質問文中のキーワードと解候補の 連想の強さをを用いた解の決定

外池 昌嗣[†] 宇津呂 武仁[†] 佐藤 理史^{††}

質問応答システムにおける処理の流れのうち、何らかの方法で抽出された複数の解候補に順位付けし、答えを選ぶステップのことを本論文では、解選択のステップと呼ぶ。本論文では、大規模かつ日々更新されるウェブを利用して、質問文中の重要語句(キーワード)と解候補の共起に基づく語彙的關係(連想)の強さを測定し、これに基づいて解選択をする。この連想の強さはウェブのサーチエンジンのヒット数から計算できる尺度で表す。本論文では、この連想を利用した解選択法を2つ提案する。1つ目の手法は、あらかじめ決めておいた語の重みに基づいて質問文からキーワードを選択した後、キーワードと解候補の連想の強さに基づいて解を決める方法である。2つ目の手法は、キーワードと解候補の連想の強さを利用して、その質問にとって最も適切なキーワードと解候補を同時に選ぶ方法である。実験の結果、これら2つの手法を統合した手法で、4択クイズ「クイズ\$ミリオネア」の約79%の問題を解くことができた。また、サーチエンジンのヒット数を用いて解選択を行う従来手法の性能を有意に改善できた。

キーワード: 質問応答, 解選択, ウェブ, サーチエンジン, 4択クイズ, 連想

Answer Validation based on Association of Keywords in a Question Sentence and Answer Candidates

MASATSUGU TONOIKE[†], TAKEHITO UTSURO[†] and SATOSHI SATO^{††}

Answer validation is a component of question answering system, which selects the most reliable answer from the answer candidates that are extracted by certain methods. In this paper, we propose a new approach of answer validation based on the strength of lexical association between keywords in a question sentence and each of the answer candidates, and examines the effectiveness of the approach; we use search engine hits to calculate the association score. We propose two types of answer selection methods based on the association score. The first method first extracts the appropriate keywords from a given question sentence using word weights and then selects the final answer by using the association score. The second method selects appropriate keywords and answer candidate simultaneously using the association score. The result of experimental evaluation shows that a good proportion (79%) of a multiple-choice quiz “Who wants to be a millionaire” can be solved by combining these two methods. The proposed method significantly outperforms a previous method for answer selection based on search engine hits.

KeyWords: *Question Answering, Answer Validation, Web, Search Engine, Multiple-choice Quiz, Association*

1 はじめに

自然言語で与えられた質問の答えを探す技術は質問応答と呼ばれ、TREC QA Task (Voorhees 2003) や NTCIR の QAC (Fukumoto, Kato, and Masui 2004) などの評価型ワークショップを中心として、研究が盛んに行われている。質問応答のプロセスは、大きく2つのステップに分けることができる。一つは解候補を集めるステップで、もう一つは得られた解候補の中から最も信頼できるものを選択するステップである。最初の、解候補を集めるステップは、これまで比較的良好に研究されてきている。このステップの標準的な手順では、まず質問文を解析して、「場所」、「人名」といった解のタイプを判定する。次に、使用可能なドキュメント集合に対して、質問文から生成されるクエリーによって検索を行い、解が書いてありそうなドキュメントを収集する。そして、得られたドキュメントに対して固有表現抽出を行い、解が書かれている部分を特定するためのパターンを用いるなどしながら、解のタイプに照合する語を解候補として抽出する。

一方、本論文では、質問応答の2つ目のステップである、解の選択に注目する。解の選択法として、これまでにいくつかの方法が提案されている。よく知られている方法の一つに、質問文中のキーワードと解候補の単語間距離や頻度を用いる方法がある。(Clarke, Cormack, and Lynam 2001) では、100 ギガバイトのウェブコーパスを利用して、キーワードと解候補の共起頻度に基づく解候補の検証を行った。この100 ギガバイトのコーパスは、質問文中の語句を利用して集められたものではない。このようなコーパスを用いるとき、質問の答えがコーパス中に存在しないということは頻繁に起りうる。このため、現在最大規模の知識源ともいえるウェブ全体を、質問応答の知識源として用いることは、カバレッジの面で有利であると考えられる。

一方、(Kwok, Etzioni, and Weld 2001) はウェブを用いた最初の質問応答システムを提案した。まず、Google を使用して質問文に関係のある文書を集め、そこから解候補を抽出する。そして、解候補の近くに現れるキーワードとその解候補の距離、IDF で計算されるそのキーワードの重要度、及び、解候補の出現頻度に基づいて、解を選択する。しかし、使用するウェブページは、フレーズ検索を利用した制限の厳しいクエリーをサーチエンジンに与えて得られる比較的少数のページであり、膨大なウェブの情報ごく一部しか利用できない。これに対して、(Magnini, Negri, Prevete, and Tanev 2002) は、サーチエンジンのヒット数を用いることによって、ウェブ全体の共起情報を利用する解の選択法を提案した。彼らの方法では、問題文中の重要語句(キーワード)、解候補、及び、両者の論理積をクエリーとする検索のヒット数から計算できる尺度によって、解候補の信頼度を推定する。この方法は、キーワードと解候補の語彙的関係の強弱を数値化することにより、解候補の信頼度を推定する手法といえる。この方法の背景にあるのは、問題文中の重要語句と正解の解候補が共起しているウェブ文書の数、重要語句と不正解の解候補が共起しているウェブ文書の数より多く、この文書数がサーチエンジンの

† 京都大学大学院情報学研究科, Graduate School of Informatics, Kyoto University

†† 名古屋大学大学院工学研究科, Graduate School of Engineering, Nagoya University

ヒット数で近似できるという仮定である。ここで, (Magnini et al. 2002) の特徴として, 単一の尺度で解候補の信頼度を測定するという点があげられる。

これに対して, 本論文では, サーチエンジンのヒット数から計算できる, 性質の異なる 2 つの尺度 (Sato and Sasaki 2003) を導入する。そして, この 2 つの尺度を組み合わせる用いることによって解選択を行う方法を提案し, (Magnini et al. 2002) の手法の性能を有意に改善できることを示す。2 つの尺度を組み合わせる手法としては, 次の 2 つの手法を提案する。1 つ目の手法は, あらかじめ決めておいた語の重みに基づいて質問文からキーワードを選択した後, 2 つの尺度を計算し, どちらの尺度がこの場合信頼できるか, 人手で作成したルールで決定するという方法である。2 つ目の手法は, 2 つの尺度を利用して, その質問にとって最も適切なキーワードと解候補を同時に選ぶ方法である。

本論文では, 質問応答のプロセス全体ではなく解の選択の問題に焦点を当てるために, 解候補があらかじめ与えられている 4 択クイズを研究の実験材料とする。本論文で題材として使う 4 択クイズは, テレビのクイズ番組「クイズ\$ミリオネア」¹(クイズ\$ミリオネア 2002) の問題である。この番組は, イギリスで始まり, 50ヶ国以上の国でローカライズされて放映されているクイズ番組 “Who wants to be a millionaire?”²の日本版である。

2 章では, 問題文中のキーワードと解候補の共起に基づく語彙的關係のことを示す「連想」を用いた質問応答を導入する。そして, 連想の強さを測定する尺度として, (Sato and Sasaki 2003) の提案する性質の異なる 2 つの尺度を導入する。3 章では, この 2 つの尺度に基づく 2 つの解選択法を提案し, さらに, それらの解選択法を組み合わせる方法を提案する。4 章では, 実験の結果を示す。5 章では, 実験結果を分析して考察を行う。6 章では, 関連する研究を紹介し本論文との相違点を述べる。

2 連想による解の選択

2.1 連想

以下に 4 択クイズの例³を示す。

Q1: 「アメリカングラフィティ」の監督は誰?

- A: ジョージ・ルーカス B: スティーブン・スピルバーグ
C: フランシス・フォード・コッポラ D: 黒澤明

いま, この問題の答えが分からず, ウェブのサーチエンジンを使って答えを見つけようとしていくとしよう。もっとも単純な方法は, クエリー「アメリカングラフィティ」をサーチエンジン

1 フジテレビ製作 <http://www.fujitv.co.jp/quiz/>

2 Celador International 製作

3 2 章で示す例は, 説明のための例であり, 「クイズ\$ミリオネア」で実際に使用された問題ではない。

表 1 問題 Q1 におけるキーワード集合 K と選択肢 c の論理積のヒット数
($K = \{\text{“アメリカングラフィティ”}\}$)

選択肢 c	$hits(K \cup \{c\})$
“ジョージ・ルーカス”	384
“スティーブン・スピルバーグ”	123
“フランシス・フォード・コッポラ”	100
“黒澤明”	82

に入力し、得られたウェブページを調べるというものである。この戦略は、キーワード「アメリカングラフィティ」を含むウェブページ上に正解が現れることを仮定している。もう少し巧妙な方法は、キーワードと選択肢の両方を含むウェブページの数を使うというものである。この数は、キーワードと選択肢の論理積をクエリーとするサーチエンジン検索のヒット数で測定できる。そして、最大のヒット数を持つ選択肢が解であると決定する。表 1 に、キーワード集合 K および選択肢 c の論理積をクエリーとする検索のヒット数を示す。ここでは、ヒット数を求めるのに goo^4 を用いた。ここで、関数 $hits$ を以下のように定義する。

$$X = \{x_1, x_2, \dots, x_n\} \quad (1)$$

$$hits(X) \equiv hits(x_1 \text{ AND } x_2 \text{ AND } \dots \text{ AND } x_n) \quad (2)$$

問題 Q1 において、キーワードと選択肢の論理積をクエリーとする検索のヒット数が最大となる選択肢を選ぶと、正解である「ジョージ・ルーカス」となる。このように、この解法はキーワードと選択肢の共起に基づく語彙的関係の強さのみに基づくので、この解法は、問題 Q1 を以下に示すような連想問題に変換して解いているといえる。

Q1': 次のうち、「アメリカングラフィティ」と最も関係の深いものはどれ？

- A: ジョージ・ルーカス B: スティーブン・スピルバーグ
C: フランシス・フォード・コッポラ D: 黒澤明

本論文では、キーワードと選択肢の共起に基づく語彙的関係のことを「連想」と呼ぶ。そして、本節では、連想の強さを表現するために、キーワード集合 K と選択肢 c の論理積のヒット数を用いた。

4 <http://www.goo.ne.jp>

表 2 問題 Q2 におけるキーワード集合 (K または K') と選択肢 c の論理積のヒット数 ($K = \{\text{“ロードオブザリング”}\}$, $K' = \{\text{“ロードオブザリング”}, \text{“監督”}\}$)

選択肢 c	$hits(K \cup \{c\})$	$hits(K' \cup \{c\})$
“イライジャ・ウッド”	699	281
“ビリー・ボイド”	105	72
“ピーター・ジャクソン”	401	365
“ケイト・ブランシェット”	127	91

2.2 どのようにキーワードを選ぶか

精度よく問題を解くためには、問題文から適切なキーワードを取り出すことが重要である。以下の問題を考えよう。

Q2: 映画「ロードオブザリング」の監督は誰？

- A: イライジャ・ウッド B: ビリー・ボイド
C: ピーター・ジャクソン D: ケイト・ブランシェット

実際にヒット数を求めた結果を表 2 に示す。ここで、 K は $\{\text{“ロードオブザリング”}\}$, K' は $\{\text{“ロードオブザリング”}, \text{“監督”}\}$ とする。映画のタイトルである「ロードオブザリング」をキーワードとしたとき、キーワードとの論理積による検索のヒット数が最大の選択肢は「イライジャ・ウッド」であり、正解の「ピーター・ジャクソン」ではない。しかし、「ロードオブザリング」と「監督」の両方をキーワードとすれば、キーワードとの論理積が最大のヒット数を持つ選択肢を選ぶことによって解ける。この例からわかるように、連想を用いた解の選択においては、適切なキーワードを選びだすメカニズムが重要となる。

2.3 前向き連想と後ろ向き連想

キーワードと選択肢の論理積からなるクエリーを生成して、単純に最大のヒット数を持つ選択肢を選ぶ方法では解けず、選択肢のヒット数を考慮しなければならない問題もある。そこで、本節では、選択肢のヒット数を考慮する尺度を導入する。以下の問題について考える。

