社団法人　情報処理学会　研究報告
IPSJ SIG Technical Report

2003－ＮＬ－158　（６）
2003／11／6

言語パターンに着目した複数文書要約

平尾 努[†]　　鈴木 潤[†]　　磯崎 秀樹[†]　　前田 英作[†]
[†] 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
〒619-0237 京都府相楽郡精華町光台 2-4
{hirao,jun,isozaki,maeda}@cslab.kecl.ntt.co.jp

概要

複数文書要約の実現のためには，文書集合から重要文を精度良く抽出することと共に複数の文書に共通して出現する冗長な表現を排除することが必要である．これまでの要約手法は，重要文抽出や冗長表現の排除のための特徴として単語や n-gram を用いている．しかし，文が木構造で表現できることを考えると連続した単語の並びだけでなく，スキップを許した単語の並びに着目することがこうしたタスクに有効であると推測される．そこで本稿では，(1) スキップを許した単語列を重要文抽出の特徴として用いること，(2) 拡張ストリング・カーネル (ESK) を利用した MMR(Maximal Meridien Relevance) を用いて冗長文の削除を行うことを特徴とする要約手法を新たに提案する．毎日，日経，読売の 3 紙より作成したテストセットを用いた評価実験から Lead 手法，TF·IDF 手法などの従来手法と比較して提案手法がより高い抽出精度を実現すること，ESK を用いた MMR が従来の MMR よりパラメータに対して頑健であることがわかった．さらに，複数文要約を目的とした重要文抽出精度の新しい評価尺度についても提案する．
キーワード: 複数文書要約，パターンマイニング，カーネル法，Maximal Marginal Relevance

Multiple Document Summarization using Sequential Pattern Mining

Tsutomu Hirao[†] Jun Suzuki[†] Hideki Isozaki[†] Eisaku Maeda[†]
[†]NTT Communication Science Laboratories, NTT Corp.
2-4, Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0237 Japan
{hirao,jun,isozaki,maeda}@cslab.kecl.ntt.co.jp

Abstract

In this paper, we propose a multiple document summarization method using a sequential pattern mining algorithm. We extract important sentences in the following way; First, extracting term patterns from target docment set by using PrefixSpan. Second, identifying significant patterns based on χ^2 statistics, Third, determining a sentence score using the patterns weighting based on TF·IDF. Moreover, we propose a kernel-based MMR (Maximal Marginal Relevance) for minimizing reduandant sentences. This method employs a similarity measure based on Extended String Subsequence kernel instead of cosine similarity. In addition, we define an evaluation measure for deta set includes redundant sentences, *i.e.*, there are many sentences whose meaning are the same. The evaluation results show that our extraction method is better than conventional methods and the kernel-based MMR outperforms conventional MMR.
Keywords: Multiple Document Summarization, Sequencial Pattern Mining, Kernel Methods, Maximal Marginal Relevance

1　はじめに

近年，TDT (Topic Detection and Tracking) などに代表される特定のトピックを持つ文書を検索する技術の発展に伴い，特定の話題に関連する文書集合を得ることができるようになってきた．しかし，検索結果である文書全てに目を通すことは手間がかかり問題である．そこで，こうした人間の負荷を軽減させる技術として複数文書要約が必要とされている．

このような要求を背景として，DUC (Document Understanding Conference)[1]や TSC (Text Summarization Challenge)[6][2] などの評価型ワークショップでも複数文書要約をタスクとして採用している．さらに，コロンビア大学の Mckeown らによって開発された Newsblaster [3] やミシガン大学の Radev らによって開発された NewsInEssense [4] のようなウェブ上のニュースサイトからデータを自動的に集めて要約するといった試みも始まっている．

一般的に，複数文書要約では，与えられた文書集合の特徴を見つけることと要約として冗長性を削減することが重要となる．本稿では，文書集合の特徴として，文内のスキップも許した単語の組み合わせを用いる手法を提案する．単語の組み合わせの抽出には，系列パターンマイニングアルゴリズムとして知られる PrefixSpan[7] を用いる．

[1] http://duc.nist.gov/
[2] http://lr-www.pi.titech.ac.jp/tsc/
[3] http://www.cs.columbia.edu/nlp/newsblaster/
[4] http://www.newsinessence.com/

また，冗長文の性削減法としては，Carbonel らによって提案された MMR (Maximal Marginal Relevance)[1] に用いる類似度尺度に拡張ストリング・カーネルを用いた手法を提案する．

以下，2 章では，従来の要約手法について述べ，3 章で，系列パターンマイニングアルゴリズムを用いて文書セットに特徴的な単語列を抽出し，それに基づき文の重要度を決定する手法について述べる．4 章で，拡張ストリング・カーネルと MMR について説明する．5 章では，ある重要文に対してそれと同じ意味を持つ文が他にも存在するような冗長なデータにおける評価指標を定式化し，12 トピックからなるテストセットを用いた評価実験の結果について述べる．6 章でまとめる．

