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

2004－NL－162 (25)
2004／7／16

文書要約の自動評価手法の提案と評価

平尾 努† 奥村 学‡ 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,isozaki,maeda}@cslab.kecl.ntt.co.jp
‡ 東京工業大学 精密工学研究所
〒226-8503 神奈川県横浜市緑区長津田町 4259
oku@pi.titech.ac.jp

概要

Document Understanding Conference (DUC) や Text Summarization Challenge (TSC) に代表される評価型ワークショップが開催されることによって，自動要約システムの評価のためのコーパスが整備されつつある．しかし，こうしたワークショップでは，人手による一度切りの主観評価法を採用しており，ワークショップに参加していないシステムが繰り返し利用することが困難な状況にある．また，人手による評価法自体には多大なコストがかかるという問題もある．本稿では，このような状況を考慮し，要約を自動的に評価する手法を提案する．提案手法では，単語のシーケンシャルパターンに基づくテキスト間の類似度を用いる．TSC3 の複数文書要約タスクにおける人間の評価結果との相関を調べたところ，従来より提案されている自動評価法である ROUGE と比較して，提案手法は，より高い相関であることがわかった．

キーワード: 自動要約，自動評価，類似度，カーネル法，相関

An Automatic Evaluation Method for Text Summarization and its Evaluation

Tsutomu Hirao† Manabu Okumura‡ Hideki Isozaki† Eisaku Maeda†
†NTT Communication Science Laboratories, NTT Corp.
2-4, Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0237 Japan
{hirao,isozaki,maeda}@cslab.kecl.ntt.co.jp
‡Precision and Intelligence Laboratory, Tokyo Institute of Technology
4259, Nagatsuda-cho, Midori-ku, Yokohama-shi, Kanagawa 226-8503 Japan
oku@pi.titech.ac.jp

Abstract

Nowadays, we can use test collections for automatic summarization. However, human evaluation demands us huge costs. A few automatic evaluation methods were proposed, but they have problems. Therefore, a better automatic evaluation method is needed. In this paper, we propose a novel automatic evaluation method for text summarization. Our method employs Extended String Subsequence Kernel (ESK) as similarity measure between reference summaries and system summaries. We conducted experimental evaluation using the results of Text Summarization Challenge 3. The results show that our method is superior to the conventional automatic evaluation method, ROUGE.

Keywords: Automatic Summarization, Automatic Evaluation, Similarity Measure, Kernel Methods, Correlation

1 はじめに

近年，TIDES プロジェクトの DUC (Document Understanding Conference)[1] や NTCIR ワークショップの一環として開催されている TSC[2] (Text Summarization Challenge) に代表されるの自動要約の評価型ワークショップが開催されており，自動要約システムの評価のためのテストコレクションが蓄積されつつある[3]．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，一度切りの評価しかできず，コストも多大にかかる．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く，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い．また，こうした一度切りの人手による評価に頼ることの問題は機

[1] http://duc.nist.gov/
[2] http://oku-gw.pi.titech.ac.jp/tsc

[3] 本稿で言及する「要約」とは，ある定められた文字数を上限として生成されたテキストを指し，要約元の文書から決まった単位 (節，文など) を抽出したものではない．

械翻訳の評価においても同様に指摘されている．

本稿ではこのような状況を考慮して，要約を自動的に評価する手法を提案し，その有効性を TSC3 での人手による評価結果を用いて検証する．以下，2 章では，DUC, TSC における人手による評価法について説明し，3 章では，従来の自動評価法としてよく知られている BLEU と ROUGE について概説する．4 章では提案手法について詳しく述べ，5 章で実験結果と考察を行い，6 章でまとめる．

2 DUC, TSC における評価法

一般的に，要約の評価は「内容」と「読みやすさ」の評価に大別することができる．DUC, TSC ともこれら双方に人間による主観評価を採用している．「読みやすさ」の評価については，まだ議論の途中にあるので本稿では扱わない．以降，要約の評価とは「内容」評価を指すこととする．

