# 質問に適応した文書要約手法とその評価

平尾 努[†] 佐々木 裕[†] 磯崎秀樹[†]

本稿では，TREC などで注目されている質問応答システムで利用するために，質問に適応した文書要約手法（*Question-Biased Text Summarization: QBTS*）を提案する．質問応答システムは，質問の解答となる語を含む短い文字列を文書から抽出し，回答として出力する．しかし，出力された文字列は短いために情報量が少ない．よって，利用者が出力結果の正誤を判定することが難しい．正確な正誤判定のためには文字列の抽出元となった文書を読む必要があり手間がかかる．本稿ではこうした手間の軽減を目的として，質問とその解答候補となる単語に着目し，それらが近接して出現するパッセージに含まれる文を要約文とする手法を提案する．また，提案手法の有効性を示すために質問応答タスクに基づく評価手法を採用し，既存の要約手法との比較評価を行った．その結果，提案手法が低い要約率で既存の要約手法より高精度であることを確認した．

# Question-Biased Text Summarization and Its Evaluation

Tsutomu Hirao,[†] Yutaka Sasaki[†] and Hideki Isozaki[†]

In this paper, we propose a *Question-Biased Text Summarization* (*QBTS*) that is useful for question-answering systems. Question-answering systems output short phrases as answers to a question by extracting phrases from given document. However, the information carried in the short phrases is insufficient for a user to judge the correctness. To read the source document in order to judge the correctness of outputs is not efficient. Therefore, we propose a text summarization method that is biased not only by the question but also by the prospective answers to the question. Our method is based on extraction of sentences from a passage in which words in a question and prospective answers appear closely. We conducted text summarization experiments based on QA tasks and confirmed the effectiveness of our method in obtaining short summaries.

## 1. はじめに

近年，ネットワーク環境の発展や大容量の記憶媒体の低価格化により大量の電子化文書が氾濫している．このため，必要とする情報を効率良く得ることが困難となっている．効率的に情報を得るための技術の 1 つとして文書要約技術が注目されている．要約された文書を利用することで，文書の概要の把握や原文を参照すべきかどうかの判断や必要な情報そのものを得ることが可能となり，人間の時間や労力を軽減することができる．

要約は一般的には generic な要約と user-focused な要約に分類されることが多い[10]．generic な要約とは原文に対する理想的な要約であり，不特定多数の読者を想定したものである．user-focused な要約とは要約を利用するユーザの興味を反映した要約であり，ユーザの興味というバイアスがかかっている．近年の情報検索システムの普及を背景として，検索要求に着目した user-focused な要約手法が提案されている[12),15),19)]．要約を利用することで，検索結果である文書の全文に目を通さずにそれが必要な情報を含むかどうかの判断が可能となり，人間の時間や労力を軽減できる．従来より，ある文書に対して唯一の理想的な要約を考えることが困難であることは指摘されており，要約を利用する状況を明確にすることは要約の在り方や評価にとって重要である[18)]．

本稿では質問応答システムに適用することを想定し，質問に適応した文書要約手法（*Question-Biased Text Summarization: QBTS*）を新たに提案する．本稿で考える質問応答システムとは，自然言語で入力された質問文に対して解答を含む文字列をある文書集合から抽出するシステムであり，近年，TREC☆などで注目

† 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

☆ Text REtrieval Conference, http://trec.nist.gov/

を集めている．さらに，質問応答タスクにとっての要約の有効性という観点から従来の要約手法との比較評価を行い，その有効性を示す．

以下，2 章で本研究の位置づけを明確にし，質問応答システムに求められる要約について説明する．3 章で提案手法である質問に適応した文書要約手法（*QBTS*）について述べ，4 章で評価実験の結果を示す．5 章では考察を行い，6 章で関連研究についてまとめる．最終章では結論を述べる．

## 2. 準　　備

### 2.1 要約の位置づけ

従来の文書要約の研究は対象となる文書に対して唯一の要約が存在することを仮定して進められているものが多い（たとえば，文献 9)，17)）．このような考え方の問題点を指摘し，要約の在り方と評価の点から議論する．

文書要約とは文書中の重要な情報を抽出したものである．しかし，何を重要とするかは一意に決まるものではなく，人間によって異なる．たとえば，情報検索で検索された文書を要約する場合，要約は検索要求の視点から行われるべきである．また，複数の話題を含む文書においても，それを読むユーザがどの話題に対して興味を持っているかによって要約は変化すべきである．このように，利用状況に応じて要約は変化すると考えた方が自然である[18]．

一方，要約の評価という観点からは，文書に対して唯一の要約を仮定できると都合よい．なぜならば，システムが出力した要約との比較により評価することが可能となるからである．しかし，要約の正解データを作成することは一般的に困難である．複数の人間が同一の文書に対して要約を作成したとしても，それらが高い一致率になるとは限らず，正解データとしての妥当性にも問題がある．また，仮に理想的な正解データを作成できたとしても，単にシステムの出力との一致率を計算するだけでは要約としての良し悪しを十分に評価できないという問題もある[6]．上記したことを考慮すると，要約の利用状況下における有効性を考えて評価すべきである．

このような考えに基づき，本稿では質問応答システムに適用するための user-focused な文書要約手法を新たに提案し，質問応答タスクにとっての有効性という観点から評価を行う．

### 2.2 質 問 応 答

TREC-8 の Q&A Track では以下の条件を満たす質問を対象として質問応答タスクを定義している[16]．

- ある文書集合中の文書に解答が存在する（ただし，解答は複数のテキストにまたがらない）．
- 特別な知識を持たない人間がその文書を読むことで解答できる．

本稿では，上記の条件を参考に作成した日本語質問応答ベンチマークテストセット NTT-QA-2000-08-23-FRUN を用いた質問応答タスクを対象とする[25]．質問には具体的な解答が文字列で表されるので，情報検索における検索要求とは異なる．

