投票型回帰モデルによる要約の自動評価法

平尾 努 † 奥村 学 †† 安田宜仁 † 磯崎 秀樹 †
† 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
{hirao,n-yasuda,isozaki}@cslab.kecl.ntt.co.jp
†† 東京工業大学 精密工学研究所 oku@pi.titech.ac.jp

1 はじめに

要約システムを評価する場合，人間による評価が最も信頼できる．しかし，時間や金銭的コストを考えると大規模に繰り返し行うことは困難である．よって，要約システムの効率的な改善のためには，自動評価法が欠かせない．この時，システム改善のための重要な知識，たとえば，どういった要約課題をシステムが苦手とするか，を知ることができれば，より効率的なシステム改善が可能となるであろう．こうしたシステム改善のための知見を得るためには，個々のシステム要約に対して，人間の評価を正確に再現する必要がある．しかし，従来より提案されている自動評価法では，個々のシステム要約に対する評価への信頼性が低いため，こうしたシステム改善への重要な知見が得られるとは限らない．より人間の評価に近づけるためには，人間の評価結果を従属変数，従来の自動評価法によるスコアを説明変数として線形回帰モデルを考えれば良い．ただし，線形回帰モデルには，多数の説明変数を用いると過学習や多重共線が起きるという問題がある．そこで，本稿ではこうした問題を解決するため，信頼性の高い n-best の回帰モデルを用いる投票型回帰モデルを提案する．与えられた説明変数の集合から可能な回帰モデルを作成し，それらのうち信頼性の N 個の回帰モデルを用いて予測を行う．提案手法を単回帰モデルをベースラインとして比較した結果，約 8%誤差を削減できた．

2 従来の自動評価法

従来より提案されている自動評価法は，内的自動評価法と外的自動評価法にわけることができる．以下，それぞれについて説明する．

2.1 内的自動評価法

一般的に，内的自動評価法は参照要約とシステム要約との間の類似度をスコアとする．多くの場合，スコアは参照要約とシステム要約との間で一致する N グラム数に基づく．

ROUGE-N [5, 4]

ROUGE-N は，参照要約中の N グラム数に対する参照要約とシステム要約間で一致する N グラム数の割合をスコアとする．以下の式で定義される．一般的には N=1,2 がよいとされる．

$$\text{ROUGE-N}(\mathcal{C},\mathcal{R})=\frac{\text{Count}_{\text{match}}(\text{gram}_{\text{N}})(\mathcal{C},\mathcal{R})}{\#\text{ of N-grams} \in \mathcal{R}} \quad (1)$$

$\text{Count}_{\text{match}}(\text{gram}_{\text{N}})(\mathcal{C},\mathcal{R})$ は，システム要約 ($\mathcal{C}$) と参照要約 ($\mathcal{R}$) との間で一致する N グラムの数を返す関数である．

ROUGE-S [4, 6]

ROUGE-2 を改良したものである，バイグラムだけでなく，スキップを許したバイグラムも考慮してスコアを計算する．以下の式で定義される．

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

ここで，$\text{R}_{\text{skip2}}(\mathcal{C},\mathcal{R})$，$\text{P}_{\text{skip2}}(\mathcal{C},\mathcal{R})$，はそれぞれ以下の式で定義される．

$$\text{R}_{\text{skip2}}(\mathcal{C},\mathcal{R})=\frac{\text{Skip2}(\mathcal{C},\mathcal{R})}{\#\text{ of skip bigrams} + \#\text{ of bigrams} \in \mathcal{R}}, \quad (3)$$

$$\text{P}_{\text{skip2}}(\mathcal{C},\mathcal{R})=\frac{\text{Skip2}(\mathcal{C},\mathcal{R})}{\#\text{ of skip bigram} + \#\text{ of bigrams} \in \mathcal{C}}. \quad (4)$$

Skip2($\mathcal{C}$,$\mathcal{R}$) は，$\mathcal{R}$ と $\mathcal{C}$ の間で共通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．

ROUGE-SU [4]

