特　集「テキストの自動評価」

テキスト要約の自動評価

Automatic Evaluation in Text Summarization

難波　英嗣
Hidetsugu Nanba
広島市立大学大学院情報科学研究科
Faculty of Information Sciences, Hiroshima City University.
nanba@hiroshima-cu.ac.jp, http://www.nlp.its.hiroshima-cu.ac.jp/~nanba/

平尾　努
Tsutomu Hirao
NTTコミュニケーション科学基礎研究所
NTT Communication Science Laboratories.
hirao@cslab.kecl.ntt.co.jp, http://www.kecl.ntt.co.jp/icl/kpro/hirao/index-j.html

Keywords: text summarization, summary evaluation.

1. はじめに

Luhn[Luhn 58] 以来，半世紀にわたって行われてきたテキスト要約研究は，さまざまな自然言語処理技術の発展により，今日，その主流がテキスト中の重要な箇所を抽出するという研究から，テキスト（要約）を生成するという研究へと移行しつつある．抽出に基づく要約を評価するために，これまでは精度や再現率といった尺度を用いて，人間の被験者が作成した抽出に基づく要約との一致度を測るのが一般的であった．しかし，生成に基づく要約の評価にこれらの尺度をそのまま適用することはできない．また，人間による評価（以後，マニュアル評価）は正確である反面，時間，金銭的コストが多大にかかるうえに評価を繰り返し行うことが困難であるという問題もある．こうしたことを背景として，近年では，テキスト要約の自動評価法に関して活発に議論されるようになってきている．本稿では，要約の自動評価に焦点を当て，これまでの研究を概観する．

コンピュータが自動的に作成した要約（以後，システム要約）は，「人間が作成した理想的な要約（以後，参照要約）とどれだけ似ているか」によって評価されることが多い．このような枠組みでの評価法は，テキスト要約に限らずテキスト生成系の研究全般に適用可能なものである．例えば，機械翻訳で一般的に用いられているBLEU[Papineni 02] という自動評価法は要約評価にもそのまま利用できる．実際に，BLEUを要約評価用に改良したROUGE[Lin 04b] は，今日，テキスト要約分野で標準的に用いられる自動評価法の一つになっている．2005年にミシガン大学で開催された自然言語処理関係で最大の国際会議であるACLにおいて，テキスト要約と機械翻訳の評価法に関するワークショップが開催され，両分野の研究者が評価法に関して活発に議論を行った．

このように「参照要約との類似性による評価」は，ほかの関連分野と協調しつつ発展してきたものの，残念ながら，まだマニュアル評価の代替となるレベルまでには至っていない．これに対し，近年，要約とそのマニュアル評価結果を大量にプーリング（蓄積）しておき，マニュアル評価結果に基づいて人間の代替となるような自動評価を実現しようとする全く別のアプローチからの研究が行われるようになってきている．本稿では，近年の要約評価研究を「参照要約との類似性による評価」と「マニュアル評価結果に基づいた評価」に分けて概観する．

本稿の構成は以下のとおりである．2章では，要約研究および要約評価の変遷について述べる．3章では参照要約との類似性による評価法について，4章ではマニュアル評価結果を使った評価法について，それぞれ述べ，5章で本稿をまとめる．

2. テキスト要約研究および要約評価の変遷

テキストを自動的に要約する研究の歴史は，1950年代に行われたLuhnの研究 [Luhn 58] から始まる．Luhnは，まず，テキスト中から単語の頻度に基づいて重要語を抽出し，次に，テキスト中の各文の重要度を，その文が重要語を含む比率によって算出し，最後に，指定された要約率に達するまで重要度の高い順に文を抽出する，というものであった．その後，さまざまな要約手法が考案されたが，いずれもテキスト中の重要文を抽出する，いわゆる重要文抽出が，長い間テキスト要約研究の中心的な課題であった．また，重要文抽出法によって作成された要約の評価は，人間の被験者が抽出した文と要約システムが自動抽出した重要文との一致度を再現率，精度，F 値などの尺度で測るのが一般的であった．

1990年代後半に入り，ある程度実用的な精度で構文解析が可能になると，要約システムのモジュールとして構文解析器を用い，文よりも小さな単位（例えば文節など）でテキスト中の重要箇所を抽出する，いわゆる重要箇所抽出と呼ばれる方法で要約を作成する研究が行われるようになった．また，重要文抽出の結果を書き換え

