自然言語処理　132－6
(1999. 7. 22)

情報検索タスクに基づいた評価による要約手法の比較

望月 源　奥村 学
北陸先端科学技術大学院大学 情報科学研究科
〒923-1292 能美郡辰口町旭台 1-1
E-mail:{motizuki,oku}@jaist.ac.jp

あらまし
要約の新しい評価方法としてタスクに基づく評価が採用され始めている．タスクに基づく評価とは，要約を利用してあるタスクを行なう際のタスクの達成率や所要時間などを間接的な要約の評価に用いるものである．本稿では情報検索をタスクとして選択し，情報検索タスクに基づいた評価方法を用いて異なる要約手法間の比較を行なう．人間の被験者による実験を行ない，タスク達成の精度から要約手法の精度を比較する．タスクベースの評価方法に関する問題点，留意点についてもいくつかのポイントから分析し，報告する．また，10種類の異なる要約作成手法の結果から我々の提案する手法が情報検索タスクにおいて，良い要約を作成することも合わせて示す．

キーワード　自動テキスト要約，タスクに基づく評価，情報検索

Evaluation of Summarization Methods based on Information Retrieval Task

MOCHIZUKI Hajime　OKUMURA Manabu
School of Information Science,
Japan Advanced Institute of Science and Technology
1-1, Tatsunokuchi, Ishikawa, 923-1292, JAPAN
E-mail:{motizuki,oku}@jaist.ac.jp

Abstract
As a new direction of evaluation method for summarization systems, task-based evaluation scheme has been adopted. It evaluates the performance of a summarization system in a given task, such as information retrieval and news analysis. This paper makes a comparison among ten different summarization systems based on the task of information retrieval. To evaluate system performance, subjects' speed and accuracy are measured in judging the relevance of texts. In order to establish a good evaluation methodology, we analyze which factors can affect evaluation results, and describe the problems that arised from the experimental design. Furthermore, we also present that our proposed summarization method can produce effective summaries for information retrieval task.

key words　automatic text summarization, task-based evaluation, information retrieval

1 はじめに

近年の電子化テキストの増大により，計算機によるテキスト処理の利便性が向上した．しかし，利用者が求めるテキストを発見するまでに処理しなければならない情報の量が増加する結果も引き起している．また，我々人間の情報処理能力の飛躍的な向上は望めないため，今後も続くと予想される情報量の増加に対処しきれないという問題がある．

この問題に対する解決策の一つとして，自動要約作成技術が注目されている．テキストを要約することにより，人間が読まなければならないテキストの量を制限し，負担を減らすことが期待できるからである．この自動要約技術を発展させていく上で，手法の評価をすることが重要であるが，要約の評価方法自体が大変に難しい問題である．これまでの要約研究の多くは，唯一の "理想的な要約" が存在すると仮定して，" 理想的な要約" とシステムの作成した要約文を比較し，再現率，適合率，および F-measure などの尺度を用いて評価を行なう．多くの場合，"理想的な要約" は，複数の人間の被験者が作成した要約を基に獲得する．しかし，人間にとっても要約作業は容易ではなく，被験者間でも要約は高い割合で一致するとは限らない．また，唯一の理想的な要約の存在を仮定すること自体が不自然であり，評価基準が曖昧であるという指摘は古くからなされている．

これに対して，最近ではタスクに基づく要約の評価方法が採用され始めている [10, 4, 5, 8]．タスクに基づく要約の評価とは，人間が要約を利用して，あるタスクを行なう際のタスクの達成率や所要時間などを間接的な要約の評価に用いるものである．この新しい評価方法は，要約の利用される状況で利用者の使用目的に合わせて動的に作成するという新しい要約作成の考え方と密接に関わっている．例えば，情報検索において，検索結果の適合性を利用者が判断するために要約を用いることを考えた場合，一般的な要約よりも，利用者が入力した検索要求に即した要約の方が良いと考えられる．このような場合に従来の "理想的な要約" との比較で要約の精度を論じるよりも，そのタスクにとってどれだけその要約が役立ったかで評価した方が良い．