DUC では，システムサマリとリファレンスサマリ (あるいは，モデルサマリとも言うが，本稿では，以降，人間による正解をリファレンスと呼ぶ) を談話構造を構成する基本単位 (Elementary Discource Unit, 以下, EDU) に分解し，リファレンスの EDU に対してシステムの EDU がどの程度情報を被覆しているかを $E = 0, 0.2, 0.4, \cdots, 1.0$ の 6 段階で判断し，以下の式でシステムサマリのスコアを決定する．

$$C = \frac{\#\text{ of marked EDUs} \times E}{\#\text{ of refernce EDUs}} \tag{1}$$

TSC でもほぼ，DUC に準拠した評価となっている．TSC では，システムサマリ，リファレンスサマリともに EDU ではなく，文に分解している．リファレンスサマリの文の数を n とし，リファレンスの i 番目に対してシステムサマリがどの程度情報を被覆しているかを $cov(i)$ として，$0, 0.1, 0.2, \cdots, 1.0$ の 11 段階で判定し，最終的に以下の式でシステムサマリのスコア (Coverage) を決定する．

$$Cov. = \frac{1}{n}\sum_{i=1}^{n} cov(i) \tag{2}$$

ここで，$cov(i)$ は，システムサマリの 1 文だけをみて決定するのではなく全体を見渡して決定する．たとえば，システムサマリの 1 文目と 3 文目を組み合わせるとモデルサマリの 2 文目の情報を 0.8 程度カバーするといったように決定する．

3 従来手法

前章での人間による主観評価を踏まえると，自動要約分野における自動評価とは，与えられたリファレンスサマリとシステムサマリとの間の類似度を測ることと等しいといえる．従来より，テキスト間の類似度を計算する手法は様々なものが提案されており，たとえば，bag-of-words 表現におけるコサイン類似度や編集距離がある．前者に属する自動評価法としては，Radev らによる Content-based evaluation [8] がある．また，後者に属する自動評価法としては，同じく Radev らによる MEAD-LCS [8] や Hori らによる SumACCY(WSumACCY)[3] がある．

一方，機械翻訳分野では，システム出力とリファレンス間における N-gram の一致率に基づく自動評価法 (BLEU) が提案されており，近年，広く用いられている．BLUE スコアは幾何平均を変化した以下の式で定義される．

$$\text{BLUE} = \text{BP} \times \exp\sum_{n=1}^{N} w_n \log p_n \tag{3}$$

p_n は，システム出力とリファレンスとの間で一致した N-gram (通常，$1 \leq n \leq 4$) が**システム出力に対して占める割合**であり，w_n は $1/n$ で与えられるその重みである．BP は，システム出力がリファレンスより短い場合に与えるペナルティ項である．

BLUE をそのまま要約の評価に利用した例 [7] もあるが，Lin らは，様々な試行の結果から，幾何平均ではなく，単なる N-gram の一致率だけを考えた方が，人間との相関が高いという事を報告している [5, 4]．この手法は Recall-Oriented Understudy for Gisting Evaluation (ROUGE) と呼ばれる．ただし，BLUE がシステム出力とリファレンスの間で一致した N-gram がシステム出力に対して占める割合に基づいていることに対して，ROUGE では，それが，**リファレンスに対して占める割合**に変更されているという特徴がある．ROUGE は以下の式で定義される．

$$\text{ROUGE-N} = \frac{\sum_{S \in Ref.}\sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in Ref.}\sum_{gram_n \in S} Count(gram_n)} \tag{4}$$

ここで，$Count_{match}(gram_n)$ は，リファレンスとシステム出力との間で一致した N-gram の数，$Count(gram_n)$ は，リファレンスに含まれる N-gram の数である．Lin らは，BLUE をそのまま使うよりも，$N = 1$ か $N = 2$ の場合に，人間の評価結果に対して高い相関が得られたことを報告している [5, 4]．

