LE-11

機械学習を用いた複数文書要約

Multi-Document Summarization using Machine Learning

平尾 努† 賀沢秀人† 磯崎 秀樹† 前田 英作† 松本 裕治††
Tsutomu Hirao† Hideto Kazawa† Hideki Isozaki† Eisaku Maeda† Yuji Matsumoto††

1 はじめに

複数文書要約は，ある話題に関連する複数の文書をまとめて要約する技術であり，Document Understanding Conference (DUC)[1] や，Text Summarization Challenge (TSC)[2] [2] といった評価型ワークショップにおいてもタスクとして設定されるなど，注目を集めている．

複数文書要約をその一例とする自動要約には，文書中から重要な情報を持つ文を抽出する重要文抽出技術を用いて，抽出された重要文の集合を要約とする手法 [4] や，その出力から不要な表現の削除や置換，あるいは，新たな表現の挿入を行い，より自然な要約にする手法がある [5]．本稿では，前者に着目し，要約を重要文抽出と捉える．

複数文書からの重要文抽出も，単一文書からの重要文抽出と同様に，ある手がかりに基いて文の重要度を決定し，重要度の高い文から順に，要約率で指定された文数までを重要文として抽出する．この際，複数の手がかりを扱うことが効果的であるが，手がかりの数が多くなると，人手によって適切な重みを見つけることが難しいという問題がある．本稿では，汎化能力が高いとされる機械学習手法の一種である Support Vector Machine (SVM) を用いて，複数の手がかりを効率的に扱い，複数文書から重要文を抽出する手法を提案する．

2 SVM に基づく複数文書からの重要文抽出手法

2.1 SVM による文のランキング

SVM は，Vapnik によって提案された 2 値分類のための教師あり学習アルゴリズムである [6]．

SVM では，学習データを $\mathbf{x}_i (1 \leq i \leq n)$ としたときに，テストデータ $\mathbf{x}$ を判別する判別関数 $f(\mathbf{x}) = \mathrm{sgn}(g(\mathbf{x}))$ が以下の式で与えられる．

$$g(\mathbf{x}) = \sum_{i=1}^{n} w_i K(\mathbf{x}_i, \mathbf{x}) + b \qquad (1)$$

w_i, b は定数である．ここで，$w_i \neq 0$ となるベクトルはサポートベクトルと呼ばれ，学習データ中の正例，負例を代表する．結局，判別関数はサポートベクトルのみで記述される．$K(\mathbf{x}_i, \mathbf{x})$ はカーネル関数と呼ばれる．様々なカーネル関数が提案されているが，本稿では多項式カーネル (式 2) を用いる．

$$K(\mathbf{x}, \mathbf{y}) = (\mathbf{x} \cdot \mathbf{y} + 1)^d \qquad (2)$$

重要文抽出は，要約率で指定された重要文の数を Num とした時に，重要度の高い上位 Num 件の文を重要文とみなし，それ以外を非重要文と見なす 2 値分類問題と考えることができる．ただし，重要文抽出では，指定された要約率に応じた特定数の文を抽出しなければならないという問題がある．判別関数 $f(\mathbf{x})$ の正負によって，重要文と非重要文の判別を行うと，重要文と判別された文の数が要約率で指定された文の数と一致するとは限らない．そこで提案手法では，入力となる複数文書集合に含まれる全ての文に対して $g(\mathbf{x})$ の値を用いてランキングを行い，指定された要約率をみたすように文を抽出する．

表 1: 用いた素性

単一文書用	複数文書用
位置 (文書，段落)	位置 (文書集合)
$TF \cdot IDF$(3 種)	$TF \cdot IDF$(2 種)
長さ	文書ジャンル
キーワード密度	
タイトルとの類似度	
固有表現の種類	
接続詞の種類	
助詞の種類	
文末表現の小・大分類種類	
修辞関係の種類	
用言の分類種類	

表 2: 被験者間の重要文の一致

組み合わせ	一致率	K 値
A ∩ B	0.465	0.40
B ∩ C	0.517	0.46
C ∩ A	0.451	0.39
A ∩ B ∩ C	0.328	0.42