Q3: ピラミッドはどこにある？

- A: カナダ B: エジプト C: 日本 D: 中国

この問題に対して、実際にヒット数を求めた結果を表 3 に示す。この場合、キーワードと選択肢の論理積をクエリーとしたときのヒット数が最大の選択肢は「日本」であり、正解の「エジプト」ではない。ここで選択肢のみをクエリーとしたときのヒット数に注目しよう。「日本」のみをクエリーとした場合のヒット数 $hits(\{\text{“日本”}\})$ は、「エジプト」のみをクエリーとした場合

表 3 問題 Q3 におけるキーワード集合 K と選択肢 c の論理積のヒット数

キーワード集合 K	$hits(K)$			
{“ピラミッド”}	63,600			
選択肢 c	$hits(\{c\})$	$hits(K \cup \{c\})$	$FA(K, c)$	$BA(K, c)$
“カナダ”	321,000	3,300	0.0159	0.0103
“エジプト”	128,000	14,300	0.225	0.112
“日本”	3,050,000	27,800	0.437	0.00911
“中国”	2,580,000	10,300	0.162	0.00399

のヒット数 $hits(\{\text{“エジプト”}\})$ の約 24 倍である。したがって、「ピラミッド」と選択肢の論理積をクエリーとした場合のヒット数は、選択肢のみのヒット数の影響を受けていると考えられる。以上の結果から、ヒット数に関する何らかの正規化が必要となると考えられる。

そこで、本稿では、佐藤らの提案する、2 つの語の関係の強さを評価するための 2 つの尺度 (Sato and Sasaki 2003) を用いる。ここで、 K はキーワード集合、 c は選択肢として、キーワード集合 K と選択肢 c の論理積からなるクエリーのヒット数を K のヒット数で正規化したものを前向き連想 (Forward Association) $FA(K, c)$ と呼ぶ⁵。また、キーワード集合 K と選択肢 c からなるクエリーのヒット数を c のヒット数で正規化したものを後ろ向き連想 (Backward Association) $BA(K, c)$ と呼ぶ。

$$FA(K, c) = hits(K \cup \{c\}) / hits(K) \quad (3)$$

$$BA(K, c) = hits(K \cup \{c\}) / hits(\{c\}) \quad (4)$$

本稿では、連想の強さを表現するために、 FA または BA の値を用いる。

表 3 には、問題 Q3 の場合における FA および BA の値を示す。この場合、最大の BA を持つ選択肢が正解である。一般に、ある問題は FA が 1 位の選択肢を選択することで解ける一方、 BA が 1 位の選択肢を選択しなければ解けない問題もある。したがって、 FA が 1 位の選択肢と BA が 1 位の選択肢のどちらを選択するかを決める仕組みを考案する必要がある。

2.4 まとめ

2.1 ～ 2.3 節での議論より、連想による解の選択を実現するために、以下の 2 つの問題を解決する必要があることがわかる。

- どのようにして問題文から適切なキーワードを抽出するか。
- どのようにして前向きと後ろ向きの連想を考慮して正解を同定すればよいか。

⁵ ここで、 K が固定されていれば、 FA は $hits(K \cup \{c\})$ に比例する。

図 1 手法 1 の全体像

3 前向き及び後ろ向き連想に基づく解選択法

本章では、連想の強さを表現する尺度 FA 及び BA を用いて解選択を行う方法を 2 つ提案し、最後に両者の統合を行う。

3.1 手法 1：語の重みに基づくキーワード選択および解選択ルールによる方法

本章で説明する解選択法の全体像を図 1 に示す。問題が与えられると、システムは問題文からキーワードを選択する。次に、このキーワードを利用して、それぞれの選択肢に対して、 FA 及び BA の値を計算する。そして、 FA が 1 位の選択肢と BA が 1 位の選択肢を取り出し、解選択ルールによってどちらを選択するか決める。この枠組みのうち、以下では、キーワード選択の部分と、解選択の部分について説明する。

3.1.1 語の重みに基づくキーワード選択

手法 1 のキーワード選択はあらかじめ決めておいた語の重みに基づいて行う。語の重みでは人名などの固有名詞や「」中の語句を重視し、重要度の高い語を 1 個または 2 個選ぶ。重みおよび、ヒット数の閾値等のパラメータは、「分析用セット」⁶を用いて人手で調整した。評価に関しては、「評価用セット」を用いて行う。

アルゴリズム

語の重みに基づくキーワード選択法のアルゴリズムを以下に示す。

- (1) 問題文が n 個の「」で括られた引用を含むならば、それらの n 個の文字列をキーワードとして選ぶ。ただし、後に説明するストップワードは除外する。

⁶ 詳細は 4.1 節で述べる

表 4 語の重み付けルール

ルール番号	条件	重み
1	n 番目の文節	$1 + 0.01 * n$
2	ストップワード	0
3	「」中の語句	3
4	人名	3
5	サ変名詞の後ろに「する」が存在	0.5
6	関係を表す語	2
7	カタカナ (重要でないものを除く)	2
8	「賞」で終わる	2
9	「時代」で終わる	0.5
10	国名	0.5
11	数字	3
12	ヒット数が 1,000,000 より大 and 1 文字	0.9
13	キーワードの後ろに「は」が存在 and	
13-a	「者」または「家」で終わる	0.1
13-b	ヒット数が 100,000 より大	0.2
13-c	ヒット数が 10,000 より小	1.1
13-d	文字数が 1	0.2
13-e	文字数が 2	0.25
13-f	文字数が 3	0.5
13-g	文字数が 4	1.1
13-h	文字数が 5 以上	1.2

(2) それ以外の場合:

- 2-1. 日本語形態素解析システム JUMAN(黒橋, 長尾 1999) 及び日本語構文解析システム KNP (黒橋 1998) を用いて, 問題文からすべての文節を抽出する. 次に, 取り出された個々の文節から形態品詞情報を参照し, 「名詞⁷」「未定義語」「接頭辞」「接尾辞」で構成される最長の形態素列を取り出す. このようにして抽出された形態素列の集合を N とする.
- 2-2. N の要素である形態素列のヒット数をそれぞれ求め, ヒット数が 15 未満の場合は, その形態素列は N から取り除き, 代わりにその形態素列のそれぞれの形態素を N の要素とする.
- 2-3. 表 4 に示す語の重み付けルールにしたがって, キーワード候補の集合 N のそれぞれの要素に重みを付与する.
- 2-4. 最大の重み及び 2 番目に大きな重みを持つキーワード候補を選択する.
- 2-5. (i) これら 2 つのキーワード候補の論理和をクエリーとする検索のヒット数が 15 以上であれば, 両方をキーワードとして選択する.
(ii) そうでなければ, 重み 1 位のキーワードのみを選択する.

⁷ ただし, 品詞細分類名が「形式名詞」「副詞的名詞」の形態素を除く.

語の重み付けルールの詳細

表 4 の語の重み付けルールの詳細について述べる.

ルール 1 の「 n 番目の文節」は, 語の重みに差がつかない場合に, 文の後ろにある語を重視することを目的としたルールである.

ルール 2 の「ストップワード」では, 分析用セットを調査することにより, ストップワードとして以下の 55 語を手で設定した. このような語は, 通常我々がサーチエンジンのクエリーとして用いないものである.

人, 人物, おなじみ, きっかけ, 現在, 往年, 自分, 全部, 物, 特徴, 種類, 場所, 1 つ, 例年, こと, いくつ, 私たち, 言葉, 昔, 一つ, 省庁, 都市, 名作, 行名, 一人, 存在, 大人気, 従来, 段階, 一度, 誰, 何, 名前, 名称, 名, 題名, 地名, 者, 人気, 意味, 国, 毎年, 正体, もと, 県, 市, 都道府県, 当時, 登場, 各界, キャラクター, 動物, 義務, 事, 題材

ルール 6 の「関係を表す語」では, 語と語の関係を表す語は重要であるとみなし, このような語を分析用セットから手で抽出した. このような語には, 通常より大きな重みを付与する.

父, 父親, 母, 母親, 兄, 弟, 祖母, 祖父, 夫, 弟子, 師匠, 親会社, 子会社, 母校

ルール 7 の「カタカナ (重要でないものを除く)」では, カタカナ語は人名や地名などの固有表現を表していることが多いので重視する. カタカナ語の認定は, KNP が形態素に与える属性に「カタカナ」という文字が含まれるかによって行う. ただし, 抽象的な語は重要でないので, 重みを変化させない. 分析用セットの調査の結果, 「クラス」, 「スポーツ」, 「タレント」, 「メンバー」の 4 語を抽象的な語と判定した.

語の重みに基づくキーワード選択法の適用例

以下に, この手法の適用例を示す.

Q4: 鯨資源の管理を目的とする国際捕鯨委員会の略称は何?

A: WBA B: IWC C: WIFI D: WHO

という問題を考える. まず問題文中の名詞を調べ, 語の重み付けルールにしたがって重みをつけると, 次のようになる. 「鯨資源 (1)」, 「管理 (1.01)」, 「目的 (1.02)」, 「国際捕鯨委員会 (1.03)」, 「略称 (0.26)」. そして, 重みの大きさが上位の 2 語をキーワードとする. この問題の場合は, 「国際捕鯨委員会」と「目的」になる.

評価実験

以上説明したキーワード選択法の精度を調べるために実験を行った。\$K\$ を選ばれたキーワードの集合 (\$K \subseteq N\$), \$c\$ を選択肢として, \$c_1^F(K)\$ を最大の \$FA(K, c)\$ を持つ選択肢, \$c_1^B(K)\$ を最大の \$BA(K, c)\$ を持つ選択肢と定義する。

$$c_1^F(K) = \operatorname{argmax}_c FA(K, c) \quad (5)$$

$$c_1^B(K) = \operatorname{argmax}_c BA(K, c) \quad (6)$$

ここで, \$c_1^F(K)\$ か \$c_1^B(K)\$ のどちらかが正解であれば, このとき選択されたキーワードは正解であるとする, 4.1 節で導入される分析用セットに対する精度は, 84.5% であった。

3.1.2 解の選択

3.1.1 節のキーワード選択部で得られたキーワード集合 \$K\$ を利用すると, サーチエンジンを利用して, 以下の値を求めることができる。

- キーワード集合 \$K\$ の論理積をクエリとする検索のヒット数: \$hits(K)\$
- 選択肢 \$c\$ のヒット数: \$hits(\{c\})\$
- キーワード集合 \$K\$ 及び選択肢 \$c\$ の論理積をクエリとする検索のヒット数: \$hits(K \cup \{c\})\$

これらのヒット数の情報を利用すれば, 式 (3), (4) に従って, 各選択肢の \$FA\$ 及び \$BA\$ が計算できる。本節では, \$FA\$ の値が 1 位の選択肢か, \$BA\$ の値が 1 位の選択肢か, どちらが正解かを同定する方法を説明する。

ここで, 本節及び付録 A で用いる記号について説明する。まず, 解候補を表す記号として, 3.1.1 節で導入した \$c_n^x\$ を用いる。これは, 尺度 \$x\$ が \$n\$ 位であった選択肢を示す記号である。ただし, \$x\$ には, 尺度が \$FA\$ であれば \$F\$ を, \$BA\$ であれば \$B\$ を用いる。次に, 選択肢 \$c\$ の \$FA\$ の値を表す \$F(c)\$ 及び, 選択肢 \$c\$ の \$BA\$ の値を表す \$B(c)\$ を導入する。この記号を利用して記述すると, 解選択においてすべきことは, \$c_1^F\$ または \$c_1^B\$ のどちらが正解かを推定することである。

本論文では, 分析用セットにおいてサーチエンジンのヒット数を調査することにより, \$c_1^F\$ または \$c_1^B\$ のどちらかを選択するルールを人手で作成した。基本的な考え方は, どちらかの尺度が僅差であれば他方を信頼し, 大差であればその尺度を信頼するというものである。

このとき, どの選択肢とどの選択肢を比べるのかという点で, 以下の 2 つが考えられる。

- (1) \$c_1^F\$ と \$c_1^B\$
- (2) \$c_1^F\$ と \$c_2^F\$, 及び, \$c_1^B\$ と \$c_2^B\$

どちらが有効かという調査の結果に関しては付録 A で述べ, ここからは, 調査の結果, 有効であった (1) の比較を用いる方法についてのみ説明する。

調査では, まず, 分析用セットの 888 問に対して, キーワードの選択を行い, それぞれの選択肢の FA と BA を求めた. 次に, ここから, c_1^F と c_1^B の両方が不正解である問題を除いた. なぜならば, ここで除外する問題は, 手法 1 の枠組みでは解くことが出来ないからである. そして, c_1^F と c_1^B が一致する問題も除いた. このような場合には, その選択肢を選べばよい. 残るのは, c_1^F または c_1^B のどちらか一方が正解となる問題である. 解選択ルールが対象とするのはこのような問題である.

