平成 14 年度

修士学位論文

記事の特徴を利用した新聞記事検索手法

A Method of Article Retrieval Utilizing Characteristics in Newspaper Articles

1055104　　友池 貴之

指導教員　　坂本 明雄

2003 年 1 月 31 日

高知工科大学大学院 工学研究科 基盤工学専攻

情報システム工学コース

要 旨

記事の特徴を利用した新聞記事検索手法

友池 貴之

現在，膨大な情報の中から，必要な情報を効率よく取り出すテキスト処理技術に対する関心が高まっている．ユーザの質問文に対して的確な答えを提示する質問応答技術や，重要な部分を損なわずにテキストをコンパクトにまとめる自動要約技術など，様々な観点からの技術研究が進められている．

本論文では，質問応答技術における文書検索手法について新聞記事の特徴を用いた検索手法を提案する．本検索手法では，検索対象に新聞記事データを用いる場合において，新聞記事に見られる特徴を用いることで検索性能の向上をねらうアプローチをとっている．新聞記事の特徴として，本文の 1 文目には結論が書かれることが多い，各段落の 1 文目は段落内で重要であることが多い，役職や年齢が添えられた人名は重要であることが多い，等を考えた．

ベースとなる検索手法として，tf-idf 法による検索索引語の重み付けを利用しているが，同手法には，検索対象の文書が長いほど優先的に検索されてしまう問題があることが知られている．本論文では，この問題を解決する方法として，重要文抽出法によるテキスト自動要約技術を用いた検索対象の新聞記事データ長の制限手法も提案している．この要約に対してもまた新聞記事の特徴を用いている．

キーワード　質問応答，情報検索，tf-idf 法，テキスト自動要約

Abstract

A Method of Article Retrieval Utilizing Characteristics in Newspaper Articles

TOMOIKE Takayuki

The concern about the text processing technology which takes out required information from huge information is increasing now. Technical research is carried out from various viewpoints, such as question answering and text summarization.

This paper describes a document retrieval method which is part of question answering system, utilizing characteristics in newspaper article. The retrieval method aims at retrieving document from newspaper articles. The examples of the characteristics in newspaper article are the first sentence of article has a conclusion in many cases, the first sentence of each paragraph is important in many cases and the name of a person to which an executive and age were attached are important in many cases.

The retrieval method is based on tf-idf weighting. However, it is known that there is a problem in the tf-idf weighting. When there is a long document in newspaper articles, it will be retrieved preferentially as compared with a short one. This paper describes the problem solution method which uses text summarization technique too.

key words Question Answering, Information Retrieval, tf-idf Weighting, Text Summarization

目次

図目次

表目次

第 1 章

はじめに

WWW の普及に伴いインターネット上で発信される情報が急増している．今や情報を検索・閲覧する手段として WWW は生活に欠かせないものとなりつつある．しかし，発信されているそれらの情報の量に比べ，知りたい情報を的確に検索・閲覧する手段は，まだ十分とは言えない状況である．

現在，膨大な情報の中から，必要な情報を効率よく取り出すテキスト処理技術に対する関心が高まっている．ユーザの質問文に対して的確な答えを提示する質問応答技術や，重要な部分を損なわずにテキストをコンパクトにまとめる自動要約技術など，様々な観点からの技術研究が進められている．

上記のようなテキスト処理技術研究の促進と研究成果の蓄積を目的として，NTCIR 評価ワークショップ [1] が開催されている．評価ワークショップは，同じ基盤の上で，どのような技術がどのような効果をもつかを調べ，互いに学びあう場をつくり出すことができる新しい研究スタイルである．研究グループはそれぞれの目的でワークショップへ参加し，研究アイデアの効果を調べることができる．

QAC-1[2] は，第 3 回 NTCIR ワークショップのサブタスクとして行われる質問応答に関する第 1 回評価会議である．QAC-1 のタスクは，大量の文書を背景に自然言語によって尋ねられた任意の質問に答えを与えることである．ドメイン依存でないこととともに，組織化されていない情報に依存していることが RDB(relational database) に対する質問と異なる点である．QAC-1 では，検索対象の知識源として 2 年分の新聞記事データを利用している．

本論文では，質問応答における文書検索手法として，新聞記事に見られる特徴を用いる手法を提案する．また，検索の際に生じる問題を解決する手法として，テキスト自動要約技術

を用いた新聞記事データの前処理手法も提案している．なお，本研究は，QAC-1 の基盤の上で研究を行ってきたものであり，検索対象文書として 2 年分の新聞記事データを用いている．

本研究の最終的なゴールは，質問応答技術において，ユーザ側の自然言語テキストで書かれた質問文に対して，システム側が膨大な文書の中から適切な答えを含むものを回答することである．これを実現することにより，ユーザは，現在主流であるキーワードによる情報検索手法に加えて，自然言語テキストによる情報検索手法というインタフェースを得ることができる．また，より良い文書を検索結果とすることで，その文書に対して情報抽出技術を用いる際に，情報の抽出結果の精度向上に寄与できる．

第 2 章

基本的事項および関連研究

2.1 基本的事項

2.1.1 質問応答

質問応答とは，自然言語で表現された質問に適切に回答する技術である [3].

1970 年代，質問応答とは RDB を自然言語で検索可能にする技術のことを指していた．しかし，近年，注目を集めているのはオープンドメインでの質問応答技術で，膨大なテキスト集合を知識源として分野を限定しない質問を受け付けるというものである．

一般的に，質問応答システムは図 2.1 に示す様に，質問解析，文書検索，回答抽出という 3 つの要素から構成される．質問解析では，ユーザの質問文を解析し，ユーザの質問意図の理解，検索に必要な情報の取得を行う．文書検索は，質問解析で得られた情報を用いて，検索対象文書の中から答えを含むと思われる文書を検索する．解答抽出は，文書検索により検索された文書からユーザの質問に対する回答を抽出する．

たとえば，ユーザの質問文「審判にボールを投げつけた巨人の投手は誰ですか」に対して正しく回答するためには，まず，「は誰ですか」という部分から，人名を問う質問であることを理解する．次に，巨人の投手の名前が書かれている事件に関係のある文書を検索する．そして，文書の中から人名である「ガルベス」や「ガルベス投手」といった回答を抽出する，といった具合となる．

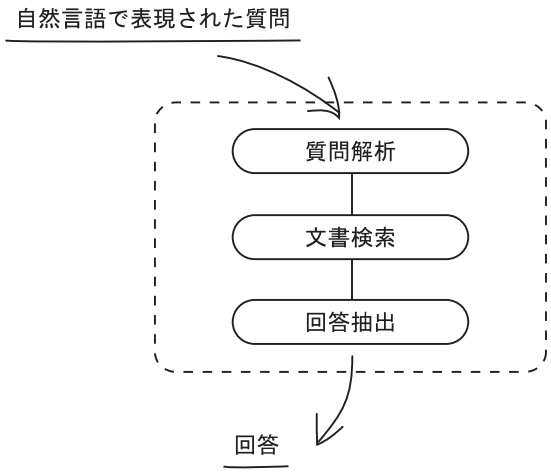

図 2.1　質問応答システムの構成例

2.1.2　形態素解析

形態素解析は，質問応答技術等の自然言語処理において第 1 段階としてよく用いられる．形態素解析の目的は，与えられた文を形態素・語の並びに分解し，それぞれの形態素・語の品詞などを決定することである [4]．このとき，形態素とは，意味を持つ最小の言語単語と定義される．また，語とは，1 つの意味のまとまりをなし，文法上 1 つの機能をもつ最小の言語単位であり，1 つ以上の形態素からなると定義される．

日本語の形態素解析を行う場合，語を区切る空白が存在しないため語の特定が困難である．各語の品詞特定には，一般に，動詞，形容詞，形容動詞，名詞，副詞，連体詞，接続詞，感動詞，助動詞，助詞の 10 品詞が用いられる．

一方，英語の場合には，語は空白で区切られた文字列と考えてよいため楽である．しかし，各語の品詞の同定は困難である．英語では，名詞は動詞としても使うことができるため品詞の曖昧性が非常に多い．

2.1.3　tf-idf 法とその問題点

質問応答技術等の情報検索分野では，一般に質問文の形態素解析により得られる索引語へ重み付けを行い，検索対象文書の重要度を求めることが多い．tf-idf 法は，文書中への索引

語の出現頻度を用いて索引語の重み付けを行うアプローチである [4].