このベンチマークテストセットは，毎日新聞 94 年の 1 年分を対象に作成されたものであり，50 問の質問と解答，解答を含む記事の例が収録されている．また，上記の条件に加え以下の条件が考慮されている．

- 質問文は正解が IREX☆の固有表現基準による固有表現および数値表現になるものに限定．具体的には，人名，地名，組織名，人工物名，日付，時間，割合，金額とその他の数値表現．
- 質問文は省略を行わない完全な文で表す．
- 質問文に関して，場所のデフォルトを「日本」，年のデフォルトを 1994 年とする．
- 質問文は正解の判定に曖昧性がないものとする．正解が複数あることは認めるが，正解かどうかの判定が曖昧になる質問文は認めない．

### 2.3 質問応答システムに求められる要約

質問応答システムの返すべき答えは文書に含まれる文字列である．一般的には，質問文を解析し，解答候補とする意味制約（たとえば，IREX における固有表現の分類）を決定した後，意味制約を満たす単語を解答とする手法が用いられる[22),25]．質問応答システムでは従来の情報検索システムのように文書中のどこに欲する情報があるかを探す手間は省くことができる．しかし，質問に対する回答が必ずしも正解であるとは限らず，仮に，システムの回答が正解であったとしても，その根拠が提示されなければ正解かどうかを判断することもできない．つまり，結局のところシステムの出力である文字列の抽出元となった文書を読んでシステム出力の正誤を判断しなければならず，従来の情報検索システムほどではないが，長い文書の場合は負担がかかる．

**図 1** を例に質問応答システムに求められる要約について説明する．「ジョン・ル・カレの出世作のタイトルは何ですか？」という質問に対して，システムが「寒い国から帰ってきたスパイ」であると答えたとする（回

☆ Information Retrieval and Extraction Exercise, http://cs.nyu.edu/cs/projects/proteus/irex/

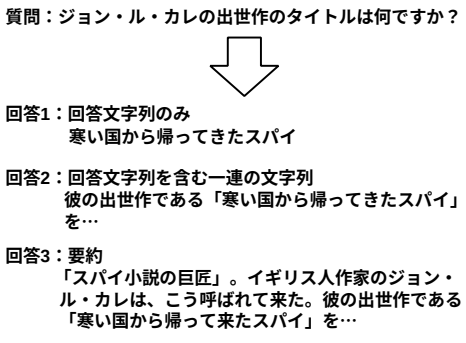


**図 1** 質問応答の例
Fig. 1 An example of question and answering.

答 1)．この際，ジョン・ル・カレの出世作が「寒い国から帰ってきたスパイ」であると知っている人間はシステムが正しく回答できたことが分かるが，そうでない場合には，システムの回答の根拠が示されない限り正しいか否かを判断することができない．また，従来の質問応答タスクで採用されているように，解答を含む文書中の連続した一部分☆で答えたとしても，「寒い国から帰ってきたスパイ」がジョン・ル・カレの出世作かどうかを判断することはできない（回答 2)．ところが，文書中の質問に対する解答候補に関連する部分と質問に関連する部分を抽出した要約である回答 3 ではその文章から「寒い国から帰ってきたスパイ」がジョン・ル・カレの出世作であることを読み取ることができる．つまり，システム出力「寒い国から帰ってきたスパイ」が正解である根拠を示している．このように質問応答システムでは，単純に解答文字列やそれを含む一部分を出力するだけでは不十分であり，解答の根拠を示す要約が必要となる．

## 3. 質問に適応した要約手法（*Question-Biased Text Summarization: QBTS*）

### 3.1 質問と解答候補を考慮した要約

従来より情報検索システムに利用するために検索要求に着目した user-focused な文書要約手法が提案されている[12),15),19)]．さらに，具体的な質問に近い検索要求に着目した要約手法も提案されている[1),3)]．これらの研究では，要約文を抽出する際の手がかりとして検索要求（あるいは質問）に含まれる単語を利用している．しかし，質問応答システムでは，解答の根拠を示すことも要求されるので，検索要求に含まれる単語に着目するだけでは不十分である．そこで，本稿では質問に含まれる単語と解答候補の単語に着目した要約手法を新たに提案する．情報検索の分野において提案されているパッセージ検索[2),5),7),14),20)] の手法を取り入れ，質問に含まれる単語と解答候補の単語が高い密度で出現するパッセージに含まれる文を要約文とする．

☆ TREC では 50 byte か 250 byte，文献 25) では 50 byte の回答を出力する．

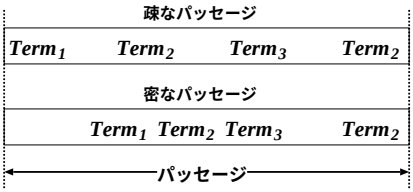


**図 2** パッセージの例
Fig. 2 Examples of passages in a text.

### 3.2 パッセージの定義

パッセージ検索に用いられるパッセージとは以下のような種類に分類される[20)]．

- 書き手が文書に付与した形式的情報によるパッセージ[14)]
- 文書中の固定長や可変長の窓によるパッセージ[2),7)]
- 文書中の意味的なまとまりによるパッセージ[5)]

書き手が文書に付与した形式的情報によるパッセージとは章，節，段落などの情報に基づき分割された文書の一部分である．文書中の固定長や可変長の窓によるパッセージとは任意の文字数からなる文書中の連続した文字列である．意味的なまとまりによるパッセージとは談話セグメントによって分割された文書の一部分である．また，検索要求が与えられた場合に検索要求に応じて動的にパッセージを作る手法も提案されている[20)]．

本稿で対象とするパッセージは，上記分類の書き手が付与した形式的なパッセージの中に設けた固定長のパッセージである．形式段落と談話セグメントは必ずしも一致しないが，こうして，文書を段落に分割することで文書中の各話題を考慮してパッセージを決定することができると考える．