ROUGE-SU は ROUGE-S に対してユニグラムを素性として追加したものである [4]．以下の式で定義される．

$$\text{ROUGE-SU}(\mathcal{C},\mathcal{R})=\frac{(1+\beta^2)\times R_{\text{su}}(\mathcal{C},\mathcal{R})\times P_{\text{su}}(\mathcal{C},\mathcal{R})}{R_{\text{su}}(\mathcal{C},\mathcal{R})+\beta^2 P_{\text{su}}(\mathcal{C},\mathcal{R})} \quad (5)$$

ここで，R_{su}，P_{su} はそれぞれ以下の式で定義される．

$$\text{R}_{\text{su}}(\mathcal{C},\mathcal{R})=\frac{\text{SU}(\mathcal{C},\mathcal{R})}{(\#\text{ of (skip) bigrams} + \#\text{ of unigrams}) \in \mathcal{R}} \quad (6)$$

$$\text{P}_{\text{su}}(\mathcal{C},\mathcal{R})=\frac{\text{SU}(\mathcal{C},\mathcal{R})}{(\#\text{ of (skip) bigrams} + \#\text{ of unigrams}) \in \mathcal{C}} \quad (7)$$

SU($\mathcal{C}$,$\mathcal{R}$) は，$\mathcal{R}$ と $\mathcal{C}$ の間で一致するユニグラム，バイグラム，スキップバイグラムの数を返す関数である．

ROUGE-L [4, 6]

ROUGE-L は最長一致部分列 (LCS) に基づきスコアを計算する [4]．以下の式で定義される．

$$\text{ROUGE-L}(\mathcal{C},\mathcal{R})=\frac{(1+\beta^2)\times R_{\text{lcs}}(\mathcal{C},\mathcal{R})\times P_{\text{lcs}}(\mathcal{C},\mathcal{R})}{R_{\text{lcs}}(\mathcal{C},\mathcal{R})+\beta^2 P_{\text{lcs}}(\mathcal{C},\mathcal{R})} \quad (8)$$

ここで，R_{lcs}，P_{lcs} はそれぞれ以下の式で定義される．

$$R_{\text{lcs}}(\mathcal{C},\mathcal{R})=\frac{1}{u}\sum_{r_i\in\mathcal{R}}\text{LCS}_{\cup}(r_i,\mathcal{C}) \quad (9)$$

$$P_{\text{lcs}}(\mathcal{C},\mathcal{R})=\frac{1}{v}\sum_{r_i\in\mathcal{R}}\text{LCS}_{\cup}(r_i,\mathcal{C}) \quad (10)$$

ここで，$\text{LCS}_{\cup}(r_i,\mathcal{C})$ は，参照要約中の文 r_i とシステム要約との間のユニオン LCS スコアをあらわす．u，v はそれぞれ参照要約に含まれる単語数とシステム要約に含まれる単語数である．ユニオン LCS については，文献 [4, 6] を参照されたい．

ESK-based method [3]

Extended String subsequence Kernel (ESK) [3] は，N グラムに加え，スキップ N グラムを考慮してスコアを計算できる．さらに，単語の意味ラベルを用いることが可能である．ただし，ROUGE-S(U) とは異なり，スキップを含む N グラムに対して，減衰パラメータ λ を用いてその重みを軽くする．ESK を用いたスコアは以下で定義される．ESK の定義に関しては文献 [3] を参照されたい．

$$F^d_{\rm esk}(\mathcal{C},\mathcal{R})=\frac{(1+\beta^2)\times R_{\rm esk}(\mathcal{C},\mathcal{R})\times P_{\rm esk}(\mathcal{C},\mathcal{R})}{R_{\rm esk}(\mathcal{C},\mathcal{R})+\beta^2\times P_{\rm esk}(\mathcal{C},\mathcal{R})} \quad (11)$$

