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

2004－SLP－54（22）
2004／12／21

〔招待講演〕

# 自動要約評価型ワークショップ：Text Summrization Challenge (TSC)

平尾　努†

† 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
E-mail: †hirao@cslab.kecl.ntt.co.jp

**あらまし**　近年，自動要約に関する研究に注目が集まっており，盛んに研究が行われている．そのなかでも，特に，個々の文書を要約する単一文書要約よりも，ある基準によって集められた複数の文書を一度に要約する複数文書要約技術に大きな期待がかかっている．本稿では，複数文書要約を主タスクとして採用した自動要約の評価型ワークショップである TSC-3 (2004 年 6 月にワークショップ開催) に関して，コーパス，タスク，評価指標，システムの評価結果についてそれぞれ詳細を述べる．

**キーワード**　複数文書要約，評価型ワークショップ，内的評価，外的評価

# Text Summarization Challege (TSC)

## –An Evaluation Workshop on Automatic Summarization–

Tsutomu HIRAO†

† Nippon Telegraph and Telepone Corp. NTT Communication Science Laboratories
Hikari-dai 2-4, Seika-cho, Soraku-gun, Kyoto, 619-0237 Japan
E-mail: †hirao@cslab.kecl.ntt.co.jp

**Abstract**　It has been said that we have too much information on our hands, forcing us to read through a great number of documents and extract relevant information from them. With a view to coping with this situation, research on automatic text summarization has attracted a lot of attention recently and there have been many studies in this field. There is a particular need to establish methods for the automatic summarization of multiple documents rather than single documents. In this paper, we introduce an evaluation workshop for multiple document summarization which called Text Summarization Challenge (TSC) and describe corpus, tasks and evaluation measures.

**Key words**　Multiple Document Summarization, Evaluation Workshop, Intrinsic Evaluation, Extrinsic Evaluation

## 1. まえがき

電子化テキストの氾濫による情報の洪水という言葉が使われて久しいが，現在でも，我々は多くのテキストに目を通し，そして情報の取捨選択を行わなければならない．こうした状況を背景として，近年，自動要約の研究に注目が集まり，盛んに研究が行われている．特に，個々の文書を独立に要約する単一文書要約よりも，ある基準によって集められた複数の文書を一度に要約する複数文書要約技術の確立に大きな期待が寄せれらている．

このような背景のもと，いくつかの自動要約の評価型ワークショップが開催されている．米国では 1998 年に TIPSTER SUMMAC [4] が開催され，その後 2001 年からは，Document Understanding Conference (DUC)(注1) が毎年開催されており，DUC は，初回より複数文書要約タスクを中心的な課題として採用している．一方，日本では，2001 年より NTCIR プロジェクトの一環として Text Summarization Challenge (TSC)(注2) が約 1 年半に 1 度開催されており，2 回目にあたる TSC-2 (2002 年に開催) では初めて複数文書要約タスクを採用している．このように複数文書要約は現在の自動要約研究の中心的課題として位置づけられる．

本稿では，複数文書要約をタスクとして採用した自動要約の評価型ワークショップである TSC-3 (2004 年 6 月にワークショップ開催) のために構築したテストセットとタスク概要，用

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

いた評価指標，要約システムの評価結果について述べる．

## 2. タスク設定とコーパス構築の指針

複数文書要約は，単一文書要約と比較すると，要約対象となるテキストの規模 (文字数や文数) が大きいため，一般的に難しい．更に，複数の情報源から得た文書を一度に要約する場合，個々の文書から得た情報に重複があるなど冗長性が問題となる．よって，自動要約システムには，こうした冗長な文をどれだけ認定し，要約として出力する際に削減できているかということが求められる．

こうした背景にもかかわらず，DUC や TSC-2 で作成されたコーパスでは冗長情報の削減が有効でに働いたかどうかを確かめる術はない．TSC-2 では，ある決まった文字数以内で人間が自由に作成した要約 (アブストラクト) しか正解データとして与えられないため，繰り返し自動的に評価することが困難であり，文抽出の評価もできない．また，DUC でも，決まった単語数以内で作成されたアブストラクトが正解として与えられるので，TSC-2 の場合と同様である．唯一，2002 年の DUC ではエクストラクトが作成されてはいるので，文抽出の評価は行えるが，先に述べたような同一意味の文に対するアノテーションがないため，冗長文の削減効果を測ることはできない．その是非はともかく，現状の自動要約システムの多くが文抽出を基盤としていることを考えると，これらのコーパスが評価用テストセットとして適しているとはいえない．

TSC-3 ではこのような状況を鑑みて，まず，複数文書要約システムの要約過程が以下のステップからなると仮定し(注3)，各ステップにおいて要約システムを評価できるコーパスを作成し，タスクを設定した．

