

ロバストな質問応答システム構築を目指した 質問の難易度とその評価手法に関する研究

平成 18 年 度

三重大学大学院工学研究科
博士前期課程 情報工学専攻

石 田 健 二

修士論文

ロバストな質問応答システム構築を目指した
質問の難易度とその評価手法に関する研究



平成 18 年度修了
三重大学大学院 工学研究科
博士前期課程 情報工学専攻

石田 健二

目次

1	序論	1
1.1	研究の背景	1
1.2	研究の目的	2
1.3	論文の構成	3
2	質問応答システムの構成と評価指標	4
2.1	質問応答システムの構成	4
2.2	質問応答システムに関する評価指標	6
3	予備調査	10
4	質問応答におけるコスト	12
3.1	クエリ操作の定義.....	12
3.2	クエリ操作の分析.....	13
5	ループ回数とシステム性能との相関	16
6	今後の課題	21
7	結論	22
	参考文献	23

第1章

序論

1.1 研究の背景

近年、インターネットの普及に伴い、膨大な電子化文書が蓄積されるようになった。これらの大量文書からユーザーの所望する情報を取り出すために有効な情報アクセス技術として、Google[13]やYahoo[14]などの情報検索エンジンが普及している。

情報検索エンジンを用いてユーザーの所望する情報を取り出す場合、まず、情報要求をキーワードとして入力し、キーワードを含む検索文書を受け取る。次に、検索文書中から所望する情報を見つけ出す。このとき、検索文書中に所望する情報があるとは限らない。ユーザーが所望する情報を得るためには、情報要求を明確化でき、かつ、どのようなキーワードで検索を行えば情報が得られるかが同定できる状態でなければならない[22]。所望する情報を得るのに適したキーワードが同定できないユーザーにとっては、検索キーワードに対して検索文書を取り出すのではなく、質問文に対して答えを取り出すことが望ましい。このようなニーズに対して「質問応答」という技術が注目されている。

質問応答とは自然言語で記述された質問文に対して、構造化されていない大量文書から、質問文に対する答えを見つけ出す技術のことである。例えば、「ダイナマイトを発明したのは誰ですか?」という質問に対しては、新聞やWebなどの構造化されていない大量文書から、「ノーベル」という答えが得られる。

質問応答システムでは、ユーザーは情報要求を質問文として入力し、その答えを受け取る。質問応答システムでは、ユーザーは検索に最適なキーワードを同定できる状態でなくとも、ユーザーの情報要求を質問文の形で明確化できる状態であれば、所望する情報を得ることができる [22]。

質問応答技術に関する研究は、主に評価型ワークショップを通して進められてきた。評価型ワークショップとは、評価データや正解データを含むテストコレクションの構築、研究コミュニティの形成、ならびに、評価手法・評価指標に関する研究の推進を目的とした参加型ワークショップを指す。質問応答技術に関する評価型ワークショップは、世界各地で開催されている。北米ではTREC(Text Retrieval Conference)[11]のQA Track, アジ

アでは NTCIR(NII-NACSIS Test Collection for IR Systems)[4]の QAC, ヨーロッパでは CLEF(Cross Language Evaluation Forum)[6]の Question Answering がある。

質問応答システムは一般的に、(1)重要語と回答タイプ¹を抽出する質問文解析部、(2)質問文解析結果をもとに、適切な文書の検索を行う文書検索部、(3) 検索文書中から回答タイプに基づく回答候補を抽出し、尤度情報に基づき、回答をランキングする回答絞り込み部の 3 つの処理部から構成される。このような質問応答システムの性能を適切に評価するために、評価技術に関する研究も進められている。

質問応答システムの性能評価は、回答のランキング性能や、精度と網羅性などを考慮したものが多い。例えば、MRR(Mean Reciprocal Rank)は、回答のランキング性能を考慮した評価指標であり、正解回答を上位にランキングするシステムに高いスコアを与える評価指標である[7][8][9][10]。回答の精度と網羅性を考慮した評価指標である MF 値は、正解を過不足無く回答したシステムに高いスコアを与える評価指標である[2][7][8]。回答の精度とランキング性能を考慮した評価指標である CWS (Confidence Weighted Score)は、回答を高精度かつ上位にランキングするシステムに高いスコアを与える[3]。システムが回答に付与した信頼度を考慮した評価指標である K-measure は、システムが付与した信頼度が高い回答が正解だった場合には高いスコアを与え、システムが付与した信頼度が高い回答が誤りだった場合には、スコアを大幅に減点する[1]。また、NTCIR5 QAC3 では、回答の詳細度が考慮された多段階評価も提案されている[9]。

1.2 研究の目的

1.1 節で述べたように、質問応答システムに関する様々な評価指標が提案されてきた。質問応答システムを、人間が質問に対する答えを見つけ出すプロセス（質問応答プロセス）の支援手段と位置づけて考えると、人間にとってコストが高く、高度な処理（ロバストな処理）に対応できるようなシステムを構築することが望ましい。このようなロバストな質問応答システムを実現するためには、人間が行うロバストな判断行動が参考となる。さらに、システム上でロバストな処理を実現するためには、どのような要素技術が必要であるかを検討すべきであろう。しかし、質問応答プロセスに求められるロバスト性とは、具体的にどのようなものなのかについての詳細な議論はなされていない。