2 従来の手法

重要文抽出に基づく要約手法では，何らかの手法で文の重要度を決定する必要がある．一般的によく用いられる手法は，下記の式で表わされる単語重要度に基づく手法である [10]．

$$\mathrm{score}(S_i) = \sum_{t \in S_i} tf(t, S_i) w(t) \qquad (1)$$

$tf(t, S_i)$ は単語 t の文 S_i での頻度，$w(t)$ は単語 t の重要度である．$w(t)$ としては，TF，IDF，TF・IDF などが用いられる．

基本的には，単一文書要約でも複数文書要約でもスコアの高い文を重要文として抽出することとなる．ただし，複数文書要約の場合では，文書セットに特徴的な単語に対して (1) 式の $w(t)$ を変更する (大きくする)，あるいは，特徴的な単語のみを重み付けの対象とするという工夫が考えられる．

たとえば，森ら [4] は，与えられた文書集合を階層的にクラスタリングし，クラスタ分割に寄与する単語に高い重みを与えるように単語重要度を決定し，それに基づき文の重要度を決定している．これは，文書セットに特徴的な単語に対して (1) 式の $w(t)$ を大きく見積もるようにしていることに等しい．また，情報量基準を用いて着目した文書集合の特徴を決定する手法も提案されている．Swan ら [9] は，単語の出現頻度分布に着目し，対象とする文書集合に偏って出現する単語を χ^2 検定を用いて決定している．Lin ら [3] は，文書集合に偏って出現する単語と 2-gram，3-gram をエントロピーゲインに基づき決定し，それらを用いて文の重要度を決定する手法を提案している．

これらの手法では，文書集合の特徴として単語やその n-gram を用いており，隣接関係にない単語の組み合わせを考慮できないという問題がある．

また，複数文書要約において必須とされる冗長文の削減手法としては，クラスタリングに基づく手法，MMR (Maximal Marginal Relevance)[1] を用いた手法がある．クラスタリングを用いた手法としては，文献 [8] などがあり，文集合をクラスタに分割し，各クラスタから (1) 式で求めた重要度の高い文を抽出することで冗長な文が少なくなるよう工夫している．また，MMR は文献 [2] など，多くのシステムで採用されている．この手法は，ある観点に基づいて文の重要度を決定し，順序付けした後に冗長な文が少なくなるよう再順序付けを行う手法である．クラスタリング，MMR を用いた手法ともに文間の類似度尺度が重要となるが，その検討が十分であるとはいえない．

3 単語列重要度に基づく重要文抽出

前章で説明した従来手法による文のスコア付けは，文書セットの特徴として単語やその n-gram を用いる手法である．しかし，これら以外にも重要文抽出に有効な単語の組み合わせは存在する．そこで，本稿では，文中に現れるスキップを許した単語の組み合わせ，単語列に重み付けを行い文の重要度を決定する手法を提案する．

3.1 単語列の抽出

まず，以下の文を例に本稿で用いる単語列について説明する．

(暗証) (番号) (を) (入力) (する) (.)

従来手法は，例文中の個々の単語，「暗証」，「番号」，「を」，「入力」，「する」，「．」や，n-gram，「暗証-番号」，「番号-を」などに対して重要度を決定し，それらの和を文の重要度[5]とする手法であった．本稿ではこのような 1 つ 1 つの単語や n-gram だけでなく，スキップを許した組み合わせに対しても重み付けを行い文の重要度を決定する．つまり，上記の例では，「暗証-入力」，「番号-する」や「暗証-を-入力」などの単語の組み合わせにも重み付けを行い文の重要度を決定する．ここで，単語列とは文中の単語の前から後ろへの組み合わせでできていることに注意されたい[6]．すなわち，文を単語の並びに意味のある系列データであると考える．大量の文が与えられた場合，ここから先のような単語列を抽出する問題は系列パターンマイニングの問題として定式化される．系列データベースに対して，与えられた最小サポート値以上の出現回数を満すアイテム列を抽出する問題である．いま，最小サポート値を n，系列データを文，アイテムを単語として考えると，n 文以上に出現する単語列をもれなく抽出するということである．本稿では，このような系列パターンマイニングアルゴリズムとして知られる PrefixSpan [7] を用いて単語列を抽出する．ただし，PrefixSpan によって抽出される単語列には長さが 1(1 つの単語) からなる単語列や n-gram も含まれている．

