WWW(World Wide Web) を利用した QA タスクにおける回答絞りこみ技術

石田　健二†　　桝井　文人†　　河合　敦夫†　　井須　尚紀†
† 三重大学工学部情報工学科
E-mail: †{ishida,masui,kawai,isu}@ai.info.mie-u.ac.jp

1. はじめに

質問応答技術が注目を集めている．質問応答とは，自然言語によって与えられた質問に対して，答えそのものを出力する技術である．例えば,「日本で一番高い山は何ですか.」という質問に対し,「富士山」と回答する技術である．現状の質問応答技術は，いくつかのステップを経て，大量の文書情報から上のような回答を見つけ出すことを基本動作としている．そのステップとは，主に以下のようなものである．
まず，与えられた質問を解析して，答えるべき回答の種類やトピックを推論する．次に，質問文解析の結果得た情報を利用して，大量文書を絞りこむ．さらに，絞りこんだ文書情報を詳しく解析して回答候補を抽出し，回答としての「尤もらしさ」に基づいて，回答を決定する．

このとき，回答絞り込みには，大まかに二種類の手法が存在する．一方は解析に基づく手法 (解析ベース) であり，他方は統計に基づく手法 (統計ベース) である．前者の例として，宮口ら [3] は質問文と構文構造の類似性と，回答候補からの重要語の係り受け階層距離を用いて，回答を絞り込む手法を提案している．後者の例として，李ら [4] は，意味的パターンを用いて回答候補を抽出し，回答候補と意味的パターンに一致した語との距離にスコアを与え，回答を絞り込む手法を提案している．

ここで，回答絞り込みの問題は「複数回答候補からの選択肢問題」と見なすことができる．外池ら [5] は WWW 検索エンジンの検索ヒット数から、4 択クイズの重要語と回答候補の関係の強さを評価し、その関係の強さが最も強いものを解として選択する手法を提案している．

しかしながら,「必ず 1 つの正解を含む 4 つの回答候補」の中から尤もらしい回答を 1 つ選択する 4 択クイズに対して，質問応答では,「質問応答システムが大量文書中から抽出した回答候補」の中から回答を選択する必要があり，選択肢問題としては，より複雑である．これらの回答候補の中に，正解が複数個存在する場合もあれば，正解が全く存在しない場合もある．これらの場合，4 択クイズのように「回答候補の中から尤もらしい回答を 1 つ選択する」だけでは不十分である．

そこで本論文では，質問応答における選択肢問題として回答を絞り込む手法を提案する．まず，WWW 検索エンジンの検索ヒット数から質問文と回答候補の関係の強さを評価する．次に，その関係の強さが適当な閾値以下の回答候補を取り除き，尤もらしい回答のみを答える．この方法は,「質問文との関係が弱い回答候補」より,「質問文との関係が強い回答候補」の方が質問に対する正しい回答である傾向が強いという前提に基づいている．

これにより,「尤もらしい回答候補を全てを答える」ことや「『回答候補中に正解がない』ことを答える」ことが可能となる．

以下，2 章で質問応答システムと質問応答タスクについて説明する．3 章で本論文で提案する回答絞り込み手法について説明する．4 章で適当な閾値の選択手法について説明する．5 章で実験環境と結果について説明する．6 章で考察を行う．

2. 質問応答システムと質問応答タスク

2.1 質問応答システム

本節では，質問応答システムについて説明する．質問応答システムは自然言語で与えられた質問に対して，答えそのものを出力するシステムである．本論文で使用する質問応答システム [2] は「質問文解析部」「文書検索部」「回答特定部」の３つの技術から成る．この質問応答システムが質問文を受け取ると，最初に「質問文解析部」で，質問文を解析し，質問文中から重要語（名詞）を抽出し，回答タイプ（地名，人名など）を決定する．続いて,「文書検索部」では先ほど抽出した回答タイプと重要語をもとに，対象文書群から回答候補を含む文章を検索する．最後に,「回答特定部」で，先ほど検索した文書から回答候補を絞り込み，特定する．

2.2 質問応答タスク

本節では，質問応答タスクについて説明する．一般的に，質問応答システムに求められることは，質問に対して漏れなく，正確な回答を出力することである．例えば，大リーグ入りした日本人野球選手には誰がいたのか？について知りたい質問応答システム利用者がいたとする．そこで，利用者が「大リーグ入りした日本人野球選手には誰がいますか.」という質問を質問応答システムに与えたとしよう．この質問に対して，質問応答システムが一般に求められることは，松井秀樹，野茂英雄，佐々木主浩，etc…という風に，回答を漏れなく，正確に列挙することである．