2.2 素性

複数文書からの重要文抽出は，話題に関する文書集合を連結して 1 文書とみなせば，従来の単一文書からの重要文抽出と同等である．しかし，文書集合中の任意の文が，それが属する文書において重要かどうかという観点だけでなく，文書集合全体において重要かどうかという観点も扱う必要がある．よって本稿では，1 文書のみで決定することのできる素性 (単一文書用素性) と文書集合が与えられたときに決定することのできる素性 (複数文書用素性) の 2 種の素性を用いる．表 1 に用いた素性を示す．

3 評価実験

3.1 コーパス

評価実験のために毎日新聞 99 年から 12 個の話題に関連する文書集合を記者経験のある人物 1 名が作成した．次に，それぞれの文書集合の総文数に対して 10%の要約率 (文単位) を設定し，重要文抽出による要約の評価用データを人手によって作成した．重要文抽出データの作成には，新聞記事の編集などに深くかかわったことのある 6 名があたり，1 つの話題に対して異なる 3 名によるデータを作成した．それぞれをセット A，セット B，セット C とよぶ．評価データの総文書数は 231 文書，総文数は 4013 文，総重要文数は 409 文である．なお，本稿での複数文書は野畑らの分類 [7] によると「Single–Event」の分類となる．

被験者が抽出した重要文間の一致率，被験者間で共通な重要

† 日本電信電話株式会社 NTT コミュニケーション科学基礎研究所
†† 奈良先端科学技術大学院大学 情報科学研究科

[1] http://www-nlpir.nist.gov/projects/duc
[2] http://lr-www.pi.titech.ac.jp/tsc

表 3: 各手法の性能評価 (Precision)

手法	セットA	セットB	セットC
Lead	39.2	38.4	45.7
$TF \cdot IDF$	39.7	36.9	37.6
SVM	**51.1**	**46.6**	**50.7**

表 4: 共通重要文に対する抽出精度

手法	A ∩ B	B ∩ C	C ∩ A	A ∩ B ∩ C
Lead	63.9	61.5	75.3	78.2
$TF \cdot IDF$	57.6	55.9	62.0	66.9
SVM(A)	**73.8**	69.5	82.0	85.1
SVM(B)	73.1	69.4	80.7	**86.7**
SVM(C)	73.2	**72.0**	**83.0**	85.5

文の割合を調べ，次に，被験者間の重要文の一致を Kappa 統計値 (K 値) で調べた．詳細を表 2 に示す．また，被験者間で一致した重要文を共通重要文とよぶ．このセットにおける被験者間での重要文の一致に対する信頼性を K で解釈すると，「A ∩ B」，「C ∩ A」が FAIR，「B ∩ C」，「A ∩ B ∩ C」が MODERATE となる [1]．これは，過去の研究におけるデータの信頼性 [8] よりも高い．

3.2 実験結果と考察

提案手法の有効性を検証するために，前節で述べたデータセットを用いて，Lead 手法，$TF \cdot IDF$ 手法との性能比較を行った．Lead 手法では，各文書の先頭から順に要約率を満すまでを重要文として抽出した．$TF \cdot IDF$ 手法では，MDL 原理に基づき文書集合に特徴的な単語を認定し，それらを用いて $TF \cdot IDF$ により文のスコアを決定し，スコアの高い文から順に要約率を満すまでを重要文として抽出した[3]．提案手法では，2 次の多項式カーネルを用い，コストパラメータを $C = 0.001$ とした．$g(\mathbf{x})$ の値が大きい文から順に要約率をみたすまでを重要文として抽出した．プログラムには TinySVM[4] を利用した．

なお，評価指標は TSC の重要文抽出タスクに従った．文書集合に対して要約率によって抽出すべき文の数を設定し，各手法がその数だけ抽出した重要文に含まれる正解重要文の数の割合 (Precision) で評価を行った．

3.2.1 各手法の性能比較

表 3 にそれぞれの手法の各要約率における重要文の抽出精度を示す．なお，SVM の評価値は A〜C の各セットを学習用 11 話題とテスト用 1 話題に分けて評価を行い，これを 12 回繰り返した結果の平均値である．

