Support Vector Machine を用いたテキスト重要個所抽出と要約文生成への応用

鈴木 大介　内海 彰
電気通信大学 電気通信学部 システム工学科
dajie@utm.se.uec.ac.jp，utsumi@se.uec.ac.jp

1　はじめに

計算機を用いて文書を自動的に要約する自動要約の研究は，電子化された文書が溢れている現在，その必要性は高い．

要約生成に機械学習を適用する研究も行われ始めている．Support Vector Machine（以下，SVM）はパターン認識分野で注目されている学習手法であり，様々な問題に適用されている．要約文生成のための重要文抽出の研究においても優れた結果が報告されている [1]．しかし，これら多くの自動要約の研究は，文単位で重要な個所を抽出する手法を用いているが，適切な要約文生成のためには，文単位の抽出では不十分である．

そこで，本研究では SVM を用いて文節単位でのテキスト重要箇所抽出を行うことを目的とする．さらに，抽出結果と原文の係り受け構造に着目して，要約文を生成する手法を提案し，その有効性を検証する．

2　SVMを用いたテキストからの重要個所抽出

2.1　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．概念図を図 1 に示す．学習データは (1) 式のベクトルとして表すことができる．

$$(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_u, y_u),\ \ \mathbf{x}_j \in \mathbf{R}^n, y_j \in \{+1, -1\} \qquad (1)$$

$\mathbf{x}_j$ を各事例の特徴ベクトル，y_j を事例 j が正例であるときに $+1$，負例であるときに -1 となる教師信号とする．SVM は正例・負例間の距離 (マージン) が最大となるような分離平面を決定する．学習事例が線形分離不可能な場合にはスラック変数 ξ_j を導入する．このとき以下のように定式化され，(3) 式の判別関数を得る．

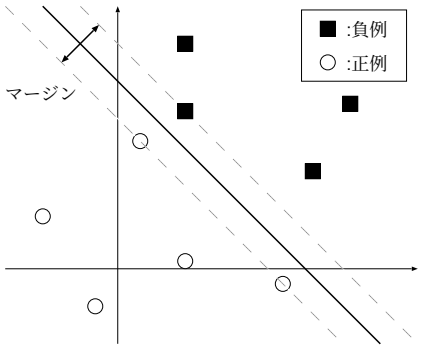

図 1: SVM の概念図

$$\min_{\mathbf{w},b,\xi} \ \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{j=1}^{u}\xi_j \qquad (2)$$
$$y_j(\mathbf{w}\cdot\mathbf{x}_j + b) \geq 1 - \xi_j$$

$$g(\mathbf{x}) = \sum_{j=1}^{u}\alpha_j y_j(\mathbf{x}_j\cdot\mathbf{x}) + b \qquad (3)$$

ここで，C は制約条件をどこまで緩めるかを指定するパラメータ，α_j は Lagrange 乗数である．(3) 式の符号を用いてテストデータを分類できる．

また，SVM は (4) 式のように内積を Kernel 関数で置き換えることで非線形の分離平面を実現できる特徴がある．本研究では，学習結果の解釈が比較的容易であるため，Kernel 関数として (5) 式に示す Polynomial 関数を用いる．

$$g(\mathbf{x}) = \sum_{j=1}^{u}\alpha_j y_j K(\mathbf{x}_j\cdot\mathbf{x}) + b \qquad (4)$$

$$K(\mathbf{x},\mathbf{y}) = (\mathbf{x}\cdot\mathbf{y}+1)^d \qquad (5)$$

2.2　SVM を用いたテキストからの重要個所抽出

入力文書を文節単位に分けて特徴ベクトルを構成し，学習・テストを行う．ただし，テストにおいて (3) 式によって抽出される文節数の割合が要約率に一致するとは限らない．そこで，学習された分離平面との距離に相当する $g(\mathbf{x})$ の値の大きい順に要約率を満たすまで文節を抽出する．

2.3　素性

各文に形態素解析・係り受け解析・固有表現抽出の各処理を行い，表 1 に示した素性に基づき各文節の特徴ベクトルを構成する．素性は，文節 P を含む文 S の性質を表現したもの（表 1 の上側）と，文節 P の性質を表現したもの（表 1 の下側）に大別される．なお，形態