### 3.3 ハニング窓関数を用いたパッセージ重要度の計算

パッセージ検索においてパッセージの重要度は，そこに含まれる検索語により決定される．一般的には $tf \cdot idf$ 法に基づく計算法が用いられる．しかし，通常，$tf \cdot idf$ 法でパッセージの重要度を計算した場合には，**図 2** の疎なパッセージと密なパッセージの重要度は等しくなる．しかし，検索語が近接して出現した場合には検索語間の関連性が高い可能性があり，出現位置を考慮して重要度を計算することが有効である[8),24)]．また，質問応答システムにおいても解答候

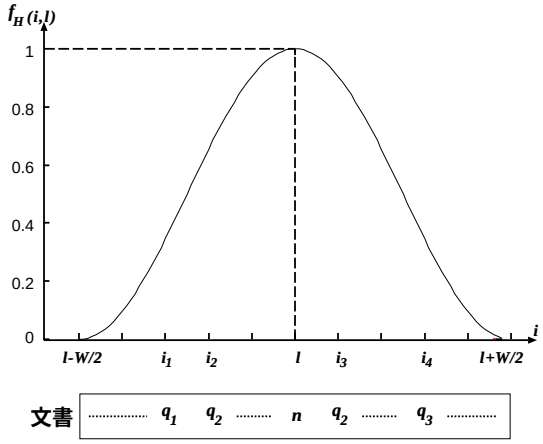


**図 3** パッセージ重要度の計算方法
Fig. 3 Calculating of the passage importance.

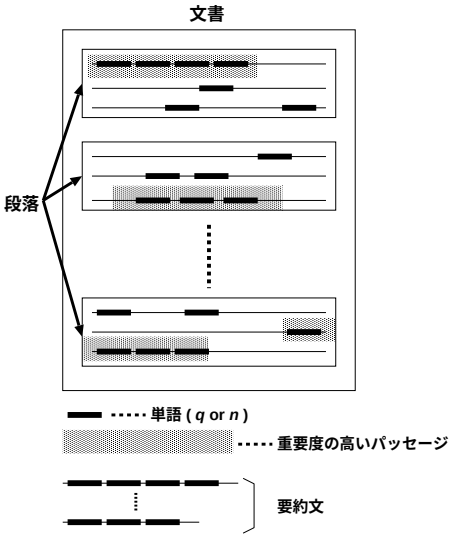


**図 4** 要約文の抽出
Fig. 4 An example of sentence extraction.

補の単語と質問文から得た単語との距離を考慮することは有効である[13]．よって，パッセージ内部に語が密集して出現している場合には重要度が高くなるようにパッセージ重要度を定義すればよい．1 つの方法として，ハニング窓関数を用いる手法が提案されている[23]．本稿では，この手法を利用してパッセージの重要度を計算する．窓の幅を $W$，中心を $l$ とすると，ハニング窓関数 $f_H(i,l)$ は以下の式 (1) により定義される．

$$f_H(i,l) \stackrel{\text{def}}{=} \begin{cases} \frac{(1+\cos 2\pi \frac{i-l}{W})}{2} & (|i-l| \leq W/2) \\ 0 & (|i-l| > W/2) \end{cases} \quad (1)$$

これを用いて，窓の中心 $l$ における重要度 $S(l)$ を以下の式 (2) で定義する（**図 3**）．

$$S(l) \stackrel{\text{def}}{=} \sum_{i=l-W/2}^{l+W/2} f_H(i,l) \cdot a(i) \quad (2)$$

ただし，$a(i)$ は以下のように定義する．質問文を形態素解析した結果から抽出した自立語の集合を $Q$，質問文を解析して得た意味制約☆を満たす対象文書中の解答候補の集合を $N$ とする．

$$a(i) \stackrel{\text{def}}{=} \begin{cases} idf(q) & \text{位置 } i \text{ を先頭として単語 } q(\in Q) \text{ が出現するとき} \\ \alpha & \text{位置 } i \text{ を先頭として解答候補 } n(\in N) \text{ が出現するとき} \\ 0 & \text{上記以外} \end{cases}$$

☆ たとえば，「～は誰ですか？」という質問であれば，人名を表す固有表現の集合を $N$ とする．

また，$idf(q)$ は以下の式 (3) で与えられる．

$$idf(q) \stackrel{\text{def}}{=} \log\left(\frac{D}{df(q)}\right) \quad (3)$$

$df(q)$ は単語 $q$ が出現する文書頻度，$D$ は文書集合に含まれる全文書数である．$\alpha$ は解答候補 $n$ の重みを表す値であり，実験により最適値を決定する．

このように質問文から得た意味制約を満たす解答候補と質問文から得た自立語に着目し，それらが近接して出現した場合の重要度を高くすることで，質問に対する解答を含むパッセージを抽出できると考える．また，質問文から得た単語が近接して出現する場合も質問に対して関連の強いパッセージであると考える．

### 3.4 パッセージ重要度に基づく要約文抽出法

前節で説明したハニング窓関数を用いてパッセージ重要度を計算し，それに基づき要約文を抽出する手法を説明する．まず，要約文抽出の概略図を**図 4** に示す．式 (2) を用いてパッセージの重要度を計算し，重要度の高いパッセージに含まれる文を要約文として文書から抽出する．手順の詳細を以下に示す．

**手順 1** 質問文を形態素解析器「茶筌」[26] を用いて解析した結果から自立語の集合 $Q$ を抽出する．質問文をパターンマッチングにより解析し意味制約を得る．質問文から得た意味制約を満たす固有表現の集合 $N$ を磯崎の固有表現抽出ルール[21] により対象文書から自動的に抽出する．

