Web 検索エンジンを用いた Why 型質問応答システムに関する研究

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
〒 680-8552 鳥取市湖山町 4-101
E-mail:{s022033,murakami,tokuhisa,ikehara}@ike.tottori-u.ac.jp

あらまし：膨大な情報の中からユーザの必要とする情報を見つけ出す技術として質問応答システムがある．本稿ではインターネット上の Web 検索エンジンを利用することにより，Why 型質問に回答する質問応答システムを試作し，出力された回答候補を人手で評価した．また明らかになった問題点に対して改善手法を示し，評価を行った結果，本システムに対して改善手法は有効であることを示し，F 値 0.352，検索成功率 54%を得た．
キーワード：質問応答システム，Why 型質問

A Study for the Why-type Question-Answering System using Web Search Engine

Yukihide Tamura , Jin'ichi Murakami , MasatoTokuhisa and Satoru Ikehara

Faculty of Engineering, Tottori University
Minami 4-101, Koyama, Tottori, 680-8552, Japan
E-mail:{s022033,murakami,tokuhisa,ikehara}@ike.tottori-u.ac.jp

Abstract : Question-Answering system is a technology to find needed information from enormous information. In this paper, we proposed Question-Answering system to answer a Why-type question by using existing Web search engine, and evaluated candidate answer by hands. In addition, we developed advance system by improvement techniques. As a result, we showed improvement techniques was effective for this system, and finally we got 0.352 for F-value, and 54% for the search success rate.
Keyword : Question-Answering system, Why-type question

1. はじめに

Google をはじめとする検索エンジンの普及により，情報検索はより身近になっている．しかし大量の検索結果から必要な情報を取り出すことはユーザにとって容易ではない．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は国立情報学研究所の NTCIR（NII Test Collection for Information Retrieval and Text Processing）の QAC[1] タスクがあり，新聞記事を対象に名称や日付・数値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，定義や説明を求める definition 型質問などが扱われている．また最近行われた QAC4（2006.4-2007.5）では，Why や How をはじめとする non-factoid 型の質問も扱われている．

Why 型質問応答技術に関する過去の研究として，諸岡らは，Why 型質問と回答間に成立する意味的な関係とその手がかりとなる語に注目し，Why 型質問の回答を抽出する手法を提案した[3]．また東中らは，Why 型，How 型質問に対して，原因表現のデータを得るために EDR コーパスを使った学習アプローチを提案した[4]．これらの研究では共に新聞記事を対象にした回答候補抽出を行っている．

ところでインターネット上には膨大な数の文書があ

り，その中には 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ある．しかし日本語において，これらを用いて Why 型質問に回答する研究は見当たらなかった．

そこで，本研究では検索エンジンから得られる検索結果を利用し，Why 型に回答する質問応答システムを試作する．また出力された回答候補を人手で評価し，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，評価する．

2. Why 型質問応答システムの構成

本研究では，検索エンジンにより得られる結果を対象に，Why 型質問に回答する質問応答システムを試作する．質問応答システムとは，自然言語で記述された質問に対し，適切な回答を出力するシステムである．例えば質問文「地震が起きるのはなぜ（どうして）?」に対して,「地震が起きるのは地球の表面を覆っているプレートが動くためです。」と回答する．

本システムの構成を**図 1** に示す．

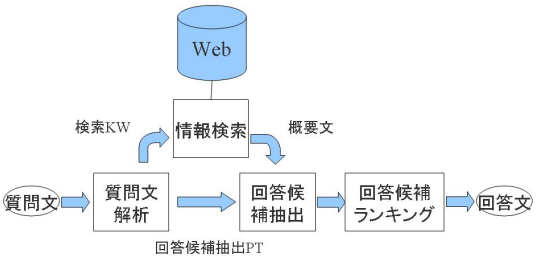

図 1 Why 型質問応答システム

本システムは（1）質問文解析,（2）情報検索,（3）回答候補抽出,（4）回答候補ランキングの 4 つのモジュールより構成されている．以下に詳細を示す．

2.1 質問文解析

入力された質問文を解析し，情報検索用キーワード（以下検索 KW）と回答候補抽出用パターン（以下回答候補抽出 PT）を作成するために，質問文の言いかえを行う．言いかえ後の質問文から検索 KW と回答候補抽出 PT を作成する．

2.1.1 質問文の言いかえ

Why 型質問文など求める情報の部分に疑問語を使う「疑問語疑問文」には大きく分けて 3 つのパターンが存在する．以下に例を示す．

（1） 疑問語が文頭にある
（例 1）どうして地球は自転しているのですか。
（2） 疑問語が文中にある
（例 2）空はどうして青いのですか。
（3） 疑問語が文末にある
（例 3）雷が鳴るのはどうしてですか。