て読みやすい要約を自動作成する研究 [Mani 99, Nanba 00]，音声認識の結果を要約しテレビのニュース番組などに字幕を自動的に付与する研究，情報検索システムが検索結果を表示する際にユーザに提示する要約（snipet）を作成する研究 [Tombros 98]，複数のテキストから一つの要約を作成する複数テキスト要約研究など，要約研究がさまざまな広がりを見せるようになってきた．この時期，国内では Text Summarization Challenge（TSC）*1，国外では Tipster プロジェクトの一環として行われた SUMMAC や Document Understanding Conference（DUC）*2 といった評価ワークショップが開催され，要約研究に取り組む研究者の裾野が広がったことも，要約研究の多様性を生み出す要因の一つになったと考えられる．さらに，評価ワークショップの成果として，要約研究のためのデータが整備されると，これらのデータを訓練データとして用い，機械学習ベースでテキスト要約システムを構築する研究も，この頃から盛んになった．

このように，重要文抽出以外のさまざまな方法で要約が作成されるようになると，要約の作成方法だけでなく評価方法にも要約研究者の注意が向けられるようになった．重要文抽出法では，人間の作成した要約（以後，参照要約）とシステム要約は，文という単位で比較していたが，単文，単語列，単語など，さまざまな言語単位で参照要約とシステム要約を比較し自動評価する方法がこれまでに提案されている．その一部は，次節で紹介する．また，マニュアル評価において，Nenkova らは SCU（Summary Content Unit）という意味的な単位に基づいて評価する方法（以後，ピラミッド法）を提案している [Nenkova 07]．ピラミッド法では，まず，人間の被験者が SCU を認定し，次に，SCU レベルで参照要約とシステム要約を比較する．ピラミッド法などのマニュアル評価は正確である反面，非常にコストがかかる．これは，例えば要約システムのパラメータをチューニングする際に要約評価を用いるといった場合，大きな問題となる．また，マニュアル評価は，比較実験のために実験環境を再現するのが困難であるという問題もある．このような背景から，マニュアル評価の代替となる自動評価法の確立を目指し，今日に至るまで，活発に研究されている．これらの評価法のうち，参照要約との類似性による自動評価法を 3 章で，マニュアル評価結果を用いた自動評価法を 4 章で，それぞれ述べる．

3. 参照要約との類似性による自動評価法

参照要約との類似性による自動評価法は，参照要約 R とシステム要約 C の間で一致する文字列に基づきそのスコア（評価値）を計算する．これは，参照要約とシステム要約との間の一種の類似度を計算することであり，参照要約との類似度が高いほどより良い要約であるという考えに基づく．類似度計算の考え方としては，テキストをシーケンスとして捉え，その部分列（N グラム，スキップを許した N グラム）の一致数に基づき計算する手法，テキストを構文木として捉え，その部分木の一致数に基づき計算する手法がある．以下，それぞれについて説明する．

3・1 ROUGE

現在，要約システムの自動評価法として最も広く用いられている ROUGE-N[Lin 03]*3 は，参照要約とシステム要約の間で一致する N グラムの割合を計算する．以下の式で定義される．

$$\text{ROUGE-N}(C,R)=\frac{\sum_{e\in n\text{-gram}(C)}\text{Count}_{\text{clip}}(e)}{\sum_{e\in n\text{-gram}(R)}\text{Count}(e)} \quad (1)$$

n-gram(C) は，システム要約に含まれる N グラム，n-gram(R) は，参照要約に含まれる N グラム集合を表す．Count(e) は，ある N グラムの出現頻度を数える関数であり，$\text{Count}_{\text{clip}}(e)$ は，システム要約に含まれる N グラムのシステム要約における出現頻度 Count$(e \in n\text{-gram}(C))$ と参照要約における出現頻度 Count$(e \in n\text{-gram}(R))$ の小さいほうの値を採用する．Lin らは，N を 1 ～ 4 まで変化させ，マニュアル評価結果との相関を調べた結果，$N = 1, 2$ が最も高い相関であったと報告している．

