平成 14 年度

学士学位論文

質問応答システムへの自動要約技術の適用

Application of Automatic Text Summarization for Question Answering System

1030260　　河内 友彦

指導教員　　坂本 明雄

2003 年 2 月 12 日

高知工科大学 情報システム工学科

要 旨

質問応答システムへの自動要約技術の適用

河内 友彦

自然言語処理は情報検索において有効な技術の１つである．質問応答は自然言語によって日常会話のような情報検索を可能とする技術であり，自然言語によって尋ねられた任意の質問に回答するものである．

本研究では，質問応答システム Prassie にテキスト簡易要約器 Posum のテキスト自動要約技術を適用することで Prassie の性能向上を試みた．まず，Prassie のアルゴリズムに回答の抽出範囲制限緩和などの改良を行った．次に，回答抽出の対象である新聞記事を改良した重要度計算を用いて要約した．これにより，索引語を含まない文からの回答抽出が可能となり，質問に対する正回答数が増加した．本研究により，質問応答に対するテキスト自動要約技術の有効性が確認できた．

キーワード　自然言語処理，質問応答，テキスト自動要約

Abstract

Application of Automatic Text Summarization for Question Answering System

Tomohiko KAWACHI

Natural language processing is one of the effective techniques in information retrieval. Question answering is a technique of information retrieval with natural language like to ask a person, and answers arbitrary questions written in natural language.

This paper proposes applying an automatic text summarization tool, called Posum, to a question answering system, called Prassie, in order to improve Prassie. First, the range of answer extraction of the Prassie was improved. Second, a newspaper article for the answer extracting was summarized with the improved way to calculate importance. As a result, the Prassie could extract right answers from a sentence not including index words for several questions, the answers of which was wrong in case of a system not making use of Posum, and the performance of extract answers was improved slightly. Therefore, it was confirmed that the automatic text summarization operate effectively to question answering.

key words natural language processing, question answering, automatic text summarization

目次

図目次

表目次

第 1 章

序論

現在，コンピュータなどを使ってインターネットやデータベースなどの大量の情報の中から欲しい情報を得るとき，欲しい情報に関連するキーワードを入力し，そのキーワードを含む情報の中から欲しい情報を探すキーワード検索などがよく使われている．しかし，このキーワード検索には問題がある．例えばどのようなキーワードを入力すれば欲しい情報が得られるのかわからないといった問題である．

その問題に対して有効と考えられるのが，自然言語処理の技術を用いた情報検索である．これによって，人にものを尋ねるように日常的な言葉で欲しい情報を得ることができるため，情報検索の知識を持たない人でも簡単に検索することができる．また，キーワード検索よりも入力される情報が多いため，それらを利用して欲しい情報をより得やすくなることが期待できる．

本研究は，質問応答システム Prassie [1] にテキスト簡易要約器 Posum [2] の自動要約技術を適用することによって，Prassie の性能向上を図るものである．具体的には，Prassie には出来るだけ手を加えず，要約を使うことによって Prassie の性能を引き出すことを考えている．

自動要約技術は現在，ある情報の内容すべてを知る前にその情報がどのような内容の情報であるかを知るといった情報検索には使われている [3]．また，質問応答の分野では，回答するために使ったテキスト中の一部を根拠として提示するための技術として利用することも考えられる．しかし，質問応答の性能を上げるための技術としては利用例がほとんどなく，その有効性は示されていない．本研究で Prassie を通して自動要約技術が質問応答の技術として有効であるかどうかを知ることが期待できる．また，有効であれば，既存の自動要約技

術を使うことによって，質問応答システムを改良，もしくは新しく構築するときの負担が軽減できると考えられる．

第 2 章

質問応答タスク QAC

質問応答とは，自然言語によって尋ねられた任意の質問に回答することである．とくに，本研究では，質問応答を質問応答タスク Question Answering Challenge (QAC) [6] のタスク定義に沿ったものとし，実験では QAC で提供された質問の中から，人名を問う質問のみ使用する．QAC はテキスト要約，情報抽出などのテキスト処理技術の研究をより発展させることを目的とした評価会議 NTCIR Workshop 3 [4][5] のタスクの一つであり，我々の研究チームはこれに参加している．ここで Prassie の位置づけを図 2.1 に示す.

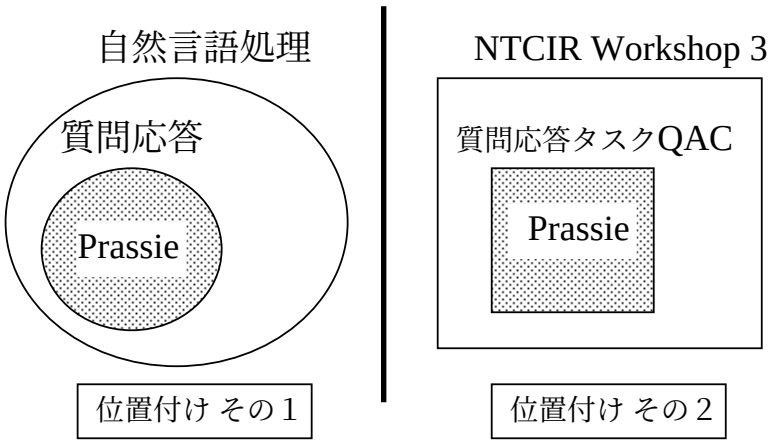

図 2.1　Prassie の位置付け

QAC の具体的な内容は，QAC によって用意された自然言語による任意の質問に対して，あらかじめ与えられた新聞 2 年分 (毎日新聞 1998-1999) のテキストデータを知識源とし，そこから回答を抽出するものである．その仕組みを Prassie の場合を例に挙げ，実際の質問例と Prassie の実際の回答と共に図 2.2 に示す.

QAC が用意した質問には人名を問うものや，日付，金額，場所を問うものなど様々な種

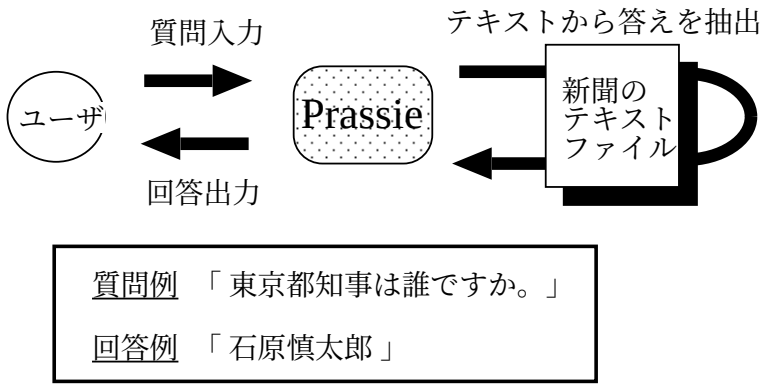

図 2.2　Prassie の質問応答の仕組みと例

類のものがあり，それに対する答えは複数ある場合やない場合もある．しかし，実験では人名を問う質問のみに対象を絞っている．

第 3 章

質問応答システム Prassie

Prassie は，自然言語で問われる質問に対して新聞記事の特徴などを利用して，新聞記事のテキストファイルから答を探し回答するシステムである [1]．Prassie は前処理，記事検索，回答抽出の 3 つの部分から構成され，基本的な流れを図 3.1 に示す．本研究の対象は回答抽出の部分である．さらに，Prassie 全体のフローチャートを図 3.2 と図 3.3 に分割して示し，以下に各部分を詳しく説明する．この Prassie のアルゴリズムを基準とし，これを Version 0.3.1.1 とする．