Step1　与えられた文書セットから重要文を抽出する．
Step2　抽出した文集合から冗長な文を削る．
Step3　Step2 までで得た文集合を与えられた文字数以内に収まるように書き換えを行う．

ここで，Step1，Step2 に対しては**抜粋作成タスク**を設定し，専用スコアラによって coverage，precision を用いた内的評価(注4)を行った．Step3 に対しては**要約作成タスク**を設定し，内的評価としては，人間による内容，読みやすさの評価を行い，外的評価として，専用スコアラを用いた擬似的質問応答による評価を行った．それぞれ，以下に詳述する．

## 3. 抜粋作成 (重要文抽出) タスク

抜粋とは原文書の一部 (文や節など) を抽出したものを指す．TSC-3 では，システムが文を抽出をするタスク (一般的には重要文抽出と呼ばれる) を抜粋作成タスクとした．抜粋作成タスクでは，正解が文として特定できるので，専用のスコアラを用意し，評価を行った．また，TSC-3 では 2 種の異なる長さ (short, long) での評価を行った．

(注3)：もちろん，要約システム作成に制約を設けるものではなく，一般的な，要約システムがこうしたステップを踏むであろうという考えに基づいている．
(注4)：内的評価とは要約そのものを直接評価することを言う．たとえば，その良し悪しを人間が評価することが該当する．一方，外的評価とは，あるタスクに要約を用いた場合のパフォーマンスの良し悪しで要約を間接的に評価することを言う．たとえば，情報検索タスクを用いて要約を評価する手法がこれにあたる．

表 1　重要文データ

| アブストラクト文 ID | 対応付けられた文集合 |
|---|---|
| 1 | $\{s_1\}_\sqcup\{s_{10},s_{11}\}$ |
| 2 | $\{s_3,s_5,s_6\}$ |
| 3 | $\{s_{21},s_{23}\}_\sqcup\{s_1,s_{30},s_{60}\}$ |

### 3.1 重要文抽出データの作成

まず，重要文のアノテーションの方針について述べる．一般的に，抜粋には以下の 2 種があると考える．

(1)　人間が文書 (セット) から直接，重要と判断した文の集合 [1], [8], [10],

(2)　人間が作文した要約 (アブストラクト) の材料として適切な文の集合．すなわち，アブストラクトに対して対応づけられた元テキストの文集合 [2], [3], [5], [9].

ここで，人間の要約作成過程を考えると，重要文を抽出し，その後，それらを書き換えて要約を作成するというよりも，文書 (セット) の重要情報を認定し，それらを自由に組み合わせ作文して要約を作成すると考えた方が自然である．よって，TSC-3 では上記 (2) が適切であると考え，これに従い抜粋を作成した．

ただし，前節でも説明したように複数文書を対象とした要約の場合，同一内容の文が複数存在するため，アブストラクトの 1 文に対応する元テキストの文集合(注5)(すなわち，アブストラクトの 1 文に対する重要文の正解) は一つとは限らない．よって，対応文としてふさわしい全ての文集合に対してアノテーションを施した．例えば，人間の作成したアブストラクトの文 $a$ の対応文は，

(1)　文書 $x$ の $s_1$，あるいは，
(2)　文書 $y$ の $s_2, s_3$ の組合せ

のようにしてアノテーションを行った．これは，$s_1$ 単独で $a$ を生成できるし，$s_2, s_3$ を組み合わることでも $a$ を生成できることを表す．

このように，あるアブストラクトの文に対してそれを生成するための素材として適した元テキストの文をもれなく抽出することで，Step2 においてシステムが冗長文を削減できたかどうかを自然に考慮できることとなる．つまり，システム出力が同じアブストラクトの文に対応付けられた元テキストの文を抽出している場合は冗長であり，そうでない場合には冗長でないといえる．先の例において，システムが $s_1, s_2, s_3$ を出力すると，これらは全てアブストラクトの文 $a$ に対応するので，冗長な例である．

### 3.2 評価指標

#### 3.2.1 システムが抽出すべき文数

抜粋の評価には一般的には Precision，Recall などが用いられ，TSC-1 においては，PR Breakeven Point (Precision=Recall の場合) で評価が行われた [1]．これは，抽出すべき文の数，すなわち，正解抜粋の文数，が既知であるとして，システムがその数だけ文を抽出した際に含まれる正解の割合である．これに従い，TSC-3 でも抽出すべき文数が既知であるとして，システムがその数だけ文を抽出した場合で評価を行うこととした．