こうした背景から，本稿では，自動要約と密接な関係にある情報検索をタスクとして選択し，情報検索タスクに基づいた評価方法を用いて異なる要約手法間の比較を行なう．評価実験では，人間の被験者に，検索要求と検索結果であると仮定したテキストのリストおよびその要約文を提示する．被験者は要約文を読むことによって，そのテキストが検索要求に適合するかどうかの判断を行う．本稿で比較する要約手法には，我々の提案する "検索要求を考慮した語彙的連鎖に基づく手法" を含め，作成される要約文が連続性を持つかどうか，検索要求を考慮するかどうかなどの違いを持つ 8 種類と全文およびタイトルのみを合わせた計 10 種類を用意する．検索要求と要約手法の異なる組み合わせについて行なわれた複数の被験者の結果から，適合性判断の精度，タスクにかかった時間および判断に迷った際に全文を参照した回数などの点に基づいて各要約手法の評価，考察を行なう．結果から我々の提案する手法が情報検索タスクにおいて，良い要約を作成することも合わせて示す．

また，タスクに基づく要約の評価はまだ行なわれ始めたばかりであり，どのような点を評価すべきかを含めて試行錯誤の段階である．そのため，今回の実験から明らかになった，評価尺度，タスク (実験) の設定，検索要求の選択などの問題点についての報告も行なう．

2 要約作成手法

本稿で比較する各要約手法について説明する．これまでの自動要約のほとんどは，テキスト中の重要個所 (文，段落，節など) を抜き出す手法を用いている [10]．そのため，本稿でも重要個所抽出による要約を扱う．比較する要約手法は次の 10 種類である (各手法の本稿での呼び名をボールド体で示す)．

- 全文 (**full**)
  要約を行なわず，全文を要約として提示する．
- タイトル (**title**)
  文の見出しのみを要約とする．
- lead 手法 (**lead**)
  見出しを含み，先頭から要約率分の文を抽出する．
- 形式段落 (**f-seg**)
  検索要求と各形式段落との類似度を計算し，最も類似度の高い形式段落を 1 つ選び要約とする．ただし，要約率を超える場合には，段落の先頭から要約率分の文を抽出する．なお，検索要求と形式段落の類似度は以下のように計算する．

  形式段落ごとに単語の重要度 w_i を基本的な $tf.idf$ [7] の式 (1) により計算し，単語の重要度

のベクトル D_j で表現する．

$$w_i = tf_i \times log\frac{N}{df_i} \quad (1)$$

ここで tf_i は段落内の単語 i の頻度，N はテキスト集合内の総段落数，df_i は単語 i の出現段落数である．

検索要求と形式段落の類似度は，それぞれのベクトル Q，D_j を用いて計算する．

$$sim(Q, D_j) = \sum_i (tf_{q_i}/log\frac{N}{df_i})^2 \times w_i \quad (2)$$

ここで，tf_{q_i} は検索語 q_i の検索要求内の頻度である．

- テキスト中の単語の重要度に基づく要約 (**tf.idf, q-tf.idf**)

テキスト中の単語の出現頻度から各単語の重要度を決定し，重要な単語を多く含む文が重要であるという考えに基づき，文の重要度を計算する．

本稿ではZechner[9]と同様の手法を用いる．まず，テキスト中の各単語の重要度 w_i を計算する．次に，各文中の単語の重要度の総和を式(3)により計算し，重要度 S_j の高い文を要約率分抽出する．

$$S_j = \sum_i w_i \quad (3)$$

このタイプの要約では，各単語の重要度 w_i を計算する際に，検索要求を考慮するかどうかの違いにより次の2種類を作成する．

1. 検索要求を考慮しない場合 (**tf.idf**)
   式(1)により，w_i を tf.idf を基に計算する．ただしこの場合の tf_i はテキスト内の単語 i の出現頻度，df_i は単語 i の出現するテキストの数，N は全テキスト数である．
2. 検索要求を考慮する場合 (**q-tf.idf**)
   検索要求内の単語 (検索語) には重み α をかける[1]．

$$w_i = \begin{cases} tf_i \times log\frac{N}{df_i} & \text{検索語でない} \\ \alpha \times tf_i \times log\frac{N}{df_i} & \text{検索語} \end{cases} \quad (4)$$