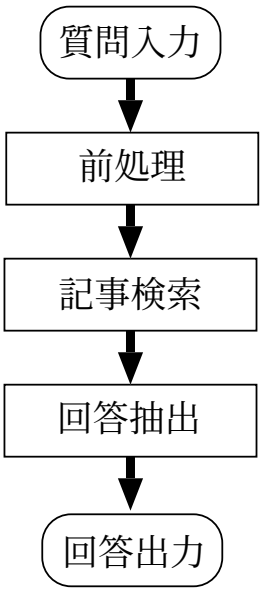

図 3.1　Prassie の基本的な流れ

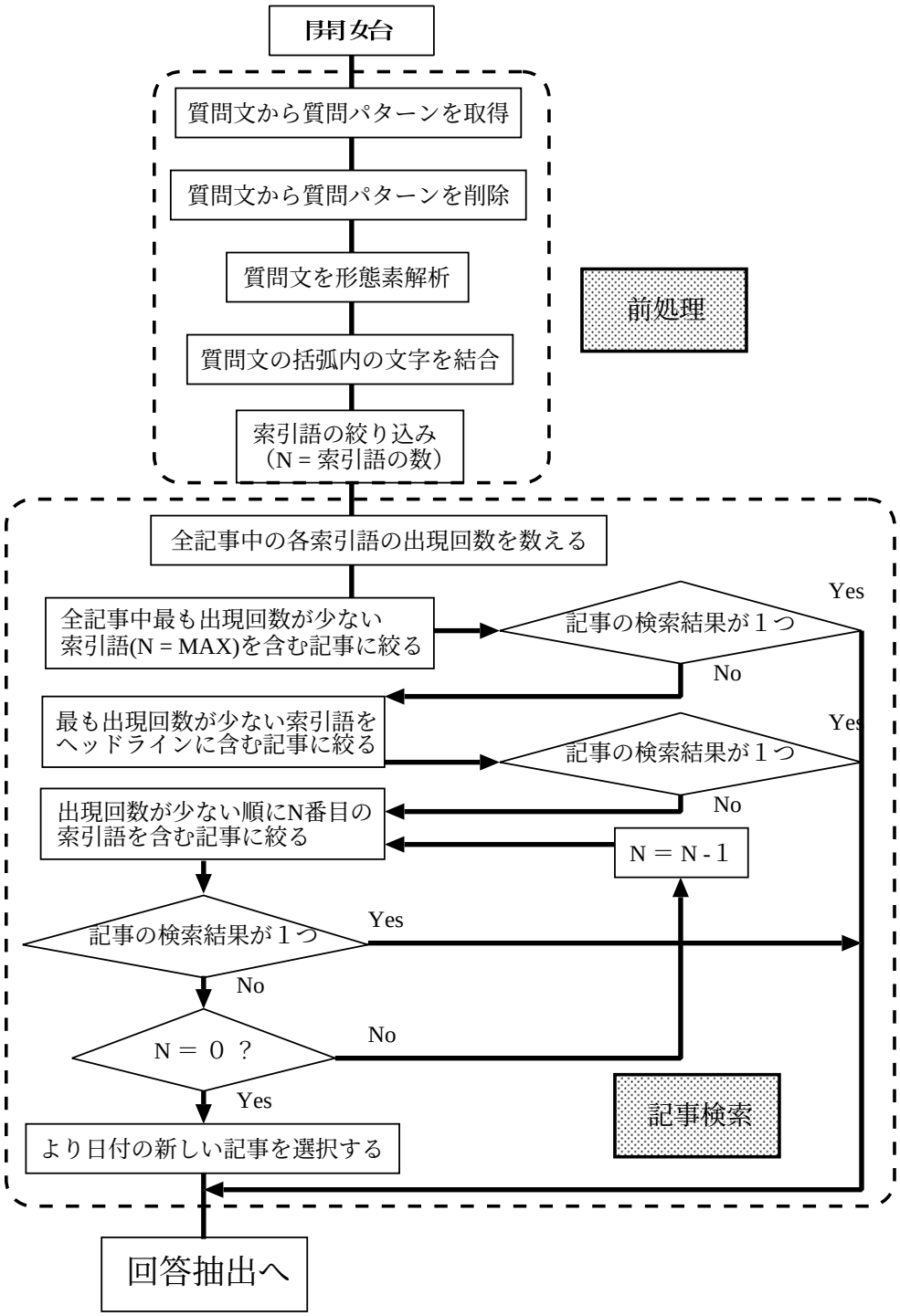

図 3.2　Prassie のフローチャート その１ (Version 0.3.1.1)

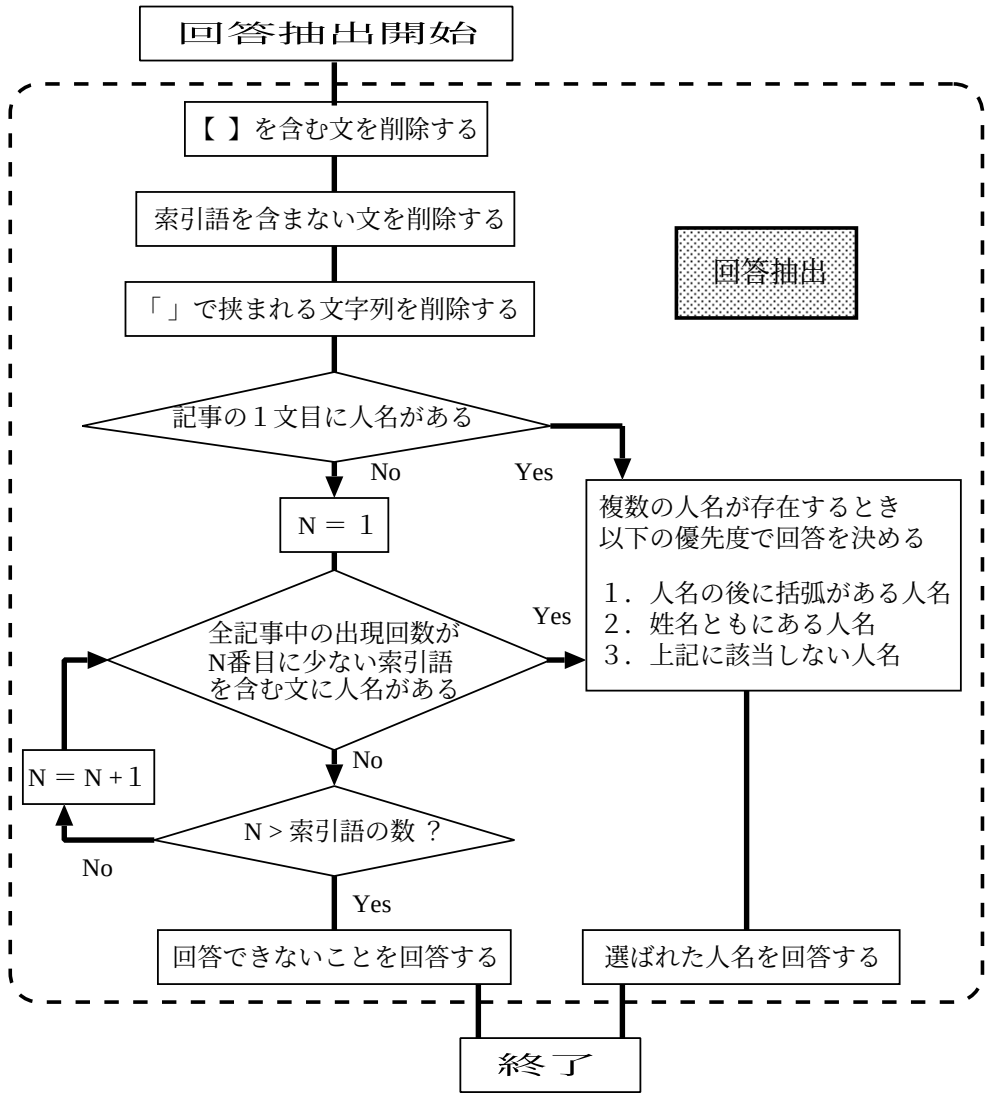

図 3.3　Prassie のフローチャート その２ (Version 0.3.1.1)

3.1　形態素解析システム ChaSen(茶筌)

ChaSen は形態素解析のツール [7] である．形態素解析とは文章を品詞ごとの単語単位に分割する技術である．Prassie, Posum 共に ChaSen を利用する．ChaSen を使った形態素解析の例を図 3.4 に示す.