表 1: 特徴ベクトルに用いた素性

文のタイトル類似度	名詞出現頻度ベクトルの余弦
文の TF・IDF 値	(6) 式
文の文書内位置	文書の先頭から文 S までの文字数
文の段落内位置	段落の先頭から文 S までの文字数
文の長さ	文 S の文字数
タイトル類似度	名詞がタイトル中に出現するかどうか
TF・IDF 値	(7) 式
文内位置	文の先頭から文節 P までの文字数
形態素（大分類）	名詞，動詞，形容詞等の計 14 種類
形態素（小分類）	普通名詞，格助詞，読点等の計 62 種類
係り方（大分類）	同格，並列等の 5 種類
係り方（小分類）	連体，連用，文末等の計 24 種類
係る文節数	直接係る文節の数．0〜5 の 6 種類
固有表現	組織名，人名等の 7 種類
用言の意味属性	日本語語彙大系用言意味属性による 36 分類

素解析には JUMAN [1]，係り受け解析には KNP [2]，固有表現抽出には NExT [3] をそれぞれ用いる．

実数を取る素性については，それぞれに $[0.0,0.1)$, $[0.1,0.2)$, $\cdots$, $[0.9,1.0]$ の 10 次元を割り当て，文書内での最大値で正規化した値に該当する次元を 1，それ以外を 0 として表現する．種類を表す素性については，それぞれに次元を割り当て，条件に合う場合に 1，そうでない場合に 0 とする．これらにより，各文節は各要素が 0 または 1 をとる 225 次元ベクトルとして表現する．

文 S のタイトル類似度は，文 S における名詞の出現頻度ベクトル $\mathbf{v}(S)$ と，文 S を含む文書のタイトル T における名詞の出現頻度ベクトル $\mathbf{v}(T)$ との余弦で定義する．また，文 S の TF・IDF 値，文節 P の TF・IDF 値は次式で定義する．

$$\mathrm{TI}_{\mathrm{sent}} = \sum_{t \in S} tf(t,S) \cdot 0.5\left(1 + \frac{tf(t,D)}{tf_{\max}(D)}\right) \cdot \log_2\left(\frac{N}{df(t)}\right) \quad (6)$$

$$\mathrm{TI}_{\mathrm{part}} = \sum_{t \in P} tf(t,P) \cdot tf(t,D) \cdot \log_2\left(\frac{N}{df(t)} + 1\right) \quad (7)$$

ここで，$tf(t,S)$，$tf(t,D)$，$tf(t,P)$ はそれぞれ文 S，文書 D，文節 P 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 は文書集合数，$df(t)$ は名詞 t を含む文書数である．

3 係り受け木からの要約文生成

3.1 係り受け木

日本語文は係り受け関係を表した木構造（係り受け木）で表すことができる．

[1]http://www.kc.t.u-tokyo.ac.jp/nl-resource/juman.html
[2]http://www.kc.t.u-tokyo.ac.jp/nl-resource/knp.html
[3]http://www.ai.info.mie-u.ac.jp/~next/next.html

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，日本語として文法上正しい文となる．また，根ノードを含まない場合にも，部分木は意味的なまとまりを持った単位になると考えられる．

3.2 要約文生成

SVM による抽出は文節単位であるため，抽出された文節を並べただけでは文法上適切になるとは限らない．

そこで，係り受け木の性質に着目し，以下のアルゴリズムで要約文を生成する．入力文書 D は係り受け解析の結果に基づき，文ごとに木構造として表現されているとする．

1. 重要文節集合 I_0 を SVM によって抽出された文節の集合，要約候補文節集合 $I_c = \phi$，要約文節集合 $I_s = \phi$ とする．
2. すべての文節 P 対し，P 自身を含む先祖に I_0 の要素を含み，かつ，P 自身を含む子孫に I_0 の要素を含むならば，I_c に加える．
3. I_c の要素からなるすべての部分木 T に対し，T が含む要素が 1 つならば，その要素を I_c から除く．
4. I_c の要素からなるすべての部分木 T に対し，要約率を満たすまで，$\frac{T \text{ に含まれる } I_0 \text{の要素数}}{T \text{ に含まれる } I_c \text{の要素数}}$ の値の大きい順に，それらの要素を I_s に加える．
5. I_s の各要素を原文で出現する順に要約文とする．

抽出例を図 2 に示す．網掛部は I_0 の要素，破線で囲まれた部分は I_c の要素を表す．文 (b) の例のように，SVM によって抽出された要素であっても，その部分木が含む要素が 1 つならば I_c の要素から除かれる．破線で囲ま

