく評価値 $D(\mathbf{x})$ は以下の式で与える．

$$D(\mathbf{x})=\tanh\left(\sum_i \lambda_i y_i K(\mathbf{x}_i,\mathbf{x})+b\right) \qquad (8)$$

ただし，$\tanh(x)$ は以下の式で定義される．

$$\tanh(x)=\frac{1}{1+\exp(-x)} \qquad (9)$$

(9) 式は，分離平面と事例 $\mathbf{x}$ 間の距離をシグモイド関数に入力した形であり，分離平面からの距離を 0 から 1 までの値に正規化したものである [7].

2.3 素性

提案手法では任意の文 S_i に対して以下の特徴を素性として用いた．素性の数 ($\mathbf{x}_i$ の次元 n) は 407 である．

文の位置

文の位置に関する特徴として，文書中の位置 F_{loc_d}，段落中の位置 F_{loc_p}，抽出すべき文数 N に対して文 S_i が文書の先頭から N 字以内にあるか否か F_{loc_l} を用いた．これらをそれぞれ次式で定義した．

$$F_{loc_d}(S_i) = \frac{C_d}{C_{S_i}}$$

$$F_{loc_p}(S_i) = \frac{C_{p_j}}{C_{S_{ij}}}$$

$$F_{loc_l}(S_i) = \begin{cases} 1 & (S_i\text{が文書の先頭から} \\ & N\text{文以内にあるとき}) \\ 0 & (\text{otherwize}) \end{cases}$$

ただし，C_d は S_i が属する文書の文字数，C_{S_i} は S_i の開始位置，C_{p_j} は S_i が属する段落 p_j の文字数，$C_{S_{ij}}$ は S_i の段落中における開始位置である．

文の長さ

$$F_{len}(S_i) = \frac{C_i}{\max_{S_i \in D} C_i}$$

ただし，C_i は S_i の文字数，D は S_i が属する文書を表わす．

単語重要度に基づく文の重要度

まず，文書中の単語 t の重要度 $w(t)$ を範囲内重要度と出現頻度を用いて定める [12]．そして，S_i の重要度を $w(t)$ の加重和として次式で定義する．

$$F_w(S_i) = \sum_t tf(t, S_i) \cdot w(t)$$

キーワードの密度

上で求めた単語重要度を利用して，文書中で重要度の高い上位3割の単語をキーワード KW とする．ここで，文 S_i における KW の密度を幅 W のハニング窓を利用して求める．ハニング窓関数 $f_H(k,l)$ は，以下の式で定義される．ただし，l は窓 W の中心，k は文 S_i における窓の位置である．

$$f_H(k,l) = \begin{cases} \frac{1}{2}\left(1 + cos2\pi\frac{k-l}{W}\right) & (|k-l| \leq W/2) \\ 0 & (|k-l| > W/2) \end{cases}$$

ここで，$f_H(k,l)$ を用いて文 S_i の密度を以下の式で定義する．

$$F_d(S_i) = \sum_{k=l-W/2}^{l+W/2} f_H(k,l) \cdot a(k)$$

ただし，$a(k)$ は以下の値をとる．

$$a(k) = \begin{cases} w(t) & \text{単語 } t(\in KW) \text{ が位置 } k \\ & \text{を先頭として出現現する} \\ & \text{とき} \\ 0 & \text{otherwize} \end{cases}$$

固有表現

IREX[2] の固有表現基準による固有表現および数値表現 (計8種類) の有無．この8種類を $NE_1, \cdots, NE_n, \cdots, NE_8$ として以下の式で定義する．

$$F_{NE_n}(S_i) = \begin{cases} 1 & (NE_n\text{が } S_i\text{に出現するとき}) \\ 0 & (\text{otherwize}) \end{cases}$$

接続詞

文 S_i における接続詞 (計53種類) の有無．接続詞を $con_1, \cdots, con_n, \cdots, con_{53}$ として以下の式で定義する．

$$F_{con_n}(S_i) = \begin{cases} 1 & (con_n\text{が } S_i\text{に出現するとき}) \\ 0 & (\text{otherwize}) \end{cases}$$

機能語

文 S_i における機能語 (計173種類) の有無．機能語を $fun_1, \cdots, fun_n, \cdots, fun_{173}$ として以下の式で定義する．

$$F_{fun_n}(S_i) = \begin{cases} 1 & (fun_n\text{が } S_i\text{に出現するとき}) \\ 0 & (\text{otherwize}) \end{cases}$$

品詞分類

文 S_i における品詞 (計66種類) の有無．品詞を $pos_1, \cdots, pos_n, \cdots, pos_{66}$ として以下の式で定義する．

$$F_{pos_n}(S_i) = \begin{cases} 1 & (pos_n\text{が } S_i\text{に出現するとき}) \\ 0 & (\text{otherwize}) \end{cases}$$