本研究における質問文の言いかえとは,（例 1）あるいは（例 2）タイプ質問文を,（例 3）タイプに変換することである．以下に質問文の言いかえ手順を示す．

言いかえの手順
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

上記の（例 1）,（例 2）,（例 3）について言いかえを行った結果を**図 2** に示す．

（例 1）地球が自転しているのはなぜ
（例 2）空が青いのはなぜ
（例 3）雷が鳴るのはなぜ

図 2 言いかえ結果

2.1.2 検索 KW の作成

検索エンジンにキーワードを与えるために，言いかえを行った質問文から疑問語部分を削除することで，検索 KW を作成する．**図 3** に質問文「空が青いのはなぜ」における検索 KW の作成例を示す．

＜質問文＞空が青いのはなぜ
↓ 疑問語部分を削除
＜検索 KW ＞空が青いのは

図 3 検索 KW を作成する例

2.1.3 回答候補抽出 PT の作成

検索結果から得られた概要文から回答候補を抽出するために，検索 KW をもとに回答候補抽出 PT を作成する．**図 4** に質問文「空が青いのはなぜ」における回答候補抽出 PT の作成例を示す．

＜検索 KW ＞空が青いのは
↓
＜回答候補抽出 PT ＞空が青いのは*（任意の文字列）

図 4 回答候補抽出 PT を作成する例

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．**図 5** は検索 KW「空が青いのは」における Google の検索結果の一部である．

空は何故青いの？
空が青いのは、主に空気の分子による光の散乱です。 空気の分子のように、光 の波長よりも長さが短い粒子による光の散乱を「レイリー散乱」と言います。一方、夕焼けは、この「レイリー散乱」に加えて、「ミー散乱」という物も関係しています。 ...

NIDEK 目のおはなし 空が青く見えるのは？
そうなると**空が青いのは**太陽が原因なのでしょうか。結論から言えばその通りで太陽の光も関係しています。 それではまず太陽の光のお話です。太陽と地球との距離は約 1 億 5000 万キロメートルあり、太陽の光が地球に届くには約 8 分 19 秒かかるそうです。 ...

パーティサファイア サファイアの魅力
また、古代ペルシャの伝説では、サファイアは大地を支える石と呼ばれ、**空が青いのは**、サファイアの輝きが空に映し出されているからだと言われてきました。実際、サファイアは空気を敏感に感じる石で、曇りの日と晴れの日では、輝き方が違い、「最も神に ...

図 5 検索結果の一部

2.3 回答候補抽出

得られた概要文と，質問文解析で作成した回答候補抽出 PT とのマッチングにより回答候補を抽出する．**図 6** 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 1 ＞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 1 ＞空が青いのは、主に空気の分子による光の散乱です。

＜適合文 2 ＞空が青いのは太陽が原因なのでしょうか。
＜回答候補 2 ＞空が青いのは太陽が原因なのでしょうか。

＜適合文 3 ＞また、古代ペルシャの伝説では、サファイアは大地を支える石と呼ばれ、空が青いのは、サファイアの輝きが空に映し出されているからだと言われてきました。
＜回答候補 3 ＞、空が青いのは、サファイアの輝きが空に映し出されているからだと言われてきました。

図 6 回答候補を作成する例

2.4 回答候補ランキング

回答候補抽出において，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「検索結果の概要文には質問に回答する情報が多く含まれている」という仮定をもとに，概要文の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 5 件，**表 2** に低頻度名詞上位 5 件を示す．

表 1 高頻度名詞上位 5 件

名詞	スコア
空	724.0
光	300.0
色	121.0
海	117.0
波長	110.0

表 2 低頻度名詞上位 5 件

名詞	スコア
世	1.0
細工	1.0
霜	1.0
範囲	1.0
視界	1.0

また各回答候補 a に対するスコア S_a は以下の式で求める．M は回答候補における名詞の数であり，ω_i は i 番目の名詞である．

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

作成した名詞頻度テーブルに基づき，各回答候補をスコアリングする．ただし,「？」,「なぜ」は回答候補に悪影響を与えるため，それらを含む回答候補のスコアに対して，閾値（score500）を減算する．最後にスコアを降順でソートすることでランキングを行う．**図 7** は回答候補ランキングの例である．

＜回答候補 1 ＞空が青いのは、主に空気の分子による光の散乱です。(score 1104)
＜回答候補 2 ＞空が青いのは太陽が原因なのでしょうか。(score 831)
＜回答候補 3 ＞、空が青いのは、サファイアの輝きが空に映し出されているからだと言われてきました。(score 772)

図 7 回答候補ランキング例

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の科学[5] なぜなぜ 110 番の自然（地球・気象）ジャンルを参考に 50 文を作成する．また，正解文は質問文の作成元から正解を含む部分を引用する．これらを用いて検索実験を行う．本研究では検索エンジンとして

