社団法人 情報処理学会 研究報告
IPSJ SIG Technical Reports

2005－NL－168（8）
2005／7／22

質問応答システムの正解順位とユーザ満足率の関係について

國分 智晴† 酒井 哲也† 齋藤 佳美† 筒井 秀樹†
真鍋 俊彦† 藤井 寛子† 小山 誠†

† （株）東芝 研究開発センター 〒212-8582 川崎市幸区小向東芝町1
E-mail:
†{tomoharu.kokubu,tetsuya.sakai,yoshimi.saito,hideki.tsutsui,toshihiko.manabe,hiroko.fujii,makoto3.koyama}@toshiba.co.jp

あらまし 近年，質問応答システムに関する研究が盛んに行われているが，システムの回答精度と，どれぐらいのユーザが満足するかとの関係が明確になっていなかった．そこで，質問応答の利用シーンとしてデスクトップ型およびモバイル型の2種類を想定し，シーン別に正解順位に対する満足の度合いを評価するアンケートを行った．この結果，回答候補を一括提示するデスクトップ型におけるユーザ満足率は質問応答の一般的な評価尺度である逆数順位と似た曲線となるが，回答候補をひとつずつ提示するモバイル型におけるそれは順位とほぼ比例することが分かった．このため，ユーザ満足率の観点からシステムの目標精度を設定するには，平均逆数順位ではなく正解順位の分布を決定することが望ましいことを示す．

キーワード 質問応答，アンケート，ユーザ満足率

The Relationship between Answer Ranking and User Satisfaction in a Question Answering System

Tomoharu KOKUBU†, Tetsuya SAKAI†, Yoshimi SAITO†, Hideki TSUTSUI†,
Toshihiko MANABE†, Hiroko FUJII†, and Makoto KOYAMA†

† R&D Center, Toshiba Corporation 1 Komukai, Toshiba-cho, Saiwai-ku, Kawasaki-shi, 212-8582 Japan
E-mail:
†{tomoharu.kokubu,tetsuya.sakai,yoshimi.saito,hideki.tsutsui,toshihiko.manabe,hiroko.fujii,makoto3.koyama}@toshiba.co.jp

Abstract Although research in effective Question Answering (QA) has become active in recent years, it was not clear how system accuracy affects user satisfaction in practice. We therefore devised two practical scenarios in which QA may be useful (namely, desktop and mobile) and conducted a questionnaire survey for each scenario. The objective was to clarify the relationship between the rank of a correct answer and the proportion of satisfied users. Results show that, while the graph of the proportion of satisfied users resembles that of Reciprocal Rank for the desktop case, it is almost proportional to the rank for the mobile case. Thus, from the viewpoint of user satisfaction, one should set a goal in terms of the distribution of the correct answers instead of a single Mean Reciprocal Rank value.

Key words Question Answering,questionnaire,user satisfaction

1. はじめに

近年，情報検索，情報抽出，自然言語処理の分野において質問応答システムが注目されている．従来の文書検索システムが文書のリストを出力するのに対し，質問応答システムは例えば「富士山の高さは？」のような事実を聞くような質問に対して回答そのものを出力するものであり，米国のTREC[1]，日本でのNTCIRのようなコンテスト型ワークショップにおいて盛んに研究されている．我々もNTCIR-4の質問応答タスク[2]（以下QAC2）への参加を通し，精度向上のための研

究開発を行ってきた[3][4]．システムの精度の評価尺度としては，NTCIR-4 QAC2 Subtask 1で用いられたMRR（Mean Reciprocal Rank：システムが出力した正解文字列の順位の逆数を質問セットについて平均したもの）を用いてきた．

質問応答システムの実用化のためには，ユーザの満足する精度の実現が必須となる．そこで本研究では，アンケート調査により，正解順位と，どれくらいのユーザが満足するかの対応づけを行うことにした．まず，デスクトップでの利用とモバイル端末上での利用という2つの利用シーンを想定し，各シーンについて仮想質問および正答をひとつ含む回答候補セットを作成した．次に，各質問とともに正答の順位をランダムに変化させた回答候補リストを被験者に提示して評価を行ってもらい，何人の被験者が満足したか（ユーザ満足率）を求め，正解順位との対応グラフを作成した．そして，このグラフをもとに精度および利用シーンとの関係について分析を行った．

以下，2章で，アンケートの内容および実施方法，3章で，正解順位とユーザ満足率の関係についての分析結果，およびアンケートのコメントから得られた知見について報告し，4章で結論と今後の課題を述べる．

2. アンケートの内容および実施方法

本章では，アンケートの内容および実施方法について説明する．

2.1 ユーザ満足率に関する要因

今回のアンケートの目的は質問応答システムの出力する正解の順位とユーザ満足率の関係を明らかにすることであるが，実際には正解順位以外にも以下のような要因がユーザの満足の度合いを左右すると考えられる．