3.2 文書セットに特徴的な単語列の決定法

PrefixSpan を用いることにより，単語列を高速に抽出することができる．しかし，抽出した単語列の全てが重要

[5] 実際には，自立語や名詞など品詞でフィルタリングを行い重み付け対象の単語を絞る場合が多い．

[6] 「入力-暗証」という組み合わせは考えない．

表 1: 2×2 分割表

	T	$\neg T$
p	n_{11}	n_{12}
$\neg p$	n_{21}	n_{22}

文抽出に有効であるとは限らない．そこで，Swan，Lin らの考えに基づき，情報量基準を用いて重要文抽出に有効であると考えられる単語列を決定する．具体的には，着目する単語列の全データに対する出現頻度と要約対象文書集合での出現頻度の偏りを χ^2 値で検定し，出現傾向に偏りのある単語列のみを重み付けの対象とする．

いま，要約対象となる文書集合を T として，T に含まれる単語列を p とすると，表 1 の分割表を得ることができる．n_{11} は，要約対象 T において，単語列 p が出現する文の数，n_{12} は，T 以外で，p が出現する文の数，n_{21} は，T において，p が出現しない文の数，n_{22} は T 以外で p が出現しない文の数である．この時，χ^2 値は以下の (2) 式で計算される．

$$\chi^2 = \frac{(n_{11}+n_{12}+n_{21}+n_{22})(n_{11}n_{22}-n_{12}n_{21})^2}{(n_{11}+n_{12})(n_{11}+n_{21})(n_{12}+n_{22})(n_{21}+n_{22})} \quad (2)$$

こうして計算した χ^2 値が大きいほど，p は T に偏って出現していると考えられる．ここで，ある閾値 θ を設けて $\chi^2 > \theta$ を満す p を重み付けの対象とすることを考えると，相関ルール抽出で同様の事を行う AprioriSMP[5] の考え方を適用でき，より高速に単語列を抽出することができる．また，χ^2 値の高い上位 N 件の単語列を抽出する場合には，同じく相関ルール抽出で同様のことを行う TidalSMP の考え方が適用でき，高速化が可能である．詳しくは文献 [11] を参照されたい．本稿では，前者の θ 以上の単語列を文書集合の特徴として抽出し，重み付けの対象とした．

3.3 単語列重要度の決定法と文のスコア付け

単語の重み付けに関しては，TF・IDF がよく用いられる．TF・IDF は，着目したデータに頻出する単語は重要であり，全データベースを見渡してめずらしい単語は重要であると考えた重み付け手法である．この考えに基づき，以下の式で単語列の重み決定する．

$$w(p) = \frac{\log(f(p,T)+1)\cdot\log(\frac{|DB|}{f(p,DB)})}{len(p)} \quad (3)$$

ここで，$f(p,T)$ は要約対象 T における p の出現文数，$f(p,DB)$ は全データにおける p の出現文数であり，$|DB|$ は全データ数 (文数)，$len(p)$ は単語列の長さ (単語数) を表わす．通常の TF・IDF では，(3) 式の分子のみであるが，長い単語列は $f(p,DB)$ が極端に小さくなる傾向があり，非常に高い重みを持つこととなる．そこで，単語数で割ることで正規化した．最終的に，こうして決定した単語列の重み $w(p)$ を用いて以下の式で文のスコアを決定する．

$A = \{\}$;
$R = \{S_1, S_2, \cdots, S_\ell\}$;
N =The number of sentences as output;
While($|A| < N$){
  $S^* = \mathrm{MMR}(A, R)$;
  $A = A \cup \{S^*\}$;
  $R = R - \{S^*\}$;
}
Output A, where

$$\mathrm{MMR}(A,R) = \begin{cases} \operatorname{argmax}_{S_i \in R} \mathrm{score}(S_i) & \text{if} \quad A = \phi \\ \operatorname{argmax}_{S_i \in R} \left(\alpha\, \mathrm{score}(S_i) - (1-\alpha) \max_{S_j \in A} \mathrm{sim}(S_i, S_j)\right) & \end{cases}$$

図 1: MMR による再順序付けアルゴリズム

$$\mathrm{score}(S_i) = \sum_{p \in S_i} w(p) \quad (4)$$

4 Kernel による類似度を用いた MMR

前章で説明したように文書集合に特徴的な単語列を用いて文のスコアを決定すると，同じような単語列を持つ文がスコアの上位を占める可能性が高い．共通の単語列を多く含む文集合は，文意が類似していることが多く，要約として冗長になるという問題がある．そこで，こうした冗長な文を減らすことが必要となる．

4.1 Maximal Marginal Relevance

MMR は，Carbonel らによって提案された再順序付けの手法であり，ある観点で点で順序付けされたデータに対して別の観点を導入し，再順序付けを行う手法である．そのアルゴリズムを図 1 に示す．R は，再順序付けの対象となる文集合を表わし，A は出力となる文集合を表わす．$\mathrm{score}(S_i)$ は式 (4) で求めた文のスコアを [0,1] の値に正規化[7]したものである．$\mathrm{sim}(S_i, S_j)$ は，文間の類似度を表わす指標であり，一般的にはコサイン類似度が良く用いられる．ここでの MMR は，文のスコアから既に重要文として選択した文との類似度をペナルティとして引いたものである．α はそれぞれの項の重みを決めるパラメータである．こうして，既に重要文として選んだ文となるべく類似していない文を採用することで冗長度の少ない文集合が得られる．