しかし，TSC-3 コーパスでは複数の重要文の正解が存在する

(注5)：一般的にアブストラクト 1 文に対して元テキストの 2 文以上が対応することも多いので「集合」という言葉を用いた．

ので，唯一の重要文の正解を持つ TSC-1 コーパスのようにシステムが抽出すべき文の数を定めることは容易ではない．そこで，以下の方法を考えた．

いま，表 1 のようにアブストラクトと原文書の文が対応付けられたとする．半角スペース「␣」は対応文集合の区切り文字である．このように複数文書要約では，あるアブストラクトの文に対して複数の対応文集合があることが多い．ここで，正解抜粋とは，アブストラクトを生成するために必要な文の集合であるから，表 1 の例では，$s_1, s_3, s_5, s_6, s_{30}, s_{60}$ や $s_{10}, s_{11}, s_3, s_5, s_6, s_{21}, s_{23}$ などがそれに該当する(注6)．複数の候補がある場合には，最小の文数で最大の情報を伝えることが望ましいので，結局，正解抜粋は「アブストラクトを生成するために必要最小限な文の集合」と定義し，システムはその要素数だけ文を抽出することと定めた．

アブストラクトを作成するために必要最小な文集合を求めることは，制約充足の問題に帰着でき，簡単に解くことができる．表 1 の例では，アブストラクトの各文から

- $s_1 \vee (s_{10} \wedge s_{11})$,
- $s_3 \wedge s_5 \wedge s_6$,
- $(s_{20} \wedge s_{21} \wedge s_{23}) \vee (s_1 \wedge s_{30} \wedge s_{60})$

という制約条件を得て，これらの連言が全て真であるという制約充足問題の最小カバーを求めれば良い．各制約条件を $C_1, C_2, C_3$ とおくと $C_1 \wedge C_2 \wedge C_3 = true$ という制約条件を満す最小カバーを考えれば良い．この場合，$\{s_1, s_3, s_5, s_6, s_{30}, s_{60}\}$ が最小カバーとなるのでシステムは 6 文抽出すればよい．実際に制約充足問題を解く際には BEM-II [7] を用いた．

TSC-3 では，上述した手法でシステムが抽出すべき文の数を定めた上で，以下の Precision と Coverage でシステムを評価する．

### 3.2.2 Precision

Precision はシステムが出力した文のうち，対応づけられた文集合に含まれる文の割合である．以下の式で定義する．

$$\text{Precision} = \frac{m}{h} \tag{1}$$

ここで，$h$ は，制約充足問題を解いて得たアブストラクトを生成するために必要な最小の文数，$m$ は，システムが $h$ 文出力したなかでの正解文数．ただし，ここでの正解文とは，表 1 にエントリされている文を指す．

いま，システムが，「$s_{10}, s_{11}, s_5, s_{17}, s_{60}, s_{61}$」を抽出したとすると，

$$\text{Precision} = \frac{4}{6} = 0.667 \tag{2}$$

となり，「$s_1, s_{10}, s_{11}, s_3, s_5, s_{60}$」を抽出した場合には，

$$\text{Precision} = \frac{6}{6} = 1 \tag{3}$$

となる．

(注6)：実際には，これら以外の文の組み合わせでもアブストラクトは生成可能である

### 3.2.3 Coverage

Coverage はシステムが出力した文集合中の冗長度合を考慮しつつ，それがアブストラクトの内容にどれだけ近いかを測る指標である．

いま，アブストラクトの $i$ 番目の文に対応する元テキストの文集合のリストを $A_{i,1}, A_{i,2}, \cdots, A_{i,j}, \cdots, A_{i,\ell}$ のように表わす．この場合，文 $i$ に対しては $\ell$ 個の対応文集合が存在することとなる．$A_{i,j}$ は元テキストの文 (番号) を要素とする集合であり，表 1 の例では，$A_{1,2} = \{s_{10}, s_{11}\}$ となる．

ここで，システム出力の文集合を $S$ として表し，$i$ 番目の文に対する評価値 $e(i)$ を以下の (1) 式で定義する．

$$e(i) = \max_{1 \leq j \leq \ell} \left( \frac{|S \cap A_{i,j}|}{|A_{i,j}|} \right) \tag{4}$$

ただし，$v(\alpha)$ は以下の式で定義する．

$$v(\alpha) = \begin{cases} 1 & \text{if the system outputs } \alpha \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

関数 $e$ は，アブストラクトの $i$ 番目の文に対する対応文集合 $A_{i,j}$ のうちいずれかを完全な形で出力していた場合には 1，部分的に出力していた場合には文数 $|A_{ij}|$ に応じて部分点を与える関数である．