- テキスト中の語彙的連鎖の重要度に基づく要約 (**cf.idf, q-cf.idf**)

語彙的連鎖 [6] とは，テキスト中で意味的つながり (語彙的結束性 [3]) を持つ語の連続のことをいう．各連鎖がテキスト中の話題を表わすと考えられるため，重要な連鎖を構成する単語を多く含む文が重要であると考え，文の重要度を決定する．

語彙的連鎖は次の手順で計算する．まずコーパスを用いた単語の共起情報から2つの単語間の類似度を式(5)のコサイン距離により計算する．

$$coscr(X, Y) = \frac{\sum_{i=1}^{n} x_i \times y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} \times \sqrt{\sum_{i=1}^{n} y_i^2}} \quad (5)$$

ここで，x_i と y_i はテキスト i に単語 X と Y が出現する数 (tf)，n はコーパスの全テキスト数を表わす．

次に，計算された類似度を基に，式(6)の最短距離法によって単語をクラスタリングする．

$$sim(C_i, C_j) = \max_{X \in C_i, Y \in C_j} coscr(X, Y) \quad (6)$$

ここで，X, Y はクラスタ C_i 内，C_j 内の単語である．

ある閾値[2]までクラスタをマージし，同一クラスタ内の単語によって語彙的連鎖を構成する．

要約の作成は，単語の重要度に基づく場合と同様に，まず各連鎖の重要度 w_i を計算し，次に式(3)により各文中の連鎖の重要度の総和を計算し，重要度 S_j の高い文を要約率分抽出する．なお，w_i を計算する際に，検索要求を考慮するかどうかの違いにより次の2種類を作成する．

1. 検索要求を考慮しない場合 (**cf.idf**)
   w_i を，連鎖 i の構成単語数 (cf) と連鎖 i の出現テキスト数 (df_i) により計算する．

$$w_i = |\, i \,| \times log\frac{N}{df_i} \quad (7)$$

   ここで，$|\, i \,|$ は連鎖 i の構成単語数，df_i は連鎖 i の出現テキスト数，N は全テキスト数である．

2. 検索要求を考慮する場合 (**q-cf.idf**)
   検索語を含む連鎖には重み α をかける[3]．

$$w_i = \begin{cases} |\, i \,| \times log\frac{N}{df_i} & \text{検索語でない} \\ \alpha \times |\, i \,| \times log\frac{N}{df_i} & \text{検索語} \end{cases} \quad (8)$$

[1] 予備的な実験において，いくつかのテキストにおいて，α を2,3,4,5と変化させ，重みをかけない場合との差の変化が最も大きかった3を今回の α の値とした

[2] 我々の以前の研究 [11] で最も良い連鎖を構成した0.25を今回の閾値とした．

[3] **q-tf.idf** の場合と同様に $\alpha = 3$ とした

- 語彙的連鎖型パッセージに基づく要約 (**lex**)
  我々の語彙的連鎖に基づくパッセージ検索 [11] により，検索要求に適合するテキスト中のパッセージを抽出し，検索要求を考慮した語彙的連鎖に基づく要約を作成する．今回は検索要求との類似度が最も高いパッセージを使用する．ただし，要約率を超える場合にはパッセージの先頭から要約率分だけを抽出する．語彙的連鎖は **cf.idfq-cf.idf** の場合と同様に計算する．

  パッセージ抽出は次の手順で行なう．まず，入力された検索要求に一致するテキストと，そのテキスト内で検索語を含む語彙的連鎖の情報を取り出す．次に出現範囲に重なりがある連鎖をまとめることによってパッセージを決定，抽出する．この時に，各パッセージと検索要求との類似度も計算する．図1にパッセージの例を示す．