4.2 カーネルによる類似度

前節で説明した MMR では，文間の類似度尺度が重要となる．理想的には，文意が同じ 2 文間では 1，文意が異なる 2 文間では 0 となる類似度が望ましい．しかし，実際にこのような類似度尺度を用いることは不可能である．ここで，コサイン類似度を用いる場合，類似度は文間の単語の

[7] 具体的には，式 (4) のスコアの最大値で割ることで正規化した．

表 2: 部分文字列とその重み (「abbab」と「abaca」)

	abbab	abaca
aaa	–	λ^2
aab	λ^2	–
aac	–	λ
aba	2λ	$1+\lambda^2$
abb	$1+2\lambda^2$	–
abc	–	λ
aca	–	λ^2+1
aa	λ^2	$2\lambda+\lambda^3$
ab	$2+\lambda+\lambda^3$	1
ac	–	$1+\lambda^2$
ba	$1+\lambda$	$1+\lambda^2$
bb	$1+\lambda+\lambda^2$	–
bc	–	λ
ca	–	1
a	2	3
b	3	1
c	–	1

重なりに依存する．確かに，文間で共通する単語が多い場合，それらは内容的にも類似する場合もある．しかし，常にそうなるとは限らない．たとえば，文献 [13] で述べられているように，要約対象となる情報源が冗長でない場合，コサイン類似度による MMR を用いると，結果的には重要文の抽出精度が下がり，有効でないという報告もある．このように MMR に用いる類似度尺度には注意を払わねばならない．

本稿では，コサイン類似度に代る尺度として，文献 [12] で提案された拡張ストリング・カーネルを用いる．拡張ストリング・カーネルは，入力として 2 文を受け取り，それらの部分文字列を基底とする空間へ写像し，内積計算することと等しく定義される．ここでの部分文字列とは，d 個以下のスキップを許した単語の組み合わせを指し，その重み (座標値) は，スキップした単語数 ℓ に対して減衰パラメータ λ を用いて λ^ℓ で与えられる．たとえば，アルファベットが単語を表わすとして T:「abaca」，U:「abbab」という文を考えると，それらから得られる部分単語列とその重みは，表 2 のとおりとなる．よって，カーネルの値 ($K_{\mathrm{esk}}(T,U)$) は以下となる．拡張ストリングカーネルの詳細については文献 [12] を参照されたい．

$$\begin{aligned} K_{\mathrm{esk}}(T,U) = {} & (1+\lambda^2)\times 2\lambda + 1\times(2+\lambda+\lambda^3) \\ & + (2\lambda+\lambda^3)\times\lambda^2 \\ & + (1+\lambda^2)\times(1+\lambda) \\ & + 3\times 2 + 1\times 3 \end{aligned} \tag{5}$$

なお，カーネルの入力となる文の長さは一定でないため，以下の正規化した値を文間の類似度として用いる．

表 3: 評価用データセット

トピック	文書数			文数		
	毎日	読売	日経	毎日	読売	日経
1	12	13	10	165	324	110
2	23	10	10	445	92	111
3	17	16	14	320	224	155
4	16	12	14	213	221	190
5	18	13	14	178	127	104
6	8	9	7	160	174	166
7	11	6	11	282	154	221
8	14	16	15	281	297	211
9	35	24	22	618	475	185
10	16	8	5	232	81	59
11	36	16	11	655	194	137
12	27	17	15	482	407	229

表 4: 各紙の重要文数

トピック	毎日	読売			日経		
		A	B	C	A	B	C
1	17	12	16	15	11	12	14
2	45	16	21	16	9	22	9
3	32	13	27	18	17	27	15
4	22	12	6	6	21	9	13
5	18	15	11	14	19	15	12
6	16	21	10	11	14	7	10
7	29	3	8	5	11	17	8
8	29	14	19	20	15	30	19
11	62	41	42	39	34	36	32
12	24	16	7	10	11	6	5
17	66	31	11	12	34	18	17
18	49	24	17	20	28	10	13

$$\mathrm{Sim}_{\mathrm{esk}}(T,U) = \frac{K_{\mathrm{esk}}(T,U)}{\sqrt{K_{\mathrm{esk}}(T,T)K_{\mathrm{esk}}(U,U)}} \tag{6}$$