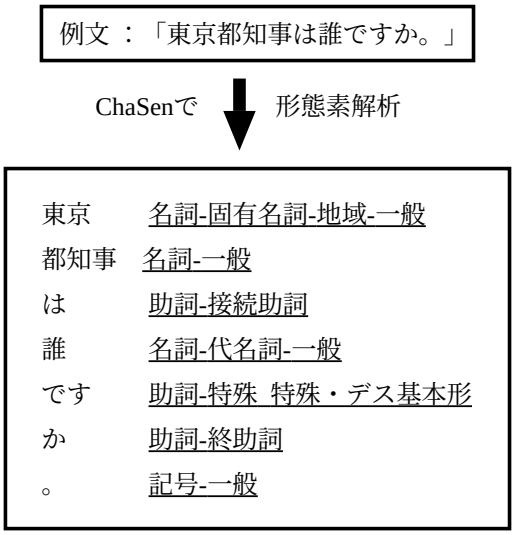

図 3.4　ChaSen による形態素解析の例

3.2　前処理

前処理では，入力された質問文が何を問う質問かを判別し，記事検索や回答抽出に役立つと考えられる索引語を取得する．

まず始めに，質問文の主に文末の表現から何を問う質問か判別する．このとき，人名を問う質問であると判断されたものが実験の対象となる．次に，質問文から何を問う質問であるか判別するのに使った文末の表現を削除し，残りの部分に対して ChaSen で形態素解析を行い索引語を探す．ただし，括弧内の単語は１つに結合する．索引語はそのほとんどが名詞である．実際の質問例と索引語の例を図 3.5 に示す．

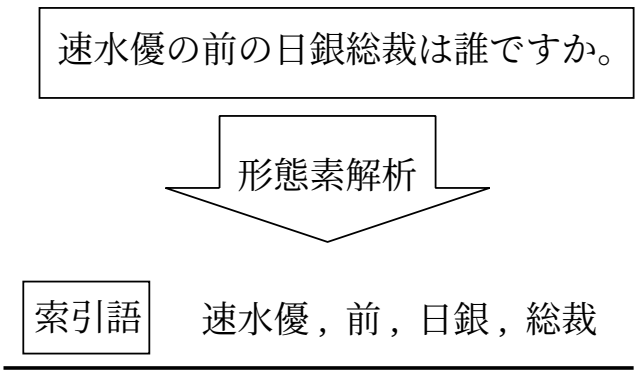

図 3.5　前処理の例

3.3　記事検索

記事検索では，各索引語と各索引語ごとの全記事中の出現回数などを使って答が含まれる記事を探し，１つの記事に絞る．

記事検索では，始めに各索引語の全記事中の出現回数を取得する．次に，この出現回数が少ないもの程記事を特定するのに役立つと考え，以下の順に記事を絞る．記事が１つに絞れた時点で終了となり，その１つの記事から回答抽出を行うこととなる．

1. 全記事の中から，最も出現回数が少ない索引語を含む記事に絞る．
2. 絞った記事から，記事のヘッドライン部分に最も出現回数が少ない索引語を含む記事に絞る．
3. 絞った記事から，次に出現回数が少ない索引語を含む記事に絞る．この時点で１つに絞れなければ，その次に出現回数が少ない索引語を含む記事という具合にすべての索引語を使って記事を絞り込む．
4. 絞った記事から，より日付の新しい記事を選択する．

3.4　回答抽出

回答抽出では，各索引語と全記事中のその出現回数，新聞記事の特徴を使って記事検索で絞った記事から答となる人名を探し，１つの人名に絞り込み回答する．Version0.3.1.1 の特徴は，必ず索引語を含む文から人名を抽出することである．まず始めに記事検索で１つに絞られた記事に以下の３つの処理を行う．

- 【 】を含む文を削除する

  【 】の中には記事を書いた人の名前が入り質問とは関係がないと考えるため．
- 索引語を含まない文を削除する

  索引語を含む文に正回答となる人名が存在すると考えるため．
- 「 」で挟まれる文字列を削除する

3.4 回答抽出

「 」の中身は正回答となる人物の発した言葉と考えられ，その人名は存在しないと考えるため．

次に，上記の処理後の記事の１文目に人名がある場合１文目から人名を抽出，回答する．１文目になければ，全記事中の出現回数が最も少ない索引語を含む文に人名がないか，記事の始めの文から順番に探す．人名があれば，その文から人名を抽出，回答する．記事の最後まで見つからなければ，次に出現回数が少ない索引語で同じ探索を行い，人名が抽出できるまですべての索引語で探す．最後まで見つからなければ回答できないことを回答する．また，１つの文に人名が複数存在する場合は以下の順の優先度で回答を決める．

1. （ ）が付いている人名
2. 姓名ともに書かれている人名
3. 上記に該当しない人名

ただし，回答抽出では質問文に含まれる人名は回答としない．

第 4 章

テキスト簡易要約器 Posum

Posum は望月によって公開されているテキスト自動要約のツールである [2]．Posum は複数のオプションにより，様々な使い方ができるが，本研究で使う範囲に関して Posum の機能を説明する．

4.1 文単位の要約

Posum は文単位での要約を行う．本研究における 1 文とは「。(句点)」によって区切られる所までを指す．要約後に残る文数は指定できる．要約によって重要度の高い文から指定文数だけ抽出されて残り，重要度の低い文は抽出されず残らない．つまり，文の一部だけが抽出，置換されたり，2 つの文が連結されるようなことはない．文が要約後に残るかどうかは，その文の重要度によって決まる．

4.2 term frequency

文の重要度の計算には term frequency(索引語頻度． 以下 tf とする．) を使う．tf とはある単語が 1 つの記事に出現する回数のことである．tf の高い単語はその記事の主題を表していると考えられ，その記事で重要な単語ということになる．例えば 1 つの記事中に，ある単語 A が 3 回出現する場合，その単語 A の重要度は 3 となる．重要度を計算する対象となる単語は ChaSen で形態素解析されたもので，名詞，動詞，形容詞の 3 種類の品詞の単語である．そして，文の重要度は，文に含まれるその対象となる単語の重要度を足したものとなる．つまり，ある文に単語 A が 2 つ含まれている場合，その文の重要度は最小でも 6 という

ことになる．

4.3　重み付け

Posum は tf とは別に，使用者が特定の単語を指定することで，その単語に特別な重みを持たせることができる．デフォルトではその指定した単語の tf を 3 倍したものがその単語の新たな重要度となる．重みは 3 倍とは限らず使用者が指定できる．