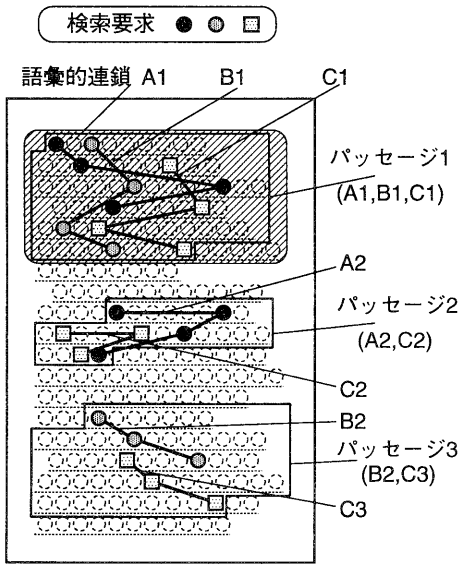

図 1: パッセージの例

図1には，検索要求内の3つの検索語と，一致するテキスト，および各検索語を含む7つの語彙的連鎖 (A1～C3) が示されている．重なりのある連鎖をまとめることによって，3つのパッセージが決定されている．この例ではパッセージ1の類似度が最高であるため，パッセージ1から要約として文が抽出される (斜線部分)．

- J社の市販ワープロソフトによる要約 (**J**)
  市販の代表的な要約機能付きワープロ3種 (J，F，M社) の要約を用いた被験者9人による予備実験から，総合的に最も精度の高かったJ社の要約を本試験に使用する．

3 評価実験

本稿の情報検索タスクに基づく要約の評価実験は，TIPSTER III のために提案された手法 [4, 5] を参考にしている．実験には『情報検索テストコレクション BMIR-J2』[12] を使用する[4]．今回は，主題を表わすA判定の正解テキストが10テキスト以上ある10種類の検索要求と，各要求毎に20テキストを選択して使用する．20のテキストは，まず各検索要求によるキーワード検索を行ない，検索結果の中から正解テキストが50%以上含まれるように選択する．

実験では，被験者30名 (日本語を母国語とする情報科学系の大学院生) に対し，検索要求と20テキストおよびその要約文を提示する．被験者は，各テキストが検索要求に適合するかどうかの判定と20テキストの判断にかかった時間を記録する．要約を読んでも判断のつかない場合に限り，そのテキストの全文を参照することを許し，全文を参照した回数を検索要求ごとに記録する．また，提示された各要約について，要約としてではなく，日本語の文章としての読み易さを主観で判断する問題も設定する．判断基準は，1．わかりやすい，2．ややわかりやすい，3．ややわかりにくい，4．わかりにくい．の4段階とする[5]．

1人の被験者は，同じ検索要求とテキストの組を1度しか判断できないので，30名を3名づつの10グループに分け，各グループが異なる要約手法と検索要求の組を10組づつ評価する．そのため，1組の検索要求と要約手法に対し3名が評価を行なう．

なお，今回の要約作成では，要約率は文を単位として20%とした．

3.1 評価基準

以下の点についての評価を行なう．

- 適合性判断の精度
- タスクにかかった時間

[4] BMIR-J2は，テキスト5080件 (毎日新聞1994年版の経済および工学，工業技術一般に関連する記事)，検索要求50種とその正解がセットとなったテストコレクションである．

[5] 判断基準を奇数にすると真中が選択される傾向が予想されるため，この4段階とした．

- 全文を参照した回数
- 要約の文章としての読み易さ

精度の評価尺度には再現率，適合率，F-measure を使用する．

$$再現率 = \frac{被験者が正しく適合と判断した数}{実際に適合するテキスト数} \quad (9)$$

$$適合率 = \frac{被験者が正しく適合と判断した数}{被験者が適合と判断したテキストの総数} \quad (10)$$

$$Fmeasure = \frac{2 \times 再現率 \times 適合率}{再現率 + 適合率} \quad (11)$$

3.2 結果

3.2.1 被験者の精度差の調整

今回の実験では被験者 30 名の言語学的，教育的背景をできるだけ揃えている．これは，被験者の知識や背景の違いに起因する検索要求の解釈や判断の違いをだけ排除する狙いがある．つまり，被験者のタスク達成精度に差がないことを仮定している．しかし，今回の 30 名の精度について，一元配置分散分析を行なったところ差がみられた ($p < 0.0039$)．そのため，F-measure を基準に差の少ない，図 2の丸で囲んだ，21 名 ($p < 0.9995$) によって以下の評価を行なった．