5 実験結果と考察

5.1 実験データ

文献 [13] で用いられた 12 トピックからなる文書集合を拡張したものを用いた．文献 [13] では，各トピックとも毎日新聞のみを対象として重要文を抽出していたが，本稿では，これに加え，読売新聞，日経新聞の同一トピックの記事も対象とした．毎日新聞に関しては 3 名の被験者 (A, B, C と呼ぶ) が，既に文の数を基準として 10%の要約率で重要文を抽出してあるので，これらの重要文に対応する文を読売新聞，日経新聞から抽出した．なお，この時，毎日新聞の文と他紙の文が 1 対 1 で対応するとは限らない．よって，毎日の 1 文に対して他紙の文が多数対応付けられることがあり得る．また，対応文がないということもあり得る．表 3 に各トピックの記事数と文数を示し，表 4 に毎日新聞の重要文の数とそれらに対して対応づけられた各紙の重要文の数を示す．

表 5: 重要文の正解例

毎日	読売	日経
a	b,c	d
e	f	g,h,i

5.2 評価指標

TSC では，抽出すべき重要文の数が既知であるとして，システムがその数だけ重要文を抽出した際の精度 (Precision) でシステムを評価している．しかし，この評価法では本稿の実験データのように同じ意味を持つ複数の文が存在するような冗長なデータには適していない．いま，抽出すべき文の数が 2 文であるとする．この時，文 S_1 と S_2 が同意の文であるという下で，システムが S_1，S_2 という 2 文を出力すると精度は 1 となってしまうので適切に評価できているとはいえない．このような場合は，S_1，S_2 のどちらかのみを評価すべきである．このように冗長なデータに対しては，それを考慮した評価法が必要となる．そこで本稿は，冗長データに対する被覆率 (coverage) という評価指標を定義し，評価に用いた．以下に詳述する．

いま，毎日新聞から重要文として抽出された文集合を $\{m_1, \cdots, m_i, \cdots, m_h\}$ と表わし，システムは h だけ文を出力するとする．ここで，m_i に対応する読売新聞の重要文集合を $\{y_{i,1}, \cdots, y_{i,j}, \cdots, y_{i,x}\}$ と表わす．すなわち，m_i に対して x 個の文が対応づけられていることとなる．日経新聞の場合も同様に m_i に対応する z 個の重要文を $\{n_{i,1}, \cdots, n_{i,k}, \cdots, n_{i,z}\}$ として表わす．この時，i 番目の要素に対する評価値 $e(i)$ を以下の (11) 式で定義する．

$$e(i) = \max(v(m_i), \sum_{j=1}^{x} v(y_{i,j})/x, \sum_{k=1}^{z} v(n_{i,k})/z) \quad (7)$$

ただし，$v(\beta)$ は以下の式で定義する．である．

$$v(\beta) = \begin{cases} 1 & \text{if the system outputs } \beta \\ 0 & \text{otherwise} \end{cases} \quad (8)$$

この関数 e は，システムが毎日新聞の i 番目の文を出力した場合には 1，他紙による解を出力した場合にはその解の数に応じたスコアを返す関数上記を用いて，coverage は以下の式で定義する．

$$\text{coverage} = \frac{\sum_{i=1}^{h} e(i)}{h} \quad (9)$$

たとえば，あるトピックに対する正解重要文が表 5 で与えられたとする．この時，毎日新聞から抽出された重要文数は 2 文であるので，システムは 2 文抽出することとなる．システムが，「a，b」を抽出したとすると，

$$e(1) = \max(1, 0.5, 0) = 1$$
$$e(2) = \max(0, 0, 0) = 0$$

となり，coverage=0.5 となる．また，「b，g」を抽出した場合には，

$$e(1) = \max(0, 0.5, 0) = 0.5$$
$$e(2) = \max(0, 0, 0.3) = 0.3$$

となるので，coverage=0.4 となる．このように coverage は，冗長な文を考慮しつつ，抽出した文集合がどの程度，正解重要文集合の情報を被覆しているかを測ることのできる指標である．

5.3 重要文抽出精度の評価

先に述べたコーパス，評価指標を用いて提案手法の重要文抽出性能を調べた．比較評価に用いた重要文抽出手法は以下のとおりである．

TF·IDF 手法 (1) 式を用いて文書集合中の全ての文の重要度を決定し，重要度の高い文から順に重要文として採用する手法．

Lead 手法 文書を時系列で並べ，各文書の先頭から順に重要文として採用する手法．

n-gram 手法 着目した文書集合において，χ^2 値が 100 以上の n-gram を対象として (4) 式で文の重要度を決定し，重要度の高い文から順に重要文として採用する手法．なお，n は 1，2 まで，3 までと制限無しの場合をそれぞれ評価した．

単語列手法 着目した文書集合において，χ^2 値が 100 以上の単語列を対象として (4) 式で文の重要度を決定し，重要度の高い文から順に重要文として採用する手法．n-gram の場合と同様に単語列の長さ (単語数) は，1，2 まで，3 までと制限無しの場合をそれぞれ評価した．付録に単語列の一例を示す．

