拡張ストリングカーネルを用いた要約システムの自動評価法

平尾 努[†] 奥村 学[††] 磯崎 秀樹[†]

近年，言語生成技術を含む自然言語処理，たとえば，自動要約や機械翻訳の評価型ワークショップがさかんに開催されており，システム評価のためのコーパスが整備されつつある．しかし，こうした自然言語処理技術の評価は，多くの場合，人間の評価に頼らざるをえない．よって，再評価実験が困難である，多大なコストがかかるという問題がある．そこで，人間の評価に匹敵する正確な自動評価法の確立に大きな期待が寄せられている．本稿では，コンボリューションカーネルの 1 つである拡張ストリングカーネル（Extended String Subsequence Kernel）を用いた要約システムの自動評価法を提案する．Text Summarization Challenge 3（TSC-3）のデータを用いて提案手法を従来手法である ROUGE と比較した結果，人間の評価結果との相関において，提案手法がより高く，頑健性に優れていることが分かった．

An Automatic Evaluation Method for Summarization Systems with Extended String Subsequence Kernel

Tsutomu Hirao,[†] Manabu Okumura[††] and Hideki Isozaki[†]

Recently, several evaluation workshops for automatic summarization are held. These evaluation workshops employ human evaluations, which are essential in terms of achieving high quality evaluations results. However, human evaluations require a huge effort and the cost is considerable. Moreover, we cannot automatically evaluate a new system even if we use the corpora built for these workshops, and we cannot conduct re-evaluation experiments. In order to promote the study of automatic summarization, we need an accurate automatic evaluation method that is close to human evaluation. In this paper, we present an evaluation method that is based on extended string subsequence kernel that measure the similarities between texts considering their substructures. We conducted an experiment using automatic summarization evaluation data developed for Text Summarization Challenge 3 (TSC-3). Our method shows higher correlation than ROUGE family with human evaluation.

1. はじめに

近年，言語生成技術をともなう自然言語処理，たとえば，自動要約や機械翻訳の研究に注目が集まっており，日米で様々な評価型ワークショップが開催されている．自動要約に関しては，米国の DUC（Document Understanding Conference）[☆]が 2001 年より毎年開催されており，日本の TSC（Text Summarization Challenge）[☆☆]が 2001 年より 1 年半に一度の割合で開催されている．このような評価型ワークショップが継続的に開催されることによって，システム評価のためのコーパスが大規模に整備されつつある．

しかし，こうした評価型ワークショップでは，人手による評価[☆☆☆]に頼っているため，1 回限りの評価しかできない．よって，ワークショップに参加していないシステムが，それらのデータを用いて性能を測ることは難しい．また，参加したシステムですら，同じ評価を再現することは難しく，蓄積されたコーパスを有効利用できないという問題がある．さらに，人手による評価には多大なコストがかかるという無視できない問題もある．

こうした状況を打開するため，人手の評価に代わる正確な自動評価法の確立が急務となっている．自動評価法の実現は，言語生成技術を包含する自然言語処理

† 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
NTT Communication Science Laboratories, NTT Corporation
†† 東京工業大学精密工学研究所
Precision and Intelligence Laboratory, Tokyo Institute of Technology

☆ http://duc.nist.gov
☆☆ http://www.lr.pi.titech.ac.jp/tsc
☆☆☆ 要約の評価には，内容と読みやすさの評価が存在するが，本稿では内容評価のみを対象とする．

の発展のためには欠かせない．

本稿では，コンボリューションカーネルの 1 つである拡張ストリングカーネルを用いた要約システムの自動評価法を提案する．拡張ストリングカーネルはテキスト間の類似度を単語列，その意味ラベル列，単語と意味ラベルの組合せの列に基づき計算する．

提案手法を TSC-3 のデータを用いて評価したところ，従来の自動評価手法である ROUGE と比較して人間の評価結果に対する相関がより高いこと，自動評価法として頑健であることが分かった．

本稿の構成は以下のとおりである．2 章では，従来の自動評価法について詳述し，その問題点を述べる．3 章では，拡張ストリングカーネルを用いた自動評価法について述べる．4 章では，評価実験の結果を示し，5 章で考察を行う．

2. 関 連 研 究

一般的に，要約の自動評価法は，システム要約と参照要約☆間の類似度を測ることで実現される．多くの場合，類似度はシステム要約と参照要約との間で一致する単語列の割合に基づいて計算する．たとえば，現状で最も広く知られている自動評価法である ROUGE-N[9),☆☆]は，システム要約と参照要約との間で一致する単語 N グラムに基づきスコアを決定する．ROUGE-N は，機械翻訳システムの自動評価法として提案された BLEU[12)] を単純化したものである．BLEU は，参照翻訳とシステム翻訳間で一致する N グラムがシステム翻訳中の N グラムに占める割合を計算するため，精度重視の指標といわれている．これに対し，ROUGE は再現率を重視した評価指標であるという特徴を持つ．たとえば，ROUGE-1 であれば，参照要約とシステム要約間で一致したユニグラムの数が参照要約のユニグラム中に占める割合，ROUGE-2 であれば，一致したバイグラムが参照要約のバイグラム中に占める割合を計算する．Lin らは，N を 1〜4 まで変化させた場合，ROUGE-1，ROUGE-2 が人間の評価結果との間の相関が最も高かったことを報告している[9)]．また，Soricut らは，再現率重視の評価指標と精度重視の評価指標を調和平均の変形を用いて統合する手法を提案している．要約だけでなく，機械翻訳，質問応答でも人間の評価結果と高い相関が得られたことを報告している[14)]．

しかし，上述した N グラムの一致率に基づく評価指標は，隣接という強い制約にある語の共起しか考慮できないという問題がある．つまり，隣接関係にはないが，かかり受け関係にあるような語の共起は考慮できない．

こうした問題に対して，Lin らは，スキップを許したバイグラム（スキップバイグラム）も考慮した手法，ROUGE-S，ROUGE-SU を提案している[7),8),10)]．ただし，スキップを許したトライグラムなどを扱うことができない．さらに，ある単語の組合せが参照要約かシステム要約のどちらか一方ではバイグラムとして出現し，もう一方ではスキップバイグラムとして出現した場合，スキップの有無を区別せずに一致数を計算するという問題もある．

また，N グラムの一致率に基づく手法以外としては，参照要約とシステム要約間における最長共通部分列（Longest Common Subsequence: LCS）に基づく手法[8),10),13)] や音声認識分野において広く用いられる単語正解率を用いた手法[6)] も提案されている．LCS を用いるとスキップを含む長い部分列を扱うことができるが，(1) 最長一致する部分列しか見ない，(2) 最長一致部分列が助詞などの機能語のみで構成される，(3) 語順が大きく入れ替わる場合には最長一致部分列が著しく短くなるのでスコアが下がるという問題がある．

さらに，上述したすべての手法が単語表記での一致を見ており，単語の言い換えがあった場合には一致率が著しく下がるという問題がある．

3. カーネル関数を用いた要約システムの自動評価法

2 章で説明した ROUGE☆☆☆では，スキップトライグラムのような長い部分単語列を有効に扱えない，スキップを許した N グラムと通常の N グラムを区別していない，単語の言い換えを吸収できないという問題がある．

そこで，本稿では，これらの問題点を解決するため，拡張ストリングカーネル[5)]（Extended String Subsequence Kernel，以下，ESK）を用いた自動評価法を提案する．

3.1 ESK

ESK[5)] は自然言語処理のために開発され，注目を集めているコンボリューションカーネル[2)] に属するカーネル関数であり，Lodhi らによって提案された String Subsequence Kernel（SSK）[11)]，Cancedda らによっ

☆ モデル要約とも呼ばれ，一般的には人間が作成した要約を指す．

☆☆ ROUGE については 4 章で詳しく説明する．

☆☆☆ 本稿で，「ROUGE」と表記した場合には，ROUGE-N，ROUGE-L，ROUGE-S，ROUGE-SU というバリエーションをすべて含んでいることを表す．

表 1 S1，S2 から抽出した部分単語列とその重み（d は部分単語列の長さを表す）
Table 1 Components of vectors corresponding to S1 and S2 (d is the length of the subsequences).