上記 $e$ とアブストラクトの文数 $n$ を用いて，coverage は以下の式で定義する．

$$\text{Coverage} = \frac{\sum_{i=1}^{n} e(i)}{n} \tag{6}$$

表 1 の例で，システムが，「$s_{10}, s_{11}, s_5, s_{17}, s_{60}, s_{61}$」を抽出したとすると，

$$\begin{aligned} e(1) &= \max(0, 1) = 1 \\ e(2) &= \max(0.33) = 0.33 \\ e(3) &= \max(0, 0.33) = 0.33 \end{aligned}$$

となり，coverage=0.553 となる．また，「$s_1, s_{10}, s_{11}, s_3, s_5, s_{60}$」を抽出した場合には，

$$\begin{aligned} e(1) &= \max(1, 1) = 1 \\ e(2) &= \max(0.67) = 0.67 \\ e(3) &= \max(0, 0.67) = 0.67 \end{aligned}$$

となるので，coverage=0.780 となる．

## 4. 要約作成 (生成) タスク

抜粋作成タスクは，システムがある決まった数の文を抽出するタスクであったことに対して，要約作成タスクでは，システムがある与えられた文字数を上限として要約を作成するタスクである．

この際，システムには文字数という制約が与えられるのみであり，それを作成する手段には何ら制限はない．つまり，重要文抽出結果をそのままこの要約作成に用いることが可能である

し，もちろん，文生成等の深い処理を行って要約を生成することでも構わない．

システムに与えられる文字数の制限は，約5%，10%の2種である．

### 4.1 評価指標

アブストラクトの評価に関しては，内的評価として，それが要約としてどの程度情報をカバーしているかという観点と文章としてどれだけ読みやすいかという観点がある．両観点の評価ともいまのところは，人手による評価に頼らざるを得ない状況にある．しかし，人間による主観評価に関しては評価の一貫性がしばしば問題となる．そこで，TSC-3では，リファレンスサマリ(注7)の作成者自信を評価者として評価の一貫性をなるべく保つようにした．

また，試験的に外的評価として，擬似的質問応答を用いた自動的な内容評価も導入した．

### 4.2 内容評価

#### 4.2.1 人間による内容評価 (Information Coverage)

基本的には，リファレンスサマリの内容をシステムサマリがどの程度含んでいるかを数値化したものである．

評価者は自らが作成したリファレンスサマリの各文に対して，システムサマリの文を用いることでそれの情報をどれだけカバーした文を生成できるかという観点から評価を行い，平均をとったものを最終的な評価とする．

いま，あるアブストラクトが$n$文で構成されており，それに対応するシステムサマリの文集合を$\Gamma$とする．この場合，評価者は以下の手順に従い，評価値を決定する．

Step1 リファレンスサマリの文$s_i(1 \leqq i \leqq n)$に対して，関連するシステムサマリの文集合$\gamma$ ($\subseteq\Gamma$)を見つける．

Step2 $s_i$の情報を$\gamma$を用いてどの程度カバーできるかを$0, 0.1, 0.2, \cdots, 1$の11段階で評価値を与える．

Step3 すべての$i$に対してStep1，Step2を適用し，その平均をシステムサマリのスコアとする．

上記手順を適用し，課題数で正規化したものを最終的なシステムスコアとした．

#### 4.2.2 擬似的質問応答による自動的な内容評価

各リファレンスサマリに対して，その作成者が重要だと判定した事項を解答とする質問をあらかじめ作成し，システムサマリが正答を含む割合を自動的に評価した．1トピックあたり，質問数はshortで約5問，longで約10問である．

また，質問を作成する際には，リファレンスサマリ内で文を越えずに連続する一連の文字列を解答とするように制限を設けた．

なお，解答文字列を完全に含むか否かの2値の場合 (exact) と解答文字列との編集距離に基づく場合 (Edit)(以下の式) とで評価を行った．

$$\mathrm{Edit}(a) = \max_{s} \frac{\mathrm{L(s)} - \mathrm{E(s,a)}}{\mathrm{L(s)}} \qquad (7)$$

ここで，$a$は解答文字列，$s$はシステムサマリの文，Lは，文の長さを表す関数，Eは，文と解答文字列との編集距離を表す関数である．

(注7)：評価時に正解として用いる要約．モデルサマリとも言う．

表2 TSC3コーパスの詳細

| | |
|---|---|
| 文書セット数 | 30 |
| 文書数 (毎日) | 175 |
| 文書数 (読売) | 177 |
| 合計 | 352 |
| 文数 | 3587 |

### 4.3 読みやすさの評価

要約には当然，文章としての読みやすさも要求される．DUCでは，文章の読みやすさに影響を与えるであろう因子に関してそのエラー数を問う質問集 (Quality Questions，以下，QQ) を用意し評価を行っている．TSC-3でもこれにならい，以下のQQを用いて読みやすさの評価を行った．これらの質問は，DUC版QQをそのまま流用したものに対し，TSC-3参加者のメーリングリストで議論を行い決定したものを加えたものである．