¹ 国名や人名など、回答として求められている語句の属性

そこで、本研究では、ロバストな質問応答システム構築を目指し、質問の難易度とその評価手法について議論する。人間のロバストな処理を文書検索におけるクエリ操作であると考え、クエリ操作コストを検索ループ回数で近似するモデルを提案する。さらに、上記モデルをシステムに反映した、ロバストな質問応答システムの構成を示す。また、上記モデルを利用して、質問応答システムの性能評価基準の有効性を明らかにする。

1.3 論文の構成

本論文は7章で構成される。第1章の序論に続いて、第2章では、質問応答システムとその評価指標について説明する。まず、質問応答システムの一般的な構成（質問文解析部、情報検索部、回答絞込み部）について説明し、次に、質問応答システムに関する主な評価指標について説明する。

第3章では、予備調査の方法について説明する。本研究では、予備調査として人間に対して質問応答タスクを実施する。人間が質問応答プロセスを実施する際、処理過程において、どの処理のコストが高く、どのような操作を実施しているのかを把握するために、質問応答プロセスを細分化した調査を実施する。

第4章では、人間にとってのコストについて論じる。まず、人間が情報検索において実施するクエリ操作について定義する。次に、クエリ操作コストと情報検索に要したループ回数との関係から、ループ回数が多い質問文ほど複雑なクエリ操作を必要とすることを確認し、人間にとっての処理コストとして、情報検索におけるループ回数が有効であることを示す。

第5章では、人間にとってのコストとシステムの性能との関係を考察する。まず、人間が要したループ回数によって質問文を複数のグループに分類する。次に、質問文のグループ毎に質問応答システムの性能を比較することで、人間にとってのコストと質問応答システムの性能との関係を示す。

第6章では、今後の課題について述べる。最後に、第7章で本論文をまとめる。

第 2 章

質問応答システムの構成と評価指標

本章では、質問応答システムの一般的な構成と、質問応答システムに関する評価指標について説明する。2.1 節では、質問応答システムの一般的な構成について説明する。2.2 節では、質問応答システムに関する主な評価指標について説明する。

2.1 質問応答システムの構成

質問応答システムは一般に、質問文解析部、情報検索部、回答候補絞り込み部の 3 つの要素技術から構成される。以下では、これら 3 つの要素技術について説明する。

質問文解析部

質問文解析部では、自然言語による任意の質問から文書検索に必要な検索クエリを生成するとともに、質問文の回答タイプを決定する。回答タイプとは、回答として求められている語句の属性（国名、人名など）である。以下に質問文と、その検索クエリ・回答タイプの例を示す。

Q1 「ダイナマイトを発明したのは誰ですか？」

検索クエリ：{ダイナマイト，発明}

回答タイプ：人名（ダイナマイトの発明者を回答する質問である）

質問文中から検索クエリを抽出するには、形態素解析や固有表現抽出などの技術が使用されている。形態素解析とは、自然言語で書かれた文を形態素（言語で意味を持つ最小単位）の列に分割し、それぞれの品詞を判別する技術を指す。固有表現抽出とは、文書中の人名、地名などの固有名詞や金額、割合といった数量表現を抽出する技術である。形態素解析結果と固有表現抽出結果を合わせたものから重要語を判断して出力する。

形態素解析ツールとしては茶筌[16]やJUMAN[17]などが挙げられる。固有表現抽出ツールとしては、NExT[18]などが挙げられる。

文書検索部

文書検索部では、先の質問文解析部で得られた検索クエリや回答タイプなどの情報を基に、新聞や WWW(World Wide Web)などの大量文書中から文書検索を実施する。

文書検索には Namazu[15]のような全文検索エンジンや、Yahoo[13]や Google[14]などの WWW 検索エンジンを用いることが多い。また、磯崎ら[19]は、質問文中のキーワードの同義語（“オリンピック”に対する“五輪”）や反意語（“夫”に対する“妻”）を用いて文書検索を行う検索エンジンを開発し、質問応答システムに組み込むことで、高い性能を有するシステムを実現している。

回答候補絞り込み部

回答絞り込み部では、文書検索部で得られた検索文書の本文中から、回答タイプと一致する語を回答候補として抽出し、さらに回答に対してスコアを与えて、スコアが高い順に順位付けする。

先ほどの例を用いて説明すると、検索文書の本文中から、回答タイプ「人名」に合致した回答候補として「ウエンツ」や「ノーベル」、「トーマス」などが列挙される。次に、それらの回答をスコア順に 1 位「ノーベル」 2 位「ウエンツ」 3 位「トーマス」というように順位付けする。最終的に、正解である「ノーベル」を回答として出力する(図 1)。

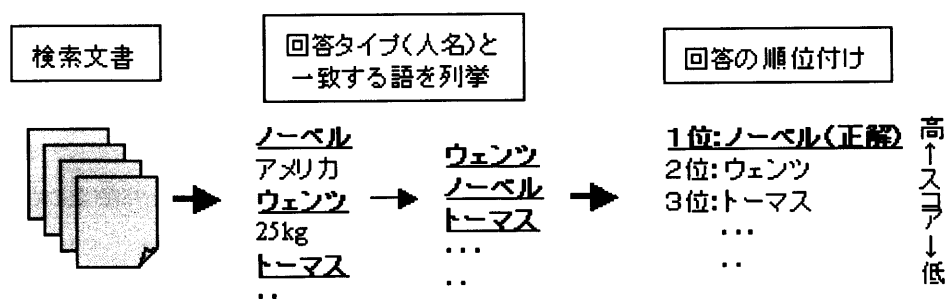


図 1. 回答候補絞り込み部の流れ

提案されている回答絞り込み手法として以下のようなものがある。石田ら[20]は WWW 上での質問文中のキーワードと回答候補の頻度情報を用いて、回答のスコアを計算する手法を提案している。また、石下らは[21]、回答候補のスコアの分布を利用することで、正解回