4 提案手法

4.1 ギャップも許した単語の共起に着目した類似度

前章で説明した BLEU, ROUGE ともテキスト間の類似度には N-gram にしか着目していない．つまり，隣接関係という強い制約にある語の共起にしか着目していない．しかし，隣接関係にない語の共起も重要であることは明白である．これは，日本語の場合，係り関係にある語が隣接しているとは限らないことからもわかる．

そこで，本稿では，テキスト中のギャップも許した語の共起を考慮し，類似度を計算する手法を用いる．こうした，テキスト中の部分構造に着目し，類似度を計算する方法としては，Convolution Kernel [2] が知られている．たとえば，テキスト中の全ての部分文字列，単語列に着目して類似度を計算する String Subsequence Kernel (SSK) [6], Word Sequence Kernel (WSK) [1] やテキス

表 1: 'abaca' と 'abbab' に対する 3 個までの部分単語列とその重み

subsequence	abaca	abbab
abb	0	$1+2\lambda^2$
aba	$1+\lambda^2$	2λ
abc	λ	0
aab	0	λ^2
aac	λ	0
aaa	λ^2	0
aca	λ^2+1	0
ab	1	$2+\lambda+\lambda^3$
aa	$2\lambda+\lambda^3$	λ^2
ac	$1+\lambda^2$	0
ba	$1+\lambda^2$	$1+\lambda$
bb	0	$1+\lambda+\lambda^2$
bc	λ	0
ca	1	0
a	3	2
b	1	3
c	1	0

トを木構造に見立てその部分木に着目して類似度を計算する Tree Kernel [2] などがこれに属する．本稿では，SSK，WSK の拡張である Extended String Subsequence Kernl (ESK) を用いる．

まず，WSK について説明する．カーネル関数 (K) とは，二つの対象 $\mathbf{x}$，$\mathbf{x}'$ に対し，それらをある関数 ϕ で写像した空間において対象間の内積計算として定義される．つまり，$\phi(\mathbf{x})\cdot\phi(\mathbf{x}')=K(\mathbf{x},\mathbf{x}')$ となる．ここで，WSK は，入力対象をテキストとし，ϕ がその d 個までの部分単語列を基底とする空間へ写像することに相当するカーネル関数である [1]．この時，基底に対する座標値は，着目した部分単語列の重みつき総和であり，重みはスキップした単語数 ℓ に応じて減衰パラメータ λ^ℓ で与えられる．いま，「abaca」というテキスト (単語列) と「abbab」というテキストを考えるとそれらの $d=3$ 個までの部分単語列とその重みは表 1 となる．よって，$K_{\mathrm{wsk}}(\text{`}abaca\text{'},\text{`}abbab\text{'})$ は，以下の式で計算される．

$$
\begin{aligned}
K_{\mathrm{wsk}}(\text{`}abaca\text{'},\text{`}abbab\text{'}) = {} & \overbrace{(1+\lambda^2)\times 2\lambda}^{aba} \\
& + \overbrace{1\times(2+\lambda+\lambda^3)}^{ab} \\
& + \overbrace{(2\lambda+\lambda^3)\times\lambda^2}^{aa} \\
& + \overbrace{(1+\lambda^2)\times(1+\lambda)}^{ba} \\
& + \overbrace{3\times 2}^{a} + \overbrace{1\times 3}^{b}. \qquad (5)
\end{aligned}
$$

ESK は，WSK における「単語」を「ノード」とみなし，複数の属性を持つことができるように拡張したものである．本稿では，ノードの持つ属性として「単語」とその「意味カテゴリ」[9] を用いた (図 1)．

4.2 ESK を用いた自動評価法

前節で説明した ESK を用いた要約の自動評価法について説明する．ROUGE は，リファレンスサマリ，システムサマリともに 1 本のシーケンスとみなしている．つまり，文境界をまたぐ N-gram も抽出している．しかし，2 章で説明した DUC や TSC での人間による主観評価法からもわかる通り，要約を N-gram の集合として考えるよりも，文あるいは EDU の集まりであると考えた方が自然である．よって，本稿では，要約を文の集合ととらえ，文間の類似度に基づいた自動評価手法を提案する．いま，リファレンスサマリの文数を n，システムサマリの文数を k として，以下の式で定義する．

$$
\mathrm{score}(Ref,Sys)=\frac{\sum_{i=1}^{n}\max_{s_j\in sys}\mathrm{sim}(s_i,s_j)}{n} \qquad (6)
$$