- ユーザの背景知識

ユーザが自分の質問に関連する知識をどれくらいもっているかによりユーザの満足の度合いは左右されるであろう．

- 扱える質問の多様さ

ユーザの様々な言い回しを正しく解釈する機能や，名称や日付，数値を解答とするようなfactoid質問だけでなく，方法，定義，Yes/No質問を扱える機能，新鮮な情報に関する質問に対処する機能の有無は当然ユーザの満足の度合いを左右するであろう．

- 提示する回答の内容

回答候補リストにおける正解順位以外にも，そもそも回答候補をいくつずつ，いくつまで見せるか，根拠文書を見せるか否かなどのインタフェースも満足の度合いを左右する可能性がある．さらに，個々の正解および不正解の質も重要であろう．例えば，人名を尋ねる質問に対して製品名を回答したり，文字列の切り出し方を間違えた固有表現抽出結果を回答したりすると，満足の度合いは低下するであろう．

- 根拠文書および知識源の質

各回答候補に根拠文書を付随させる場合には，適切な根拠文書を適切に切り出して提示する必要がある．また，根拠文書の抽出元である知識源自体の質や信頼性も，ユーザの満足の度合いに影響する可能性がある．例えば，根拠文書として企業や自治体等の公式サイトから得られたものを提示する場合と，個人の日記や掲示板から得られたものを提示する場合では，ユーザの満足の度合いが異なるかも知れない．

しかし今回は，NTCIR-4 QAC2 Subtask 1[2]のタスク定義を参考に，正解順位以外の要因に関しては以下のように条件を固定することにした．

- 質問はfactoid型に限定する
- ひとつの質問に対し，回答中に必ずひとつ正解を含む
- 質問に対する回答候補数は5つとする
- 各回答候補とともに1件の根拠文書を提示する
- 根拠文書の長さは300文字以内とする
- 出力する回答候補は，固有表現として正しい文字列に限定する．（例：「富士山」が正解である場合に，「士山」は出力不可だが，「山田太郎」は出力可）

さらに各根拠文書に日付情報を付与し，各回答の新鮮さが被験者に分かるようにした．また被験者の背景知識に関しては今回のアンケートでは考慮しないことにした．被験者に一度に提示する回答候補の数は利用シーン毎に決定した．

2.2 利用シーンの決定

質問応答システムの利用シーンとしてWeb等を対象に汎用的な検索を行う「デスクトップ型」と，携帯電話など画面が小さいインタフェースで利用する「モバイル型」の2つのシーンを想定した．なお，質問応答の利用シーンとしてはマニュアルやFAQサイトを対象に，家電の操作方法などを分かりやすく説明することを目的とした家電ヘルプ[5]なども考えられるが，これは方法を聞く質問などを扱う必要があるので今回は対象外とした．デスクトップ型とモバイル型の利用シーンを被験者が実感できるように，以下のようにインタフェースの違いを持たせた．

- デスクトップ型では回答および根拠文書を一括提示し，モバイル型では回答および根拠文書をひとつずつ提示する
- モバイル型では被験者が旅行先などの利用シーンを想像しやすいように，質問とともに，現在地として想定している地名を表示する

2.3 質問，回答セットの作成

アンケートの質問，回答セットは以下の手順で作成した．

（1） 質問の作成

質問の内容が偏らないように，デスクトップ型では人名，場所，数値，その他の4種類，モバイル型では地名，時間，その他の3種類の回答タイプを用意した．その上で，本研究の著者（7人）が実際にシステムに聞きたい内容を質問文にした．各回答タイプ毎に最低5種類の質問を用意した．

（2） 正解，不正解の作成

Webを検索して質問に対する正解回答を1つ，不正解回答を4つ作成した．

（3） 根拠文書の作成

正解回答，不正解回答ともに，Web上に実際にあるテキストデータを参考に，300文字以内に収まるように根拠文書を独自に作成した．その際，正解回答の根拠文書を読むことで，その回答が正解であることが分かるように留意した．