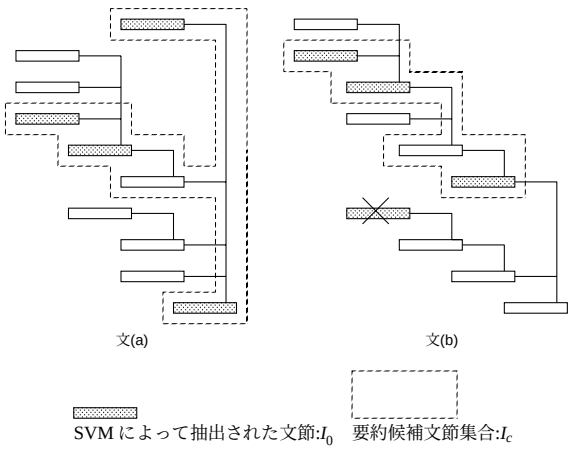

図 2: 要約文生成アルゴリズムによる抽出例

表 2: 5 分割交差検定の結果（F 値）

	$d=1$		$d=2$	
	SVM	SVM+係り受け	SVM	SVM+係り受け
報道 20%	0.385	0.391	0.430	0.433
報道 40%	0.488	0.493	0.491	0.502
社説 20%	0.243	0.257	0.288	0.296
社説 40%	0.480	0.486	0.483	0.486

れたそれぞれの部分木で，SVM によって抽出された文節の割合の大きい順に，それらの要素を I_s に加える．

4 評価実験と考察

4.1 コーパス

国立情報学研究所主催の評価型ワークショップ NTCIR2 の自動要約タスク TSC(Text Summarization Challenge) で作成された重要個所抽出データを基に，文節単位での正解データを構成した．重要個所抽出データは，毎日新聞 94，95，98 年の 180 記事からなり，20%・40%の要約率が設定されている．これらのうち，報道 69 記事・社説 43 記事を用いて以下の実験を行った．

なお，(6)，(7) 式の計算には，上記の 180 記事を文書集合として用いた．

4.2 実験方法と結果

報道記事，社説記事ごとに，実験データを 5 分割し，4 セットで学習，1 セットでテストを行う実験を繰り返す交差検定を行った．

表 2 に 5 回の実験を繰り返したときの結果を示す．

表中の値は以下に示す F 値である．

$$F \text{ 値} = \frac{2PR}{P+R}$$

$$\text{適合率 } P = \frac{\text{システム正解数}}{\text{システム出力数}}$$

$$\text{再現率 } R = \frac{\text{システム正解数}}{\text{正解数}}$$

また,「SVM」は SVM による抽出結果を,「SVM+係り受け」は要約文生成アルゴリズムによる抽出結果をそれぞれ表す．コストパラメタについては，予備実験から良好な結果を得た $C=0.1(d=1)$，$C=0.0005(d=2)$ とした．なお，SVM のプログラムは TinySVM [4] を用いた．

表 5 に要約例を示す．

4.3 有効な素性の分析

$d=2$ の場合，$\mathbf{x}=(x[1],\cdots,x[n])$ の各次元が二値ベクトルであることに注意すると，$g(\mathbf{x})$ 以下のように展開できる．

$$\begin{aligned} g(\mathbf{x}) &= b+\sum_{i=1}^{u} w_i + 2\sum_{i=1}^{u} w_i \sum_{k=1}^{n} x_i[k]x[k] \\ &\quad + \sum_{i=1}^{u} w_i \sum_{h=1}^{n}\sum_{k=1}^{n} x_i[h]x_i[k]x[h]x[k] \\ &= W_0 + \sum_{k=1}^{n} W_1[k]x[k] + \sum_{h=1}^{n-1}\sum_{k=h+1}^{n} W_2[k,h]x[h]x[k] \end{aligned} \tag{8}$$

ここで，$w_i=\alpha_i y_i$，$W_0=b+\sum_{i=1}^{u} w_i$，$W_1[k]=3\sum_{i=1}^{u} w_i x[k]$，$W_2[h,k]=2\sum_{i=1}^{u} w_i x_i[h]x_i[k]$ である．これは $W_1[k]$ は単独の素性 ($x[k]$)，$W_2[h,k]$ は組み合わせの素性 ($x[h]x[k]$) に対するスコア関数として解釈できる．すなわち，この絶対値が大きいほど分類に有効な素性であると考えられる．そこで，絶対値の大きい順に 5 位までを表 3，4 に示す．

