抜粋による複数文書要約を評価するためのコーパスと評価指標

平尾 努[†1] 奥村 学[†2] 福島孝博[†3]
難波英嗣[†4] 野畑 周[†5] 磯崎秀樹[†1]

複数文書要約の対象となる文書群には，ある文に対して，意味的に似通った文やまったく同じ文が含まれていることが多い．こうした傾向は，要約のための文書群を複数の情報源から得た場合に特に顕著である．しかし，従来のコーパスには，このようなよく似た文，あるいは同一の文の間に注釈付けが存在しない．これは，抜粋を評価するための指標を定義するうえで致命的な問題となる．本稿では，こうした冗長性を考慮したコーパスへの注釈付けの枠組みを提案し，それに基づき，抜粋の情報量を測る指標である被覆率，抜粋に含まれる重要文の冗長度を測る指標である重要文冗長率を提案する．これらの指標による抜粋の順位付けと被験者による順位付けとの間の順位相関係数は，ともに 0.7 以上であり，人間の順位付けとの間に高い相関があることが分かった．

Corpus and Evaluation Measures for Extractive Multiple Document Summarization

Tsutomu Hirao,[†1] Manabu Okumura,[†2] Takahiro Fukushima,[†3]
Hidetsugu Nanba,[†4] Chikashi Nobata[†5] and Hideki Isozaki[†1]

In multiple document summarization, input documents have many similar (or even identical) sentences. However, conventional corpora for multiple document summarization do not include links between similar sentences. This is a critical problem with regard to the definition of evaluation measures for sentence extraction. In this paper, we propose both annotation scheme for corpus and evaluation measures, “coverage” and “redundancy.” “Coverage” measures the content information of the system extract and “redundancy” measures the redundancy of the important sentences contained in system extract. We evaluate “coverage”and “redundancy” by comparing their ranking correlation coefficients with subjective human rankings. The results show that both measure attained enough high correlation coefficients, which were more than 0.7 correlation coefficients.

1. はじめに

現在の自動要約研究において，複数文書要約は重要な課題の 1 つとしてとらえられており，評価型ワークショップにおいてもそれが中心的な課題として採用されている．米国では 2001 年から，Document Understanding Conference（DUC）[☆]が毎年開催されており，初回より複数文書要約タスクをその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）[☆☆]が約 1 年半に 1 度開催されており，2 回目にあたる TSC2（2002 年に開催）より，複数文書要約タスクを採用している．

現状の要約システムは，人間のようなアブストラクト（生成に基づく要約）を自動的に生成するまでには至っておらず，多くの場合，抜粋[☆☆☆]を作成した後，文

†1 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
NTT Communication Science Laboratories, NTT Corporation
†2 東京工業大学精密工学研究所
Precision and Intelligence Laboratory, Tokyo Institute of Technology
†3 追手門学院大学国際教養学部英語コミュニケーション学科
Department of English, Faculty of International Liberal Arts, Otemon Gakuin Univsersity
†4 広島市立大学情報科学研究科
Faculty of Information Science, Hiroshima City Univsersity
†5 マンチェスター大学計算機科学科
School of Computer Science, University of Manchester, UK

☆ http://duc.nist.gov
☆☆ http://www.lr.pi.titech.ac.jp/tsc/
☆☆☆ 要約の元となるテキストの構成要素（文，句，単語など）を抽出したものを抜粋と呼ぶ．なお，本稿における抜粋とは文を抽出したものを指す．

短縮技術などを併用することで要約を作成している．つまり，抜粋を作成するためのモジュールは，現状の要約システムにおいて中心的役割を担っている．しかしながら，従来の評価型ワークショップで用いられたコーパス，採用された評価指標は，抜粋を評価するという観点から，必ずしも適正であるとはいえない．なぜなら，要約対象となる文書群に冗長な情報が多いこと，つまり，ある文に対して意味的に類似した文，あるいはまったく同一の文が存在することを考慮した注釈付けがされておらず，評価指標も冗長性を考慮していないからである．

そこで本稿では，こうした冗長性を考慮したコーパスへの注釈付けの枠組みを提案し，それに基づき，抜粋の情報量を測る指標である被覆率，抜粋が含む重要文の冗長度を測る指標である重要文冗長率を提案する．これらの指標による抜粋の順位付けと被験者による順位付けの間の順位相関係数はともに 0.7 以上であり，人間の順位付けとの間に高い相関があることが分かった．