Google[6] 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[7] を用いる．

3.2 評　　価

3.2.1 人手による評価

本システムの精度を確認するために，人手による評価を行う．

評価は 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” で示され，以下の評価基準を用いる．また**図 8** に評価例を示す．

評価基準

評価 A　出力された回答は正解である

評価 B　出力された回答は正解を含むが，他の内容も含む

評価 C　出力された回答は正解の一部を含む

評価 D　出力された回答は不正解である

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 8　評価例

3.2.2 F 値による評価

人手による評価をもとに，F 値を以下の式で求める．ただし $Question$ は質問数，$Correct$ は正答（評価 A または評価 B）を出力した数であり，$Output$ は本システムが回答を出力した数である．

$$Recall = \frac{Correct}{Question} \tag{2}$$

$$Precision = \frac{Correct}{Output} \tag{3}$$

$$F - value = \frac{2 * Recall * Precision}{Recall + Precision} \tag{4}$$

3.2.3 平均逆順位による評価

人手による評価をもとに，上位 100 件までの平均逆順位[8](MRR : $MeanReciprocalRank$) を以下の式で求める．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．

$$MRR = \frac{1}{N}\sum_{k=1}^{N}\frac{1}{rank_k} \tag{5}$$

3.2.4 検索成功率による評価

人手による評価をもとに，回答候補ランキング上位 10 件に正答を含む割合として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，また**表 4** に F 値，MRR，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，検索成功率の結果

F 値	MRR	検索成功率
0.20	0.285	40%

回答候補ランキング 1 位の回答候補の約 8 割は不正解または出力なしであった．また検索成功率も 4 割にとどまっていた．

3.4 システム出力の分析

回答候補の評価を行った結果，回答候補ランキング 1 位の回答候補が正答にならない理由として大きく 2 つあると考えている．

第 1 に「回答候補抽出の失敗」である．検索成功率が 5 割以下であり，また評価 “なし” が 2 割以上であることから回答候補の中に正解となる回答候補が存在しない場合がある．

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場合がある．

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」「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

回答候補抽出の失敗の原因として，回答候補中に正解となる回答候補が存在しないことが考えられる．例えば「雨が降っても海がいっぱいにならないのはなぜ」という質問から「雨が降っても海がいっぱいにならないのは」で検索した結果，回答候補は出力されなかった．これは検索 KW が長いため，回答候補が得られなかった例である．これに対して質問を解析し，検索 KW を「従属節の名詞+主節」とすることで検索 KW のスリム化をはかり回答候補を出力する．

主節の判定には助詞の「は」または「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,「のはなぜ」を除いた文に対して，最も後ろにある「は」,「が」を元に判定を行う．同時に主節部分を回答候補抽出 PT とする．**図 9** に「雨が降っても海がいっぱいにならないのはなぜ」における検索 KW のスリム化の例を示す．

雨が降っても海 が いっぱいにならない
(最も後ろにある) 助詞

雨が降っても（従属節） 海がいっぱいにならない（主節）

（検索 KW）雨＆海がいっぱいにならない＋のは
（回答候補抽出 PT）海がいっぱいにならないのは

図 9 検索 KW のスリム化の例

4.2 名詞頻度テーブルの最適化

「回答候補ランキングの失敗」の原因として，名詞頻度テーブルの不備が考えられる．例えば，名詞頻度テーブルに余計な語が含まれていたり，原因・理由を表す語（「から」など）を評価していないことが挙げられる．以下は名詞頻度テーブルを最適化するために行った対策である．

4.2.1 低頻度削除

Web 検索エンジンの結果には必ず一定の不必要な情報（ノイズ）が存在する．例えば，検索 KW「雪が白いのは」で検索した結果の「雪が白いのは神がそうしたから。」は明らかに質問の意図に反する場合である．ノイズの影響を減らすために，頻度 1 以下の名詞を削除する．

4.2.2 スコア加算

本システムの評価過程において，正解には「から（助詞）」をはじめとする原因・理由を表す語が含まれることが分かった．これらの語を含む文は正解である可能性が他よりも高くなるため，回答候補のスコアに対して，閾値（score500）を加算する．**図 10** はスコア加算の対象文例である．

（例 4）海の水がしょっぱいのは、海の水にふくまれる塩の せい。
（例 5）火山が噴火するのは、水分などがガスになって容積がふえる勢いで溶岩や火山灰をおしあげる ため です。
（例 6）空が青いのは大気を形作る分子によって青い光が散乱される から です。