しかし，ROUGE-N は，隣接という強い制約にある単語共起のみに着目し，スコアを計算する．よって，隣接していないが，係り受け関係にあるような単語の共起を考慮することができない．このような問題を解決するため，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）を考慮して一致率を計算する ROUGE-S を提案した [Lin 04a, Lin 04b]．以下の式で定義される．

$$\text{ROUGE-S}(C,R)=\frac{(1+\gamma^2)P_S(C,R)R_S(C,R)}{R_S(C,R)+\gamma^2 P_S(C,R)} \quad (2)$$

ここで，$P_S(C,R)$，$R_S(C,R)$ は，それぞれ以下の式で定義される．

$$P_S(C,R)=\frac{\sum_{e\in \text{bigram}(C)\cup\text{skip-bigram}(C)}\text{Count}_{\text{clip}}(e)}{\sum_{e\in \text{bigram}(C)\cup\text{skip-bigram}(C)}\text{Count}(e)} \quad (3)$$

$$R_S(C,R)=\frac{\sum_{e\in \text{bigram}(C)\cup\text{skip-bigram}(C)}\text{Count}_{\text{clip}}(e)}{\sum_{e\in \text{bigram}(R)\cup\text{skip-bigram}(R)}\text{Count}(e)} \quad (4)$$

さらに，ROUGE-S に対しユニグラムを素性として追加した ROUGE-SU[Lin 04a, Lin 04b] も提案されている．

ただし，ROUGE-S，ROUGE-SU ではスキップを許したトライグラムを扱うことはできない．さらに，ある

*1 http://www.lr.pi.titech.ac.jp/tsc/
*2 http://duc.nist.gov/

*3 先に説明した DUC で自動評価指標として採用されている．

単語の組合せが参照要約，システム要約のどちらか一方ではバイグラム，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がある．さらに，ROUGE-Nも含め，単語の表記での一致しか見ておらず，単語の言換えがあった場合には一致数が著しく低下するという問題もある．

3・2 ESK

先述したROUGEの問題点を解決するため，[平尾 06]では，拡張ストリングカーネル（Extended String Subsequence Kernel: ESK）を用いた自動評価法が提案された．

ESKは，Lodhiらによって提案されたString Subsequence Kernel（SSK）[Lodhi 02]，Canceddaらによって提案されたWord Sequence Kernel（WSK）[Cancedda 03]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も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．詳しい計算法に関しては[平尾 06]を参照されたい．このとき，ノードのスキップに対しては，λ（$0 \leqq \lambda \leqq 1$）という減衰パラメータを用いてその重みを小さくする．例えば，ノードを一つスキップした場合には，重みがλとなり，二つスキップした場合には，λ^2となる．

例として，以下の参照要約，システム要約を用いて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．

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

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

ここで，“cosmonaut”と“astronaut”は共通の意味ラベル“*SPACEMAN*”をもち，“ambition”と“dream”は共通の意味ラベル“*DREAM*”をもつ．

“Becoming-*DREAM*”という部分列は，Rでは“a”，“cosmonaut:*SPACEMAN*”，“is”，“my”，“great”という五つのノードをスキップしており，Cでは同様に四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．

$\mathrm{ESK}^{d=2}(C, R)$は，C，Rから得た重み付きベクトルの内積であるので，C，R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．

$$\begin{aligned}\mathrm{ESK}^{d=2}(C,R) &= 1+1+1+1+1+\lambda^9+\lambda^2+\lambda^4+\lambda^6+\lambda^5+1 \\ &\quad +\lambda^2+\lambda^3+1+\lambda \\ &= 7+\lambda+2\lambda^2+\lambda^3+\lambda^4+\lambda^5+\lambda^6+\lambda^9\end{aligned}$$

表1に例としてあげたR，Cに対しROUGE-1，2，ROUGE-S，SU，$\mathrm{ESK}^{d=2}$を用いて一致する部分列を計算した結果をまとめる．表より，単語の意味ラベルを考慮できないROUGEは，ESKと比較してかなり低い値をとる傾向にあることがよくわかる．特にバイグラムの一致数は1しかなく，極端に低いスコアである．これに対し，意味ラベルを考慮できるESKは高いスコアであり$\lambda=0$としてもROUGEより十分高い．これは，人間の直観に合致している．[平尾 06]によると，λは，0.5程度がよいとされている．