d	subsequence	S1	S2	d	subsequence	S1	S2	d	subsequence	S1	S2
1	**Becoming**	1	1	2	**Becoming–is**	λ^2	λ^2	2	astronaut–DREAM	0	λ^2
	DREAM	1	1		**Becoming–my**	λ^3	λ^3		astronaut–ambition	0	λ^2
	SPACEMAN	1	1		***SPACEMAN–DREAM***	λ^3	λ^2		astronaut–is	0	1
	a	1	0		SPACEMAN–ambition	0	λ^2		astronaut–my	0	λ
	ambition	0	1		SPACEMAN–dream	λ^3	0		cosmonaut–DREAM	λ^3	0
	an	0	1		SPACEMAN–great	λ^2	0		cosmonaut–dream	λ^3	0
	astronaut	0	1		***SPACEMAN*–is**	1	1		cosmonaut–great	λ^2	0
	cosmonaut	1	0		***SPACEMAN*–my**	λ	λ		cosmonaut–is	1	0
	dream	1	0		a–DREAM	λ^4	0		cosmonaut–my	λ	0
	great	1	0		a–SPACEMAN	1	0		great–DREAM	1	0
	is	1	1		a–cosmonaut	1	0		great–dream	1	0
	my	1	1		a–dream	λ^4	0		**is–*DREAM***	λ^2	λ
2	**Becoming–*DREAM***	λ^5	λ^4		a–great	λ^3	0		is–ambition	0	λ
	Becoming–*SPACEMAN*	λ	λ		a–is	λ	0		is–dream	λ^2	0
	Becoming–a	1	0		a–my	λ^2	0		is–great	λ	0
	Becoming–ambition	0	λ^4		an–DREAM	0	λ^3		**is–my**	1	1
	Becoming–an	0	1		an–SPACEMAN	0	1		**my–*DREAM***	λ	1
	Becoming–astronaut	0	λ		an–ambition	0	λ^3		my–ambition	0	1
	Becoming–cosmonaut	λ	0		an–astronaut	0	1		my–dream	λ	0
	Becoming–dream	λ^5	0		an–is	0	λ		my–great	1	0
	Becoming–great	λ^4	0		an–my	0	λ^2				

て提案された Word Sequence Kernel（WSK）[1] を拡張したものである．ESK では，まずテキストを単語とその意味ラベルを属性としたノード列として考える．そして，テキストを d 個までの部分ノード列に対応する軸を持つ高次元空間へと写像する．ESK は，その空間における内積として定義できる．ただし，陽にテキストを高次元空間へ写像することなく内積を効率的に計算できる．このとき，ノードのスキップに対しては，λ（$0 \leq \lambda \leq 1$）という減衰パラメータを用いてその重みを小さくする．たとえば，ノードを 1 つスキップした場合には，重みが λ となり，2 つスキップした場合には，λ^2 となる．

例として，下記のテキスト，S1，S2 を入力として，ESK の値を計算する．なお，単語の意味ラベルはカッコ内に示す．

S1 Becoming a cosmonaut:{*SPACEMAN*} is my great dream:{*DREAM*}

S2 Becoming an astronaut:{*SPACEMAN*} is my ambition:{*DREAM*}

ここで，“cosmonaut” と “astronaut” は共通の意味ラベル “*SPACEMAN*” を持ち，“ambition” と “dream” は共通の意味ラベル “*DREAM*” を持つ．このような単語の意味ラベルは日本語の場合には日本語語彙大系[15]，英語の場合には WordNet から得ることができる．後述の実験では，語の意味ラベルの獲得に日本語語彙大系を用いた．語の多義解消は行わず，単語に対して可能なすべての意味ラベルを用いた．

S1，S2 において，$d = 2$ とした場合のすべての部分ノード列とその重み付き出現回数を**表 1** に示す．なお，S1，S2 に共通する部分列を太字で表している．たとえば，“Becoming–*DREAM*” という部分列は，S1 では “a”，“cosmonaut:*SPACEMAN*”，“is”，“my”，“great” という 5 つのノードをスキップしており，S2 では同様に 4 つのノードをスキップして出現している．よってその重みは，それぞれ，λ^5，λ^4 となる．$\mathrm{ESK}^{d=2}(\mathrm{S1},\mathrm{S2})$ は，S1，S2 から得た重み付きベクトルの内積であるので，S1，S2 に共通する 15 の部分列の重みの積として以下の式で計算される．$\mathrm{ESK}^{d=2}(\mathrm{S1},\mathrm{S2}) = 1+1+1+1+1+\lambda^9+\lambda^2+\lambda^4+\lambda^6+\lambda^5+1+\lambda^2+\lambda^3+1+\lambda = 7+\lambda+2\lambda^2+\lambda^3+\lambda^4+\lambda^5+\lambda^6+\lambda^9$．参考までに，S1，S2 に共通する単語ユニグラムは 3 個，バイグラムは 1 個，トライグラムは存在しない．

正確には，ESK は以下の式で定義される．

$$\mathrm{ESK}^d(S1,S2) = \sum_{m=1}^{d} \sum_{s_i \in S1} \sum_{s_j \in S2} K_m(s_i, s_j) \quad (1)$$

$$K_m(s_i, s_j) = \begin{cases} val(s_i, s_j) & \text{if} \quad m = 1 \\ K'_{m-1}(s_i, s_j) \cdot val(s_i, s_j) & \text{otherwise} \end{cases} \quad (2)$$

ここで，s_i は，S1 の i 番目のノードを指し，s_j は，S2 の j 番目のノードを指す．いま，s_i に含まれる単語を t_i，それに対応する意味ラベル集合を M_i，s_j に

含まれる単語を t_j，それに対応する意味ラベル集合を M_j とすると $val(s_i, s_j)$ は以下の式で定義される．

$$val(s_i, s_j) = v + |M_i \cap M_j| \quad (3)$$

ここで，v は以下の式で定義される．

$$v = \begin{cases} 1 & \text{if} \quad t_i = t_j \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

また，$K'_m(s_i, s_j)$ は以下の式で定義される．

$$K'_m(s_i, s_j) = \begin{cases} 0 & \text{if} \quad j = 1 \\ \lambda K'_m(s_i, s_{j-1}) + K''_m(s_i, s_{j-1}) & \text{otherwise} \end{cases} \quad (5)$$

さらに，$K''_m(s_i, s_j)$ は以下の式で定義される．

$$K''_m(s_i, s_j) = \begin{cases} 0 & \text{if} \quad i = 1 \\ \lambda K''_m(s_{i-1}, s_j) + K_m(s_{i-1}, s_j) & \text{otherwise} \end{cases} \quad (6)$$

なお，類似度としてカーネルの値を 0〜1 の間に収めるため，下記の式で正規化を行う．

$$\mathrm{Sim}^d_{\mathrm{esk}}(S1, S2) = \frac{\mathrm{ESK}^d(S1, S2)}{\sqrt{\mathrm{ESK}^d(S1, S1)\,\mathrm{ESK}^d(S2, S2)}}. \quad (7)$$

3.2 ESK を用いた自動評価法

いま，$\mathcal{C}$ を ℓ 文からなるシステム要約とし，$\mathcal{R}$ を m 文からなる参照要約とする．また $\mathcal{C}$ 中の文を c_i とし，$\mathcal{R}$ 中の文を r_j とする．このとき，下記の精度重視の指標 $P^d_{\mathrm{esk}}(\mathcal{C}, \mathcal{R})$ と再現率重視の指標 $R^d_{\mathrm{esk}}(\mathcal{C}, \mathcal{R})$ を定義する．

$$P^d_{\mathrm{esk}}(\mathcal{C}, \mathcal{R}) = \frac{1}{\ell} \sum_{i=1}^{\ell} \max_{1 \le j \le m} \mathrm{Sim}^d_{\mathrm{esk}}(c_i, r_j) \quad (8)$$

$$R^d_{\mathrm{esk}}(\mathcal{C}, \mathcal{R}) = \frac{1}{m} \sum_{j=1}^{m} \max_{1 \le i \le \ell} \mathrm{Sim}^d_{\mathrm{esk}}(c_i, r_j) \quad (9)$$

最終的に，上記スコアの重み付き調和平均を評価指標として定義する．

$$F^d_{\mathrm{esk}}(\mathcal{C}, \mathcal{R}) = \frac{(1+\beta^2) \times R_{\mathrm{esk}}(\mathcal{C}, \mathcal{R}) \times P_{\mathrm{esk}}(\mathcal{C}, \mathcal{R})}{R_{\mathrm{esk}}(\mathcal{C}, \mathcal{R}) + \beta^2 \times P_{\mathrm{esk}}(\mathcal{C}, \mathcal{R})} \quad (10)$$

ここで，β は，R^d_{esk} と P^d_{esk} のどちらを優先するかを調整するパラメータである．ここで，システムが参照要約に含まれる 1 文を繰り返すだけの冗長な要約を出力したときに，P^d_{esk} は 1 となり過大評価される．よって，β を大きくとって R^d_{esk} を重視すべきであろう．

3.3 複数参照の場合の拡張

次に，複数の参照要約が与えられた場合への拡張法について説明する．

いま，R を参照要約の集合，つまり，$R = \{\mathcal{R}_1, \cdots, \mathcal{R}_n\}$，が与えられたとする．このとき，システム要約のスコアは，各参照要約に対して求めたスコアの平均として，下記の式で定義する．

$$F^{\mathrm{avg}}_{\mathrm{esk}}(\mathcal{C}, R) = \frac{1}{n} \sum_{i=1}^{n} F_{\mathrm{esk}}(\mathcal{C}, \mathcal{R}_i) \quad (11)$$