答と誤った回答を分離する手法を提案している。

2.2 質問応答システムに関する評価指標

質問応答システムに関する評価指標は、主に TREC や NTCIR や CLEF などの評価型ワークショップで提案されている。以下では、質問応答システムに関する主な評価指標について説明する。

MRR(Mean Reciprocal Rank)

MRR は QA システムが出力した回答のランキング性能を考慮した評価尺度であり、正解回答をより上位にランキングするシステムを高く評価する。

QA システムに対する *MRR*(*MRR(sys)*) は以下の式で求められる。

$$MRR(sys) = \frac{\sum_{i \in \text{questions}} RR_i(sys)}{\#\text{questions}}$$

$$RR_i(sys) = \frac{1}{rank(sys, i)}$$

rank(sys, i): 質問 *i* に対するシステム回答の中で、正解だった回答の順位

MRR は 1 つしか正解を持たない質問文を評価するのに適した評価尺度である。従って、*MRR* は基本的に 1 つしか正解を持たない質問文を評価対象とする QAC2-subtask1 などを用いられ、複数正解を持つ質問文を評価対象とする QAC2 subtask2 では別の評価尺度 (*Mean F-measure*) が用いられている。

例)

Q1 「日本で一番高い山は何ですか? (正解: 富士山)」

システム回答

1 位 阿蘇山

2 位 富士山 (正解)

3 位 三霊山

この質問文 Q1 の場合、正解が 2 位に現われたので、順位の逆数である RR は 1/2 となる。
他の質問文に対しても同様に RR を計算し、それらを平均したものが MRR となる。

MF 値(Mean F-measure)

MF 値 はシステムが出力した回答の精度を示す *presicion* と網羅性を示す *recall* を考慮した評価指標であり、正解を過不足無く回答したシステムに高いスコアを与える評価指標である。

QA システムに対する *MF 値(MeanF-measure(sys))* は以下の式で求められる。

$$MeanF - measure(sys) = \frac{\sum_{i=1}^{\text{質問文の総数}} F - measure_i(sys)}{\text{質問文の総数}}$$

$$F - Measure_i(sys) = \frac{recall_i(sys) \times precision_i(sys)}{\beta \times recall_i(sys) + (1 - \beta) precision_i(sys)}$$

$$recall_i(sys) = \frac{\text{質問 } i \text{ の回答に含まれる正解数}}{\text{質問文 } i \text{ の正解総数}}$$

$$precision_i(sys) = \frac{\text{質問 } i \text{ の回答に含まれる正解数}}{\text{質問 } i \text{ における回答数}}$$

F-measure は β の値によって、精度と網羅性の重みを変えるが、質問応答システムの評価では一般的に、 $\beta=0.5$ としたもので、つまり精度と網羅性の調和平均としたものが用いられている。

例)

Q2 「色の三原色は何色と何色と何色？ (答え：{赤，緑，青})」

システム回答 {グレー，赤，青，ピンク}

上記の例の場合、3 つの正解 {赤，青，緑} のうち、回答として 2 つの正解 {赤，青} を

出力しているので、回答の網羅性を示す *recall* は 2/3 となる。また、4 つの回答 {グレー, 赤, 青, ピンク} のうち、2 つの正解 {赤, 青} を出力しているので、回答の精度を示す *precision* は 2/4 となる。F 値は 1/2 と 2/4 の調和平均であり、計算するとおよそ 0.571 となる。他の質問文についても F 値を算出し、それらを平均したものが MF 値となる。

***K*-measure**

K-measure は QA システムが回答に対して与えた自信度 を考慮した評価尺度である。*K-measure* はシステムが自信を持って出力した回答が正解だった場合、高いスコアを与え、逆に自信を持って出力した回答が不正解だった場合、大幅に減点する。

各 QA システムに対する *K-measure* ($K(sys)$) は以下の式で求められる。

$$K(sys) = \frac{1}{\#questions} \cdot \sum_{i \in questions} \frac{\sum_{r \in answers(sys, i)} score(r) \cdot eval(r)}{\max \{R(i), answered(sys, i)\}}$$

#questions: 評価対象となる質問文の総数

$R(i)$: 質問 i の正解総数

$answered(sys, i)$: 質問 i に対する回答数

$score(r)$: システムの回答 r に対する自信度 (システムが回答 r に与えたスコア)

$eval(r)$: 査定者が回答 r を判断した結果

$$eval(r) = \begin{cases} 1 & \text{回答 } r \text{ が正解} \\ 0 & \text{回答 } r \text{ が重複正解} \\ -1 & \text{回答 } r \text{ が不正解} \end{cases}$$

***K1*-measure**

K1-measure は *K-measure* を一問一答形式の質問文評価に対応させた評価尺度である。各 QA システムに対する *K1-measure* ($K1(sys)$) は以下の式で求められる。

$$K1(sys) = \frac{\sum_{r \in answers(sys)} score(r) \cdot eval(r)}{\#questions}$$

#questions: 評価対象となる質問文の総数

score(r): システムが回答 **r** に与えたスコア (システムの回答 **r** に対する自信度)

eval(r): 査定者が回答 **r** を判断した結果

$$eval(r) = \begin{cases} 1 & \text{回答} r \text{ が正解} \\ -1 & \text{回答} r \text{ が不正解} \end{cases}$$