$\mathrm{sim}(s_i,s_j)$ は ESK を用いて以下の式で定義する．

$$
\mathrm{sim}(s_i,s_j)=\frac{K_{esk}(s_i,s_j)}{\sqrt{K_{esk}(s_i,s_i)K_{esk}(s_j,s_j)}} \qquad (7)
$$

5 評価実験と考察

提案手法の有効性を確認するため，TSC3 の評価結果データを用いて提案手法と ROUGE を比較評価した．

5.1 実験データ

TSC3 の複数文書要約タスクは，30 トピックに対して文字数制限が short，long という長さが異る 2 種の要約を作成するという設定である．リファレンスサマリは 5 名の人間がそれぞれ，6 トピックを担当し作成した．また，要約作成者自らが (2) 式を用いて，タスクに参加した 9 システムとオーガナイザが作成したベースライン (LEAD 手法)，リファレンスとは異なる人間による要約を加えた計 11 要約を評価した．最終的なシステムスコアは 30 トピックの平均値で与えられる．TSC3 コーパスの詳細については，文献 [10] を参照されたい．

5.2 実験の設定

要約の表現としては，以下の 2 つモデルを考え，N-gram の一致率による類似度，ESK による類似度をそれぞれ組み合わせた．

モデル 1　要約を 1 本のシーケンスとみなす

モデル 2　要約を文の集合とみなす

5.3 評価指標

評価指標としては，TSC3 で付与されたシステムスコアと先に述べた自動評価手法で付与したシステムスコア間の

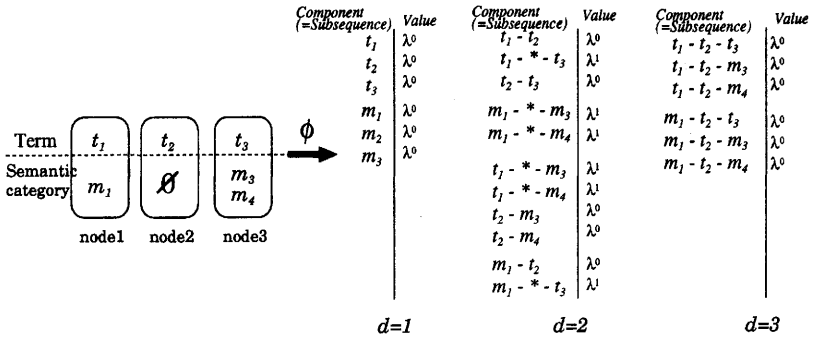

図 1: ノード列を入力とする ESK

ピアソンの相関係数と TSC3 で付与されたスコアにおいて，有意水準 5%で Tukey の多重比較手法を用いて検定した結果，有意な差があると判定されたペアを自動評価法がどの程度正確に当てることができたかを F 値を用いて評価する．いま，TSC3 における評価結果から，Tukey の多重比較手法を用いて検定した結果，有意な差が認められたシステムペアの数を a，自動評価手法を用いた結果から，同じく Tukey の多重比較手法を用いて検定した結果，有意な差が認められたシステムペア数を b，自動評価手法を用いた場合に，正しく有意差を認めたシステムペア数を c とすると，F 値は以下の式で表される．

$$\text{F-measure} = \frac{2 \times c/a \times c/b}{c/a + c/b} \tag{8}$$

ここで，F 値を評価尺度に用いた理由を以下に示す．

ピアソンの相関係数は，30 トピックで平均をとった後に計算している．このような場合では，高い相関が得られたとしても，人間によって付与されたシステム間のスコアの差 (間隔) を正しく反映しているとは限らない．自動評価手法には，人間の採点結果との相関が高いことも当然要求されるが，このように要約システム間の性能の差を正しく反映することも望まれる．そこで，本稿では上に示した F 値を用いて間隔尺度としての良さも測ることとした．