3・3 Basic Elements

これまでに紹介したROUGE，ESKがテキストをシーケンスとして捉えていることに対し，Basic Elements（BE）を用いた自動評価法[Hovy 06]はテキストを構文木として捉えている点で大きく異なる．参照要約，システム要約に含まれる文を構文木へと変換し，その部分構造であるBEへと分解し，それらの間で一致するBEの数に基づきスコアを計算する．ここで，BEは，文中の名詞，動詞，形容詞などの単語とそれらとの間に依存関係（係り受けの関係）をもつ単語の組，それらの組の間の関係名のタプル〈関係名，単語1，単語2〉として表現される．

R 天気は雨になりそうだ．

C 雨の天気になりそうだ．

上記の参照要約，システム要約を例にするとBEは以下のとおりである．

R：〈は，天気，なりそうだ〉，〈に，雨，なりそうだ〉

C：〈の，雨，天気〉，〈に，天気，なりそうだ〉

なお，関係名に関しては格解析を用いて詳細に決定すべきであるが，本稿では，[福本 06]と同様，単純に二つの単語を結ぶ単語で表した．例より，係り受け関係にある単語の組に着目するため，隣接した単語の組だけでなく，自然とスキップした単語の組を扱えていることがわかる．しかし，この例では，Rの一つ目BEとCの二つ目のBEが関係名が異なっているため，一致するBEの数は0であり，我々の直観に反する．このように，関係名を入れることで制約がより強くなり，問題となる場合もある．また，単語の言換えを吸収することができないという点ではROUGEと同様の問題を抱えている．

表1 各手法による参照要約とシステム要約との間の部分列の一致数

手法	一致数
ROUGE-1	3
ROUGE-2	1
ROUGE-S	3
ROUGE-SU	4
$\mathrm{ESK}^{d=2}$	$7+\lambda+2\lambda^2+\lambda^3+\lambda^4+\lambda^5+\lambda^6+\lambda^9$

4. マニュアル評価結果を用いた自動評価法

マニュアル評価は人間の被験者が要約を評価するため正確であるが，コストがかかるという問題がある．これに対し，3章で述べた参照要約との類似性による評価法はプログラムやツールなどを用いるためコストはかからないが，現状ではマニュアル評価のような正確な評価を行うまでには至っていない．この両者の問題を解決するため，近年では，マニュアル評価結果を大量にプーリングしておき，それらを用いて自動評価するという新しい評価法が検討されるようになってきている．本章では，マニュアル評価結果を用いたいくつかの自動評価法について述べる．

4・1 プーリングデータを用いた自動評価法

一般に，もし二つの要約が類似していれば，その評価結果も類似していると考えられる．そこで，あるテキストから生成された要約（以後，プーリング要約）とそのマニュアル評価結果を大量にプーリングしておけば，任意の要約のマニュアル評価スコアを，内容の類似したプーリング要約から推測できる．このような考え方に基づいて，賀沢ら [賀沢 03] は次に述べる手法で要約の自動評価を行っている．

あるテキストに対し n 種類の要約手法を用いて要約 $C_1 \sim C_n$ を生成し，それらをマニュアル評価しておく．ここで，要約 $C_1 \sim C_n$ に対するマニュアル評価スコアを $H(C_1,R) \sim H(C_n,R)$ と表記する．R は参照要約を示す．また，要約 $C_1 \sim C_n$ とその評価スコア $H(C_1,R) \sim H(C_n,R)$ をここではプーリングデータと呼ぶ．ある要約システムXが生成した要約の評価スコアは，その要約と各プーリング要約 C_i との類似度[*4] sim_i を考慮し，$\sum_{i=1}^{n} H(C_i,R) \cdot sim_i$ によりマニュアル評価スコアを足し合わせることで計算される．もし，システムXの要約と非常に類似した要約がプーリングデータ中に存在すれば，類似度 sim_i の値が大きくなるため，システムXの要約の評価スコアは，そのプーリング要約の評価スコアに近くなる．また，もし非常に類似する要約がプーリングデータ中に存在しなくても，ほかのプーリング要約との類似度を考慮して，評価スコアを算出できる．なお，賀沢らが提案する実際の評価法では，プーリング要約を作成する要約手法の信頼度，評価スコアを0～1の間におさまるよう調整するためのパラメータなども考慮している．

