人工知能学会研究会資料
SIG-SLUD-B402-07

対話を通じた情報獲得のための期待効用に基づく質問選択

Question Selection based on Expected Utility for Information Acquisition through Dialogue

大塚 嗣巳 [1*] 駒谷 和範 [2] 佐藤 理史 [1] 中野 幹生 [3]
Tsugumi Otsuka[1] Kazunori Komatani[2] Satoshi Sato[1] Mikio Nakano[3]

[1] 名古屋大学大学院工学研究科
[1] Graduate School of Engineering Nagoya University
[2] 大阪大学産業科学研究所
[2] The Institute of Scientific and Industrial Research, Osaka University
[3] ホンダ・リサーチ・インスティチュート・ジャパン
[3] Honda Research Institute Japan Co., Ltd.

Abstract: We are developing a framework where a dialogue system asks questions for users and thus acquires information (cuisine of unknown restaurants) through dialogue. Such a question would be better if it is more concise and concrete. We propose a method to select an appropriate question on the basis of expected utility calculated for four question types: Yes/No, binary, ternary, and Wh- questions. We first define utility for the four types by considering their appropriateness for users. We then define a probability that represents how likely a question contains a correct cuisine by using confidence measure (CM) of cuisine estimation. Then the sum totals of products of the utility and the probability are calculated as expected utility. As experiments, we compare several ways to integrate two basic CMs we proposed before. We also discuss adequacy of the utility values by using questionnaire results for 15 subjects.

1 はじめに

対話システムは，システム自身の知識にない単語や表現をうまく扱える必要がある．これに対し，テキスト入力の結果を正しく形態素解析して固有表現を抽出できたり，超大語彙の音声認識器 [1] により正しい音声認識結果が得られたとしても，十分に対処できない場合がある．例えば，音声認識や固有表現抽出が出来ても，その内容がシステムの知識 (e.g. 検索対象データベース) にない場合，ユーザ要求を満たす応答は行えない．図 1 上部に示すように，レストラン検索タスクで，検索対象データベースにない「牡丹亭」をユーザが発話した場合には，システムは「知りません」としか応答できない．

この問題を回避するには，現状では，開発者がオフラインで検索対象データベースを拡張するしかない [2][3][4]．この作業にはコストがかかる．例えば，レストラン検索用のデータベースは，飲食店が日々新たに開店するため，その都度，更新する必要がある．

*連絡先：〒 464-8603 愛知県名古屋市千種区不老町 C3-1(631) 名古屋大学大学院工学研究科電子情報システム専攻 佐藤理史研究室
E-mail: t_ootuka@nuee.nagoya-u.ac.jp

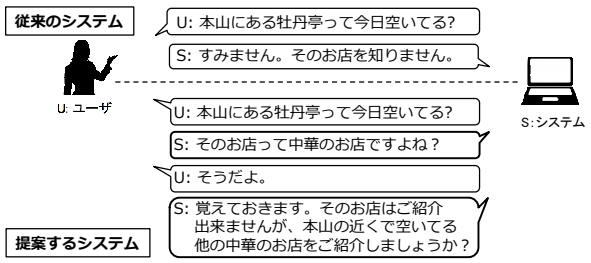

図 1: 従来の/提案システムにおける対話例

この問題の緩和を目標として，対話中にユーザに質問することで，情報を獲得する手法を提案している [5]．現在は一例として，対象タスクをレストラン検索とし，獲得対象をデータベース (以下 DB) 中にない店舗のジャンル，としている．ジャンルの獲得により，図 1 下部に示すように，入力と類似した店舗の推薦も可能となる．

本稿では，期待効用に基づき，適切な質問を選択する手法について述べる．適切な質問とは，ユーザが賢いと感じる度合を考慮して，質問内容が正しく，かつ具体的な質問とする．このような質問を選択するために，まず四種類の質問形式を用意し，その内容の正誤それぞれに，ユーザが賢いと感じる度合に対応する効

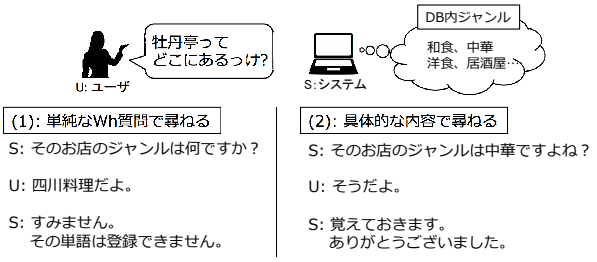

図 2: 店名：牡丹亭/ジャンル：中華である時の対話例

用を与える．次に，各質問に含まれる内容が正しい確率を，ジャンル推定結果の確信度を用いて得る．これら 2 つを乗じて総和をとることで期待効用を計算し，この値が最大となる質問形式を選択する．

2 情報獲得のための適切な質問

情報獲得のための質問は，内容が煩雑でなく，可能な限り具体的であることと，誤った内容を尋ねないこと双方を満たすことが望ましい．具体的，つまり選択肢を限定するほど，システムが受理できるユーザ回答が得られやすく，ユーザがシステム発話を聞く負担も軽減される．しかし，選択肢を限定しすぎれば，誤った内容を尋ねる可能性も高まる．

