6P-04

対話を通じた情報獲得のための質問選択とその実験的評価

大塚 嗣巳 † 駒谷 和範 ‡ 佐藤 理史 † 中野 幹生 §

†名古屋大学大学院工学研究科 ‡大阪大学産業科学研究所 §ホンダ・リサーチ・インスティチュート・ジャパン

1. はじめに

現状の対話システムでは，システム自身の知識にない単語が発話された時に，うまく対応できない．例えば，その単語を音声認識・言語理解できても，その内容がシステムの知識 (e.g. 検索対象データベース (以下，DB)) 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，DB に情報が登録されていない「牡丹亭」が発話された場合，システムはユーザ要求を満たす応答ができないという問題が生じる．

この問題を回避するには，現状では，対応できなかった単語に関して，人手でシステムの知識を拡張するしかない [1]．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 DB 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．

我々はこの問題に対して，オンラインで情報獲得することを目指しており，そのためにユーザに対して適切な質問をする手法を提案している [2][3]．具体的には，推定結果の確からしさと，質問形式に対するユーザの印象を考慮して，表 1 に示す四種類から，最適な質問形式を選択する．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 DB 中にない店舗のジャンル，としている．これにより，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．

2. 質問選択手法

質問選択は，確信度と効用を用いて算出した期待効用に基づき，適切な形式の質問を表 1 の四種類から選択する．その流れを図 2 に示す．詳細は [3] を参照されたい．

まず，入力された店舗名から，ジャンル推定結果の確信度 (Confidence Measure: *CM*) を生成する．確信度とは，ある店舗名が与えられた際の確からしさを，事後確率で表した値である．確信度を用いて，各質問形式の内容に正解が含まれる確率を算出する．具体的には，選択ジャンル分，確信度を足し合わせる．例えば，図 2 の確信度を用いた二択質問が正解する確率，つまり，居酒屋か和食のどちらかが正解である確率は，85%($0.64+0.21=0.85$) である．

次に，確信度と，効用 (Utility) を用いて，各質問形式に対する期待効用を算出する．本稿での効用とは，質問を聞いた時，ユーザがどの程度賢い/煩わしいと感じたかを数値化したものである．表 2 に定めた効用を示す．$U^{\{+,-\}}_{x\in\{1,2,3,\mathrm{Wh}\}}$ の x は，Wh 質問を除き，表 1 の提示ジャンル数を表す．また，$\{+,-\}$ は質問の正誤それぞれでの効用を表す．質問の正誤は，正解ジャンルが含まれるか否かで定める．正解時の効用は，$U^+_1 \geq U^+_2 \geq U^+_3 \geq U^+_{\mathrm{Wh}}$ となるように定めた．これは，質問が正解となる時は，Wh 質問よりも選択肢を提示する質問の方が賢く，選択肢を限定できるほど賢いと感じるであろう，という直感に基づく．誤り時の効用は，$U^-_{\mathrm{Wh}} \geq U^-_1 \geq U^-_2 \geq U^-_3$ となるように定めた．これは，誤るくらいなら Wh 質問の方が煩わしくなく，誤る時は選択肢が少ないほど煩わしくないであろう，という直感に基づく．

最後に，期待効用が最大となる質問形式を，最適な質問として選択する (式 (1))．本稿における期待効用は，推定結果の確からしさを踏まえた上で，質問内容がユーザに与える印象を見積もった値である．つまり，期待効用が最大となる質問形式は，推定結果を考慮した上で，ユーザに最も良い印象を与えるであろう質問となる．期待効用の算出方法を表 2 に示す．なお，式 (1) において Wh 質問の期待効用は 0 として，各期待効用と比較する．

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

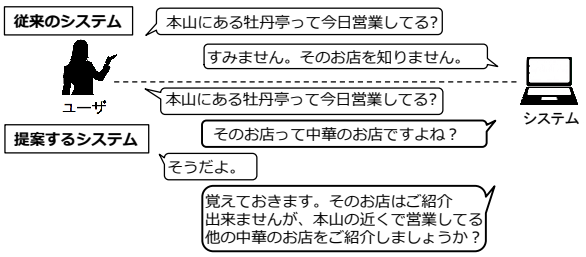

図 1: 従来の/提案システムにおける対話例

表 1: 選択に用いる四種類の質問形式

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?