なお，qq01〜qq15に関しては，qq00を適用し，不要な文を除いた後に適用する．

qq00 同一の、あるいはほぼ重複する文はいくつあるか？
qq01 (ゼロ) 代名詞化、指示表現化すべき箇所はいくつあるか？
qq02 先行詞のない指示表現はいくつあるか？
qq03 固有表現に対する修飾語の出現箇所の誤りがいくつあるか？
qq04 同一事物を参照する表現の一貫性という観点から修正すべき表現はいくつあるか？
qq05 (前後の文脈も踏まえた上で) 必須要素が欠如している箇所はいくつあるか？
qq06 接続詞が必要・不必要な箇所はいくつあるか？
qq07 副詞・形容詞などで不要な語はいくつあるか？
qq08 時系列の関係が矛盾してないか？
qq09 敬体 (「〜です。」等) と常体 (「〜である」「〜だ」等) が混在する等、文末表現の不統一はいくつあるか？
qq10 不適切な格要素の重複はいくつあるか？
qq11 呼応表現で不適切なものはいくつあるか？
qq12 不自然な語順の文はいくつあるか？
qq13 活用形に不備のある語はいくつあるか？
qq14 分割した方が良い重文・複文はいくつあるか？
qq15 統合した方が良い文集合はいくつあるか？

## 5. コーパスの詳細

毎日新聞，読売新聞98年〜99年の2年分を情報源として，先に述べた指針に従い，30の文書セット (トピック) に対してエクストラクトとアブストラクトを作成した．各文書セットの総文字数に対して約5%，10%の2種の要約率を設定し，アブストラクトを作成，それをもとにエクストラクトを作成した．データの詳細を表2に示す．

1文書セットあたり，平均で約10記事から成っており，毎日新聞と読売新聞の記事が割合はほぼ均等である．なお，トピックはそのほとんどがMcKeownらの分類[6]に従うとsingle-eventに分類される．下記に全てのトピックを挙げる．

0310 ２５０万年前の新種猿人の化石がエチオピアで発見されたことに関する記事群
0320 ＮＴＴ（とC&W）のＩＤＣ買収に関する記事群
0340 ゲームソフトの中古販売は適法であると東京地裁が判断したことに関する記事群
0350 インディペンデンス艦載機の夜間離着陸訓練（ＮＬＰ）に関する記事群
0360 タンザニア、ケニアでの米国大使館同時爆破事件に関する記事群
0370 スハルト大統領辞任に関する記事群
0480 ロシアの首相にプーチン氏が指名されたことに関する記事群
0400 オサマ・ビン・ラディン氏がアフガニスタンでタリバン政権にかくまわれ

表 3　抜粋作成タスクの評価結果

| ID | Short | | Long | |
|---|---|---|---|---|
| | Cov. | Prec. | Cov. | Prec. |
| F0301(a) | 0.376 | 0.471 | 0.429 | 0.535 |
| F0301(b) | 0.419 | 0.591 | 0.433 | 0.587 |
| F0303(a) | 0.251 | 0.314 | 0.371 | 0.432 |
| F0304 | 0.384 | 0.496 | 0.392 | 0.535 |
| F0306 | 0.372 | 0.449 | 0.432 | 0.545 |
| F0307 | 0.385 | 0.567 | 0.461 | 0.680 |
| F0309 | 0.367 | 0.505 | 0.416 | 0.585 |
| F0310 | 0.207 | 0.266 | 0.267 | 0.442 |
| LEAD | 0.264 | 0.426 | 0.315 | 0.539 |

ているとされることに関する記事群
0410　中田のペルージャ移籍に関する記事群
0420　ドリームキャスト発売に関する記事群
0440　ニホンカワウソの生存証拠が見付かったとされることに関する記事群
0450　京セラが三田工業を子会社化することに関する記事群
0460　台風によって壊れた室生寺（五重塔）に関する記事群
0470　ＹＳ－１１の引退に関する記事群
0480　天体望遠鏡「すばる」の試験観測開始に関する記事群
0500　クローン羊ドリーに関する記事群
0510　ニュートリノに質量があるとされることに関する記事群
0520　ヒトゲノムプロジェクト、第２２番染色体の解読完了に関する記事群
0530　９９年末の北アイルランド和平協議に関する記事群
0540　新型新幹線（７００系）デビューに関する記事群
0550　青島幸男氏が知事選不出馬を決めたことに関する記事群
0560　関西大学の入試ミスに関する記事群
0570　スペースシャトル、エンデバーの打ち上げから帰還までに関する記事群
0580　京大の研究グループがミャンマーで４０００万年前の新種サルの化石を発見したことに関する記事群
0590　ジョージ・マロリー氏の遺体がエベレストで発見されたことに関する記事群
0600　ＡＩＢＯ（アイボ）発売に関する記事群
0610　ｉＭａｃのそっくりさんｅ－ｏｎｅに関する記事群
0630　キトラ古墳の調査再開に関する記事群
0640　パプアニューギニアの地震による津波被害に関する記事群
0650　ＮＡＴＯの中国大使館誤爆に関する記事群