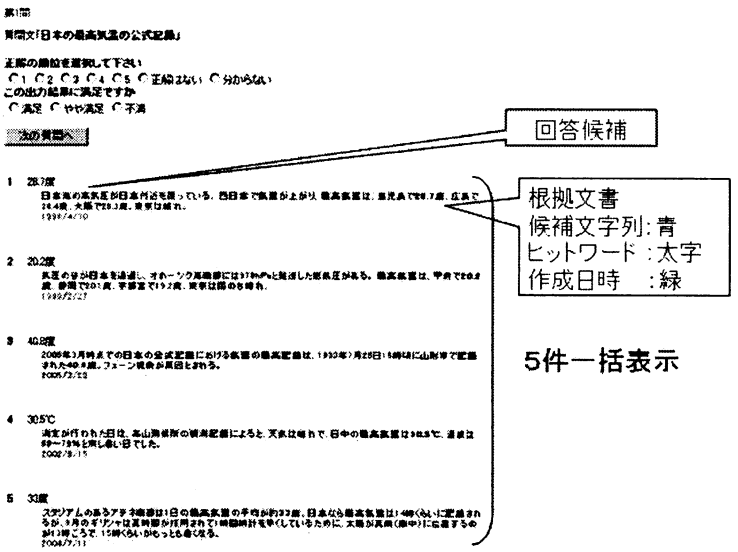

図 1 デスクトップ型のアンケートの画面例

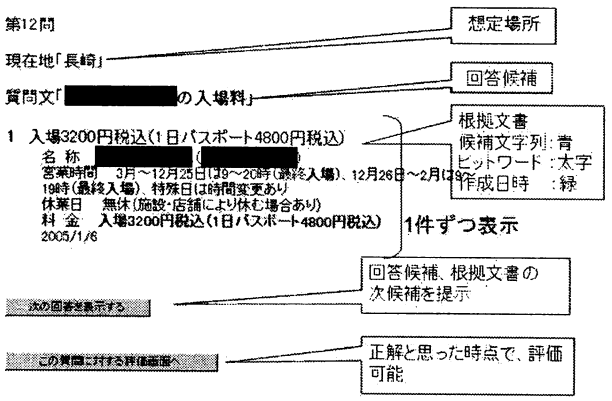

図 2 モバイル型のアンケートの画面例

（4） 質問の選択

被験者がアンケートで正解と判断する回答が実際の正解と異なっていた場合には，根拠文書の分かりづらさや，紛らわしい不正解の存在など，正解順位以外の要因が被験者の評価に影響を与える可能性が高い．そのためあらかじめ著者内でアンケートの予備実験を行い，正解選択率（被験者が正解と思い選択した回答が本当に正解であった割合）の低い質問を除き，最終的にデスクトップ型 20 問，モバイル型 15 問を選択した（各回答タイプあたり 5 問）．

（5） 正解の出現順位，質問の出題順の決定

各質問の正解の出現する順位は，各回答タイプ毎に正解の順位が 1 位から 5 位まで 1 度ずつ出てくるという条件の下でランダムに決定した．各質問の出題順はランダムに決定した．

2.4 実施方法

アンケートは Web 上で実施した．被験者数は，デスクトップ型で 27 人，モバイル型で 25 人であった．モバイル型の方が人数が 2 人少ないが，これは，時間がなくデスクトップ型のアンケートのみ提出した被験者が 1 人，モバイル型では 2 位以下の回答候補を見ることができないと誤解してしまった被験者が 1 人いたためである．全ての被験者に対してデスクトップ型 20 問，モバイル型 15 問を評価してもらった．質問の出題順，および各質問に対する回答候補の提示順は，全被験者共通にした．

デスクトップ型は，図1に示した例のように，質問文に対して回答，根拠文書，および根拠文書の日付情報のセットを5つ一括で提示する．根拠文書はヒットワードを太字，回答文字列を青色で表示した．正解が何番であるか（なければ「ない」分からなければ「わからない」）を選択するチェックボックスと，結果に対して「満足」，「やや満足」，「不満」を選択するチェックボックスをつけた．このチェックをしてもらうことにより各質問の順位に対する満足の度合いを計測した．「次へ」ボタンを押すと次の質問が提示される．

モバイル型は，図2に示した例のように，各質問文に対して，想定している現在地，回答，根拠文書，および根拠文書の日付情報のセットを1件ずつ提示する．「次の回答を表示する」ボタンを押すことで次候補の回答，根拠文書等を見ることができ，最大5位まで提示する．また，途中で正解が見つかった場合，全ての回答候補を見なくても評価できるように，常に「この質問に対する評価画面へ」ボタンを用意した．

またアンケートの最後に意見・感想を自由記述するフォームを用意した．

3. アンケート結果の分析

本章では実際のアンケート結果の分析を報告する．3.1節で，各正解順位に対するユーザ満足率グラフの傾向分析について，3.2節で，得られたユーザ満足率とRR（Reciprocal Rank：システムが出力した正解文字列の順位の逆数），MRRとの関係について，3.3節で，開発中の質問応答システムASKMi[3][4]の平均ユーザ満足率について述べる．また，3.4節では，アンケートのコメントから得られる知見について，3.5節では，アンケートでユーザが正解の選択を間違った例について述べる．

3.1 各精度に対する満足率

前述したように，被験者の正解選択率が低い質問（75％以下）は，正解順位以外の要因が被験者の評価に影響を与える可能性が高いため分析対象から除外した．そのため実際にはデスクトップ型，モバイル型の質問数はそれぞれ，18問，11問となった．残った質問に対する正解選択率はそれぞれ93.2%，91.6%であった．各正解順位に対するユーザ満足率は以下のように求めた．