4. 評価実験

提案手法の有効性を確認するため，TSC-3 のデータを用いて評価実験を行った．以降，データの詳細，比較した評価法，実験結果を詳述する．

4.1 TSC-3 データ

TSC-3 は，NTCIR プロジェクトの一環として，2004 年に開催された複数文書要約システムの評価型ワークショップである．システムは，あるトピック（出来事）に関連する一連の文書セットを入力とし，文書セットの総文字数に対して 5%，10%の長さの要約を出力する．以降，前者を short，後者を long と呼ぶ．トピック数は 30 で，参加システム数は 10 である．うち 1 つはオーガナイザが用意したベースラインシステムである．詳細については，文献 4) を参照されたい．

TSC-3 では，以下の手順で人間による主観評価が行われた．

Step 1 参照要約中のそれぞれの文 r_j（$\in \mathcal{R}$）に対して以下の Step 2 と Step 3 を適用する．

Step 2 評価者は，文 r_j に対して，システム要約から最も関連する文集合 S_j を抽出する．

Step 3 評価者は，S_j が r_j の情報をどの程度包含しているかという観点から $0, 0.1, \cdots, 1.0$. の 11 段階（1.0 なら S_j は r_j の情報をすべて含む）で評価を行う．この値を $e(r_j, S_j)$ と表す．

Step 4 システム要約 $\mathcal{C}$ の参照要約 $\mathcal{R}$ に対する主観的スコアを $H(\mathcal{R}, \mathcal{C}) = \sum_j e(r_j, S_j)/|\mathcal{R}|$ で求める．

各システムの主観的スコアは，すべてのトピックに対し，上記手続きを適用した後，トピック数で平均したもの $\sum_{t=1}^{30} H(\mathcal{R}_t, \mathcal{C}_t)/30$ で与える．また，複数の参照要約が利用可能な場合には，各参照要約に対する $H(\mathcal{R}, \mathcal{C})$ の平均値を用いる．

4.2 被験者による評価の信頼性

TSC-3 のフォーマルランでは，30 トピックを 6 ト

表 2 各データセットと被験者の関係
Table 2 The relationship between topics and reference summary creators, *i.e.*, human assessors.

topic-ID	$\mathcal{D}_1$	$\mathcal{D}_2$	$\mathcal{D}_3$	$\mathcal{D}_4$	$\mathcal{D}_5$
1 – 6	A	E	D	C	B
7 – 12	B	A	E	D	C
13 – 18	C	B	A	E	D
19 – 24	D	C	B	A	E
25 – 30	E	D	C	B	A

表 3 各データセットにおける人間の評価結果
Table 3 Evaluation results by human judgments on each data set.

short

	sys1	sys2	sys3	sys4	sys5	sys6	sys7	sys8	sys9	sys10
$\mathcal{D}_1$	.319	.215	.236	.318	.290	.365	.271	.280	.151	.273
$\mathcal{D}_2$	.304	.213	.229	.287	.290	.311	.248	.255	.146	.238
$\mathcal{D}_3$	.302	.204	.264	.323	.280	.299	.290	.282	.149	.248
$\mathcal{D}_4$	.294	.208	.249	.316	.300	.300	.305	.282	.159	.268
$\mathcal{D}_5$	.304	.212	.243	.336	.286	.316	.310	.329	.147	.251

long

	sys1	sys2	sys3	sys4	sys5	sys6	sys7	sys8	sys9	sys10
$\mathcal{D}_1$	.298	.221	.311	.322	.330	.392	.273	.300	.261	.278
$\mathcal{D}_2$	.285	.185	.290	.298	.290	.319	.272	.275	.230	.259
$\mathcal{D}_3$	.307	.245	.313	.339	.324	.356	.299	.336	.234	.277
$\mathcal{D}_4$	.316	.221	.313	.321	.330	.322	.304	.306	.265	.277
$\mathcal{D}_5$	.328	.243	.300	.343	.337	.334	.308	.330	.251	.293

ピックずつ 5 つのトピックセットに分け，それぞれのトピックセットに対し，1 名の被験者☆があらかじめ参照要約を作成しておき，それに基づきすべてのシステム要約を評価した．トピックセットが 5 つなので，被験者は 5 名（A，B，C，D，E）である．

TSC-3 の評価では，1 つのシステム要約に対して，1 名の被験者しか評価を行っていない．信頼性を向上させるため，我々はトピックセットと被験者の組を変化させ，各トピックに対して，異なる 5 名の被験者が参照要約の作成とシステム評価を行うように追加実験を行った（**表 2** を参照）．たとえば，$\mathcal{D}_2$ では，トピック 1〜6 に対して被験者 E が参照要約を作成し，すべてのシステムの評価を行った．以降，これをデータセットと呼ぶ．なお，TSC-3 のフォーマルランにおけるデータセットは $\mathcal{D}_1$ である．さらに，すべてのトピックに対し，A〜E の 5 名の平均点をシステムの評価結果としたデータセット $\mathcal{D}_{\text{avg}}$ も作成した．

各データセットにおけるシステムスコアを**表 3** に示す．システムスコア間のピアソンの積率相関係数（以下，r）とスペアマンの順位相関係数（以下，ρ）を**表 4** に示す．表 3 より，システムスコアはデータセット間で大きな違いはなく比較的安定していることが分かる．

☆ 被験者はすべて記者を経験したことのある人間である．

表 4 各データセット間の相関
Table 4 Correlations between human judgments.

ピアソンの積率相関係数　スペアマンの順位相関係数

short

	$\mathcal{D}_1$	$\mathcal{D}_2$	$\mathcal{D}_3$	$\mathcal{D}_4$	$\mathcal{D}_5$	$\mathcal{D}_1$	$\mathcal{D}_2$	$\mathcal{D}_3$	$\mathcal{D}_4$	$\mathcal{D}_5$
$\mathcal{D}_1$	1.00	.968	.903	.902	.888	1.00	.976	.842	.697	.758
$\mathcal{D}_2$	—	1.00	.916	.910	.878	—	1.00	.830	.733	.733
$\mathcal{D}_3$	—	—	1.00	.972	.962	—	—	1.00	.842	.879
$\mathcal{D}_4$	—	—	—	1.00	.954	—	—	—	1.00	.818
$\mathcal{D}_5$	—	—	—	—	1.00	—	—	—	—	1.00

long

	$\mathcal{D}_1$	$\mathcal{D}_2$	$\mathcal{D}_3$	$\mathcal{D}_4$	$\mathcal{D}_5$	$\mathcal{D}_1$	$\mathcal{D}_2$	$\mathcal{D}_3$	$\mathcal{D}_4$	$\mathcal{D}_5$
$\mathcal{D}_1$	1.00	.908	.864	.822	.786	1.00	.964	.915	.939	.855
$\mathcal{D}_2$	—	1.00	.896	.963	.903	—	1.00	.915	.952	.879
$\mathcal{D}_3$	—	—	1.00	.862	.938	—	—	1.00	.842	.891
$\mathcal{D}_4$	—	—	—	1.00	.923	—	—	—	1.00	.903
$\mathcal{D}_5$	—	—	—	—	1.00	—	—	—	—	1.00

表 4 より，r に関しては short，long ともに高い相関であり，ρ に関しては，short における $\mathcal{D}_1$，$\mathcal{D}_4$ 間の相関が例外的にやや低いが，全体的には，r と同様高い相関である．

さらに，データセット間でシステムの順位がどの程度一致しているかを，ケンドールの一致度係数 W を用いて調べたところ，short で 0.849，long で 0.924 という高い一致であった．

以上より，W が高いこと，r，ρ も十分に高いことから，トピックセットと被験者の組を変更したことの効果は小さく，信頼性の高いデータセットであることが分かる．これは，DUC における追加実験の結果[3)]ともよく合致しており，同じ背景を持つ被験者であれば，要約システムを評価するという観点からは，その評価の差異が小さいことが分かる．

4.3 比較した自動評価法

本稿では，ESK とよく似たカーネルである WSK と ROUGE の各バリエーションを比較手法として評価実験を行った．

WSK-based method

ESK の代わりに WSK を用いた自動評価法．式 (8)〜(10) の ESK を WSK で置き換えたもの．ESK におけるノードの属性として単語しか許さない場合が WSK なので，式 (3) を下記に変更すればよい．

$$val(s_i, s_j) = \begin{cases} 1 & \text{if} \quad t_i = t_j \\ 0 & \text{otherwise} \end{cases} \tag{12}$$

ROUGE-N

N グラムの一致率に基づく自動評価法．以下の式で定義される．

$$\text{ROUGE-N}(\mathcal{C},\mathcal{R}) = \frac{\text{count}_{\text{N-gram}}(\mathcal{C},\mathcal{R})}{\#\text{ of N-grams} \in \mathcal{R}} \quad (13)$$

ここで，$\text{count}_{\text{N-gram}}(\mathcal{C},\mathcal{R})$ は，$\mathcal{C}$ と $\mathcal{R}$ の間で一致する N グラムの数を返す関数である．