[2] Information Retrieval and Extraction Exercise, http://cs.nyu.edu/cs/projects/proteus/irex/

表 1: TSC で提供されたデータセット

	A	B	C
文書 (記事) 数	30	30	120
全文数	716	994	2811
重要文数 (10%)	108	102	287
重要文数 (30%)	324	301	851
重要文数 (50%)	540	508	1437

意味カテゴリの深さ

文 S_i に出現する名詞 *noun* のシソーラスにおける階層の深さ．シソーラスとしては日本語語彙体系を用いた [11]．日本語語彙体系では，名詞は 11 階層のノードのいずれかに分類される．階層を $c_1, \cdots, c_n, \cdots, c_{11}$ として以下の式で定義する．

$$F_{c_n}(S_i) = \begin{cases} 1 & (S_i\text{に出現する } noun \text{ が} \\ & \text{階層 } c_n\text{に属するとき}) \\ 0 & (\text{otherwize}) \end{cases}$$

3 評価実験

3.1 コーパス

評価実験には，TSC で提供された重要文抽出データを利用した．このデータは，毎日新聞 94 年，95 年，98 年の報道，社説，解説などの全 180 記事から成る．各記事を構成する文の数に対して，10%，30%，50%の要約率に応じた重要文があらじめ与えられている[3]．Dryrun 時に 30 文書 (以下，セット A)，Formalrun 時に 30 文書 (以下，セット B) が公開され，Formalrun 後に 120 文書 (以下，セット C) が公開された．各セットの詳細を表 1 に示す．

3.2 評価方法

TSC の重要文抽出タスクでは各文書の各要約率に対して抽出すべき文数が提示されている．従って，その文数に応じた抽出をした場合，Precision，Recall，F-measure は同じ値をとる．本稿ではこの値を一致率 (Accuracy) として，抽出結果の評価を行なった．システムによって抽出された重要文の数を a，a に含まれる正解文の数を b とすると，一致率は以下の式で表わされる．

[3] たとえば，10 文からなる文書では，10%の要約率では 1 文，30%の要約率では 3 文，50%の要約率では 5 文が重要文として与えられる．

表 2: A を訓練に用いた場合の抽出精度

要約率	SVM	C4.5	Lead
10%	0.278 ($d=2$)	0.178	0.284
30%	0.430 ($d=1$)	0.350	0.432
50%	0.591 ($d=1$)	0.543	0.586

表 3: A+C を訓練に用いた場合の抽出精度

要約率	SVM	C4.5
10%	0.362 ($d=2$)	0.297
30%	0.475 ($d=1$)	0.387
50%	0.604 ($d=1$)	0.558

$$\text{Accuracy} = \frac{b}{a} \tag{10}$$

SVM のパラメータのうち，Polynomial 関数の次数 d は，1〜4，式 (2) の C は，0.01〜1 まで変化させ，最適値を決定した．SVM のプログラムは Tiny SVM を用いた．

また，最も一般的な重要文抽出手法である Lead 手法，決定木学習による手法を比較対象とした．決定木学習には C4.5([8]) を利用した．以降，この手法を C4.5 と略記する．C4.5 では提案手法と同じ素性を用いた．

3.3 結果と考察

3.3.1 抽出精度に関して

提案手法，Lead 手法，C4.5 の抽出精度を比較した．A の 30 文書，A+C の 150 文書をそれぞれ訓練セットとし，B の 30 文書を評価した結果を表 2，表 3 に示す．ただし，各要約率に対して与えられた重要文を正例，非重要文を負例として学習を行なった．

A を訓練データとして用いた場合 (表 2) には，Lead 手法と 提案手法がほぼ同等の精度であり，C4.5 の精度を上回った．提案手法や C4.5 などの機械学習手法を利用した手法は，訓練データを増すことによって，抽出精度の向上が期待できる．そこで，訓練データを A+C に増やし (文数としては約 5 倍)，同じテストデータ (B) を用いて評価を行なった (表 3)．この時，提案手法，C4.5 ともに抽出精度が向上し，提案手法では，Lead 手法よりも高い精度が得られた．一

表 4: 交差検定による抽出精度

要約率	SVM	C4.5	Lead
10%	0.416 ($d=1$)	0.374	0.372
30%	0.509 ($d=1$)	0.416	0.446
50%	0.648 ($d=1$)	0.519	0.588