質問応答システムが出力した回答に漏れが存在する場合，網羅性を持った回答とはいえない．例えば，上記の質問に対

して，質問応答システムが “野茂英雄” のみを出力した場合，この回答は網羅性に欠ける．また，上記の質問に対して，質問応答システムが “織田信長” などの誤った回答を出力した場合，この回答は正確さに欠ける．さらに，質問応答システムが正しい回答を出力したとしても，それが “野茂” のように名字だけであった場合，“野茂英雄” のようにフルネームで答えた回答と比べると，“野茂” という回答は正確さに欠ける．

上記のような質問応答の定義をタスクとして実施した例に，NTCIR QAC [1] の subtask2 がある．この subtask2 は，提示された質問に対し，考えられる回答を過不足なく網羅的に出力することが求められる．

3. 質問文と回答候補の関係に基づく回答候補絞りこみ

3.1 基本的な考え方

本論文で説明する回答絞り込み手法は,「重要語句と回答候補の関係が強いほど，それらが同一文章中に共起することが多い」という仮定に基づいている．この仮定に基づくと,「大量文書における重要語句と回答候補の共起頻度が高いほど，それらの関係は強い」と言える．この共起頻度は WWW 検索エンジンの AND 検索結果から推測出来る．しかし，このヒット数は回答候補と重要語句のヒット数に影響される．回答候補のヒット数が高いほど，その回答候補と重要語句の AND 検索のヒット数も高くなる．同様に，重要語句のヒット数が高いほど，その重要語句と回答候補の AND 検索のヒット数も高くなる．これらを考慮する為に，“重要語句と回答候補の AND 検索のヒット数” を “重要語句のみのヒット数” と “回答候補のみのヒット数” で割る必要がある．

また，重要語句と回答候補の関係の強さを評価する際，重要語句を重要語 “2 語” の組み合わせとしている．これには以下の理由がある．まず，重要語 “1 語” と回答候補の AND 検索は，検索条件が緩すぎて，結果を絞り込めないことがある．この場合，質問文とは関係のない記事ばかりが検索され，検索結果は信頼性が低い．逆に，重要語句が重要語 “3 語” 以上の組み合わせから成る場合，“重要語句” と “回答候補” の AND 検索は，検索条件が厳しすぎてヒット数が 0 になることがある．

そこで，重要語 2 語の全ての組み合わせにおいて，“重要語 2 語” と “回答候補” の関係の強さを評価し，それらの算術平均値を “質問文” と “回答候補” の関係の強さとする手法を提案した．

さらに，適当な閾値を選択し，この関係の強さが閾値以下の回答候補を取り除く手法を提案した．

3.2 提案手法と具体例

ここに，QAC2 subtask2 の質問例がある．

Q1：エリツィン大統領によって
解任された首相とはだれですか.

この質問に対して，質問応答システムが出力した回答候補と質問文中から抽出した重要語を以下の表 1 に示す.

表 1： 重要語と回答候補

重要語	エリツィン，大統領，解任された，首相
回答候補	プリマコフ，橋本

これらの回答候補を絞り込み，誤った回答候補を取り除く方法について説明する．この方法とは，以下のようなものである．まず，各回答候補において，WWW 検索エンジン goo (注 1)で以下の検索クエリを検索し，そのヒット数を調べる．

検索クエリ 1：重要語句と回答候補（ヒット数：$hit(K\ and\ A)$）
検索クエリ 2：重要語句（ヒット数：$hit(K)$）
検索クエリ 3：回答候補（ヒット数：$hit(A)$）

検索結果を以下の表 2，3 に示す.