このように、質問応答システムに関する様々な評価指標が提案されている。しかし、質問応答システムを人間の質問応答プロセスの支援手段として位置づけて考えた場合、人間にとっての処理コストを考慮した評価指標を提案する必要がある。

そこで、本研究では人間にとってのコストを把握するための調査を実施する。

第 3 章

予備調査

本章では、人間に対して実施した質問応答タスク（以下、人間QAと呼ぶ）について説明する。

対象データとして NTCIR4 subtask2 Formal run[7]の質問文のうち 100 問を使用し、4 人の被験者を対象として人間QAを実施した。人間QAは、一般的な質問応答システムの構成（質問文解析、文書検索、回答絞り込み）を考慮し、2 種類の間 outputs を記録した(図 1)。

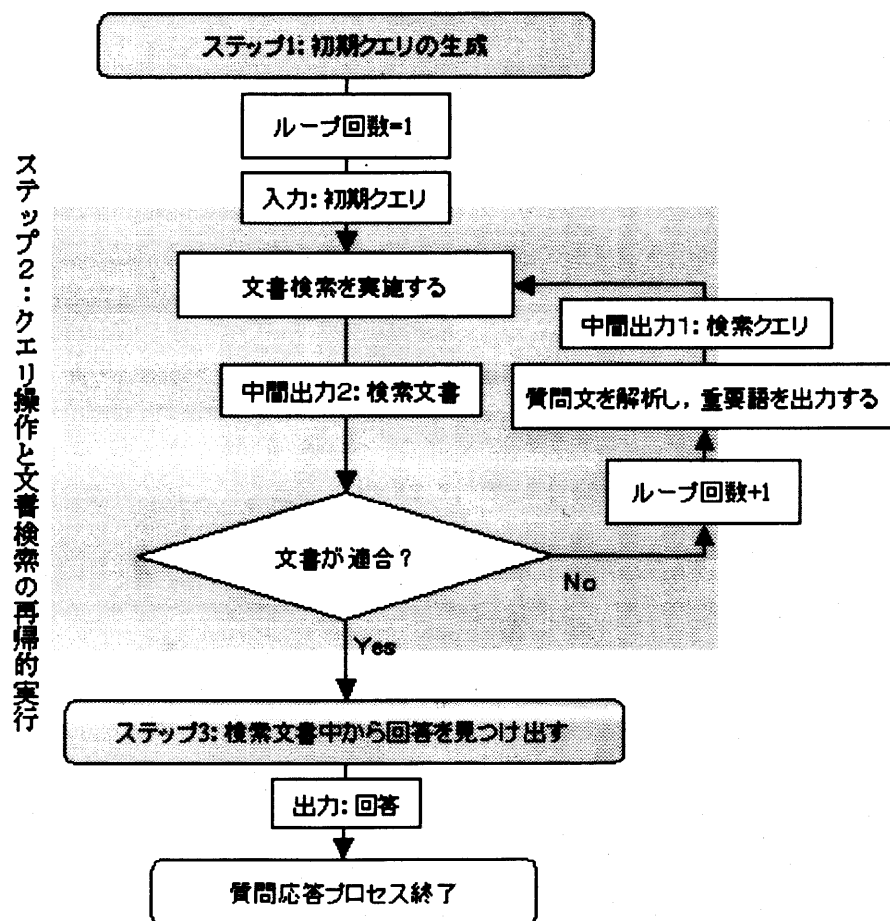


図 1. 人間の質問応答プロセスの流れ

ステップ 1：初期クエリの生成

このステップでは、ステップ 2 への入力（初期クエリ）を用意する。検索クエリは人間が質問文を読んで生成する。検索クエリの生成は質問文の答えを探すために検索エンジンを利用することを前提として行う。

ステップ 2：クエリ操作と文書検索の再帰的実行

クエリ操作と文書検索を再帰的に実施する。検索には検索エンジン[5]を用いる。検索対象として 1998 年と 1999 年の毎日新聞と読売新聞の述べ 4 年分を用いる。中間出力として、検索クエリと検索文書を記録する。検索文書が質問文に適合するまで検索を繰り返す。ループ回数が 10 を越えた質問文は調査対象から除外する。ステップ 2 の具体的な処理の流れを以下に示す。

処理 1. 初期クエリを受け取る。

処理 2. 初期クエリを用いて文書検索を実施する。

処理 3. 中間出力として、文書検索結果上位 10 件を出力する。

処理 4. 検索文書全文を参照し、検索文書の適合性判定を行う。適合性判定の基準は検索文書中に質問文に対する明確な答えを含むか否かである。検索文書が適合していると判断された場合、ステップ 2 を終了する。検索文書が適合していないと判断された場合、処理 5 に移る。

処理 5. ループ回数を 1 インクリメントする。

処理 6. 質問文を解析（参照）し、検索クエリを操作する。

処理 7. 中間出力として検索クエリを出力し、処理 2 に戻る。

ステップ 3：検索文書中から回答を見つけ出す

検索文書中から質問文に対する回答を見つけだして出力する。

第 4 章

質問応答におけるコスト

本章では、2 章で説明したステップ 2 で実施される処理のコストについて議論する。

ステップ 2 の処理コストは主に以下の 2 つが考えられる。1 つは検索文書の適合性を判定するコストであり、2 つ目は、検索クエリを検索文書に適合させるためのクエリ操作コストである。その中でも、クエリ操作コストが重要と考える。適合性判定は 2 値判断であり、コストの差異は生じにくいと考える。これに対して、クエリ操作コストは、クエリの数や曖昧性などによってコストに差異が生じると考えられる。