4・2 単回帰モデルに基づく自動評価法

賀沢らの評価法とはアプローチが異なるが，機械翻訳の分野でもプーリングデータを用いてシステムを評価する手法が提案されている [Yasuda 03]．著者ら [Nanba 06] はこの手法（以後，Yasuda手法）を用い，要約評価におけるYasuda手法の有効性を検証している．以下に，Yasuda手法を用いた要約評価について述べる．

一般に，A，Bの二つの要約システムのうち，AがBよりも優れていれば，Aが高品質の要約を生成する割合はBよりも高い．また，AとBの性能差が大きいほど，その割合は大きくなる．逆に，AとBの性能差がほとんどなければ，割合はほぼ0.5になると考えられる．そこで，性能の高いものから低いものまでさまざまなレベルの要約システムを準備しておき，それらをあらかじめマニュアル評価しておけば，「ある要約システムXを評価する」という問題は，「さまざまな性能の要約システムの中から，要約システムXと同品質の要約を生成するシステムを見つける」という問題として解くことができる．以下に，評価の具体的な手順について述べる．

準備段階

(1) 複数の要約システム $S_1 \sim S_n$ を用意する．
(2) テキスト $T_1 \sim T_m$ を各要約システムを使って要約する．
(3) これらの要約をマニュアル評価し，その値を各要約システムの評価スコア（性能）とみなす．

評価段階

(1) 評価対象の要約システムXでテキスト $T_1 \sim T_m$ を要約する．
(2) 3章で述べたROUGEなどの尺度を用いてシステムXの要約を評価する．
(3) 要約システム $S_1 \sim S_n$ を用いてテキスト $T_1 \sim T_m$ を要約する．
(4) ROUGEなどの尺度を用いてシステム $S_1 \sim S_n$ の要約を評価する．
(5) 手順2の結果と手順4の結果を比較し，システムXよりも高品質な要約が作成した割合を要約システム $S_1 \sim S_n$ ごとに計算する．
(6) グラフの横軸に要約システムの性能（プーリングデータの準備の手順3で得られた値），縦軸に手順5の割合をプロットする．
(7) 回帰直線を引き，その直線上で割合（グラフの縦軸）が0.5となる点の横軸の値をシステムXの評価スコアとして出力する．

著者らはTSC2のデータを用いてYasuda手法とROUGEを比較する実験を行い，Yasuda手法がROUGEよりも優れていることを確認している．

4・3 重回帰モデルに基づく自動評価法

前節で説明したYasuda手法は，単回帰（説明変数が一つしかない）モデルである．しかし，一般的には，複数の説明変数を組み合わせたほうが回帰モデルの予測性能はより向上する．そこで，[平尾 07] では，重回帰モ

*4 類似度計算には，情報検索などで一般的に使われているコサイン距離のほかに，3章で紹介した参照要約との類似性による評価法が利用できる．

デルを改良した自動評価法が提案された．

訓練データとしてn個のシステム要約に対してマニュアル評価$H(C_i, R)$が得られており，ROUGE-1, ROUGE-S, ESKスコアがそれぞれ求まっているとする．ここで，重回帰モデルでは，$H(C_i, R)$と自動評価スコアの間に以下の関係を仮定する．

$$\left.\begin{aligned} H(C_1,R) &= \beta_0+\beta_1\text{ROUGE-1}(C_1,R) \\ &\quad +\beta_2\text{ROUGE-S}(C_1,R) \\ &\quad +\beta_3\text{ESK}(C_1,R)+\varepsilon_1 \\ H(C_2,R) &= \beta_0+\beta_1\text{ROUGE-1}(C_2,R) \\ &\quad +\beta_2\text{ROUGE-S}(C_2,R) \\ &\quad +\beta_3\text{ESK}(C_2,R)+\varepsilon_2 \\ &\vdots \\ H(C_n,R) &= \beta_0+\beta_1\text{ROUGE-1}(C_n,R) \\ &\quad +\beta_2\text{ROUGE-S}(C_n,R) \\ &\quad +\beta_3\text{ESK}(C_n,R)+\varepsilon_n \end{aligned}\right\} \quad (5)$$