表 2： 検索ヒット数（プリマコフ:A）

重要語句:K	$hit(K\ and\ A)$	$hit(K)$	$hit(A)$
“エリツィン” “解任された”	28	89	528
“エリツィン” “大統領”	214	3630	528
“エリツィン” “首相”	200	1990	528
“解任された” “大統領”	35	835	528
“解任された” “首相”	34	853	528
“大統領” “首相”	323	98600	528

表 3： 検索ヒット数（橋本:A）

重要語句:K	$hit(K\ and\ A)$	$hit(K)$	$hit(A)$
“エリツィン” “解任された”	15	89	307000
“エリツィン” “大統領”	667	3630	307000
“エリツィン” “首相”	584	1990	307000
“解任された” “大統領”	86	835	307000
“解任された ” “首相”	107	853	307000
“大統領” “首相”	8,850	98600	307000

次に，これらの値を用いて重要語句と回答候補の関係の強さを評価する．この評価には以下の式 1 を用いる．

$$\text{関係の強さ} = \frac{hit(K\ and\ A)}{hit(K) \times hit(A)} \quad (1)$$

さらに，各回答候補において，各重要語句との関係の強さの算術平均値を計算する．計算結果を表 4 に示す.

最後に，適当な閾値を選択し，算術平均値が閾値以下の回

(注 1)：http://www.goo.ne.jp

表 4： 各回答候補と質問文の関係の強さ

回答候補	関係の強さ（$\times 10^{-6}$）
プリマコフ	176.49
橋本	0.52

答候補を取り除く．例えば，閾値として 10（$\times 10^{-6}$）を選択すると，“橋本” が取り除かれる．事実として，エリツィン大統領によって解任された首相は “プリマコフ” であり，“橋本” は誤りである．

4. 閾値の選択

本章では，適当な閾値の選択手法について説明する．その手法とは以下のようなものである．

まず，上記の方法で「質問文と正しい回答の関係の強さ」と「質問文と誤った回答の関係の強さ」を評価する．この評価結果を訓練データとして用い，これらの回答を様々な閾値で絞り込む．絞り込んだ回答に対する評価が最も高い閾値を「最適な閾値」として選択する．

回答の評価には F 値という値が用いられる．F 値 を計算するために必要な要素を示すと，$Answer$ は実験に用いた質問文における全正解数，$Output$ はシステムが出力した回答数，$Correct$ はシステムが正解回答を絞り込んだ数，F 値，再現率，適合率はそれぞれ以下の式 2〜4 を用いて計算できる．

$$F\text{ 値} = \frac{2 \times \text{再現率} \times \text{適合率}}{\text{再現率} + \text{適合率}} \quad (2)$$

$$\text{再現率} = Correct \,/\, Answer \quad (3)$$

$$\text{適合率} = Correct \,/\, Output \quad (4)$$

再現率は正解のカバー率を示し，適合率はシステムの精度を示す．F 値 はこれらをひとまとまりにしたものである．F 値 が高いほど，システムの評価は高くなる．

5. 実験環境と結果

5.1 実 験 環 境

質問応答システムが出力した回答に対して，提案手法を実行した．今回の実験で使用した質問文は NTCIR QAC2 subtask2 の質問文を用いた．本実験では，質問文 200 文中，50 文を用いた．その 50 文の中で，25 文を訓練データとして用いた．そして，残りの 25 文を評価データとして用いた．

5.2 実 験 結 果

訓練データと評価データに対して，前述の回答絞り込み手法を実行した．訓練データから適当な閾値として，0.300（$\times 10^{-6}$）が得られた．表 5 は評価データに対する回答絞り込み結果を示し，表 6 はそれに対する評価結果を示す．

評価データに対して，上記の回答絞り込み手法を実行した結果を表 7 に示す．それに対して評価を行った結果を表 8 に示す．また，参考として「質問文と回答個補の関係の強さが最高となる回答候補 1 つを選択する方法（以後，表記を簡単にするため回答選択法と呼ぶ）」を実行した結果を表 7 に示す．それに対する評価結果を表 8 に示す．

表 5： 絞り込み前後の結果（訓練データ）

	$Answer$	$Output$	$Correct$	閾値（$\times 10^{-6}$）
絞り込み前	88	76	17	—
絞り込み後	88	50	16	0.300

表 6： 絞り込み前後の評価結果（訓練データ）

	再現率	適合率	F 値
絞り込み前	0.193	0.224	0.207
絞り込み後	0.182	0.340	0.232

表 7： 絞り込前後の結果（評価データ）

	$Answer$	$Output$	$Correct$	閾値（$\times 10^{-6}$）
絞り込み前	84	65	9	—
絞り込み後	84	45	8	0.300
回答選択法	84	25	6	—

表 8： 絞り込前後の評価結果（評価データ）

	再現率	適合率	F 値
絞り込み前	0.107	0.138	0.121
絞り込み後	0.095	0.178	0.124
回答選択法	0.071	0.240	0.110