$P^d_{\rm esk}(\mathcal{C},\mathcal{R})$, $R^d_{\rm esk}(\mathcal{C},\mathcal{R})$ はそれぞれ以下の式で定義される．

$$P^d_{\rm esk}(\mathcal{C},\mathcal{R})=\frac{1}{\ell}\sum_{i=1}^{\ell}\max_{1\le j\le m}{\rm Sim}^d_{\rm esk}(c_i,r_j) \quad (12)$$

$$R^d_{\rm esk}(\mathcal{C},\mathcal{R})=\frac{1}{m}\sum_{j=1}^{m}\max_{1\le i\le \ell}{\rm Sim}^d_{\rm esk}(c_i,r_j) \quad (13)$$

ここで，${\rm Sim}^d_{\rm esk}$ はカーネルの値を [0,1] の間に納まるように正規化したものであり，ℓ はシステム要約の文数，m は参照要約の文数を表わす．

2.2 外的自動評価法

外的評価とは，要約をあるタスクに適用し，そのタスクの遂行度合いで要約を間接的に評価することである．タスクとしては，質問応答がしばしば用いられる [8, 7]. あらかじめ用意した質問集に対し，被験者がシステム要約を読むことでどの程度正しく回答できたかで評価を行う．ただし，こうした被験者による質問応答タスクはコスト的に高いことから，以下のように，擬似的に外的評価を行う手法が提案されている．

Exact Match [2]

質問集に対する回答文字列をシステム要約が完全一致で含む割合．以下の式で定義される．

$$\mathrm{PR}_{\rm exact}(\mathcal{C},\mathcal{R})=\frac{\#\text{ of exact answers}\in\mathcal{C}}{\#\text{ of questions}} \quad (14)$$

Edit Distance [2]

質問集に対する回答文字列と似た文字列を含む割合．以下の式で定義される．

$$\mathrm{PR}_{\rm edit}(\mathcal{C},\mathcal{R})=\frac{\sum_a \mathrm{edit}(a)}{\#\text{ of questions}} \quad (15)$$

edit は編集距離であり，以下の式で定義される．

$$\mathrm{edit}(a)=\max_s\frac{\mathrm{L}(s)-\mathrm{E}(s,a)}{\mathrm{L}(s)} \quad (16)$$

$\mathrm{L}(s)$ はシステム要約中の文 s の文字数，$\mathrm{E}(s,a)$ は，s と回答文字列 a との編集距離をあらわす．

2.3 従来の自動評価法の問題点

前節で説明した自動評価法は人間の評価と高い相関があることが知られている．いま，被験者を i ($1\le i\le I$)，要約システムを j ($1\le j\le J$)，要約課題を t ($1\le t\le T$) として説明する．

要約課題 t に対する j 番目のシステムの要約を $\mathcal{C}_{t,j}$ とし，t に対する参照要約集合を $\boldsymbol{\mathcal{R}}_t=\{\mathcal{R}_{t,1},\ldots,\mathcal{R}_{t,i},\ldots,\mathcal{R}_{t,I}\}$ とする．このとき，システム要約 $\mathcal{C}_{t,j}$ に対する自動評価スコアは以下の式で定義される．

$$f(\mathcal{C}_{t,j},\boldsymbol{\mathcal{R}}_t)=\frac{1}{I}\sum_{i=1}^{I}f(\mathcal{C}_{t,j},\mathcal{R}_{t,i}). \quad (17)$$

f は前節で説明した自動評価法のいずれかである．同様にして，被験者による主観スコアは以下の式で定義される．

$$H(\mathcal{C}_{t,j},\boldsymbol{\mathcal{R}}_t)=\frac{1}{I}\sum_{i=1}^{I}h_i(\mathcal{C}_{t,j},\mathcal{R}_{t,i}). \quad (18)$$

h_i は，i 番目の被験者による評価をあらわす．従来の自動評価法は，J 次元の主観スコアのベクトル $\mathbf{v}_h$ と自動評価スコアのベクトル $\mathbf{v}_a$ との間の相関で評価される．$\mathbf{v}_h$ の要素は，T 個の要約課題におけるシステムの自動評価スコアの平均である．すなわち，j 番目の要素は，$\sum_{t=1}^{T}f(\mathcal{C}_{t,j},\boldsymbol{\mathcal{R}}_t)/T$ となる．同様に，$\mathbf{v}_h$ の要素は，T 個の要約課題における主観スコアの平均なので，その j 番目の要素は，$\sum_{t=1}^{T}h_i(\mathcal{C}_{t,j},\boldsymbol{\mathcal{R}}_t)/T$ となる．