$F(c_1^F)$ と $F(c_1^B)$ を比較するために, 以下に示す比の値を用いる.

$$r(F) = \frac{F(c_1^B)}{F(c_1^F)} \quad (7)$$

同様に, $B(c_1^F)$ と $B(c_1^B)$ を比較するために, 以下に示す比の値を用いる.

$$r(B) = \frac{B(c_1^F)}{B(c_1^B)} \quad (8)$$

これらの値が 1 に近い場合はその尺度が僅差であることを示しているので, この場合は, その尺度は無視し, もう一方の尺度の支持する選択肢を採用する. 逆に 0 に近い場合はその尺度が大差であることを示しているので, この場合は, その尺度の支持する選択肢を採用し, もう一方の尺度は無視する.

一方, $F(c_1^B)$, $F(c_1^F)$, $B(c_1^F)$, $B(c_1^B)$ に, 上記の特徴が見られない場合は, キーワード集合の論理積をクエリとしたときのヒット数 $hits(K)$ を調べる. $hits(K)$ が大きい場合は, K が一般的な語である可能性が高いので, FA の値には, それぞれの選択肢のヒット数 $hits(\{c\})$ の大小の傾向がそのまま反映されてしまう可能性がある. それよりは選択肢のヒット数 $hits(\{c\})$ で正規化した尺度である BA によって選ばれた c_1^B の方が信頼できると考えられるので, c_1^B を採用するものとする.

以上のような考え方に基づいてパラメータを決定し, 以下の解選択ルールを作成した.

- (1) if $c_1^F = c_1^B$ then c_1^F
- (2) else if $r(F) \geq 0.8$ then c_1^B
- (3) else if $r(F) \leq 0.2$ then c_1^F
- (4) else if $r(B) \geq 0.53$ then c_1^F
- (5) else if $hits(K) \geq 1300$ then c_1^B
- (6) else if $r(F) \geq 0.6$ then c_1^B
- (7) else c_1^F

このルールを前述の問題 Q4 に対して適用する. 選ばれたキーワードは「国際捕鯨委員会」と「目的」であった. ヒット数を求めた結果を表 5 に示す. c_1^F は「WHO」で, c_1^B は「IWC」である. この場合, ルール (1) には該当しない. 次にルール (2) を調べると, $r(F) = 713/851 \simeq 0.838$

表 5 問題 Q4 において FA と BA を求めた結果 ($K = \{ \text{“国際捕鯨委員会”, “目的”} \}$)

$hits(K)$	851			
選択肢 c	$hits(\{c\})$	$hits(K \cup \{c\})$	FA	BA
“WBA”	7350	2	0.002350	0.027211
“IWC”	18600	713	0.837838	3.833333
“WIFI”	5410	0	0.000000	0.000000
“WHO”	325000	851	1.000000	0.261846

表 6 解選択の精度 (語の重みに基づくキーワード選択との組み合わせ)

手法	精度
c_1^F を選択	70.8%
c_1^B を選択	67.6%
解選択ルール	77.3%

なので条件を満たす。したがって、 $BA1$ 位の選択肢である「IWC」が選ばれる。

表 6 に、分析用セットに対して、この解選択法を適用した場合の精度を示す。ここで、キーワードは 3.1.1 節の語の重みに基づく手法で選んだものを用いた。この表では、上記の解選択ルールの結果に加えて、単純に c_1^F または c_1^B を選ぶ、ベースラインとなる手法の結果も示した。この結果より、ここで述べた解選択ルールはベースラインの手法の精度を大きく上回ることが分かる。

3.2 手法 2：連想の強さの比を用いる方法

手法 1 では、キーワードを選択するにあたって、キーワードと選択肢の間の連想を用いず、あらかじめ決めた語の重みに基づいてキーワードを選択した。一方、手法 2 では、キーワードと選択肢の間の連想を用いてキーワードを選択する。以下では、連想に基づいてキーワードと解を同時に選ぶ手法をいくつか述べる。なお、手法 2 を設計するにあたって、本章で述べる方法以外にもいくつかの方法を設計したが、精度または適用できる問題の条件に問題があったので採用しなかった。これらの手法に関しては、付録 B で述べる。

まず、選択すべきキーワード集合の候補を作成する。その作成法では、まず、日本語形態素解析システム JUMAN を用いて、問題文からすべての名詞を抽出する。このとき、名詞の列が複合語を構成するとき、最も長い複合語のみを抽出し、その構成要素である名詞は抽出しない。このようにして抽出された名詞や複合語の集合を N とする。次に、 N の巾集合 2^N を作成する。このようにして作成された 2^N のそれぞれの要素が、選択すべきキーワード集合の候補となる。

本章で提案する手法は、 FA や BA といった尺度の 1 位と 2 位の比を用いて、最も適切なキーワード集合と選択肢の組を同時に選択するものである。この手法の根底にある仮説は、適切なキーワード集合が選ばれているとき、そのキーワード集合と正解の間の連想の強さは、そ

のキーワード集合と他の選択肢の間の連想の強さよりも十分大きく、逆に、不適切なキーワード集合が選ばれているときは、そのキーワード集合と正解の間の連想の強さは、そのキーワード集合と他の選択肢の間の連想の強さと差がない、という考え方である。ここでは、連想の強さが1位と2位の選択肢の連想の強さを比較し、1位の選択肢の連想の強さは、2位のそれより十分に大きければ、キーワード集合と1位の選択肢は信頼できると考える。

以下ではまず、基本的な手法として、 FA の比を用いる手法及び、 BA の比を用いる手法を導入する。

FA の比 FA の比とは、あるキーワード集合に対して、2 番目に大きな FA を持つ選択肢の FA を、最大の FA を持つ選択肢の FA で割った値である。 FA の比を用いる手法では、この比を用いてキーワード集合を評価する。実際の手順ではまず、このような FA の比をすべてのキーワード集合に対して計算し、最小の比の値を持つキーワード集合を選ぶ。そして、このときの FA 最大の選択肢を解とする。選ばれたキーワード集合を $\hat{K}$ 、選ばれた選択肢を $\hat{c}$ とすると、この方法は以下のように定式化できる。

$$c_2^F(K) = \underset{c}{\text{arg-secondmax}} FA(K, c) \quad (9)$$

$$\hat{K} = \underset{K \in 2^N}{\text{argmin}} \frac{FA(K, c_2^F(K))}{FA(K, c_1^F(K))} \quad (10)$$

$$\hat{c} = c_1^F(\hat{K}) \quad (11)$$

ただし、**arg-secondmax** は関数の値が2 番目に大きくなるときの、引数を返す関数とする。

以上の方法の適用例を示す。Q1(「アメリカングラフィティ」の監督は誰?) の場合について、キーワード集合と選択肢の組、その組の FA 、及び FA の比を表7に示す。ここでは、 FA の比の値が最小のキーワード集合、{「アメリカングラフィティ」、「監督」} が選ばれ、 FA が最大である選択肢、「ジョージ・ルーカス」が解となる。

同様に BA の比に基づく手法を以下に示す。

BA の比

$$c_2^B(K) = \underset{c}{\text{arg-secondmax}} BA(K, c) \quad (12)$$

$$\hat{K} = \underset{K \in 2^N}{\text{argmin}} \frac{BA(K, c_2^B(K))}{BA(K, c_1^B(K))} \quad (13)$$

$$\hat{c} = c_1^B(\hat{K}) \quad (14)$$

次に、 FA の比や BA の比に基づく方法とは異なり、 FA と BA の両方を考慮する方法を述べる。

表 7 FA の比を用いてキーワード集合と選択肢を選ぶ手法の例

キーワード集合 K	選択肢 c	FA	比 ($FA2$ 位/ $FA1$ 位)
{“アメリカングラフィティ”}	“ジョージ・ルーカス”	0.165	0.344 ($\simeq 0.0569/0.165$)
	“ステイーブン・スピルバーグ”	0.0569	
	“フランシス・フォード・コッポラ”	0.0408	
	“黒澤明”	0.0341	
{“アメリカングラフィティ”, “監督”}	“ジョージ・ルーカス”	0.457	0.324 ($\simeq 0.148/0.457$)
	“ステイーブン・スピルバーグ”	0.148	
	“フランシス・フォード・コッポラ”	0.135	
	“黒澤明”	0.0876	
{“監督”}	“ジョージ・ルーカス”	0.00604	0.996 ($\simeq 0.00809/0.00813$)
	“ステイーブン・スピルバーグ”	0.00813	
	“フランシス・フォード・コッポラ”	0.00252	
	“黒澤明”	0.00809	

FA が 1 位の選択肢と 2 位の選択肢の BA の比 それぞれのキーワード集合に対して, FA が 1 位の選択肢と 2 位の選択肢を選び, それらの選択肢の BA の比を計算する. そして, この比の値が最も小さかったキーワードを選び, $\hat{K}$ とする. そして, $c_1^F(\hat{K})$ をこの手法の選んだ解とする.

$$\hat{K} = \underset{K \in 2^N}{\operatorname{argmin}} \frac{BA(K, c_2^F(K))}{BA(K, c_1^F(K))} \quad (15)$$

$$\hat{c} = c_1^F(\hat{K}) \quad (16)$$

以上の方法の適用例を示す. Q1(「アメリカングラフィティ」の監督は誰?) の場合について, キーワード集合と選択肢の組, その組の FA , 及び BA の比を表 8 に示す. ここでは, BA の比の値が最小のキーワード集合, {“アメリカングラフィティ”, “監督”} が選ばれ, FA が最大である選択肢, 「ジョージ・ルーカス」が解となる.

同様に, BA が 1 位の選択肢と 2 位の選択肢の FA の比を用いる手法を以下に示す.

BA が 1 位の選択肢と 2 位の選択肢の FA の比

$$\hat{K} = \underset{K}{\operatorname{argmin}} \frac{FA(K, c_2^B(K))}{FA(K, c_1^B(K))} \quad (17)$$

$$\hat{c} = c_1^B(\hat{K}) \quad (18)$$

これら 4 つの手法のカバレッジと精度を表 9 に示す. カバレッジは, 問題セットに含まれる問題のうち, 比が 1 以下の問題の割合によって測定する⁸. したがって, FA の比を用いる手法及び BA の比を用いる手法のカバレッジは 100%, FA が 1 位の選択肢と 2 位の選択肢の BA の比を用いる手法及び BA が 1 位の選択肢と 2 位の選択肢の FA の比を用いる手法の場合は,
⁸ この比は FA と BA の両方を考慮するので, 比が 1 より大であることは FA と BA が指し示す選択肢が異なることを意味する.