5.4 モデル 1 の場合の評価結果

モデル 1 の場合に，類似度として N-gram の一致率 (ROUGE-N)，ESK を用いた場合のそれぞれを表 2，表 3 に示す．なお，参考までに WSK を用いた場合の結果もあわせて示す．N-gram は自立語のみを対象とした場合と全ての単語を対象とした場合の双方を評価した．なお，ESK(WSK) では，全ての単語を用い，考慮する組み合わせ数を 2 から 4 まで変化させ，それぞれの場合において，スキップした単語数に応じて与える減衰パラメータ λ を 0.1 きざみで 0.1 から 1 まで変化させた場合の最大値，最小値，平均値を示す．

表 2，3 より，要約を 1 本のシーケンスとみなすのであれば，N-gram の一致率に基づく類似度を用いた方が

表 2: N-gram の一致率を類似度として用いた場合の相関係数と F 値 (モデル 1)

	short		long	
	cor	F(5%)	cor	F(5%)
ROUGE-1(自立語のみ)	.777	.565	.847	.760
ROUGE-1(全ての単語)	.808	.638	**.880**	**.824**
ROUGE-2(自立語のみ)	.738	.524	.765	.667
ROUGE-2(全ての単語)	**.820**	**652**	.873	.808
ROUGE-3(自立語のみ)	.663	.474	.718	.667
ROUGE-3(全ての単語)	.769	.564	.829	.739
ROUGE-4(自立語のみ)	.557	.485	.628	.545
ROUGE-4(全ての単語)	.698	.438	.765	.684

EKS(WSK) よりも良いことがわかる．Lin の観察のとおり，uni-gram，bi-gram の場合に人間との相関が高く，F 値の成績も良い．short の場合には自立語のみを対象とした bi-gram を用いた場合が最も良く，long の場合には全ての単語を対象とした uni-gram を用いた場合が最も良い．これに対し，ESK は組み合わせ数を多くすると平均値では下がる傾向にあり，最大値と最小値の差が大きくなる．すなわち，減衰パラーメタに対し敏感になる．また，F 値も相関係数と同様で下がる傾向にある．これは，間隔尺度として良くないことを示している．ESK(WSK) は，相関係数，F 値とも類似度に N-gram の一致率を用いる場合よりも低い値であるが，これは，要約を 1 本のシーケンスとみなすことで，無意味な長距離の共起を考慮してしまうことが原因であると考える．

5.5 モデル 2 の場合の評価結果

次に，モデル 2 の場合に，類似度として N-gram の一致率を用いた場合，すなわち，式 (6) における $\text{sim}(s_i, s_j)$ を式 (4) におきかえて計算した場合，ESK を用いた場合のそれぞれを表 4，表 5 に示す．カーネルパラメータは，モデル 1 の場合と同じ設定で実験を行った．

表 2 と表 4，表 3 と表 5 を比較すると相関係数，F 値

表 3: ESK を用いた場合の相関係数と F 値 (モデル 1)

	short						long					
	cor			F(5%)			cor			F(5%)		
	mean	max	min	mean	max	min	mean	max	min	mean	max	min
eskd02	.647	.702	.629	.502	.571	.462	.692	.763	.678	.530	.583	.490
eskd03	.617	.726	.581	.484	.537	.474	.567	.733	.533	.512	.583	.489
eskd04	.596	.750	.531	.449	.556	.400	.503	.727	.436	.479	.612	.429
wskd02	.649	.705	.631	.508	.537	.488	.702	.732	.660	.551	.553	.542
wskd03	.623	.736	.585	.500	.550	.476	.641	.720	.577	.546	.565	.511
wskd04	.592	.760	.528	.472	.513	.387	.597	.733	.519	.531	.612	.476

