Web 検索エンジンを用いた Why 型質問応答システム

田村元秀　村上仁一　徳久雅人　池原悟
鳥取大学 工学部 知能情報工学科
{s022033,murakami,tokuhisa,ikehara}@ike.tottori-u.ac.jp

1　はじめに

新聞記事をはじめとする大量の文書から必要な情報を取り出すことは容易ではない．そこで自然言語で記述された質問に対し，情報検索や情報抽出の技術を組合せ，最適な回答を得る質問応答技術が注目されている．

質問応答技術を評価するワークショップとして国立情報学研究所主催の NTCIR において QAC[1] があり，新聞記事を利用し名称や日付など事実に基づく回答を求める factoid 型質問 [2] や，複数の回答を求める list 型質問などが扱われている．また最近行われた QAC4 では，Why などを問う non-factoid 型質問も挑戦されている．

ところで Web 上には新聞記事に比べ膨大な数の文書があり，Why 型をはじめとする non-factoid 型質問に回答する文が存在する可能性が高い．しかし日本語において，Web 文書を用いて Why 型質問に回答する研究は見当たらなかった．本研究では検索エンジンから得られる結果を利用し，Why 型に回答する質問応答システムを試作する．また得られた結果より本システムの問題点を明らかにすると同時に，改善手法を提案し評価を行う．

2　Why 型質問応答システムの構成

本研究における Why 型質問応答システムとは，自然言語で記述された「なぜ」または「どうして」を含む質問に対し，適切な回答を出力するシステムである．

図 1 に本システムの構成を示す．

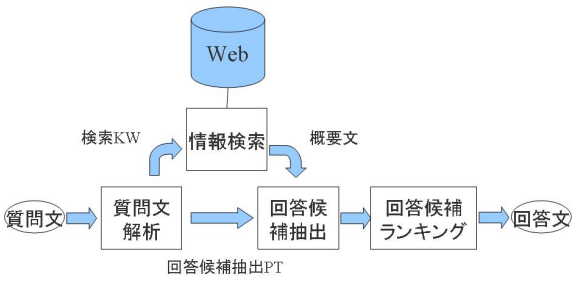

図 1　Why 型質問応答システム

本システムは質問応答技術として代表的な「質問文解析」，「情報検索」，「回答候補抽出」，「回答候補ランキング」という 4 つのモジュールで構成している．

2.1　質問文解析

入力された質問文の言い換えを行い，検索キーワード（以下検索 KW）と回答候補抽出パターン（以下回答候補抽出 PT）を作成する．以下に具体的に説明する．

（1）質問文の言い換え

検索 KW と回答候補抽出 PT を作成するために，まず質問文の言い換えを行う．疑問語を用いた疑問文はその位置（文頭，文中，文末）により大きく 3 つに分類できる．

本研究における質問文の言い換えとは，疑問語の位置を “文末” にすることである．以下に質問文の言い換え手順を示す．

step1.　疑問語 (なぜ、どうして) を削除
step2.　文末表現 (の、のですか。) を削除
step3.　助詞「は」を助詞「が」に置換
step4.　文末に「のはなぜ」を追加

（2）検索 KW の作成

検索エンジンにクエリを与えるために，言い換え後の質問文の疑問語を削除することで検索 KW を作成する．

（3）回答候補抽出 PT の作成

検索結果から回答候補を抽出するために，検索 KW に任意記号を追加し回答候補抽出 PT を作成する．

ここで図 2 に質問解析における具体例を示す．

＜入力質問文 (文中) ＞空はどうして青いのですか。
＜言い換え後質問文 (文末) ＞空が青いのはなぜ
＜検索 KW ＞空が青いのは
＜回答候補抽出 PT ＞空が青いのは*

図 2　質問文解析の例

2.2　情報検索

質問文解析によって得られた検索 KW をクエリとし，検索エンジンに与える．そして検索エンジンから得られた結果から概要文 (snippet) を取り出す．

2.3 回答候補抽出

得られた概要文と，回答候補抽出 PT との照合により回答候補を抽出する．図 3 は，回答候補抽出 PT「空が青いのは*」で抽出される回答候補の例である．

＜適合文＞それゆえに空が青いのは、主に空気の分子による光の散乱です。
＜回答候補抽出 PT ＞空が青いのは*
＜回答候補＞空が青いのは、主に空気の分子による光の散乱です。