2. 既存コーパスにおける注釈付けと評価指標の問題点

先に述べたとおり，複数文書要約における要約対象文書群には，同意味の文や意味内容が大きく重複している文が多く含まれている．これは，文書群の情報源が複数ある場合に特に顕著である．よって要約システムには，要約対象となる文集合から同一の関係にある文の組を認定することが望まれる．このような同等関係にある文の組を認定する研究としては，文献 2)，5)，6)，17) などがあり，異なる文書中の文間の関係を解析する CST（Cross-document Structure Theory）[14] の枠組みでも扱われている．ここで，同一の関係にある文の組を認定できるなら，要約システムは，そのどちらか一方のみを抜粋に含めればよい．つまり，要約システムには，同意味な文の組を認定するだけでなく，冗長な情報を排除することも望まれる．このような認識は研究者の間でも広く共有されている．これは，多くのシステムが冗長な情報を最小化する技術，たとえば，情報検索や複数文書要約などで使われる Maximal Marginal Relevance（MMR）[3] という技術や文のクラスタリングなどを用いていることからも分かる．

しかし，現状では，要約対象に意味的に似た文が存在すること（冗長な情報を含むこと）を前提として注釈付けを行ったコーパスは存在しない．さらにそれを前提として定義された評価指標も存在しない．

たとえば，**表 1** のように抜粋として選択された文（以降，これを重要文と呼ぶ）とそれに対して同意味の文があったと仮定する．ただし，m_* だけが重要文として注釈付けされており，y_* は，それに対して，同意味の文であるが，注釈付けはされていない文であるとする．また，y_1，y_2 はそれら 2 文をあわせて m_1 と同等ということを表す．以下のような例を想定されたい．

表 1 重要文として注釈付けされた文とそれに対して隠れた同意味の文
Table 1 Annotated important sentences and their hidden alternatives.

重要文	同意味の文
m_1	y_1，y_2
m_3	y_{10}
m_4	y_{21}

m_1： 野球の試合は台風がもたらした豪雨によって，中止となった．
y_1： 台風は豪雨をもたらした．
y_2： よって，野球の試合は中止となった．

ここで，以下の 3 つの抜粋を考える．

抜粋 A m_1，m_3，y_{10}，
抜粋 B m_1，m_3，m_4，
抜粋 C m_1，m_3，y_{21}

抜粋 A を抜粋に含まれる重要文の割合で評価すると，m_1，m_3 は重要文として注釈付けされていることから，その値は 2/3 となる．一方，抜粋 B を同様に評価すると 3 文とも重要文として注釈付けされているので，その値は 3/3 となり，抜粋 A よりも抜粋 B が優れていることが容易に分かる．この例は，我々の直感に合致する．しかし，抜粋 C を同様に評価するとその値は 2/3 となり，抜粋 A と同じ評価となる．m_4 と y_{21} が同意味であることに注意すると，抜粋 B と抜粋 C は等しく評価されるべきである．しかしながら，y_{21} が重要文として注釈が付けられていないことから，その評価は，抜粋 A と同じ評価を得てしまうこととなる．この例からも分かるとおり，重要文とそれに対する同意味の文の双方に注釈付けを行わなければならないことは明白である．

また，上記のように重要文とそれに対する同意味の文にも注釈付けを行うことは，評価尺度にも大きな影響を及ぼす．表 1 の例で y_* に対しても重要文であるという注釈付けが存在することを仮定する．この場合，抜粋 A と抜粋 B を先の例のとおりシステム出力中に占める重要文の割合で評価すると，ともにその値は 1 となってしまう．m_3 と y_{10} は同意味なので，明らかに前者は冗長であり，後者はそうでない．この区別が

できないことは評価指標として致命的である．よって単純にシステム抜粋に含まれる重要文の割合で評価することはできない．さらに，重要文とそれに対する同意味の文は 1 対 1 対応になるとは限らないので，表 1 のような対応関係がある場合には，長さ（文数）が異なる複数の正解抜粋が存在することとなる．たとえば，m_1，y_{10}，m_4 や y_1，y_2，m_3，y_{21} などがそうである．正解抜粋の長さを一意に決定できなければ，再現率を定義することができない．よって，既存の評価指標をそのまま適用することは不可能である．このように，既存コーパスには，注釈付け，評価指標の双方に大きな問題が存在する．