図 10 スコア加算の対象文例

4.2.3 スコア減算

本システムの評価過程で「当たり前」や「当然」をはじめとする多数の禁止語が，回答候補ランキングに悪影響を及ぼすことが分かった．回答候補ランキングへの悪影響を減らすため「当たり前」,「当然」,「常識」を禁止語に加え，回答候補のスコアに対して，閾値（score500）を減算する．**図 11** はスコア減算の対象文例である．

（例 7) 雪が白いのは 当たり前
（例 8）今では地球が丸いのは 常識 ですが、当時は非常識…

図 11 スコア減算の対象文例

4.2.4 名詞限定

名詞頻度テーブルは回答候補をランキングするのに大きくかかわっているため，登録する単語は慎重に選ぶ必要がある．例えば，代名詞「あれ」自体は意味の無い単語である．名詞頻度テーブルに登録するか否かを調査する必要がある．

そこで名詞頻度テーブルにおける登録対象の名詞は，意味のある単語である「一般名詞」「固有名詞」のみにする．

4.3 結　　果

表 5，**表 6** に各手法別人手による評価の結果，また F 値，MRR，検索成功率を示す．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．またスリム化や低頻度削除をはじめとする各手法は改善前のシステムに対して個別に適用している．

表 5　各手法別人手による評価の結果

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，検索成功率の結果

手法	F 値	MRR	検索成功率
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 改善手法に対する考察

4.4.1 検索 KW のスリム化の影響

表 3 より評価 “なし” の割合が 4%減少し，検索 KW のスリム化は有効であることが確認できた．**図 12** に．評価 “なし” から評価 “A” になった例を示す

＜質問文＞山に登るとどうして空気が薄くなるのですか。
＜スリム化後 KW ＞「山」,「空気が薄くなるのは」
＜正解文＞高いところでは、地球の引力が弱いためにたくさんの空気を引きつけておくことができないことになります。したがって、山の上では空気の量がうすくなるのです。
＜回答候補＞空気が薄くなるのは、そこより上にある空気の重さが少なくなるからです。

図 12　評価が上がった例

4.4.2 名詞頻度テーブルの最適化の影響

表 4 より，名詞頻度テーブルの最適化にもっと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．

また「名詞限定」についても同程度の効果が期待できるという結果になった．しかし名詞頻度テーブルに加える品詞として「一般名詞」,「固有名詞」だけに限定するのではなく，動詞や副詞などを考慮する必要がある．

「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頻度を変えて調べる必要があると考えている．

4.4.3 評価における別解の扱い

本研究では正解文をもとに人手評価を行い，別解を正解にしなかった．そのため実質的には，表 5, 表 6 より良い結果が得られている．**図 13** に別解の例を示す．

＜質問文＞雪はどうして降るのですか。
＜正解文＞雲の中で水のつぶは、はじめ小さな氷のつぶになります。そして、その氷のつぶのまわりに空気中にある水蒸気がくっついて、だんだん大きな氷の結晶となっていくのです。
＜回答候補＞雪が降るのは、シベリアってトコロから冷たい空気がやってきて北海道の高い山にぶつかるからなんだって。

図 13　別解の例

5. おわりに

本稿では，インターネット上に存在する膨大な数の文書から，検索エンジンを用いることで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した．また出力された回答候補を人手で評価し，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，評価した．その結果，本システムに対して改善手法は有効であることを示し，F 値 0.352，検索成功率 54%を得た．

しかし，未だ評価 D や評価なしの割合が大きいことから,「回答候補抽出」,「回答候補ランキング」について更に調べる必要がある．

今後は，本システムにおける正答の回答候補，誤答の回答候補について特有の表現を抽出したり，それぞれのスコア分布を調べ，回答候補の絞り込みを行い，より精度の高い Why 型質問応答システムの構築を目指す．

参 考 文 献

1) J.Fukumoto, T.Kato, F.Masui and T.Mori: An overview of NTCIR-6 QAC4. *In Proc. of the 6th NTCIR WorkshopMeeting*, pp.433–440, (2007)
2) 田村, 村上, 徳久, 池原: 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) K.Morooka and J.Fukumoto: Answer extraction method for why-type question answering system. In IEICE Technical Report, volume 105, pp.7–12, (2006)
4) R.Higashinaka and H.Isozaki: NTT's Question Answering System for NTCIR-6 QAC-4 *In Proc. of the 6th NTCIR WorkshopMeeting*, pp.460–463, (2007)
5) 学研サイエンスキッズ http://kids.gakken.co.jp/kagaku/index.html
6) Google Search Engine: http://www.google.co.jp
7) MeCab: Yet Another Part-of-Speech and Morphological Analyzer http://www.chasen.org/~taku/software/mecab/
8) 水野, 秋葉: 任意の回答を対象とする質問応答のための実世界質問の分析と回答タイプ判定法の検討, 言語処理学会 第 13 回年次大会発表論文集 pp. 1002–1005, (2007)