表 8 $FA1$ 位の選択肢と $FA2$ 位の選択肢の BA の比が最小となるキーワード集合と選択肢を選ぶ手法の例

キーワード集合 K	選択肢 c	FA	BA	比 ($FA2$ 位の BA / $FA1$ 位の BA)
{“アメリカングラフィティ”}	“ジョージ・ルーカス”	0.165	0.0499	0.200 ($\simeq 0.0100/0.0499$)
	“ステイーブン・スピルバーグ”	0.0569	0.0100	
	“フランシス・フォード・コッポラ”	0.0408	0.0358	
	“黒澤明”	0.0341	0.00672	
{“アメリカングラフィティ”, “監督”}	“ジョージ・ルーカス”	0.457	0.0455	0.189 ($\simeq 0.00858/0.0455$)
	“ステイーブン・スピルバーグ”	0.148	0.00858	
	“フランシス・フォード・コッポラ”	0.135	0.0392	
	“黒澤明”	0.0876	0.00570	
{“監督”}	“ジョージ・ルーカス”	0.00604	0.815	1.12 ($\simeq 0.714/0.639$)
	“ステイーブン・スピルバーグ”	0.00813	0.639	
	“フランシス・フォード・コッポラ”	0.00252	0.992	
	“黒澤明”	0.00809	0.714	

表 9 連想の強さの比に基づく手法の評価 (精度・カバレッジ)(%)

		1 位, 2 位	
		FA	BA
比	FA	63.1/100	70.6/95.0
	BA	75.8/93.2	67.6/100

比が 1 以下の問題の割合となる。精度は、比が 1 以下の問題のうち選ばれた選択肢 $\hat{c}$ が正解である問題数を、比が 1 以下の問題数で割った値によって測定される。最もよい精度を示したのは、 FA が 1 位の選択肢と 2 位の選択肢の BA の比を用いる手法であった。以下の章では、この比を「連想の強さの比」と呼ぶ。比の値の範囲とカバレッジ・精度の関係を図 2 に示す。比が 0.25 以下であるとき、全体の約 60 % の問題が 90 % 近くの精度で解ける。この 0.25 という閾値は、3.3 節で連想の強さの比と語の重みを組み合わせるときに用いる。

3.3 手法 1 と手法 2 の統合手法

3.2 節で示した連想の強さの比を用いる手法では、比が 0.25 以下であれば高い精度で問題を解くことができた。一方、3.1.1 節で示した語の重みに基づくキーワード選択法と 3.1.2 節で示した解選択ルールを組み合わせた手法は、分析用セットの問題すべてを解いたときには、連想の強さの比を用いる手法よりも精度がよかった。そこで、両者の長所を生かすために、以下の手順で 2 つの手法を統合する。

(1) 比が 0.25 以下のとき:

FA が 1 位の選択肢と 2 位の選択肢の BA の比によって選ばれたキーワード集合を選

図 2 FA が 1 位の選択肢と 2 位の選択肢の BA の比：カバレッジ・精度

図 3 統合手法の概略図

択し、最大の FA を持つ選択肢を選ぶ。

(2) それ以外の場合:

語の重みに基づく方法で選ばれたキーワード集合を使い、3.1.2 節の手順によって、解選択を行う。

4 実験

4.1 データセット

本論文では、4 択クイズとして、トミー社から発売されているカードゲーム版の「クイズ \$ ミリオネア」の問題を用いた。カードゲーム版の問題は、1,960 問あり、それらは、賞金額別に A(¥10,000)~O(¥10,000,000) の 15 クラスからなる。それぞれの問題は解候補として 4 つの選択肢を持つ。出題範囲は、アニメ・マンガ、ことば、スポーツ、ファッション、映画、音楽、科学・数学・天文学、芸術・美術、芸能、雑学、時事・社会、社会風俗、政治経済、地理、動植物、童謡・童話、文学、料理、歴史と多岐にわたる。以下に「クイズ \$ ミリオネア」の問題例を示す。

10,000 円クラス

[A39] エジプトやケニアなどの国がある大陸はどこ？

A: アフリカ大陸 B: ユーラシア大陸 C: 北アメリカ大陸 D: 南アメリカ大陸

[正解: アフリカ大陸]

1,000,000 円クラス

[J39] コロンブスが新大陸発見の航海をしたときに乗っていた船の名前は何？

A: アトランティス号 B: アルゴ号 C: サンタマリア号 D: ノーチラス号

[正解: サンタマリア号]

10,000,000 円クラス

[O4] 夏期オリンピック大会で参加国数が初めて 100 ヶ国を越えた大会はどれ？

A: ローマ五輪 B: 東京五輪 C: メキシコ五輪 D: ミュンヘン五輪

[正解: メキシコ五輪]

ここでは、それぞれのクラスを 2 つに分割し、半分 (980 問) を分析用の問題セットとして、残り半分を評価用の問題セットとして使用する。ただし、比較表現を含む問題 (例えば、「次のうち、最も州の数が多い国はどれ？」) と否定の表現を含む問題 (例えば、「次のうち、フランスの国旗に使われていない色はどれ？」) は連想で解くのにふさわしくないので除外した。比較表現を含む問題を除く際には、問題文に「次の～最」というパターンを含む問題、及び、「最も」または「もっとも」を含む問題を機械的にすべて除外した。同様に、否定表現を含む問題を除く際には、問題文に「ない」を含む問題を機械的にすべて除外した。その結果、分析用セットの問題数は 888 問に、評価用セットの問題数は 906 問になった。ここで、賞金額別のそれぞれの問題セットの問題数を表 10 に示す。

ここで、「クイズ \$ ミリオネア」の問題は、TREC 2003 QA トラックや NTCIR の QAC2 タスクに比べて易しいのかどうかを考える。「クイズ \$ ミリオネア」のすべての問題は事実を問う問題である。これは、TREC 2003 QA トラックのメインタスクに含まれる問題のうち、事実を

表 10 賞金額別の問題数

賞金額 (円)	問題数		
	全問題	分析用	評価用
10,000	160	71	74
20,000	160	71	77
30,000	160	67	70
50,000	160	75	71
100,000	160	73	73
150,000	160	76	72
250,000	160	71	77
500,000	160	74	77
750,000	160	78	71
1,000,000	160	73	76
1,500,000	120	53	58
2,500,000	90	38	42
5,000,000	70	30	32
7,500,000	50	24	21
10,000,000	30	14	15
合計	1960	888	906

問う問題に対応する。また、NTCIR の QAC2 タスクの問題も事実を問う問題である。NTCIR QAC2 サブタスク 1 の問題と「クイズ\$ミリオネア」の問題を比較したところ、筆者の主観では難しさに差はないと感じられた。実際、表 11 に示した例のように、NTCIR QAC2 サブタスク 1 の問題と似た問題を「クイズ\$ミリオネア」から見つけることができる。ここで、問題の難しさを測定する一指標として、これらのセットの問題の問題文の長さを比較した。「クイズ\$ミリオネア」と NTCIR QAC2 サブタスク 1 に関しては、個々の問題の問題文の文節数の平均を計算し、TREC 2003 QA トラックの事実を問う問題では、問題文の単語数の平均を計算した。その結果は、「クイズ\$ミリオネア」は平均 7.24 文節⁹、TREC 2003 QA トラックは平均 7.76 ワード、NTCIR QAC2 サブタスク 1 は平均 6.19 文節であった。したがって、「クイズ\$ミリオネア」の問題は TREC 2003 QA トラックや NTCIR QAC2 サブタスク 1 と比べても短いというわけではないことがわかった。

4.2 分析用セット及び評価用セットを用いた実験結果

分析用セットと評価用セットに対して、3.3 節で示した手法 1 と手法 2 の統合手法の評価の結果を表 12 に示す。この表では、3.2 節で示した手法 2(連想の強さの比を用いる手法) と、3.1 節で示した手法 1(語の重みに基づくキーワード選択法、及び、解選択ルールを組み合わせた手法) という、2 つのベースラインの性能も示している。ここで、 r は手法 2 における、 FA が 1 位の選択肢と 2 位の選択肢の BA の比を表す。カバレッジは、問題セットに含まれる問題のう

⁹ なお、賞金額毎に問題文の平均文節数を調べた結果、問題文の平均文節数は賞金額に依存せず、ほぼ 7 文節であった。

表 11 NTCIR QAC2 の問題とクイズ \$ ミリオネア の問題

NTCIR QAC2 サブタスク 1		クイズ \$ ミリオネア	
QAC2-10176-01	パキスタンの首都はどこですか.	D60	ルーマニアの首都はどこ?
QAC2-10196-01	無着陸世界一周飛行に挑戦した熱気球の名前は何ですか.	J39	コロンブスが新大陸発見の航海をしたときに乗っていた船の名前は何?
QAC2-10051-01	米ラスベガス, テーマホテルが林立するラスベガス大通りは通称何と呼ばれていますか.	A76	とげぬき地蔵で有名な東京・巣鴨 (すがも) は, 通称「誰の原宿」?

表 12 最終的な実験結果

手法	分析用セット (888 問)				評価用セット (906 問)			
	#	Cov	○	精度	#	Cov	○	精度
手法 1: 語の重み + 解選択ルール	888	100%	686	77.3%	906	100%	665	73.4%
手法 2: 連想の強さの比 ($r \leq 1$)	828	93.2%	628	75.8%	848	93.6%	633	74.6%
統合手法	888	100%	698	78.6%	906	100%	688	75.9%
手法 2 ($r \leq 0.25$)	536	60.4%	466	86.9%	557	61.5%	479	86.0%
手法 1 ($r > 0.25$)	352	39.6%	232	65.9%	349	38.5%	209	59.9%

#: 手法が適用された問題数, Cov: カバレッジ, ○: 正解数

ち, 括弧中の r の条件を満たす問題の割合を表す. ただし, 条件が与えられていない手法のカバレッジは 100% となる. 精度は, 括弧中の r の条件を満たす問題のうち, 正解であった問題の割合を示す. ただし, 条件が与えられていない手法の精度は, 問題セットに含まれる問題のうち, 正解であった問題の割合となる. 3.3 節で示した統合手法は, それらのベースラインの性能を上回った. 結果として, 約 79% (分析用セット) あるいは約 76% (評価用セット) の問題を, 3.3 節で示した統合手法により正解することができた.

2 つの問題セットを比較すると, 手法 1 では, 評価用セットにおいて精度が約 4% 低下した. これは, 語の重み付けルールと解選択ルールが分析用セットに過適応していることを示している. 一方, 手法 2 においては, 2 つの問題セットの間で精度の差は 1.2% であり, 手法 1 における 2 つの問題セット間の差よりも少なかった. これより, 手法 2 は, データセットへの依存性が低いといえる.

ここで, 人間にとっての問題の難しさと, 計算機にとっての問題の問題の難しさに相関があるのかを調べるために, 分析用セットに対して, 賞金額別に精度を求めた. その結果を図 4 に示す. 賞金額が大きくなるほど精度も下がる傾向が見られるが, 賞金額が大きい問題の精度が極端に低いわけではない.

図 4 賞金額別の精度

4.3 パラメータの評価

ここからは、システム中のパラメータの評価を行っていく。

4.3.1 統合手法における比の上限值

まずは、統合手法において、比の上限値を 0.25 としたのは妥当であったかについて調べる。この比の上限値は、統合手法において、手法 2 が適用される上限値であり、図 2 に基づいて、精度の高さとカバレッジの大きさのバランスがよくなる値を目測で決定したものである。ここでは、分析用セットに対して、上限値を変化させ、精度の推移を観察した。結果を図 5 に示す。この結果、上限値 0.3 を境に精度が急激に低下していることが分かる。したがって、比の上限値を 0.25 としたのは妥当であったと言える。

4.3.2 語の重みに基づくキーワード選択法における寄与度の分析

次に、語の重みに基づくキーワード選択法において、どの部分が精度に寄与しているかを調べるために、以下に示す 4 つの要素をそれぞれ外した状態の精度を調べた。

- (1) 語の重み表
- (2) ヒット数が 15 未満の複合語を分解するルール
- (3) 「」中の語句の特別処理
- (4) キーワードの個数制限

図5 閾値を変化させたときの統合手法の精度

図6 語の重みに基づくキーワード選択法の分析

その結果を図6に示す。最もよい結果であったのは、すべての要素がそろっている状態であった。ここから、一つだけ要素を外したときに最も影響が大きかったのは語の重み表であった。逆に貢献が少ないのは、複合語の分解ルールであった。これは、このルールの対象となる場合が少ないことが理由である。「」中の語句の特別処理に関しては、語の重み表が使える場合は大き

表 13 それぞれの解選択ルールの評価 (語の重みに基づくキーワード選択を用いる)

ルール	答え	精度
1	$c_1^F = c_1^B$	88.5% (479/541)
2 ~ 6	-	60.3% (207/343)
合計	-	77.6% (686/884)
2	c_1^B	65.3% (32/49)
3	c_1^F	61.8% (68/110)
4	c_1^F	53.6% (37/69)
5	c_1^B	60.3% (35/58)
6	c_1^B	66.7% (12/18)
7	c_1^F	59.0% (23/39)

く精度に貢献しているとは言えない。個数制限に関しては、これを外すと、精度が2.8%低下するので、個数の制限は有効であった。最も悪い結果となったのは、キーワードをランダムに少数個選ぶ「xxxo」の場合であった。

4.4 解選択ルールの分析

解選択ルールのそれぞれに対して、表 13 にその精度を示す。分析用セットにおいて¹⁰、 c_1^F と c_1^B が一致する問題は 541 問 (約 60%) あり、選ばれた選択肢の 88.5% が正解であった。したがって、この c_1^F と c_1^B が一致すれば、選ばれた解は信頼できる。

なお、この解選択ルールを goo 以外の他のサーチエンジンの返すヒット数に対して適用するためには、 $hits(K)$ の様にヒット数の絶対値を判断条件にしている部分に関しては修正する必要があると考えられる。

4.5 サーチエンジンの影響

次に、サーチエンジンそのものが、本手法にどのような影響を与えているかを調べる。goo は 2003 年 12 月 1 日よりエンジンを、google に変更した。ここでは、旧 goo と新 goo を使ったときで、精度がどのように変化するかを調べた。

まず、手法 1 に対して実験を行った。サーチエンジンによって、蓄えているウェブページの絶対数が違うことに注意すると、ヒット数の絶対値を用いている部分が影響を受けることが考えられる。手法 1 において、ヒット数の絶対値を用いているのは、キーワード選択の部分、及び、解選択ルールにおいてキーワードのヒット数の絶対値の大きさを評価している部分である。そこで、特に解選択ルールに対する影響を調べるために、実験では、ヒット数のデータに旧 goo を用いるか新 goo を用いるか、及び、解選択ルールに旧 goo 用にパラメータを調整したものをを用いるか新 goo 用に調整したものをを用いるかの組み合わせで、合計 4 通りの実装の精度を求め

10 キーワード集合 K と選択肢 c の論理積をクエリとする検索のヒット数 $hits(K \cup \{c\})$ が 0 であった 4 問は除く

表 14 手法 1 においてサーチエンジン及び解選択ルールを変更した場合の精度

		解選択ルール	
		新 goo 用に調整	旧 goo 用に調整
サーチエンジン	新 goo	77.3 %	74.5 %
	旧 goo	74.5 %	76.3 %

た. 実験の都合上, 問題セットは分析用セットから旧 goo でヒット数のデータを取得していない 7 問を除いた 881 問の問題セットで行った. 実験の結果を表 14 に示す. この結果より, 解選択ルールのパラメータの調整を行わなくても 74.5 % という比較的高い精度を達成できることがわかる. また, パラメータの調整を行えば, サーチエンジンが違っていても同程度の精度を達成できることがわかる.