**手順 2** 文書を形式段落により $p_1, \cdots, p_n$ に分割する $(P = \{p_1, \cdots, p_n\})$．

**手順 3** 各段落 $p_i(\in P)$ に対して，**手順 4**～**手順 5** を

用いる．

**手順 4** 段落の先頭から末尾に向かって幅 $W$ のハニング窓を 1 文字単位で移動させ，その中心 $l$ におけるパッセージの重要度 $S(l)$ を式 (2) により求める．ただし，段落 $p_i$ の文字数を $|p_i|$ とすると，$l$ は $1 \leq l \leq |p_i|$ を満たす．

**手順 5** **手順 4** で求めた $S(l)$ の $p_i$ における最大値を $S_{p_i}$ とし，$S_{p_i}$ を与える位置 $l$ を中心とした幅 $W$ の窓に全体，あるいは一部が含まれる文を段落 $p_i$ における要約文候補 $I_{pi}$ とする．窓が複数の文にまたがる場合には複数文を，単一文の場合にはその 1 文を要約文候補とする．

**手順 6** 指定された要約率に最も近くなるように $S_{p_i}$ の高いものから順に要約文候補 $I_{pi}$ を要約文として採用する．

このような手順をとることで，質問に対する解答を含むことが期待できるパッセージと質問と関連性が強いパッセージを抽出することができ，そこに含まれる文を要約文とすることで質問に対する解答の根拠として利用可能な要約を作成できると考える．

## 4. 評価実験

本稿では，質問に対する解答とその根拠を含んだ要約を作成することを目標としている．よって，要約に解答文字列が含まれているかを評価するだけでは不十分であり，その根拠を含んでいるかどうかも評価しなければならない．そこで本稿では，人間による要約を利用した質問応答タスクの回答精度に基づき間接的に要約手法を評価する．つまり，人間が精度良く回答できるということは，要約が解答とその根拠を含んでいることを意味し，質問応答システムに適した要約であると考える．

### 4.1 評価対象とする要約手法

評価対象とする要約手法は，既存の要約手法である lead 手法と単語重要度に基づく手法，提案手法である．なお，参考として全文（以下，**Full**）も評価実験の対象とした．以下，それぞれについて説明する．

**lead 手法**

文書の先頭から順に要約文として採用する手法．文書の文字数 $C_t$，要約の文字数 $C_s$ を用いて要約率 $R_s$ を以下の式 (4) で定義し，要約率 10%，30%，50%のそれぞれの場合を評価する☆．以下，それぞれ**L(10)**，**L(30)**，**L(50)** と表す．

$$R_s \stackrel{\text{def}}{=} \frac{C_s}{C_t} \times 100 \tag{4}$$

たとえば，要約率 10% ($R_s = 10$) とは原文の文字数に対して 1/10 の文字数に圧縮することである．つまり，要約率が低いことは原文に対する圧縮率が高いことを表す．

**単語重要度に基づく要約手法**

重要語を多く含む文が重要であると考え，単語の重要度 $w(t)$ の加重和を文の重要度とする手法[17]．文書中の文 $s$ の重要度 $Sc(s)$ は，文 $s$ に出現する単語集合を $T_s$ とすると以下の式 (5) で表される．

$$Sc(s) \stackrel{\text{def}}{=} \sum_{t \in T_s} n(t,s) \cdot w(t) \tag{5}$$

ここで，$n(t,s)$ は文 $s$ における単語 $t$ の出現頻度である．$w(t)$ は以下のように定義する．$tf(t,d)$ は要約対象となる文書 $d$ 全体における単語 $t$ の出現頻度である．

$$w(t) \stackrel{\text{def}}{=} \begin{cases} \beta \cdot tf(t,d) \cdot idf(t) & t \in Q \\ tf(t,d) \cdot idf(t) & \text{上記以外} \end{cases}$$

$\beta$ は単語 $t$ が質問文に含まれる場合に考慮する重み係数である．今回の評価実験では，いくつかの文書に対し $\beta$ を変化させて $\beta = 1$ の場合と抽出される文の違いが大きかった $\beta = 7$ を採用した．質問に含まれる単語を考慮しているので，user-focused な要約手法であるといえる．要約率 10%，30%，50%のそれぞれの場合を評価する．以下，**T(10)**，**T(30)**，**T(50)** と表す．

**質問に適応した要約手法（*QBTS*）**

3 章で提案した手法．文献 25) の質問応答タスクに対して，本手法と同じスコア計算法を採用し，窓幅 $W$ と解答候補単語の重み $\alpha$ をパラメータとして最高スコアを与える位置を中心とした 50 byte の出力を評価した結果，最も高精度であった $W = 50$，$\alpha = 2.1$ を採用した．要約率 10%，30%，50%のそれぞれの場合を評価する．以下，**QT(10)**，**QT(30)**，**QT(50)** と表す．

先述した lead 手法，単語重要度に基づく手法は，文書中のすべての文がゼロでない重要度を持つ．これに対し，提案手法では重要度がゼロとなる文が発生するため，指定の要約率よりも低いことがある．

### 4.2 実験の設定

前節で説明した各手法による要約を用いて人間を対象に質問応答タスクを行い，その精度から要約手法を評価する．

質問は日本語質問応答ベンチマークテストセット

☆ ただし，要約率は文字に基づき決定しているので，指定された要約率に最も近くなるように文を抽出する．

**表 1** 質問と解答
Table 1 Examples of questions and answers.

| 質問番号 | 質問 | 正解例 |
|---|---|---|
| 3 | APEC の首脳会議で採択されたのは何という宣言ですか？ | ボゴール宣言 |
| 6 | 日本初の純国産航空機は何という名前ですか？ | YS11 |
| 11 | 細川護熙の後を継いで首相になったのは誰ですか？ | 羽田 孜 |
| 13 | 新進党が結成された時点での党首は誰ですか？ | 海部 俊樹 |
| 14 | 12 月に日銀総裁になったのは誰ですか？ | 松下 康雄 |
| 16 | セルビアの大統領は誰ですか？ | ミロシェビッチ |
| 22 | 三菱銀行が支援を行い子会社とした銀行は何と言いますか？ | 日本信託銀行 |
| 35 | コンコルドが関西国際空港にやって来たのはいつですか？ | 9 月 5 日 |
| 36 | 金日成が亡くなったのは何月何日ですか？ | 7 月 8 日 |
| 43 | 政府が決定した 94 年度の予算案は総額いくらですか？ | 七十三兆八百十六億円 |