図 3　回答候補を抽出する例

2.4 回答候補ランキング

回答候補抽出では，通常複数の回答候補が得られる．得られた複数の回答候補をランキングするために，各回答候補をスコアリングする必要がある．本研究では「検索結果には質問に回答する情報が多く含まれている」という仮定をし，スコアリングには，概要文の名詞出現頻度を用いた名詞頻度テーブルを利用する．

名詞頻度テーブルとは，概要文について名詞（ただし非自立語を除く）を抽出し，その出現頻度を示したテーブルである．表 1 に検索 KW「空が青いのは」における名詞頻度テーブルの高頻度名詞上位 3 件，表 2 に低頻度名詞上位 3 件を示す．

表 1　高頻度名詞上位 3 件

名詞	スコア
空	724.0
光	300.0
色	121.0

表 2　低頻度名詞上位 3 件

名詞	スコア
世	1.0
細工	1.0
霜	1.0

作成した名詞頻度テーブル T に基づき，各回答候補 a に対するスコア S_a を以下の式で求める．M は回答候補 a における名詞の数であり，ω_i は i 番目の名詞である．

$$S_a = \sum_{i=1}^{M} T(\omega_i)$$

最後に，スコアリングされた各回答候補は降順にソートし，ランキングする．

3 実験

3.1 実験条件

使用する Why 型質問文は，学研サイエンスキッズ [3] の科学なぜなぜ 110 番の自然ジャンルを参考に 50 文を作成する．また正解文は同サイトから正解を含む部分を引用する．本研究では検索エンジンとして Google[4] を使用し，検索結果最大 500 件から概要文を抽出する．また形態素解析器として MeCab[5] を用いる．

3.2 評価

3.2.1 人手評価

本システムの精度を確認するために，人手評価を行う．評価基準は QAC[1] に従って，ランキング 1 位の回答候補に対して “A”，“B”，“C”，“D”，“なし” で評価する．以下に評価基準を示す．

評価 A　出力された回答は正解とほぼ等しい内容である
評価 B　出力された回答は正解に加え他の情報を含む
評価 C　出力された回答は正解の一部を含む
評価 D　出力された回答は正解の内容を含まない
評価 なし　出力なし

（評価 A）
＜質問文＞虹はどうしてできるのですか。
＜正解文＞プリズムににた形の雨のつぶに光が当たると、何色にも分かれて見えるようになるのです。
＜回答候補＞虹ができるのは、雨のあとなど空気中にうかぶ水蒸気の粒に太陽の光があたると、水滴がプリズムの働きをするため光が反射・屈折して光の帯となって見える自然現象なのです。

（評価 B）
＜質問文＞虹はどうして 7 色なのですか。
＜正解文＞虹が 7 色なのは、虹のもとになる太陽の光が、およそ 7 色の光からできているからなのです。
＜回答候補＞虹が 7 色なのは太陽の光が水蒸気などでそれぞれの波長に分解され見えるのですが、人間には見えない波長も存在します。

（評価 C）
＜質問文＞雷はどうして光るのですか。
＜正解文＞雷の電気が流れたところの空気の温度はかなり高くなります。すると、熱くなった空気は光ります。
＜回答候補＞雷が光るのは、もちろんこれによって発生した光によるものです。

（評価 D）
＜質問文＞雪はどうして冷たいのですか。
＜正解文＞雪が溶けるとき融解熱を奪っていくために、雪は冷たいものだと感じるのです。
＜回答候補＞雪が冷たいのは、とても当たり前のことなのに、今までこんな風に俳句に作った人はいなかったみたいです。