## 6. 評価結果と考察

これまでに説明したタスク設定，コーパスを用いて TSC-3 のフォーマルランを 2003 年 11 月 17 日～11 月 25 日にわたって開催した．この際，参加者に開示される情報は，各トピックのタイトルと記事集合，抜粋作成タスクにおける文数 (2 種)，要約作成タスクにおける文字数制限 (2 種) に加え，4.2.2 節での質問集 (この情報を使うか否かは任意)，4.3 節の QQ である．

抜粋作成タスクには，7 チーム，9 システムに加え，オーガナイザが用意したベースラインシステム (Lead 手法) が参加し，要約作成タスクには，9 チーム，9 システムに加え先述したベースラインシステム，リファレンスサマリとは異なる人間が作成した要約を加えた．以下の各結果は，各要約システムが出力した 30 個の要約を評価し，その平均をとったものである．

### 6.1 抜粋作成タスクの評価結果

表 3 に抜粋作成タスクの評価結果を示す．一般的に新聞記事を対象とした重要文抽出では，Lead 手法が良い成績であることが知らせているが，表 3 より，ほとんどのシステムが Lead 手法を上回っている．複数の文書，情報源を対象とした場合には Lead 手法では多くの情報がカバーできないことが原因であると考える．また，各システムとも precision に比べて coverage が大幅に小さい．これは，システム出力に冗長な文が多く残されていることを示している．また，short，long とも似たような結果となっている．

表 4　人間による要約作成タスクの評価結果 (内容)

| ID | Short | Long |
|---|---|---|
| F0301 | 0.319 | 0.298 |
| F0303 | 0.236 | 0.311 |
| F0304 | 0.318 | 0.322 |
| F0306 | 0.290 | 0.330 |
| F0307 | 0.365 | 0.392 |
| F0308 | 0.271 | 0.273 |
| F0309 | 0.280 | 0.300 |
| F0310 | 0.151 | 0.261 |
| F0311 | 0.273 | 0.278 |
| LEAD | 0.215 | 0.221 |
| HUMAN | 0.474 | 0.522 |

表 5　擬似的質問応答による要約作成タスクの評価結果 (内容)

| ID | Short | | Long | |
|---|---|---|---|---|
| | exact | edit | exact | edit |
| F0301 | 0.394 | 0.677 | 0.399 | 0.706 |
| F0302 | 0.257 | 0.556 | 0.266 | 0.602 |
| F0304 | 0.367 | 0.653 | 0.356 | 0.677 |
| F0306 | 0.342 | 0.614 | 0.327 | 0.630 |
| F0307 | 0.439 | 0.710 | 0.442 | 0.751 |
| F0308 | 0.321 | 0.601 | 0.313 | 0.611 |
| F0309 | 0.390 | 0.684 | 0.356 | 0.633 |
| F0310 | 0.133 | 0.427 | 0.201 | 0.549 |
| F0311 | 0.304 | 0.579 | 0.308 | 0.628 |
| LEAD | 0.300 | 0.589 | 0.275 | 0.602 |
| HUMAN | 0.461 | 0.716 | 0.426 | 0.721 |

### 6.2 要約作成タスクの評価結果

#### 6.2.1 人間による内容評価

表 4 に要約作成タスクにおける人間による内容評価の結果を示す．傾向としてはほぼ表 3 と同様であり，ほとんどのシステムが Lead 手法を上回っている．これは多くのシステムが重要文抽出に基づいていることを考えると妥当である．また，人間の要約と比較するとシステム要約は明らかに質が劣っていることがわかる．

#### 6.2.2 擬似的質問応答による内容評価

表 5 に擬似的質問応答タスクによる要約の内容評価の結果を示す．これも，全体的な傾向は先に述べた 2 つの内容評価の結果と似ている．ただし，この評価手法はシステム要約中に解答文字列やそれに近い文字列があるか否かを評価するだけであって，それが質問に対する解答となる文脈を保持しているかどうかは評価していない．そのため，非常に粗い評価となっている．このことを考えると精確な評価とはなっていないが，おおざっぱな傾向をつかむ程度にはこうした外的評価も利用可能であると考える．