ある索引語の文書中への出現回数を tf(term frequency)，全文書数 N に対する索引語の出現回数を df(document frequency) で表す．idf(inverse document frequency) を $idf = 1 + \log\left(\frac{N}{df}\right)$ と定義すると，tf-idf 法による索引語のある文書における重み w は

$$w = tf \cdot idf = tf \cdot \left(1 + \log\left(\frac{N}{df}\right)\right) \tag{2.1}$$

で表される．そして，各文書における索引語の重みの総和をその文書の重要度としている．

しかし，tf-idf 法には検索対象の文書が長いほど索引語の出現回数が暗に大きくなる欠点が知られている．似た内容の文書があった場合，内容が長い文書が短い文書と比較して優先的に検索されてしまう問題が生じる．

2.1.4 テキスト自動要約

要約とは，一般にあるひとまとまりのテキストが表している意味内容を非常に短いテキストで簡潔に表現することを指す [4]．計算機で要約を行う場合，人間の要約プロセスをシミュレーションすることは現在の技術では非常に困難であると言われている．そのため，元の文書の中から重要な文だけを残し，その他の部分を削除する重要文抽出による要約作成手法がある．この手法を用いることにより実用レベルに近い要約を作成することが可能であるとされている．

本論文では，重要文抽出手法を実装したテキスト簡易要約器 Posum[5] を利用してテキスト自動要約を行う．Posum は，テキスト中の単語の重要度や，単語間のつながりを利用した単語の重要度を元にする手法によって重要文抽出を実現している．オプションが多く存在し，組み合わせることで様々な重要度計算を行うことができるが，今回は，基本的な重要文抽出型の要約を用いることとする．

2.2 関連研究

2.2.1 NTCIR

NTCIR (NII Test Collection for Information Retrieval and Text Processing: エンティサイル) は，情報検索，言語横断検索，テキスト自動要約，質問応答など情報アクセスに関わるテキスト処理技術の評価ワークショップである [1].

NTCIR の目的は，大規模テストコレクションと共通の評価枠組みの提供による情報アクセスに関わる研究の発展を図る，研究アイデアの交換などをするための研究者フォーラムの構築等とされている.

情報アクセスに関わるテキスト処理技術の研究開発では，複数の異なるシステムやアルゴリズムの有用性の客観的な比較評価が必要不可欠である．テストコレクションは，これらの評価実験に用いるデータセットのことであり，文書データの集合，設問群，各設問に対する正解の 3 つからなる.

評価ワークショップは，共通のテストコレクション・研究課題・評価の基盤と意見交換の場を用意し，参加する研究グループは共通の研究課題を各々のアプローチで遂行し，成果を相互比較し，議論を深めていくという主催者と参加者が協力して研究を盛り立てていく新しい研究スタイルである．近年の情報アクセスに関わるテキスト処理技術研究では大規模文書データを使用することが多いが，大規模文書データに対して全数調査で正解を見つけることは困難である．多数の研究グループが同一課題を遂行する評価ワークショップは，正解候補をより網羅的に効率よく収集する良い機会である.

2.2.2 QAC-1

QAC-1 は，第 3 回 NTCIR ワークショップのサブタスクとして行われる質問応答に関する第 1 回評価会議である [2].

QAC-1 の目的は，膨大な文書を背景に自然言語によって尋ねられた任意の質問に答えを与えることである．ドメイン依存でないこととともに，組織化されていない情報に依存して

いることが RDB に対する質問と異なる点である．

検索対象文書として，毎日新聞の 1998, 99 年の 2 年分の新聞記事データを利用しており，回答として求められるものは，人名や組織名等の固有表現，金額や温度等の数値表現，作品名，日付け，種やカテゴリの名称等である．回答方法は以下に示すタスクによって異なる．

タスク 1

システムは，与えられた質問文に対して，その回答と考えられるものひとつを優先順位をつけて 5 つ返す．複数の回答が考えられる場合でもそのうちひとつ返すものとする．たとえば，正解が，山田と鈴木のふたつであるような質問に対して，システムは，第一候補 佐藤，第二候補 鈴木，第三候補 田中，第四候補 山田，第五候補 山本，の様に回答する．

タスク 2

システムは，与えられた質問文に対して，質問文の回答と判断されたものをすべて列挙して返すものとする．例えば，正解が，山田と鈴木のふたつであるような質問に対して，システムは，たとえば，(山田，鈴木)，(鈴木)，(佐藤，山田，鈴木) 等のいずれかを回答する．

タスク 3

連続して入力されたと想定される複数の質問文（枝問）を対象とする．後に続く枝問には，それ以前の質問文の一部もしくは回答を参照する表現を含むものとする．

著者らの研究チームも QAC-1 に参加しており，タスク 2 について研究・開発を行ってきた [6].

QAC-1 のタスク 2 で想定される質問応答の例を図 2.2 に示す．質問解析では，質問文から検索に必要な情報を取得する．文書検索では，検索対象文書の新聞記事データから質問文に最も関連のある新聞記事を検索する．そして，回答抽出では，検索された新聞記事から質問文への回答となる単語を抽出し回答する．

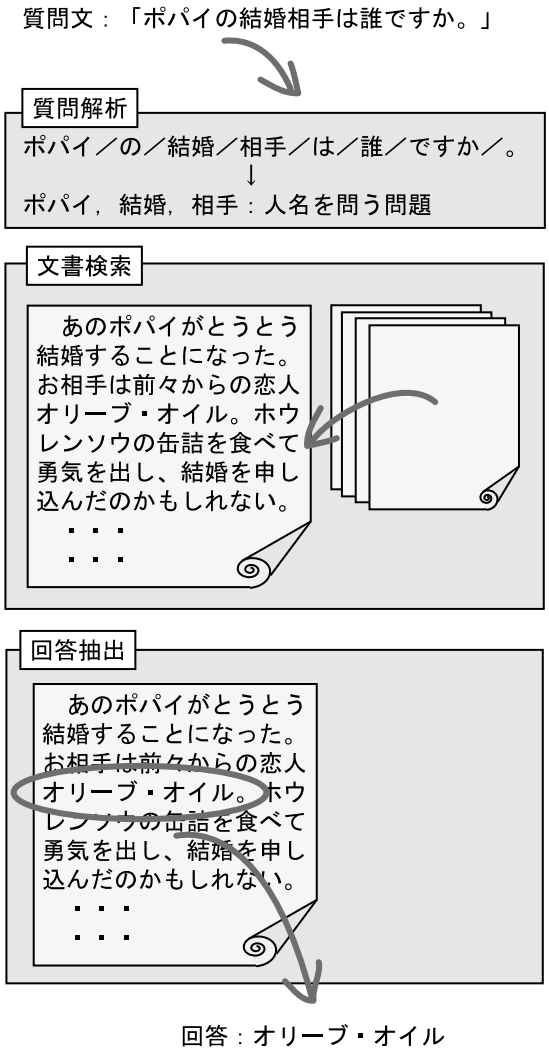

図 2.2　QAC-1 のタスク 2 で想定される質問応答の例

2.3　本研究の戦略

2.1.3 で述べたように，ユーザの質問文を形態素解析して得られた索引語に対して，tf-idf 法を用いて重み付けを行い情報検索を行う際には，検索対象のデータ長を何らかの方法で制限する必要がある．