ここで，情報獲得のために不適切な質問の例を挙げる．例えば，和食や居酒屋など，候補となる全てのジャンルを用いた択一質問を生成すれば，内容を誤ることなく，かつ，ジャンル獲得は可能である．しかし，選択肢が多くなるほど，ユーザにとって煩雑極まりないものとなる．人間の短期的記憶は多くの内容を覚えられないという有名な心理学実験 [6] からも，多くの選択肢を提示するのは避けるべきであると言える．また,「そのお店のジャンルを教えて下さい.」と単純な Wh 質問を生成すれば，内容を誤ることなく煩雑でない質問が可能となる．しかし，ユーザがシステムの受理出来ない表現を用いてしまい，ジャンルを獲得できない可能性が高まる (図 2(1))．実際に，我々が用いる DB には「中華」というジャンルは存在するが,「四川料理」というジャンルは存在しない．

具体的かつ正しい内容の質問が生成できれば，情報獲得が可能である．例えば，図 2(2) の「そのお店は中華ですよね?」と特定のジャンルを提示した Yes/No 質問をすれば，ユーザ回答は「はい」または「いいえ」などの肯定表現か否定表現に限定できる．また,「そのお店は中華か和食のどちらですか?」という択一質問をすれば，ユーザ回答を「中華です」または「和食です」など，システムが受理できる内容を含む質問に誘導しやすい．これにより，システムの受理出来る表現をユーザが用いる可能性が Wh 質問よりも高まり，情報獲得をより確実に行える．また，ユーザに煩雑さを感じさせることもない．

表 1: 4 種類の質問形式

質問形式	提示ジャンル数	質問例
Yes/No 質問	1	中華ですよね?
二択質問	2	中華と和食のどちらですか?
三択質問	3	中華、和食、居酒屋のどれですか?
Wh 質問	—	そのお店のジャンルは何ですか?

質問内容は，可能な限り選択肢を限定した具体的なものが望ましい．例えば,「そのお店は中華，和食，居酒屋のどれですか？」という三択の質問よりも「そのお店は中華ですよね？」と Yes/No 質問が出来た方が良い．これは，ユーザは冗長な質問を聞かされるのが煩雑だと感じることが想定されるためである．

ただし，提示ジャンルを限定するほど，内容に正解を含まない質問を行うリスクも高まる．例えば，提示ジャンルを三つまで増やせば正しい内容で質問出来たが，提示ジャンルを一つに限定してしまった結果，誤った内容を聞いてしまうという状況も考えられる．したがって，提示ジャンルを限定し過ぎたために誤った内容で質問してしまうよりは，ジャンルの選択肢を増やす，Wh 質問をするなどの選択を取った方が良い場合もある．

このように，質問が煩雑でなく，可能な限り具体的であることと，誤った内容を尋ねないこととの間には，トレードオフの関係がある．これを上手く考慮した質問生成を目指す．本研究では，表 1 に示すように，Yes/No 質問，二択質問，三択質問，Wh 質問の四種類の質問形式から選択する．四択質問以上が存在しないのは，四択以上の候補を聞かされるのは，ユーザにとって煩雑であるだろうと考えたためである．各々のジャンルの確信度を考慮して，Yes/No 質問では一つのジャンルを，二択質問では二つ，三択質問では三つのジャンルを提示する．Wh 質問ではジャンルを提示しない．

3 期待効用に基づく質問選択

2 章で述べたような適切な質問を選択する方法について説明する．本節ではまず 3.1 節で，2 章で述べた四種類の各質問形式が，ユーザにとってどの程度賢い (煩わしくない) かを表す効用について述べる．次に 3.2 節で，推定結果がどの程度確からしいかを表す確信度について述べる．3.3 節では，この確信度を質問内容が正解である確率として適用することで，四種類の質問形式から得られる効用の期待値，つまり期待効用を計算する．これが最大である質問形式を選ぶことで，最適な質問選択を実現する．

表 2: 各質問形式に対する効用

質問形式	正解時の効用	誤り時の効用
Yes/No 質問	U_1^+	U_1^-
二択質問	U_2^+	U_2^-
三択質問	U_3^+	U_3^-
Wh 質問	U_{Wh}^+	U_{Wh}^-

3.1 質問形式ごとの効用

システムが行う質問が，ユーザにとって煩わしくないという度合を，効用として表す．効用はゲーム理論において参加者の利益を表す値である．

効用は各質問形式毎，その正誤毎に設定する．正誤とは，選択した質問内に正解ジャンルが含まれたか否かで定める．例えば，二択質問で提示した二つのジャンルの中に，当該店舗に関する正解ジャンルが含まれれば，その質問は正解とする．効用は，正解の場合は正の値，誤った場合は負の値とする．表 2 の，$U_x^{\{+,-\}}(x \in \{1,2,3,\mathrm{Wh}\})$ の U は効用 (Utility) の頭文字である．x は表 1 の提示ジャンル数を表し，$+,-$ は正解/誤りを表す．例えば，U_2^- は二択質問で誤った場合の効用を表す．