3. 抜粋を評価するための指標

本稿では，参照要約（人間が作成した正解要約）を生成するために必要な元テキストの文集合を抜粋として定義する．ここで，2 章で述べた注釈付けと評価指標の問題点を解決するため，参照要約中の 1 文に対し，それを生成するために必要な元テキストの文集合☆をもれなく対応付ける☆☆．この注釈付けに基づき，

- 情報の冗長性を考慮し，システム抜粋の情報量を評価する指標
- システム抜粋に含まれる重要文がどの程度冗長であるかを評価する指標

を定義する．

以下，正解抜粋の長さを定義した後，上記それぞれに対応する評価指標である**被覆率**，**重要文冗長率**の定義について説明する．

なお，抜粋を単なる文の集合と見なすのではなく，要約として利用するのであれば，抽出した文をどのような順で出力するかを考えなければならない．特に複数文書要約の場合，結束性を確保するために必須である[1]．ただし，本稿では要約としての抜粋を評価するのではなく，要約システムが最終的に要約を生成するために必要な文をどの程度抽出できたかを評価するという立場をとった．よって，文の順序については評価の際に考慮しないことに注意されたい．

3.1 抜粋の長さの決定

抜粋の評価には一般的には，精度，再現率などが用いられる．たとえば，TSC1 においては，PR Breakeven Point（精度=再現率の場合）で評価が行われた[4]．これは，抽出すべき文の数，すなわち正解抜粋の長さ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，そこに含まれる重要文の割合を表す．

しかし，先に述べたとおり，重要文とそれに対して同意味の文に対して注釈付けを行う場合，長さの異なる複数の正解抜粋が存在することがある．よって，TSC1 のように唯一の正解抜粋に基づき，システムが抽出すべき文の数を定め，精度と再現率で評価することはできない．そこで，以下の方法で正解抜粋の長さを決定した．

表 2 参照要約の文とそれに対応する重要文
Table 2 An example of alignment between an abstract sentence and extract sentences.

参照要約文 ID	対応付けられた文集合
a_1	$\{s_1\}\sqcup\{s_{10},s_{11}\}$
a_2	$\{s_3,s_5,s_6\}$
a_3	$\{s_{20},s_{21},s_{23}\}\sqcup\{s_1,s_{30},s_{60}\}$

表 2 のように参照要約と元テキストの文が対応付けられていることを想定する．半角スペース「␣」は対応文集合の区切り文字である．ここで，抜粋とは，参照要約を生成するために必要な元テキストの文集合であるから，表 2 の例では，$\{s_1, s_3, s_5, s_6, s_{20}, s_{21}, s_{23}\}$ や $\{s_{10}, s_{11}, s_3, s_5, s_6, s_{20}, s_{21}, s_{23}\}$ などがそれに該当する☆☆☆．このように複数の正解抜粋候補がある場合，要約という観点からは，最小の文数で最大の情報を伝えることが望ましいので，正解抜粋を「参照要約を生成するために必要最小限な文の集合」と定義した．

上に述べた参照要約を生成するために必要最小限な文集合を求めることは，制約充足の問題に帰着できる．表 2 の例では，参照要約の各文から

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

という制約条件を得て，これらの連言がすべ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よい．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よい．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 文抽出すればよいこととなる．実際に制約充足問題を解く際には BEM-II [10] を用いた．

3.2 被 覆 率

参照要約の i 番目の文 a_i に対応する元テキスト

☆ 一般的に参照要約の 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．

☆☆ 参照要約のうち約 19%が元テキストと対応付けが不可能であったという報告[8] があるが，後述するコーパスにおいて対応付けをとる際には特にこうした問題は起こらなかった．これは，要約作成者に対して，参照要約を作成した後に元テキストとの間で対応付けを行うことをあらかじめ知らせていたことが影響していると考える．

☆☆☆ 実際には，これら以外の文集合でも参照要約は生成可能である．

の文集合のリストを $A_{i,1}, A_{i,2}, \cdots, A_{i,j}, \cdots, A_{i,\ell}$ のように表す．この場合，文 a_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，表 2 の例では，$A_{1,1} = \{s_1\}$，$A_{1,2} = \{s_{10}, s_{11}\}$ となる．ここで，システム出力の文集合を E として表し，文 a_i に対する評価値 $e_i(E)$ を，以下の式 (1) で定義する．

$$e_i(E) = \max_{1 \le j \le \ell} \left(\frac{|E \cap A_{i,j}|}{|A_{i,j}|} \right) \quad (1)$$