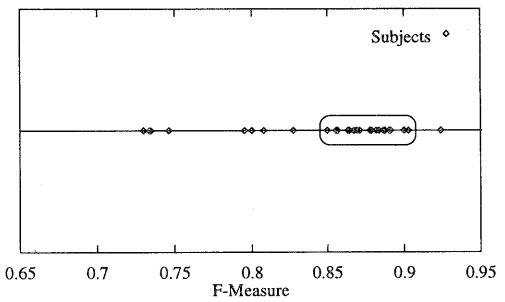

図 2: F-measure

21 名での実験結果を表 1に示す．

表には検索要求 (20 テキスト) ごとの結果の平均が示されている．表中の「読み易さ」は，被験者が 4 段階で判断した読み易さを数値化した値である．この数値は，被験者が全てのテキストが「わかりやすい」と判断した場合に 10 点,「ややわかりやすい」と判断した場合に 5 点,「ややわかりにくい」,「わかりにくい」の場合それぞれ，-5 点，-10 点となるように計算した．また,「時間比」は各要約での作業に要した時間を全文に要した時間で割った値である.「平均文間数」は要約内の隣接 2 文間の全文での距離 (文間数) の平均を表わしている．なお，今回の要約率の決定は文を単位として 20%としているが，語を単位として各手法の要約率を計算したものを「要約率 (語)」として示す．

3.2.2 被験者の判断の統計的信頼性

Jing ら [4] と同様に，帰無仮説を「あるテキストが検索要求に適合すると被験者達によって判断される数はランダムである」とした Cochran の Q-test[1] を行なった．結果，全 10 種の検索要求それぞれについて $p < 10^{-5}$であり，帰無仮説は棄却された．つまり，今回の実験で，被験者達は適合するテキストをランダムに選択しているのではないとの検定結果を得た．

3.3 考察

10 種類の要約手法全体，連続性の有無，検索要求の考慮の有無という作成される要約の持つ特徴の違いごとに，精度，時間，読み易さ，および，各要因間の相関というポイントで考察を行なう．また，各要約手法により作成された要約の類似性と検索要求の難易度についても考察する．

3.3.1 全要約手法間の比較

F-measure で比較すると，語彙的連鎖 (**lex**) と **J** が全文 (**full**) での精度を上回り，他のものは **full** と同じかやや低い．

ただし，今回の 10 手法での F-measure について，一元配置分散分析を行なった結果，全体として平均値に統計的な有意差は見られなかった ($p < 0.9725$)．しかし，ダンカンの方法 [2] によって，10 手法内の各 2 手法間の比較を行なったところ，我々の語彙的連鎖パッセージ **lex** と **tf.idf,q-tf.idf,cf.idf,q-cf.idf** の間，および，**J** と **tf.idf,q-tf.idf,cf.idf,q-cf.idf** の間にはそれぞれ有意な差が見られた．そのため，**lex** と **J** は **tf.idf,q-tf.idf,cf.idf,q-cf.idf** の各手法よりも統計的に有意な差で良い精度を示しているといえる．

作業時間に関しては，**full** が最もかかっており，どの要約手法によっても時間短縮の効果はあるといえる．ただし，表 1の「時間比」と語レベルでの要約率「要約率 (語)」を対応させると，**title** の場合が一番効率が悪い．他の場合についてははっきりとした差が見られない．

	full	title	lead	f-seg	tf.idf	q-tf.idf	cf.idf	q-cf.idf	lex	J
再現率	87.1%	86.7	85.9	87.2	86.3	89.7	87.0	85.3	90.5	86.5
適合率	89.0%	89.1	88.9	88.8	89.7	85.3	85.3	87.0	88.5	91.3
F-measure	88.0%	87.9	87.4	88.0	88.0	87.4	86.1	86.1	89.5	88.9
読み易さ	4.1	1.8	4.6	3.7	3.8	4.0	4.3	3.9	4.1	5.5
時間 (分:秒)	15:38	7:54	9:47	10:55	10:37	9:54	10:54	10:29	10:41	10:52
時間比	100%	50.5	62.6	69.8	67.9	63.3	69.7	67.1	68.3	69.5
参照回数	0.6	4.8	3.8	2.8	2.6	1.7	2.0	1.5	1.9	2.0
平均文間数	0.0	0.0	0.0	0.0	3.6	3.5	3.8	3.6	0.0	1.4
要約率 (語)	100.0%	5.3	19.1	21.7	32.1	31.5	30.5	30.2	23.6	27.7