本論文では，まず，この問題を克服する手法として，テキスト自動要約を用いた検索対象文書長の制限手法を提案する．テキスト自動要約を用いることにより，検索対象文書データは本来の意味を維持しつつを短くまとめることができる．なお，検索対象の文書は，新聞記事データを想定している．

次に，新聞記事データを検索対象文書とする文書検索手法として，新聞記事の特徴を用いた検索手法を提案する．提案する手法は，索引語に対して tf-idf 法による重み付けを行うことを基本としており，この重み付けを新聞記事の特徴を用いることで拡張したものである．質問応答技術における文書検索手法の 1 つのアプローチとして提案する．

第 3 章

テキスト自動要約による新聞記事の前処理

本章では，質問応答において検索対象の文書長を制限する手法を提案する．なお，検索対象の文書として新聞記事データを想定している．

tf-idf 法を用いて索引語に対して重み付けをし文書の重要度を求める文書検索手法では，検索対象文書の内容が長い文書が短い文書と比較して優先的に検索されてしまう問題が生じることを先に述べた．そこで，検索対象の新聞記事データに対して Posum によるテキスト自動要約を用いるアプローチにより，この問題の解決を試みる．

3.1 記事データの構造と問題点

まず，検索対象の新聞記事データの構造を明らかにする．QAC-1 では，各新聞記事の各項目ごとにタグ付けがされた新聞記事データを利用している．図 3.1 に，新聞記事データの例を示す．そして，それぞれのタグで囲まれる項目の意味を表 3.1 に示す．

図 3.1 の例を見ると，本文は 5 文で構成されていることがわかる．他の新聞記事を見ると，本文は，10 文であったり，20 文であったりと実にまちまちである．

tf-idf 法を用いて索引語に対して重み付けをし文書の重要度を求める文書検索手法では，検索対象文書の内容が長い文書が短い文書と比較して優先的に検索されてしまう問題が生じる．

```
<DOC>
<DOCNO>JA-980902017</DOCNO>
<LANG>JA</LANG>
<SECTION>２面</SECTION>
<AE>無</AE>
<WORDS>372</WORDS>
<HEADLINE>自民党、連合と“復縁”――「縁切り」もう限界？政策協議を再開へ</HEADLINE>
<DATE>1998-09-02</DATE>
<TEXT>
　自民党執行部と連合幹部との会談が１日、国会内で行われ、森喜朗幹事長は３月以降中断していた連合との政策協議を再開する考えを伝えた。連合の「民主党支持」方針に当時の自民党執行部が猛反発し、一方的に「縁切り」宣言していたが、金融再生法案の早期成立を図るうえで、民主党に影響力を持つ連合との関係修復は不可欠とみて一転、矛を収めたようだ。
　会談では連合の鷲尾悦也会長が「協議を受け入れてほしい」と求め、森氏は「協力をいただきたい」と応じる考えを表明した。村岡兼造幹事長代理は、金融再生法案をめぐる野党との修正協議について「時間との勝負であり、応援してほしい」と低姿勢に徹した。
　連合としても、政府・自民党とのパイプが切れたままでは要求が政策に反映されないため、苦慮していた。連合幹部の一人は「政策協議再開は双方のあうんの呼吸」と説明している。【桜井茂】
</TEXT>
</DOC>
```

図 3.1　新聞記事データの例

3.2　テキスト自動要約の適応

検索対象文書の内容が長い文書が短い文書と比較して優先的に検索されてしまう問題を解決するために，検索対象文書に対してテキスト自動要約技術の適応を試みる．

テキスト自動要約には，重要文抽出手法の Posum を利用することにする．Posum は，要約後に生成される文の数を指定して要約を実行することができる．たとえば，30 文や 50 文で構成される文書を，20 文等の任意の指定した文数に要約することができる．もし，指定した文数に満たない文書を要約しようとすると，要約は行われずに元の文書のままとなる．したがって，Posum によるテキスト自動要約を行う際には，要約後に生成される適切な文の数を設定する必要がある．

そこで，検索対象文書の 2 年分の全新聞記事データ 236,664 記事について各記事の本文が何文で構成されているかを調査した．表 3.2 に調査結果を，図 3.2 に全記事における本文の長さの分布を示す．なお，文の区切りには，「。(句点)」を使用した．その他の改行コード

表 3.1　新聞記事データの各項目の意味

DOCNO	記事間でユニークな ID
LANG	使用言語
SECTION	紙面情報
AE	写真，図の有無
WORDS	文字数
HEADLINE	見出し
DATE	発行年月日
TEXT	本文

等では文の区切りとしていない．

表 3.2　記事を構成する文の数

記事数	236,664
1 記事の平均文数	10.63
1 記事の最大文数	202
1 記事の最小文数	1

表 3.2，図 3.1 より，Posum による要約後の記事文数は 10 文程度にすると良いと考えられる．分布を見ると，比較的なだらかな曲線を描いており，しきい値を設けることが可能であることが解る．なお，文数が 1 である新聞記事から平均文数である 10 までの新聞記事は，全体の約 70%を占めている．また，文数が 20 文を越える新聞記事は全体の約 15%となっており，これらには要約による効果が顕著に表れるのではないかと期待される．

次に，以下の 3 点を新聞記事の特徴として考え，記事の長さの調査結果を元に表 3.3 で示す要約手法を提案する．

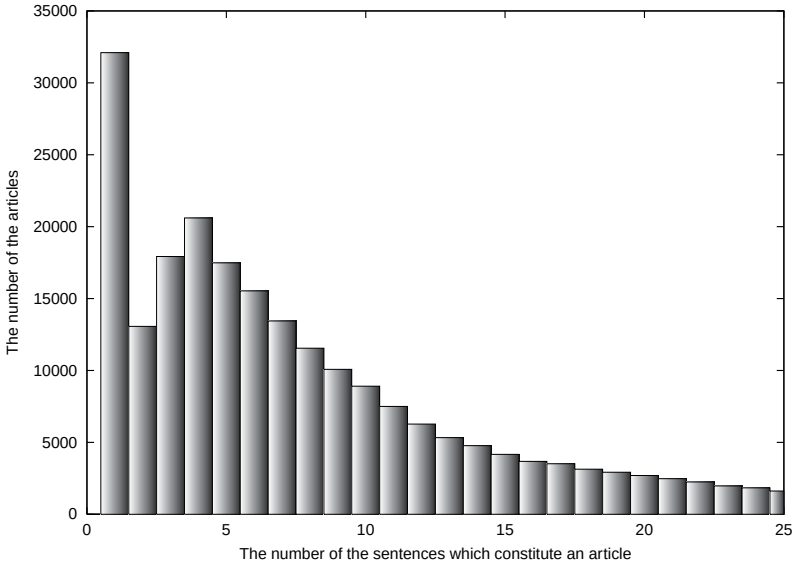

図 3.2　記事を構成する文の数

本文の 1 文目は結論が書かれることが多い

本文の 1 文目には，その新聞記事全体を短くまとめた結論が書かれることが多いように見受けられる．これは，ユーザの質問文の答えであることが多いと考えられる．

記事の先頭に近い段落ほど重要であることが多い

記事の先頭に近い段落ほど，その新聞記事の核となる話題が書かれることが多いように見受けられる．先頭から遠い段落ほど，記事の内容の背景であったり関連事項が書かれることが多いように見受けられる．