#### 6.2.3 QQ による読みやすさの評価

表 6 に QQ による読みやすさの評価結果を示す．人間とシステムを比較すると，特に qq00，qq01，qq04，qq08 において差が大きく開いている．qq00 より，システム出力に多くの冗長，不要文が存在していることを示してることがわかる．これは，抜粋作成タスクの結果と合致している．また，qq01 より，システムが人間のように効率的にゼロ代名詞などを扱えないことを示している．q04 では，同一の事象を指しながら異なる単語が要約中に多く混在している示し，qq08 では，要約として文のならびに問題があることを示している．上にあげた例は特に複数の情報源を対象とした複数文書要約に特徴的な例であり，今

表 6 QQ による読みやすさの評価結果

Short

| ID | q00 | q01 | q02 | q03 | q04 | q05 | q06 | q07 | q08 | q09 | q10 | q11 | q12 | q13 | q14 | q15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| F0301 | 0.333 | 1.033 | 0.600 | 0.333 | 2.333 | 0.900 | 0.633 | 0.933 | −0.567 | 0.500 | 1.767 | 0.100 | 0.000 | 0.000 | 0.100 | 0.133 |
| F0303 | 0.033 | 0.567 | 0.700 | 0.667 | 1.567 | 1.400 | 0.500 | 0.267 | −0.500 | 0.100 | 0.367 | 0.033 | 0.000 | 0.033 | 0.000 | 0.100 |
| F0304 | 0.200 | 1.333 | 0.533 | 0.333 | 3.067 | 0.467 | 0.733 | 1.067 | 0.000 | 0.033 | 2.467 | 0.000 | 0.000 | 0.033 | 0.067 | 0.233 |
| F0306 | 0.067 | 0.700 | 0.433 | 0.300 | 2.433 | 0.933 | 0.933 | 0.500 | −0.133 | 0.100 | 1.267 | 0.000 | 0.000 | 0.033 | 0.067 | 0.100 |
| F0307 | 0.700 | 0.633 | 1.200 | 0.600 | 2.367 | 1.267 | 0.767 | 0.567 | −0.300 | 0.200 | 0.967 | 0.067 | 0.000 | 0.000 | 0.067 | 0.100 |
| F0308 | 0.100 | 1.067 | 0.433 | 0.400 | 2.433 | 0.500 | 0.567 | 0.867 | 0.200 | 0.267 | 1.633 | 0.100 | 0.000 | 0.000 | 0.067 | 0.100 |
| F0309 | 0.167 | 1.100 | 0.133 | 0.300 | 1.433 | 0.667 | 0.667 | 0.867 | 0.133 | 0.033 | 1.867 | 0.000 | 0.000 | 0.033 | 0.033 | 0.133 |
| F0310 | 1.967 | 0.200 | 1.767 | 0.400 | 0.633 | 3.800 | 1.333 | 0.167 | −0.600 | 0.233 | 0.000 | 0.200 | 0.033 | 0.000 | 0.000 | 0.133 |
| F0311 | 0.167 | 1.233 | 0.767 | 0.267 | 2.567 | 2.800 | 0.600 | 0.667 | −0.600 | 0.567 | 1.833 | 0.000 | 0.000 | 0.067 | 0.033 | 0.233 |
| LEAD | 1.500 | 1.267 | 0.267 | 0.267 | 1.667 | 0.067 | 0.767 | 1.533 | 0.267 | 0.067 | 1.667 | 0.000 | 0.000 | 0.033 | 0.033 | 0.200 |
| HUMAN | 0.033 | 0.267 | 0.000 | 0.000 | 0.433 | 0.400 | 0.400 | 0.000 | 0.933 | 0.500 | 0.033 | 0.000 | 0.000 | 0.033 | 0.033 | 0.033 |