表 2 において，正解時の効用は $U_1^+ \geq U_2^+ \geq U_3^+ \geq 0$，誤り時の効用は，$0 \geq U_1^- \geq U_2^- \geq U_3^-$ となるように設定する．これは，正解時はより候補ジャンルが限定されている方が良く，誤り時も候補ジャンルが多いほどユーザにとって煩わしいだろうという直感に基づく．この直感に関する検証は 5 章で行う．また，Wh 質問には正解も誤りもないことや，効用の基準値とすることから，$U_{\mathrm{Wh}}^+ = U_{\mathrm{Wh}}^- = 0$ とする．

3.2 質問内容の正しさを表す確信度

期待効用計算のために，質問内に正解ジャンルが含まれる確率を算出する．この確率を，店舗のジャンル推定がどの程度確からしいかを表す確信度 (Confidence measure: CM) を用いて表現する．確信度は，システムが検索対象とする DB 内のジャンル全てに対して算出され，その総和は 1 となるように正規化される．このため，これはある店舗名が与えられた際の事後確率と見なすことができる．

質問内に正解ジャンルが含まれる確率を，提示したジャンルの確信度の和で表現する．各ジャンルの確信度の中で，i 番目に大きいものを $CM(g_i)$ とし，その時のジャンルを g_i とする．例えば g_1 は，全ジャンルの中で一番確信度が高いジャンルを表し，それが正解である確率は $CM(g_1)$ となる．また，上位二つのジャンルのどちらか一方が正解でありそうな確率は $CM(g_1)+CM(g_2)$ となる．

表 3: 確信度と効用に基づく質問形式毎の期待効用

質問形式	正解時の効用	誤り時の効用	正解を含む確率	期待効用
Yes/No	U_1^+	U_1^-	P_1	$U_1^+P_1 + U_1^-(1-P_1)$
二択	U_2^+	U_2^-	P_2	$U_2^+P_2 + U_2^-(1-P_2)$
三択	U_3^+	U_3^-	P_3	$U_3^+P_3 + U_3^-(1-P_3)$
Wh	U_{Wh}^+	U_{Wh}^-	–	0

$P_x = \sum_{i=1}^{x} CM(g_i), \quad U_{\mathrm{Wh}}^+ = U_{\mathrm{Wh}}^- = 0$

ただし，g_i：$CM(g_i)$ の値が i 番目に大きいジャンル

過去に提案した二つの確信度を統合することで，より質問選択に適した確信度生成を行う．二つの確信度とは，検索対象とする DB 内の文字分布を利用した確信度 CM_D と，Web ページ上のジャンルの出現頻度を利用した確信度 CM_W である [5].

3.3 期待効用の計算

期待効用が最大となる質問形式を，ユーザが最も賢いと感じる最適な質問として選択する．期待効用とは，効用の期待値であり，効用と，質問内に正解が含まれる確率から算出する．これにより，正解/誤り双方を加味した上で，質問形式毎にどの程度賢いと感じられるかの見込みを数値化する．

例として，表 3 の Yes/No 質問の期待効用を計算する過程を，具体的に説明する．まず，一位のジャンル g_1 が正解となる確率は，$P_1 = CM(g_1)$ である．したがって，一位のジャンル g_1 が正解でない確率は $1-P_1$ である．この確率と効用 U_1^+，U_1^- を用いて，期待効用 $U_1^+P_1 + U_1^-(1-P_1)$ を算出する．また，二択質問時は，提示ジャンル二つのうちどちらかが正解である確率 $P_2 = \sum_{i=1}^{2} CM(g_i)$ と，そうでない確率 $1-P_2$，効用 U_2^+，U_2^- から，期待効用 $U_2^+P_2 + U_2^-(1-P_2)$ を算出する．表 3 で算出する期待効用 $U_x^+P_x + U_x^-(1-P_x)$ が最大となる $x \in \{1,2,3,\mathrm{Wh}\}$ をから質問形式を選択する (式 (1))．なお，Wh 質問の効用が 0 であることから，P_{Wh} は定義しなくても良い．

$$(\text{質問形式}) = \underset{x \in \{1,2,3,\mathrm{Wh}\}}{\arg\max} \{U_x^+P_x + U_x^-(1-P_x)\} \quad (1)$$