次に, 手法 2 に対して実験を行った. 手法 2 には, サーチエンジンに合わせて調整すべき箇所はない. 問題セットは, 旧 goo のヒット数のデータを用意することができた 222 問である. この 222 問の問題セットは, 以下の手順で作成した. まず, 比較表現を含む問題と, 否定語を含む問題を除く前の分析用セット (980 問) から, 4 問おきに問題を抽出し, 245 問を得た. 次に, この 245 問から, 旧 goo のヒット数のデータを用意出来ない問題を除いた. このようにして作成した問題セットに対する実験の結果, 比が 0.25 以下の問題の精度は共に 86.5 % で, カバー率は新 goo で 56.8 %, 旧 goo で 53.6 % であった. つまり, 精度には影響がなく, カバー率の低下も許容範囲であった. この結果より, 手法 2 は新 goo, 旧 goo のどちらにも依存せず有効であることが確認できた.

これらの結果から, サーチエンジンを変更してもほぼ同等の精度を出すことができると期待できる.

5 考察

次に, 実験結果を詳細に分析し, そこで生じている現象をとらえ, 改善の糸口を探ることを行った. 今回調査対象としたのは, 分析用セット (888 問) を統合手法で解いた結果である.

5.1 提案手法で解くことが困難な問題

まず, 誤りの原因を切り分けるために, 分析用セット (888 問) を統合手法で解いた結果, 不正解であった 190 問を提案手法で解くことが困難な問題とそうでないものに分離する. そして, 解くことが困難な問題を, 最も致命的な原因で分類した. その結果を表 15 に示す. 提案手法で解くことが困難な問題例を以下に示す.

表 15 不正解であった問題の分類

原因	個数	割合
提案手法で解くことが困難な問題	29	15 %
ウェブに知識がない	13	6.8 %
「初」を含む問題	9	5 %
算術問題	6	3 %
否定語を含む問題	1	0.5 %
提案手法で解くことが困難な問題に該当しない問題	161	84.7 %
合計	190	100 %

ウェブに知識がない N25, オリンピックに初めて登場したマスコットの名称は何？

A:ロニ B:ミーシャ C:アミック D:バルディ 正解:バルディ

「初」を含む問題 L11, 日本人選手が、オリンピックで初めてメダルを獲得した競技は何？

A:体操 B:水泳 C:テニス D:陸上競技 正解:テニス

算術問題 C41, 円グラフで25%の割合を示すとき、扇形の中心角は何度？

A:25度 B:45度 C:60度 D:90度 正解:90度

否定語を含む問題 D24, 次のうち、懐かしのアイドル「おニャン子クラブ」のメンバーでなかったのは誰？

A:浅香唯 (あさかゆい) B:工藤静香 C:国生さゆり D:ゆうゆ 正解:浅香唯

提案手法で解くことが困難な問題は29問存在した。ウェブに知識がない問題とは、次の3種類の問題の総称である。1つ目は、ヒット数が極めて小であり、それらのページのどれにもその問題の答えが書いていない問題である。2つ目は、ヒット数が極めて小であるが、それらのページのどれかにその問題の答えが書いてある問題である。3つ目は、ヒット数は十分な数があるが、それらのページのどれにもその問題の答えが書いていない問題である。ウェブに知識がない問題が解けないのはやむをえない。「初」を含む問題は、分析用セット(888問)全体では60問存在し、そのうち40問は正解した。このタイプの問題は、「初」という語を考慮すべき問題と、「初」の考慮が不要な問題に分けられる。問題数は、それぞれ、29個と31個であった。「初」を考慮すべき問題の例を以下に示す。

J10, メジャーリーグプロ野球で、最初に三冠王を獲得したバッターは誰？

A:タイ・カップ B:ペーブ・ルース C:ルー・ゲーリック D:テッド・ウィリアムス

正解:タイ・カップ

上記の問題では、三冠王を獲得したバッターは何人もいるので、「最初」という語を考慮しなければ正解を導くことはできない。次に、「初」の考慮が不要な問題の例を以下に示す。

E20, 1969年に人類初の月面着陸を果たしたのは、アポロ何号？

A:11号 B:13号 C:15号 D:17号 正解:11号

上記の問題では、「初」という語を考慮しなくても、「1969年」というキーワードから一意に正解を導くことができる。正解率は、「初」を考慮すべき問題が48%、「初」の考慮が不要な問題が84%であった。このことから、「初」を含む問題のうち、解くことが困難な問題は、「初」を考慮すべき問題のみであることがわかる。算術問題は連想を利用して解くのは不可能である。このタイプの問題は分析用セット全体では11問存在し、そのうち5問は正解した。否定語を含む問題は、4.1節で述べた、問題文に「ない」を含む問題を除くという方法で除外できなかった問題である。

5.2 正解を選択できるキーワードが存在しない問題

次に、どのようなキーワードを選んでも、手法の特性上解くことができない問題は、どのくらい存在するのかを調べた。分析対象は、不正解の問題のうち、上記の「提案手法で解くことが困難な問題」に属さない161問題で、その内訳は、手法2が59問、手法1が102問である。

手法2において不正解であった問題のうち、FA1位の選択肢が正解となるキーワードが存在する問題は35問(59%)で、存在しない問題は24問(41%)であった。この41%の問題は手法の特性上、手法2では解けない問題である。

同様に、手法1において不正解であった問題のうち、FA1位の選択肢またはBA1位の選択肢の少なくともどちらか一方が正解となるキーワードが存在する問題は80問(78%)で、存在しない問題は22問(22%)であった。この22%の問題は手法の特性上、手法1では解けない問題である。

5.3 キーワードの妥当性評価

次に、不正解の原因を調べるために、提案手法で選ばれたキーワードがどのようなものであったかを、手法1と手法2のそれぞれに対して調査する。分析対象は、分析用セット(888問)を統合手法で解いた結果のうち、不正解であった190問から提案手法で解くことが困難な29問を除いた161問である。その内訳は、手法2が59問、手法1が102問である。なお、比較のため、正解であった問題に対しても同様の調査を行う。内訳は、手法2が466問、手法1が232問である。

キーワードの妥当性の判定は、人手により以下のどれかに分類することによって行った。

妥当 人間が解くときに選ぶであろうキーワードと一致

不当 人間が解くときに選ぶであろうキーワードと一致しない。一致しない原因別に以下の細分類を設ける。

不足 重要な語が不足している

余分 余分な語が含まれている

不足&余分 重要な語が不足し、かつ余分な語が含まれている

複合語の区切り 別々のキーワードであると判定すべき2つの語を複合語と誤認定

キーワード選択不可 手法1では、問題文中に「」が存在する場合、「」中の語のみをキーワードとするが、このとき、「」中の語句がストップワードとして認定された場合、キーワードが1つも選ばれない。手法1に対してのみ判定。

ヒット数の取得ミス 本来ヒット数が0でないはずなのに、サーチエンジンがヒット数として誤って0を返したことが原因で、重要な語がキーワードに選ばれなかった。手法1に対してのみ判定。

ここで、分類の例として、実験結果から、該当する問題と選ばれたキーワード、及び、選ばれた選択肢を示す。なお、「妥当」、「不当(不足)」、「不当(余分)」、「不当(不足&余分)」の例は手法2の結果であり、「不当(複合語の区切り)」は手法1の結果である。

妥当 E56, 1895年、日清戦争の講和会議が行われた都市はどこ？

A:下関 B:南京 C:ベルサイユ D:ポーツマス 正解:下関

選ばれたキーワード: 日清戦争, 講和会議, 選ばれた選択肢: ポーツマス

不当(不足) M18, 作家、渡辺淳一の直木賞受賞作はどれ？

A:「失楽園」 B:「光と影」 C:「かりそめ」 D:「ガラスの結晶」 正解:「光と影」

選ばれたキーワード: 渡辺淳一, 選ばれた選択肢: 失楽園

正解が得られるキーワード: 渡辺淳一, 直木賞受賞作, 作家

不当(余分) J75, アラビア語で「巨人の腋(わき)の下」という意味を持つ、オリオン座のアルファ星は何？

A: リゲル B:シリウス C:プロキオン D:ベテルギウス 正解:ベテルギウス

選ばれたキーワード: 巨人の腋の下, オリオン座, アルファ星, 意味

選ばれた選択肢: プロキオン

正解が得られるキーワード: 巨人の腋の下

不当(不足&余分) C67, 1962年に堀江謙一さんが太平洋単独横断に成功していたときに乗っていたヨットの名前は何？

A: よど号 B:ウラヌス号 C:マーメイド号 D:ロンドンブーツ1号 正解:マーメイド号

選ばれたキーワード: 1962年, 太平洋単独横断, 名前, ヨット, 成功

選ばれた選択肢: よど号

正解が得られるキーワード: 堀江謙一

不当(複合語の区切り) C3, iMAC (アイ・マック) などのマッキントッシュ・シリーズを販売するパソコンメーカーはどれ？

A:ピーチ B:アップル C:オレンジ D:ストロベリー 正解:アップル

選ばれたキーワード: マッキントッシュ・シリーズ

選ばれた選択肢: アップル

正解が得られるキーワード: マッキントッシュ

妥当なキーワードを選択できたかの検討 手法 2, 手法 1, それぞれについて, 妥当なキーワードを選択できたか調べた. 分析対象は, 手法 2 においては, 手法 2 で解けなかった問題から提案手法で解くことが困難な問題を除いた 59 問と, 手法 2 で解くことができた 466 問を合わせた 525 問である. 手法 1 においても同様に, 手法 1 で解けなかった問題から提案手法で解くことが困難な問題を除いた 102 問と, 手法 1 で解くことができた 232 問を合わせた 334 問である. 調査の結果を表 16 に示す. この表より, 以下のことがわかる.

- 妥当なキーワードが選べた問題は, 手法 2 で 51.4 %, 手法 1 で 66.5 %
- 手法 2 で不当の場合は, 余分なキーワードを含む問題 (「余分」及び「不足&余分」の合計) が計 40.9 % と多い
- 手法 1 はキーワードが不足する場合 (「不足」及び「不足&余分」) と余分なキーワードを含む場合 (「余分」及び「不足&余分」) がほぼ同数

表 16 妥当なキーワードを選択できたか

キーワード	手法 2		手法 1	
	個数	割合	個数	割合
妥当	270	51.4 %	222	66.5 %
不当	255	48.6 %	106	31.7 %
不足	(32)	(6.1 %)	(30)	(9.0 %)
余分	(134)	(25.5 %)	(27)	(8.1 %)
不足&余分	(81)	(15.4 %)	(34)	(10.2 %)
複合語の区切り	(8)	(1.5 %)	(15)	(4.5 %)
キーワードの選択不可	-	-	3	0.9 %
ヒット数の取得ミス	-	-	3	0.9 %
合計	525	100 %	334	100 %

キーワードが妥当な場合, 不当な場合の精度に関する検討 キーワードが妥当な場合, 不足がある場合, 及び, 余分なものがある場合のそれぞれに対して, 手法別に精度を求めた. 分析対象は, 以下の問題である.