前章で述べた人間 QA に関して、ステップ 2 において実施したクエリ操作モデルを考える。一連のクエリ操作を行う 5 種類の基本操作を定義し、ループにおいてどのような組み合わせが選択されているのかを調査する。

3.1 クエリ操作の定義

まず、クエリ操作における基本操作を以下のように定義する。

削除

クエリに含まれるキーワードを削除する操作。

例：{日本, 内閣, 首相} → {日本, 首相}

※キーワード“内閣”を削除している。

追加

クエリにキーワードを追加する操作。

例：{ユーロ, 加盟国} → {フランス, ユーロ, 加盟国}

※キーワード“フランス”を追加している。

言い換え

クエリに含まれるキーワードを同じ意味を持つ語で言い換える解析操作。複数のキー

ワードを1つに結合する操作や(例 2), キーワードを複数に分割する操作 (例 3) も言い換え操作に分類する.

例 1: オリンピック → 五輪

例 2: {関東, 大震災} → 関東大震災

例 3: 関東大震災 → {関東, 大震災}

具体化

クエリに含まれるキーワードを具体化する操作. or 検索を実行したキーワード群の一部を削除する操作(例 2)も具体化操作に分類する.

例 1: 公判 → 82 回公判

※キーワード“公判”を“82 回公判”に具体化している.

例 2: {98 年 or 九十八年} → 九十八年

抽象化

クエリに含まれる重要語を抽象化する操作. キーワードに対して, or 検索を実行する操作も抽象化操作に分類する(例 2).

例 1: 82 回公判 → 公判

例 2: 九十八年 → {98 年 or 九十八年}

3.2 クエリ操作の分析

本節では, ステップ 2 におけるループ回数とクエリ操作の頻度との相関を分析する.

ここでは, 最終クエリを質問文に対する適合状態と見なし, 初期クエリと最終クエリを比較し, 初期クエリにはどのようなクエリ操作が必要だったのかを調査する.

質問文 Q1, Q2 に対して, 初期クエリと最終クエリとの比較した例を以下に示す.

Q1: 橋を構造で分類するとどんな種類がありますか。

最初の質問文解析結果：{橋，構造，種類，分類}

最終的な質問文解析結果：{橋，構造，分類}

Q1 における検索クエリを比較すると，“種類”を削除したことが解り，この質問文で必要となるクエリ操作は削除となる．

Q2: 国連安保理で拒否権を握っている国はどこですか。

最初の質問文解析結果：

{{国連安保理 or 国連安全保障理事会}，拒否権}

最終的な質問文解析結果：

{国連安保理，拒否権，フランス，常任理事国}

Q2 における検索クエリを比較すると，“国連安保理 or 国連安全保障理事会”が“国連安保理”に具体化され，“フランス”と“常任理事国”が追加されたことが解り，この質問文で必要となるクエリ操作は具体化と追加となる．

2 章で述べたとおり，ステップ 2 は複数回ループされる可能性があり，ループ回数が多いほど，多くのクエリ操作が必要になる．そこで，質問文をループ回数で分類した(表 2)．

表 2. 必要としたループ回数毎に質問文を分類

	ループ 1 回で終了	ループが 2 回必要	ループが 3 回以上必要	ループが 10 回	合計
質問数	69 問	18 問	12 問	1 問	100 問

表 2 を見ると，ループ 1 回で終了したケースは 69 問あった．2 回のループを要したケースは，18 問あり，3 回以上のループを要したケースは，12 問あった．また，10 回のループを要した質問文は 1 問あった．

さらに上記の結果において，クエリ操作とループ回数の関係を調査した(表 3)．10 回のループを要した 1 問²については質問文に対する適合文書が得られなかったため，調査対象から除外した．

² 除外した 1 問は正解がない質問文だった．

表 3. クエリ操作とループ回数の関係

クエリ操作	ループ 1 回で終了	ループが 2 回必要	ループが 3 回以上必要	合計
削除	—	10(48%)	9(29%)	19(37%)
言い換え	—	3(14%)	6(19%)	9(17%)
追加	—	0(0%)	6(19%)	6(12%)
抽象化	—	1(5%)	2(6%)	3(6%)
具体化	—	7(33%)	8(26%)	15(29%)
操作の合計	—	21	31	52

表 2 の結果をクエリ操作別（合計）で見ると，削除は 19，言い換えは 9，追加 6，抽象化は 3，具体化は 15 である．全体的に見れば，人間はクエリ操作において削除と具体化を多用していることが解る．

次に，クエリ操作とループ回数（1 回，2 回，3 回以上）の関係について考える（表 3）．表 3 を見ると，ループ回数 2 のケースで主に実施されるクエリ操作は削除や具体化であり，10（48%）に対して削除を実施し，7（33%）に対しては具体化操作を実施している．ループ回数 2 のケースにおいて必要なクエリ操作は主に削除や具体化であることが解る．ループ回数 3 以上ケースでは，全てのクエリ操作(削除，言い換え，拡張，具体化，抽象化)を実施している．

以上から，ループ回数の多さはクエリ操作処理の複雑さを反映しており，多くのループ回数を要する質問文は人間にとって処理コストが高い質問文であろうと推測出来る．

この知見を踏まえた評価をシステムに対して実施することで，人間にとって処理コストの高い質問に対応できるシステムをより高く評価出来る可能性がある．

第 5 章

ループ回数とシステムの性能との相関

本章では、ループ回数と質問応答システムの性能との相関を調査する。

質問応答システムの性能評価には、QAC2 subtask2 で使用された評価尺度である、*MF 値 (Mean F-measure)*を用いる。*MF 値* は各質問文につき求めた *F-measure* の平均であり、次の式(1)で求められる。

$$MeanF-measure = \frac{\sum_{i=1}^{\text{質問文の総数}} F-measure_i}{\text{質問文の総数}} \cdots \text{式 (1)}$$