表 5: ESK を用いた場合の相関係数と F 値 (モデル 2)

	short						long					
	cor			F(5%)			cor			F(5%)		
	mean	max	min	mean	max	min	mean	max	min	mean	max	min
eskd02	.865	.872	.855	.763	.769	.750	.926	.927	.923	.907	.917	.898
eskd03	.853	.862	.810	.697	.722	.684	.895	.908	.865	.880	.917	.780
eskd04	.812	.834	.728	.580	.647	.370	.845	.879	.779	.794	.889	.667
wskd02	.839	.844	.835	.674	.737	.632	.910	.911	.910	.881	.898	.875
wskd03	.833	.841	.801	.706	.757	.684	.895	.904	.869	.873	.917	.818
wskd04	.807	.831	.730	.641	.722	.414	.866	.896	.797	.822	.889	.757

表 4: N-gram の一致率を類似度として用いた場合の相関係数と F 値 (モデル 2)

	short		long	
	cor	F(5%)	cor	F(5%)
1-gram(自立語のみ)	.743	.558	.785	.683
1-gram(全ての単語)	.573	.421	.566	.450
2-gram(自立語のみ)	**.837**	**.684**	.876	.837
2-gram(全ての単語)	.818	.615	**.885**	**.837**
3-gram(自立語のみ)	.792	.545	.867	.769
3-gram(全ての単語)	.798	.611	.872	.780
4-gram(自立語のみ)	.682	.414	.809	.629
4-gram(全ての単語)	.760	.516	.843	.737

ともに向上していることがわかる．これは，要約を 1 本のシーケンスとみなし，N-gram や部分単語列の集合として表わすよりも，文の集合として表わすことが適切であることを示している．特に，ESK を類似度として用いた場合には，性能の向上が著しい．モデル 1 では，N-gram の一致率を類似度として用いた場合よりも成績が悪かったが，モデル 2 では，大きく上回っている．組み合わせの数を 2 に設定した場合では，short の場合，相関係数で 2 から 4 ポイント，F 値では 7 から 8 ポイント上回り，long の場合もほぼ同様の向上が見られる．特に F 値での改善が大きいことから，類似度に ESK を用いる方が，間隔尺度としても優れていることを示す．また，ESK の場合に考慮すべき組合せの数 (d) に関しては，2 から増やすに従って，相関係数，F 値ともに，最小値と最大値の差が大きくなる傾向にあり，減衰パラメータの影響を受けやすいことがわかる．よって，減衰パラーメタの影響が小さく，人間との相関，F 値とも高い $d = 2$ を用いるのが良いと考える．また，WSK と ESK を比較すると ESK の方がやや成績が良い．これは，単語の意味情報を考慮することで語彙の言いかえを吸収できている部分があることを示している．

以上より，要約を文の集合と捉え，文間の類似度を ESK を用いて計算することで従来の手法よりも高い相関，F 値を得ることができた．

5.6 トピック毎の相関係数

表 4 と表 5 において最も相関係数が高かった手法に対して，各トピック毎の相関係数を調べた．結果を表 6 に示す．表 6 より，short，long ともに約 20 トピックにおいて，ESK を用いた方が N-gram の一致率を用いるよりも高い相関を得ることができている．

各トピック毎との相関係数に着目すると，良い場合で，0.9 程度，悪い場合では，ほぼ 0 の無相関に近い結果となっている．さらに，N-gram の一致率を用いた場合では，long で負の相関となったトピックがいくつか存在する．この原因は，自動表化手法は，基本的には文や節，句といったテキスト中のかたまりに対する意味を扱うことができないことにあると考える．たとえば，意味をなさない非文のようなテキストであっても，リファレンスに対して表層的に類似していれば，自動評価法は高いスコアを付与する傾向がある．一方，人間は，意味を考えたスコアリングを行って