**表 2** 質問と要約に対する被験者の割当て
Table 2 Examples of human resource mapping.

| | Full | L(10) | T(10) | ··· | QT(50) |
|---|---|---|---|---|---|
| 質問 3 | $G_1$ | $G_2$ | $G_3$ | ··· | $G_{10}$ |
| 質問 6 | $G_{10}$ | $G_1$ | $G_2$ | ··· | $G_9$ |
| 質問 11 | $G_9$ | $G_{10}$ | $G_1$ | ··· | $G_8$ |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 質問 43 | $G_2$ | $G_3$ | $G_4$ | ··· | $G_1$ |

NTT-QA-2000-08-23-FRUN より，正解とその根拠を含む文書の数が多い 10 問を選んで使用する（**表 1**）．各質問に対して 10 文書を選び，各手法により要約する．10 文書は，質問文から得た単語をキーワードとして $tf \cdot idf$ 法によりスコアリングした結果の上位文書から正解文字列を含む文書が過半数程度になるように選択した．

被験者は日本語を母国語とする関東圏の 7 大学の大学院生 80 名とした．8 名を 1 グループとして 10 グループに分ける．各グループが各要約手法を 1 度だけ評価するように，10 種の要約を各質問に対して割り当てる（**表 2**）．被験者は，与えられた質問と要約に対して，各要約を読むことで質問に回答できると判断した場合にのみ回答を作成する（**図 5**）．

### 4.3 評価指標

評価指標としては，適合率（Precision），再現率（Recall），適合率と再現率の調和平均である F 値（F-measure）を用いる．被験者が回答を作成した文書数を $a$，**Full** において正解文字列とその根拠を含む文書数を $b$，被験者が正解した文書数を $c$ とすると，それぞれ以下の式で与えられる．

$$\text{適合率 } (P) = \frac{c}{a} \tag{6}$$

$$\text{再現率 } (R) = \frac{c}{b} \tag{7}$$

$$F \text{ 値} = \frac{1+\gamma^2}{\frac{1}{P}+\gamma^2\frac{1}{R}} \tag{8}$$

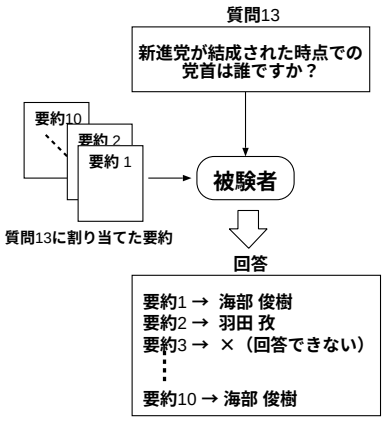


**図 5** 評価実験
Fig. 5 An example of the experiment.

$\gamma$ は重みづけ係数であり，今回は $\gamma = 1$ とした．

また，1 つの質問に対して回答にかかった時間も計測した．

## 5. 実験結果と考察

### 5.1 実験結果

今回の評価実験では被験者間の大きな能力差がないことを前提としている．そこで，被験者の F 値を式 (8) により計算し，グループ間の被験者の F 値の平均に有意差があるかどうかを一元配置分散分析を用いて調べる．帰無仮説を「各グループ間における被験者の F 値の平均は等しい」，対立仮説を「各グループ間における被験者の F 値の平均は等しくない」として，有意水準 5%で検定した．その結果，$p = 0.00065$ ($< 0.05$) となり，帰無仮説が棄却された．よって，グループ間に被験者の偏りがあることが分かった．そこで，各グループから F 値が低い被験者を除いた 45 名を対象として評価を行う．これにより，各グループを構成する被験者は 3 名から 6 名となった．グループ間の

**表 3** 評価結果
Table 3 Experimental results.

| | Full | L(10) | T(10) | QT(10) | L(30) | T(30) | QT(30) | L(50) | T(50) | QT(50) |
|---|---|---|---|---|---|---|---|---|---|---|
| F 値 | 0.87 | 0.58 | 0.40 | **0.65** | **0.74** | 0.51 | **0.74** | **0.76** | 0.69 | 0.71 |
| 適合率 | 0.94 | 0.91 | 0.81 | **0.92** | **0.94** | 0.87 | **0.94** | **0.93** | 0.92 | 0.86 |
| 再現率 | 0.84 | 0.47 | 0.28 | **0.54** | **0.66** | 0.39 | **0.66** | **0.68** | 0.57 | 0.64 |
| 文字数 | 1324 | 143 | 149 | 108 | 397 | 398 | 202 | 667 | 661 | 266 |
| 文数 | 30.95 | 3.32 | 1.83 | 2.06 | 9.05 | 5.85 | 3.90 | 15.18 | 10.84 | 5.15 |
| 時間（分：秒） | 7:41 | 2:21 | 2:28 | 2:49 | 3:56 | 4:23 | 3:09 | 5:13 | 5:08 | 4:03 |

