TSC3 コーパスの構築と評価

平尾 努 † 奥村 学 †† 福島 孝博 ††† 難波 英嗣 ††††
† 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@cslab.kecl.ntt.co.jp)
†† 東京工業大学 精密工学研究所 (oku@pi.titech.ac.jp)
††† 追手門学院大学 文学部 (fukusima@res.otemon.ac.jp)
†††† 広島市立大学 情報科学部 (nanba@its.hiroshima-cu.ac.jp)

1 はじめに

電子化テキストの氾濫による情報の洪水ということが言われて久しいが，現在もなお我々は多くのテキストに目を通し，情報の取捨選択を行わなければならい状況におかれている．こうした状況を打破する重要技術として，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．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の確立に大きな期待が集まっている．近年，このような背景のもとに自動要約の評価型ワークショップが行われている．たとえば，米国では 1998 年に TIPSTER SUMMAC が開催され，その後 2001 年からは，Document Understanding Conference (DUC)[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)[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予定) のために構築したテストセットと用いた評価指標の詳細について述べ，簡単な複数文書要約システムによる予備実験の結果について報告する．

2 TSC3 コーパス

2.1 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，複数文書要約の特有の難しさとして，システムが冗長な情報をどれだけ削減できるかということが挙げられる[3]．単一の文書では，基本的にはある文に対して同一内容を表わす文は少ないが，複数の文書を集めた場合には同一の内容でありながら異なる字面の文や全く同一の文などが多数存在することとなる．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ことがほとんどなので，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[4]，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することとした．

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．

Step1，Setp2 に対しては，文書セット中の重要文に対するアノテーションを行い，それらのうち同一の意味をなす文に対してもアノテーションを行った．Step3 に対しては，人間が定められた文字数に従いアブストラクトを作成した．

実際に TSC-3 では，Step1，2 では文を単位とした正解データが存在するので，スコアリングプログラムによる自動評価，Step3 では，人間による主観評価を採用し，現在，参加システムの評価を行っている．以下では Step1，2 のために作成したエクストラクトとその評価指標について詳細を述べる．TSC-3 全体での評価結果に関しては別の機会に譲る．

2.2 重要文抽出データの作成

まず，Step1 での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，エクストラクトには以下の 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 [2]，
(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

[1] http://duc.nist.gov
[2] http://www.lr.pi.titech.ac.jp/tsc/
[3] 当然，同一事象を指す語の表記統一などの読みやすさに関連する重要な技術も必要となるが，ここでは扱わない．
[4] 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．

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	$\{s_{21},s_{23}\}_\sqcup\{s_1,s_{30},s_{60}\}$

応づけられた元テキストの文集合 [3].

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，これに従いエクストラクトを作成した．

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[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 評価指標

3.1 システムが抽出すべき文数

エクストラクト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 [2]．これは，抽出すべき文の数，すなわち，正解エクストラクトの文数，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，正解エクストラクトとは，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[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，正解エクストラクトは「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．

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 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．

3.3 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 ℓ 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，$A_{i,j} = \{\theta_{i,j,1}, \theta_{i,j,2}, \cdots, \theta_{i,j,k}, \cdots\}$ となる．表 1 の例では，$A_{1,2} = \{\theta_{1,2,1}, \theta_{1,2,2}\}$ であり，$\theta_{1,2,1} = s_{10}, \theta_{1,2,2} = s_{11}$ である．

ここで，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．

[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．

[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

$$e(i) = \max_{1 \le j \le \ell} \left(\frac{\sum_{k=1}^{|A_{i,j}|} v(\theta_{i,j,k})}{|A_{i,j}|} \right) \quad (4)$$

ただし，$v(\alpha)$ は以下の式で定義する．

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

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．

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

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．

3.4 Weighted Coverage

TSC3 のデータでは，人間が作成したアブストラクトの各文に対して「A，B，C」の順に重要度ランクが付与されている．この情報を用いて先に定義した Coverage に重要度を導入した Weighted Coverage を定義する．(6) 式を変形した以下の式で定義する．

$$\text{W.C.} = \frac{\sum_{i=1}^{n} w(r(i)) \times e(i)}{w(A)n(A) + w(B)n(B) + w(C)n(C)} \quad (7)$$

ここで，$r(i)$ は，アブストラクトの i 番目の文のランクを表し，$w(r(i))$ はランクに応じた重みを表す．$n(rank)$ はアブストラクト中の $rank$ の文の数を表す．表 1 の例で 1 文目がランク A，2 文目がランク B，3 文目がランク C であったとして，その重みを $w(A) = 1$，$w(B) = 0.5$，$w(C) = 0.3$ と設定する[7]．ここで，前節の例と同様に，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，

$$\text{WC} = \frac{1 \times 1 + 0.5 \times 0.3 + 0.3 \times 0.3}{1 \times 1 + 0.5 \times 1 + 0.3 \times 1} = 0.689 \quad (8)$$