ただし、質問 i に対するシステムの F 値($F-measure_i$)は以下の通り

$$F-measure_i = \frac{recall_i \times precision_i}{\beta \times recall_i + (1 - \beta) precision_i}$$

$$recall_i = \frac{\text{質問 } i \text{ の回答に含まれる正解数}}{\text{質問文 } i \text{ の正解総数}}$$

$$precision_i = \frac{\text{質問 } i \text{ の回答に含まれる正解数}}{\text{質問 } i \text{ における回答数}}$$

F-measure は回答の *recall* (網羅性) と、回答の *precision* (精度) によって求められる。*F-measure* は β の値によって、精度と網羅性の重みを変えるが、QAC2 subtask2 では、 $\beta=0.5$ としたもので、つまり精度と網羅性の調和平均としたものが用いられている。

ループ回数 (1 回, 2 回, 3 回以上) で分けた質問群に対して QAC2 参加システム上位 6 システムの評価結果 (*MF 値*) との対応を調べた。その結果を図 3 に示す。

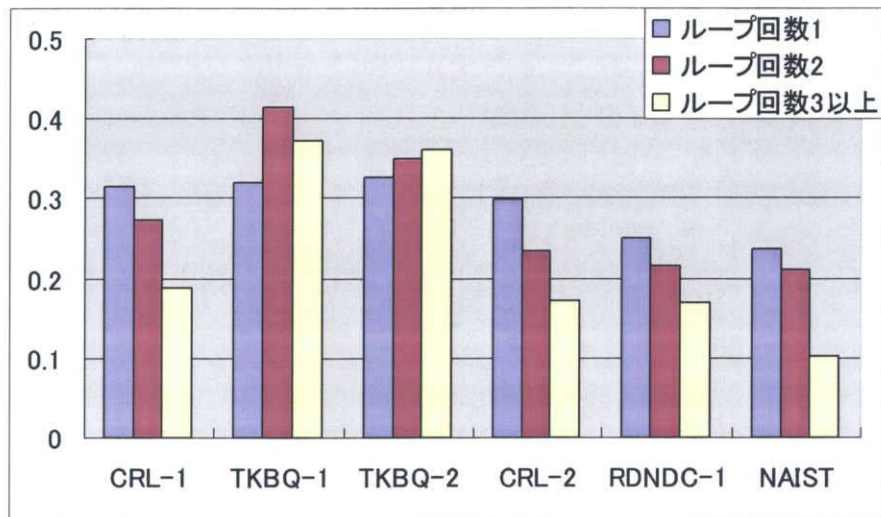


図 3. 質問文解析回数で分類した質問群におけるQAシステムのMF値

図3を見ると、全体的な傾向として、ループ回数が増えるにつれて、MF値が低下する傾向が見られた。このことから人間にとって処理コストの高い質問文はシステムにとっても苦手であることが考察できる。

個々のシステムについて見ると、TKBQ-1、TKBQ-2については、ループ1回目より、2回目、3回目のMF値が同等以上の結果を示した。したがって、TKBQ-1、TKBQ-2は人間にとってコストの高い処理に対応するロバストなシステムであると言える。[9]によれば、TKBQ-1、TKBQ-2は回答タイプを考慮したクエリ拡張処理や、検索文書の内部評価を実装しており、これらの処理がロバスト性をもたらしたと思われる。

多くのループ回数を要する質問文は、クエリ操作コストが高く、難易度が高い質問文である。クエリ操作コストを上げる要因として、初期クエリに含まれる言い換え表現の有無を考えた。例えば、“関西学院大学の校歌は誰が作りましたか。”という質問文の場合、重要語“関西学院大学”に対して“関学”などの言い換え表現が考えられる。初期クエリに言い換え表現を含み、or検索実施した質問文では、それらを検索のために最適化する操作が必要となり、ループ回数が多くなると考えた。

そこで、初期クエリにor検索を含むかどうかで、質問文を2つに分類し、それぞれのループ回数を調査した(表3)。また、or検索の有無で分類した質問群における、QAC参加システムのMF値を調査した(図4)。

表 3. or 検索の有無で分類した質問群における平均ループ回数

	or無し	or有り
質問数	39 問	61 問
平均ループ回数	1.28 回	1.85 回

表 3 を見ると，or 検索無しの質問文の平均ループ回数は 1.28 回，or 検索有りの質問文の平均ループ回数は 1.85 回であった．この結果から，人間にとっての質問文のコストを上げる要因の 1 つとして，言い換え表現を含むか否かが考えられる．