表 1: 実験結果

読み易さについては，**J**, **lead**,**cf.idf** が **full** よりも高く，**lex** は同じであり，それ以外は **full** よりも低い．**title** が一番低い値だった．後述するように要約文を元のテキストの先頭から選ぶ傾向のある手法 (**lead**, **lex**, **J**) の値が高いといえる．

全モデル間の相関 (自由度 208 の 5%有意水準 = 0.136) を計算すると，再現率と読み易さに正の相関 (0.186)，F-measure と読み易さに正の相関 (0.162)，読み易さと参照回数に負の相関 (−0.237)，がそれぞれ見られた．

3.3.2 要約の連続性による比較

作成される要約の連続性という観点からの比較を行なう．表 1で,「平均文間数」が 0 に近いもの程連続性が高い．表 1から，**full**, **title**, **f-seg**, **lead**, **lex**, **J** は連続性が高く，**tf.idf**, **cf.idf**, **q-tf.idf**, **q-cf.idf** は連続性が低いといえる．

精度に関しては，全体的に連続性がある方がないものよりも F-measure が高い．しかし，**lead** は連続性があるが精度が低く，**tf.idf** は連続性がないが精度が高いというように若干のばらつきもある．

読み易さに関しては，全体的に連続性がある方が読み易さも高いが，形式段落の場合のように良くないものもあるため，連続性だけでなく元のテキストでの位置も関係するものと推測される．

全体としては連続性のある要約の方が読み易く，精度も高いと考えられる．

相関を計算すると，連続性のあるモデル[6]だけの場合 (自由度 82 の 5%有意水準 = 0.215) には，特に有意な相関は見られなかった．

一方，連続性のないモデル[7]だけの場合 (自由度 82 の 5%有意水準 = 0.215) には，再現率と読み易さに正の相関 (0.236)，再現率と参照回数に負の相関 (−0.285)，F-measure と読み易さに正の相関 (0.218) が見られた．

また，要約率を語レベルで考えた場合に，連続性のないモデルでは，適合率と要約率に正の相関 (0.356)，F-measure と要約率に正の相関 (0.317) が見られる．このことは，作成される要約に連続性がない (つまり，ばらばらに抜き出された) 場合には，長めの文を選択する方が適合率が上がる傾向があることを示している．

3.3.3 検索要求の考慮の違いによる比較

要約の作成に検索要求を考慮する (検索語に重みをかける) 場合としない (重みをかけない) 場合についての比較を行なう．今回の手法の内，検索語に重みをかける手法は，**q-tf.idf**, **q-cf.idf**, **lex**, **f-seg** であり，かけない手法は，**tf.idf**,**cf.idf**,**lead**,**J** である．この内，次の 2 組は直接比較できる．

1. **tf.idf** と **q-tf.idf**
   重みをかけることで，(1)F-measure が下った，(2) 再現率が上がった．(3) 読み易さが上がった．(4) 作業時間が減った．(5) 参照回数が減った．

2. **cf.idf** と **q-cf.idf**
   重みをかけることで，(1)F-measure は変わらない．(2) 再現率が下った．(3) 読み易さが下がった．(4) 作業時間が減った．(5) 参照回数が減った．

[6] **full**,**title** は除き，**lead**, **f-seg**, **lex**, **J** とした．

[7] **tf.idf**,**q-tf.idf**,**cf.idf**, **q-cf.idf**.