表 6: トピック毎の相関係数

	short		long	
	ESK	N-gram	ESK	N-gram
	$d=2, \lambda=0.8$	N = 2(自立語のみ)	$d=2, \lambda=0.5$	N = 2(全ての単語)
0310	**.829**	.672	.698	**.731**
0320	.740	**.784**	**.614**	.424
0340	**.667**	.623	.774	**.781**
0350	**.349**	.135	**.643**	.636
0360	**.896**	.823	**.887**	.874
0370	**.591**	.569	**.714**	.607
0380	.765	**.804**	**.963**	.860
0400	**.687**	.660	**.588**	.253
0410	**.483**	.421	.597	**.706**
0420	**.818**	.816	**.868**	.828
0440	**.739**	.656	**.748**	.561
0450	**.571**	.415	**.234**	−.008
0460	**.900**	.815	**.866**	.837
0470	.537	**.675**	**.594**	.323
0480	**.749**	.701	**.944**	.896
0500	**.734**	.500	**.777**	.609
0510	.458	**.630**	**.776**	−.037
0520	**.897**	.754	**.631**	.530
0530	.804	**.843**	**.155**	−.041
0540	.840	**.851**	**.718**	.685
0550	.847	**.858**	.818	**.833**
0560	.768	**.845**	**.954**	**.954**
0570	.023	**.190**	**.544**	.434
0580	**.858**	.757	.898	**.932**
0590	**.665**	.586	**.967**	.932
0600	.674	**.722**	**.873**	.804
0610	**.939**	.916	.937	**.946**
0630	.616	**.653**	**.783**	.758
0640	**.631**	.521	.830	**.893**
0650	**.655**	.562	**.519**	.430

表 7: すべての要約に対する相関係数

	short	long
N-gram	.646	.636
ESK	.696	.684

いるので，自動評価法との間には大きな隔がある．これは大きな問題点であるが，現在の自然言語処理技術では対処できない問題であると考える．

5.7 全要約に対する相関

続いて，全 330 要約 (人間の要約も含めた 11 システム ×30 トピック) に対する人間のスコアと自動要約手法との間の相関係数を調べた．散布図を図 2 に相関係数を表 7 に示す．散布図より，ESK を用いた方が点のばらつきがや や小さいことがわかるが，実際の相関係数は，5 ポイント程度の差がある．ただし，30 トピックで平均をとった場合の相関 (表 4，5) よりも相関係数の値は大幅に小さくなっている．これは，表 6 からもわかるとおり，トピック毎の相関係数のばらつきが原因である．

6 まとめと今後の課題

本稿では，要約を文の集合ととらえ，文間の類似度を Extended String Subsequence Kernel (ESK) を用いて計算する自動評価手法を提案し，従来より提案されている N-gram の一致率に基づく自動評価手法 (ROUGE) との比較評価を Text Summarization Challenge 3 (TSC3) における人手の評価結果を用いて行った．その結果，提案手法が従来の手法よりも人間の評価に対して高い相関であるこ