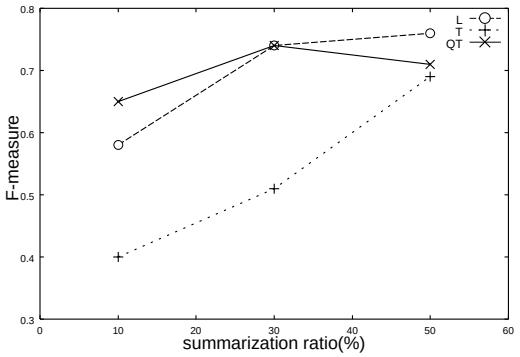


**図 6** 要約率と F 値の関係
Fig. 6 A performance/summarization-rate relation.

差を先述と同様に一元配置分散分析で調べたところ $p = 0.099$ ($> 0.05$) となり，有意差がないことを確認した．

各質問ごとに求めた適合率，再現率，F 値の平均値を**表 3** に示す．また，各要約の文字数の平均値と要約文の数の平均値もあわせて示す．**図 6** には要約率と F 値の関係を示す．

### 5.2 考　　察

#### 5.2.1 各要約率での比較

各要約率における手法間の違いを考察する．

10%の要約率で比較した場合，QT が F 値，適合率，再現率のすべてにおいて高精度である．F 値では L より 0.07，T より 0.25 精度が良い．再現率では，L より 0.07，T より 0.26 精度が良い．QT の再現率が高い理由は，解答と根拠をともに保持する割合が高いからであると考える．文字数，文数ともに L より少なく提案手法の有効性が分かる．回答にかかる時間は，文字数，文数と読みやすさに左右される．L は，文字数は多いが抽出した文間に連続性があるため読みやすく，回答にかかった時間が短い．一方，T と QT では連続性が低い．よって，これらの手法は低い要約率では，lead 手法よりも読みやすさがそこなわれ，時間がかかると考える．

30%の要約率で比較した場合，QT と L が F 値，適合率，再現率ともに並び，T より F 値で 0.23，適合率で 0.07，再現率で 0.27 精度が良い．10%の場合と同様に適合率よりも再現率の差が大きい．10%では提案手法と lead 手法に差があったが要約率を上げる（圧縮率を下げる）ことで同等の精度となった．これは，各文書の先頭から 30%程度の文字列に質問に対する解答とその根拠が多く含まれていることを示す．この点に関しては後に詳しく考察する．回答にかかった時間では QT が最も短い．これは，文字数，文数ともに少ないこと，L より劣るが T よりは連続性があることによると考える．

50%の要約率で比較した場合には，L が F 値，適合率，再現率のすべてにおいて高精度である．しかし，F 値の比較では，10%や 30%の要約率の場合ほど 3 手法間に大きな差がない．各手法とも，正解文字列と根拠を抽出する確率が高くなるためであると考える．また，回答にかかった時間では抽出文字数が少ないため QT が短い．L，T では，Full の約 7 割程度の時間がかかっており，時間の効率が悪い．

#### 5.2.2 要約率による変化

次に，各手法間で要約率を変化させた場合を考察する．

lead 手法では，要約率を上げるに従い F 値，再現率が向上している．10%から 30%の精度向上率が 30%から 50%の精度向上率よりも大きい．回答にかかる時間とともに長くなるが，その上昇率は 10%から 30%の変化が大きい．

単語重要度に基づく手法では，要約率を上げるに従い F 値，適合率，再現率が向上している．これは，要約率を上げることで，正解とその根拠を含む割合が多くなり，人間の判断の的確性が増していることを示す．ただし，すべての要約率において他の 2 手法よりも精度は低い．回答にかかった時間は，lead 手法と同様の傾向で 10%から 30%の場合の上昇率が大きい．

提案手法では，10%から 30%では，F 値，適合率，再現率ともに向上がみられたが，30%と 50%は精度の低下がみられた．これは，関連性の低い文によって被験者が混乱を起こしたためであると考える．提案手法では，文書中の質問に関連する段落から文を抽出す

**表 4** 正解文字列の分布
Table 4 Distribution of correct strings.

| 出現位置 | 文書の割合 |
|---|---|
| 10%未満 | 0.56 |
| 10%以上 30%未満 | 0.24 |
| 30%以上 50%未満 | 0.07 |
| 50%以上 | 0.13 |

るため文間の連続性が低いことが原因であると考える．50%に対し新たに抽出した文が原因であることはQT(30) と QT(50) を比較すると抽出文字数の増加率が低いにもかかわらず，タスクにかかった時間の上昇率が大きいことからも分かる．

また，Full では正解文字列とそれを正解と認定するための文脈はすべて文書中に含まれている．しかし，文字数が多いことが読み落しや書き手と異なる解釈を生ずる原因となり F 値，適合率，再現率のすべてにおいて 1.0 の精度は得られていない．

以上，F 値，適合率，再現率から考察した結果，要約率 10%では提案手法，要約率 30%では提案手法と lead 手法，要約率 50%では lead 手法が優れていることが分かった．また，回答にかかった時間では，要約率 10%では lead 手法，30%では提案手法，50%では提案手法が短かった．ここで，実際に要約を利用する場合には文字数の制限などもあり，より低い要約率で高精度なものが望まれることを考慮すると提案手法が精度，時間の効率の両方の面で優れているといえる．

### 5.3 正解文字列の分布について

前節で述べたとおり，lead 手法が単純な手法にもかかわらず高い精度である．この原因として，質問に対する正解とその根拠が文書の先頭付近に分布していることが推測できる．そこで，質問に対する正解の出現位置を調べたところ**表 4** を得た．文書の先頭から30%までの位置に約 8 割の正解が含まれている．ただし，正解だけではなくその根拠がなければ正解と認定できないため，L(30) における再現率は 66%であり，正解を含む文書の割合よりは低い．

評価実験では，正解とその根拠を含む文書が多い質問を選んだ．これは，質問に関連する事項が複数の文書で述べられていることを示す．新聞記事の場合，複数の文書で言及される話題はその文書の主題になることが多い．たとえば,「APEC の首脳会議」というイベントは多くの文書で述べられており，その多くは文書の主題として取り上げている．このように，利用したベンチマークテストセットでは，正解とその根拠を含む文書が多い質問は文書の主題との関連が強いことが分かった．よって，新聞記事では，文書の先頭付近