4 効用および確信度に関する議論

3 章で述べた効用と確信度に関してどのようなものが適当であるのかを議論する．実験では，愛知県内の 1,656 件の店舗名とそのジャンルの組が記載されたレス

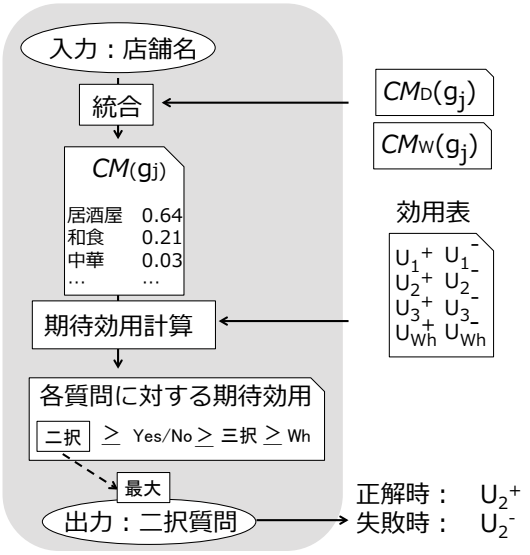

図 3: 正誤とその時の効用を得るための実験フロー

表 4: 検討する効用

効用パターン	U_1^+	U_2^+	U_3^+	U_1^-	U_2^-	U_3^-
(i)	1	1	1	-1	-2	-3
(ii)	1	1/2	1/3	-1	-1	-1
(iii)	2	3/2	4/3	-2	-3	-4
(iv)	2	3/2	4/3	-1	-2	-3
(v)	2	3/2	4/3	-3	-4	-5

トラン DB を用いる．この 1,656 件に対し算出された確信度を用いて評価する．

本節では，DB 内のあるエントリが未知であるとみなしたうえで，それに対する質問を選択し，その効用の値や正誤に基づいて手法を評価する．この過程を図 3 に示す．まず，DB 内のあるエントリの店舗名を入力とし，CM_D と CM_W を算出する．なお，オープンテストを実施するために，CM_D は 1,656 件の店舗名において，10 分割交差検定を行うことにより算出する [5]．

これらを入力として統合した確信度を算出する．統合方法は 4.2 節で検証する．統合した確信度と，効用の値が与えられれば，期待効用最大の質問形式を求められる．効用の値については 4.1 節で検証する．

DB 内の当該店舗のジャンルを参照することにより，その質問が正解を含むかどうかが分かる．その正誤を DB 全体 1,656 件についてカウントする．また，選択された質問形式の効用の総和も求める．

4.1 効用に関する議論

質問内容の正誤数や Wh 質問生成数を基にした二つの観点から，効用の適切さを決める．一つ目は，誤りの数である．誤りの数は少ないことが望ましい．二つ目は，Yes/No 質問の数と Wh 質問の数である．具体的には，Yes/No 質問の数に比べて，Wh 質問の数が多過ぎない効用を適切であるとする．これは，Wh 質問ばかり行うのは，確信度をほぼ利用しない場合に相当するからである．

表 5: 5 種類の効用の比較結果

		(i)	(ii)	(iii)	(iv)	(v)
	Yes/No	**381**	**597**	**520**	**756**	**393**
正解	二択	221	229	224	251	166
	三択	416	160	258	198	263
	Yes/No	42	76	61	181	39
誤り	二択	24	40	28	43	17
	三択	55	32	46	43	33
	正解	1018	986	1002	1205	822
総数	**誤り**	**121**	**148**	**135**	**267**	**89**
	Wh	**517**	**522**	**519**	**184**	**745**

適切な効用について議論するため，5 種類の効用パターン (i)〜(v) を用意した (表 4)．効用パターン (i) は，ジャンルを多く提示して誤るほどユーザが煩わしいと感じることを表現した効用である．具体的には，正解時の効用は等しい値とし，誤り時に効用に差をつけている．このパターン (i) は，一つのジャンルあたり 1 点を賭ける場合に相当する．つまり，Yes/No 質問では 1 点，二択質問では，2 点，三択質問では 3 点をかけ，正解ジャンルに賭けていれば賭けた分の 1 点を効用として獲得し，誤れば賭けた分全ての効用を失うことを表現している．

効用パターン (ii) は，正解する時に，提示ジャンルを限定出来ているほどユーザが賢いと感じることを表現した効用である．具体的には，誤り時の効用は等しくし，正解時の効用に差をつけたものである．このパターン (ii) は，持ち点 1 点を提示ジャンル数で分配することに等しい．つまり，Yes/No 質問には持ち点全てを，二択質問では二分割，三択質問では三分割した点を賭け，正解すれば賭けた分の効用が貰え，誤れば 1 点の効用全てを失うことを表現している．

効用パターン (iii) は，正解時と誤り時のユーザの感じ方双方を表現した効用である．具体的にはパターン (i) と (ii) の各効用を足し合わせた場合に相当する．また，効用パターン (iv) は，(iii) の誤り時の効用を減少させたもの，(v) は増加させたものである．

5 種類の効用に関し，それぞれで生成した質問での正誤，Wh 質問の数を比較した結果を表 5 に示す．表 5 は三段構成となっている．一段目が Yes/No，二択，三択質問で正解した件数の内訳，二段目がそれらで誤った件数の内訳，三段目が，正解，誤り，Wh 質問の総数を示している．ここでは簡単のため，確信度を $CM_D(g)$ と $CM_W(g)$ の単純な加算 (後述する $CM1(g)$) として期待効用を計算した．