各段落の 1 文目は段落内で重要であることが多い

各段落の 1 文目には，その段落を短くまとめた結論が書かれることが多いように見受けられる．この考えは，前述の本文の 1 文目と似た考えである．しかし，本文の 1 文目と比較するとその重要度は劣るだろう．

同じ内容の記事の場合，日付が新しいものが重要であることが多い

新聞記事には，同じ内容の記事が，翌日，または 1 週間後，1 ヶ月後，と複数回掲載されることがある．前日に速報として掲載したものを翌日に詳しく記事にする場合や，過去に起きた事件・事故が解決し，それを記事にする場合などがこれにあたる．つまり，

同じ様な内容の新聞記事が複数検索された場合，より新しい記事を選択するとよいと考えられる．

表 3.3　新聞記事データの要約手法

	要約文数	要約手法
手法 0	不定	要約を行わない
手法 1	最大 10	全文が Posum の出力
手法 2	最大 10	先頭 4 文をそのまま利用し，残りの 6 文は Posum の出力
手法 3	最大 10	第 1 段落をそのまま利用し，第 2 段落以降は 各先頭の 1 文と Posum の出力 1 文の計 2 文の繰り返し

3.3　質問解析

日本語の質問文には，検索の意図が文末表現に表れる場合が多いという特徴がある．例えば，文末表現に “誰ですか” とあれば，それは人名を問う質問であるといった検索の意図が読みとれる．この文末表現のパターンを質問パターンと呼ぶことにし，検索の意図を質問タイプと呼ぶことにする．質問パターンならびに質問タイプを取得することにより，後の回答抽出の際にどの品詞を抽出すべきであるかが理解できる．

質問タイプは，あらかじめ表 3.4 に示す質問パターン辞書を用意し，優先順位に従い質問文とのマッチングを行うことにより決定する．質問文が質問パターンとマッチすると，その質問パターンに対応するタイプが質問タイプとなる．

質問文から質問タイプを取得すると，質問パターンにはそれ以上の情報が含まれない．このため，質問文から質問パターンを削除する．

次に，残った質問文の形態素解析を行う．形態素解析ツールには，計算機による日本語の解析の研究を目指す多くの研究者に共通に使える形態素解析ツールを提供するために開発された茶筌 [7] を利用する．茶筌を利用することにより，質問文が形態素に分解されるだけで

表 3.4　質問パターン

優先度	質問パターン	質問タイプ
1	でしたか	（削除）
2	ですか	（削除）
3	という名前	Who
4	名前は	Who
5	本名は	Who
6	は誰が	Who
7	誰が	Who
8	は誰	Who
9	誰	Who

なく，分解された形態素の品詞情報を得ることができる．

質問文の形態素解析を行ったとき，“「” や “」” のような括弧内の文字列が分解されてしまう問題点が出てくる可能性がある．本来，括弧内の文字列は本の名前等の固有名詞であることが多く，1 語として扱うのが自然である．このため，括弧内の文字列については，形態素解析を行わずに 1 語として扱うことにする．

最後に，文書検索を行うための索引語の絞り込みを行う．平仮名しか含まない形態素には，漢字を含む形態素に対して索引語としての価値が劣る可能性が高い．これは，平仮名のみの場合，助詞，助動詞等の文を構成するために必要な形態素である可能性が高いからである．また，記号に関しても索引語としての価値が低いと考えられる．このため，平仮名しか含まない形態素と記号に関しては質問文から削除し，残った形態素を文書検索を行うための索引語とする．

3.4　文書検索

質問解析で得られた索引語を用いて，新聞記事データを検索対象とした文書検索を行う．文書検索では，膨大な新聞記事データの中からユーザの要求を最も満たす新聞記事を 1 つ検索することを目的とする．

ある索引語の新聞記事中への出現回数を tf，全新聞記事数 N に対する索引語の出現回数を df で表す．idf を $idf = 1 + \log\left(\frac{N}{df}\right)$ と定義し，tf-idf 法による索引語のある文書における重み w を

$$w = tf \cdot idf = tf \cdot \left(1 + \log\left(\frac{N}{df}\right)\right) \tag{3.1}$$

で表す．

索引語の重みを求めたら，全新聞記事データの重要度を計算する．全新聞記事データに対して式 3.1 で求めた索引語の重みを求め，その総和により各新聞記事の重要度を求める．重要度が一番高い新聞記事を回答とする．ただし，このとき重要度が同じ新聞記事が複数ある場合，より新しい新聞記事を回答とする．

3.5　実験

テキスト自動要約を行った 3 種類の新聞記事データと，要約を行っていない新聞記事データとの検索性能を比較する．

表 3.5　要約文書に対する検索比較実験結果

	手法 0	手法 1	手法 2	手法 3
検索成功数	7/43	5/43	4/43	10/43
	(16.3%)	(11.6%)	(9.3%)	(23.3%)

検索に用いる質問文には，QAC-1 のデータセットを用いる．QAC-1 では，Formal Run のための質問文として 200 問が用意されている．このうち，本論文が対象としている人名を問う質問は 43 問であった．この 43 問を本実験に用いることとする．表 3.5 に各検索対象に対する実験結果を示す.

実験結果より，手法 1, 2 では正解数が減ったが，手法 3 を用いると検索手法はまったく同じであっても正解となる回答が増えることがわかった.

第 4 章

提案する新聞記事検索手法

本章では，新聞記事データを検索対象とする質問応答システムにおいて，膨大な新聞記事データの中から適切な答えを含む記事を検索する際に新聞記事の特徴を活用する手法を提案する．

提案する新聞記事検索手法は，索引語に対して tf-idf 法による重み付けを行うことを基本としており，この重み付けを新聞記事の特徴を用いることで拡張したものである．

図 4.1 に本検索手法の概要を示す．ユーザから自然言語で書かれた質問文を受け付けると，まず，質問解析を行う．質問解析では，質問文からユーザの検索意図である質問タイプと新聞記事検索に用いる際に必要となる索引語を取得する．次に，新聞記事検索を行う．tf-idf 法を用いて索引語に重み付けを行う際に，新聞記事の特徴による重みを付加し，最終的な新聞記事の重要度を求める．最も重要度が高かった新聞記事がユーザに回答される．

なお，検索対象の新聞記事データには，3.2 で提案した手法により作成された手法 3 のデータを使用する．このデータを用いることで，本文が長い記事が優先的に検索されてしまう問題を解決できる．

4.1 質問解析

ユーザから質問文を受け付けると，まず，質問解析を行う．質問解析では，質問文からユーザの検索意図である質問タイプと新聞記事検索に用いる際に必要となる索引語を取得する．