図 4　人手評価例

3.2.2 F 値による評価

人手評価をもとに，F 値を以下の式で求める．ただし $Recall$ は正答（評価 A または評価 B）数を質問数で割った値であり，$Precision$ は正答数を出力数で割った値である．

$$F\text{-}value = \frac{2 * Recall * Precision}{Recall + Precision}$$

3.2.3 MRR による評価

別の評価指標として，ある質問に対して最上位正答順位の逆数を平均した $MRR(Mean\ Reciprocal\ Rank)$ を用いる．ただし N は問題数，$rank_k$ は問題 k における正答の最高順位である．

$$MRR = \frac{1}{N}\sum_{k=1}^{N}\frac{1}{rank_k}$$

3.2.4 累積検索成功率による評価

また別の指標として，回答候補ランキング上位 10 件に正答を含む割合として累積検索成功率を用いる．

3.3 実験結果

表 3 に Why 型質問文 50 問における人手評価の結果，表 4 に F 値と MRR と累積検索成功率の結果を示す．

表 3　人手による評価の結果

A	B	C	D	なし
10% (5/50)	8% (4/50)	4% (2/50)	56% (28/50)	22% (11/50)

表 4　F 値，MRR，累積検索成功率の結果

F 値	MRR	累積検索成功率 (〜10)
0.20	0.285	40%

回答候補ランキング 1 位が正答 (評価 “A”，“B”) であるのが 2 割にも満たないのに対し，誤答 (評価 “C”，“D”) と出力なしが全体の 8 割を占める結果となった．また累積検索成功率も 4 割にとどまっている．

3.4 システム出力の分析

評価を行った結果，回答候補ランキング 1 位が正答にならない理由として大きく 2 つあることが分かった．

第 1 に “回答候補抽出の失敗” である．評価 “なし” が 2 割以上であることから，「Web 検索結果が存在しない」，「Web 検索結果は存在するが，回答候補が存在しない」ことが考えられる．

第 2 に “回答候補ランキングの失敗” である．回答候補の中に正解となる回答候補は存在するが，他の回答候補に埋もれてしまうことが考えられる．

4 改善手法

“回答候補抽出の失敗” と “回答候補ランキングの失敗” について，それぞれ「検索用 KW のスリム化」と「名詞頻度テーブルの最適化」により，本システムの改善を行う．

4.1 検索 KW のスリム化

検索エンジンに与えるクエリが長い場合，検索結果が無いあるいは数件になり，回答候補が抽出されないことがある．そこで，検索結果を増やすために検索 KW のスリム化を行う．以下に，具体的な手法を述べる．

- 検索 KW のスリム化
  検索 KW を「従属節の名詞＆主節」とする

主節の判定には助詞の「は」，「が」を利用する．日本語において主節は文の後半に置かれることが多いことから，検索 KW から「のは」を除いた文に対して，最も後ろにある「は」，「が」をもとに判定を行う．

4.2 名詞頻度テーブルの最適化

名詞頻度テーブルは回答候補ランキングを大きく左右するため，最適化する必要がある．以下に最適化のための 4 つの手法を述べる．

1. 低頻度削除
   Web 検索結果に存在する不必要な情報 (ノイズ) の影響を減らすために，出現頻度 1 以下の名詞を削除
2. スコア加算
   Why 型質問の正解には「から」をはじめとする原因，理由を表す表現が含まれることが多いため，それらを含む回答候補に対してスコアを加算
3. スコア減算
   Why 型質問の回答として「当然」などは回答候補ランキングに悪影響を与えるため、それらを含む回答候補に対してスコアを減算
4. 名詞限定
   代名詞「あれ」などの意味の無い単語による回答候補ランキングの影響を防ぐため，「一般名詞」と「固有名詞」のみを登録

4.3 実験結果

表 5，表 6 に各改善手法別の人手評価の結果，及び F 値と MRR と累積検索成功率を示す．ただし，スリム化や低頻度削除など各手法の値は改善前のシステムに対して個別に得ている．また表中の (all) は改善手法すべてを適用した後のシステムの結果である．

表 5　各改善手法別の人手評価の結果

手法＼評価	A	B	C	D	なし
改善前	10% (5/50)	8% (4/50)	4% (2/50)	56% (28/50)	22% (11/50)
スリム化	12% (6/50)	8% (4/50)	4% (2/50)	58% (29/50)	18% (9/50)
低頻度削除	12% (6/50)	8% (4/50)	2% (1/50)	56% (28/50)	22% (11/50)
スコア加算	18% (9/50)	4% (2/50)	4% (2/50)	52% (26/50)	22% (11/50)
スコア減算	10% (5/50)	8% (4/50)	6% (3/50)	54% (27/50)	22% (11/50)
名詞限定	16% (8/50)	8% (4/50)	4% (2/50)	50% (25/50)	22% (11/50)
(all)	20% (10/50)	12% (6/50)	2% (1/50)	48% (24/50)	18% (9/50)