表 5 と以下に示す二点の理由から，以降では，パターン (iii) を適切な効用として用いる．まず，適切な効用

かを判断する二つの観点を双方共に満たすのはパターン (ii) と (iii) であった．パターン (iv) に関しては誤りの総数から，パターン (i)，(v) に関しては Yes/No 質問と Wh 質問のバランスから (ii)，(iii) の方が良いとし，この二つのうち誤りの数が少ない (iii) を採用した．

表 4 の値も踏まえた上で，表 5 の結果に関して定性的に議論する．(i)〜(iii) では Wh 質問をしない場合，つまり，Yes/No，二択，三択質問の数のみ変化した．これは，効用の正誤の値の差が小さいことにより，それが期待効用の値の大小にあまり影響しなかったためである．つまり，確信度の値のみから期待効用の大小関係 (正か負か) が定まるため，効用の値を変化させても Wh 質問の数はほぼ変化しなかった．これに対し，(iv)，(v) では，効用の値，例えば，U_1^+ と U_1^- の差が大きくなり，Wh 質問を含めた各質問の数が変化した．

(ii)，(iii)，(iv) では，(i) に比べ，Yes/No 質問が選択されやすかった．これは，(i) では正解時の効用に変化がないのに対し，(ii)，(iii)，(iv) では Yes/No 質問の効用が一番高かったためである．また，(iii) では，(ii)，(iv) に比べ Yes/No 質問が生成されすぎず，二択，三択が選択されやすかった．例えば，(iii) は (ii) に比べ三択質問を選択した数が 100 件ほど多い．これは，(iii) の正解時の効用の比が相対的に小さいためである．さらに，(iv) は Yes/No 質問の数が多い分誤りが多く，逆に (v) は Yes/No 質問が少ない分 Wh 質問の数が多いこれは，(iv)，(v) の正誤の効用のバランスが悪かったことが原因である．

4.2 確信度の統合方法に関する議論

表 4 の効用パターン (iii) を用いて，適切な確信度の統合方法を議論する．具体的には，1,656 件分の推定により得られた効用の総和を，統合方法ごとに比較する．質問内容の正誤に対応する効用が，その推定結果における質問の賢さを表す．したがって，効用の 1,656 件分の総和が大きいほど，質問の正誤を踏まえたうえで，賢い質問が選択されていたことを表す．以下，まずは比較する確信度の具体的な説明をした後に，比較結果を示す．

4.2.1 比較する確信度

本稿では，統合元となる 2 種類の確信度 CM_D，CM_W と，3 種類の統合した確信度，計 5 種類を比較する．なお，統合とは，$CM_D(g)$ と $CM_W(g)$ の値を統合するものである．3 種類の確信度統合法を，式 (2)〜(4) に示す．

$$CM1(g) = \frac{1}{Z_1}(CM_D(g) + CM_W(g)) \tag{2}$$

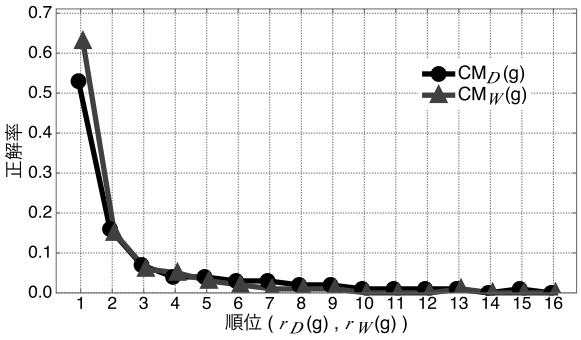

図 4: 順位毎の正解率

$$CM2(g) = \frac{1}{Z_2}\{ \sum_{k\in\{D,W\}} (1 + \frac{1}{r_k(g)})CM_k(g)\} \tag{3}$$

$$CM3(g) = \frac{1}{Z_3}\{ \sum_{k\in\{D,W\}} (1 + \frac{Acc_k(g)}{r_k(g)})CM_k(g)\} \tag{4}$$

(ただし，Z_1〜Z_3 は総和を 1 にするための g に関する正規化係数)

$CM1(g)$ は，$CM_D(g)$ と $CM_W(g)$ を足し合わせる統合法である．これにより，両確信度を考慮した確信度の生成が期待できる．

$CM2(g)$ は，$CM1(g)$ に順位の情報 $r_k(g)$ の重みをボーナスとして反映させたものである．$r_k(g)$ により，推定結果上位のジャンルほど大きなボーナスが得られるようにする．これは，降順に並べた確信度において，上位ほど正解ジャンルが存在する傾向があることを考慮している．図 4 は，1,656 件のレストランに関して推定を行った時の $CM_D(g)$，$CM_W(g)$ の順位別での正解率を示したものである．