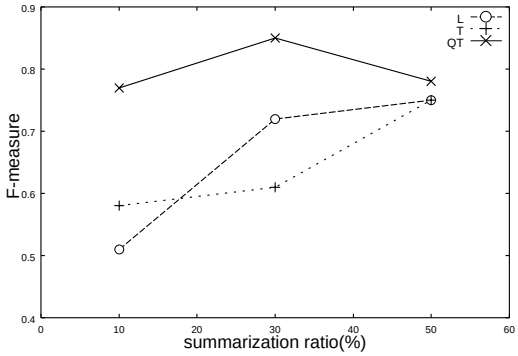


**図 7** 4 問での要約率と F 値の関係
Fig. 7 A performance/summarization-rate relation for four questions.

が主題を表していることが多いので，正解が文書の先頭付近に偏ったと考える．

正解の分布が文書の先頭付近に偏っている場合では，generic な要約手法と user-focused な要約手法との差がつきにくい．そこで，正解文字列の出現位置が文書中に分散している 4 問（質問 3，16，22，36）だけの評価結果を**表 5** に示す．**図 7** には要約率と F 値の関係を示す．当然ではあるが，正解が比較的先頭に偏っていないことで lead 手法の精度は表 3 と比較すると低下している．これに対し，単語重要度に基づく手法，提案手法ともに精度が向上しており，正解文字列の分布が文書の先頭に偏っていない場合には質問から得た情報を考慮して要約を作成することの有効性が分かる．さらに，user-focused な要約手法である提案手法と単語重要度に基づく手法を比較した場合には提案手法が高精度であり，解答候補を考慮した効果が現れていると考える．

ユーザが自由に質問を入力できる質問応答システムでは，質問に対する解答とその根拠は文書の先頭付近につねに出現するわけではない．よって，どのような位置に正解が出現しても対応できることが望ましい．提案手法は解答が文書の先頭に偏っていない場合でも高精度であるので，既存の要約手法よりも質問応答システムに適しているといえる．

## 6. 関 連 研 究

本稿の主張点は 2 つある．1 つは要約手法であり，もう 1 つは評価法である．

要約手法の特徴としては，質問文に含まれる単語と質問文から得た意味制約を満たす固有表現を重要と考え，ハニング窓によってその分布を考慮していることがあげられる．重要語が近接している箇所に着目して

**表 5** 4 問による評価結果
Table 5 Experimental results for four questions.

| | Full | L(10) | T(10) | QT(10) | L(30) | T(30) | QT(30) | L(50) | T(50) | QT(50) |
|---|---|---|---|---|---|---|---|---|---|---|
| F 値 | 0.90 | 0.51 | 0.58 | **0.77** | 0.72 | 0.61 | **0.85** | 0.75 | 0.75 | **0.78** |
| 適合率 | 0.93 | 0.90 | **0.97** | 0.91 | **0.95** | 0.92 | 0.93 | 0.93 | **0.94** | 0.92 |
| 再現率 | 0.88 | 0.39 | 0.42 | **0.70** | 0.62 | 0.48 | **0.79** | 0.65 | 0.64 | **0.73** |

重要度を計算する手法は文書要約の最初の研究ともいわれる Luhn の研究でも提案されている[9]．Luhn は，1 文中で重要語が 4 語以上離れずに出現する部分に着目し，そこに含まれる重要語と非重要語の割合に着目して文の重要度を計算している．しかし，ハニング窓のように単語間の距離をより細かく考慮できず，1 文に閉じて重要度を計算する点も異なる．ハニング窓関数を用いて単一の単語の出現距離を考慮する手法は黒橋らによって提案されており[23]，本稿では，この研究で提案された手法を応用している．ただし，単一の単語のみに着目するのではなく，質問に含まれる単語と質問に対する解答候補の単語に着目しそれらの単語の重みを考慮している点が異なる．

評価法の特徴としては，質問応答タスクに基づく評価法を採用したことがあげられる．このようなタスクに基づく要約の評価法は近年になっていくつか提案されている．Jing ら[6]，Hand[4]，TIPSTER SUMMAC プロジェクト[11]，望月ら[19] は要約を情報検索タスクにおける検索結果の適合性判断に用いて評価する手法を提案している．質問応答タスクも一種の情報検索タスクであるといえるが，文書単位で答えるか文字列（パッセージなど）で答えるかという回答の粒度に違いがある．

なお，TIPSTER SUMMAC プロジェクトでは，ある検索要求に対してその検索結果である文書が含むであろう重要事項をたずねる質問をいくつか作成し，検索要求に応じて作成した要約がそれらの質問に対する正解を含む割合をもとに評価する手法も提案されている．これも，質問応答タスクに基づく要約の評価法であり，本稿での評価法と似通っている．しかし，情報検索の結果，適合した文書に求められる要約を評価するために提案された手法であり，本稿のように質問応答における解答の根拠を示す要約を評価するための手法ではない．さらに，要約が正解を含むかどうかの判断は，少数の人間の判断に委ねられており，本稿での評価実験のように多数の被験者の判断に基づくものではない点も異なる．

## 7. おわりに

本稿では，質問応答システムで利用するために質問に適応した文書要約法を提案した．評価手法として質問応答タスクに基づく評価法を採用し，既存の要約手法との比較を行った．評価対象は，lead 手法，単語重要度に基づく手法，提案手法である．その結果，低い要約率（原文に対して高い圧縮率）では提案手法が既存の要約手法より高精度であることを確認した．