妥当 表 16 で「妥当」に分類された問題 (手法 2 で 270 問, 手法 1 で 222 問)

不足あり 表 16 で「不足」または「不足&余分」に分類された問題 (手法 2 で 113 問, 手法 1 で 64 問)

余分あり 表 16 で「余分」または「不足&余分」に分類された問題 (手法 2 で 215 問, 手法 1 で 61 問)

分析結果を表 17 に示す. この表より, 以下のことがわかる.

- 手法 2, 手法 1 共に, キーワードが妥当であれば高い精度で解ける

- 手法 2, 手法 1 共に, キーワードが不当の場合は, 余分なものがあるよりも不足がある方が, 精度が悪い

表 17 キーワードが妥当な場合, 及び, 不当な場合の精度

	手法 2			手法 1		
	妥当	不足あり	余分あり	妥当	不足あり	余分あり
解ける	94.4 %	69.0 %	84.2 %	81.1 %	39 %	69 %
解けない	5.6 %	31.0 %	15.8 %	18.9 %	61 %	31 %
合計	100 %	100 %	100 %	100 %	100 %	100 %

5.4 ヒット数と精度の相関の分析

次は, ヒット数と精度の関係について分析する. 手法 2 に対しては, 888 問の分析用セットに統合手法を適用したとき, 手法 2 が適用された問題 536 問すべてを対象とし, 手法 1 に対しては, 同様に手法 1 が適用された問題 352 問すべてを対象とする. 本節では, 結果及び考察の重要な部分だけ述べ, 詳細は付録 C で説明する.

まずは, キーワード集合と選択肢の論理積をクエリーとする検索のヒット数と精度の関係の分析結果について述べる.

手法 2 キーワード集合と選択肢の論理積をクエリーとする検索のヒット数が 10,000 以上となる場合, 精度は 42 % と悪いので, そのキーワード集合と選択肢のペアは選択しないという戦略が考えられる.

手法 1 キーワード集合と選択肢の論理積をクエリーとする検索のヒット数が 10,000 以上となる場合, 精度が 52 % 以下と悪いので, キーワード選択をやり直すという戦略が考えられる.

次に, キーワード集合の論理積をクエリーとする検索のヒット数と精度の関係の分析結果について述べる.

手法 2 キーワード集合の論理積をクエリーとする検索のヒット数が 100,000 以上となる場合, 精度が 56 % と悪いので, そのキーワード集合と選択肢のペアは選択しないという戦略が考えられる.

手法 1 キーワード集合の論理積をクエリーとする検索のヒット数が 1,000,000 以上または 10 未満となる場合, 精度は 50 % 以下と悪いので, キーワード選択をやり直すという戦略が考えられる.

一方, 選ばれた選択肢をクエリーとする検索のヒット数と精度の関係からは, 特筆すべき現象は確認できなかった.

6 関連研究

本章では, まず, ウェブを利用した関連研究を取り上げ, 本論文との違いを述べる. チュートリアル (Lin and Katz 2003) を参考にすると, ウェブを利用した質問応答システムは, 知識アノテーション型と知識マイニング型の2つに分けることができる.

知識アノテーション型とは, ウェブ上のテキストを構造化して, データベースのように利用する技術に基づく. このタイプのシステムは, データの扱い方により, さらに2つのタイプに分類できる. 一つは, 映画データベースや人物データベースのウェブページなど良質のウェブページをあらかじめ探しておき, このようなページに対してSQLのようなデータベース制御言語でアクセスできるようにするためのラッパーを用意するものである. もう一つは, ウェブから知識を収集してローカルデータベースを作成するというものである. 知識アノテーション型のシステムとして, “START and Omnibase” (Katz, Felshin, Yuret, Ibrahim, Lin, Marton, McFarland, and Temelkuran 2002) がある. これは, ウェブ上にある, “biography.com” (人名に関する情報), “World Factbook” (CIA による国に関する情報), “Merriam-Webster” (英英辞典), “IMDb” (映画のデータベース) などを, 専用のクエリー言語で検索できるようなラッパーを準備するという研究である. このような方法はあらかじめ想定したタイプの問題であれば, 精度よく解くことができるが, 用意されていない知識に関する問題は解くことができない. なお, 知識アノテーションの技術と後述の知識マイニングの技術を併用する研究も存在する (Lin, Fernandes, Katz, Marton, and Tellex 2002).

一方, 知識マイニング型とは, 構造化されていないテキストを効率的に利用する技術に基づく. 本論文で提案した手法はウェブをそのままの形で利用するので, このタイプに属する. 以下このタイプの研究について述べる.

(Brill, Lin, Banko, Dumais, and Ng 2001) によって提案された方法は, サーチエンジンの検索結果のサマリーの部分から解候補を抽出しスコアリングして解を決めるというものである. まず, 質問文を平叙文に直し, 「書き換え」と呼ばれる〈文字列, $L/R/-$, 重み〉の3つ組を生成する. 第2項は解の出現場所を示す. 次に, 「書き換え」をクエリーとしてGoogleで検索し, 検索結果上位100件のサマリーから n -gram を収集しスコアを与える. スコアは出現文書数と「書き換え」の重みの積で計算する. その後, 解のタイプを考慮してスコアを修正し, 重複している n -gram をまとめることを行う. 最後に, 得られた解を含む文書をコーパスから探す. この手法は, ウェブの知識を利用するためにサーチエンジンを使用するが, 本論文で提案した手法のようにサーチエンジンのヒット数は利用しない.

(Magnini et al. 2002) は我々の解選択法と同様に, サーチエンジンのヒット数を使う解の選択法を提案した. 彼らは AltaVista の OR と NEAR の演算子を使ってサーチエンジンのクエリーを構成する. Magnini らの手法と我々の手法の主要な違いは, キーワードの選択法にある. Magnini らの手法において, キーワード集合の初期候補は質問文中にあるすべての内容語

を含む。キーワード集合のヒット数がある閾値以下であれば、人手で作成した少数の規則に従い、キーワード集合から最も重要でない語を除く。この手続きをキーワードのヒット数が閾値を超えるまで繰り返す。Magnini らのキーワード選択法は、我々の手法 1 の語の重みに基づくキーワード選択法と似た方法である。これに対して、我々は、手法 2 において、問題文と解候補の両方を考慮してキーワードを選択する方法も提案した点が異なる。また、Magnini らの実験では、解候補にスコアをつける 3 つの尺度のうち、我々の FA に近い “Corrected Conditional Probability” という尺度が最もよい性能であった。我々もこの尺度を実装し、分析用セットに対して実験したところ、3.3 節で示した統合手法よりも 4.5% 低い精度であった。この精度の差は、母比率の検定により、有意水準 5% で統計的に有意である。また、解くことが困難な問題の主要な割合を占める、「初」を考慮すべき問題 29 問の精度は 55% と、提案手法の精度 (48%) より若干高かったが、母比率の検定により、有意水準 5% で統計的に有意でない。したがって、解くことが困難な問題に対する精度に違いがあるとはいえない。この結果から、我々の実装では、Magnini らの提案する “Corrected Conditional Probability” よりも、我々の用いた FA 及び BA を併用する方法の方が有効であることがわかった。

次に、ウェブページ本文を利用する方法と、本論文の手法を直接比較する。本論文の方法は、サーチエンジンのヒット数のみを用いており、サーチエンジンによって得られるウェブページ本文は用いていない。一方、(外池, 佐藤 2003) は、サーチエンジンを利用してウェブページを 25 ページ取得し、得られたウェブページ本文における、キーワードと解候補の近接の頻度を利用する方法で解を決定する。この方法と本論文の方法を、(外池, 佐藤 2003) で用いた評価用セットを用いて比較したところ、本論文の方法の方が 1.6% 精度が高かった。この精度の差は、母比率の検定により、有意水準 5% で統計的に有意であるとは言えないが、本論文で提案したサーチエンジンのヒット数のみを用いる方法は、ウェブページ本文を取得し、キーワードと解候補の近接の頻度を利用する方法と同程度の性能を達成できることがわかった。なお、(外池, 佐藤 2003) では、本論文の手法 1 における、単純にヒット数 1 位の選択肢を選ぶ手法を変形させることにより、本論文で除外した、否定語「ない」を含む問題を約 40% の精度で解ける方法も提案した。この方法は、本論文で提案したシステムにも適用できると考えられる。

次に、本論文で用いた尺度が、これまでに提案されたコーパス中の共起頻度を利用した様々な尺度に対して、どのように位置付けられるのかを述べる。コーパス中の共起頻度を利用して、2 語の関係を調べる研究は数多くなされてきた (Manning and Schütze 1999)(Matsumoto and Utsuro 2000)(宇津呂 1999)。相互情報量, t 検定, χ^2 検定, ϕ^2 統計値, Dice coefficient, など、2 語を対等に扱う尺度である。その中でも、相互情報量に関しては、(Magnini et al. 2002) において、質問応答ではうまく機能しないことが報告されている。質問応答では、2.3 節で述べたように、前向きと後ろ向きの両方の連想を考慮する必要がある。しかし、これら 2 つの連想をうまく一つの尺度にまとめるのは難しい。Association Score や Maximal Likelihood Ratio,

(Magnini et al. 2002) の Corrected Conditional Probability など, 非対称の尺度も提案されているが, 片方向の連想しか考慮していない. その中でも, Maximal Likelihood Ratio に関しては, (Magnini et al. 2002) において, 質問応答における解の検証タスクにおいては, 共起を発見する古典的なタスクと違い, 相互情報量を用いた場合と同程度の精度しか得られないことが報告されている. そこで, 本論文では, FA と BA という 2 つの簡単な尺度に, 解選択ルールや連想の強さの比というチューニングを施すことによって, 期待する効果が得られることを実証した. また, 本論文で用いた尺度は, 共起に基づく条件付きの単語出現確率と同じ形の式であるが, 2 つの語が共起した回数ではなく, 2 つの語が共起するウェブページ数を数えている点が異なる. すべてのウェブページにおいて, 2 つの語が共起した回数を数えるのは困難であるので, 本尺度を採用した.

最後に紹介するのが, TREC QA Task(Voorhees 2003) において毎年, トップクラスの成績を納めている, Moldovan (Moldovan, Harabagiu, Girju, Morarescu, and Lacatusu 2002) らの研究である. Moldovan (Moldovan et al. 2002) らの解の選択法は, テキストの深い理解に基づく方法である. 深い理解に基づく解の選択法では, まず, 問題文, および, 解候補を含むパラグラフを論理表現に変換する. そして, 推論の操作によって両者が対応しているかを調べる. この方法の欠点の一つは, WordNet などの語彙的知識と, 推論ルールセットなどの世界知識という豊富な情報を必要とすることである. このため, WordNet に知識がないような専門分野の問題には対応できない. また, 言語解析システムや英語における WordNet に相当する語彙的知識が整備されていない言語にも対応することはできない. これに対して, 本論文で提案した方法は, 自然言語処理のために設計された知識源が存在しない専門分野の問題や, 言語解析システムが実現されていない言語に対応できる.

7 結論と今後の課題

本稿では, 問題文から抽出されたキーワードと解候補の間の共起に基づく語彙的な関係(連想)の強さに基づく解の選択法を 2 つ提案し, 両者の統合を行った. 手法 1 は, 語の重みに基づいて適切なキーワードを選択し, そのキーワードを利用してサーチエンジンのヒット数を求めることにより, キーワードと選択肢の連想の強さを求め, $FA1$ 位の選択肢と $BA1$ 位の選択肢のどちらかを解選択ルールによって決める手法である. 手法 2 は, 連想の強さに基づいて, 最適なキーワードと解候補を同時に選ぶという手法である. そして, 統合手法は, まず手法 2 を適用し, そこで十分なスコアが得られない問題には手法 1 を適用するというものである.