（1） 各質問に対して「満足」「やや満足」「不満」をカウントし，全被験者の平均値を求める

（2） 上記結果の正解順位毎の平均を求める

この内訳を表1に，デスクトップ型，モバイル型毎に百分率であらわした図をそれぞれ図3，図4に示す．さらにデスクトップ型とモバイル型それぞれについて，「満足」のみを集計したもの（以下「満足」）と，「満足」と「やや満足」の合計を集計したもの（以下「満足＋やや満足」）の2種類のレベルのユーザ満足率を定義した．「満足」はユーザが真に満足するための要求精度の指標．「満足＋やや満足」はユーザの不満を少なくするために必要な要求精度の指標となると考えられる．正解順位-ユーザ満足率グラフを図5に示す．

図5の正解順位-ユーザ満足率グラフから，利用シーンや満足のレベルによらず，正解順位が上がるとユーザ満足率も向上

順位/シーン「レベル」	1	2	3	4	5
デスクトップ型「満足」	0.17	0.32	0.33	0.40	0.85
デスクトップ型「やや満足」	0.48	0.44	0.50	0.5	0.13
デスクトップ型「不満」	0.35	0.23	0.16	0.1	0.02
モバイル型「満足」	0.18	0.36	0.54	0.62	0.89
モバイル型「やや満足」	0.5	0.48	0.36	0.34	0.07
モバイル型「不満」	0.32	0.16	0.1	0.04	0.04

表1 利用シーン毎のユーザ評価の内訳

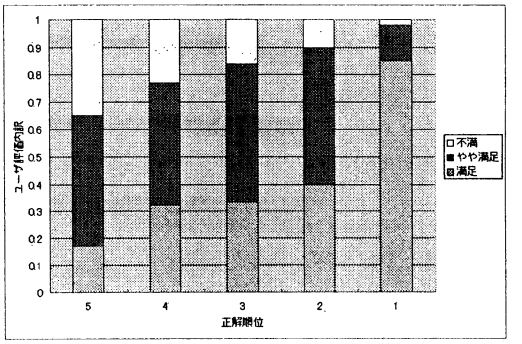

図3 正解順位-ユーザ評価内訳（デスクトップ型）

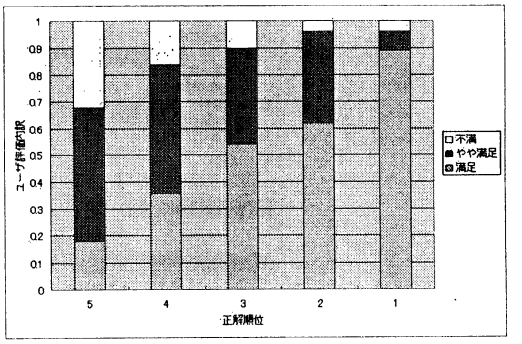

図4 正解順位-ユーザ評価内訳（モバイル型）

することが分かる．

「満足」に関しては次のことが言える．

- デスクトップ型の「満足」のユーザ満足率は，2位以下の正解の順位はそれほど影響しない
- モバイル型の「満足」のユーザ満足率は，順位にほぼ比例して低下する

この結果はデスクトップ型では，正解らしきものが1位にない場合に結局2〜5位の回答候補まで見ている可能性があるためと考えられる．またモバイル型では，正解らしきものより下位を見る必要がないことを反映していると考えられる．このためデスクトップ型では2位以下の精度向上はユーザ満足率向上

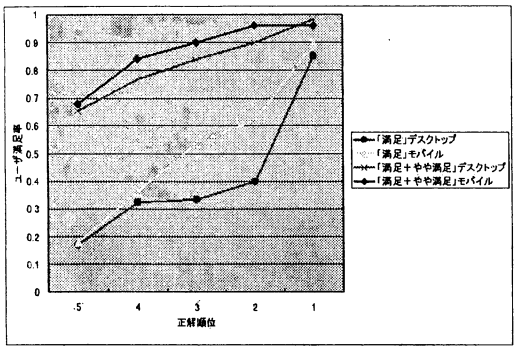

図 5　正解順位-ユーザ満足率のグラフ

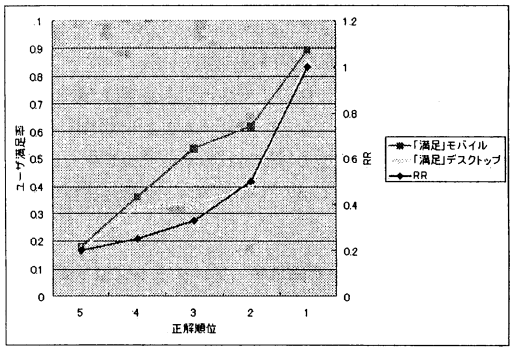

図 6 「満足」のユーザ満足率及び RR の，順位との相関グラフ

につながらない可能性がある．一方，モバイル型では満足率が正解順位にほぼ比例しているため，2 位以下の精度向上もユーザ満足率向上につながる可能性がある．このように回答の提示方法もユーザ満足率に影響を与えることが考えられる．