表 6 に評価結果を示す．表 6 より，全てのデータセットに対して Lead 手法，TF·IDF 手法よりも n-gram，単語列手法が良い成績である．このことから，与えられた文書集合に特徴的な単語やその組み合わせに基づき文の重要度を決定することが有効であることがわかる．なお，一般的に報道記事を対象とすると Lead 手法が有効であることが知られているが，本稿の場合，1 つのトピックあたりの文書数が多いこと，情報源が冗長であることが影響して良い結果は得られていない．更に提案手法と n-gram 手法を詳しく比較する．単語数が 1 の時には，共に個々の単語しか重み付けの対象とならないので，TF·IDF 法において重み付けの単語を χ^2 を用いて選抜したことに等しい．この時，TF·IDF 法よりも成績が良いことを考えると，χ^2 値による単語の選抜が有効に働いていることがわかる．更に，単語数を変化させた場合，提案手法は成績が向上する傾向があり，単語数に制限を設けない場合が最も良い成績である．これに対し n-gram 手法では，単語数を増やしても成績の変化は小さく，制限を設けない場合にはやや成績が低下す

表 6: 重要文抽出の評価結果

セット \ 手法	TF·IDF	Lead	n-gram				単語列			
			1	2	3	∞	1	2	3	∞
A	0.253	0.258	0.290	0.298	0.289	0.285	0.290	0.316	0.312	**0.331**
B	0.238	0.237	0.268	0.260	0.265	0.262	0.268	0.272	0.272	**0.282**
C	0.239	0.261	0.269	0.280	0.279	0.276	0.269	0.291	0.299	**0.307**

る傾向が見られる．このことから，重要文抽出に有効な単語の組み合わせは n-gram のように隣接したものだけではなく，スキップがあるような組み合わせを考慮しなければならないことがわかる．また，単語数については，制限を設けない場合に一番良い成績が得られていることから，有効な単語列は長さに依存しないことがわかる．

以上より，トピックに特徴的な単語列に基づいて文の重要度を決定する手法が有効であることがわかった．

5.4 冗長文削減の評価

次に，単語列を用いて (4) 式により決定した文の重要度で順序付けたデータを図 1 のアルゴリズムを用いて再順序付けを行った．なお，計算量の観点から再順序付けの対象は，各トピックにおける毎日新聞の正解重要文数の 3 倍，すなわち，上位 $3h$ の文を対象に行った．また，MMR に用いる類似度には従来からよく用いられるコサイン類似度 (以下，MMR(COS)) と ESK による類似度 (以下，MMR(ESK)) を用いてこれらを比較した．なお，ESK に与えるパラメータは，組み合わせの数 $d = 3$, 減衰パラメータ $\lambda = 0.7$ とした．図 2 に α を 1 から 0.20 まで，0.05 刻で変化させた場合の coverage の平均を示す．

A，B，C のセットに共通して MMR(ESK) では，初期値 ($\alpha = 1$) を上回る場合が必ずあるが，MMR(COS) では，初期値を越えることはない．また，$\alpha > 0.5$ の区間では，MMR(ESK) は，MMR(COS) よりも良い成績であり，初期値の成績と比較して性能の劣化も小さい．このことから，$1 > \alpha > 0.5$ の区間では，MMR(ESK) を用いる方が適しているといえる．

しかし，MMR(ESK) で初期値を越える場合があるとはいえ，その差は小さく，この結果から MMR による再順序付けが有効に働いているとは言えない．そこで，更に詳しく各トピックごとの MMR の効果を調べた．表 7 に α を変化させた時の最も良い成績を記す．なお，カッコ内の値は α を表わす．表 7 より，多くのトピックで MMR を用いることによって成績が向上している．初期値に対しても，最大で 10 ポイント程度の成績の向上が見られ MMR による再順序付けが有効であることがわかる．特にセット A と C では，MMR(COS) と MMR(ESK) の差は大きい．しかし，最適な α は，セット，トピックによって大きく異なっている．MMR は，再順序付けを行うだけなので，再順序付けの対象となる文集合に既に選んでいない正解重要文が含まれていなければ，成績が向上することは見込めな