表 1 に，TSC-3 のデータ ($I=5, J=10, T=30$) を用いた場合の $\mathbf{v}_a$ と $\mathbf{v}_h$ の間のピアソンの積率相関係数を示す．表 1 より，自動評価スコアの平均と主観スコアの平均の間の相関係数は非常に高く，その有効性がよくわかる．しかし，この結果は，個々のシステム要約に与える自動評価スコアの信頼性が高いことを保証しない．いま，j 番目のシステムに対して，T 次元の自動評価ベクトル $\mathbf{v}_{aj}$ を主観評価ベクトル $\mathbf{v}_{hj}$ との間の相関係数を計算する．$\mathbf{v}_{aj}$ と $\mathbf{v}_{hj}$ の t 番目の要素は，j 番目のシステムの t 番目の要約課題に対する自動評価スコアと主観評価スコアなのでそれぞれ，$\mathbf{v}_{aj}=(f(\mathcal{C}_{1,j},\boldsymbol{\mathcal{R}}_1),f(\mathcal{C}_{2,j},\boldsymbol{\mathcal{R}}_2),\cdots,f(\mathcal{C}_{30,j},\boldsymbol{\mathcal{R}}_{30}))$, $\mathbf{v}_{hj}=(H(\mathcal{C}_{1,j},\boldsymbol{\mathcal{R}}_1),H(\mathcal{C}_{2,j},\boldsymbol{\mathcal{R}}_2),\ldots,H(\mathcal{C}_{30,j},\boldsymbol{\mathcal{R}}_{30}))$ となる．表 2 に各自動評価手法における 10 システムの平均相関係数と標準偏差を示す．

表 2 より，個々のシステム要約に対して与えられた主観スコアと自動評価スコアの間の相関係数は驚くほど低いことがわかる．よって，冒頭で述べたようなシステム改善のための知見をここから得ることは難しい．

表 1: 主観スコアの平均と自動評価スコアの平均の間の相関係数

手法	相関係数
ROUGE-1	.901
ROUGE-2	.946
ROUGE-3	.905
ROUGE-L	.884
ROUGE-S	.943
ROUGE-SU	.943
ESK($d=2$)	.971
ESK($d=3$)	.957
PR(exact)	.933
PR(edit)	.884

表 2: システムごとの主観スコアと従来の自動評価スコアの相関係数

手法	相関係数
ROUGE-1	.470 ± .168
ROUGE-2	.592 ± .129
ROUGE-3	.614 ± .111
ROUGE-L	.352 ± .130
ROUGE-S	.489 ± .133
ROUGE-SU	.494 ± .132
ESK($d=2$)	.589 ± .102
ESK($d=3$)	.623 ± .087
PR(exact)	.302 ± .101
PR(edit)	.314 ± .119

3 投票型回帰モデルによる自動評価

前節で説明したように，従来の自動評価法が個々のシステム要約に与えるスコアは信頼性が低い．これを改善するためには，人間が与えた主観スコアを従属変数，従来の自

動評価スコアを説明変数とした線形回帰モデルを考えればよい．ただし，線形回帰モデルには，説明変数を増やすと過学習や多重共線が起るという問題がある．そこで，本稿では投票型回帰モデルを提案し，この問題を解決する．

3.1 線形回帰モデル

訓練データとして，$\{y_1, \mathbf{x}_1\}, \{y_2, \mathbf{x}_2\}, \ldots, \{y_n, \mathbf{x}_n\}$ が与えられると線形回帰モデルは．以下の式で定義される．

$$y_\ell = \mathbf{x}'_\ell \boldsymbol{\beta} + \epsilon_\ell, \;\; \epsilon_\ell \sim \text{i.i.d. } N(0, \sigma^2), \quad \ell = 1, 2, \cdots, n \tag{19}$$