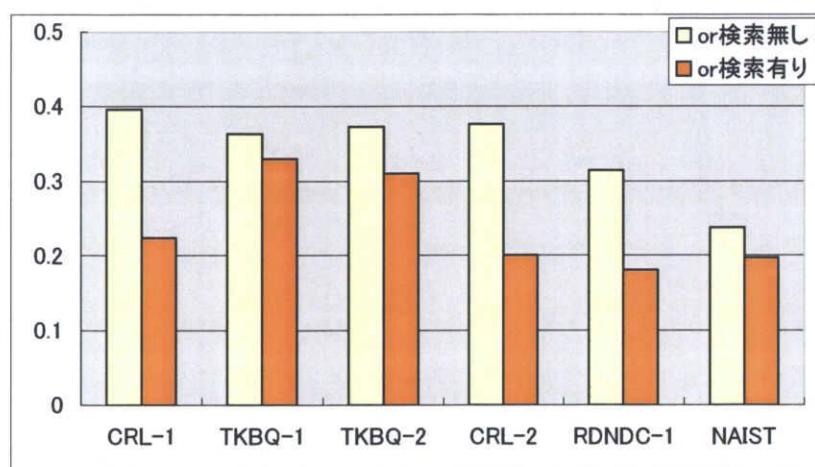


図 4. or 検索の有無で分類した質問群における
QAシステムのMF 値

図 4 を見ると，全体の傾向として初期クエリで or 検索を実施した質問文の方が，or 検索を実施しなかった質問文よりも MF 値が低いことが解る．言い換え表現を含む質問文は，

システムにとっても難しい質問文であると言える。

以上の知見を踏まえて、人間にとっての難易度を考慮したシステム評価のパラメータとして、ループ回数を考慮した評価尺度 LWF (Loop Weighted F-measure) を考えた。 LWF は以下の式で求められる (式 2)。

$$LWF_i = F-measure_i \times \frac{L_i}{\sum_{i=1}^{questions} L_i} \times questions \cdots \text{式(2)}$$

$F-measure_i$: 質問 i に対するシステムの F 値

L_i : 人間が質問 i に要したループ回数

$questions$: 評価対象となる質問数

各 QAC2 参加システムに対して、各々の LWF 値 の平均 ($MLWF$ 値) を計算する(式(3))。

$$MLWF = \frac{\sum_{i=1}^{questions} LWF_i}{questions} \cdots \text{式(3)}$$

QAC2 参加システム上位 6 位の $MLWF$ 値 と MF 値 を図 5 に示す。

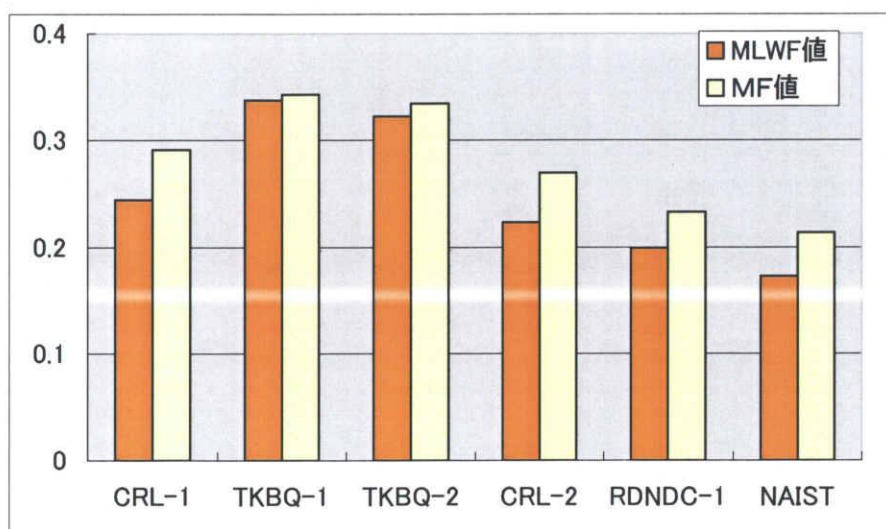


図 5. QAC2 参加システムの MLWF 値と MF 値

全体的な傾向として、MLWF 値は MF 値よりも低い傾向が見られた。前に述べたように全体的にロバスト性が低いと言える。何故かと言うと、ロバスト性が低いシステムは多くのループ回数を要した質問に対して有効に働かないからである。ところが、その中で TKBQ-1, 2 は MLWF 値と MF 値の差が極端に小さい。これは、TKBQ-1, 2 がループ回数の影響を受けないロバスト性の高いシステムであることを示唆している。

人間にとってのループ回数を考慮したシステムの評価を実施することにより、人間にとって処理コストが高い質問に対してロバストに対応出来るかどうかを定性的に評価することが出来ると言える。

第 5 章

今後の課題

今後の課題として、以下の点が挙げられる。

1. 検索エンジンによる結果の差異

今回の実験では、質問文に対する適合文書を得るために要するクエリ操作やループ回数を検索エンジン Kabayaki[5]のみを用いて調査した。しかし、検索エンジンの性能によって、各質問に要するクエリ操作やループ回数に差異が生じる可能性がある。例えば、自動的にキーワードの言い換え検索を行う検索エンジンを用いた場合、キーワードの言い換えを行わなくても質問文に適合した文書が得られ、言い換え操作の回数や必要とするループ回数が減少する可能性がある。

このような検索エンジンの性能差を考慮した実験結果を得るためには、複数の検索エンジンを用いて再実験をする必要がある。

2. 人間にとってのコストを考慮した質問応答システムの構築

本論文では、人間にとってのコストを考慮した評価尺度を考察した。人間にとってコストの高い質問文に対応可能な質問応答システムを構築することが今後の課題として挙げられる。今回得られた知見から、このような質問を解くために必要となる技術を検討し、システムに組み込むことで、よりロバストな質問応答システムを構築する予定である。

3. 被験者の層別調査

今回は人間 QA の被験者として情報検索に熟練した人間を用いた。しかし、質問応答システムのユーザーを想定した場合、情報検索熟練者の他にも、あまり情報検索を利用したことがない人間が考えられる。想定するユーザー層が異なれば、実験結果にも差異が生じることが予想される。質問応答システムのユーザーを様々な観点から層別に調査する必要がある。