検索語に重みをかけることで，少なくとも適合性判断のしやすさは向上すると言える．しかし判断の精度が向上するとは限らない．

相関については，重みをかけない手法，重みをかける手法，それぞれの場合にわけて計算したが，どちらについても有意な相関は見られなかった．

3.3.4 要約の類似度

作成される要約の類似性についての考察を行なう．類似性は，要約手法毎に用意したベクトル間の cosine 距離によるベクトル間類似度として計算した．各要約手法毎のベクトルは，ベクトルの要素を元テキスト (全文) の各文とし，値をその文が重要文として選択されれば 1，されなければ 0 とする．計算された類似度を用いて，最短距離法と平均距離法の 2 種類のクラスタ間距離による階層型クラスタリングを行なったところ，どちらの場合も図 3のようになった．図 3で上段の数字が最短距離法，下段が平均距離法での類似度を示す．なお，**full** と **title** は他の文に比べて長さに差があるので類似度計算からは除いある．

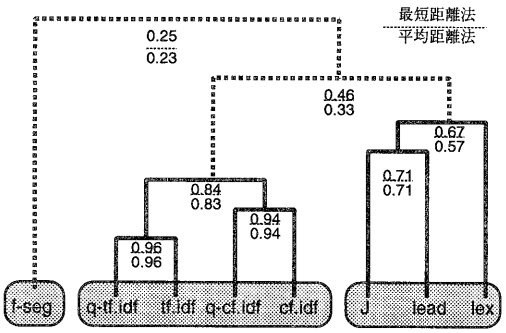

図 3: 要約間の類似度

J と **lead** と **lex** の類似度が高く 1 つのグループを形成する．また，**tf.idf**, **cf.idf**, **q-tf.idf**, **q-cf.idf** の類似度も高く別のグループを形成する．**f-seg** はどちらのグループにも近くない．全体として 3 つのグループにわかれた．

各グループ毎に見ると，最初のグループの **J** と **lead** および，**lex** と **lead** はそれぞれ類似度が高いが，**J** と **lex** はそれほど類似度が高くない (0.47)．これは **J**, **lex** のどちらの手法でも元のテキストの先頭部分をある程度含むが，**J** では先頭と離れた文もある程度の割合で含んでいることを示している．

また 2 番目のグループに含まれる 4 つの手法では，どの手法間の類似度も非常に高い．これは今回のテキストについて 4 つの手法によって作成される要約にそれほど差がなかったことを示す．

3.3.5 検索要求の難しさ

検索要求毎の平均精度と標準偏差を表 2に示す．

検索要求	再現率	適合率	F-m.
マンションの販売	96.1%	98.3%	97.2%
(標準偏差)	6.2	3.5	0.4
国内航空大手 3 社	95.2	95.5	95.4
	14.3	6.0	10.9
株価動向	92.5	91.0	91.7
	6.9	11.5	6.3
女性の雇用問題	92.9	88.2	90.5
	9.2	8.6	6.3
携帯電話またはパー	88.6	91.6	90.1
ソナルハンディホン	9.9	7.3	5.6
銀行の経営計画	88.9	87.2	88.0
	11.4	4.9	5.7
行政機関が関係	79.1	90.4	84.4
する不況対策	10.4	7.4	7.3
減税	89.8	74.9	81.7
	8.3	7.0	5.1
円高対策のための	67.8	94.0	78.8
メーカーの海外進出	18.9	7.8	12.8
所得税の減税	81.3	71.7	76.2
	9.9	8.1	6.2

表 2: 検索要求別の結果

精度が高く，標準偏差の小さい検索要求ほど簡単であると考えられる．表を見ると比較的容易な検索要求と難しいものがあることがわかる．少なくとも 1 番目の検索要求は平均精度が高く，標準偏差も小さいことから簡単であると予想される．あまり簡単な要求だけで評価をすると要約手法間の差がわかりにくくなるなるため，適度に難しい要求を選択する必要もある．しかし，Jing ら [4] も指摘しているように，どのようにして，良い検索要求を選択するかについては今後の課題である．

3.4 評価実験における問題点

3.4.1 全文参照の影響

今回の実験では被験者が適合性の判断に迷う場合に全文の参照を許したが，この点が評価の際に問題となる．まず，全文の参照が精度にどのくらい影響するかがはっきりしない．全文参照を許さなかったとしたら実際の精度がどのくらいになるかを知る手がかりがない限りは，実験での全文の参照を許さない方が良い．また，タスクにかかる時間も，全文の参照によって長くなると思われるが，影響ははっきりしない．この点からも参照を許さない方が良い．しかし，実際には要約だけでは判断のつけようがない場合も確かに存在するため難しい問題である．