関数 e_i は，参照要約の 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，$|A_{i,j}|$ に応じて部分点を与える関数である．なお，$A_{i,j}$ に対して重み付けを行うと，より詳細な評価が可能となる．たとえば，$A_{i,1}$，$A_{i,2}$ ともに要素は 1 文であり，それらが a_i の情報を完全に含んでいるのなら，短い文を出力した方がより良いであろう．しかし，こうした重み付けは被験者への負担が大きいので，実現することは難しいと考える．

関数 e_i と参照要約の文数 n を用いて，被覆率を以下の式で定義する．

$$\text{被覆率 } (E) = \frac{\sum_{i=1}^{n} e_i(E)}{n} \quad (2)$$

表 2 の対応関係が与えられた場合に以下の抜粋を考える．

$E1 = \{s_{20}, s_{21}, s_{23}, s_{30}, s_{60}, s_{70}\}$

$E2 = \{s_1, s_3, s_5, s_6, s_{30}, s_{70}\}$

抜粋 $E1$ の場合，

$$e_1(E1) = \max\left(\frac{0}{|\{s_1\}|}, \frac{0}{|\{s_{10}, s_{11}\}|} \right) = 0$$

$$e_2(E1) = \max\left(\frac{0}{|\{s_3, s_5, s_6\}|} \right) = 0$$

$$e_3(E1) = \max\left(\frac{|\{s_{20}, s_{21}, s_{23}\}|}{|\{s_{20}, s_{21}, s_{23}\}|}, \frac{|\{s_{30}, s_{60}\}|}{|\{s_1, s_{30}, s_{60}\}|} \right) = 1$$

となり，被覆率は 0.33 となる．

また，抜粋 $E2$ の場合，

$$e_1(E2) = \max\left(\frac{|\{s_1\}|}{|\{s_1\}|}, \frac{0}{|\{s_{10}, s_{11}\}|} \right) = 1$$

$$e_2(E2) = \max\left(\frac{|\{s_3, s_5, s_6\}|}{|\{s_3, s_5, s_6\}|} \right) = 1$$

$$e_3(E2) = \max\left(\frac{0}{|\{s_{20}, s_{21}, s_{23}\}|}, \frac{|\{s_1, s_{30}\}|}{|\{s_1, s_{30}, s_{60}\}|} \right) = 0.67$$

となるので，被覆率は 0.89 となる．これら 2 つの抜粋をシステム出力中に占める重要文（表 2 にエントリされた文）の割合で評価するとともに 5/6 = 0.83 であるが，被覆率に関しては，冗長な $E1$ は $E2$ よりも低い評価となっている．すなわち，被覆率が情報の冗長性を考慮できていることを示している．

3.3 重要文冗長率

前節より，関数 $e_i(E)$ は，参照要約の i 番目の文 a_i を抜粋 E がどれほど充足するか，冗長性（重複）を考慮して評価する．これに対し，冗長性を考慮せずに E が a_i を充足する文集合は，a_i の対応文集合 $A_{i,1}, A_{i,2}, \ldots, A_{i,j}$ の和集合を L_i とすると，$E \cap L_i$ として表すことができる．ここで，$e_i(E) = e_i(S)$ となる E の部分集合 S を考え，その部分集合の中で最も要素数が少ないものを $S_i^{\min}$ とする．これは，E のすべての文を用いなくても $S_i^{\min}$ の文を用いるだけで $e_i(E)$ という値を得ることができることを示している．よって，a_i に関して，重要文でかつ，$e_i(E)$ を得るために貢献していない文の数，つまり冗長な文の数は以下の式となる．

$$f_i(E) = |E \cap L_i| - |S_i^{\min}| \quad (3)$$

よって，抜粋 E に含まれる重要文がどの程度冗長であるかを以下の式で定義する．なお，重要文冗長率は 0〜1 の間の値に収まらないことに注意されたい．

$$\text{重要文冗長率 } (E) = \frac{\sum_{i=1}^{n} f_i(E)}{n} \quad (4)$$

抜粋 $E1$ を例にすると，

$L_1 = \{s_1, s_{10}, s_{11}\}$

$L_2 = \{s_3, s_5, s_6\}$

$L_3 = \{s_1, s_{20}, s_{21}, s_{23}, s_{30}, s_{60}\}$

であるから，$e_1(E1) = e_2(E1) = 0$ である．また，$e_3(E1) = 1$ であることから，$e_3(S) = 1$ となる $E1$ の最小部分集合は，$S_3^{\min} = \{s_{20}, s_{21}, s_{23}\}$ となる．