「満足＋やや満足」に関しては次のことが言える．

- 「満足＋やや満足」ではデスクトップ型，モバイル型とも正解順位が下がるとユーザ満足率も緩やかに低下していく傾向を示し，5 位に正解を提示した場合でも，ユーザ満足率が 0.6 以上である

また，両レベルを比較すると次のことが言える．

- 「満足」と「満足＋やや満足」では，1 位が正解のとき以外は大きな差がある

ある質問に対してユーザが真に満足するために必要な要求精度と，不満を少なくするために必要な要求精度には大きな差がある．ユーザが真に満足する精度を実現するには 1 位に正解を出力できるようにタスクを限定するか，正解が 2 位以下になっても満足が得られるようなインタフェース上の工夫（例えば回答提示方法の改良）等が必要であろう．

さらに，利用シーンで比較すると次のことが言える．

- 各満足レベルにおける 1 位に正解を出力した際のユーザ満足率は，デスクトップ型，モバイル型で同等である

これは，5 件の回答候補を一括提示するデスクトップ型であっても，正解が 1 位にある場合は 2 位以下の回答候補をチェックする労力がほとんどかからないため，候補を 1 件ずつ提示するモバイルの場合と同等の評価が得られたものと考えられる．

3.2 RR，MRR とユーザ満足率の関係

この節では前節で得られたユーザ満足率と従来用いてきた評価尺度である RR，MRR との関係について考察する．NTCIR-4 QAC2 Subtask1 等では質問に対する精度として RR（Reciprocal Rank:システムの出力した正解文字列の順位の逆数）を用いていた．正解順位に対する RR と前節で得られた「満足」の正解順位-ユーザ満足率グラフを重ねたものを図 6 に示す．この図から以下のことが言える．

- デスクトップ型の「満足」のユーザ満足率と RR は正解順位が 2 位で値が落ち込み，その後は緩やかに値が下がるという点で傾向が似ている
- モバイル型の「満足」のユーザ満足率は正解順位にほぼ比例し，RR とは傾向が異なる

すなわち，RR はデスクトップ型における「満足」のユーザ満足率の比較的よい近似となっているが，モバイル型については，正解順位に比例した尺度の方が適している可能性がある．

次に，ユーザ満足率とシステムの評価尺度として広く用いられている MRR との関係について考察する．MRR は以下の式で表される．

$$MRR = \frac{1}{C}\sum_{i=1}^{5} RR_i C_i \qquad (1)$$

$RR = 1/i(i \leqq 5)$，C_i はシステムにおける正解順位 i の質問数，C は全質問数である．

ここで以下の 2 つのシステムを仮定する．

- システム A　1 位に正解を出力したものが 1 件，正解が出せないものが 1 件
- システム B　2 件とも 2 位に正解を出力

システム A，B の MRR は 1 式に代入することにより，

$$\text{システム } A \text{ の } MRR = \frac{1}{2}(1*1+0*1) = 0.5$$

$$\text{システム } B \text{ の } MRR = \frac{1}{2}(0.5*2) = 0.5$$

となり，ともに 0.5 となる．このときユーザ満足率の観点から，システム A，システム B の優劣を比較してみたい．

ユーザ満足率が正解順位により一意に定まると仮定すると，各質問に対するシステムの平均的なユーザ満足率（以下，平均ユーザ満足率）は 1 式にならって，以下の式で求められる．

$$\text{平均ユーザ満足率} = \frac{1}{C}\sum_{i=1}^{5} Si_i C_i \qquad (2)$$

S_i は正解順位 i におけるユーザ満足率である．ここで RR とユーザ満足率の関係として以下の 3 つのケースを考えてみる．

（1） RR とユーザ満足率が常に一致する

（2） RR とユーザ満足率の曲線の傾向が似ている

（3） RR とユーザ満足率が曲線の傾向が似ていない

ケース 1 のシステム A，B の平均ユーザ満足率は MRR と等しくなり，ともに 0.5 となる．次にケース 2 として RR と傾向の似ているデスクトップ型の「満足」のユーザ満足率を適用した場合，表 1 よりシステム A，B の平均ユーザ満足率は以下のようになる．

システム A の平均ユーザ満足率（デスクトップ型「満足」）

$$= \frac{1}{2}(0.85 * 1 + 1 * 0) = 0.43$$

システム B の平均ユーザ満足率（デスクトップ型「満足」）

$$= \frac{1}{2}(0.40 * 2) = 0.40$$

このように，絶対値としては MRR より低い値になっているものの，システム A とシステム B の平均ユーザ満足率はほぼ等しい値となっている．またケース 3 として，RR ではなく，正解順位に対してユーザ満足率がほぼ比例するモバイル型の「満足」のユーザ満足率を適用した場合の，システム A，B の平均ユーザ満足率は，同様に表 1 より，