質問解析の方法は 3.3 で述べたものと同じであるため割愛するが，質問文が人名を問うも

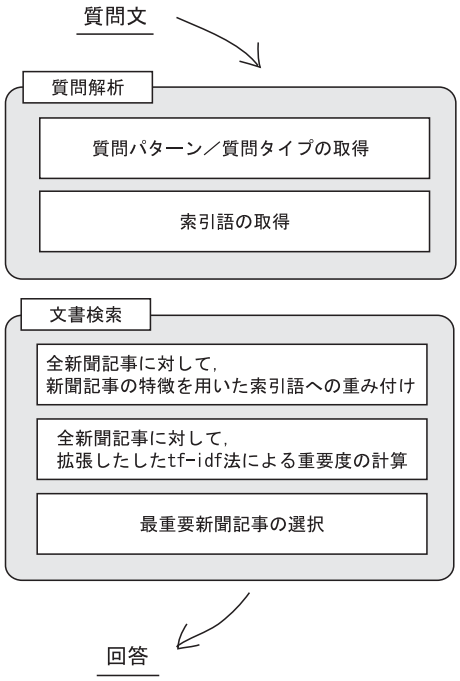

図 4.1　提案する新聞記事検索手法の概要

のであるのかどうかを判断する質問タイプの取得と，後の文書検索の際に用いる索引語の取得を行う．

4.2　文書検索手法

質問解析で得られた索引語を用いて，新聞記事データを検索対象とした文書検索を行う．文書検索では，膨大な新聞記事データの中からユーザの要求を最も満たす新聞記事を 1 つ検索することを目的とする．

提案する手法は検索対象として新聞記事データを使用し，そのデータに見られる特徴を用いて検索を行うものである．本論文では，質問パターンが人名を問う質問，つまり，質問タイプが “Who” であるものに絞ってその特徴を活用する手法について述べる．

まず，文書検索に用いる新聞記事の特徴を以下のように考えた．

本文の 1 文目は結論が書かれることが多い

本文の 1 文目には，その新聞記事全体を短くまとめた結論が書かれることが多いよう

に見受けられる．これは，ユーザの質問文の答えであることが多いと考えられる．つまり，本文の 1 文目に索引語が含まれる場合，その索引語は答えを導くために重要であると考えられる．したがって，索引語が本文の 1 文目に含まれる場合，その索引語に対して大きな重みを付加することとする．

各段落の 1 文目は段落内で重要であることが多い

各段落の 1 文目には，その段落を短くまとめた結論が書かれることが多いように見受けられる．この考えは，前述の本文の 1 文目と似た考えである．しかし，本文の 1 文目と比較するとその重要度は劣るだろう．したがって，索引語が各段落の 1 文目に含まれる場合，その索引語に対して少々重みを付加することとする．

役職や年齢が添えられた人名は重要であることが多い

新聞記事に出現する人名には，役職や年齢等が人名の後に添えられたものがある．これらの人名は，それが添えられていない人名と比較して，その新聞記事のキー・パーソンであることが多いように見受けられる．つまり，キー・パーソンが出現する文に索引語が含まれる場合，その索引語は答えを導くために重要であると考えられる．したがって，索引語が出現する文に役職や年齢等が添えられた人名が含まれる場合，その索引語に対して大きな重みを付加する．

本論文では，このような役職や年齢等が添えられた人名を優位人名と呼ぶこととする．優位人名の定義を表 4.1 に示す．文を形態素解析した結果，該当する形態素の品詞列が存在する場合，そこに含まれる人名が優位人名である．

見出しと本文両方に出現する優位人名は重要であることが多い

優位人名が，本文ばかりでなく見出しにも出現する場合，それは新聞記事の内容を表す人名であることが多いように見受けられる．したがって，索引語が本文の優位人名と同じ文に出現し，かつ，その優位人名が見出しにも出現する場合，その索引語に対して少々重みを付加することとする．

同じ内容の記事の場合，日付が新しいものが重要であることが多い

新聞記事には，同じ内容の記事が，翌日，または 1 週間後，1 ヶ月後，と複数回掲載さ

れることがある．前日に速報として掲載したものを翌日に詳しく記事にする場合や，過去に起きた事件・事故が解決し，それを記事にする場合などがこれにあたる．つまり，同じ様な内容の新聞記事が複数検索された場合，より新しい記事を選択するとよいと考えられる．

これらの新聞記事の特徴を用いて付加する重み B を求める方法を図 4.2 に示す．

次に，新聞記事の特徴を用いて拡張した tf-idf 法により，各索引語の重みを求める．ある索引語の新聞記事中への出現回数を tf，全新聞記事数 N に対する索引語の出現回数を df で表す．また，新聞記事の特徴を用いて得られた重みを B で表す．idf を $idf = 1 + \log\left(\frac{N}{df}\right)$ と定義し，拡張した tf-idf 法による索引語のある文書における重み w を

$$w = B \cdot tf \cdot idf = B \cdot tf \cdot \left(1 + \log\left(\frac{N}{df}\right)\right) \tag{4.1}$$

で表す．

索引語の重みを求めたら，全新聞記事データの重要度を計算する．全新聞記事データに対して式 4.1 で求めた索引語の重みを求め，その総和により各新聞記事の重要度を求める．ただし，価値の低い索引語の重みの加算による検索性能の低下を防ぐために以下のルールを作成した．このルールを適応し，各新聞記事に対して，動的に重要な索引語と価値の低いそれとを認識した上で記事の重要度を求めることがねらいである．

1. df 値の昇順で索引語をソート
2. (tf 値) = 0 の索引語を削除
3. 残った索引語の tf-idf 値を合計する

最後に，重要度が一番高い新聞記事を回答とする．ただし，このとき重要度が同じ新聞記事が複数ある場合，より新しい新聞記事を回答とする．

4.3　実験

提案した新聞記事データを検索対象とする文書検索手法と，一般的な tf-idf 法を用いた文書検索手法との検索性能を比較する．

検索対象の新聞記事データには，3.2 で提案した手法により作成された手法 3 のデータを使用する．

検索に用いる質問文には，QAC-1 のデータセットを用いる．QAC-1 では，Formal Run のための質問文として 200 問が用意されている．このうち，本論文が対象としている人名を問う質問は 43 問であった．この 43 問を本実験に用いることとする．

検索によって回答された新聞記事の正否判定が必要となるが，正否の判定には，同じく QAC-1 のデータセットに含まれるスコアリング・ツールを用いる．

実験結果を表 4.2 に示す．提案した手法を用いることにより，一般的な手法を用いたときと比較して 2.4 倍の検索成功数を出すことができた．

本検索手法は，一般的な検索手法と比較して高い検索成功数を出すことがわかった．また，以前に提案した検索手法 [6] と比較しても優れていることがわかった．

表 4.1 形態素解析による優位人名の定義

優位人名を含む文字列の例 （優位人名）	形態素の品詞列
山田太郎前首相（ （山田太郎）	（名詞-固有名詞-人名-姓）（名詞-固有名詞-人名-名） （接頭詞-名詞接続）（名詞-一般）（記号-括弧開）
山田太郎前首相 （山田太郎）	（名詞-固有名詞-人名-姓）（名詞-固有名詞-人名-名） （接頭詞-名詞接続）（名詞-一般）
山田太郎首相（ （山田太郎）	（名詞-固有名詞-人名-姓）（名詞-固有名詞-人名-名） （名詞-一般）（記号-括弧開）
山田太郎さん（ （山田太郎）	（名詞-固有名詞-人名-姓）（名詞-固有名詞-人名-名） （名詞-接尾-人名）（記号-括弧開）
山田太郎首相 （山田太郎）	（名詞-固有名詞-人名-姓）（名詞-固有名詞-人名-名） （名詞-一般）
山田太郎さん （山田太郎）	（名詞-固有名詞-人名-姓）（名詞-固有名詞-人名-名） （名詞-接尾-人名）
山田太郎（ （山田太郎）	（名詞-固有名詞-人名-姓）（名詞-固有名詞-人名-名） （記号-括弧開）
山田首相 （山田）	（名詞-固有名詞-人名-姓）（名詞-一般）
山田さん （山田）	（名詞-固有名詞-人名-姓）（人名-姓）
太郎首相 （太郎）	（名詞-固有名詞-人名-名）（名詞-一般）
太郎さん （太郎）	（名詞-固有名詞-人名-名）（名詞-接尾-人名）