Long

| ID | q00 | q01 | q02 | q03 | q04 | q05 | q06 | q07 | q08 | q09 | q10 | q11 | q12 | q13 | q14 | q15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| F0301 | 0.200 | 1.600 | 1.133 | 0.433 | 5.533 | 2.100 | 1.000 | 1.900 | −0.833 | 0.733 | 2.967 | 0.500 | 0.000 | 0.000 | 0.067 | 0.300 |
| F0303 | 0.300 | 0.500 | 2.100 | 0.333 | 2.667 | 3.600 | 1.500 | 0.467 | −0.900 | 0.133 | 0.233 | 0.000 | 0.067 | 0.000 | 0.033 | 0.233 |
| F0304 | 0.433 | 1.700 | 0.933 | 0.100 | 5.133 | 1.300 | 1.100 | 1.800 | −0.500 | 0.167 | 3.900 | 0.033 | 0.033 | 0.033 | 0.033 | 0.333 |
| F0306 | 0.167 | 1.367 | 0.833 | 0.400 | 4.200 | 1.100 | 1.100 | 1.300 | −0.533 | 0.133 | 3.000 | 0.133 | 0.000 | 0.033 | 0.100 | 0.200 |
| F0307 | 1.100 | 1.200 | 1.367 | 0.633 | 4.667 | 1.233 | 1.133 | 1.233 | −0.633 | 0.167 | 3.067 | 0.033 | 0.000 | 0.033 | 0.033 | 0.300 |
| F0308 | 0.333 | 1.567 | 1.067 | 0.533 | 4.567 | 1.000 | 0.933 | 1.967 | −0.133 | 0.200 | 3.300 | 0.000 | 0.000 | 0.000 | 0.033 | 0.100 |
| F0309 | 0.367 | 1.500 | 0.600 | 0.467 | 2.833 | 2.033 | 0.967 | 1.567 | −0.267 | 0.067 | 3.533 | 0.100 | 0.000 | 0.067 | 0.033 | 0.233 |
| F0310 | 2.367 | 0.133 | 2.500 | 0.500 | 2.567 | 4.667 | 1.833 | 0.333 | −0.733 | 0.333 | 0.300 | 0.233 | 0.067 | 0.067 | 0.000 | 0.233 |
| F0311 | 0.700 | 1.433 | 1.400 | 0.400 | 5.133 | 4.500 | 0.767 | 1.967 | −0.933 | 0.933 | 4.367 | 0.067 | 0.033 | 0.033 | 0.033 | 0.400 |
| LEAD | 2.833 | 2.100 | 0.633 | 0.367 | 4.300 | 0.300 | 1.033 | 4.333 | −0.333 | 0.067 | 5.133 | 0.000 | 0.000 | 0.000 | 0.067 | 0.433 |
| HUMAN | 0.033 | 0.167 | 0.100 | 0.000 | 1.133 | 0.467 | 0.433 | 0.067 | 0.800 | 0.567 | 0.000 | 0.033 | 0.000 | 0.000 | 0.033 | 0.100 |

後，システムの改善が望まれる．

## 7. まとめ

本稿では，複数文書要約の評価型ワークショップである TSC-3 についてタスク概要，評価結果の詳細を述べた．複数文書要約は，異なる文書から得た情報をまとめるという作業である．よって，重要な情報を抽出することだけでなく，文章としての読みやすさと要約としての内容の双方に整合性が取れていることが必要とされる．評価結果からもわかるとおり人間による要約とシステムの要約では質の差が大きい．今後はこうした問題点に焦点をあてた研究開発が必要であると考える．

## 謝辞

データの使用を許諾いただいた毎日新聞社，読売新聞社に感謝致します．共に TSC の運営に尽力いただいたオーガナイザの方々，東京工業大学の奥村学氏，追手門学院大学の福島孝博氏，広島市立大学の難波英嗣氏に感謝致します．また，常日頃より有益なコメントをいただくシャープ株式会社の野畑周氏に感謝致します．さらに，TSC-3 の参加者みなさまに感謝致します．

## 文献

[1] T. Fukushima and M. Okumura. Text summarization challenge: Text summarization evaluation in japan. In *Proc. of the NAACL2001 Workshop on Automatic summarization*, pages 51–59, 2001.

[2] H. Jing and K. McKeown. The Decomposition of Human-Written Summary Sentences. *Proc. of the 22nd ACM-SIGIR*. pages 129–136, 1999.

[3] J. Kupiec, J Petersen. and F. Chen. A Trainable Document Summarizer. In *Proc. of the 18th SIGIR*, pages 68–73, 1995.

[4] 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.

[5] D. Marcu. The automatic construction of large-scale corpora for summarization research. *Proc. of the 22nd Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 137–144, 1999.

[6] K. McKeown, R. Barzilay, D. Evans, V. Hatzivassilogou, M. Y. Kan, B. Schiffman, and S. Teufel. Columbia multi-document summarization: Approach and evaluation. In *Porc. of Document Understanding Conference 2001*, 2001.

[7] S. Minato. BEM-II:an arithmetic boolean expression manipulator using bdds. *IEICE Trans. Fundamentals*, E76-A(10):1721–1729, 1993.

[8] C. Paice. Constructing Literature Abstracts by Computer: Techniques and Prospects. *Information Processing and Management*, 26(1):171–186, 1990.

[9] S. Teufel and M. Moens. Sentence Extraction as a Classification Task. In *Proc. of the ACL Workshop on Intelligent Scalable Text Summarization*, pages 58–65, 1997.

[10] K. Zechner. Fast generation of abstracts from general domain text corpora by extracting relevant sentences. In *Proc. of the 16th International Conference on Computational Linguistics*, pages 986–989, 1996.