システム A の平均ユーザ満足率（モバイル型「満足」）

$$= \frac{1}{2}(0.89 * 1 + 0 * 1) = 0.45$$

システム B の平均ユーザ満足率（モバイル型「満足」）

$$= \frac{1}{2}(2 * 0.62) = 0.62$$

となり，MRR が等しくてもシステムの平均ユーザ満足率は一致しないことが分かる．上記の例のように MRR の値から平均ユーザ満足率を一意に定めることは出来ない．したがって，平均ユーザ満足率の観点からシステム精度の目標設定を行うには，MRR という単一の値ではなく正解順位の分布を定めることが望ましい．

3.3 ASKMi の平均ユーザ満足率の見積り

本節では我々が研究開発している質問応答システム ASKMi [3] [4] の平均ユーザ満足率を見積もってみる．NTCIR-4 参加後も精度向上を継続して行っており，質問解析と回答選択を改良することにより，MRR では 0.16 向上している（0.45 から 0.61）．この NTCIR-4 の質問セットにおけるワークショップ参加時（以下「提出時」）の精度，およびその後の精度向上結果（以下「改良後」）がどの程度の平均ユーザ満足率で，精度向上が平均ユーザ満足率にどの程度貢献しているかを 3.1 の正解順位-ユーザ満足率の対応結果と 3.2 の 2 式から評価する．ここではアンケート質問セットの正解順位とユーザ満足率の対応結果を用いているが，アンケートで利用した仮想出力と ASKMi の出力とは，少なくとも以下の点が大きく異なる．

- 根拠文書

アンケートの根拠文書は人手で作成しているのに対し，実際のシステムでは自動で根拠文書を作成するため，後者の方がユーザにとって分かりづらい可能性がある．

	MRR	1	2	3	4	5	5 位以下
提出時	0.454	75	17	7	6	6	84
改良後	0.613	102	23	9	6	8	47

表 2 ASKMi の精度

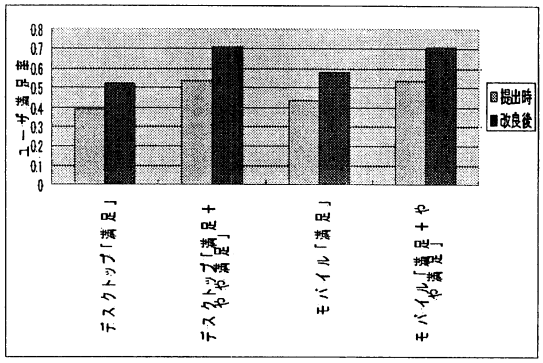

図 7 ASKMi の満足率の見積り

- 不正解の間違い方

アンケートの不正解の回答には固有表現の部分文字列など，単独で意味を成さない文字列は出力していないが，実際のシステムにはこのような間違いが存在する．

今回の分析は，平均ユーザ満足率の目安を知ることが目的のため，これらの相違点は無視して換算を行った．

各システムの NTCIR-4 QAC2 Subtask1 の質問セットに対する精度は表 2 のとおりである．ここで MRR は各システムの MRR の値，1 − 5 位以下は，システムが該当順位に正解を出力した質問の数である．表 2 の各正解順位の質問数と，表 1 のユーザ満足率を 2 式に代入して求めた，平均ユーザ満足率を図 7 に示す．図 7 から以下のことが言える．

- 改良後の平均ユーザ満足率

「満足」ではデスクトップ型が 0.525，モバイル型が 0.582 と，ともに 0.6%未満であった．「満足＋やや満足では」デスクトップ型，モバイル型ともにほぼ 0.7 であった．

- デスクトップ型とモバイル型の比較

「満足」のユーザ満足率グラフでは，正解順位 2 位および 3 位におけるユーザ満足率がデスクトップ型とモバイル型で大きな差があったが，ASKMi では 1 位および 5 位以下の正解が多かったため平均ユーザ満足率の差はそれ程大きくならなかった（提出時で 0.043，改良後で 0.057）．「満足＋やや満足」のユーザ満足率はデスクトップ型とモバイル型で似た傾向を示していたため，平均ユーザ満足率もほぼおなじ値であった．

- 提出時と改良後の比較

「満足」ではデスクトップ型で 0.135，モバイル型で 0.149 改善している．「満足＋やや満足」ではデスクトップ型，モバイル型ともに 0.179 改善し，デスクトップ型，モバイル型ともこのとき不満の人が 10 人中，5 人から 3 人に減ったことを示し

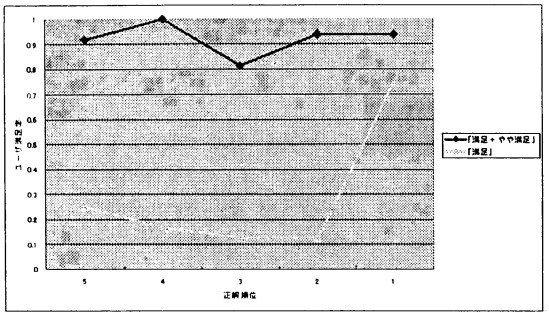

図 8　デスクトップでは満足の度合いには順位によらないというユーザの正解順位-ユーザ満足率グラフ

ており，精度向上が平均ユーザ満足率に貢献していることが分かった．

3.4　アンケートのコメント

本節ではアンケートの自由記述のコメントを列挙し，知見を述べる．

3.4.1　正解順位に関して

デスクトップ型では，「順位は重要ではない」や「5位以内に入っていれば満足である」という意見が多かった．そこで該当ユーザ（4人）のみを対象として，正解順位-ユーザ満足率グラフを作成した．図8に示す．