第 5 章

実験

本章では大きく 3 つの実験に分けて述べる．始めの実験は Prassie(Version 0.3.1.1) のアルゴリズムに手を加えず，Posum で基本的な要約を行った．次の実験では，Posum の要約を活かすために Prassie のアルゴリズムを改良し，基本的な要約を行った．最後に，同じく改良したアルゴリズムで，特定の単語に特別な重みを付ける要約を用いた実験を行った．各実験の詳細を述べる前にすべての実験が満たす基本事項を述べておく．

基本事項その１　実験で使う質問数は全 42 問である．

この 42 問は QAC で提供された質問のうち，以下の条件をすべて満たすものである．

1. 人名を問う質問であり，かつ Prassie が前処理部分で人名を問う種類の質問であると判断する．
2. Prassie の記事検索部分で絞り込んだ記事が正回答となる人名を含んでいる．
3. Prassie の記事検索部分で絞り込んだ記事中の正回答となる人名のうち，ChaSen で正しく抽出できるものが 1 つでも含まれる．

また，上記の条件を満たしていても，次の条件にあてはまる質問は使用していない．

- Prassie の記事検索部分で絞り込んだ記事に 1 種類の人名しか存在せず，その人名を含むいずれかの文に索引語が 1 つ以上含まれている．

なぜなら，この条件にあてはまる場合，Prassie(Version0.3.1.1) は記事検索で正しい記事を選んでいれば必ず正回答となり，要約することが性能低下にしかつながらないためである．

基本事項その２　要約のための各文の重要度計算は記事検索の直後，すなわち記事のテキス

トデータに何も処理を施していない時点で行う．

基本事項その３　要約後残す文数は 10 文以下を目標とする．

これは，２年分の新聞記事の平均文数が 10 文であること，また，実験に使う 42 問の中にも記事検索で 10 文以下の記事を選ぶ質問が 13 問も存在することを考慮し，自動要約技術を有効に働かせるための目標である．

5.1　要約導入

5.1.1　目的

この実験は，Posum で記事を要約することにより Prassie のアルゴリズムを変えることなく性能を上げることが目的である．

5.1.2　内容

まず始めに，Prassie の記事検索の部分と回答抽出の部分の間に Posum を導入し要約後に残す文数を変えて実験した．つまり，回答抽出のアルゴリズムは Version0.3.1.1 のままである．このアルゴリズムの Prassie を Version 0.4.1 とし，基本的な流れを図 5.1 に示す．

5.1.3　結果

実験結果を表 5.1 に示す．表の「＋」の列は Version0.3.1.1 のとき誤回答であったものがこの実験で正回答に変わった回答数を表し，「－」の列は Version0.3.1.1 のとき正回答であったものがこの実験で誤回答に変わった回答数に「－」を付けて表している．「合計」の行はその２つの合計値である．以降に掲載する表 5.2，表 5.3 においても同じ表記を使用する．なお，Version0.3.1.1 の正回答数は 18 である．

表 5.1 より，残す文数が少ないほど誤回答が増え，正回答数が減るという結果となった．その原因は，Prassie は索引語を含む文からしか回答抽出しないが，要約導入前に人名を抽出していた索引語を含む文が要約後残らなかったため，正回答から誤回答に変わるものは

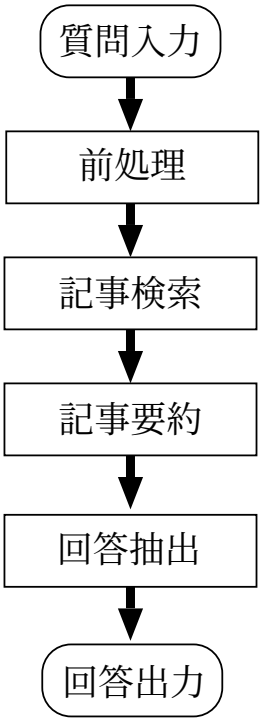

図 5.1　Version0.4.1 の基本的な流れ

あったが，誤回答から正回答になるものがなかったからである．この結果から，単に記事を要約するだけでは性能を向上させることができないとわかった．またこのとき，要約後の記事の 1 文目に，索引語を含んでいないが質問に対する正回答となる人名を含む文がいくつか見受けられたため，5.2 で要約を活かし索引語を含まない文からも回答を抽出できるよう Prassie のアルゴリズムを改良した．

5.2　Prassie の改良

5.2.1　Version0.4.1-saku

目的

5.1.3 の実験結果をふまえ，Prassie のアルゴリズムを改良し Posum を使うことで，索引語を含まない文からも正回答となる人名を抽出することで，性能を上げることが目的である．

表 5.1 Version 0.4.1 の実験結果

要約後残す文数	Version0.3.1.1 と比較した正回答数の増減		
	＋	－	合計
1 文	1	-13	-12
2 文	1	-9	-8
3 文	0	-8	-8
4 文	0	-5	-5
5 文	0	-4	-4
6 文	0	-3	-3
7 文	0	-3	-3
8 文	0	-2	-2
9 文	0	-2	-2
10 文	0	-1	-1
11 文	0	-1	-1
12 文	0	-1	-1

内容

Version0.3.1.1 の回答抽出のアルゴリズムから「索引語を含まない文を削除する」という部分を取り除いた．これに要約を加えたものを Version0.4.1-saku とし，要約後に残す文数を変えて実験した．改良した回答抽出部分のアルゴリズムを図 5.2 に示す．