表 3 より，どのセットにおいても提案手法の抽出精度が最も高く，続いて Lead，$TF \cdot IDF$ の順となる．提案手法はどのセットに対しても安定して他の 2 手法よりも成績が良い．Lead と $TF \cdot IDF$ の抽出精度の差はわずかであるが，セット C においては，抽出精度の差は大きい．これは，セット C が Lead 文を多く含んでいることを示す．

一般的に，報道記事では Lead 手法が有効である．評価データは，ある話題に関連する記事であるため，報道記事が占める割合が多い．それにもかかわらず，提案手法は，Lead 手法よりも十分に抽出精度が高い．また，利用したデータや素性が異なるため，正確な比較とは言えないが，Lead と SVM の差は単一文書 (TSC の報道記事) の場合 [3] と比較して大きくなっている．

また，共通重要文に対して，各手法の抽出精度を調べた．表 4 にその結果を示す．なお，SVM(A)，SVM(B)，SVM(C) はそれぞれ A，B，C による正解を学習に用いた結果を表す．各表より，どの正解セットの組み合わせに対しても共通重要文に対する抽出精度は SVM が最も高く他の 2 手法との差も大きい．また，セット A，C を学習に用いた場合が，セット B を用いた場合よりもやや成績が良いことがわかる．今回のデータセットの中で K 値の高い組み合わせ，「B ∩ C」，「A ∩ B ∩ C」においても SVM の抽出精度は高く，他の手法との差も大きい．

以上より，提案手法は Lead 手法，$TF \cdot IDF$ と比較して成績が良いことがわかった．さらに，共通重要文に対しても，提案手法は高い精度で抽出できることがわかった．

表 5: 単一文書用素性のみを用いた抽出精度

手法	セットA	セットB	セットC
SVM	48.5	46.3	50.0

3.2.2 複数文書用の素性の有効性

表 1 に示した素性のうち単一文書用の素性のみを用いた場合の抽出精度を表 5 に示す．

表 3 と表 5 を比較すると，複数文書用の素性を除くことで成績が低下しており，複数文書用素性が有効に働いていることがわかる．ただし，改善幅はそれほど大きくない．これは，対象とした文書集合に含まれる文書がすべて話題を主題としている文書であったということが考えられる．このような場合，単一文書での重要文がそのまま複数文書での重要文になることが多くなり，複数の文書であることを特に考慮することの効果が小さい．

4 まとめ

本稿では，機械学習手法の 1 種である Support Vector Machine を用いた複数文書要約手法について述べた．評価実験によって，Lead 手法，$TF \cdot IDF$ 手法との比較を行い，提案手法が最も抽出精度が高いことを実証した．さらに，複数の文書を考慮した素性を用いることで重要文の抽出精度が向上することを示した．

謝辞

評価法について有益なコメントをいただいた通信総合研究所の竹内和広氏に感謝いたします．また，データの使用を許諾してくださった毎日新聞社に感謝いたします．

参考文献

[1] Carletta, J., Isard, A., Isard, J., Kowtko, J., Doherty-Sneddon, G. and Anderson, A.: The Reliability of A Dialogue Structure Coding Scheme, Vol. 23, No. 1, pp. 13–31 (1997).

[2] Fukushima, T. and Okumura, M.: Text Summarization Challenge Text summarization evaluation in Japan, *Proc. of the NAACL2001 Workshop on Automatic summarization*, pp. 51–59 (2001).

[3] Hirao, T., Isozaki, H. and Maeda, E.and Matsutomo, Y.: Extracting Important Sentences with Support Vector Machines, *Proc. of the 19th International Conference on Computational Linguistics* (2002).

[4] Luhn, H.: The Automatic Creation of Literature Abstracts., *IBM Journal of Research and Development*, Vol. 2, No. 2, pp. 159–165 (1958).

[5] Nanba, H. and Okumura, M.: Producing More Readable Extracts by Revising Them, *Proc. of the 18th International Conference on Computational Linguistics*, pp. 1071–1075 (2000).

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

[7] 野畑周, 関根聡: 複数記事に対する要約や情報抽出に関する一考察, 言語処理学会第 8 回年次大会発表論文集, pp. 547–550 (2002).

[8] 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).

[3] この $TF \cdot IDF$ 手法は複数文書用素性の一つとして用いている．

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