表 3: 分類に有効な素性 (正の重み)

報道 20%	文書内位置 [0.0,0.1)，文節のタイトル類似度 1 文書内位置 [0.0,0.1)，形態素 (小分類)：サ変名詞 文のタイトル類似度 [0.7,0.8) 文書内位置 [0.0,0.1) 文のタイトル類似度 [0.4,0.5)，文書内位置 [0.0,0.1)
報道 40%	文書内位置 [0.0,0.1) 文のタイトル類似度 [0.4,0.5) 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)，形態素 (大分類)：名詞 文書内位置 [0.0,0.1)，形態素 (大分類)：助詞
社説 20%	文の TF・IDF 値 [0.2,0.3)，文書内位置 [0.0,0.1) 文書内位置 [0.0,0.1)，文の長さ [0.7,0.8) 文の TF・IDF 値 [0.3,0.4)，文書内位置 [0.9,1.0] 文書内位置 [0.9,1.0]，文の長さ [0.3,0.4) 文のタイトル類似度 [0.2,0.3)，文書内位置 [0.9,1.0]
社説 40%	文書内位置 [0.9,1.0] 文書内位置 [0.9,1.0]，形態素 (大分類)：助詞 文書内位置 [0.9,1.0]，係り方 (大分類)：一般 文書内位置 [0.9,1.0]，形態素 (大分類)：名詞 文の TF・IDF 値 [0.3,0.4)，文書内位置 [0.9,1.0]

4.4 考察

表 2 から，$d=1$ の場合に比べて $d=2$ の場合の方が良い結果が得られた．特に 20%要約の場合に結果が向上しており，要約率が高い場合（20%要約）には組み合わせ素性が重要であると言える．

要約文生成アルゴリズムを用いた方が結果が向上しており，学習結果を適切に反映したものだと言える．また，表 5 の要約例を見ると，特に「刑事訴訟法は〜」で始まる文では，アルゴリズムが有効に働いていることが分かる．

表 3 からは，社説記事においては，報道記事とは逆に位置が末尾であることに高い重みが与えられているなど，ジャンルに応じて学習していることが分かる．

[4] http://cl.aist-nara.ac.jp/~taku-ku/software/TinySVM/

表 4: 分類に有効な素性 (負の重み)

報道 20%	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)，文の長さ [0.5,0.6) 文書内位置 [0.0,0.1)，文節の TF・IDF 値 [0.1,0.2) 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 [0.5,0.6)，文の長さ [0.9,1.0]
報道 40%	文のタイトル類似度 [0.0,0.1) 文のタイトル類似度 [0.0,0.1)，文書内位置 [0.0,0.1) 文のタイトル類似度 [0.1,0.2) 文のタイトル類似度 [0.3,0.4)，文の長さ [0.6,0.7) 文のタイトル類似度 [0.1,0.2)，文の長さ [0.4,0.5)
社説 20%	文のタイトル類似度 [0.1,0.2)，文書内位置 [0.9,1.0) 文のタイトル類似度 [0.0,0.1)，文書内位置 [0.9,1.0) 文のタイトル類似度 [0.2,0.3)，文の長さ [0.4,0.5) 文のタイトル類似度 [0.3,0.4)，文の TF・IDF 値 [0.6,0.7) 文のタイトル類似度 [0.3,0.4)，文の長さ [0.5,0.6)
社説 40%	文のタイトル類似度 [0.0,0.1)，文の長さ [0.5,0.6) 文のタイトル類似度 [0.0,0.1) 文の TF・IDF 値 [0.5,0.6)，文の長さ [0.4,0.5) 文書内位置 [0.5,0.6)，文の長さ [0.7,0.8) 文のタイトル類似度 [0.0,0.1)，文の長さ [0.3,0.4)

なお，厳密には比較できないが，同様のコーパスによる遺伝的アルゴリズムを用いた重要文節抽出の研究 [2] では，20%要約において F 値が 0.388（報道）・0.238（社説），40%要約において 0.565（報道）・0.520（社説）という結果を得ている．20%要約においては，本研究の方が良い結果を得られた．

5 おわりに

本研究では，SVM を用いて文節単位でのテキスト重要個所抽出を行い，要約率が高い場合には組み合わせ素性が重要であることや，係り受け構造に着目して文節を抽出することが効果的であることなどが分かった．

しかし，本研究で学習に用いた素性の多くは表層的なものであり，より意味的な情報を扱う等，素性を見直す必要があると考えられる．

参考文献

[1] 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).