結果

実験結果を表 5.2 に示す．

上述のアルゴリズムの改良によって，要約後に残った記事の 1 文目に限り，索引語を含ん

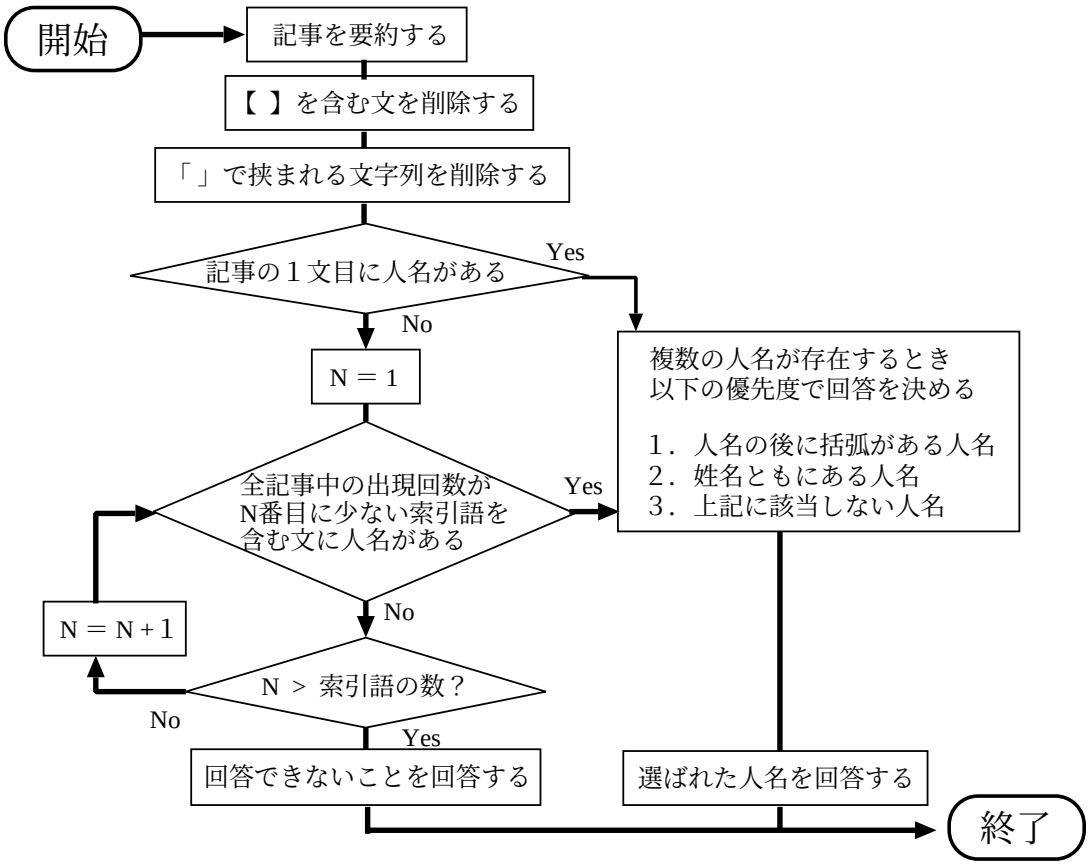

図 5.2　Version0.4.1-saku の回答抽出部分のアルゴリズム

でいなくても回答抽出が可能となり，要約後に残す文数によっては Version0.3.1.1 より正回答数が増えるものもみられた．しかし，その改良によって逆に既存の正回答を誤回答にしてしまう質問が複数みられた点が Version0.4.1 と同じく問題点である．例えば 6 文に要約したときの結果は Version0.3.1.1 では誤回答であったものを 5 つ正回答に変えたものの，既存の正回答を 4 つ誤回答にしてしまった結果，正回答数が 1 増加したということである．そのため，既存の正回答を維持することを考えなければならない．また，Version0.4.1-saku でも要約後の記事の 1 文目以外の索引語を含まない文からは回答抽出を試みない．つまり，要約を行ってもまだ正回答を抽出することにつながらない文が多量に含まれていることになる．そこで，要約後の記事に回答抽出に有効な文が多く含まれるように 5.2.2 のような改良を行った．

表 5.2 Version0.4.1-saku の実験結果

要約後残す文数	Version0.3.1.1 と比較した正回答数の増減		
	＋	－	合計
1 文	4	-12	-8
2 文	3	-8	-5
3 文	4	-8	-4
4 文	3	-6	-3
5 文	4	-5	-1
6 文	5	-4	+1
7 文	4	-5	-1
8 文	4	-3	+1
9 文	3	-3	0
10 文	3	-3	0
11 文	3	-3	0
12 文	1	-3	-2

考察

この実験では，Version0.3.1.1 に対して「索引語を含まない文を削除する」というアルゴリズムを取り除き，記事は要約するという 2 つの改良を加えている．そのため実験結果が要約によるものなのか判断しにくい．そこで確認のために，Version0.3.1.1 から「索引語を含まない文を削除する」というアルゴリズムを取り除き，要約は行わず実行した．その結果，正回答数は Version0.3.1.1 より 3 つ少ない 15 となった．これにより要約を加えることの有効性が確認された．

5.2.2　Version0.10.0, Version0.10.5

目的

要約後の記事中に正回答となる人名を含む文が残るよう，また，その割合が多くなるように回答抽出のアルゴリズムを改良し性能を上げることが目的である．

内容

以下の Version0.10.0, Version0.10.5 を提案し，要約後に残す文数を変えて実験した．ただし，42 の記事において人名を含む文の平均文数が約９文であり，９文以下の記事も複数あるため９文まで実験した．また，Version0.10.5 については，回答内容が変化しないため５文までしか実験していない．

- Version 0.10.0

  記事検索で１つに絞られた記事から人名を含む文だけを残し，残りの文は削除する．その残った文に対して要約を行い回答抽出を行う．このアルゴリズムを図 5.3 に示す．
- Version 0.10.5

  記事検索で１つに絞られた記事から人名を含む文だけを残し，残りの文は削除する．その残った文に対し，要約のための重要度が高い順に文を並び変えてから要約し，回答抽出を行う．このアルゴリズムを図 5.4 に示す．

結果

実験結果を表 5.3 に示す．

表のとおり Version0.10.0 において３文または６文以上に要約した場合が最も性能がよく，Version0.3.1.1 より正回答数が２つ増加し 20 となった．一方，Version0.10.5 は重要度の高い順に並び変えているため，要約後に残す文数を変えても各文の前後関係に変化がなく

5.2　Prassie の改良

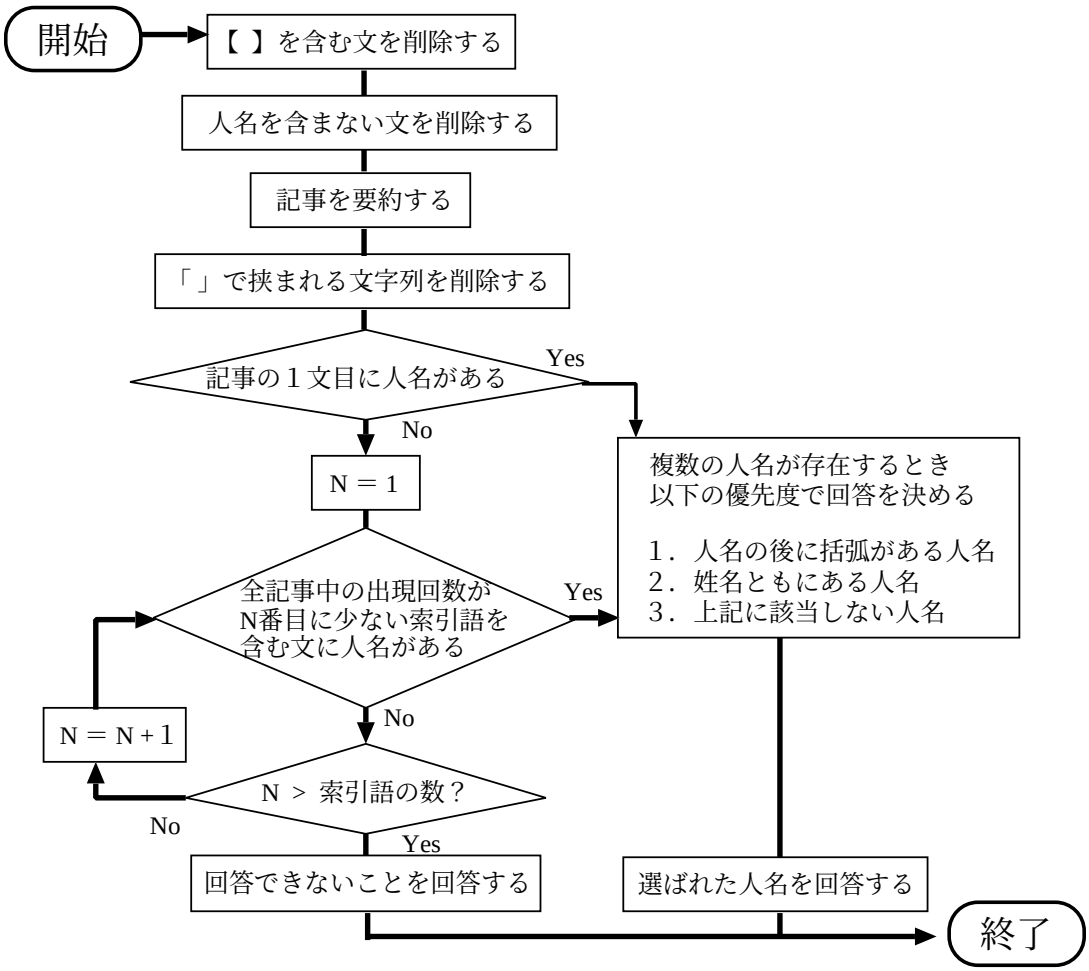

図 5.3　Version0.10.0 の回答抽出部分のアルゴリズム

回答内容が変化しなくなった．この結果からわかることは，要約後に正回答となる人名を含む文を残すことは可能であるが，要約のための文の重要度が高いほど正回答となる人名を含むとは限らないということである．それはつまり，正回答となる人名を含む文の重要度が特に高くなるような重要度計算ができていないということである．