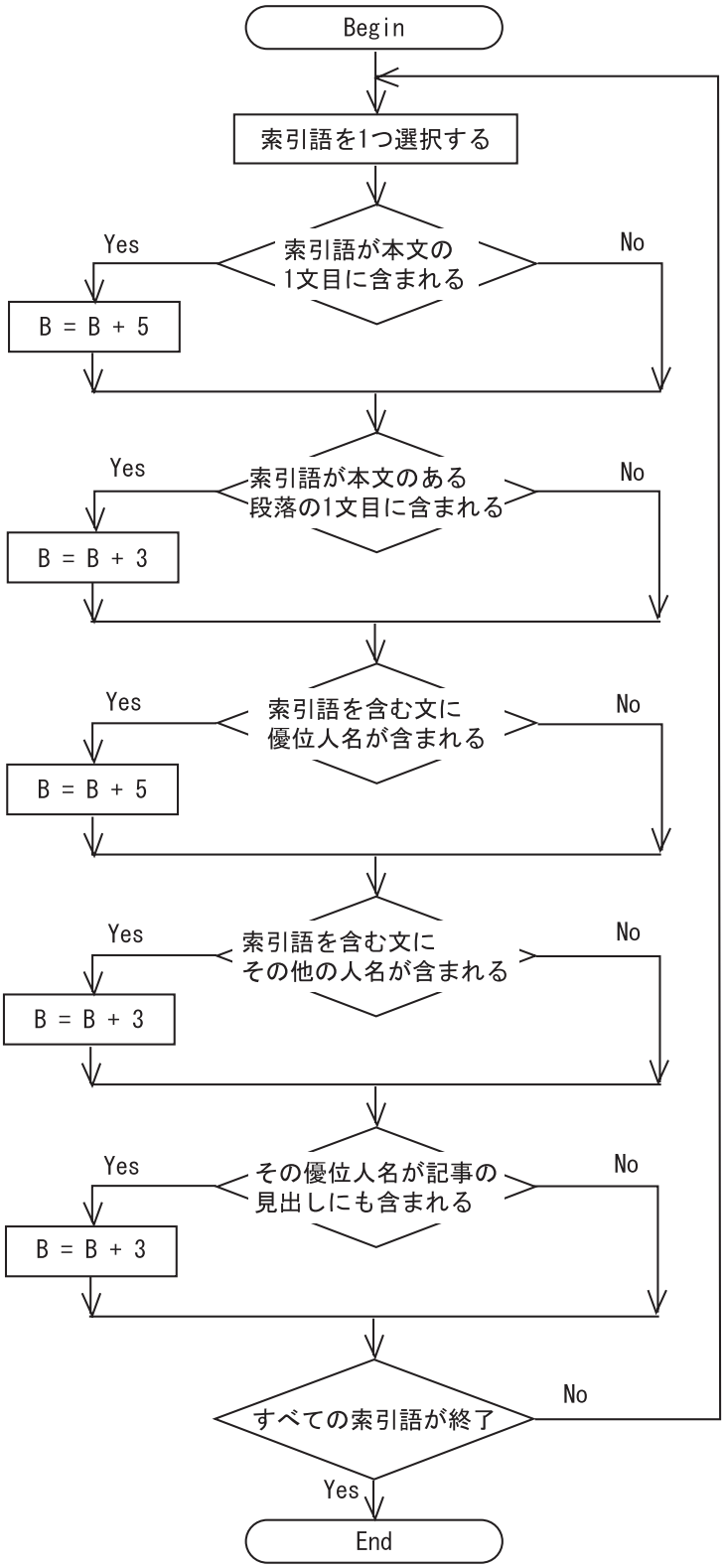

図 4.2　記事の特徴を用いた索引語への重み付けのアルゴリズム

表 4.2 提案した検索手法と一般的な検索手法の比較実験結果

	一般的な手法	提案した手法
正解数	10/43	24/43

第 5 章

考察

テキスト自動要約を用いる tf-idf 法の問題解決法の検索結果を見ると，新聞記事に対して要約を行ったことによりそれまで検索に失敗していた質問を検索成功できるようになったものが見受けられた．たとえば，図 5.1 に示すように，質問文：“自民党のニューリーダー・トリオ「ＹＫＫ」といえば誰ですか。” に対して，新聞記事に要約を行わない場合，期待される回答を含まない「CS デジタル放送」に関する記事を回答したために検索に失敗していた．原因は，38 文という非常に多い文数で構成された記事に対して “ニュー” という索引語に重みが片寄ってしまったためであった．新聞記事に要約を行った結果，この「CS デジタル放送」の記事での索引語 “ニュー” の出現回数が減少し，検索結果として選択されなくなった．そして，無事に政治に関係する正解の記事を回答することができた．これにより，テキスト自動要約を用いて tf-idf 法の問題解決ができることが示せた．

また，提案した検索手法では，新聞の特徴を取り入れたことによりそれまで検索に失敗していた質問を検索成功できたことがわかった．新聞の特徴を用いずに検索を行うと回答候補 4 位に位置づけられ検索できなかった質問において，新聞の特徴を取り入れて検索を行うと回答候補 1 位となり検索成功となった．これは，重要な索引語に対して適切に重みを付加できたことを意味している．