$f_i(E1)$ は以下のとおりである．

$f_1(E1) = 0$

$f_2(E1) = 0$

$f_3(E1) = 5 - 3 = 2$

よって，重要文冗長率は以下のとおりである．

$$\text{重要文冗長率 } (E1) = \frac{0 + 0 + 2}{3} = 0.67$$

また，抜粋 $E2$ を例にすると，$e_1(E2) = 1$，$e_2(E2) = 1$ であり，$e_3(E2) = 0.67$ であるから，$S_1^{\min} = 1$，$S_2^{\min} = 3$，$S_3^{\min} = 2$ である．

$f_1(E2) = 1 - 1 = 0$

$f_2(E2) = 3 - 3 = 0$

$f_3(E2) = 2 - 2 = 0$

よって，重要文冗長率は以下のとおりである．

$$\text{重要文冗長率 } (E2) = \frac{0+0+0}{3} = 0$$

a_3 に対応する文を多く含む $E1$ の重要文冗長率は，参照要約のどの文に対しても冗長な文を含まない $E2$ の重要文冗長率よりも高い．

なお理想的な抜粋は，被覆率が 1，重要文冗長率が 0 となる．

4. 評価実験の設定

4.1 コーパス

3 章での注釈付けの枠組みに基づき構築された TSC3 コーパス[7] より無作為に選んだ 25 トピックを評価実験に用いた．このコーパスは，読売新聞，毎日新聞の 98 年，99 年を対象として作成されており，各トピックは約 10 記事程度からなる．毎日新聞と読売新聞の比率はほぼ同等である．TSC3 では各トピックに対して 1 名の要約作成者が short，long という長さの異なる抜粋を作成しているが，本稿での評価実験には short のみを用いた．文書セットはそのほとんどが McKeown らの分類[9] に従うと single-event ☆に分類される．**図 1** に実験に用いたトピックを示す．

4.2 評価実験に用いた要約システム

評価実験には，TSC3 に参加した 4 システム[11)~13),16)]，オーガナイザが用意したベースラインシステムである Lead 手法，クラスタリングに基づく手法の 6 システムを用いた．

4.3 抜粋の情報量を評価する指標の比較

各トピックに対し 20 名の被験者を割り当て，各被験者は，元テキスト集合の重要情報をどの程度含んでいるかという観点に基づきシステム抜粋の順位付けを行う．システムスコアを 20 名による順位の平均値とし，人間によるシステムの順位付けと「被覆率」，下記に説明する「精度」，「正解率」によるシステムの順位付けとの間のスピアマンの順位相関係数を計算する．

なお，各トピックごとに，「20 名の順位付けが一致していない」という帰無仮説のもとフリードマン検定を行った結果，すべてのトピックにおいて p 値は 0.01 未満であり，帰無仮説が棄却された．この結果，参照要約を作成し，それに対して元テキストの文を対応付けを行ったのは同じ 1 名ではあるが，20 名の被験者間の順位付けが有意に一致したので，信頼性の高いデータであると考える．

精　　度

抜粋の長さ h を決定するために求めた最小カバーを唯一の正解抜粋として注釈付けを行い，システム抜粋にそれらが含まれる割合で定義する．最小カバーとシステム抜粋に共通に含まれる文の数を k として以下の式で定義する．

$$\text{精度 } (E) = \frac{k}{h} \tag{5}$$

正　解　率

システムが出力した文のうち，重要文として注釈付けられた文が占める割合を冗長性を考慮せずに計算する．以下の式で定義する．

$$\text{正解率 } (E) = \frac{m}{h} \tag{6}$$

ここで，h は，制約充足問題を解いて得た最小カバーの文数，m は，システムが出力した重要文の数である．ここでの重要文とは，参照要約に対応付けされたすべての文を指す．

なお，正解率，被覆率，精度の関係は **図 2** を参照されたい．

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
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
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
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
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

図 1　実験に用いたトピック
Fig. 1　Topics used for experimental evaluation.