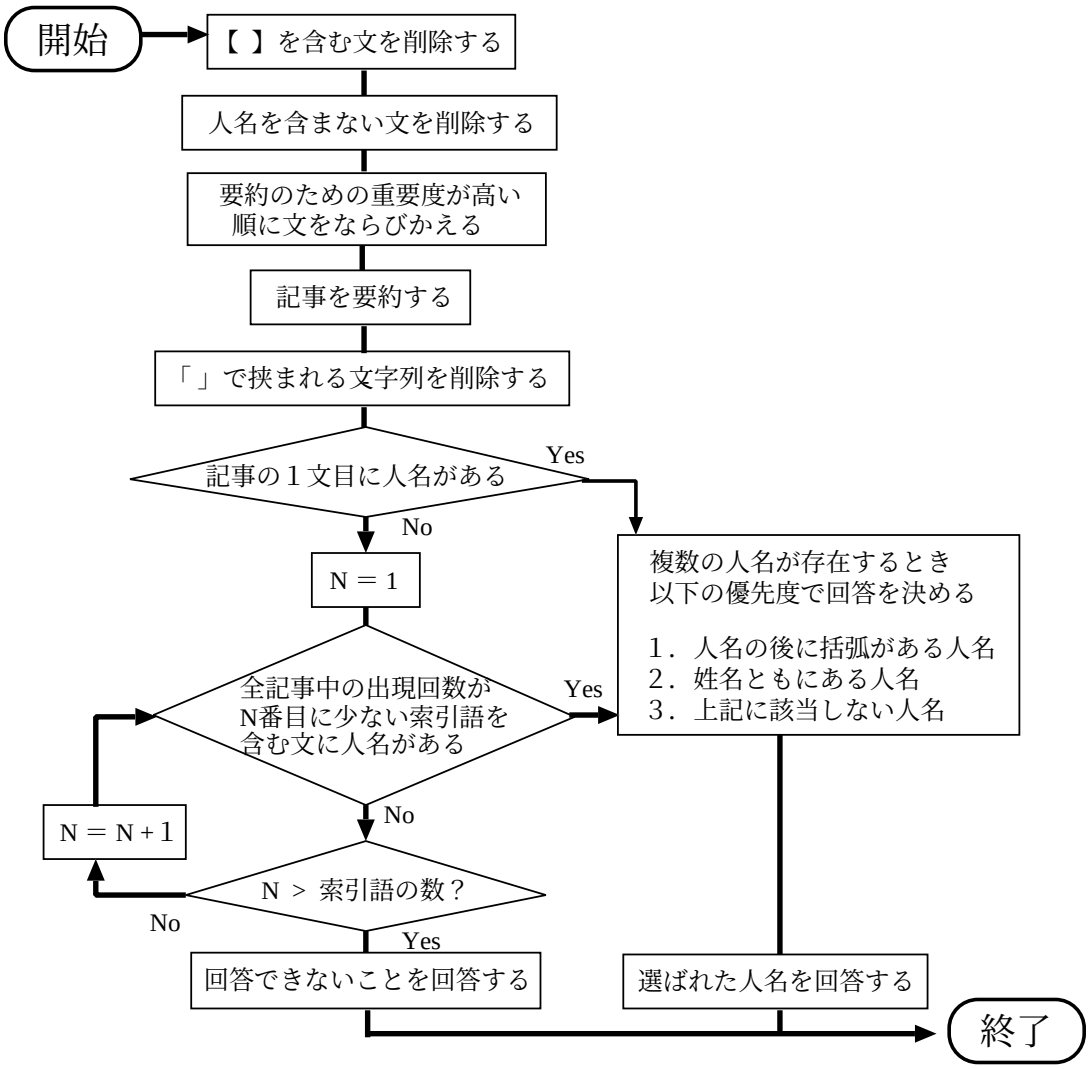

図 5.4　Version0.10.5 の回答抽出部分のアルゴリズム

5.3　重み付き要約

5.3.1　目的

特定の単語に重みを付ける重要度の計算を行い，より回答抽出に有効な重要度に修正し性能を向上させることが目的である．

表 5.3　Version0.10.0, Version0.10.5 の実験結果

要約後残す文数	Version0.3.1.1 と比較した正回答数の増減					
	Version0.10.0			Version0.10.5		
	＋	−	合計	＋	−	合計
1 文	5	-7	-2	5	-7	-2
2 文	6	-7	-1	5	-7	-2
3 文	8	-6	+2	5	-7	-2
4 文	6	-6	0	5	-7	-2
5 文	6	-5	+1	5	-7	-2
6 文	6	-4	+2			
7 文	6	-4	+2			
8 文	6	-4	+2			
9 文	6	-4	+2			

5.3.2　内容

5.2.2 の結果をふまえ，Version0.10.0 において要約後に残す文数を 3 文，6 文にするものに対し，さらに要約のための重要度を求めるとき，以下の 4 種類の単語に特別な重みを付け実験した．特別な重みはそれぞれ 2 倍から 8 倍まで変化させた．

1. 質問文に含まれるすべての索引語
2. 記事に 1 番多く出現する人名
3. 「(」「)」
4. 1 番出現回数が少ない索引語

5.3.3　結果

要約後に 3 文残す場合の実験結果を表 5.4 に，6 文残す場合の実験結果を表 5.5 に示す．表中の値は Version0.10.0 の正回答数と比較したときの正回答数の増減である．なお，Version0.10.0 (3 文残す場合，6 文残す場合共に) の正回答数は 20 である．表中の重みを付ける単語の表記の意味は以下のとおり対応している．また，重みは例えば「× 2」の場合 4.3 で述べたとおり，単語の重要度が tf から 2 倍の tf に変わることを意味する．

saku　質問文に含まれるすべての索引語

most　記事に 1 番多く出現する人名

kakko　「(」「)」

least　1 番出現回数が少ない索引語

表 5.4　Version0.10.0(3 文残す) と比較した正回答数の増減

重み	重みを付ける単語			
	saku	most	kakko	least
×2	+1	+1	0	0
×3	+1	+1	0	0
×4	+1	+1	0	0
×5	+1	+1	0	0
×6	+1	+1	0	0
×7	+1	+2	0	0
×8	+1	+2	0	0

記事に 1 番多く出現する人名に 7 または，8 倍の重みを付け 3 文に要約したとき，Version0.10.0 と比べ正回答数が 2 増え 22 となった．この正回答数の内容は，Version0.10.0 の正回答をすべて維持し，新たに 2 つの正回答ができたものであった．そのうち 1 つは，正回

表 5.5　Version0.10.0(6 文残す) と比較した正回答数の増減

重み	重みを付ける単語			
	saku	most	kakko	least
×2	0	+1	0	0
×3	0	-1	0	0
×4	+1	0	0	0
×5	+1	0	0	0
×6	+1	0	0	0
×7	+1	0	0	0
×8	+1	0	0	0