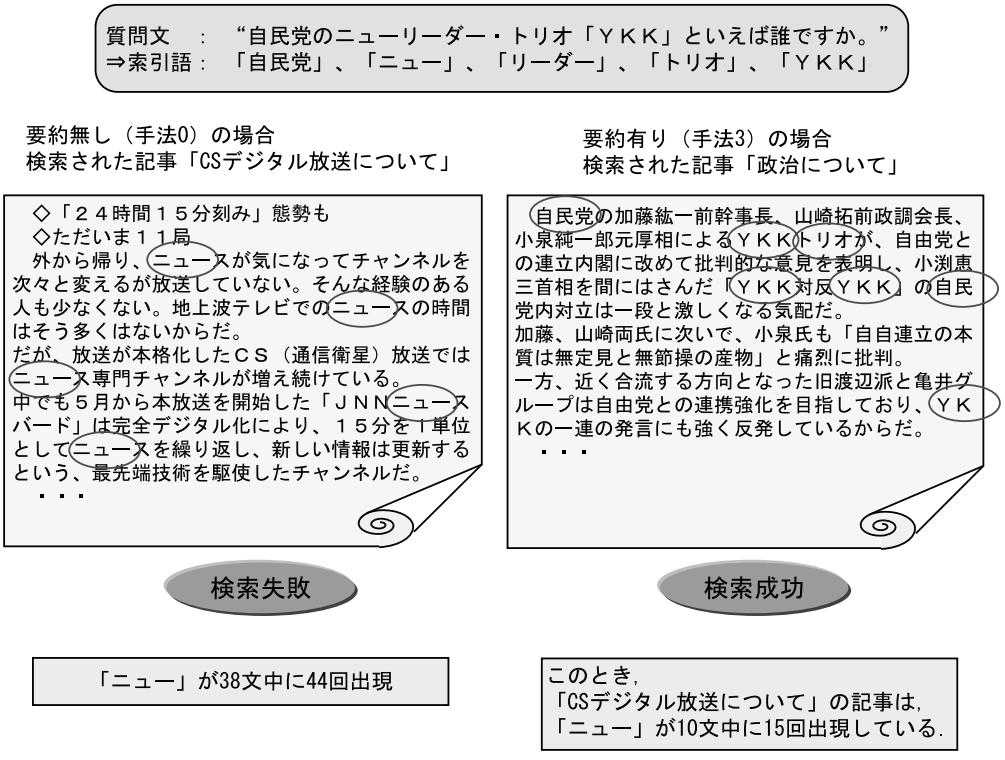

図 5.1　要約により記事検索が改善された例

第 6 章

おわりに

本論文では，質問応答の文書検索における，新聞記事の特徴を利用した検索手法を提案した．まず，新聞記事の特徴とテキスト自動要約を用いて，検索手法に tf-idf 法を用いる際に起こる検索対象文書の長さによる問題の解決法を提案し評価した．そして，記事の特徴を利用した新聞記事検索手法を提案し評価した．

テキスト自動要約の適応による tf-idf 法の検索対象文書の長さによる問題の解決法では，問題を解決することにより，記事が検索できなかった質問文から検索できるようになったばかりでなく，誤った記事検索をしていた質問文に対しても正しく検索できるようになったものが見られた．問題を解決するばかりでなく，検索成功数も向上させることができ，本手法は有用であることがわかった．本論文では，要約手法として確立した手法と言われている重要文抽出法を用いた．文書要約技術には，さらに高度な自由作成要約があり，TSC[8] 等で研究が進められている．自由作成要約等の高度な要約を適応することで，問題解決とさらなる検索成功数の向上が見込めると考えられる．

また，記事の特徴を用いた新聞記事検索手法についても，特徴を用いることにより検索が改善され，記事が検索できなかった質問文から検索できるようになったばかりでなく，誤った記事検索をしていた質問文に対しても正しく検索できるようになったものが見られた．今回用いたものに加えてさらに多くの新聞記事の特徴を用いることで，検索成功数の向上が見込めると考えられる．しかし，絶対的な実用レベルである 90%以上の検索成功率には遠く，さらなる改良が必要である．

本論文では，特に，人名を問う質問文に絞って記事の特徴を考案し検索手法に取り入れた．

質問文が問うものは，なんらかの名称もしくは値であると考えられ，人名や組織名等の固有表現，金額や温度等の数値表現，作品名，日付け，種やカテゴリの名称等が想定される．今後は，人名に関する以外の質問文に対しても回答できるように改良することが課題となる．

謝辞

本研究を進めるにあたり懇切丁寧に御指導くださいました坂本 明雄教授に心より御礼申し上げます.

突然の出来事であったにも関わらず，快く輪講に参加させていただき，また，研究活動においても貴重な御意見と適切なアドバイスをいただきました福本 昌弘助教授に心より御礼申し上げます.

遠隔地でありながら密な御指導をいただきました立命館大学理工学部情報学科の Ruck Thawonmas 助教授にも心から御礼申し上げます.

また，研究活動ばかりでなく，学生生活を送るうえで支えになってくれた同輩の登 伸一氏に感謝いたします.

さらに，研究室活動において，同輩の平山 純一郎氏，福永 諭氏，学部 4 年の河内 友彦氏，赤間 寛氏，河野 兼祐氏，西村 章氏に種々の面でお世話いただいたことに感謝致します.

最後に，本論文に対して審査してくださる坂本 明雄教授，竹田 史章教授，任 向実講師，ならびに，著者が本大学院入学時から今まで過ごしやすい環境を整えていただいた情報システム工学コースならびに情報システム工学科の諸先生方に感謝の意を表します.

参考文献

[1] 神門典子, “NTCIR とその背景 ー情報アクセス技術の評価ワークショップとテストコレクションー, ” 人工知能学会誌 Vol.17 No.3, pp.296-300, May 2002.

[2] http://www.nlp.cs.ritsumei.ac.jp/qac/

[3] 福島孝博, 奥村学, 加藤恒昭, “テキスト処理研究の動向 ー情報抽出・自動要約・質問応答における評価ワークショップの重要性ー, ” 人工知能学会誌 Vol.17 No.3, pp.301-305, May 2002.

[4] 長尾真, 自然言語処理, 岩波書店, 1996.

[5] 望月源, “テキスト簡易要約器 Posum version1.50.2 マニュアル, ” 北陸先端科学技術大学院大学情報科学研究科, 2002.

[6] Takayuki TOMOIKE, Tomohiko KAWACHI, Ruck THAWONMAS, Akio SAKAMOTO., “Article Retrieval and Answer Extraction Exploiting Characteristics in Newspaper Articles for the QAC Task2,” Working Notes of the Third NTCIR Workshop Meeting Part IV: Question Answering Challenge, pp.101-105, Oct. 2002.

[7] 松本祐治ほか, “形態素解析システム『茶筅』version 2.2.9 使用説明書, ” 奈良先端科学技術大学院大学 松本研究室, 2002.

[8] http://lr-www.pi.titech.ac.jp/tsc/

付録 A

実験に用いた質問文

表 A.1　実験に用いた質問文 その 1

質問文 ID	質問文
QAC1-2008-01	１９９８年と１９９９年の２年間に横綱に昇進した力士の名前は何ですか。
QAC1-2013-01	「Ｅ．Ｔ.」「ジュラシック・パーク」「ジョーズ」「未知との遭遇」 「シンドラーのリスト」といったら誰が監督した作品ですか。
QAC1-2018-01	福岡国際女子柔道選手権で１０連覇を達成したのは誰ですか。
QAC1-2026-01	北方領土を訪問した初めての官僚は誰ですか。
QAC1-2033-01	速水優の前の日銀総裁は誰ですか。
QAC1-2041-01	「震災文化」とは誰がつくった言葉ですか。
QAC1-2054-01	自民党のニューリーダー・トリオ「ＹＫＫ」といえば誰ですか。
QAC1-2058-01	ノーベル物理学賞を受賞した日本人は誰ですか。
QAC1-2060-01	アマゾン川をいかだで川下り中、ペルー軍兵士に殺害されたのは誰ですか。
QAC1-2063-01	源頼朝の弟は誰ですか。
QAC1-2071-01	ポパイの結婚相手は誰ですか。
QAC1-2074-01	モスバーガーを創業したのは誰ですか。
QAC1-2079-01	１９９７年に、ＩＢMの「ディープブルー」と対戦したチェスの 世界チャンピオンは誰ですか。
QAC1-2081-01	小渕恵三の前に総理大臣だった人は誰ですか。
QAC1-2085-01	菅原道真と誕生日が同じ首相は誰ですか。