ROUGE-S

ROUGE-S は ROUGE-2 の拡張であり，下記の式で定義される[8]．

$$\text{ROUGE-S}(\mathcal{C},\mathcal{R}) = \frac{(1+\beta^2)\times R_{\text{skip2}}(\mathcal{C},\mathcal{R})\times P_{\text{skip2}}(\mathcal{C},\mathcal{R})}{R_{\text{skip2}}(\mathcal{C},\mathcal{R})+\beta^2 P_{\text{skip2}}(\mathcal{C},\mathcal{R})} \quad (14)$$

R_{skip2}，P_{skip2} は以下の式で定義される．

$$R_{\text{skip2}}(\mathcal{C},\mathcal{R}) = \frac{\text{Skip2}(\mathcal{C},\mathcal{R})}{U+V} \quad (15)$$

$$P_{\text{skip2}}(\mathcal{C},\mathcal{R}) = \frac{\text{Skip2}(\mathcal{C},\mathcal{R})}{W+X} \quad (16)$$

ここで，Skip2 は，$\mathcal{R}$ と $\mathcal{C}$ に共通して出現するバイグラムとスキップバイグラムの数を返す関数である．U，V は，それぞれ，$\mathcal{R}$ に出現するバイグラムとスキップバイグラムの数であり，W，X は，それぞれ，$\mathcal{C}$ に出現するバイグラムとスキップバイグラムの数である．

ROUGE-SU

ROUGE-SU は，ROUGE-S の拡張であり，バイグラム，スキップバイグラムに加え，ユニグラムも素性とした以下の式で定義される[8]．

$$\text{ROUGE-SU}(\mathcal{C},\mathcal{R}) = \frac{(1+\beta^2)\times R_{\text{su}}(\mathcal{C},\mathcal{R})\times P_{\text{su}}(\mathcal{C},\mathcal{R})}{R_{\text{su}}(\mathcal{C},\mathcal{R})+\beta^2 P_{\text{su}}(\mathcal{C},\mathcal{R})} \quad (17)$$

R_{su}，P_{su} は以下の式で定義される．

$$R_{\text{su}}(\mathcal{C},\mathcal{R}) = \frac{\text{SU}(\mathcal{C},\mathcal{R})}{U+V+Y} \quad (18)$$

$$P_{\text{su}}(\mathcal{C},\mathcal{R}) = \frac{\text{SU}(\mathcal{C},\mathcal{R})}{W+X+Z} \quad (19)$$

ここで，SU は，$\mathcal{R}$ と $\mathcal{C}$ に共通して出現するバイグラム，スキップバイグラム，ユニグラムの数を返す関数である．Y は，$\mathcal{R}$ に出現するユニグラムの数，Z は，$\mathcal{C}$ に出現するユニグラムの数である．

なお，式 (8)，(9)，式 (11)，(12) から分かるが，ROUGE-S や ROUGE-SU では，参照要約とシステム要約の単語数がほぼ同じ場合には調和平均をとる効果がほとんどない．TSC-3 のタスク設定では，参照要約とシステム要約の単語数は近いので，後述の評価実験の際には，ROUGE-S，ROUGE-SU の β を変化させなかった．

ROUGE-L

ROUGE-L は LCS に基づく自動評価法であり，下記の式で定義される[8]．

$$\text{ROUGE-L}(\mathcal{C},\mathcal{R}) = \frac{(1+\beta^2)\times R_{\text{lcs}}(\mathcal{C},\mathcal{R})\times P_{\text{lcs}}(\mathcal{C},\mathcal{R})}{R_{\text{lcs}}(\mathcal{C},\mathcal{R})+\beta^2 P_{\text{lcs}}(\mathcal{C},\mathcal{R})} \quad (20)$$

R_{lcs}，P_{lcs} は以下の式で定義される．

$$R_{\text{lcs}}(\mathcal{C},\mathcal{R}) = \frac{1}{u}\sum_{r_i\in\mathcal{R}} \text{LCS}_{\cup}(r_i,\mathcal{C}) \quad (21)$$

$$P_{\text{lcs}}(\mathcal{C},\mathcal{R}) = \frac{1}{v}\sum_{r_i\in\mathcal{R}} \text{LCS}_{\cup}(r_i,\mathcal{C}) \quad (22)$$

ここで，$\text{LCS}_{\cup}(r_i,\mathcal{C})$ は，参照要約の文 r_i とシステム要約 $\mathcal{C}$ の間のユニオン LCS の長さを返す関数である．また，u と v は，$\mathcal{R}$ と $\mathcal{C}$ に含まれる単語の数を表す．ユニオン LCS の詳細については，文献 8) を参照されたい．ROUGE-L も ROUGE-S，SU と同様に参照要約とシステム要約の単語数がほぼ同じ場合には調和平均をとる効果がほとんどないので，後述の実験では β を変化させなかった．

なお，WSK，ROUGE に対して，複数参照要約を用いる場合は，式 (11) の右辺 F_{esk} をそれぞれの関数に置き換えればよい．

4.4 評価指標

各自動評価法をピアソンの積率相関係数（r），スペアマンの順位相関係数（ρ）を用いて人間の評価結果とどの程度の相関があるかで評価した．まず，自動評価法を用いてシステムスコアのベクトル $\mathbf{x} = (x_1, x_2, \cdots, x_i, \cdots, x_{10})$ を作成する．システム数が 10 なのでベクトルの次元は 10 である．ここで，i 番目のシステムのスコアは，$x_i = 1/30\sum_{t=1}^{30} f(\mathcal{R}_t, \mathcal{C}_{i,t})$ となる．$\mathcal{R}_t$ は t 番目のトピックにおける参照要約を表し，$\mathcal{C}_{i,t}$ は i 番目のシステムの t 番目のトピックにおける要約を表す．また，f は ROUGE，WSK，提案手法のいずれかの自動評価法を表す．次に，同様にしてベクトル $\mathbf{y} = (y_1, y_2, \cdots, y_i, \cdots, y_{10})$ を人間の評価結果を用いて作成する．ここで，i 番目のシステムに対して人間が与えたスコアは，$y_i = 1/30\sum_{t=1}^{30} H(\mathcal{R}_t, \mathcal{C}_{i,t})$ となり，主観的システムスコアと一致する．なお，複数参照要約を用いる場合には，ベクトルを作成する際の関数 f，H をそれぞれ，f^{avg}，H^{avg} に置き換えればよい．最終的に，$\mathbf{x}$ と $\mathbf{y}$ の間の r，ρ を計算する．

4.5 実験結果

表 5，**表 6** に各データセット（short）に対する r，ρ とそれらの全データセットに対する平均値を示し，**表 7**，**表 8** に各データセット（long）に対する r，ρ とそれらの全データセットに対する平均値を示す．

表 5 ピアソンの積率相関係数による評価結果（short）
Table 5 Results obtained with Pearson's correlation coefficient (short).

	$\mathcal{D}_1$		$\mathcal{D}_2$		$\mathcal{D}_3$		$\mathcal{D}_4$		$\mathcal{D}_5$		$\mathcal{D}_{\text{avg}}$		Average	
	case	stop	case	stop	case	stop	case	stop	case	stop	case	stop	case	stop
ROUGE-1	.965	.884	.931	.888	.952	.903	.937	.879	.967	.894	.968	.901	.953	.891
ROUGE-2	.943	.960	.836	.880	.889	.932	.861	.906	.907	.949	.907	.946	.890	.929
ROUGE-3	.906	.936	.759	.814	.834	.882	.786	.846	.859	.906	.858	.905	.834	.882
ROUGE-4	.877	.914	.725	.752	.810	.844	.729	.793	.827	.861	.836	.875	.801	.840
ROUGE-L	.777	.919	.683	.789	.860	.882	.867	.875	.908	.897	.884	.908	.830	.878
ROUGE-S*	.914	.934	.887	.805	.955	.890	.937	.872	.947	.910	.943	.893	.930	.884
ROUGE-S9	.938	.926	.890	.765	.942	.844	.906	.789	.941	.875	.949	.860	.928	.843
ROUGE-S4	.945	.930	.865	.772	.918	.852	.889	.810	.934	.885	.936	.871	.914	.853
ROUGE-SU*	.914	.934	.888	.805	.955	.890	.938	.872	.947	.910	.943	.893	.931	.884
ROUGE-SU9	.935	.929	.899	.783	.949	.854	.917	.808	.945	.887	.953	.871	.933	.855
ROUGE-SU4	.943	.936	.891	.802	.939	.869	.917	.839	.948	.902	.952	.889	.932	.878
$F_{\text{esk}}^{d=2}(\beta=2)$	.942		.927		.952		.921		.961		.963		.944	
$F_{\text{esk}}^{d=2}(\beta=3)$	.929		.943		.958		.928		.968		.971		.950	
$F_{\text{esk}}^{d=3}(\beta=2)$	.939		.923		.943		.919		.944		.963		.939	
$F_{\text{esk}}^{d=3}(\beta=3)$	.927		.933		.944		.920		.948		.968		.940	
$F_{\text{esk}}^{d=4}(\beta=2)$	.921		.900		.926		.897		.925		.959		.921	
$F_{\text{esk}}^{d=4}(\beta=3)$	.909		.900		.924		.888		.927		.960		.918	
$F_{\text{wsk}}^{d=2}(\beta=2)$	.939		.900		.938		.897		.949		.948		.928	
$F_{\text{wsk}}^{d=2}(\beta=3)$	.928		.921		.949		.909		.959		.961		.938	
$F_{\text{wsk}}^{d=3}(\beta=2)$	.938		.902		.928		.886		.943		.947		.924	
$F_{\text{wsk}}^{d=3}(\beta=3)$	.928		.922		.937		.895		.952		.959		.932	
$F_{\text{wsk}}^{d=4}(\beta=2)$	.929		.896		.914		.874		.934		.944		.915	
$F_{\text{wsk}}^{d=4}(\beta=3)$	.918		.915		.920		.879		.942		.955		.921	

表 6 スペアマンの順位相関係数による評価結果（short）
Table 6 Results obtained with Spearman's ranking correlation coefficient (short).