2.4 節で提示した 2 つの問いに対する本論文の答えは, まず, キーワード選択に関しては, 手法 1 の語の重みに基づくキーワード選択法と手法 2 の連想の強さに基づいて解候補と同時に選択する手法を提示した. そして, $FA1$ 位の選択肢と $BA1$ 位の選択肢のどちらを選ぶのかについては, 手法 1 で解選択ルールを提示した. 手法 2 に関しては, $FA1$ 位の解を選ぶ尺度を採用

した。評価実験の結果、4択クイズ「クイズ\$ミリオネア」の問題のうち、連想でとける可能性のない否定及び比較の表現を含む問題を除いた問題は全体の91%であり、このうち79%の問題を解くことができた。そして、上記の方法でFAとBAという性質の異なる2つの尺度を組み合わせて用いる提案手法が、(Magnini et al. 2002)の手法の性能を有意に改善できることを示した。質問応答の研究分野全体における本論文の貢献は、FAとBAという意味付けのしやすい共起に基づく尺度を「解選択ルール」または「連想の強さの比」によってチューニングすることにより、解選択ができることを示したことと、キーワード選択も連想の強さを利用して行うことができることを示したことである。

今後の課題について述べる。まず、手法1におけるキーワード選択における重みや解選択ルールにおける閾値、統合手法における連想の強さの比の閾値を、機械学習により調整することが挙げられる。次に、4択の与えられない一般的な質問応答への応用が挙げられる。実際の質問応答における解候補の抽出においては、4択クイズの選択肢のような解候補が抽出されるとは限らない。まず、解候補が大量に抽出される場合が考えられる。このような場合は、連想の強さを表すFAまたはBAの絶対値が小さい解候補を排除することにより、個数を絞り込むことができると考えられる。抽出された解候補の中に正解がない場合においても、同様に、連想の強さの絶対値がある値を越えなければ選択しないという条件を加えておくことにより、答えを選ばないことが求められる状況にも対応できると考えられる。実際、このような方法は、(Magnini et al. 2002)においても実現されている。また、実際の解候補抽出においては、解候補を探すと同時にスコアを計算してしまう場合も多いが、この時に求められたスコアと、本手法を併用し精度を上げるといった方法も考えられる。

参考文献

- Brill, E., Lin, J., Banko, M., Dumais, S., and Ng, A. (2001). "Data-Intensive Question Answering." In *Proc. TREC 2001*, p. 393.
- Clarke, C. L. A., Cormack, G. V., and Lynam, T. R. (2001). "Exploiting Redundancy in Question Answering." In *Proceedings of SIGIR'01*.
- Fukumoto, J., Kato, T., and Masui, F. (2004). "Question Answering Challenge for Five ranked answers and List answers – Overview of NTCIR4 QAC2 Subtask 1 and 2 –." In *Proc. 4th NTCIR Workshop Meeting*.
- Katz, B., Felshin, S., Yuret, D., Ibrahim, A., Lin, J., Marton, G., McFarland, A. J., and Temelkuran, B. (2002). "Omnibase: Uniform Access to Heterogeneous Data for Question Answering." In *Proc. NLDB 2002*.
- クイズ\$ミリオネア (編) (2002). 完全攻略! クイズ\$ミリオネア. フジテレビ出版.
- 黒橋禎夫 (1998). 日本語構文解析システム KNP version 2.0 b6 使用説明書.

- 黒橋禎夫, 長尾真 (1999). 日本語形態素解析システム JUMAN version 3.62 使用説明書.
- Kwok, C. C. T., Etzioni, O., and Weld, D. S. (2001). “Scaling Question Answering to the Web.” In *Proc. the 10th WWW Conf.*, pp. 150–161.
- Lin, J., Fernandes, A., Katz, B., Marton, G., and Tellex, S. (2002). “Extracting Answers from the Web Using Knowledge Annotation and Knowledge Mining Techniques.” In *Proc. TREC 2002*, p. 447.
- Lin, J. and Katz, B. (2003). “Question Answering Techniques for the World Wide Web.” Tutorial presentation at The 11th Conference of the European Chapter of the Association of Computational Linguistics(EACL-2003).
- Magnini, B., Negri, M., Prevete, R., and Tanev, H. (2002). “Is It the Right Answer? Exploiting Web Redundancy for Answer Validation.” In *Proc. 40th ACL*, pp. 425–432.
- Manning, C. D. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. The MIT Press.
- Matsumoto, Y. and Utsuro, T. (2000). “Lexical Knowledge Acquisition.” In Dale, R., Moisl, H., and Somers, H. (Eds.), *Handbook of Natural Language Processing*, chap. 24, pp. 563–610. Marcel Dekker Inc.
- Moldovan, D., Harabagiu, S., Girju, R., Morarescu, P., and Lacatusu, F. (2002). “LCC Tools for Question Answering.” In *Proc. TREC 2002*, p. 388.
- Sato, S. and Sasaki, Y. (2003). “Automatic Collection of Related Terms from the Web.” In *ACL-03 Companion Volume to the Proceedings of the Conference*, pp. 121–124.
- 外池昌嗣, 佐藤理史 (2003). “ウェブを用いて4択クイズを解く.” 言語処理学会第9回年次大会発表論文集, pp. 641–644.
- 宇津呂武仁 (1999). “知識の獲得.” 田中穂積 (編), 自然言語処理 — 基礎と応用 —, 6.1 節, pp. 185–204. 電子情報通信学会.
- Voorhees, E. M. (2003). “Overview of the TREC 2003 Question Answering Track.” In *Proc. TREC 2003*, p. 54.

付録

A 解選択ルールに関する検討

本章では、3.1.2節の解選択ルールに関して、 c_1^F (FA1位の選択肢)と c_1^B (BA1位の選択肢)のどちらを選ばよいか決定するための指標について検討する。3.1.2節で提案した手法では、 c_1^F 及び c_1^B 以外の選択肢は全く考慮しない。しかし、 $F(c_2^F)$ (ただし、 c_2^F はFA2位の選択肢)と $F(c_1^F)$ を比較することにより c_1^F の信頼性を評価したり、同様に $B(c_2^B)$ と $B(c_1^B)$ を比較することにより c_1^B の信頼性を評価するといった尺度も考えられる。

そこで、3.1.2節で提案した手法に対する比較手法として、 $F(c_1^F)$ と $F(c_2^F)$ の比、及び、 $B(c_1^B)$ と $B(c_2^B)$ の比を用いる手法を設計し、提案手法との比較を行った。ただし、条件を同じにするために提案手法の簡略化を行った。以下に、比較手法、及び、提案手法の簡略版のルールを示す。

比較手法 $F(c_1^F)$ と $F(c_2^F)$ の離れ方が、 $B(c_1^B)$ と $B(c_2^B)$ の離れ方より大きい場合は、 c_1^F を選び、そうでない場合は、 c_1^B を選ぶ手法。ルールを以下に示す。

- (1) if $\frac{F(c_2^F)}{F(c_1^F)} \leq \frac{B(c_2^B)}{B(c_1^B)}$ then c_1^F
- (2) else c_1^B

提案手法の簡略版 $F(c_1^F)$ と $F(c_1^B)$ の離れ方が、 $B(c_1^F)$ と $B(c_1^B)$ の離れ方より大きい場合は、 c_1^F を選び、そうでない場合は、 c_1^B を選ぶ手法。ルールを以下に示す。

- (1) if $\frac{F(c_1^B)}{F(c_1^F)} \leq \frac{B(c_1^F)}{B(c_1^B)}$ then c_1^F
- (2) else c_1^B

比較のため、元の提案手法及び、単純に c_1^F を選ぶ手法、 c_1^B を選ぶ手法に対しても調べた。問題セットには、888問からなる分析用セットのうち、手法1を適用したとき c_1^F または c_1^B のどちらか一方のみが正解となる271問を用いた。

実験の結果を表18に示す。提案手法の簡略版の精度は69.7%であったのに対し、比較手法の精度は61.3%と低かった。この結果より、 c_2^F や c_2^B も考慮するよりは、 c_1^F と c_1^B のみを考慮する方が有効であることが分かった。

さらに、提案手法の簡略版で解けた問題を調べると、その92.1%の問題は提案手法でも解けることがわかった。したがって、提案手法の簡略版は、提案手法をうまく簡略化できていると言える。

B 連想による解の選択に関する検討

本章では、3.2節で述べるができなかった、連想に基づく単純なキーワード選択法をいくつか説明する。

表 18 解選択ルールの比較

手法	正解数	精度
提案手法	207	76.4 %
提案手法の簡略版	189	69.7 %
比較手法	166	61.3 %
c_1^F を選択	150	55.3 %
c_1^B を選択	121	44.6 %

表 19 最大の FA を持つ (キーワード集合, 選択肢) の組を選ぶ方法

(キーワード集合, 選択肢) の組	FA
({"アメリカングラフィティ"}, "ジョージ・ルーカス")	0.165
({"アメリカングラフィティ"}, "スティーブン・スピルバーグ")	0.0569
({"アメリカングラフィティ"}, "フランシス・フォード・コッポラ")	0.0408
({"アメリカングラフィティ"}, "黒澤明")	0.0341
({"アメリカングラフィティ", "監督"}, "ジョージ・ルーカス")	0.457
({"アメリカングラフィティ", "監督"}, "スティーブン・スピルバーグ")	0.148
({"アメリカングラフィティ", "監督"}, "フランシス・フォード・コッポラ")	0.135
({"アメリカングラフィティ", "監督"}, "黒澤明")	0.0876
({"監督"}, "ジョージ・ルーカス")	0.00604
({"監督"}, "スティーブン・スピルバーグ")	0.00813
({"監督"}, "フランシス・フォード・コッポラ")	0.00252
({"監督"}, "黒澤明")	0.00809

最初の手法は, 以下のように最大の FA を持つキーワード集合と選択肢の組 $\langle \hat{K}, \hat{c} \rangle$ を選ぶものである.

$$\langle \hat{K}, \hat{c} \rangle = \underset{c, K \in 2^N}{\operatorname{argmax}} FA(K, c) \quad (19)$$

これは, FA が最も大きいキーワード集合と選択肢の組が最も連想が強い組であろうという仮説に基づき, その組の選択肢を選ぶ方法である. 以上の方法の適用例を示す. 「Q1: 「アメリカングラフィティ」の監督は誰?」という問題であれば, キーワード集合と選択肢の組, 及びその FA は表 19 のようになる. したがって, ここでは, 最大の FA を持つ, ({"アメリカングラフィティ", "監督"}, "ジョージ・ルーカス") が選ばれ, 選ばれた解は「ジョージ・ルーカス」となる. 分析用セットに対して, この手法で選ばれた選択肢の精度を計算すると 71.3 % であった.

同様に, 最大の BA または最大のヒット数を持つキーワード集合と選択肢の組を選ぶ別の手法を以下に示す.

$$\langle \hat{K}, \hat{c} \rangle = \underset{c, K \in 2^N}{\operatorname{argmax}} BA(K, c) \quad (20)$$

$$\langle \hat{K}, \hat{c} \rangle = \underset{c, K \in 2^N}{\operatorname{argmax}} hits(K \cup \{c\}) \quad (21)$$

精度は分析用セットに対して, それぞれ, 36.1 % と 35.7 % であった. これまでに示した 3 つの

表 20 FA/BA /ヒット数が最大のキーワード集合と選択肢の組を選ぶ手法

手法	精度
FA	71.3 %
BA	36.1 %
ヒット数	35.7 %

表 21 FA 最大となる選択肢と BA 最大となる選択肢が一致するキーワード集合の集合 $\hat{S}$ の分析

条件	問題数
$ \hat{S} = 1$	70 (7.9 %)
$ \hat{S} \geq 2$	700 (78.8 %)
$ \hat{S} = 0$	118 (13.3 %)
合計	888 (100 %)

手法を分析用セットに適用した結果を表 20 にまとめる.

次に, あるキーワード集合のもとで, FA が最大となる選択肢と BA が最大となる選択肢が一致するとき, そのキーワード集合は信頼できると仮定し, そのようなキーワード集合を選択する手法について検討する. この手法を数式化したものを以下に示す. ここで, $\hat{S}$ を選択されたキーワード集合の集合とする.

$$\hat{S} = \{K | \hat{c}_1 = \operatorname{argmax}_c FA(K, c), \hat{c}_2 = \operatorname{argmax}_c BA(K, c), \hat{c}_1 = \hat{c}_2\} \quad (22)$$

分析用セットのそれぞれの問題に対して, この方法で選択されたキーワード集合の個数を調べた結果を表 21 に示す. 表の「条件」は, FA 最大の選択肢と BA 最大の選択肢が一致するキーワード集合の個数に関する条件である. この結果, 選択されたキーワード集合が 1 つであった問題は 70 問 (7.9 %) であった. 一方, 2 つ以上のキーワード集合が選択された問題は 700 問 (78.8 %) であった. このうち, それぞれのキーワード集合と同時に選ばれた選択肢がすべて一致した問題は 415 問で, 全体の 46.7 % (415/888) であった. 選択されたキーワード集合が 1 つであった場合 (70 問) と, 選択されたキーワード集合は 2 つ以上であったが選ばれた選択肢がすべて一致した場合 (415 問) は, 選択肢の中から答えを 1 つに決めることができる. 答えを 1 つに決めることができた 485 問 (全体の 54.6 %) の精度は 82.1 % (398/485) であった. したがって, この方法は精度は高いものの, 全体の約半分程度の問題に対してしか, 答えを決めることができない.