$CM3(g)$ は，$CM1(g)$ に $r_k(g)$ だけでなく，ジャンル毎の正解率 $Acc_k(g)$ の重みもボーナスとして反映させたものである．$Acc_k(g)$ により，正解しやすいジャンルほど大きなボーナスが得られるようにする．これは，$CM_D(g)$，$CM_W(g)$ それぞれで正解しやすいジャンルが異なることを考慮している．図 5 は，1,656 件のレストランに関して推定を行った時に，ジャンル g が最大の確信度を持つ場合のジャンル毎の正解率を示したものである．図 5 から，ジャンルごとや確信度 $CM_D(g)$，$CM_W(g)$ ごとに，正解率が異なる傾向が分かる．特定のジャンルを示す文字列，例えばイタリアンのお店であることを示す「オステリア」や「リストランテ」という文字列を含む店名の場合，CM_D によって正しくイタリアンだと推定できる．CM_W の場合，CM_D で上手く推定できない店舗でも，Web 上の情報を利用して正しく推定できることもある．このような情報源の違いによるジャンル正解率の傾向を考慮したボーナスにより，より適切な確信度生成を目指す．

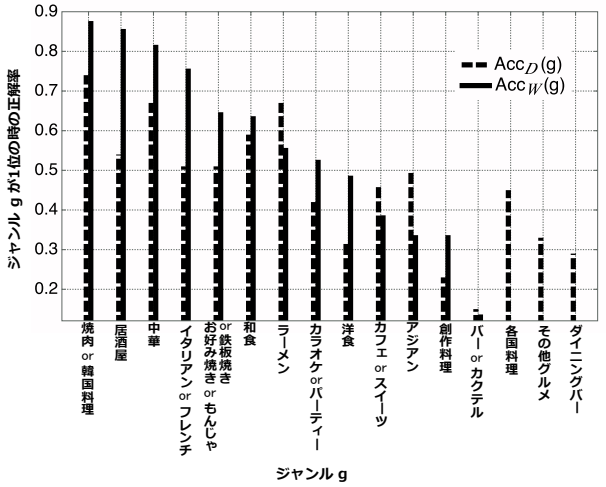

図 5: ジャンル g が 1 位の時の正解率

表 6: 正誤の数および効用の総和の比較

		CM_D	CM_W	$CM1$	$CM2$	$CM3$
正解	Yes/No	556	659	522	660	**650**
	二択	155	149	226	335	**276**
	三択	186	238	254	277	**267**
誤り	Yes/No	281	156	61	107	**90**
	二択	50	23	28	76	**45**
	三択	61	51	46	66	**55**
総数	正解	897	1046	1002	1292	**1193**
	誤り	392	230	135	249	**190**
	Wh	367	380	519	146	**273**
効用の総和		636	1273	1331	1515	**1534**

4.2.2 効用の総和に基づく確信度の比較

5 種類の確信度での効用総和の比較結果を表 6 に示す．表 6 は，三種類のデータを比較している．質問形式ごとの正解/誤りの内訳，正解/誤り/Wh 質問の総数，効用の総和である．

表 6 より，$CM3(g)$ の効用の総和が最大となった．順位の情報 $r_k(g)$ とジャンルごとの正解率 $Acc_k(g)$ 双方を考慮することで，より適切な確信度が得られたことを示している．

正解/誤り/Wh 質問の数から，$CM3(g)$ の適切さに関して議論する．$CM3(g)$ は，$CM_D(g)$，$CM_W(g)$ と比較して，正解数は増加し，誤り数は減少した点で適切な統合方法であったと言える．$CM1(g)$ と比較すると，誤り数が増加するものの，正解数も大きく増加したため，$CM3(g)$ の効用の総和が大きくなった．具体的には，Yes/No 質問での正解数が，$CM3(g)$ では $CM1(g)$ よりも 128 件多いことが総和の増加に効いている．また，$CM2(g)$ と比較しても $CM3(g)$ の効用の総和が大きくなった．これは，$CM2(g)$ よりも慎重になり，誤っていそうな確信度に対して Wh 質問を選択する傾向があったと言える．具体的には，「バー or カクテル」などの正解率の低いジャンルが一位になった場合に $r_k(g)$ では大きな値を与えようとするが，$CM3(g)$ では，それを回避できた．

表 7: 統合前後での正解，誤り，Wh 質問の内訳

		CM_D，CM_W での結果			
		両方共で正解	一方のみ正解	良くて Wh 質問	共に誤り
$CM3(g)$ の結果	正解	**620**	**535**	**38**	0
	Wh	6	115	151	1
	誤り	3	35	61	91
太字の割合		98.6%	78.1%	13.6%	-

4.2.3 統合前後の結果の比較による検証

統合前後での正誤の変化から，$CM3(g)$ が適切な確信度であることを確認する．表 7 は，各店舗に関し，統合前の確信度と $CM3(g)$ において，正解，誤り，Wh 質問のどれであったか分類したものを 1,656 件分集計したものである．具体的には，CM_D と CM_W で共に正解したもの，どちらか一方で正解したもの，双方で Wh 質問を生成するか誤っていたもの，双方共に誤りであったものである．

表 7 から，$CM3(g)$ により双方の確信度が上手く考慮出来ていることが分かる．まず，双方共に正しく推定出来た 629 件のうち，Wh 質問を選択してしまうのは 6 件，誤ってしまうのは 3 件のみに留まった．次に，どちらか一方のみでしか正しく推定出来ない 685 件のうち，535 件 (78.1%) が $CM3(g)$ により正しく推定出来た．一方，CM_D，CM_W 双方ともに Wh 質問をするか誤った質問を生成していた場合，つまり双方ともに適切な確信度付与が出来ない 249 件のうち，38 件 (13.6%) に対して新たに正しい内容の質問を選択することができるようになった．