	$\mathcal{D}_1$		$\mathcal{D}_2$		$\mathcal{D}_3$		$\mathcal{D}_4$		$\mathcal{D}_5$		$\mathcal{D}_{\text{avg}}$		Average	
	case	stop	case	stop	case	stop	case	stop	case	stop	case	stop	case	stop
ROUGE-1	.988	.964	.842	.891	.952	.915	.842	.855	.915	.806	.915	.915	.909	.891
ROUGE-2	.927	.976	.770	.794	.927	.964	.855	.842	.891	.903	.842	.867	.867	.891
ROUGE-3	.879	.927	.588	.697	.806	.927	.818	.818	.806	.891	.721	.842	.770	.850
ROUGE-4	.818	.879	.721	.697	.806	.891	.746	.746	.721	.746	.709	.721	.753	.780
ROUGE-L	.830	.927	.600	.661	.818	.927	.818	.806	.952	.915	.842	.830	.810	.844
ROUGE-S*	.939	.939	.818	.673	.915	.855	.818	.794	.891	.879	.903	.782	.881	.820
ROUGE-S9	.964	.879	.745	.600	.927	.758	.794	.721	.867	.758	.879	.697	.863	.736
ROUGE-S4	.988	.879	.745	.600	.927	.794	.770	.721	.891	.806	.867	.697	.865	.749
ROUGE-SU*	.939	.939	.818	.673	.915	.855	.818	.794	.891	.903	.903	.782	.881	.824
ROUGE-SU9	.952	.879	.745	.600	.927	.758	.794	.721	.915	.806	.879	.697	.869	.744
ROUGE-SU4	.964	.891	.794	.600	.964	.867	.794	.794	.915	.879	.879	.758	.885	.798
$F_{\text{esk}}^{d=2}(\beta=2)$	.952		.879		.903		.855		.927		.879		.899	
$F_{\text{esk}}^{d=2}(\beta=3)$	.952		.915		.903		.891		.915		.952		.921	
$F_{\text{esk}}^{d=3}(\beta=2)$	.964		.867		.952		.867		.927		.927		.917	
$F_{\text{esk}}^{d=3}(\beta=3)$	.964		.891		.952		.915		.927		.927		.929	
$F_{\text{esk}}^{d=4}(\beta=2)$	.927		.830		.952		.867		.915		.927		.903	
$F_{\text{esk}}^{d=4}(\beta=3)$	.927		.842		.988		.842		.927		.927		.909	
$F_{\text{wsk}}^{d=2}(\beta=2)$	.976		.794		.855		.830		.903		.867		.871	
$F_{\text{wsk}}^{d=2}(\beta=3)$	.952		.842		.867		.830		.915		.891		.883	
$F_{\text{wsk}}^{d=3}(\beta=2)$	.976		.794		.867		.818		.903		.879		.873	
$F_{\text{wsk}}^{d=3}(\beta=3)$	.976		.879		.915		.855		.903		.879		.901	
$F_{\text{wsk}}^{d=4}(\beta=2)$	.964		.794		.952		.818		.903		.867		.883	
$F_{\text{wsk}}^{d=4}(\beta=3)$	.964		.867		.952		.855		.903		.927		.911	

表 7 ピアソンの積率相関係数による評価結果（long）
Table 7 Results obtained with Pearson's correlation coefficient (long).

	$\mathcal{D}_1$		$\mathcal{D}_2$		$\mathcal{D}_3$		$\mathcal{D}_4$		$\mathcal{D}_5$		$\mathcal{D}_{\text{avg}}$		Average	
	case	stop	case	stop	case	stop	case	stop	case	stop	case	stop	case	stop
ROUGE-1	.906	.876	.919	.916	.945	.798	.897	.891	.935	.892	.931	.932	.922	.884
ROUGE-2	.886	.930	.788	.941	.959	.976	.834	.616	.811	.938	.859	.951	.856	.892
ROUGE-3	.873	.909	.717	.849	.961	.975	.826	.431	.763	.848	.837	.902	.892	.819
ROUGE-4	.850	.890	.651	.787	.953	.965	.836	.292	.707	.774	.815	.873	.802	.763
ROUGE-L	.840	.917	.812	.861	.875	.946	.829	.847	.832	.901	.870	.932	.843	.901
ROUGE-S*	.864	.811	.954	.743	.932	.893	.547	.707	.989	.855	.944	.814	.872	.804
ROUGE-S9	.904	.829	.948	.705	.955	.924	.586	.701	.953	.782	.950	.805	.883	.791
ROUGE-S4	.921	.868	.928	.730	.971	.947	.620	.785	.944	.793	.950	.844	.889	.828
ROUGE-SU*	.863	.812	.954	.744	.932	.894	.547	.709	.989	.856	.944	.815	.872	.805
ROUGE-SU9	.903	.840	.951	.735	.953	.932	.617	.730	.960	.802	.953	.824	.890	.810
ROUGE-SU4	.920	.876	.945	.778	.966	.951	.663	.814	.960	.826	.959	.865	.902	.852
$F_{\text{esk}}^{d=2}(\beta=2)$	.941		.957		.987		.967		.945		.978		.962	
$F_{\text{esk}}^{d=2}(\beta=3)$	.939		.962		.959		.959		.952		.974		.958	
$F_{\text{esk}}^{d=3}(\beta=2)$	.926		.954		.971		.953		.930		.975		.951	
$F_{\text{esk}}^{d=3}(\beta=3)$	.920		.947		.938		.904		.928		.957		.932	
$F_{\text{esk}}^{d=4}(\beta=2)$	.900		.932		.949		.890		.906		.962		.923	
$F_{\text{esk}}^{d=4}(\beta=3)$	.892		.921		.911		.819		.897		.936		.896	
$F_{\text{wsk}}^{d=2}(\beta=2)$	.931		.923		.983		.936		.938		.960		.945	
$F_{\text{wsk}}^{d=2}(\beta=3)$	.932		.939		.967		.950		.950		.967		.951	
$F_{\text{wsk}}^{d=3}(\beta=2)$	.924		.921		.977		.934		.923		.962		.940	
$F_{\text{wsk}}^{d=3}(\beta=3)$	.920		.929		.953		.919		.931		.957		.935	
$F_{\text{wsk}}^{d=4}(\beta=2)$	.910		.913		.962		.908		.903		.955		.925	
$F_{\text{wsk}}^{d=4}(\beta=3)$	.903		.913		.930		.866		.905		.941		.910	

表 8 スペアマンの順位相関係数による評価結果（long）
Table 8 Results obtained with Spearman's ranking correlation coefficient (long).