となる．

[7] $w(r(i))$ の値は他にも考えられるが，ここでは 1/ランク順位で与えた．

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 予備実験

TSC3 で作成したコーパスの性質を明らかにするため，簡単な重要文抽出システムと抽出した重要文から冗長文を削減する手法を実装し，評価を行った．コーパスの詳細，重要文の抽出精度比較，冗長文削減の有効性を以下に述べる．

4.1 コーパスの詳細

TSC3 では 2 章で述べたコーパス構築の指針に従い，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 [4] 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

表 3: 重要文の抽出精度

手法	長さ	Prec.	Cov.	W.C.
Lead	Short	.426	.212	.326
	Long	.539	.259	.369
TF·IDF	Short	.497	.292	.397
	Long	.604	.325	.434

表 4: 冗長文削減の効果

手法	長さ	Prec.	Cov.	W.C.
クラスタリング	Short	.417	.299	.389
	Long	.528	.358	.461
MMR	Short	.454	.305	.399
	Long	.553	.374	.435

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

4.2　重要文の抽出精度比較

4.2.1　重要文抽出法

以下の手法を用いて重要文の抽出精度を比較した．

Lead 手法

文書セット中の文書を時系列の昇順でソートし，各文書の先頭から順に 1 文ずつ重要文として抽出する手法

TF·IDF 手法

文のスコアを単語重要度の和で定義し，スコアの高い文から順に重要文として抽出する手法 [5]．文のスコア $S(s_i)$ は以下の式で定義される．

$$\mathrm{S}(s_i) = \sum_{t \in s_i} w(t, DS) \tag{9}$$

$w(t, DS)$ は以下の式で定義する．

$$w(t, DS) = tf(t, DS) \cdot \log \frac{|DB|}{df(t)} \tag{10}$$

$tf(t, DS)$ は，単語 t の要約対象文書セット中での出現頻度，$df(t)$ は，t が出現する文書頻度，$|DB|$ は全文書数を表し，実際には，毎日新聞，読売新聞の 98 年から 99 年の全ての記事を用いて計算した．

4.2.2　評価結果

評価結果を表 3 に示す．両手法とも Precision の値に対して Coverage，Weighted Coverage が低い値となっており，抽出した文集合が冗長であることを示している．また，一般的に，新聞記事を対象とした場合，Lead 手法が良い性能であることが知られているが，TSC3 コーパスではそうした傾向はみられなかった．

4.3　冗長文削減の効果

前節の結果より，単純に文書セットから文を抽出するだけでは，冗長な文が多く含まれていることがわかる．そこで，冗長な文をなるべく避けて文を抽出する手法として，クラスタリングに基づく手法と Maximal Marginal Relevance (MMR)[1] に基づく手法を TF·IDF 法に適用し，その有効性を調べた．なお，両手法とも文間の類似度には，bag-of-words 表現によるコサイン類似度を用いた．

クラスタリングによる冗長文削減法

(9) 式を用いて文の重要度を決定した後，上位 $3h$ 文を対象に h 個のクラスタを得るまで ward 法によって階層的クラスタリングを行う．その後，各クラスタから最も高いスコアを持つ文を 1 文ずつ抽出する．

MMR による冗長文削減法

(9) 式を用いて文の重要度を決定した後，上 $3h$ 文を対象に MMR を用いてリランキングを行った後，1 位から h 位までの文を抽出する．

4.3.1　評価結果

評価結果を表 4 に示す．表 3 の TF·IDF の値と比較すると，クラスタリング，MMR ともに Precision の値は低くなっているが，Coverage の値は高くなっている．特に，Long の場合の改善幅が大きい．このことより，確かに冗長な文を削減することができ，結果的に Coverage を改善できていることがわかる．しかし，Weighted Coverage でみた場合，MMR では必ずしも改善されているとは言えない．

5　おわりに

本稿では，TSC3 で作成したコーパスと評価指標に関して，主に文抽出に関連する部分について詳細を述べた．複数文書要約を対象として，重要文とそれらの間で意味的に同一の文にアノテーションを施したデータは，これまでの要約コーパスにない特徴である．TSC3 コーパスは TSC3 参加者でなくとも，今後，国立情報学研究所と覚え書きを交すことで利用可能となる予定である．このコーパスが自動要約に興味を持つ研究者の役に立てば幸いであるし，分野の発展に貢献できればと願っている．

謝辞

コーパス作成の際に多大な協力と有益なご助言をいただいた (株) 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．また，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前田英作グループリーダ，磯崎秀樹氏，泉谷知範氏，賀沢秀人氏に感謝いたします．

参考文献

[1] J. Carbonell and J. Goldstein. The use of mmr,diversity-based reranking for reordering documents and producing summaries. In *Porc. of the 21th Annual International ACM SIGIR Conference on Research and Development in the Information Retrieval*, pages 335–336, 1998.

[2] 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.

[3] 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.

[4] 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.

[5] 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.