3.4.2 読み易さの判定

今回の実験では，タスクにかかった時間の要約手法間での差がはっきりしなかったが，これは読み易さの判定を同時に行なったためであると思われる．読み易さの判定は，適合性判断に比べてかなり時間がかかるようであり，各手法間の時間差がほとんどなくなってしまうため，情報検索タスクに基づく要約の評価実験とは別の実験にした方が良い．

3.4.3 検索要求とテキストの選択

検索要求に関係する単語が比較的均等に散らばっているテキストと要求の組み合わせでは，テキストのどの部分を取り出しても検索要求との判断がつきやすい要約になる可能性がある．また，分布が均等でなくても，検索語の出現頻度が高いテキストでは，検索要求を考慮した手法としない手法で作成される要約の差がつきにくくなる可能性がある．

要約手法間の相違をよりはっきりさせるために，以上の点を考慮して検索要求とテキストを選択する必要がある．今回は正解テキストにはA判定を用いたが，B判定[8]の使用についても今後検討する必要がある．最適な検索要求とテキストの選択は，検索要求の難易度とも関わる難しい問題でもあり今後の課題である．

4 おわりに

本稿では，情報検索タスクに基づいた要約手法間の比較を行なった．結果から語彙的連鎖に基づくパッセージ抽出による手法が情報検索タスクにおいて，良い要約を作成できることを示した．また，今回の実験から明らかになった，評価尺度，実験の設定，検索要求とテキストの選択などの問題点と今後の課題についても報告した．

[8] B判定では，検索要求の内容を少しでも記述している記事を正解としている．

謝辞

本研究では，(社) 情報処理学会・データベースシステム研究会が，新情報処理開発機構との共同作業により構築したBMIR-J2を利用させていただきました．感謝致します．

参考文献

[1] M. H. Degroot, et al., editors. *Encyclopedia of Statistical Sciences*, Vol. 2, pp. 24–26. A Wiley-Interscience Publication, 1981.

[2] D. B. Duncan. Gives details of the test procedure, and explains the reasons for using modified significance levels. *Biometrics*, Vol. 11, pp. 1–42, 1955.

[3] H.A.K. Halliday and R. Hasan. *Cohesion in English*. Longman, 1976.

[4] H. Jing, R. Barzilay, K. McKeown, and M. Elhadad. Summarization Evaluation Methods: Experiments and Analysis. In *Intelligent Text Summarization*, pp. 51–59. AAAI Press, 199.

[5] I. Mani, et al. The tipster summac text summarization evaluation. Technical Report MTR 98W0000138, MITRE Technical Report, 1998.

[6] J. Morris and G. Hirst. Lexical Cohesion Computed by Thesaural Relations as an Indicator of the Structure of Text. *Computational Linguistics*, Vol. 17, No. 1, pp. 21–48, 1991.

[7] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, Vol. 24, pp. 513–523, 1988.

[8] A. Tombros and M. Sanderson. Advantages of Query Biased Summaries in Information Retrieval. In *Proc. of 21th Annual International ACM Special Interest Group on Information Retrieval Conference on Research and Development in Information Retrieval*, pp. 2–10, 1998.

[9] K. Zechner. Fast Generation of Abstracts from General Domain Text Corpora by Extracting Relevant Sentences. In *Proc. of 16th International Conference on Computational Linguistics*, pp. 986–989, 1996.

[10] 奥村学, 難波英嗣. テキスト自動要約に関する研究動向. 自然言語処理, Vol. 6, No. 6, pp. 1–25, 1999.

[11] 望月源, 岩山真, 奥村学. 語彙的連鎖に基づくパッセージ検索. 自然言語処理, Vol. 6, No. 3, pp. 101–126, 1999.

[12] 木谷ほか. 日本語情報検索システム評価用テキストコレクション BMIR-J2. 情報処理学会研究会資料 DBS-144-3, pp. 15–22, 1998.