☆ ある特定の事柄について記述された文書の集合．

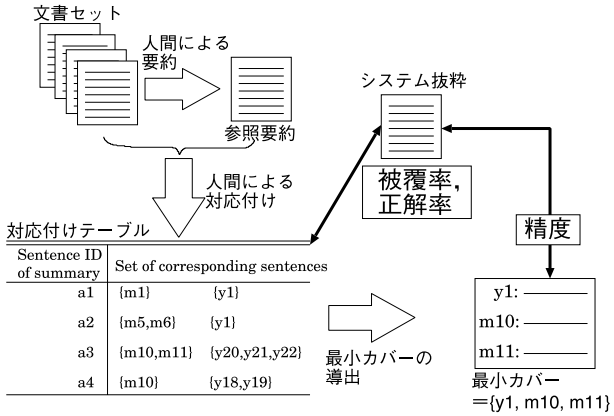


図 2 正解率，被覆率，精度の間の関係
Fig. 2 The relationship among automatic evaluation measures of extracts.

4.4 抜粋に含まれる重要文の冗長度を評価する指標の比較

各トピックに対し 1 名の被験者がシステム抜粋に含まれる重要文がどれだけ冗長かという観点から順位付けを行う．抜粋の情報量評価の場合と同様に人間によるシステムの順位付けと「重要文冗長率」，下記に説明する「正解率に対する被覆率の割合」によるシステムの順位付けとの間のスピアマンの順位相関係数を計算する．

正解率に対する被覆率の割合

正解率は重要文間の冗長性を考慮せず，システム抜粋が重要文をどの程度を含むかを測る指標である．一方，被覆率は冗長性を考慮し，システム抜粋が参照要約にどの程度近いかを測る指標である．よって，正解率が高く被覆率が低い場合には，抜粋に冗長な重要文が含まれることを示すと考えることができる．したがって，下記の式で定義する正解率に対する被覆率の割合を比較評価の対象とした．

$$\text{正解率に対する被覆率の割合 } (E) = 1 - \min\left(1, \frac{\text{被覆率 } (E)}{\text{正解率 } (E)}\right) \quad (7)$$

5. 評価結果と考察

表 3 に各評価指標による順位付けと人間による順位付けとの間のスピアマンの順位相関係数を示す．被覆率，正解率，精度を比較すると，被覆率の性能が他の 2 手法よりも優れている．25 トピックの平均でも 0.740 の相関係数を得ており，抜粋の順位付けという観点から優れた指標であることが分かる．一方，正解率，精度は 25 トピックの平均順位相関係数が 0.5 以下であり，抜粋の順位付けに対する人間との間の相関は低い．精度の成績が被覆率よりも大幅に低いことからは，従来のように唯一の重要文セットに対して注釈付けを行うだけでは，不十分であることが分かる．さらに，正解率の成績も同程度に悪いことから，重要文とそれに対する同意味の文に注釈付けを行うだけでなく，冗長性を考慮した評価指標を定義しなければならないことも分かる．

表 3 人間による順位付けと被覆率，正解率，精度による順位付けとの間のスピアマンの順位相関係数
Table 3 Spearman's ranking correlation coefficients between the human ranking of extracts and the automatic ranking by coverage, accuracy and precision.

トピック番号	被覆率	正解率	精度
0310	**.971**	.926	.309
0320	**.725**	.455	.207
0350	.754	.530	**.759**
0360	**.783**	.647	.606
0370	.338	**.655**	.293
0400	**.986**	.926	.778
0410	**.348**	.058	.235
0450	.912	.609	**.956**
0460	.772	−.131	**.926**
0470	**.530**	.353	.525
0480	**.618**	−.334	.309
0500	.441	**.471**	.216
0510	.943	**.971**	.765
0520	.845	.736	**.926**
0530	**.829**	.754	.802
0540	**.464**	−.353	−.339
0550	.319	.265	**.441**
0560	**.926**	.145	-.359
0570	.899	.463	**.971**
0580	**.986**	.698	.516
0590	**.577**	.257	.353
0600	**.928**	.395	.617
0610	.853	**.926**	.507
0640	**.928**	.353	.239
0650	**.829**	.754	.706
平均	.740	.461	.491

表 4 人間による順位付けと重要文冗長率，正解率に対する被覆率の割合との間のスピアマンの順位相関係数
Table 4 Spearman's ranking correlation coefficients between the human ranking of extracts and the automatic ranking by redundancy and coverage/accuracy.