ここで，$\mathbf{x}_\ell$ は p 個の説明変数からなるベクトルであり，$\mathbf{x}_\ell$=$(1, x_{\ell,1}, x_{\ell,2}, \cdots, x_{\ell,p-1})$ となる．$\boldsymbol{\beta}$ は，回帰ベクトルをあらわし，$\boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_{p-1})$ となる．y_ℓ $(1{\leq}\ell{\leq}n)$ は従属変数，$\epsilon_{\ell,t}$ は，平均 0，分散 σ^2 に従う誤差である．回帰ベクトルは，訓練データにおける二乗誤差を最小化することで求められる．

よって，人間の評価により近い正確な自動評価法を実現するには，従属変数として人間の主観スコアを採用し，説明変数ベクトルとして従来の自動評価スコアを採用すれば良い．ただし，訓練データ数が少ない場合，説明変数の数が多いと過学習が起りやすい．また，相関の高い説明変数の組がある場合には予測性のが著しく低下するという問題がある．

3.2 モデル選択

一般的に，p 個の説明変数が与えられた場合，$2^p{-}1$ の回帰モデルを構成することが可能である．この時，多くのモデルは先に説明した問題を抱えている．こうした場合，情報量基準を用いることで良い回帰モデルを選択することができる．情報量基準としては，AIC (Akaike Information Criterion) [1] がよく知られている．AIC の値が低ければ低いほど良いモデルである．ある回帰モデル $\mathcal{M}$ の AIC は以下の式で定義される．

$$\text{AIC}(\mathcal{M}) = -2\text{MLL} + 2p, \tag{20}$$
$$\text{MLL} = -\frac{n}{2}\{(1 + \log 2\pi\sigma^2) + \log(\frac{R_p}{n})\}$$

R_p は残差平方和であり，モデル $\mathcal{M}$ における y_ℓ の予測値を $\hat{y_\ell}$ とすると，$\sum_{\ell=1}^{n}(y_\ell - \hat{y_\ell})^2$ となる．すなわち，訓練データにおける二乗誤差である．AIC の第 1 項は回帰モデルにおける誤差を評価し，第 2 項はモデルの複雑さに対するペナルティである．つまり，誤差が小さくともモデルが複雑 (説明変数が多い) な場合には，AIC は大きな値をとる傾向にある．AIC は，サンプル数が多い場合にはうまく働くことがよく知られている．しかし，サンプル数が小さい場合には，AIC を改善した correct-AIC [9] がより良い．Correct-AIC は以下の式で定義される．

$$\text{AIC}_c(\mathcal{M}) = \text{AIC}(\mathcal{M}) + \frac{2p(p+1)}{n-p-1} \tag{21}$$

AIC との違いは，サンプル数を考慮したペナルティになっている点である．文献 [9] によると，サンプル数が 20 以下の場合には correct-AIC を用いることが推奨されている．

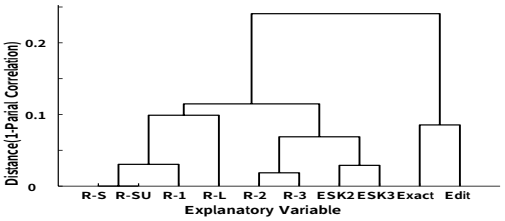

図 1: 説明変数間の関係

表 5: システムごとの主観スコアと自動評価スコアの相関係数 (単回帰モデル)

説明変数	相関係数
ROUGE-1	.677 ± .0735
ROUGE-2	.744 ± .0894
ROUGE-3	.722 ± .0722
ROUGE-L	.721 ± .0424
ROUGE-S	.727 ± .0929
ROUGE-SU	.727 ± .0943
ESK($d=2$)	**.765 ± .0498**
ESK($d=3$)	.744 ± .0730
PR(exact)	.636 ± .1217
PR(edit)	.636 ± .0696

3.3 投票型回帰モデルを用いた自動評価

一般的に，AIC や correct AIC の値がある程度の幅に納まるモデルの間には差がない．そこで，これらの回帰モデルが予測した値を統合する投票型回帰モデルを提案する．以下にその手続きを示す．