	$\mathcal{D}_1$		$\mathcal{D}_2$		$\mathcal{D}_3$		$\mathcal{D}_4$		$\mathcal{D}_5$		$\mathcal{D}_{\text{avg}}$		Average	
	case	stop	case	stop	case	stop	case	stop	case	stop	case	stop	case	stop
ROUGE-1	.818	.830	.903	.806	.891	.867	.867	.855	.818	.770	.867	.891	.861	.837
ROUGE-2	.721	.891	.721	.855	.927	.952	.794	.648	.745	.794	.867	.939	.796	.847
ROUGE-3	.758	.842	.636	.745	.964	.964	.806	.564	.673	.709	.794	.927	.772	.792
ROUGE-4	.685	.794	.564	.612	.952	.952	.830	.455	.600	.576	.697	.842	.721	.705
ROUGE-L	.770	.842	.612	.576	.891	.903	.709	.636	.661	.782	.915	.915	.760	.776
ROUGE-S*	.879	.770	.818	.636	.939	.867	.553	.697	.879	.770	.903	.879	.829	.770
ROUGE-S9	.806	.745	.758	.576	.939	.891	.564	.612	.758	.697	.939	.891	.794	.735
ROUGE-S4	.855	.758	.794	.576	.952	.915	.612	.709	.745	.612	.927	.891	.814	.744
ROUGE-SU*	.879	.770	.818	.636	.939	.867	.527	.697	.879	.770	.903	.879	.824	.770
ROUGE-SU9	.806	.758	.806	.576	.939	.891	.564	.673	.758	.709	.952	.891	.804	.750
ROUGE-SU4	.842	.709	.770	.576	.976	.915	.733	.770	.745	.697	.939	.891	.834	.760
$F_{\text{esk}}^{d=2}(\beta=2)$	.842		.927		.976		.903		.818		.952		.903	
$F_{\text{esk}}^{d=2}(\beta=3)$	.855		.903		.976		.903		.855		.952		.907	
$F_{\text{esk}}^{d=3}(\beta=2)$	.818		.927		.988		.879		.830		.964		.901	
$F_{\text{esk}}^{d=3}(\beta=3)$	.758		.903		.976		.709		.842		.952		.857	
$F_{\text{esk}}^{d=4}(\beta=2)$	.661		.903		.964		.733		.782		.927		.828	
$F_{\text{esk}}^{d=4}(\beta=3)$	.588		.903		.976		.673		.758		.879		.796	
$F_{\text{wsk}}^{d=2}(\beta=2)$	.818		.867		.964		.806		.770		.952		.863	
$F_{\text{wsk}}^{d=2}(\beta=3)$	.818		.867		.939		.794		.830		.915		.861	
$F_{\text{wsk}}^{d=3}(\beta=2)$	.806		.855		.952		.733		.746		.952		.841	
$F_{\text{wsk}}^{d=3}(\beta=3)$	.806		.818		.952		.794		.806		.915		.849	
$F_{\text{wsk}}^{d=4}(\beta=2)$	.806		.855		.952		.697		.770		.915		.833	
$F_{\text{wsk}}^{d=4}(\beta=3)$	.746		.855		.952		.770		.782		.879		.831	

ROUGEに対しては，名詞，動詞，形容詞，未知語のみを用いてスコアを計算した場合（stop），とすべての単語を用いた場合（case）の評価を行った．ROUGE-S，ROUGE-SUに関しては文献8)に従い，スキップする単語数を4個までに制限する場合，9個までに制限する場合，制限なしの場合のそれぞれを評価した．また，提案手法とWSKを用いた手法に対しては，すべての単語を用い，単語の組合せ数 d を2〜4まで変化させ，調和平均のパラメータ β は2と3の場合の評価を行った．減衰パラメータ λ は，0.5に設定した．なお，λ と β が提案手法に与える影響については，5.3節で考察する．

5. 考　　察

5.1 各手法との性能比較

まず，shortにおける r に関して議論する．表5より，提案手法は，$d=2$，$\beta=3$ の場合に安定して高い相関を得ている．d を増やすと相関は低くなる傾向にあり，平均的には $\beta=3$ の方が $\beta=2$ の場合よりも良い．WSKを用いた手法と比較すると，$d=2$，3の場合に差が大きく，$d=4$ では差がやや小さくなり，WSKを用いた手法が提案手法を上回る場合がある．これは，提案手法が単語の意味ラベルを素性としているため，$d=4$ の場合には素性数が膨大になることで性能の劣化を招いていると考える．

一方，ROUGEに関しては，ROUGE-1（case）の成績が最も良い．提案手法との差も小さく，$\mathcal{D}_1$，$\mathcal{D}_4$ では提案手法より勝っており，全データの平均も提案手法より良い．次いで，ROUGE-2，ROUGE-S，ROUGE-SUが同程度の成績で良く，ROUGE-3，4，Lの成績はそれらよりもやや落ちる．ROUGE-S，SUにおけるスキップの制限に関しては，データセットによって最適値が異なっているため，一概にはいえないが，全データの平均を見る限りは，大きな違いはない．なお，ROUGE-1，ROUGE-S，ROUGE-SUの場合には，caseの場合が，それら以外では，stopの方が良い傾向にある．

続いてshortにおける ρ について議論する．表6より，全体的に ρ は r よりも低い値をとる傾向にあることが分かる．表5と同じく，提案手法はおおむねWSKを用いた手法よりも良いが，データセットによっては，ROUGE-1（case）よりも劣る場合がある．ただし，表5ほどの差はなく，全データの平均では提案手法が最も良い．さらに，d を増やしていった場合の相関係数の変化が表5とは異なっている．データセットにもよるが，$d=3$ が最も良い場合が多い．

次にlongにおける r について議論する．表7を見るとROUGEが表5と比較して成績が大きく下がっているのに対して，カーネル関数を用いた手法は特に $d=2$，3の場合に成績が向上していることが分かる．さらに，提案手法は，全データセットでROUGE-1（case）に勝っておりその有効性がより明確である．また，表5とは異なり，$\beta=2$ が $\beta=3$ よりも全体的に良い結果を得る傾向にある．WSKを用いた手法との比較では $d=4$ 以外では，提案手法の方が良い成績である．ROUGEに関しては，ROUGE-L（stop）の成績が大きく向上しており，ROUGE-SU(4)（case）とほぼ同等となっている．表5とは異なり，ROUGE-SUがROUGE-Sよりやや良い成績であり，双方ともにスキップする単語数を少なく設定した方が成績が向上する傾向にある．

続いてlongにおける ρ について議論する．表8より，shortの場合と同様 ρ は，r よりも低い値をとる傾向にある．全データセットの平均を見た場合，ROUGEではROUGE-1（case）の成績が最も良く0.861であるのに対して，提案手法は $d=2$，$\beta=3$ の場合には0.907とその差は大きい．$d=2$ の場合には，$\beta=2$ と $\beta=3$ の差は小さいが，$d=3$，4の場合には $\beta=2$ の方が成績が良い．WSKを用いた手法との比較では，どの d においてもおおむね提案手法の方が成績が良い．また，表6と同じく，$d=3$ が最も良い成績である場合がある．

以上より，WSK，ESKといったコンボリューションカーネルを用いることで，全体的には，ROUGEよりも良い結果を得る傾向にあることが分かる．特にROUGEがshortでは良い成績であるが，longではやや悪い成績であるのに対し，WSK，ESKは長さによらず安定して良い成績である．また，WSKがROUGE-S，SUよりも良いことから，スキップNグラムに対しては，通常のNグラムよりも重みを小さくした方が効果的であることが分かる．さらに，提案手法が $d=2$，3において，WSKよりも良い成績を得る傾向にあることから，語の意味ラベルを用いたことの有効性も分かる．ただし，先にも述べたとおり，$d=4$ の場合には，意味ラベルを用いることによって素性数が爆発するため成績が悪くなる傾向にある．

5.2 ROUGE-1の問題点

今回の実験において，ROUGEの中で最も成績が良かったのはROUGE-1であるが，これに関しては，スコアをだますことが容易であるという問題がある．ROUGE-1では語順をまったく考慮しないので，文書セット中でIDFが高い単語，固有名詞，出現頻度の高

表 9 自動評価法がシステムに与えたスコアの平均
Table 9 Average scores assigned by automatic evaluation methods.

	ROUGE-1	ROUGE-2	ROUGE-3	$F_{\mathrm{esk}}^{d=2}(\beta=2)$
sys1	.4335	.2001	.1168	.3328
sys2	.3583	.1589	.0906	.2830
sys3	.3540	.1298	.0601	.2917
sys4	.4481	.2091	.1188	.3434
sys5	.4092	.1780	.0969	.3082
sys6	.4061	.1685	.0890	.3223
sys7	.4254	.2019	.1193	.3127
sys8	.4070	.1912	.1114	.3251
sys9	.3118	.1058	.0509	.2417
sys10	.3591	.1340	.0667	.2952
sys11	.3550	.0054	.0002	.0841

い助詞などを並べることで，ある程度のスコアを稼ぐことが容易に予測できる．実際に，単語 t の重要度を $\mathrm{TF}(t, DS)\cdot\mathrm{IDF}(t)$ ☆で求め，その値の高いものから順に指定された文字数を満たすまで単語を出力するシステムを sys11 として作成し，各データセット $\mathcal{D}_1, \cdots, \mathcal{D}_5$ について，ROUGE-1，2，3，$F_{\mathrm{esk}}^{d=2}$ $(\beta=2)$ のスコアを計算した．また，同様にして TSC-3 参加システムのスコアも計算した．5 つのデータセットに対するスコアの平均値を**表 9** に示す．なお，紙面の都合上 short の結果のみ掲載する．表 9 より，ROUGE-1 が sys11 に対して非常に高いスコアを与えていることが分かる．そのスコアは，TSC-3 参加の中程度のシステムとほぼ同等である．単なる語の羅列であるシステムに対してこうした高いスコアを与えることは，致命的な問題であると考える．これに対して，ROUGE-2，3，$F_{\mathrm{esk}}^{d=2}$ $(\beta=2)$ は，語順を考慮するので，そのスコアは非常に小さく，TSC-3 に参加した最下位のシステムと比較しても十分小さい．よって，語順を考慮した ROUGE や提案手法は ROUGE-1 と比較すると頑健である．