6. 考　　察

本章では，実験で得られた比較結果について考察を行う．まず，回答選択法と提案手法に対する比較を行う．表 8 から，再現率は提案手法の方が約 34 %高い．一方，適合率は提案手法より回答選択法の方が約 35 %高い．その結果，F 値は提案手法の方が約 10 %高くなった．この結果は提案手法の優位性を示している．

次に，質問応答システムが出力した回答に対して，提案手法を実行する前後に対する比較を行う．表 8 から，再現率が約 13 %低下し，適合率が約 22 %向上したことが解る．その結果，F 値が 3 %向上した．

適合率が向上した一方で，再現率が低下した原因として，「質問に対する正しい回答であるにも係わらず，質問文との関係が弱い回答が存在する」ことが考えられる．このような回答が存在すると，4 章で述べた閾値を低く設定しなければならない．この結果，回答が上手く絞り込めず，適合率向上の妨げになると考えられる．また，回答を絞り込む際に，このような回答は取り除かれる．この結果，再現率が低下すると考えられる．

質問文と正しい回答との関係が弱くなる例として，次の質問文と回答を挙げる．

Q2 ：異種格闘技のルーツと言われた一戦を戦ったのは誰．
回答：アリ

本論文で使用した質問応答システム [2] は Q2 に対して，“アリ”(注 2)という回答を出力した．QAC2 subtask2 では，この回答は正しいとされている．しかし，この “アリ” と質問文との関係の強さ（表 9）と，他の誤った回答と質問文の関係の強さ（表 3）を比べても，大きな差は見られない．

表 9： “アリ” と各重要語句の関係の強さ

重要語句	関係の強さ（10^{-6}）
“一戦” “異種”	0.96
“一戦” “格闘技”	0.46
“一戦” “ルーツ”	0.60
“異種” “格闘技”	0.53
“異種” “ルーツ”	0.61
“格闘技” “ルーツ”	0.53
算術平均値	0.61

正しい回答である “アリ” と Q2 との関係が弱くなるのは，“アリ” が多義語であることが原因と考えられる．この “アリ” は，世界中の “アリ” という名前の人を指すこともあれば，昆虫の “アリ” のことを指すこともある．我々が知りたいのはボクサーのモハメッド・アリ氏と Q2 との関係の強さであり，昆虫の “アリ” と Q2 との関係の強さではない．実際に，“アリ” の WWW エンジン検索結果と，“ボクサーのアリ” の検索結果を比較すると，“ボクサーのアリ” の検索結果の多くは Q2 の意図するモハメッド・アリ氏についての文章で占められている．一方，“アリ” の検索結果の多くは，モハメッド・アリ氏とは無関係と考えられる文章で占められている．

そこで，“アリ” がボクサーのモハメッド・アリ氏の事を指すように，文章を特定する必要があると考えられる．

7. おわりに

本論文では，質問文と回答の関係の強さに注目し，質問応答システムが出力した回答を絞り込む手法を提案した．質問応答システムが出力した回答に対して提案手法を実行した結果，実行前と比較して F 値が 3 %向上した．したがって，質問文と回答の関係の強さを用いた回答絞り込み手法は有効であると言える．

今後は，さらに提案手法を改良し，より精度の高い回答絞り込み手法について検討する．

(注 2)：モハメッド・アリ氏とアントニオ・猪木氏の一戦は一般に，異種格闘技戦のルーツと言われている．

参考文献

[1] J. Fukumoto , T. Kato and F. Masui, “Question Answering Challenge(QAC1): Question answering evaluation at NTCIR Workshop 3,”, In Working Notes of the Third NTCIR Workshop Meeting:QAC1，2002.

[2] 日高直哉，桝井文人，“質問応答における回答絞り込み手法の比較,” 第１７回人工知能学会全国大会論文集，CD-ROM, 2003.6.

[3] 宮口正行，桝井文人，“構文構造を考慮した質問応答のための重要文抽出,” 信学技報 NLC202-38，Vol.102，No.414, pp.1–6，2002.10.

[4] S. Lee, G. G. Lee,“ SiteQ/J:A Question Answering for-Japanese, ”In Working Notes of the Third NTCIRWorkshop Meeting:QAC1，2002.

[5] 外池晶嗣，佐藤理史，宇津呂武仁，“4 択クイズを連想問題として解く,” 言語処理学会第 10 回年次大会発表論文集, pp301–304, 2003.3.16.