表 5: 抽出精度の違い

要約率	Dryrun	Formalrun
10%	0.464	0.308
30%	0.546	0.421
50%	0.769	0.584

方，C4.5 は要約率30%，50%の双方で Lead 手法よりも精度が低い．

より信頼の高い評価を行なうため，全ての文書 (180 文書) を 5 等分し，5 回の交差検定を行なった．抽出精度の平均値を表 4 に示す．提案手法が最も高精度で次いで Lead 手法，C4.5 の順であった．3 手法間の性能比較をするため，統計的に有意な差があるかどうかを多重比較法である Tuky の方法を用いて検定を行なった．10%の要約率においては，3 手法間に有意差は認められなかったが，30%，50%の要約率においては，提案手法が有意水準 1%で他の 2 手法と比較して優れていることがわかった．なお，C4.5 と Lead 手法の間には有意差はなかった．

以上より，Lead 手法，決定木学習に基づく手法と比較して，提案手法が優れているといえる．

3.3.2 訓練データと評価データに関して

一般的に，機械学習手法を利用する場合には，より大量の訓練データを用いることにより精度を上げることが可能である．表 2 と表 3 を比較すると訓練に用いる文書数を増加させることで精度の向上を確認できる．しかし，文書数を 120 文書 (文としては 2811 文) 追加しても C4.5 による手法の 10%の要約率を除くとわずかな精度の向上しかみられない．この原因の一つとして，訓練データの数が十分でないという可能性がある．提案手法では，訓練データに対する抽出精度はテストデータよりもはるかに高く，訓練データを増やすことでテストデータに対する精度が向上する可能性がある．

提案手法を用いて C を訓練データとし，A，B を評価した結果を表 5 に示す．表 5 より，同じ訓練データを用いているにもかかわらず，A と B を評価した場合にその精度に大きな差がでていることがわかる．この差は，交差検定時の各セット間の値の揺れよりも大きく，しかも，B での精度は，表 2 と比較しても精度向上がみられない．これらのことから，C に出現する重要文と B に出現する重要文との間には特徴の相違があると考えられる．

4 まとめ

本稿では，Support Vector Machine (SVM) を用いた重要文抽出手法を提案した．TSC のコーパスを用いて提案手法と Lead 手法，決定木に基づく手法との比較評価を行なった．その結果，提案手法が統計的に有意な差で従来の手法より高精度であることを確認した．

謝辞

、研究を進めるにあたって，Tiny SVM のプログラムを提供していただくとともに有益なコメントをいただいた奈良先端科学技術大学院大学の工藤拓氏に感謝いたします．

参考文献

[1] Aone, C., Okurowski, M. and Gorlinsky, J.: Trainable Scalable Summarization Using Robust NLP and Machine Learining, *In Proc. of the 17th COLING and 36th ACL*, pp. 62–66 (1998).

[2] Edmundson, H.: New methods in automatic abstracting, *Journal of ACM*, Vol. 16, No. 2, pp. 246–285 (1969).

[3] Joachims, T.: Text Categorizatin with Supprt Vector Machines: Learing with Many Relevant Features, *In Proc. of European Conference on Machine Learing* (1998).

[4] Kudo, T. and Matsumoto, Y.: Chunking with Support Vector Machine, *In Proc.*

of the 2nd Meeting of the NAACL, pp. 192–199 (2001).

[5] Mani, I. and Bloedorn, E.: Machine Learnig of General and User-Focused Summarization, *In Proc. of the 15th National Conference on Artifical Intteligence*, pp. 821–826 (1998).

[6] Nobata, C. et al.: Sentence Extraction System Assembling Multiple Evidence, *In Proc. of the 2nd NTCIR Workshop Meeting*, pp. 213–218 (2001).

[7] Platt, J.: *Probabilistics Outputs for Support Vector Machine and Compairsons to Regularized Likelihood Methods. In Advances in Large Margin Classifires*, MIT Press (2000).

[8] Quinlan, J.: *C4.5: Programs for Machine Learning*, Morgan Kaufman (1993).

[9] Vapnik, V.: *The Nature of Statistical Learning Theory*, New York (1995).

[10] 奥村学, 原口良胤, 望月源: 決定木学習を用いたテキスト自動要約に関するいくつかの考察, 情報処理学会第 59 回全国大会講演論文集 (分冊 5), pp. 393–394. 5N–2 (1999).

[11] NTT コミュニケーション科学基礎研究所監修: 日本語語彙体系, 岩波書店 (1999).

[12] 原正巳, 中島浩之, 木谷強: テキストのフォーマットと単語の範囲内重要度を利用したキーワード抽出, 情報処理学会論文誌, Vol. 38, No. 2 (1997).

[13] 工藤拓, 松本裕治: Support Vector Machine による日本語係り受け解析, 情報処理学会研究報告 NL-138, pp. 79–86 (2000).

[14] 野本忠司, 松本裕治: 人間の重要度判定に基づいた自動要約の試み, 情報処理学会研究報告 NL-120-11, pp. 71–76 (1997).

[15] 平博順, 春野雅彦: Support Vector Machine によるテキスト分類における属性選択, 情報処理学会論文誌, Vol. 39, No. 4, pp. 1113–1123 (2000).