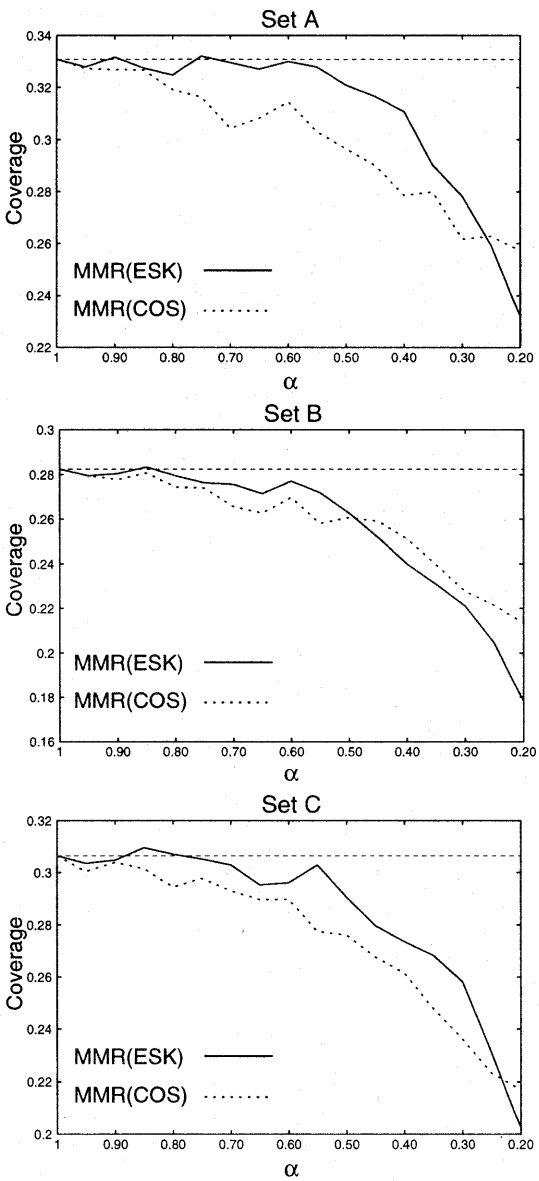

図 2: α を変化させた場合の coverage の変化

表 7: MMR の効果

トピック	A			B			C		
	初期値	MMR(COS)	MMR(ESK)	初期値	MMR(COS)	MMR(ESK)	初期値	MMR(COS)	MMR(ESK)
1	**0.382**	0.382 (1.00)	0.382 (1.00)	**0.265**	0.265 (1.00)	0.265 (1.00)	0.294	**0.353** (0.40)	0.294 (1.00)
2	0.256	0.256 (1.00)	**0.278** (0.50)	0.200	**0.267** (0.50)	0.222 (0.70)	0.289	0.289 (1.00)	**0.311** (0.70)
3	0.312	0.344 (0.65)	**0.375** (0.55)	0.234	0.250 (0.60)	**0.281** (0.50)	0.219	0.250 (0.65)	**0.312** (0.40)
4	0.330	**0.375** (0.60)	**0.375** (0.75)	0.333	**0.379** (0.60)	**0.379** (0.75)	0.273	0.318 (0.60)	**0.364** (0.55)
5	0.380	**0.435** (0.40)	0.417 (0.15)	0.333	**0.389** (0.40)	0.361 (0.15)	**0.389**	0.389 (1.00)	0.389 (1.00)
6	0.375	0.375 (1.00)	**0.438** (0.45)	**0.438**	0.438 (1.00)	0.438 (1.00)	**0.438**	0.438 (1.00)	0.438 (1.00)
7	**0.293**	0.293 (1.00)	0.293 (1.00)	**0.293**	0.293 (1.00)	0.293 (1.00)	**0.241**	0.241 (1.00)	0.241 (1.00)
8	0.241	**0.276** (0.85)	**0.276** (0.60)	0.138	**0.184** (0.25)	0.172 (0.60)	0.310	**0.345** (0.85)	**0.345** (0.60)
9	0.323	**0.347** (0.70)	0.331 (0.50)	0.210	**0.231** (0.40)	0.210 (1.00)	0.306	0.306 (1.00)	**0.323** (0.75)
10	0.396	0.396 (1.00)	**0.417** (0.90)	**0.417**	0.417 (1.00)	0.417 (1.00)	0.375	**0.417** (0.90)	**0.417** (0.85)
11	0.316	0.316 (1.00)	**0.346** (0.55)	**0.242**	0.242 (1.00)	0.242 (1.00)	0.197	**0.242** (0.20)	0.197 (1.00)
12	0.367	0.378 (0.80)	**0.388** (0.55)	0.286	**0.327** (0.50)	**0.327** (0.50)	0.347	**0.367** (0.60)	**0.367** (0.55)
平均	0.331	0.348	**0.360**	0.282	**0.307**	0.301	0.306	0.330	**0.333**

い．さらに，再順序付け対象の文集合の冗長度が低い場合にもその効果は見込めない．既に選ばれていない正解を含む割合と冗長な文を含む割合によって α の最適値は定まると考えるが，これを自動的に決定することは難しく，今後の課題である．

6 おわりに

本稿では，トピックに偏って出現する単語列を抽出し，重み付けを行い，それに基づき文の重要度決定する手法を提案した．また，冗長な文を削減する手法として，MMR の類似度に拡張ストリング・カーネルを用いることを提案した．さらに，同意の文が複数ある冗長なデータを用いた場合の重要文抽出の評価指標を定式化した．評価実験の結果，提案手法が従来手法よりも良い成績であることを確認した．

謝辞

系列パターンマイニングアルゴリズムに関して有益な議論をさせて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．また，日頃さまざまなコメントをいただく NTT コミュニケーション科学基礎研究所の知識処理研究グループの皆様に感謝いたします．さらに，データの使用を許諾してくださった毎日新聞社，読売新聞社，日経新聞社に感謝いたします．

参考文献

[1] Carbonell, J. and Goldstein, J.: The Use of MMR,Diversity-Based Reranking for Reordering Documents and Producing Summaries, *Porc. of the 21th Annual International ACM SIGIR Conference on Research and Development in the Information Retrieval*, pp. 335–336 (1998).