この結果より正解順位は重要でないとコメントした被験者に関して，「満足＋やや満足」では正解順位とユーザ満足率に相関がないことが分かった．また「満足」では2位から5位は正解順位とは相関がないが，1位に正解を提示することにより満足率が向上することが分かった．

また「モバイル型の方は，順番に上から読んでいくので，早めに正解が出てくると，満足度が高い」というコメントがあり，正解順位にユーザ満足率の比例するモバイル型の「満足」の結果と一致している．

3.4.2　質　　問

- 「今日○○は勝ったか」というような質問もしたい（○○は野球チーム）

アンケートは factoid 質問のみを対象にしたが，Yes/No 型の質問に対するニーズや，新鮮情報を知りたいというニーズが伺える．

- 以前見たことがあるが記憶が曖昧な情報に関する質問は回答を見ただけで判断できるが，知らない情報に関する質問は根拠文書も見ないと，正解/不正解は判断できない

例えば何かを思い出したいときには回答のみを提示し，新しいことを知りたいときには詳細な根拠まで提示するなど，利用する用途やユーザの背景知識に応じて提示する情報を切り替えるような技術が必要になってくると考えられる．

3.4.3　回　　答

- 時間と時刻の区別ができた方がうれしい
- 海外の地名に国内の地名が入っていると不満

具体的には「「閉演時間」を聞いているのに回答に「15分」とあると不満」や，「海外の地名を質問しているのに，国内の地名が回答にあるとがっかり」という意見があった．これは回答タイプの詳細化によりある程度解決できる問題であると考えられる．

- 場所が答えとなる質問では，実際に行きたいのか，単に場所を知りたいのかで，ほしい情報の粒度が変わる

これを解決するにはシステムを利用するユーザのコンテキストを用いた情報提示が課題となる．

- トンチンカンな候補が混ざっていると，あまり賢くないと感じる
- 紛らわしい答えが多くなると不満が大きくなる

この二つの問題を同時に解決するのは非常に難しい．どのようなときに前者のように感じ，どのようなときに後者のように感じるのかを，今後詳細に調査する必要があるが，アンケートでは以下のような例があり，回答タイプもひとつの要因と考えられる．

－　人名のなかに人名ではない回答があると満足率が低かった

質問「青色ダイオード発明をめぐって会社を訴えたのは誰だっけ」に対して，1位に正解を提示し，2位に明らかに人名でない回答候補を提示したところ，正解が1位に出力された他の質問に比べ，ユーザ満足率が低かった．

－　単位の等しい数値が列挙されていると満足率が低かった

質問「○年△月の××首相の支持率」に対して，正解を5位に含む「XX%」という形式の回答候補を5件提示したところ，正解が5位に出力された他の質問に比べ，ユーザ満足率が低かった．

- 回答が複数ありそうなリスト型の質問には複数の正解があるとうれしい

「神戸市内で神戸牛の食べられるお店は？」等の複数の回答が期待される質問に関しても，今回のアンケートでは正解は1つに限定していた．今後正解の個数の同定や，提示方法などの課題を解決していく必要がある．

- 不正解の回答候補をいかに気づきやすくするかが使いやすさのポイント

例えば回答タイプ毎に色分けすることで，人名を聞いた質問に地名を答えている場合にすぐに間違いと気づかせるなど，質問応答における回答の成否判定の容易さの工夫に関して検討する必要がある．

3.4.4　根拠文書

- 公式サイトなどに載っているような様式の整った文章の方が信用できる
- blog や個人サイトのような様式の根拠文書だと信頼性を疑う

非常に多い意見であった．現在は，ヒットワードとの距離をベースに算出した，回答候補文字列のスコアが最も高い記事を根拠文書にしているが，根拠文書自体の信頼度や Authority を

利用した根拠文書の選出が課題として挙げられる.

- 結局中身を読んで正誤を判断することになるので，文書検索と比較したときの有効性が分からなかった

通常の文書検索で得られる文書のリストと，質問応答の各回答に付随する根拠文書のリストは異なるので，今後これらのリストの有用性の比較を行いたい．もし後者のリストの方が有用であれば，たとえユーザが根拠文書を読むことになっても質問応答には意義があることになる．もし文書リストとしての有用性に差がないのならば，正答に到達するまでのコストを比較し，文書検索との差を明らかにする必要がある.