なお，$H(C_i, R)$は目的変数と呼ばれる．また，ε_iは，誤差を表す．ここで，母回帰係数ベクトルを$\boldsymbol{\beta}=(\beta_0, \beta_1, \beta_2, \beta_3)^T$，目的変数ベクトルを$\boldsymbol{y}=(H(C_1,R), H(C_2,R), \cdots H(C_n,R))^T$，誤差ベクトルを$\boldsymbol{\epsilon}=(\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n)$，計画行列$\boldsymbol{X}$を

$$\boldsymbol{X}=\begin{bmatrix} 1 & \text{ROUGE-1}(C_1,R) & \text{ROUGE-S}(C_1,R) & \text{ESK}(C_1,R) \\ 1 & \text{ROUGE-1}(C_2,R) & \text{ROUGE-S}(C_2,R) & \text{ESK}(C_2,R) \\ \vdots & \vdots & \vdots & \vdots \\ 1 & \text{ROUGE-1}(C_n,R) & \text{ROUGE-S}(C_n,R) & \text{ESK}(C_n,R) \end{bmatrix}$$

とすると，式(5)は，以下の式で表せる．

$$\boldsymbol{y}=\boldsymbol{X\beta}+\boldsymbol{\epsilon} \quad (5)'$$

ここで，マニュアル評価スコアと回帰モデルによって推測したスコアの二乗誤差を最小にする回帰パラメータベクトル$\hat{\boldsymbol{\beta}}$は以下の式で求まる．

$$\hat{\boldsymbol{\beta}}=(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} \quad (6)$$

このように，$\hat{\boldsymbol{\beta}}$を求めることができれば，マニュアル評価スコアが与えられていないシステム要約に対し，ROUGEなどの自動評価スコアのみを用いてそのマニュアル評価スコアを予測することができる．

ただし，重回帰モデルでは，強い相関にある説明変数の組が含まれていると，多重共線性という現象が起こり，式(6)において，逆行列が存在しない，あるいは，存在したとしても汎化性能が著しく悪化するという問題が生じる．3.2節で説明したような自動評価スコアは，参照要約とシステム要約との間の文字列の一致割合に基づき計算されるため，それらの間の相関は高い．よって，これらを説明変数として用いる場合には注意しなければならない．そこで，[平尾 07]では，情報量基準の観点から良い回帰モデルを複数選択し，それらの回帰モデルが予測したスコアの平均を最終的な予測スコアとする手法を提案している．

先の例を用いて説明する．まず，与えられた説明変数集合$\boldsymbol{E}$ = {ROUGE-1, ROUGE-S, ESK}から下記の空でない可能な部分集合を求める．

表2 各モデルに対するAIC$_c$, ΔAIC$_c$, 予測値

モデル	変数セット	AIC$_c$	ΔAIC$_c$	予測値
M_1	S_1	-17.4	13.8	0.82
M_2	S_2	-27.1	4.1	0.71
M_3	S_3	-19.1	12.1	0.99
M_4	S_4	-26.6	4.6	0.68
M_5	S_5	-29.7	1.5	0.70
M_6	S_6	-19.3	11.9	0.61
M_7	S_7	-31.2	0	0.79

- {ROUGE-1}
- {ROUGE-S}
- {ESK}
- {ROUGE-1, ROUGE-S}
- {ROUGE-1, ESK}
- {ROUGE-S, ESK}
- {ROUGE-1, ROUGE-S, ESK}

次に，それぞれの説明変数セットを用いて回帰モデル$M_1, M_2, \cdots, M_7$を構築し，以下の式を用いてAIC [Akaike 73]を改良したAIC$_c$ (Corrected-AIC) [Sugiura 78]を求める．ここで，pは説明変数の数である．

$$\text{AIC}_c(M)=\text{AIC}(M)+\frac{2p(p+1)}{n-p-1} \quad (7)$$

なお，AIC(M)は以下の式で定義される．

$$\left.\begin{aligned} &\text{AIC}(M)=-2\text{MLL}+2p \\ &\text{MLL}=-\frac{n}{2}\left\{(1+\log(2\pi\sigma^2))+\log\left(\frac{R_p}{n}\right)\right\} \end{aligned}\right\} \quad (8)$$