答となる人名が記事中最も出現回数の多い人名であった．この重み付けは記事検索で，正回答となる人名を主題とする記事を検索できていれば有効に働くと考えられる．Version0.10.0 の正回答を維持できたのは，出現回数が多い人名なので複数の文に含まれており，複数の文の重要度が平行して高くなり，重みを付けても文同士の重要度の大小関係にあまり影響がなかったためである．

質問文に含まれる索引語すべてに重みを付けた場合は，複数の場合で 1 つ正回答数が増え 21 となったが，内容は Version0.10.0 と比べ新しい 2 つの正回答と 1 つの誤回答ができたというものである．この新しい 2 つの正回答は Version0.3.1.1 では正回答であったものであり，新しい誤回答は Version0.3.1.1 では誤回答で，要約を使うことで正回答に変わっていたものである．つまり，この重み付けは索引語を重要視し，要約を使っていない Version0.3.1.1 の正回答を維持するのに有効な重み付けであるといえる．

そして，残りの 2 つの重み付けはどの場合も正回答数は Version0.10.0 と変わらず，正回答となる質問もまったく同じものであった．単純に出現回数が少ないことが原因である．もともと少ない出現回数のものに少し重みを加えても影響がほとんどないということである．しかし，過剰な重みを付けると，単語の重みがその単語を除いた文の重要度より高くなるこ

ともある．そうなった場合，重み付けというより明らかにその単語を含む文から回答を探す索引語と同じような存在になってしまう．そのため，この実験のように１桁倍程度の重み付けが妥当と考え，この２つの重み付けは正回答となる人名を含む文の重要度を高めるのに有効ではないと考える．また，「(」「)」に関しては記事に複数存在する時もあるが，複数存在するときは複数の文に分散していて，１文だけに集中することはほとんどない．

第 6 章

結論

本研究では，質問応答システム Prassie に，テキスト簡易要約器 Posum の自動要約技術を適用し，その性能向上を試みた．

その結果 Prassie Version0.3.1.1 の回答抽出のアルゴリズムから「索引語を含まない文を削除する」という部分を取り除き，記事検索で選ばれた記事から人名を含まない文を削除した後残った文に対し，記事に 1 番多く出現する人名に 7 倍，または 8 倍の重みを付けて要約の重要度計算を行い，要約後の文数が 3 文になるように要約することで正回答数を要約導入前の 18 から 22 へ 4 つ増やすことができた．このことから，自動要約技術は細かく調整して質問応答システムに導入すれば，性能を向上させることが可能であることがわかった．

また，要約導入前は索引後と新聞記事の特徴という二つの手がかりから回答抽出を行っていたのに対し，記事の索引語以外の言葉も使う重要度計算による要約を加えることによって，索引語を含まない文からも回答抽出ができるようになった．

結果として大幅な性能向上はみられなかったが，実験とは別の単語に特別な重みをつけるなど重要度の求め方を工夫すればさらに性能向上が望めるものと考えている．記事や質問に合わせたより有効な重要度の計算方法を考案し，既存の正回答を維持しつつ，性能を向上させることが今後の課題である．具体的には要約によって正回答になる質問，記事，誤回答になる質問，記事に共通な特徴があれば，記事または質問ごとに要約する文数を変えたり，あるいは要約を行わないといったことが考えられる．

一方，正解の人名を含む文から人名を抽出するとき，その文に複数の人名が存在した場合の回答抽出の成否は Posum に依存しないため，さらなる性能向上のためには要約だけでなく，Prassie の 1 つの文に複数の人名が存在するときの回答抽出のアルゴリズムを改良する

ことも必要であると考える.

謝辞

この論文を書くまでの長い間，それ以前からの研究活動，研究室での活動など様々な場面での丁寧な御指導をして下さった，坂本明雄先生，ラックターウォンマット先生に大変感謝しています（教授や助教授といった言葉は他人行儀なので使いません）．僕の質問攻めにずっと耐えてきた，一番お世話になった友池さんありがとうございます．友池さんの次に僕の質問攻めにあいながらも，いろいろ面倒をみてくれた登さんありがとうございます．折橋さんの「手伝えることがあったら言ってね」という言葉には感動しました，４年生のことを心配してくれてありがとうございます．赤間君，河野君，西村君，私生活ではみんなと一緒にエンジョイできなかったのが残念ですが，研究室活動はみんなのおかげでがんばれました．就職してもがんばって下さい．坂本研の３年生のみなさん頼りない先輩で申し訳なかったです．その分自分でがんばれる力が付いたことでしょう，これからもがんばって下さい．福本先生と福本研のみなさん，壁の向うから毎日楽しい話題提供ありがとうございました．それを聞きながら密かに笑ってました．

４年間お世話になった情報システム工学科の先生方，学生の皆さん，高知工科大学関係者の皆さん，そして土佐山田町，本当にありがとうございました．

参考文献

[1] 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, 2002.

[2] テキスト簡易要約器 Posum Home Page :
http://www.tufs.ac.jp/ts/personal/motizuki/software/posumcl/index.html

[3] NewsInEssence Home Page :
http://www.newsinessence.com/

[4] NTCIR Home Page :
http://research.nii.ac.jp/ntcir/index-ja.html

[5] NTCIR Workshop 3 Home Page :
http://research.nii.ac.jp/ntcir/workshop/index-ja.html

[6] QAC Home Page :
http://www.nlp.cs.ritsumei.ac.jp/qac/

[7] ChaSen Home Page :
http://chasen.aist-nara.ac.jp/index.html.ja

付録 A

質問集

実験に使った質問 42 問である．質問文以外の数字や記号は QAC のフォーマットである．[QAC1-2008-01] などと書かれている数字の部分が各質問固有の番号である．

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。”
QAC1-2018-01: ”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。”
QAC1-2033-01: ”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。”
QAC1-2058-01: ”ノーベル物理学賞を受賞した日本人は誰ですか。”
QAC1-2074-01: ”モスバーガーを創業したのは誰ですか。”
QAC1-2099-01: ”審判にボールを投げつけた巨人の投手は誰ですか。”
QAC1-2123-01: ”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。”
QAC1-2146-01: ”ＮＵＵの「青いドレス」を作詞した高井良斉とは誰ですか。”
QAC1-2172-01: ”「ビビビッ！」で結婚したタレントは誰ですか。”
QAC1-2178-01: ”「めだかの学校」の作詞者は誰ですか。”
QAC1-20021-01: ”「平成おじさん」とは誰のことを指しますか。”
QAC1-20037-01: ”禅宗の黄檗宗の開祖は誰ですか。”
QAC1-20039-01: ”「東風吹かば匂ひおこせよ梅の花主なしとて春な忘れそ」で有名な人は誰ですか。”
QAC1-20055-01: ”ペルーで憲法違反をしてまで再選した大統領は誰ですか。”
QAC1-20085-01: ”「淡路夢舞台」の設計を手がけた建築家は誰ですか。”
QAC1-20086-01: ”「梅田スカイビル」を設計した建築家は誰ですか。”
QAC1-20121-01: ”現在までで最も年齢の高い宇宙飛行士は誰ですか。”
QAC1-20123-01: ”童謡「めだかの学校」を作詞したのは誰ですか。”

QAC1-20126-01: ”長野五輪スキー・ジャンプのラージヒルで銅メダルを取った日本人は誰ですか。”

QAC1-20142-01: ”世界長者番付の第１位は誰ですか。”

QAC1-20143-01: ”東ティモールの紛争問題で、ノーベル平和賞受賞者は誰と誰ですか。”