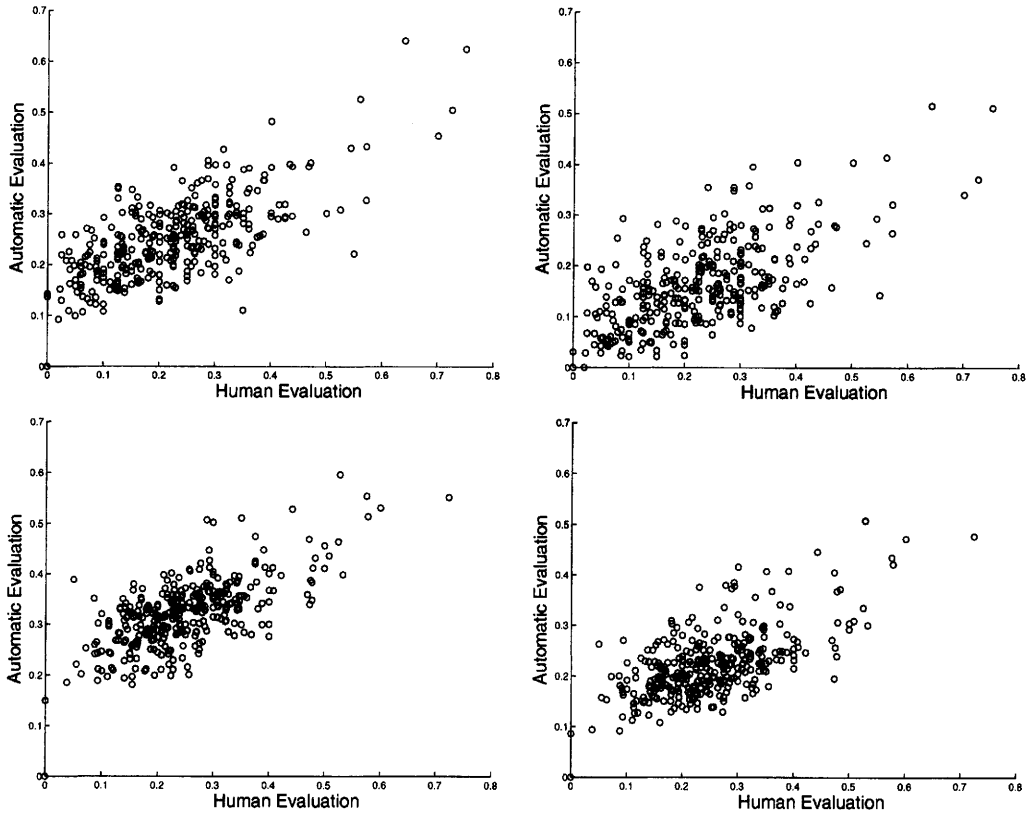

図 2: 人間による評価と自動評価法による評価の関係.short における人間と ESK を用いた場合の散布図 (左上), 人間と N-gram 一致率を用いた場合の散布図 (右上), long における人間と ESK を用いた場合の散布図 (左下), 人間と N-gram 一致率を用いた場合の散布図 (右下).

とを確認し, 間隔尺度としてもより優れていることを確認した.

今後の課題としては, 提案手法が ROUGE よりも優れている原因が言語の違いに由来しているかどうかを調べる必要がある. これについては, DUC のデータを用いて実験を行う予定である. また, ESK を用いた場合, 機能語同士の組合せも考慮するという問題があるので, 品詞の組合せによって重みを変化させるような仕組を取り入れたい. さらには, 機械翻訳結果の評価への適用を行い, BLUE との比較評価を行いたいと考える.

謝辞

NTCIR ワークショップ, TSC の運営に関係された全ての皆様に感謝致します. また, TSC3 の参加者の皆様, TSC3 の評価に関して有益なコメントをいただいた株式会社アイアール・アルトの河野香織氏, 山田薫氏に感謝致します. さらに, 常日頃より議論いただく NTT コミュニケーション科学基礎研究所の賀沢秀人氏, 鈴木潤氏に感謝いたします.

参考文献

[1] Cancedda, N., Gaussier, E., Goutte, C. and Rnders, 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. of Neural Information Processing Systems (NIPS'2001)* (2001).

[3] Hori, C. and Furui, S.: Evaluation Mehtods for Automatic Speech Summarization, *Proc. of the Eurospeech 2003* (2003).

[4] Lin, C.-Y.: Looking for a Good Metrics: ROUGE and its Evaluation, *Proc. of the 4th NTCIR Workshops (open submission)*, pp. 1–8 (2004).

[5] Lin, C.-Y. and Hovy, E.: Automatic Evaluation of Summaries Using N-gram Co-Occurrence Statistics, *Proc. of the 4th Meeting of the North American Chapter of the Association for Computational Linguistics and Human Language Technology* (2003).

[6] 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).

[7] Pastra, K. and Saggion, H.: Colouring Summaries BLEU, *EACL 2003. Workshop on Evaluation Initiatives in Natural Language Processing* (2003).

[8] Saggion, H., Radev, D., Teufel, T. and Lam, W.: Meta-Evaluation of Summaries in a Cross-Lingual Environment Using Content-Based Metrics, *Proc. of the ?th International Conference on Computational Linguistics* (2002).

[9] 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).

[10] 平尾努, 奥村学, 福島孝博, 難波英嗣: TSC3 コーパスの構築と評価, 第 10 回言語処理学会年次大会 (2004).