5 被験者実験による効用の評価

定めた効用の値の妥当性を検証するために被験者実験を行った．定めた効用の値とユーザが実際に受ける印象それぞれの大小関係を比較し，効用の妥当性を検証する．

実験には，前章までに説明した，期待効用に基づく質問選択法を実装したシステムを用いた．被験者は 15 名であり，そのうち 9 名が研究室内の学生，6 名が外部の被験者である．前者は本研究の内容について知っていた可能性が高いため，別に集計した．

実験では，ある店舗名とそのジャンルを提示し，それに関する二種類の状況それぞれで四種類の質問を合成音声により聴取してもらい，各質問への印象の得点付けをしてもらった．二種類の状況とは，正解時の印

検索店舗:東寿司　ジャンル:和食

正解時とWh質問との印象の比較に用いる四種類の質問

・そのお店のジャンルは和食ですよね？
・そのお店のジャンルは和食か中華のどちらですか？
・そのお店のジャンルは和食、中華、居酒屋のどれですか？
・そのお店のジャンルは何ですか？(Wh質問)

誤り時とWh質問との印象の比較に用いる四種類の質問

・そのお店のジャンルは中華ですよね？
・そのお店のジャンルは中華か居酒屋のどちらですか？
・そのお店のジャンルは中華、居酒屋、洋食のどれですか？
・そのお店のジャンルは何ですか？(Wh質問)

質問に対する印象の採点方式

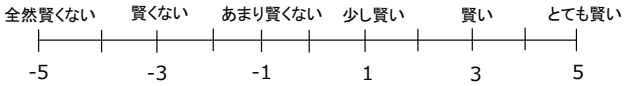

図 6: 賢さの印象を比較する質問とその採点方式

表 8: 15 名に対する質問への印象調査結果

		正解時				誤り時			
	被験者	Yes/No	二択	三択	Wh	Yes/No	二択	三択	Wh
研究室	A	5	4	3	-5	-4	-4	-4	-5
	B	5	4	3	-3	-5	-5	-5	-1
	C	3	1	-1	-5	-1	-3	-5	0
	D	2	1	1	-2	-3	-3	-4	-1
	E	5	1	0	-3	-5	-5	-5	-3
	F	4	1	0	0	0	0	0	1
	G	2	-1	-2	0	2	-1	-2	0
	H	5	3	1	-5	-1	-1	-3	-5
	I	3	1	0	0	-5	-5	-5	0
外部	a	5	3	2	0	-5	-5	-5	-5
	b	3	1	0	-1	-5	-4	-3	-1
	c	3	1	0	1	-2	-2	-3	-3
	d	5	4	2	3	-1	-3	-5	-5
	e	5	3	1	-5	-1	-3	-5	-1
	f	5	3	1	0	-2	-3	-4	0
平均	研究室	3.8	1.7	0.6	-2.6	-2.4	-3.0	-3.7	-1.6
	外部	4.3	2.5	1.0	-0.3	-2.7	-3.3	-4.2	-2.5
	全体	4.0	2.0	0.7	-1.7	-2.5	-3.1	-3.9	-1.9

象を比較するものと，誤り時の印象を比較するものである．図 6 に,「東寿司」という「和食」のお店を検索した時に聴取してもらう質問の例を示す．具体的には，正解時には，Yes/No 質問，二択質問，三択質問，Wh 質問の四種類の場合を相対的に評価してもらった．同様に誤り時には，Yes/No 質問，二択質問，三択質問，Wh 質問の四種類の場合を相対的に評価してもらった．被験者に聴取してもらう四種類の質問の順序は随時変化させた．印象調査では「質問に対し，どの程度賢いと感じたか」という教示の下，図 6 下部に示す，−5 点〜5 点でのリッカートスケールによる採点をしてもらった．

15 名の被験者に対する質問への印象調査の結果を表 8〜10 に示す．表 8 は，正解/誤り時の各質問形式に被験者が何点をつけたかを示すものである．表 9，10 は，表 8 に関して，被験者の各質問への採点結果の大小関係が，我々の効用設定時の想定と一致しているか否かを示したものである．具体的には，表 2 において，正解時の効用 U_1^+〜U_{Wh}^+ の値は $U_1^+ \geq U_2^+ \geq U_3^+ \geq U_{\mathrm{Wh}}^+$ とし，誤り時の効用 U_1^-〜U_{Wh}^- の値は，$U_{\mathrm{Wh}}^- \geq U_1^- \geq U_2^- \geq U_3^-$ とした．これは，正解できる時は，Wh 質問よりも選択肢を提示できる質問の方が良く，質問に正しいジャンルが含まれない時は，ジャンルを提示するよりも Wh 質問の方が良いという考えによるものである．表 9 は，Wh 質問を除いた，Yes/No 質問，二択質問，三択質問の三種類での一致数を示したものであり，表 10 は Wh 質問を含めた場合の一致数を示したものである．また，表 9，10 共に，質問への採点結果が同点の場合を含めるか否か二つの場合の結果を載せている．例えば表 8 の被験者 B は，誤り時の採点を Yes/No 質問，二択質問，三択質問で同点にしている．同点を許す場合は，被験者 B の結果を大小関係が一致するとし，同点を許さない場合は一致しないとした．