表 6　F 値，MRR，累積検索成功率の結果

手法＼評価	F 値	MRR	累積検索成功率 (〜10)
改善前	0.202	0.285	40%
スリム化	0.220	0.308	44%
低頻度削除	0.225	0.283	42%
スコア加算	0.247	0.333	50%
スコア減算	0.202	0.299	42%
名詞限定	0.270	0.325	44%
(all)	0.352	0.409	54%

各手法のうち，最も MRR の増加が大きかったのは「スコア加算」であり，最も増加が小さかったのは「低頻度削除」であった．しかしいずれの手法も効果があることを確認した．

4.4　考察

4.4.1　最適化の影響について

表 6 より，各改善手法のうち，最も有効であったのは「スコア加算」であった．これは，Why 型質問応答において「から」などの原因・理由の表現は，回答を抽出する上で大きな手がかりとなることを意味している．

「低頻度削除」,「スコア減算」については，あまり良い結果が得られなかった．しかし，「当たり前」や「当然」など回答候補ランキングに悪影響を与える禁止語を増やすことで「スコア減算」の効果が上がると考えている．また「低頻度削除」においては，頻度 2 以下，頻度 3 以下など削除対象となる名詞の出現頻度を変えて調べる必要があると考えている．

4.4.2　回答候補のスコア分布について

ランキング上位 5 件の回答候補において，スコアと頻度の関係を正答誤答別に比較したものを図 5 に示す．

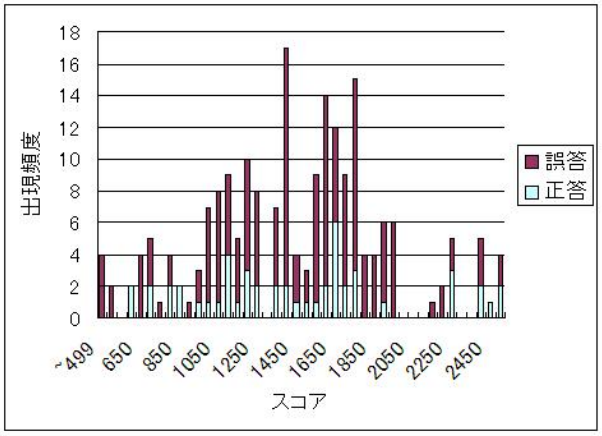

図 5　回答候補のスコア分布

図 5 よりスコアと正答，誤答間に関連性が低いことが確認できた．なお，ランキング上位 5 件に占める割合は正答 24.6%(50/203)，誤答 75.4%(153/203) であった．

5　おわりに

本研究では，Web 上に存在する膨大な数の文書を対象に，検索エンジンを用いて得られる結果を利用し，Why 型質問に回答する質問応答システムを試作した．また得られた結果より本システムの問題点を明らかにすると同時に，改善手法を提案し，評価した．その結果，本システムに対して各種最適化は有効であることを示し，F 値 0.352，累積検索成功率 54% を得た．

ところで本研究では，回答候補は引用した正解文をもとに評価を行ったため，正解文に記述されている内容以外正解にしなかった．そこで，著者の基準で評価を行ったところ累積検索成功率で約 20% の増加が見られ，実質的には表 5，表 6 よりも良い結果が得られている．

今後は，本システムにおける正答または誤答の回答候補について特有の表現を抽出による回答候補の絞り込みや，回答候補抽出 PT のバリエーションの追加によって精度の高い Why 型質問応答システムの構築を目指す．

参考文献

[1] J.Fukumoto, et al. : An overview of NTCIR-6 QAC4. In Proc. of the 6th NTCIR Workshop Meeting, pp.433-440, (2007).

[2] 田村ほか: Web を知識源とする質問応答システムにおけるパターン方式とキーワード方式の比較, 電子情報通信学会ソサイエティ大会講演論文集 p.195, (2007).

[3] 学研サイエンスキッズ
http://kids.gakken.co.jp/kagaku/index.html

[4] Google Search Engine:
http://www.google.co.jp

[5] MeCab:Yet Another Part-of-Speech and Morphological Analyzer
http://www.chasen.org/~taku/software/mecab/