表 A.2　実験に用いた質問文 その 2

質問文 ID	質問文
QAC1-2090-01	「怪談」の作者が日本に帰化する前の名前は何ですか。
QAC1-2096-01	日本神話で「天孫」とは誰のことを指しますか。
QAC1-2098-01	北野武監督の「ＨＡＮＡ－ＢＩ」で主演は誰でしたか。
QAC1-2099-01	審判にボールを投げつけた巨人の投手は誰ですか。
QAC1-2103-01	「おかあさんといっしょ」の「うたのおにいさん」として活躍し、「だんご３兄弟」のヒットを生んだ歌手は誰ですか。
QAC1-2110-01	オールスターファン投票の最終結果で一位に輝いた選手は誰ですか。
QAC1-2111-01	横綱貴乃花の本名は何ですか。
QAC1-2115-01	プロ野球選手の中で、最高年俸の選手は誰ですか。
QAC1-2122-01	「ブリキの太鼓」を代表作に持つ作家は誰ですか。
QAC1-2123-01	完全試合を達成した、米大リーグ、ニューヨークヤンキースの選手は誰と誰ですか。
QAC1-2128-01	テニスの全仏オープン女子シングルスで３年ぶりの優勝を果たしたのは誰ですか。
QAC1-2139-01	梅原猛さんと同時に文化勲章を受賞した４人は誰ですか。
QAC1-2142-01	秀吉の家臣で関ヶ原の戦いの直後、熊本城を築造した人物は誰ですか。
QAC1-2146-01	ＮＵＵの「青いドレス」を作詞した高井良斉とは誰ですか。
QAC1-2148-01	バント「ＢＩＮＧＯ　ＢＯＮＧＯ」のボーカルをしていたのは誰ですか。
QAC1-2153-01	電子楽器「テルミン」は誰が考えましたか。
QAC1-2156-01	映画「魔女の宅急便」を監督した人は誰ですか。
QAC1-2158-01	元首相でＡ級戦犯になった人物は誰ですか。
QAC1-2149-01	柔道の井上康生の父親は何という名前ですか。

表 A.3　実験に用いた質問文 その 3

質問文 ID	質問文
QAC1-2164-01	流行語大賞の「凡人・軍人・変人」とは誰のことを指していますか。
QAC1-2165-01	茶道表千家家元は誰ですか。
QAC1-2172-01	「ビビビッ！」で結婚したタレントは誰ですか。
QAC1-2174-01	国民栄誉賞を受賞した映画監督は誰ですか。
QAC1-2176-01	１９９７年の国会議員の所得で１３位だったのは誰ですか。
QAC1-2178-01	「めだかの学校」の作詞者は誰ですか。
QAC1-2188-01	エドワード王子の婚約相手は誰ですか。
QAC1-2197-01	ドラマ「ＧＴＯ」（フジ系）で教頭役を演じた俳優は誰ですか。
QAC1-2198-01	プレイステーション用ソフト「トゥームレイダー３」の主人公は誰ですか。

付録 B

3.5 の実験結果

表 B.1　3.5 の実験結果 その 1

	手法 0		手法 3	
質問 ID	回答した DOCNO	正解	回答した DOCNO	正解
QAC1-2008-01	991005028		980525121	○
QAC1-2013-01	991210285		980225160	
QAC1-2018-01	980918107		991213010	○
QAC1-2026-01	980119202		980129039	
QAC1-2033-01	990811078		980317039	
QAC1-2041-01	990205098		980322226	
QAC1-2054-01	980724195		990125013	○
QAC1-2058-01	991011152	○	991101062	
QAC1-2060-01	990306155		991129179	
QAC1-2063-01	980701331		990415289	
QAC1-2071-01	990124138		990112001	○
QAC1-2074-01	980925100		980925101	
QAC1-2079-01	991013267		991013267	
QAC1-2081-01	990811078		990824018	
QAC1-2085-01	980918107		980825060	
QAC1-2090-01	990811078		990816036	

表 B.2　3.5 の実験結果 その 2

	手法 0		手法 3	
質問 ID	回答した DOCNO	正解	回答した DOCNO	正解
QAC1-2096-01	980702150		980706006	
QAC1-2098-01	991210285		990621238	
QAC1-2099-01	990619178	○	980802150	○
QAC1-2103-01	991210286		990412212	
QAC1-2110-01	980717097		991007357	
QAC1-2111-01	980318276		980912299	
QAC1-2115-01	980217049		980706216	
QAC1-2122-01	991026082		980116255	
QAC1-2123-01	991230072		990719318	○
QAC1-2128-01	991230072	○	990808100	
QAC1-2139-01	980703344		991026178	○
QAC1-2142-01	980315167		990908188	
QAC1-2146-01	980310263	○	980310263	○
QAC1-2148-01	980820141		990220126	
QAC1-2149-01	980912030		990817125	
QAC1-2153-01	980105123		980606330	
QAC1-2156-01	991210285		980907263	
QAC1-2158-01	991103116		980614230	
QAC1-2164-01	990820208		980106236	

表 B.3　3.5 の実験結果 その 3

	手法 0		手法 3	
質問 ID	回答した DOCNO	正解	回答した DOCNO	正解
QAC1-2165-01	981001230	○	981001230	○
QAC1-2172-01	990124138		980605357	
QAC1-2174-01	991210285		981101128	
QAC1-2176-01	980630357	○	980630395	
QAC1-2178-01	981116226	○	980603379	
QAC1-2188-01	980415119		990107147	○
QAC1-2197-01	990401259		990401259	
QAC1-2198-01	980722215		991202086	

付録 C

4.3 の実験結果

表 C.1　4.3 の実験結果 その 1

	一般的な手法		提案した手法	
質問 ID	回答した DOCNO	正解	回答した DOCNO	正解
QAC1-2008-01	980525121	○	980525121	○
QAC1-2013-01	980225160		980325075	○
QAC1-2018-01	991213010	○	991213010	○
QAC1-2026-01	980129039		990819015	
QAC1-2033-01	980317039		980317039	
QAC1-2041-01	980322226		990111256	○
QAC1-2054-01	990125013	○	991029008	○
QAC1-2058-01	991101062		990220177	○
QAC1-2060-01	991129179		980105214	○
QAC1-2063-01	990415289		980926283	
QAC1-2071-01	990112001	○	990112001	○
QAC1-2074-01	980925101		981223079	○
QAC1-2079-01	991013267		990706037	○
QAC1-2081-01	990824018		980928015	○
QAC1-2085-01	980825060		990205181	○
QAC1-2090-01	990816036		991113171	

表 C.2　4.3 の実験結果 その 2

	一般的な手法		提案した手法	
質問 ID	回答した DOCNO	正解	回答した DOCNO	正解
QAC1-2096-01	980706006		980706006	
QAC1-2098-01	990621238		990202113	○
QAC1-2099-01	980802150	○	980802150	○
QAC1-2103-01	990412212		990312159	
QAC1-2110-01	991007357		980101246	
QAC1-2111-01	980912299		980912299	
QAC1-2115-01	980706216		991217099	
QAC1-2122-01	980116255		991001034	○
QAC1-2123-01	990719318	○	990219119	○
QAC1-2128-01	990808100		990704102	
QAC1-2139-01	991026178	○	991127201	
QAC1-2142-01	990908188		990627076	
QAC1-2146-01	980310263	○	980310263	○
QAC1-2148-01	990220126		990216276	○
QAC1-2149-01	990817125		990430105	
QAC1-2153-01	980606330		980606330	
QAC1-2156-01	980907263		980419068	
QAC1-2158-01	980614230		990819359	○
QAC1-2164-01	980106236		981202178	

表 C.3　4.3 の実験結果 その 3

	一般的な手法		提案した手法	
質問 ID	回答した DOCNO	正解	回答した DOCNO	正解
QAC1-2165-01	981001230	○	990107333	○
QAC1-2172-01	980605357		991231139	
QAC1-2174-01	981101128		980907303	○
QAC1-2176-01	980630395		980225211	
QAC1-2178-01	980603379		981102161	○
QAC1-2188-01	990107147	○	990107147	○
QAC1-2197-01	990401259		980919193	
QAC1-2198-01	991202086		990226107	○