Step 0: 与えられた p 個の説明変数に対し，可能な部分集合を求める．これを $\boldsymbol{\mathcal{F}}_{\mathbf{all}}$: $\{\boldsymbol{\mathcal{F}}_1, \boldsymbol{\mathcal{F}}_2, \cdots, \boldsymbol{\mathcal{F}}_{2^p-1}\}$ とする．

Step 1: $\boldsymbol{\mathcal{F}}_{\mathbf{all}}$ のすべての要素を用いて訓練を行い，回帰モデルを得る．これを $\mathcal{M}_{all}$:$\{\mathcal{M}_1, \mathcal{M}_2, \cdots, \mathcal{M}_{2^p-1}\}$ とする．

Step 2: 以下のとおり，すべてのモデルに対して correct AIC を計算し，最良のモデルを決定する．
$\mathcal{M}_{best} = \text{argmin}_{\mathcal{M}_x \in \mathcal{M}_{all}} \text{AIC}_c(\mathcal{M}_x)$

Step 3: Step 2 で決定した最良モデル $\mathcal{M}_{best}$ とそれ以外のモデルとの correct AIC の差 (ΔAIC_c) を以下ので計算し，その値がある閾値 $\mathcal{T}$ 以下のモデル集合を $\boldsymbol{\mathcal{M}}_{\text{cand}}$ を得る．
$\Delta\text{AIC}_c(\mathcal{M}){=}\text{AIC}_c(\mathcal{M}){-}\text{AIC}_c(\mathcal{M}_{best})$

Step 4: モデル $m_k \in \boldsymbol{\mathcal{M}}_{\text{cand}}$ を用いて予測を行い $\hat{y_k}$ を得る．$\hat{y_k}$ の平均，つまり，$\frac{1}{|\boldsymbol{\mathcal{M}}_{\text{cand}}|}\sum_{k=1}^{|\boldsymbol{\mathcal{M}}_{\text{cand}}|}\hat{y_k}$ を最終的な予測とする．

4 評価実験

4.1 実験の設定

評価実験には TSC-3 のデータを用いた．先にも述べたが，各要約課題に対して，参照要約の作成とシステム要約の評価を行った被験者は 5 名 ($I = 5$)．要約システム数は 10($J = 10$)，要約課題 (トピック) 数は 30($T = 30$) である．各システム要約に対し人間の主観スコアをどれだけ忠実に再現できたかを評価するため，主観スコアと自動評価スコアの誤差の絶対値，$|y - \hat{y}|$ を用いた．また，システムを固定したときの主観スコアと自動評価スコア間の相関も用いた．なお，比較手法として，説明変数としてただ一つの従来の自動評価スコアを用いる単回帰モデルを用いた．提案手法，比較手法とも各課題 t に対して， leave-one-out 法で訓練，テストを行った．

提案手法に対しては，初期状態として与える説明変数を

表 3: 主観スコアに対する誤差の絶対値 (単回帰モデル)

説明変数	s1	s2	s3	s4	s5	s6	s7	s8	s9	s10	mean $\pm \sigma$
ROUGE-1	.0391	.0527	.0604	.0607	.0572	.0510	.0508	.0501	.0687	.0508	.0542 $\pm$ .0080
ROUGE-2	**.0347**	.0490	**.0467**	.0482	**.0529**	.0548	.0486	.0394	**.0467**	.0400	**.0461** $\pm$ .0063
ROUGE-3	.0433	.0639	.0492	.0483	.0558	.0581	.0520	.0429	.0551	.0446	.0513 $\pm$ .0069
ROUGE-L	.0561	.0613	.0556	.0577	.0622	.0565	.0542	.0488	.0759	.0454	.0574 $\pm$ .0083
ROUGE-S	.0390	.0426	.0506	.0558	.0566	.0564	.0448	.0405	.0504	.0444	.0481 $\pm$ .0067
ROUGE-SU	.0391	**.0425**	.0507	.0558	.0565	.0563	.0447	.0405	.0505	.0442	.0481 $\pm$ .0067
ESK($d=2$)	.0413	.0437	.0541	.0448	.0548	.0516	.0524	**.0351**	.0583	**.0377**	.0474 $\pm$ .0079
ESK($d=3$)	.0471	.0490	.0497	**.0439**	.0615	**.0506**	.0548	.0421	.0722	.0379	.0509 $\pm$ .0100
PR(exact)	.0643	.0666	.0581	.0592	.0621	.0723	**.0415**	.0429	.0593	.0623	.0588 $\pm$ .0097
PR(edit)	.0516	.0724	.0623	.0601	.0592	.0622	.0493	.0505	.0665	.0554	.0589 $\pm$ .0074