QAC1-20202-01: ”国際オリンピック委員会（ＩＯＣ）の「オリンピック・レビュー」にて、２０世紀最高のスポーツ選手と評された日本人は誰ですか。”

QAC1-20205-01: ”信楽鉄道事故で罪が問われ、最終弁論で無罪を主張したのは誰ですか。”

QAC1-20330-01: ”ノーベル文学賞を受賞したことのある日本人は、川端康成ともう一人誰ですか。”

QAC1-20336-01: ”最年少で世界７大陸の最高峰を制覇したのは誰ですか。”

QAC1-20343-01: ”経済企画庁長官になった作家は誰ですか。”

QAC1-20345-01: ”「『ＮＯ』と言える日本」の著者は誰ですか。”

QAC1-20359-01: ”テレビドラマ「古畑任三郎」で西園寺守刑事を演じているのは誰ですか。”

QAC1-20386-01: ”誰の提唱で、国際宇宙ステーションを建設していますか。”

QAC1-20389-01: ”米国ハワイ州出身の横綱といえば誰ですか。”

QAC1-20422-01: ”閣僚として初めて北方領土を訪れたのは誰ですか。”

QAC1-20443-01: ”「マトリョーシカ」という芝居に出演しているのは誰ですか。”

QAC1-20456-01: ”夏目漱石の長男は誰ですか。”

QAC1-20633-01: ”９９年の中国の首相は誰でしたか。”

QAC1-20638-01: ”「ハリー・ポッターと賢者の石」を翻訳したのは誰ですか。”

QAC1-20639-01: ”坂本龍一と矢野顕子の娘の名前は何ですか。”

QAC1-20649-01: ”坂本美雨の両親は誰ですか。”

QAC1-20663-01: ”北朝鮮の総書記は誰でしたか。”

QAC1-20708-01: ”「不思議の国のアリス」の挿絵を描いたのは誰ですか。”

QAC1-20710-01: ”長野五輪の開会式で合唱を指揮したのは誰ですか。”

QAC1-20737-01: ”第４８期王将戦で羽生善治と対戦したのは誰ですか。”

付録 B

解答集

質問 42 問に対する答えである．答えが複数あるものもあり，また，同じ人物を表していても姓と名ともにあるもの，片方しかないものなどあるが，どれを回答しても正回答となる．

QAC1-2008-01 武蔵丸 若乃花

QAC1-2018-01 田村亮子 ヤワラ

QAC1-2033-01 松下康雄 松下

QAC1-2041-01 小林郁雄 小林

QAC1-2058-01 江崎玲於奈

QAC1-2074-01 桜田慧 桜田

QAC1-2099-01 バルビーノ・ガルベス ガルベス

QAC1-2123-01 デービッド・ウェルズ ウェルズ

QAC1-2146-01 秋元康

QAC1-2172-01 松田聖子

QAC1-2178-01 茶木滋 ちゃきしげる 茶木

QAC1-20021-01 小渕恵三 小渕

QAC1-20037-01 隠元

QAC1-20039-01 菅原道真

QAC1-20055-01 フジモリ アルベルト・フジモリ

QAC1-20085-01 安藤忠雄 安藤

QAC1-20086-01 原広司

QAC1-20121-01 ジョン・グレン グレン

QAC1-20123-01 茶木滋 ちゃきしげる 茶木

QAC1-20126-01 原田 原田雅彦

QAC1-20142-01 ビル・ゲイツ ゲイツ

QAC1-20143-01 カルロス・ベロ ジョゼ・ラモス・ホルタ

QAC1-20202-01 加藤沢男 加藤 山下泰裕 山下

QAC1-20205-01 里西孝三 里西 山本長生 山本 八木沢守 八木沢

QAC1-20330-01 大江健三郎

QAC1-20336-01 野口健 野口

QAC1-20343-01 堺屋太一 堺屋

QAC1-20345-01 盛田昭夫 盛田 石原慎太郎 石原

QAC1-20359-01 石井正則 石井

QAC1-20386-01 レーガン

QAC1-20389-01 武蔵丸 曙

QAC1-20422-01 鈴木 鈴木宗男

QAC1-20443-01 松本幸四郎 幸四郎 市川染五郎 松本紀保 染五郎 紀保

QAC1-20456-01 夏目純一

QAC1-20633-01 江沢民

QAC1-20638-01 松岡佑子

QAC1-20639-01 坂本美雨 美雨

QAC1-20649-01 坂本龍一 矢野顕子

QAC1-20663-01 金正日 キムジョンイル 金

QAC1-20708-01 ジョン・テニエル テニエル

QAC1-20710-01 小沢征爾 小沢

QAC1-20737-01 森下卓 森下

付録 C

回答集

「Version0.3.1.1」の列は Prassie Version0.3.1.1 の回答内容である．

「Version0.10.0+重み」の列は Prassie Version0.10.0 で要約のための重要度計算のとき，記事に 1 番多く出現する人名に 7 倍の重みを付け，残す文数を 3 文に要約したときの回答内容である．

●は正回答であることを表す．

表 C.1　回答集その 1

	Version0.3.1.1	Version0.10.0+重み
QAC1-2008-01	武蔵丸 ●	武蔵丸 ●
QAC1-2018-01	田村亮子 ●	楢崎教子
QAC1-2033-01	松下康雄 ●	松下康雄 ●
QAC1-2041-01		小林郁雄●
QAC1-2058-01	福井謙	福井謙
QAC1-2074-01	桜田慧 ●	杉村春子
QAC1-2099-01	高原須美子	高原須美子
QAC1-2123-01	ウェルズ ●	ウェルズ ●
QAC1-2146-01	松島菜々子	松島菜々子
QAC1-2172-01	松田聖子 ●	松田聖子 ●
QAC1-2178-01	小川	茶木滋●
QAC1-20021-01	小沢一郎	小渕●
QAC1-20037-01	鄭	鄭
QAC1-20039-01	菅原道真 ●	完
QAC1-20055-01	フジモリ ●	フジモリ ●
QAC1-20085-01	安藤忠雄 ●	安藤忠雄 ●
QAC1-20086-01	梅田	梅田

表 C.2　回答集その 2

	Version0.3.1.1	Version0.10.0+重み
QAC1-20121-01	向井	向井
QAC1-20123-01	小川	茶木滋●
QAC1-20126-01	船木	船木
QAC1-20142-01	スティーブン	ゲイツ●
QAC1-20143-01	エルステッド	エルステッド
QAC1-20202-01	ペレ	ペレ
QAC1-20205-01	西孝	西孝
QAC1-20330-01	赤松大麓	川端
QAC1-20336-01	野口健 ●	野口健 ●
QAC1-20343-01	堺屋 ●	堺屋 ●
QAC1-20345-01		石原慎太郎●
QAC1-20359-01	田村正和	石井●
QAC1-20386-01	レーガン ●	斉藤邦彦
QAC1-20389-01	武蔵丸 ●	武蔵丸 ●
QAC1-20422-01	鈴木 ●	鈴木 ●
QAC1-20443-01	松本幸四郎 ●	松本幸四郎 ●
QAC1-20456-01	夏目純一 ●	夏目純一 ●
QAC1-20633-01	縫	縫
QAC1-20638-01	松岡佑子 ●	松岡佑子 ●
QAC1-20639-01	矢野顕子	矢野顕子
QAC1-20649-01	小島直子	小島直子
QAC1-20663-01	姜錫柱	姜錫柱
QAC1-20708-01	アリス	キャロル
QAC1-20710-01	宮	小沢征爾●
QAC1-20737-01	森下卓 ●	森下卓 ●
正回答数	18	22