トピック番号	重要文冗長率	正解率に対する被覆率の割合
0310	**1.00**	.853
0320	**.496**	.308
0350	**.438**	**.438**
0360	.078	**.553**
0370	**1.00**	.775
0400	.500	**.612**
0410	**.575**	.566
0450	**1.00**	.857
0460	**.850**	.657
0470	.718	**.857**
0480	.899	**.985**
0500	.696	**.816**
0510	**.858**	.338
0520	**.938**	.721
0530	**.920**	.541
0540	**.824**	.580
0550	**.851**	.426
0560	**.986**	.924
0570	**.984**	.705
0580	**.567**	.381
0590	**.708**	.577
0600	**.853**	.883
0610	N/A	N/A
0640	.767	**.866**
0650	**.309**	.135
平均	.743	.640

また，被覆率であっても 0.5 以下の相関しか得られていないトピックがいくつかある．これらのトピックの共通点は，被覆率では同順位になるシステム抜粋が多いが，被験者の評価ではそれらのシステム抜粋が同順位にならないことであった．これは，被覆率が重要文として注釈付けされた文のみを対象として評価することに対し，人間は，たとえ重要文でなくてもトピックに関連する何らかの情報を持つ文であれば，評価することが原因であると考える．これを避けるには，Utility [15] のように元テキスト中のすべての文に対して重みを与えなければならない．しかし，複数文書要約では元テキストに含まれる文の数は非常に多いので現実的ではない．

表 4 に抜粋中の重要文の冗長性に関して，人間の順位付けと正解冗長率，被覆率と正解率の比による順位付けとの間の相関係数を示す．なお，トピック 0610 については，人間の順位付けにおいて，6 システムの順位がすべて同じだったので，スピアマンの順位相関係数は計算できなかった．

表 4 より，重要文冗長率の相関係数は平均で 0.74 程度の十分高い相関を得ており，その有効性がよく分かる．また，被覆率と正解率の比と比較しても，優れている．相関係数が 0.5 未満のトピックは 4 つだけであり，全体的に良い成績である．正解率に対する被覆率の割合でも良い相関を得ているトピックはいくつか存在するが，重要文冗長率を超える相関係数を得たトピック数は 6 つだけでしかない．

以上より，抜粋の情報量を測る指標として「被覆率」が有効であること，抜粋に含まれる重要文の冗長度を測る指標として「重要文冗長率」が有効であることを示した．

6. おわりに

本稿では，複数の文書から得た抜粋を評価するため，コーパスへの注釈付けの枠組みとそれに基づく評価指標である「被覆率」と「重要文冗長率」を提案した．これらの評価指標の有効性を示すため，TSC3 コーパスを用い，システム抜粋の人間による順位付けと被覆率，重要文冗長率による順位付けとの間の順位相関係数を調べた．その結果，被覆率との間の順位相関係数の平均は約 0.74 であり，従来の注釈付けとそれに基づく評価指標である精度の順位相関係数よりも大幅に良いことが分かった．また，重要文冗長率との間の相関係数も約 0.74 であり十分高いことを確認した．

謝辞 データの使用を許諾いただいた毎日新聞社，読売新聞社に感謝いたします．システム抜粋をご提供くださった横浜国立大学の森辰則氏，東京大学の岡崎直観氏，豊橋技術科学大学の酒井浩之氏に感謝いたします．

参考文献

1) Barzilay, R., Elhadad, N. and McKeown, K.: Inferring Strategies for Sentence Ordering in Multi-Document News Summarization, *Journal of Artificial Intelligence Research*, Vol.17, pp.33–55 (2002).
2) Barzilay, R., McKeown, K. and Elhadad, N.: Information Fusion in the Context of Multi-Document Summarization, *Proc. 38th ACL*, pp.550–557 (1999).
3) Carbonell, J. and Goldstein, J.: The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries, *Proc. 21st ACM-SIGIR*, pp.335–336 (1998).
4) Fukusima, T. and Okumura, M.: Text Summarization Challenge: Text Summarization Evaluation in Japan, *Proc.NAACL 2001 Work-*