R_pは残差平方和（訓練データに対する二乗誤差の和）である．AIC$_c$は小さければ小さいほどよい．ここで，AIC$_c^{\text{best}}$ ($=\min_i$AIC$_c(M_i)$)を求め，これとの差ΔAIC$_c$ ($=$AIC$_c(M_i)-$AIC$_c^{\text{best}}$)がある一定の幅τにおさまるすべての回帰モデルを選択する．そして，これらの回帰モデルがそれぞれ独立に予測した値の平均値を最終的な予測値とする．

回帰モデル$M_1 \sim M_7$に対して，AIC$_c$, ΔAIC$_c$, 予測値が表2のとおり得られ，$\tau=5$であるとすると最終的な予測値は，モデルM_2, M_4, M_5, M_7が予測した値の平均であるので，$(0.71+0.68+0.70+0.79)/4=0.72$となる．

[平尾 06]では，TSC3, DUC2004のデータを用いて評価実験を行った結果，単回帰モデルをはじめとする従来の回帰モデルよりも性能が良く，また，説明変数セットとして，ROUGE-2, ESK, 擬似的質問応答（詳しくは[Hirao 04]を参照されたい）などの組合せが良いと報告している．

5. おわりに

本稿では，テキスト要約におけるこれまでの自動評価法を「参照要約との類似性による自動評価法」と「マニ

ュアル評価結果を用いた自動評価法」に分類し，概観した．前者の評価法は，参照要約のみ用意すればシステム要約が評価可能であり，自動評価のコストはあまりかからない．しかし，各システム要約に与えられる自動評価スコアそのものの信頼性は低いため，システム間の性能差を粗く見積もりたい場合には有効であるが，あるシステムがどのようなトピックスを得意，不得意とするかというような特性を知るためには有効ではないという特徴をもつ．一方，後者はマニュアル評価結果をある程度プーリングする必要があることから，前者よりもコストはかかるが，評価法としての信頼性は前者よりも高いという特徴がある．利用者は，自動評価法としてのこのような特性の違いを理解し，目的によって適切に使い分ける必要がある．

ただし，マニュアル評価結果を用いた自動評価法には，まだ解決すべき課題が残されている．その一つは，「どの程度の種類と分量のシステム要約とそのマニュアル評価をプーリングしておけば十分であるか」である．テキスト中でどの情報が重要と考えるかは，要約率，要約の読み手の立場，テキストが客観的事実のみを述べたものか意見文も含んだものなのかなど，多くの事項が要因としてあげられる．例えば，要約の読み手の立場に関して，ある年に行われたテニスのウィンブルドン選手権に関する報道記事から要約を作成する場合を考えると，一般的には「誰が優勝したか」という観点で要約が作成される．しかし，この選手権に日本人選手が出場している場合には，要約の読み手が日本人であれば「誰が優勝したか」のほかに，「その日本人選手がどこまで勝ち進んだか」という点にもおそらく興味をもつであろう．この場合，この観点からの要約も作成可能である．このように，プーリングデータの種類は誰が要約の読み手であるのかによって変わってくると思われる．また，プーリングデータの種類と分量は，おそらく要約の課題が単一テキスト要約であるか，複数テキスト要約であるかにも依存するであろう．

上記のような問題は，要約固有の課題として解決する必要があるが，同時に，要約評価には一般的なテキスト評価の問題として扱えるものも存在する．実際に，要約の自動評価がほかの分野から影響を受けていることは，すでに本稿で述べたとおりである．これらのほかにも，本稿では取り上げなかったが，テキストの読みやすさに関する問題は，要約に限らず，テキスト生成系の研究全般に関するものである．今後のテキスト要約の評価法は，テキスト要約固有の問題と一般的なテキストの評価に関する問題に切り分けて議論し，後者に関しては可能な限り関連分野と連携していく必要があろう．

◇ 参 考 文 献 ◇

[Akaike 73] Akaike, H.: Information theory as an extension of the maximum likelihood principle, *Proc. 2nd International Symposium on Information Theory*, pp. 267-281 (1973)

[Cancedda 03] 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)