表 9: 被験者の印象と効用の値の関係 (Wh 質問を除く)

パターン	正解時の一致数			誤り時の一致数		
	研究室	外部	全体	研究室	外部	全体
同点許す	8/9	6/6	14/15	8/9	5/6	13/15
同点許さず	8/9	6/6	14/15	3/9	4/6	7/15

表 10: 被験者の印象と効用の値の関係 (Wh 質問を含む)

パターン	正解時の一致数			誤り時の一致数		
	研究室	外部	全体	研究室	外部	全体
同点許す	8/9	4/6	12/15	5/9	3/6	8/15
同点許さず	5/9	4/6	9/15	1/9	1/6	2/15

表 9 から，Wh 質問を除き，同点を許せば，正解/誤り双方の場合で効用の設定基準は妥当であったと言える．具体的には，正解時は 93%(14/15)，誤り時は 87%(13/15) の一致率が得られていた．つまり，正解時は提示ジャンルを限定した質問の方が賢く感じられ，誤り時は提示ジャンルが多い質問の方が賢くないと感じる傾向があったと言える．一方，同点を許さない場合，誤り時の一致率が低下することが分かる．例えば，被験者 A，B，E，I，a は，Yes/No 質問，二択質問，三択質問全てを同点としていた．

表 10 から，正解する質問と Wh 質問の比較において，効用の設定基準が概ね妥当であることが確認できる．具体的には，80%(12/15) で被験者の採点結果と効用の大小関係は合致した．ただし，被験者 c，d のように，三択質問よりも Wh 質問の方が賢いと判定する被験者がいた．一方，誤った質問と Wh 質問の比較では，効用の設定基準が妥当だとは言い難い．例えば，被験者 A，H，d は，誤り時でも Wh 質問が最も賢くない印象を持っていたことが表 8 よりわかる．これは，Wh 質問が何も推定しない質問であるという印象を与え，推定した結果生成された質問の方が，その正誤に依らず，賢そうな印象を与えた可能性がある．今後の実験では，

得点付けの理由を回答してもらい，この結果の分析を行う．

表 9，10 から，同点を許さない場合での，採点結果と効用との一致数が低下したことは，今後分析する必要がある．この傾向は，ユーザがそもそも推定の正誤のみにしか関心がなく，質問内で提示するジャンル数が多かろうと少なかろうと賢さの違いがないと感じている可能性がある．この点に何らかの結論を出すために，今後は被験者数を増やすだけでなく，「質問を聞いていてどの程度煩わしかったか」という尺度でも印象調査を実施予定である．

6　おわりに

本稿では，DB 外のレストランのジャンルをユーザから獲得するための最適な質問を選択する手法を述べ，被験者実験による手法の妥当性に関して報告した．今後はさらに被験者実験を実施し，ユーザのふるまいを詳細に分析したい．具体的には，質問の賢さの採点時に効用の値と一致しなかったものに対し，何故そうなったのか分析を行う．また，生成した質問に対しユーザがどのようなふるまいをするのかも分析予定である．

参考文献

[1] Preethi Jyothi, Leif Johnson, Ciprian Chelba, and Brian Strope. Large-scale discriminative language model reranking for voice-search. Proceedings of the NAACL-HLT 2012 Workshop: Will We Ever Really Replace the N-gram Model? On the Future of Language Modeling for HLT, pp. 41–49, 2012.

[2] 竹内翔大, 川波弘道, 猿渡洋, 鹿野清宏. 音声情報案内システムにおける質問応答データベース構築コスト削減の検討. 情報処理学会研究報告, Vol. 2009-SLP-15 pp. 1–6, 2009.

[3] 成松宏美, 中野幹生, 船越孝太郎, 長谷川雄二, 辻野広司. 音声質問応答システムのユーザ発話を用いた質問応答データベース拡張支援. 情報処理学会研究報告, Vol. 2009-SLP-77, pp. 1–6, 2009.

[4] Hidetsugu Nanba, Ryuta Saito, Aya Ishino, and Toshiyuki Takezawa. Automatic Extraction of Event Information from Newspaper Articles and Web Pages. ICADL 2013, LNCS 8279, pp. 171–175, 2013.

[5] Tsugumi Otsuka, Kazunori Komatani, Satoshi Sato and Mikio Nakano. Generating More Specific Questions for Acquiring Attributes of Unknown Concepts from Users. Proceedings of the 14th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pp. 70–77, 2013.

[6] George A. Miller, The magical number seven, plus or minus two: some limits on our capacity for processing information. Psychological review, Vol. 63-2, pp. 81–97, 1956.