第 6 章

結論

本論文では、人間が質問応答処理を実施する際、どのような操作、および処理コストが存在するかを把握するため、人間に対して質問応答タスクを実施した。

人間が行う操作の中でも、情報検索過程で実施される質問文解析(検索クエリ)に注目し、クエリ操作の分析を行った。クエリ操作は人間にとっての処理コストに直結していることが解った。また、人間にとっての質問文の処理コストについても考察し、人間にとっての処理コストをシステムの MF 値の相関を調査した。全体として、人間にとって処理コストが高い質問文はシステムにとっても難易度が高い質問文であることが解った。

今後は、今回の分析結果をシステムに反映し、よりロバストなシステムを構築する予定である。また、複数の検索エンジンを用いた再実験を実施し、検索エンジンの性能差を考慮した考察を行う予定である

参考文献

- [1] A. Vallin et al. Overview of the CLEF 2004 Multilingual Question Answering Track. In Proceedings of the CLEF 2004 Workshop, Bath, United Kingdom, September 2004.
- [2] E. M. Voorhees. Overview of the TREC 2003 Question Answering Track. In Proceedings of the Twelfth Text REtrieval Conference (TREC 2003), volume 500-255 of NIST Special Publication, pages 54-68, 2003.
- [3] E.M.Voorhees. Overview of the TREC 2002 Question Answering Track. In E.M.Voorhees and L.P.Buckland, editors, Proceedings of the Eleventh Text Retrieval Conference (TREC 2002), volume 500-251 of NIST Special Publication, 2002.
- [4] NTCIR (NII-NACSIS Test Collection for IR Systems) Project.
URL: <http://research.nii.ac.jp/ntcir/>
- [5] Kabayaki :: 日本語全文検索環境 Kabayaki: オフィシャルサイト.
URL: <http://www.kabayaki.jp/ntcir/>
- [6] The Cross-Language Evaluation Forum (CLEF). URL: <http://www.clef-campaign.org/>
- [7] Jun'ichi Fukumoto, Tsuneaki Kato, Fumito Masui. Question Answering Challenge for Five ranked answers and List answers Overview of NTCIR-4 QAC2 Subtask 1 and 2, pp.283-290
- [8] J.Fukumoto, T.Kato, and F.Masui. Question Answering Challenge (QAC-1). An Evaluation of Question Answering Task at NTCIR Workshop 3. In Keizo Oyama, Emi Ishida, and Noriko Kando, editors, Proceedings of the Third NTCIR Workshop on Research in Information Retrieval, Automatic Text Summarization and Question Answering. National Institute of Informatics, 2003.
- [9] Tsuneaki Kato, Jun'ichi Fukumoto and Fumito Masui. An Overview of NTCIR-5 QAC3
- [10] Tetsuya Sakai, Yoshimi Saito, Yumi Ichimura, Toshiba ASK Mi at NTCIR-4 QAC2. NTCIR-4 QAC Experiments at Matsushita. NTCIR Workshop 4 Meeting, pp.387-394
- [11] Text REtrieval Conference (TREC) Home Page. URL: <http://trec.nist.gov/>
- [12] B. Magnini, S. Romagnoli, A. Vallin, J. Herrera, A. Penas, V. Peinado, F.

- Verdejo, and M.deRijke.TheMultiple Language Question Answering Track at CLEF2003.InC.Peters, J. Gonzalo, M. Braschler, and M. Kluck, editors, Comparative Evaluation of Multilingual Information Access Systems. Results of the CLEF2003 Evaluation Campaign , volume 3237 of LNCS , pp479-pp.495.Springer-Verlag, 2004.
- [13]Google, <http://www.google.co.jp/>
- [14]Yahoo, <http://www.yahoo.co.jp/>
- [15]全文検索システム Namazu:
URL: <http://www.namazu.org/>
- [16]形態素解析システム茶筌
URL: <http://chasen.naist.jp/hiki/ChaSen/>
- [17]日本語形態素解析システム JUMAN
URL: <http://nlp.kuee.kyoto-u.ac.jp/nl-resource/juman.html>
- [18] NExT - a Named Entity Extraction Tool
URL: <http://www.ai.info.mie-u.ac.jp/~next/next.html>
- [19]Hideki Isozaki,NTT's Question Answering System for NTCIR QAC2,NTCIR Workshop 4 Meeting Working Notes of the Fourth Workshop Meeting National Institute of Informatics June 2-4 2004,pp326-pp332.
- [20] 石田健二, 梶井文人, 河合敦夫: "World Wide Web を利用した質問応答における回答絞り込み技術", 第 11 回言語処理学会年次大会発表論文集, C2-4, 2005.3.
- [21] 石下円香, 森辰則, 解候補スコアの分布を利用したリスト型質問応答, 言語処理学会第 11 回年次大会, D5-6, 2005.
- [22]徳永健伸, 言語と計算 5 情報検索と言語処理, 財団法人 東京大学出版会, pp1-pp5,1999.

謝辞

本研究を遂行するにあたり，日頃からご指導，ご鞭撻頂きました井須尚紀教授，河合敦夫助教授，榊井文人助手に深く御礼申し上げます．また，多岐にわたり便宜を図って頂きました岡保子事務官並びに，田中みゆき事務官に深く感謝します．ここに書ききることは決してできませんが，人工知能研究室で同じ時間を共有してきた先輩，同輩，後輩の全員に心より感謝致します．