shop on Automatic summarization, pp.51–59 (2001).
5) Hatzivassiloglou, V., Klavans, J.L. and Eskin, E.: Detecting Text Similarity Over Short Passage: Explording Linguistic Feature Combinations via Machine Learning, *Proc. EMNLP*, pp.203–212 (1999).
6) Hatzivassiloglou, V., Klavans, J.L., Holocombe, M.L., Barzilay, R., Kan, M.-Y. and McKeown, K.: Simfinder: A Flexible Clustering Tool for Summarization, *Proc. NAACL Workshop on Autmatic Summarization*, pp.41–49 (2001).
7) 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).
8) Jing, H. and McKeown, K.: Cut and Paste based Text Summarization, *Proc. 1st NAACL*, pp.178–185 (2000).
9) McKeown, K., Barzilay, R., Evans, D., Hatzivassilogou, V., Kan, M.Y., Schiffman, B. and Teufel, S.: Columbia Multi-Document Summarization: Approach and Evaluation, *Proc. Document Understanding Conference 2001* (2001).
10) Minato, S.: BEM-II: An Arithmetic Boolean Expression Manipulator Using BDDs, *IEICE Trans. Fundamentals*, Vol.E76-A, No.10, pp.1721–1729 (1993).
11) Mori, T., Nozawa, M. and Asada, Y.: Multi-Document Summarization Using a Question-Answering Engine, *Proc. 4th NTCIR-Workshop* (2004).
12) Nobata, C., Sekine, S., Uchimoto, K. and Isahara, H.: Comparison of feature usage at TSC-3 summarization tasks, *Proc. 4th NTCIR-Workshop* (2004).
13) Okazaki, N., Matsuo, Y. and Ishizuka, M.: TISS: An Integrated Summarization System for TSC-3, *Proc. 4th NTCIR-Workshop* (2004).
14) Radev, D.: A Common Theory of Information Fusion from Multiple Text Sources, Step One: Cross-document Structure, *Proc. SIGDIAL*, pp.74–83 (2000).
15) Radev, D., Jing, H. and Budzikowska, M.: Centroid-based Summarization of Multiple Document Summarization: Sentence Extraction, Utility-based Evaluation and User Studies, *Proc. ANLP/NAACL2000 Workshop on Automatic Summarization*, pp.21–30 (2000).
16) Sakai, H. and Masuyama, S.: A Multiple Document Summarization System intoroducing User Interaction for Reflecting User's Need, *Proc. 4th NTCIR-Workshop* (2004).
17) 宮部泰成，高村大也，奥村 学：異なる文書中の文間関係の特定，情報処理学会研究報告自然言語処理研究会 NL-169，pp.35–42 (2005).

(平成 19 年 3 月 19 日受付)
(平成 19 年 7 月 5 日採録)

(担当編集委員 岸田 和明)

平尾 努（正会員）

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．言語処理学会，ACL 各会員．

奥村 学（正会員）

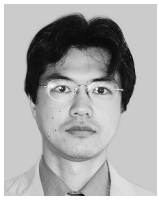
1989 年東京工業大学大学院情報理工学研究科計算工学専攻博士後期課程修了．1989 年より東京工業大学大学院情報理工学研究科助手．1992 年より 2000 年北陸先端科学技術大学院大学助教授．1997 年より 1998 年トロント大学客員助教授．2000 年より東京工業大学精密工学研究所助教授．自然言語処理，自動テキスト要約，コンピュータによる語学学習支援，テキストデータマイニングに関する研究に従事．工学博士．AAAI，ACL，JSAI，JCSS 各会員．

福島　孝博（正会員）

1984 年大阪外国語大学英語科卒業．1990 年ニューヨーク州立大学大学院コンピュータ・サイエンス研究科修士課程修了．1990 年から 1993 年ニューメキシコ州立大学付属 Computing Research Lab 研究員，1994 年英国シェフィールド大学コンピュータ・サイエンス Research Associate．1996 年日本電気（株）入社．同年通信放送機構にて研究員．2000 年より追手門学院大学文学部英語文化学科．2007 年同大学国際教養学部英語コミュニケーション学科．自然言語処理，情報抽出，要約筆記に関する研究に従事．電子情報通信学会，言語処理学会，ACL 各会員．

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 年北陸先端科学技術大学院大学情報科学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，人工知能学会，ACL，ACM 各会員．



野畑　　周（正会員）

2000 年東京大学大学院理学系研究科博士課程修了．博士（理学）．同年郵政省通信総合研究所関西先端研究センター知的機能研究室非常勤研究員．2004 年シャープ株式会社情報通信事業本部技術企画室主事．2007 年マンチェスター大学 Research Associate．言語処理学会，ACL 各会員．

磯崎　秀樹（正会員）

1983 年東京大学工学部計数工学科卒業．1986 年同工学系大学院修士課程修了．同年日本電信電話（株）入社．1990〜1991 年スタンフォード大学ロボティクス研究所客員研究員．現在，NTT コミュニケーション科学基礎研究所知識処理研究グループリーダ．博士（工学）．平成 15 年度情報処理学会論文賞・山下記念研究賞受賞．人工知能・自然言語処理の研究に従事．電子情報通信学会，人工知能学会，言語処理学会，ACL 各会員．