- 根拠文書自体を読むのが面倒
- 回答文字列は明らかな誤りでも，どこかに正解が書いてある気がして，結局根拠文書まで読んでしまう

現状ではすべての利用シーンで回答だけを提示してユーザを説得するのは困難であり，根拠文書のオンデマンド提示などのインタフェースの上の工夫が課題となる.

3.4.5 利用シーン

デスクトップ型，モバイル型に対して以下のようなコメントがあった.

- デスクトップ型の場合は見出しになっている回答は読まずに，いきなり根拠文書中のハイライト部分を読み始める

この被験者においては回答文字列を別出しで提示する機能は効果がなくなっており，根拠文書中の回答文字列をハイライトする機能が従来の文書検索に比べ追加されていることになる．根拠文書中の回答文字列のハイライトがユーザ満足率にどれくらいの効果があるか今後分析してみたい.

- 全体的に，モバイル型の方が答えを探しやすかった
- モバイル型の方が提示がコンパクトにまとまっていた

質問応答システムがずばり回答を提示するものであるとすれば，デスクトップ型の一括提示よりもモバイル型の提示方法の方がユーザのイメージに近い可能性がある.

3.5 被験者が正解選択を間違えたケース

アンケートには被験者が正解の選択を間違えた例がいくつかある．実際に質問応答システムを研究開発していく上での参考になると考えられるので，典型的な間違え方を列挙しておく.

- 根拠文書の不適切な抜粋による選択誤り

例えば正解が「製品 A」である質問「ノンカフェインのお茶は?」に対し，誤って不正解「製品 B」を選択した被験者が複数いた．これは「製品 B」の根拠文書が,

```
<--
  製品 C の説明
  製品 B
  製品 B の説明
-->
```

という形の切り出し方が不適切なものであり，かつ,「製品 C の説明」中に「ノンカフェイン」という語が出現していたためであった（なお「製品 C」はお茶ではないため正解ではない）.

- 別の地名に関する情報は間違えやすい

「神戸市内で神戸牛の食べられるお店は?」というように場所を指定した質問に対して，根拠文書に「東京都」とあるにもかかわらず東京にある店の名前を正解として選択した.

- 質問作成者の意図とアンケート被験者の解釈のずれ

実際にアンケートを取ってみると質問作成者の意図と被験者の解釈がずれてしまうことがあった．ずれの要因は以下の 2 種類に大別できた.

– 単語の解釈のずれ

例えば質問 「アカデミー賞を最も多く受賞している俳優は誰?」に対して正解が「人名 A」であったが，根拠文書では「俳優」ではなく「女優」という言葉で説明されていたため，半数以上の人が「ない」「分からない」を選択した.

– 文としての解釈のずれ

例えば質問 「紅葉の名所」に対して，正解は「場所 A」であったが,「ツアー名 B」という紅葉の名所めぐりのツアーの名前を選択する被験者が多かった.

4. おわりに

質問応答システムの回答精度とユーザ満足率の関係を明らかにするために，回答候補を一括提示するデスクトップ型と，回答候補をひとつずつ提示するモバイル型の 2 種類の利用シーンを想定し，シーン別に正解順位に対するユーザの満足の度合いを評価するアンケートを行った．この結果，デスクトップ型におけるユーザ満足率は質問応答の一般的な評価尺度である逆数順位と似た曲線となるが，回答候補をひとつずつ提示するモバイル型におけるそれは順位とほぼ比例することが分かった．このため，ユーザ満足率の観点からシステムの目標精度を設定するには，平均逆数順位ではなく正解順位の分布を決定することが望ましいことが分かった.

また得られたグラフから開発システムの平均ユーザ満足率を見積もったところ「満足＋やや満足」では 70%程度,「満足」では 60%未満であり，現状の精度では十分なものではないと考えられる.

さらに今回のアンケート結果から，ある質問に対するユーザが真に満足するために必要な要求精度と，不満を少なくするために必要な要求精度には大きな隔たりがあることが分かった．正解順位や回答の提示方法がこの原因の主要因となっている可能性は高いが，その他の要因に何があるのかは明確になっておらず，また各要因がどれくらい影響しているかも明確にはなっていない．今後，上記要因の洗い出し，各要因の影響度の明確化を行い，その結果をもとに真に満足できる質問応答システムの実現へ向けた改良に取り組みたい.

文　献

[1] TREC: http://trec.nist.gov

[2] NTCIR4 QAC2 Subtask1: http://www.nlp.is.ritsumei.ac.jp/qac/qac2/index-j.html

[3] Sakai, T. *et al.*: ASKMi: A Japanese Question Answering System based on Semantic Role Analysis, Proceedings of RIAO 2004, pp. 215-231. 2004

[4] Sakai, T. *et al.*: Toshiba ASKMi at NTCIR-4 QAC2, Proceedings of NTCIR-4, 2004

[5] 浦田 他: 質問応答技術に基づくマルチモーダルヘルプ, 情報処理学会研究報告, FI-74-4, 2004