表 4: 主観スコアに対する誤差の絶対値 (投票型回帰モデル)

	$\mathcal{T}$	s1	s2	s3	s4	s5	s6	s7	s8	s9	s10	mean $\pm \sigma$
$\mathcal{F}_1$	14	.0402	.0467	.0459	.0424	.0542	.0486	.0430	.0356	.0502	.0351	.0442 $\pm$ .0006
$\mathcal{F}_2$	12	.0386	.0456	.0468	.0442	.0536	**.0479**	.0418	.0341	**.0467**	.0357	.0435 $\pm$ .0007
$\mathcal{F}_3$	13	.0390	.0462	.0459	.0420	.0530	.0483	.0427	**.0330**	.0485	.0328	.0431 $\pm$ .0005
$\mathcal{F}_4$	15	**.0383**	.0433	**.0435**	**.0394**	**.0525**	.0481	.0426	.0338	.0505	**.0312**	**.0423 $\pm$ .0004**
$\mathcal{F}_5$	12	.0397	**.0431**	.0443	.0428	.0542	.0506	**.0417**	.0356	.0481	.0351	.0435 $\pm$ .0005
$\mathcal{F}_6$	11	.0386	.0470	.0470	.0421	.0545	.0488	.0415	.0332	.0511	.0380	.0442 $\pm$ .0013

表 6: システムごとの主観スコアと自動評価スコアの相関係数 (投票型回帰モデル)

説明変数	相関係数
$\mathcal{F}_1$	.779$\pm$ .0500
$\mathcal{F}_2$	.778 $\pm$ .0486
$\mathcal{F}_3$	.787 $\pm$ .0597
$\mathcal{F}_4$	**.800 $\pm$ .0444**
$\mathcal{F}_5$	.780 $\pm$ .0611
$\mathcal{F}_6$	.781 $\pm$ .0594

決定するため，説明変数間の偏相関に基づき階層クラスタリング (平均法) を行い，その関係を調べた．偏相関の計算には, 1 つのシステムを取り除いた 270(9 システム $\times$ 30 課題) サンプルを用いた．図 1 に結果を示す．図より，説明変数は，5 つのクラスタ：{ROUGE-1,S,SU}, {ROUGE-L}, {ROUGE-2,3}, {ESK(d=2, 3)},{$\mathrm{PR_{exact}}$,$\mathrm{PR_{edit}}$} に分割できることがわかる．よって，以下の説明変数セットを用いた．

$\mathcal{F}_1$: ROUGE-2, ESK(d=2), ROUGE-S,$\mathrm{PR_{exact}}$,ROUGE-L

$\mathcal{F}_2$: ROUGE-1,ROUGE-2, ESK(d=2), ROUGE-S,$\mathrm{PR_{exact}}$

$\mathcal{F}_3$: ROUGE-2, ROUGE-3,ESK(d=2), ROUGE-S,$\mathrm{PR_{exact}}$

$\mathcal{F}_4$: ROUGE-2, ESK(d=2), ESK(d=3),ROUGE-S,$\mathrm{PR_{exact}}$

$\mathcal{F}_5$: ROUGE-2, ESK(d=2), ROUGE-S, SU, $\mathrm{PR_{exact}}$

$\mathcal{F}_6$: ROUGE-2, ESK(d=2), ROUGE-S,$\mathrm{PR_{exact}}$,$\mathrm{PR_{edit}}$