[2] Goldstein, J., Mittal, V., Carbonell, J. and Kantrowitz, M.: Multi-Document Summarization by Sentence Extraction, pp. 40–48 (2000).

[3] Lin, C.-Y. and Hovy, E.: The Automated Acquisition of Topic Signatures for Text Summarization, *Proc. of the 18th International Conference on Computational Linguistics (COLING2000)*, pp. 495–501 (2000).

[4] Mori, T. and Sasaki, T.: Information Gain Ratio meets Maximal Marginal Relevance — A method of Summarization for Multiple Documents —, *In Proc. of NTCIR Workshop 3 Meeting — Part V: Text Summarization Challenge 2 (TSC2)*, pp. 25–32 (2002).

[5] Morishita, S. and Sese, J.: Traversing Itemset Lattices with Statistical Metric Pruning, *Proc. of ACM SIGACT-SIGMOD-SIGART Symp. on Database Systems(PODS)*, pp. 226–236 (2000).

[6] Okumura, M., Fukusima, T. and Nanba, H.: Text Summarization Challenge 2 - Text summarization evaluation at NTCIR Workshop 3, *HLT-NAACL 2003 Workshop: Text Summarization (DUC03)*, pp. 49–56 (2003).

[7] Pei, J., Han, J., Mortazavi-Asl, B. and Pinto, H.: PrefixSpan: Mining Sequential Patterns Efficiently by Prefix-Projected Pattern Growth, *Proc. of 17th International Conference on Data Engineering (ICDE 2001)*, pp. 215–224 (2001).

[8] Stein, G., Strazalkowski, T. and Wise, G.: Summarizing Multiple Documents using Text Extraction and Interactive Clustering, *Proc. of the Pacific Association for Computational Linguistics 1999*, pp. 200–208 (1999).

[9] Swan, R. and Allan, J.: Extracting Significant Time Varying Features from Text, *Porc. of the 8th ACM CIKM*, pp. 38–45 (1999).

[10] Zechner, K.: Fast Generation of Abstracts from General Domain Text Corpora by Extracting Relevant Sentences, *Proc. of the 16th International Conference on Computational Linguistics*, pp. 986–

989 (1996).
[11] 磯崎秀樹, 平尾努, 鈴木潤: 機械学習のための組み合わせ素性の選択基準について, 情報処理学会研究報告 NL-158-10 (2003).
[12] 鈴木潤, 平尾努, 磯崎秀樹, 前田英作: 素性選択機能をもった String Kernel の提案, 情報処理学会研究報告 NL-157-6, pp. 41–48 (2003).
[13] 平尾努, 賀沢秀人, 磯崎秀樹, 前田英作, 松本裕治: 機械学習による複数文書からの重要文抽出, 自然言語処理, Vol. 10, No. 1, pp. 81–108 (2003).

付録:単語列の例

トピック番号 10 の「インドのミサイル実験成功」に関して抽出した単語列 (χ^2 値でランキングした結果の上位 10 件) を以下に示す. なお, 対象とした品詞は, 名詞, 動詞, 助動詞である.

長さ	単語列	χ^2 値
1	実験	2432.45
	ミサイル	1891.27
	インド	1520.62
	発射	911.655
	弾道	783.621
	射程	761.954
	中距離	760.801
	核	637.032
	アグニ 2	626.009
	連立	516.709
2	実験 た	1654.5
	ミサイル 実験	1528.45
	実験 し	1363.43
	インド 実験	1269.57
	発射 実験	1256.58
	ミサイル た	1253.5
	インド た	1092.04
	ミサイル し	1026.44
	インド ミサイル	1008.5
	実験 実験	897.039
3	実験 し た	1214.8
	ミサイル 実験 た	1053.26
	インド 実験 た	989.634
	発射 実験 た	940.874
	ミサイル 実験 し	918.344
	実験 た た	887.193
	ミサイル し た	884.597
	インド ミサイル 実験	805.783
	発射 実験 し	784.455
	インド ミサイル た	761.954
4 以上	ミサイル 実験 し た	805.783
	発射 実験 し た	716.967
	インド ミサイル 実験 た	649.526
	インド 実験 し た	626.389
	実験 た し た	605.659
	発射 実験 た た	603.567
	実験 し た た	594.074
	ミサイル 発射 実験 た	558.706
	インド 発射 実験 た	514.784
	中距離 ミサイル 発射 実験	513.875