**謝辞** 研究を進めるにあたり様々なご助言をいただいた，奈良先端科学技術大学院大学の松本裕治教授に感謝いたします．また，評価手法に関して有益なコメントをいただいた，北陸先端科学技術大学院大学の望月源氏に感謝いたします．

## 参 考 文 献

1) Berger, A. and Mittal, V.M.: Query-Relevant Summarization using FAQs, *Proc. 38th Annual Meeting of the Association for Computational Linguistics*, pp.294–301 (2000).
2) Callan, J.P.: Passage-Level Evidence in Document Retrieval, *Proc. 17th Annual International ACM SIGIR Conference on Research and Development in the Information Retrieval*, pp.302–310 (1994).
3) Chali, Y., Matwin, S. and Szpakowicz, S.: Query-Biased Text Summarization as a Question-Answering Technique, *Proc. AAAI Fall Symposium on Question Answering Systems*, pp.52–56 (1999).
4) Hand, T.: A Proposal for Task-based Evaluation of Text Summarization Systems, *Proc. ACL Workshop on Intelligent Scalable Text Summarization*, pp.31–38 (1997).
5) Hearst, M.A. and Plaunt, C.: Subtopic Structuring for Full-Length Document Access, *Proc. 16th Annual International ACM SIGIR Conference on Research and Development in the Information Retrieval*, pp.59–68 (1993).
6) Jing, H., Bazilay, R., McKeown, K. and Elhadad, M.: Summarization Evaluation Methods: Experiments and Analysis, *AAAI Intelligent Text Summarization Workshop*, pp.51–59 (1998).
7) Kaszkiel, M. and Zobel, J.: Passage Retrieval Revisited, *Proc. 20th Annual International ACM SIGIR Conference on Research and Development in the Information Retrieval*,

pp.302–310 (1997).
8) Keen, E.: The Use of Term Position Devices in Ranked Output Experiments, *The Journal of Document*, Vol.47, No.1, pp.1–22 (1991).
9) Luhn, H.: The automatic creation of literature abstracts, *IBM Journal of Research and Development*, Vol.2, No.2, pp.159–165 (1958).
10) Mani, I. and Bloedorn, E.: Machine Learnig of General and User-Focused Summarization, *Proc. 15th National Conference on Artifical Intelligence*, pp.821–826 (1998).
11) Mani, I., House, D., Klein, G., Hirschman, L., Obrst, L., Firmin, T., Chrzanowski, M. and Sundheim, B.: The TIPSTER SUMMAC text summarization evaluation final report, Technical Report MTR98W0000138, The MITRE Corporation (1998).
12) Miike, S., Ito, E., Ono, K. and Sumita, K.: A Full-Text Retrieval System with a Dynamic Abstract Generation Function, *Proc. 17th Annual International ACM SIGIR Conference on Research and Development in the Information Retrieval*, pp.152–161 (1994).
13) Prager, J., Brown, E. and Coden, A.: Question-Answering by Predictive Annotation, *Proc. 23rd Annual International ACM SIGIR Conference on Research and Development in the Information Retrieval*, pp.184–191 (2000).
14) Salton, G., Allan, J. and Buckley, C.: Approaches to passage retrieval in full text information systems, *Proc. 16th Annual International ACM SIGIR Conference on Research and Development in the Information Retrieval*, pp.49–56 (1993).
15) Tombros, A. and Sanderson, M.: Advantages of Query Biased Summaries in Information Retrieval, *Proc. 21st Annual International ACM SIGIR Conference on Research and Development in the Information Retrieval*, pp.2–10 (1998).
16) Voorhees, E.M. and Tice, D.: Building a question-answering test collection, *Proc. 23rd Annual International ACM SIGIR Conference on Research and Development in the Information Retrieval*, pp.192–199 (2000).
17) Zechner, K.: Fast Generation of Abstracts from General Domain Text Corpora by Extracting Relevant Sentences, *Proc. 16th International Conference on Computational Linguistics*, pp.986–989 (1996).
18) 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.1–26 (1999).
19) 望月 源，奥村 学：語彙的連鎖に基づく要約の情報検索タスクを用いた評価，自然言語処理，Vol.7, No.4, pp.63–77 (2000).
20) 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.101–125 (1999).
21) 磯崎秀樹：固有表現抽出のための可読性の高い規則の自動生成，情報処理学会研究報告 NL-140-10, pp.69–76 (2000).
22) 賀沢秀人，加藤恒昭：意味制約を用いた日本語質問応答システム，情報処理学会研究報告 NL-140-24, pp.173–180 (2000).
23) 黒橋禎夫，白木伸征，長尾 真：出現密度分布を用いた語の重要説明箇所の特定，情報処理学会論文誌，Vol.38, No.4, pp.845–853 (1997).
24) 高木 徹，木谷 強：共起単語間の関連性を考慮した文書重要度付与，情報処理学会論文誌：データベース，Vol.40, No.SIG 8, pp.74–84 (1999).
25) 佐々木裕ほか：質問応答システムの比較と評価，電子情報通信学会研究報告 NLC-2000-24, pp.17–24 (2000).
26) 松本裕治，北内 啓，山下達雄，平野善隆，浅原正幸，松田 寛：日本語形態素解析システム『茶筌』version 2.0 使用説明書 第二版，Information Science Technical Report NAIST-IS-TR99012, Nara Institute of Science and Technology (1999).

(平成 13 年 3 月 23 日受付)
(平成 13 年 6 月 19 日採録)

**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，NTT データ通信株式会社（現，株式会社 NTT データ）入社．2000 年より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に所属．自然言語処理の研究に従事．ACL 会員．

**佐々木 裕**

1986 年筑波大学第三学群情報学類卒業．1988 年同大学院修士課程理工学研究科修了．同年，日本電信電話株式会社入社．現在，NTT コミュニケーション科学基礎研究所に所属．工学博士．1995〜1996 年サイモン・フレーザー大学（カナダ）客員研究員．主として自然言語処理，機械学習に関する研究に従事．人工知能学会，言語処理学会各会員．

**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，日本電信電話株式会社入社．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特別研究員．博士（工学）．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，日本ソフトウェア科学会，言語処理学会，AAAI，ACL 各会員．