また，$\Delta\mathrm{AIC}_c$ に対する閾値 $\mathcal{T}$ は，1,2,. . .,15 まで変化させた．

4.2 実験結果

表 3 に単回帰，表 4 に提案手法における誤差の絶対値の平均を示す．また，表 5 には，単回帰モデル，表 6 には，投票型回帰モデルにおけるシステムごと主観スコアと自動評価スコアの相関係数を示す．

単回帰モデルでは，誤差で評価した場合，ROUGE-2 が最も良い成績である．次いで，ESK(d=2)，ROUGE-S(U) の順に成績が良い．システムごとの主観スコアと自動評価スコアの相関係数で評価した場合には，ESK(d=2) が最も良く，次いで，ESK(d=3)，ROUGE-2 が良い．双方の場合において，外的評価法である，$\mathrm{PR_{exact}}$，$\mathrm{PR_{edit}}$ はともに成績が悪い．表 2 と比較すると，どの手法も大幅に相関係数が改善されており，回帰モデルを用いる有効性がわかる．

一方，提案手法では，表 4 より，どの説明変数セットも単回帰で最も成績の良い ROUGE-2 を上回っている．最も成績の良い $\mathcal{F}_4$ を用いた場合には，単回帰で最も良い ROUGE-2 と比較して，誤差の絶対値を約 8%小さくできている．また，6 より，主観スコアとの間の相関係数についても同様であり，$\mathcal{F}_4$ を用いた場合は，単回帰で最も良い ESK(d=2) より 3.5 ポイント程度高い．評価実験に用いた説明変数セットは，単回帰では，成績の良くない ROUGE-L や $\mathrm{PR_{exact}}$ を含んでいるにもかかわらず成績が良いことから，correct-AIC を用いたモデル選択の有効性がよくわかる．

5 まとめ

本稿では，correct-AIC を用いて決定した N 個の回帰モデルを組合せる投票型回帰モデルを用いた要約の自動評価法を提案した．TSC-3 のデータを用いて提案手法を評価した結果，ベースラインである単回帰モデルに対し，誤差の絶対値を約 8%削減できた．また，システムを固定した場合の主観スコアとの相関の平均は 0.8 を達成した．

参考文献

[1] K. Burnham and D. Anderson. *Model Selection and Multimodel Inference*. Springer, 1998.

[2] T. Hirao, M. Okumura, T. Fukushima, and H. Nanba. Text Summarization Challenge 3. In *Proc. of the NTCIR-4*, pages 407–411, 2004.

[3] T. Hirao, M. Okumura, and H. Isozaki. Kernel-based Approach for Automatic Evaluation of Natural Language Generation Technologies: Application to Automatic Summarization. In *Proc. of the HLT/EMNLP*, pages 145–152, 2005.

[4] C-Y. Lin. ROUGE: A Package for Automatic Evaluation of Summaries. In *Proc. of Workshop on Text Summarization Branches Out*, pages 74–81, 2004.

[5] C-Y. Lin and E. Hovy. Automatic Evaluation of Summaries Using N-gram Co-occurrence Statistics. In *Proc. of the NAACL/HLT*, pages 150–157, 2003.

[6] C-Y. Lin and F.J. Och. Automatic Evaluation of Machine Translation Quality Using Longest Common Subsequence and Skip-Bigram Statistics. In *Proc. of the ACL*, pages 606–613, 2004.

[7] I. Mani, G. Klein, D. House, L. Hirschman, T. Firman, and B. Sundheim. SUMMAC: a text summarization evaluation. *Natural Language Engineering*, 8(1):43–68, 2002.

[8] A. H. Morris, G. M. Kasper, and D.A. Adams. The Effects and Limitations of Automatic Text Condensing on Reading Comprehension. *Information System Research*, 3(1):17–35, 1992.

[9] N. Sugiura. Further Analysis of the Data by Akaike's Information Criterion and the Finite Corrections. *Communication in Statistics*, Theory and Methods(A7):13–26, 1978.