上記より，提案手法や ROUGE-2，3 など語順を考慮する自動評価法は，ROUGE-1 よりも頑健性に優れていることが分かる．

5.3 パラメータの影響

評価実験では，$d=2$，3，4 のそれぞれに対し，λ を 0.5 に固定し，$\beta=2$，3 の場合のみを評価したが，ここでは，λ と β が r と ρ に与える影響を議論する．データセット $\mathcal{D}_1, \ldots, \mathcal{D}_{\mathrm{avg}}$ に対し，λ は，0，0.1，0.2，…，1 まで変化させ，β は，0.5，1，1.5，…，4.5 まで変化させ，提案手法の r と ρ を計算し，平均を求めた．このときの最大値をそれぞれ**表 10**，**表 11** に示し，

☆ $\mathrm{TF}(t, DS)$ は文書セット中での t の出現頻度を表す．

表 10 λ，β を変化させた場合の r の最大値
Table 10 Best scores of r for various values of λ and β.

	short			long		
d	β	λ	r	β	λ	r
2	3	.5	.950	2.5	.3	.963
3	3	.3	.941	2	.2	.955
4	2	.3	.925	2	.1	.938

表 11 λ，β を変化させた場合の ρ の最大値
Table 11 Best scores of ρ for various values of λ and β.

	short			long		
d	β	λ	ρ	β	λ	ρ
2	3	.5	.921	2.5	.6	.911
3	4	.8	.931	2	.3	.901
4	4	.5	.921	2	.1	.877

値の変化をそれぞれ**図 1**，**図 2** に示す．なお，すべての β をグラフに掲載すると煩雑になるため，short に関しては β =1，2，3，4，long に関しては $\beta=1.5$，2.5，3.5，4.5 の場合を掲載する．

表 10 より，r に対する最適パラメータは，short の場合では $d=2$，$\lambda=0.5$，$\beta=3$，long の場合では $d=2$，$\lambda=0.3$，$\beta=2.5$ である．評価実験の際に決定したパラメータとはやや異なるが，表 5，表 7 の「Average」カラムの値と比較すると，その差は小さい．また，表 11 より，ρ に対する最適パラメータは，short の場合では $d=3$，$\lambda=0.8$，$\beta=4$，long の場合では $d=2$，$\lambda=0.6$，$\beta=2.5$ である．これも r の場合と同様，評価実験の際に設定したパラメータとは異なるが，ρ の値の差は小さい．

図 1 より，short の r に関しては，$\beta=2$，3，4 の間の差は小さい．$d=2$ の場合には，$\lambda=0.8$ 付近，$d=3$ の場合には，$\lambda=0.6$ 付近，$d=4$ の場合には，$\lambda=0.4$ 付近で急激に r が下がる傾向にある．ρ については，全体的な傾向は r と似ているが，$\beta=2$ の性能がやや劣っている．ただし，ρ が急激に劣化する λ の値は r の場合ほど異ならない．r，ρ ともに $\lambda=1$ に設定すると相関係数が大きく下がることからもスキップ N グラムの重みを通常の N グラムよりも小さくすることの有効性が分かる．また，r の場合は，$d=2$，3 の $\lambda=0.5$ 付近，ρ の場合は，$d=3$，4 の $\lambda=0.5$ 付近において，相関係数は十分高いことが分かる．

図 2 より，long の r に関しては，$d=2$，3 の場合には $\beta=2.5$ が良く，$d=4$ の場合には，$\beta=1.5$ が良い傾向にある．β を変化させた場合の性能差は short の場合よりも大きい．また，r が急激に劣化する λ に関しては，short とほぼ同様の傾向である．ρ に関し

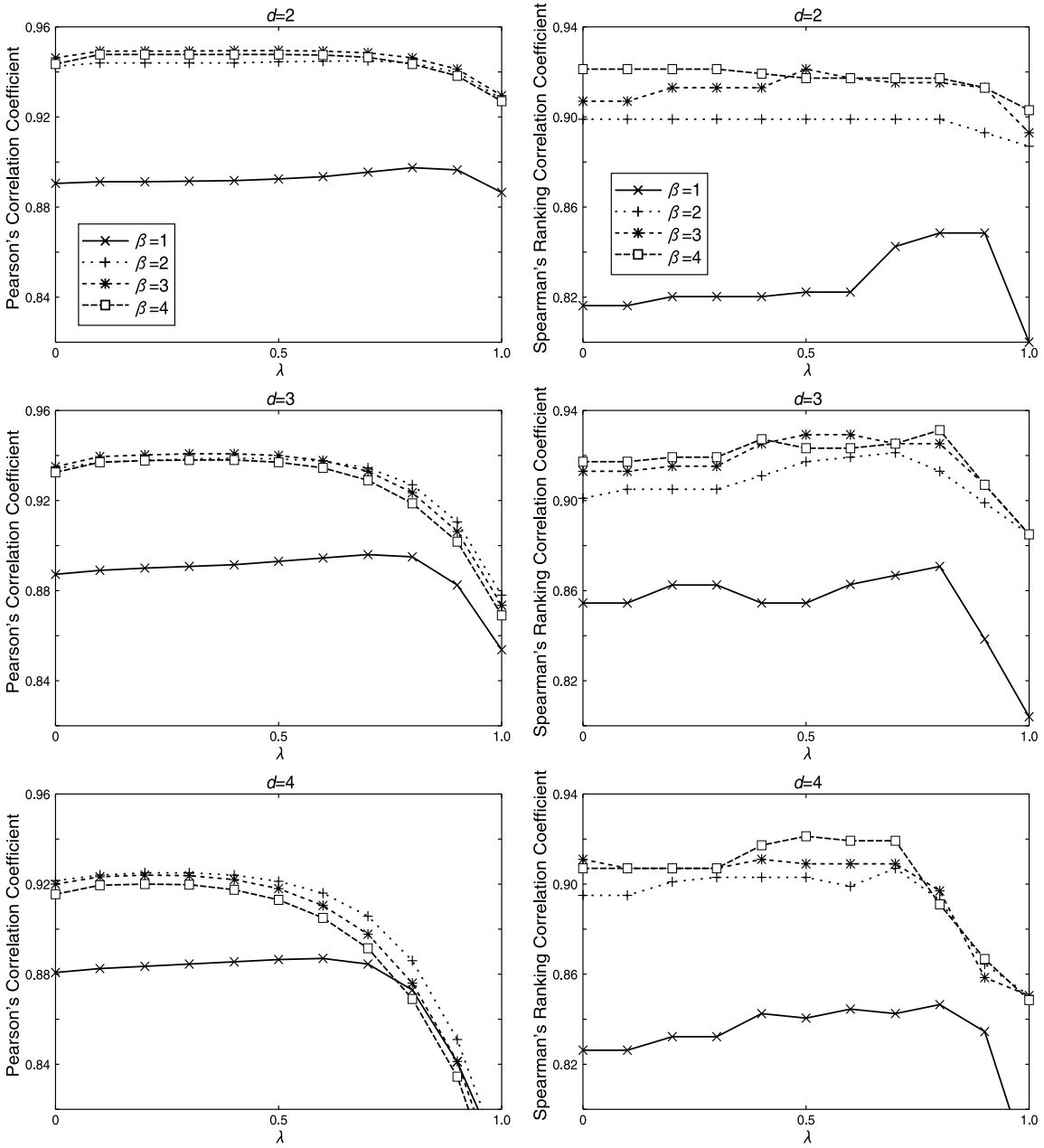

図 1 様々な λ と β に対する相関係数（short）
Fig. 1 Correlation coefficients for various values of β and λ (short).