[2] 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).

表 5: 要約例（報道，要約率 40%）

原文

タイトル：店主強殺事件で予備的訴因認め殺害場所を特定せず「無期懲役」－－大津地裁判決　【大阪】

滋賀県日野町で十一年前、酒店経営の女性が殺され店内の手提げ金庫が奪われた事件で、強盗殺人罪に問われた同町中山、元工員、阪原弘被告（６０）に対する判決が三十日、大津地裁であった。中川隆司裁判長は殺害場所を特定せずに、求刑通り無期懲役を言い渡した。自供以外に物証がないため、大津地検が求刑前に追加した、殺害の場所などをぼかす内容の予備的訴因を認めた判決で、殺人などの重大犯罪では極めて異例。被告側は「事実認定の根拠があいまい」として即日控訴した。

当初の起訴事実は、一九八四年十二月二十八日午後八時四十分ごろ、日野町豊田の酒店内で池元はつさん（当時六十九歳）を絞殺、現金五万円と手提げ金庫を奪った、とされた。遺体は同町内の造成地で発見された。

しかし、被告は公判で、捜査員の暴行などで自供した－－と無罪を主張。犯行の日時や場所、被害品を示す物証がなく、地検は今年二月、予備的訴因を追加。殺害場所と、被害品をあいまいなかたちにした。

判決では、自白の信用性を否定したが、店内に残された被告の指紋などから「被告の犯行と認められる」とした。

刑事訴訟法は、公訴事実でできる限り日時、場所などを特定するよう定めているが、最高裁は六二年、中国への密出国の時期の特定をめぐって争われた白山丸事件で、特別な事情がある場合は日時や場所を具体的に示さなくてもよいとの判断を示している。

板倉宏・日大法学部教授（刑法）の話　殺人事件の裁判でここまで広く殺害現場を認定した例は聞いたことがない。今回の判決は、被告の防御をより困難にする恐れがある。

正解

滋賀県日野町で十一年前、酒店経営の女性が殺され店内の手提げ金庫が奪われた事件で、強盗殺人罪に問われた阪原弘被告に対する判決が三十日、大津地裁であった。中川隆司裁判長は殺害場所を特定せずに、求刑通り無期懲役を言い渡した。殺人などの重大犯罪では極めて異例。被告側は「事実認定の根拠があいまい」として即日控訴した。犯行の日時や場所、被害品を示す物証がなく、地検は今年 二月、予備的訴因を追加。殺害場所と、被害品をあいまいなかたちにした。刑事訴訟法は、公訴事実でできる限り日時、場所などを特定するよう定めているが、

SVM による抽出結果

滋賀県日野町で十一年前、酒店経営の女性が殺され店内の手提げ金庫が奪われた事件で、強盗殺人罪に問われた同町中山、元工員、阪原弘被告に判決が大津地裁であった。中川隆司裁判長は殺害場所を特定せずに、求刑通り無期懲役を言い渡した。自供以外にない大津地検が追加した、殺害の場所などを内容の予備的訴因を判決で、殺人などの重大犯罪では極めて異例。控訴した。発見された。追加。殺害場所と、した。した。刑事訴訟法は、できる日時、場所などを特定する定めているが、最高裁は特定を白山丸事件で、ある場合は場所を判断を示している。ある。

要約文生成アルゴリズムによる抽出結果

滋賀県日野町で十一年前、酒店経営の女性が殺され店内の手提げ金庫が奪われた事件で、強盗殺人罪に問われた同町中山、元工員、阪原弘被告に対する判決が大津地裁であった。中川隆司裁判長は殺害場所を特定せずに、求刑通り無期懲役を言い渡した。自供以外にないため、大津地検が追加した、殺害の場所などをぼかす内容の予備的訴因を認めた判決で、殺人などの重大犯罪では極めて異例。殺害場所と、した。刑事訴訟法は、できる限り日時、場所などを特定するよう定めているが、最高裁は特定をめぐって争われた白山丸事件で、ある場合は場所を示さなくてもよいとの判断を示している。