さらに条件を厳しくして, 最大の FA を持つキーワード集合と選択肢の組と, 最大の BA を持つキーワード集合と選択肢の組が一致する場合のみ, そのキーワード集合を採用するという

表 22 最大の FA を持つキーワード集合と選択肢の組と
最大の BA を持つキーワード集合と選択肢の組が一致する場合の分析

条件	問題数
$ \hat{S} = 1$	13 (1.5 %)
$ \hat{S} = 0$	875 (98.5 %)
合計	888 (100 %)

方法も調査した。この方法を以下に式で示す。

$$\hat{S} = \{\hat{K} | \langle \hat{K}, \hat{c}_1 \rangle = \underset{c, K \in 2^N}{\operatorname{argmax}} FA(K, c), \langle \hat{K}, \hat{c}_2 \rangle = \underset{c, K \in 2^N}{\operatorname{argmax}} BA(K, c), \langle \hat{K}, \hat{c}_1 \rangle = \langle \hat{K}, \hat{c}_2 \rangle\} \quad (23)$$

この方法でどの程度キーワード集合を選べるのかを分析用セットを対象に調べた結果を表 22 に示す。しかしながら、わずかに 13 問しか一致しなかった。

したがって、このような手法では、キーワード集合を選ぶ条件が厳しすぎるため、適用できる問題が限られてしまう。

C ヒット数と精度の相関の分析

本章では、5.4 節で述べた、ヒット数と精度の相関に関する分析の詳細を説明する。手法 2 に対しては、888 問の分析用セットに統合手法を適用したとき、手法 2 が適用された問題 536 問すべてを対象とし、手法 1 に対しては、同様に手法 1 が適用された問題 352 問すべてを対象とする。ここで、手法 1 が適用されるのは手法 2 が適用できなかった問題であるため、比較のために、分析用セットの 888 問すべてに対して手法 1 を適用した結果も併せて示す。

キーワード集合及び選択肢の論理積をクエリーとする場合 選ばれたキーワード集合及び選択肢の論理積をクエリーとする検索のヒット数と精度の関係を表 23 に示す。ここで、カバレッジは、対象となる手法が適用された問題のうち、ヒット数がある範囲にある問題の割合を示し、精度はヒット数がある範囲にある問題の精度を示す。

まず、カバレッジに関して検討すると、以下のことがわかる。

- 手法 2 はヒット数が 1 以上 1,000 未満の問題が全体の 86.9 % を占める。
- 手法 1 はヒット数が 10 以上 10,000 未満の問題が全体の 86.1 % 以上 (352 問で 86.1 %, 888 問で 88.3 %) を占める。

手法 2 と手法 1 のカバレッジの傾向に、このようなずれがある原因は、手法 1 ではキーワードの個数を制限しているのに対して、手法 2 では制限していないのが原因と考えられる。

次に、ヒット数の範囲と精度に関して検討すると、以下のことがわかる。

- ヒット数が 10,000 以上の問題は、手法 2 が 42 %, 手法 1 が 52 % 以下 (352 問で 35 %, 888 問で 52 %) と、共に精度が悪い。手法 2 においてヒット数が 10,000 以上の問題を例にと

表 23 キーワード集合と選択肢の論理積をクエリとする場合

ヒット数の範囲	手法 2 (536 問)		手法 1 (352 問)		手法 1 (888 問)	
	カバレッジ	精度	カバレッジ	精度	カバレッジ	精度
10,000 ≤ ヒット数	2.2%	42%	6.0%	35%	3.5%	52%
1,000 ≤ ヒット数 < 10,000	10.8%	81%	20.7%	64%	15.1%	72%
100 ≤ ヒット数 < 1,000	25.0%	90.3%	37.8%	71%	40.1%	81.5%
10 ≤ ヒット数 < 100	35.3%	89.9%	27.6%	71%	33.1%	80.3%
1 ≤ ヒット数 < 10	26.7%	86.0%	6.3%	59%	7.1%	76%
ヒット数 = 0	0%		1.7%	0%	1.0%	0%
合計	100%	86.9%	100%	65.9%	100%	77.3%

表 24 キーワード集合の論理積をクエリとする場合

ヒット数の範囲	手法 2 (536 問)		手法 1 (352 問)		手法 1 (888 問)	
	カバレッジ	精度	カバレッジ	精度	カバレッジ	精度
1,000,000 ≤ ヒット数	0.7%	50%	2.8%	33%	1.5%	23%
100,000 ≤ ヒット数 < 1,000,000	3.9%	57%	6.5%	65%	4.7%	76%
10,000 ≤ ヒット数 < 100,000	9.0%	88%	22.4%	62%	18.0%	71.3%
1,000 ≤ ヒット数 < 10,000	17.2%	89%	29.0%	63%	28.3%	74.5%
100 ≤ ヒット数 < 1,000	27.2%	87.0%	27.0%	78%	31.8%	85.5%
10 ≤ ヒット数 < 100	31.2%	91.6%	10.8%	63%	15.0%	81.2%
1 ≤ ヒット数 < 10	9.7%	88%	0.6%	50%	0.3%	33%
ヒット数 = 0	1.1%	33%	0.9%	0%	0.5%	0%
合計	100%	86.9%	100%	65.9%	100%	77.3%

ると、選ばれたキーワード集合は、答えと関係のある語は選ばれずに、「階」や「作品」という一般的な語が選ばれていた。

- ヒット数が 10 以上 1,000 未満の問題は、手法 2 が 90.0%，手法 1 が 70.9% 以上 (352 問で 70.9%，888 問で 80.9%) と、共に精度がよい。
- ヒット数が 1,000 以上 10,000 未満の問題は、手法 2、手法 1 共に上記の範囲の問題よりも精度が 10% 程度低下する。

以上の分析より、精度向上の為には、次に示す戦略が考えられる。

手法 2 キーワード集合と選択肢の論理積をクエリーとする検索のヒット数が 10,000 以上となる場合、そのキーワード集合と選択肢のペアは選択しない

手法 1 キーワード集合と選択肢の論理積をクエリーとする検索のヒット数が 10,000 以上となる場合、キーワード選択をやり直す

キーワード集合の論理積をクエリーとする場合 選ばれたキーワード集合の論理積をクエリーとする検索のヒット数と精度の関係を表 24 に示す。

まず、カバレッジに関して検討すると、以下のことがわかる。

- 手法 2 では、ヒット数が 1 以上 100,000 未満の問題が全体の 94.2% を占める。

- 手法1では、ヒット数が10以上100,000未満の問題が全体の89.2%以上(352問で89.2%, 888問で93.1%)を占める。

「選ばれたキーワード集合及び選択肢の論理積をクエリーとする検索のヒット数」と同様、手法2では手法1に比べてヒット数が小さい問題が多く見受けられた。

ヒット数の範囲と精度に関して検討すると、以下のことがわかる。

- 手法2では、ヒット数が100,000以上の問題で精度が56%と悪い一方、手法1では、キーワード集合のヒット数が1,000,000以上の問題で精度が33%以下(352問で33%, 888問で23%)と悪い。
- 手法2では、ヒット数が100,000より小さい問題では精度は約90%とほぼ一定である一方、手法1では、ヒット数が100以上1,000未満のとき精度は78%以上(352問で78%, 888問で85.5%)と最もよい。
- ヒット数が10未満の問題では、手法1の精度は50%以下(352問で50%, 888問で33%)と悪い。

以上の分析より、精度向上の為には、次に示す戦略が考えられる。

手法2 キーワード集合の論理積をクエリーとする検索のヒット数が100,000以上となる場合、そのキーワード集合と選択肢のペアは選択しない

手法1 キーワード集合の論理積をクエリーとする検索のヒット数が1,000,000以上または10未満となる場合、キーワード選択をやり直す

なお、手法2において、キーワード集合の論理積をクエリーとする検索のヒット数が0であるにもかかわらず正解している問題が存在するのは、以下の2つの要因による。

- これらの問題において、キーワード集合の論理積をクエリーとする検索のヒット数が0となったのは、ヒット数の取得ミスが原因である。
- 比の計算をするときには、キーワード集合の論理積をクエリーとする検索のヒット数は不要であるため、このような場合でも比の値が計算できるので、選択肢を選ぶことができる。

選択肢をクエリーとする場合 選ばれた選択肢をクエリーとする検索のヒット数と精度の関係を表25に示す。

まず、カバレッジに関して検討すると、以下のことがわかる。

- どちらの手法が適用された問題も、90%近くの問題は、ヒット数が100以上1,000,000未満の範囲にある。
- ヒット数が0以上100未満の選択肢が選ばれた問題は、いずれの手法においても2%程度とわずかであった。このような選択肢には、「床屋さんの看板」などがあった。

次に、ヒット数の範囲と精度に関して検討すると、以下のことがわかる。

表 25 選択肢をクエリとする場合

ヒット数の範囲	手法 2 (536 問)		手法 1 (352 問)		手法 1 (888 問)	
	カバレッジ	精度	カバレッジ	精度	カバレッジ	精度
1,000,000 ≤ ヒット数	0.7 %	25 %	9.1 %	50 %	4.2 %	45 %
100,000 ≤ ヒット数 < 1,000,000	17.4 %	87 %	32.1 %	67 %	24.0 %	73.2 %
10,000 ≤ ヒット数 < 100,000	34.3 %	87.5 %	31.5 %	68 %	32.9 %	79.1 %
1,000 ≤ ヒット数 < 10,000	34.7 %	87.1 %	19.3 %	71 %	28.3 %	83.3 %
100 ≤ ヒット数 < 1,000	11.4 %	90 %	5.4 %	68 %	8.6 %	84.2 %
10 ≤ ヒット数 < 100	1.1 %	83 %	1.4 %	60 %	1.0 %	78 %
1 ≤ ヒット数 < 10	0.2 %	100 %	0.3 %	100 %	0.3 %	67 %
ヒット数 = 0	0.2 %	0 %	0.9 %	0 %	0.7 %	0 %
合計	100 %	86.9 %	100 %	65.9 %	100 %	77.3 %

- ヒット数が 1,000,000 以上の問題は、手法 2 では 25 %、手法 1 では 50 %以下 (352 問で 50 %、888 問で 45 %) と、共に精度が悪い。このような問題の一部は、選択肢が一般的な語となっているので、提案手法で解くのは困難である。このような選択肢には、「A」(A20, 地下を表すアルファベットを選ばせる問題), 「事件」(O3, 映画の名前を選ばせる問題), 「や」(D47, 相撲の休場を表すひらがなを選ばせる問題) などがあつた。

略歴

外池 昌嗣: 2001 年京都大学 工学部 情報学科卒業。2003 年 同大学大学院情報学研究科修士課程 知能情報学専攻修了。現在、同大学大学院情報学研究科博士後期課程在学中。自然言語処理の研究に従事。

宇津呂 武仁: 1989 年京都大学工学部 電気工学第二学科 卒業。1994 年同大学大学院工学研究科 博士課程電気工学第二専攻 修了。京都大学博士(工学)。奈良先端科学技術大学院大学情報科学研究科助手、豊橋技術科学大学工学部情報工学系講師を経て、2003 年より 京都大学 情報学研究科 知能情報学専攻講師。自然言語処理の研究に従事。

佐藤 理史: 1983 年京都大学工学部電気工学第二学科卒業。1988 年同大学院博士課程研究指導認定退学。京都大学工学部助手、北陸先端科学技術大学院大学情報科学研究科助教授、京都大学情報学研究科助教授を経て、2005 年より名古屋大学大学院工学研究科教授。工学博士。自然言語処理、情報の自動編集等の研究に従事。

(2004 年 10 月 14 日 受付)

(2005 年 2 月 12 日 再受付)

(2005 年 6 月 18 日 再々受付)

(2005 年 7 月 13 日 採録)