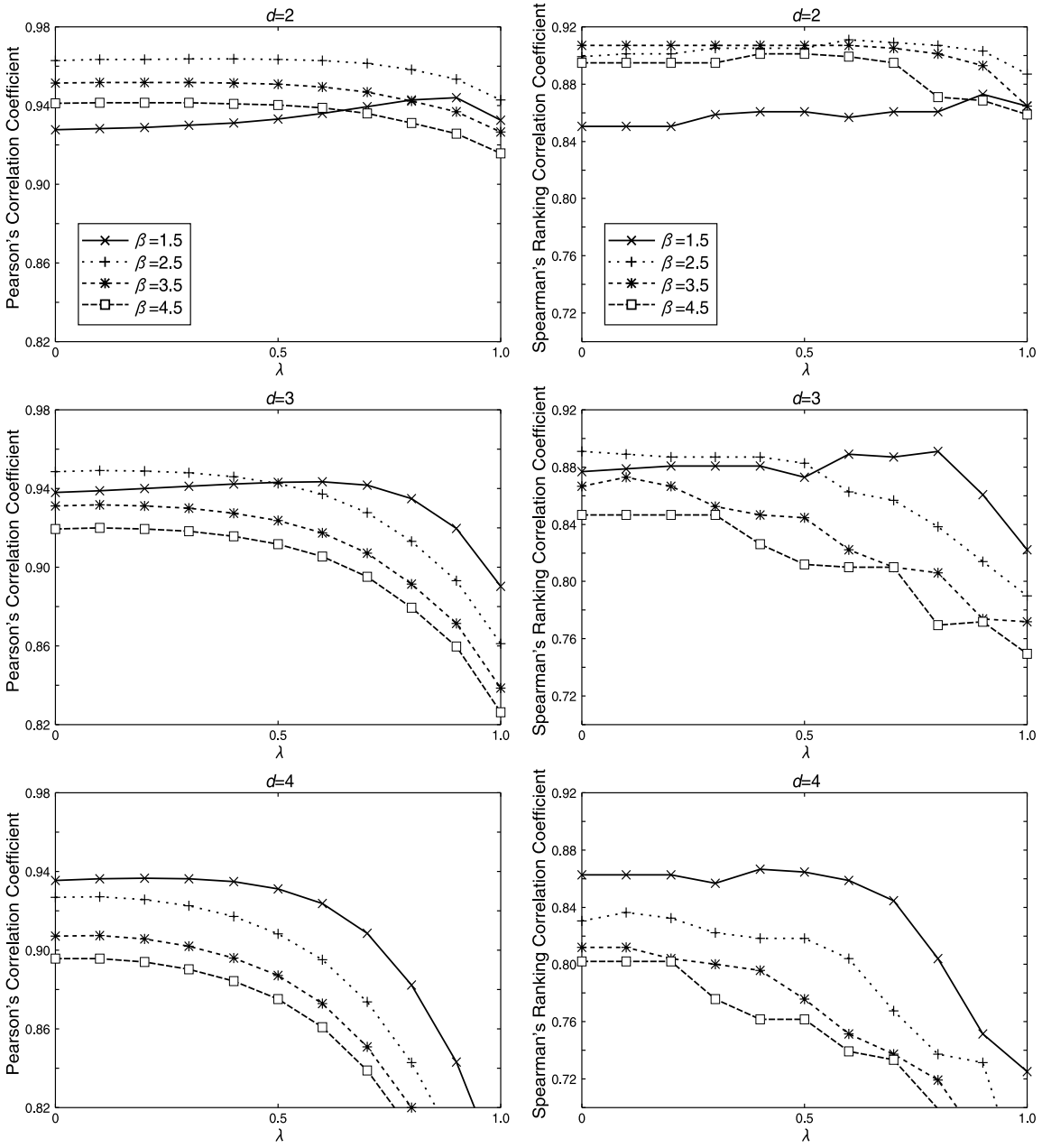

図 2 様々な λ と β に対する相関係数（long）
Fig. 2 Correlation coefficients for various values of β and λ (long).

ては，$d = 2$ の場合には，$\beta = 2.5$，3.5 が良い傾向にあり，$d = 3$，4 の場合には $\beta = 1.5$，2.5 が良い傾向にある．また，d を大きくするに従って，性能が急激に劣化する λ の値が short の場合より小さくなる傾向にある．short の場合と同様，$\lambda = 1$ に設定すると r，ρ の値が大きく下がっており，スキップ N グラムの重みを通常の N グラムよりも小さくすることの有効性が分かる．また，short とは異なり，r，ρ ともに $d = 2$ で最も良い結果が得られる．このとき，$\lambda = 0.5$ 付近の相関係数は十分高い．

以上より，β を 2〜3 に設定し，λ を 0.5 付近に設定すると，最適とは限らないが，比較的安定して良い成績であることが分かった．

6. ま と め

本稿では，拡張ストリングカーネルを用いた要約システムの自動評価法を提案した．TSC-3 のデータを用いて評価実験を行った結果，ピアソンの積率相関係数は平均で 0.95 程度，スペアマンの順位相関係数は平均で 0.92 程度であり，人間の評価結果に対し，非常に高い相関を得た．また，提案手法は，従来より提案されている自動評価法である ROUGE より，参照要約の長さに依存せずに人間の評価結果との間の相関が高いこと，評価指標として頑健であることが分かった．

参 考 文 献

1) Cancedda, N., Gaussier, E., Goutte, C. and Renders, J.-M.: Word Sequence Kernels, *Journal of Machine Learning Research*, Vol.3, No.Feb, pp.1059–1082 (2003).
2) Collins, M. and Duffy, N.: Convolution Kernels for Natural Language, *Proc. Neural Information Processing Systems* (*NIPS2001*) (2001).
3) Harman, D. and Over, P.: The Effects of Human Variation in DUC Summarization Evaluation, *Proc. Workshop on Text Summarization Branches Out, Post Conference Workshop of ACL 2004*, pp.10–17 (2004).
4) Hirao, T., Okumura, M., Fukushima, T. and Nanba, H.: Text Summarization Challenge 3 — Text Summarization Evaluation at NTCIR Workshop 4, *Working Notes of the Fourth NTCIR Workhop Meeting*, pp.407–411 (2004).
5) Hirao, T., Suzuki, J., Isozaki, H. and Maeda, E.: Dependency-based Sentence Alignment for Multiple Document Summarization, *Proc. 20th International Conference on Computational Linguistics*, pp.446–452 (2004).
6) Hori, C., Hori, T. and Furui, S.: Evaluation Methods for Automatic Speech Summarization, *Proc. Eurospeech2003*, pp.2825–2828 (2003).
7) Lin, C.-Y.: Looking for a Good Metrics: ROUGE and its Evaluation, *Proc. 4th NTCIR Workshops* (*open submission*), pp.1–8 (2004).
8) Lin, C.-Y.: ROUGE: A Package for Automatic Evaluation of Summaries, *Proc. Workshop on Text Summarization Branches Out, Post Conference Workshop of ACL 2004*, pp.74–81 (2004).
9) Lin, C.-Y. and Hovy, E.: Automatic Evaluation of Summaries Using N-gram Co-occurrence Statistics, *Proc. 4th Meeting of the North American Chapter of the Association for Computational Linguistics and Human Language Technology*, pp.150–157 (2003).
10) Lin, C.-Y. and Och, F.: Automatic Evaluation of Machine Translation Quality Using Longest Common Subsequence and Skip-Bigram Statistics, *Proc. 42nd Annual Meeting of the Association for Computational Linguistics*, pp.606–613 (2004).
11) Lodhi, H., Saunders, C., Shawe-Taylor, J., Cristianini, N. and Watkins, C.: Text Classification using String Kernel, *Journal of Machine Learning Research*, Vol.2, No.Feb, pp.419–444 (2002).
12) Papineni, K., Roukos, S., Ward, T. and Zhu, W.-J.: BLEU: a Method for Automatic Evaluation of Machine Translation, *Proc. 40th Annual Meeting of the Association for Computational Linguistics*, pp.311–318 (2002).
13) Saggion, H., Radev, D., Teufel, T. and Lam, W.: Meta-Evaluation of Summaries in a Cross-Lingual Environment Using Content-Based Metrics, *Proc. 19th International Conference on Computational Linguistics* (2002).
14) Soricut, R. and Brill, E.: A Unified Framework for Automatic Evaluation using N-gram Co-occurrence Statistics, *Proc. 42nd Annual Meeting of the Association for Computational Linguistics*, pp.614–621 (2004).
15) 池原 悟，宮崎正弘，白井 諭，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，岩波書店 (1999).

(平成 17 年 10 月 14 日受付)
(平成 18 年 4 月 4 日採録)

平尾　努（正会員）

1995 年関西大学工学部電気工学科卒業．1997 年奈良先端科学技術大学院大学情報科学研究科博士前期課程修了．同年 NTT データ通信株式会社（現，株式会社 NTT データ）入社．2000 年より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に所属．2002 年奈良先端科学技術大学院大学情報科学研究科博士後期課程修了．博士（工学）．自然言語処理の研究に従事．言語処理学会，ACL 各会員．

奥村　学（正会員）

1989 年東京工業大学大学院情報理工学研究科計算工学専攻博士後期課程修了．1989 年より東京工業大学大学院情報理工学研究科助手．1992〜2000 年北陸先端科学技術大学院大学助教授．1997〜1998 年トロント大学客員助教授．2000 年より東京工業大学精密工学研究所助教授．自然言語処理，自動テキスト要約，コンピュータによる語学学習支援，テキストデータマイニングに関する研究に従事．工学博士．AAAI，ACL，JSAI，JCSS 各会員．

磯崎　秀樹（正会員）

1983 年東京大学工学部計数工学科卒業．1986 年同工学系大学院修士課程修了．同年日本電信電話（株）入社．1990〜1991 年スタンフォード大学ロボティクス研究所客員研究員．現在，NTT コミュニケーション科学基礎研究所知識処理研究グループリーダ．博士（工学）．平成 15 年度情報処理学会論文賞・山下記念研究賞受賞．人工知能・自然言語処理の研究に従事．電子情報通信学会，人工知能学会，言語処理学会，ACL 各会員．