[福本 06] 福本淳一，加藤恒昭，桝井文人，森 辰則，神門典子：Basic Element を用いた質問応答の自動評価，情報処理学会情報学基礎研究会報告 FI-084, pp. 71-78 (2006)

[Hirao 04] Hirao, T., Okumura, M., Fukushima, T. and Nanba, H.: Text summarization challenge 3 . Text summarization evaluation at NTCIR Workshop 4, *Working Notes of the 4th NTCIR Workhop Meeting*, pp. 407-411 (2004)

[平尾 06] 平尾 努，奥村 学，磯崎秀樹：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，情報処理学会論文誌，Vol. 47, No. 6, pp. 1753-1766 (2006)

[平尾 07] 平尾 努，奥村 学，安田宜仁，磯崎秀樹：投票型回帰モデルによる要約の自動評価法，人工知能学会論文誌，Vol. 22, No. 2B, pp. 115-126 (2007)

[Hovy 06] Hovy, E., Lin, C.-Y., Zhou, L. and Fukumoto, J.: Automated summarization evaluation with basic elements, *Proc. 5th Conference on Language Resources and Evaluation* (2006)

[賀沢 03] 賀沢秀人，Arrigan, T.，平尾 努，前田英作：文書要約における抽出単位と評価法についての考察，情報処理学会自然言語処理研究会報告 NL-158，pp. 25-30 (2003)

[Lin 03] Lin, C.-Y. and Hovy, E.: Automatic evaluation of summaries using N-gram cooccurrence statistics, *Proc. 4th Meeting of the North American Chapter of the Association for Computational Linguistics and Human Language Technology*, pp. 150-157 (2003)

[Lin 04a] 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)

[Lin 04b] 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)

[Lodhi 02] 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)

[Luhn 58] Luhn, H.: The automatic creation of literature abstracts, *IBM Journal of Research and Development*, Vol. 2, No. 2, pp. 159-165 (1958)

[Mani 99] Mani, I., Gates, B. and Bloedom, E.: Improving Summaries by revising them, *Proc. 37th Annual Meeting of the Association for Computational Linguistics*, pp. 558-565 (1999)

[Nanba 00] Nanba, H. and Okumura, M.: Producing more readable extracts by revising them, *Proc. 18th International Conference on Computational Linguistics*, pp. 1071-1075 (2000)

[Nanba 06] Nanba, H. and Okumura, M.: An automatic method for summary evaluation using multiple evaluation results by a manual method, *Proc. COLING/ACL 2006 Main Conference Poster Sessions*, pp. 603-610 (2006)

[Nenkova 07] Nenkova, A., Passonneau, R. and McKeown, K.: The pyramid method: Incorporating human content selection variation in summarization evaluation, *ACM Trans. on Speech and Language Processing*, Vol. 4, Issue 2 (2007)

[Papineni 02] Papineni, K., Roukos, S., Ward, T. and Zyu, 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)

[Sugiura 78] Sugiura, N.: Further analysis of the data by Akaike's information criterion and the finite corrections, *Theory and*

Methods, Vol. A7, No. 1, pp. 13-26 (1978)
[Tombros 98] Tombros, A. and Sanderson, M.: Advantages of query biased summarization in information retrieval, *Proc. 21st Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pp. 2-10 (1998)
[Yasuda 03] Yasuda, K., Sugaya, F., Takezawa, T., Yamamoto, S. and Yanagida, M.: Applications of automatic evaluation methods to measuring a capability of speech translation system, *Proc. 10th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 371-378 (2003)

2007年11月2日 受理

著者紹介

難波 英嗣（正会員）

1996年東京理科大学理工学部電気工学科卒業．1998年北陸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．2001年同研究科博士後期課程修了．同年，日本学術振興会特別研究員．2002年，東京工業大学精密工学研究所助手．同年，広島市立大学情報科学部講師．現在に至る．博士（情報科学）．テキストマイニング，情報検索，自動要約に関する研究に従事．言語処理学会，情報処理学会，ACL，ACM各会員．

平尾 努

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，(株)NTTデータ入社．2000年より，NTTコミュニケーション科学基礎研究所に所属．博士（工学）．自然言語処理の研究に従事．情報処理学会，言語処理学会，ACL各会員．