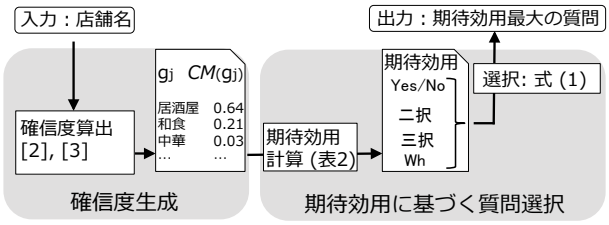

図 2: 質問選択の流れ

表 2: 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	$U^+_1=2$	$U^-_1=-2$	P_1	$U^+_1P_1+U^-_1(1-P_1)$
二択	$U^+_2=\frac{3}{2}$	$U^-_2=-3$	P_2	$U^+_2P_2+U^-_2(1-P_2)$
三択	$U^+_3=\frac{4}{3}$	$U^-_3=-4$	P_3	$U^+_3P_3+U^-_3(1-P_3)$
Wh	$U^+_{\mathrm{Wh}}=0$	$U^-_{\mathrm{Wh}}=0$	–	0

$P_x = \sum_{i=1}^{x} CM(g_i)$ (g_i : $CM(g_i)$ の値が i 番目に大きいジャンル)

Wh 質問ではジャンルを提示しないため，P_{Wh} は定義しない

Question Selection for Information Acquisition through Dialogue and Its Experimental Evaluation: *Tsugumi Otsuka (Nagoya Univ.), Kazunori Komatani (Osaka Univ.), Satoshi Sato (Nagoya Univ.), and Mikio Nakano (Honda Research Institute Japan Co., Ltd.)*

組	回答する印象	質問セット	
(A)	賢さ	Yes/No質問内に正解あり 三択質問内に正解あり	二択質問内に正解あり Wh質問
(B)	煩わしさ	Yes/No質問内に正解なし 三択質問内に正解なし	二択質問内に正解なし Wh質問
(C)	賢さ 煩わしさ	Yes/No質問内に正解あり 二択 or 三択質問内に正解あり	Yes/No質問内に正解なし Wh質問

図 3: 採点する質問セット

3. 質問選択に関する実験的評価

以下の 2 点を検証した．

1. 定めた効用が，人間の感じる実際の印象と比較して妥当なのか．
2. DB 内の店舗を用いた交差検定でなく，オープンデータである DB 外の店舗に関して，正しい内容を含む質問が選択できるのか．

1 点目では，質問に対するユーザの印象を，被験者実験により収集し，効用の設定基準と比較する．我々は，質問を聞いた際にユーザが感じる賢さ/煩わしさを基準に正解/誤り時の効用を定めた．[3] では，賢さのみを尺度に用いていたが，本稿では両方の尺度で妥当性を検証する．2 点目では，DB 外の店舗を用いて質問選択を行い，その性能を評価する．その性能が，従来の性能，すなわち DB 内の店舗を 10 分割交差検定により DB 外店舗と見なした際の性能，と同程度の水準となるかを検証する．

1 点目の検証のために行った，被験者実験の手順を述べる．被験者には，こちらが用意した質問を実際に聞いてもらい，感じた印象を数値化してもらった．具体的には，どの程度賢く/煩わしく感じたかという，二つの印象をそれぞれ採点してもらった．採点には，−5〜5 点でのリッカートスケールによる採点方式を用いた．採点は，図 3 に示す三種類の質問セットの組に関して行った．各質問セットには四つの質問を用意し，それらを相対的に採点してもらった．対象被験者は 12 名である．

被験者実験の結果から，効用の値の設定基準は概ね妥当であることを確認した．つまり，正解/誤り時の効用は賢さ/煩わしさを数値化したもので良いことを確認した．表 3 は，各質問形式に関して，我々が定めた効用の値の順位と，被験者の採点結果の順位との一致数を示したものである．(C) は，賢さと煩わしさそれぞれの印象を回答してもらった結果を示した．

まず，組 (A) から，質問内容が正解となるならば，Wh 質問をするよりも選択肢を提示した方が賢い印象を与えることを確認できた．また，選択肢を限定できている方が，より賢い印象を与えることが確認できた．具体的には，正解時の効用の妥当性を検証した組 (A) では，11/12(92%) の一致数であった．これは，表 2 で $U_1^+ \geq U_2^+ \geq U_3^+ \geq U_{\mathrm{Wh}}^+$ となるように定めたことが妥当であったことを意味する．

次に，組 (B) から，質問内容が誤るのであれば，選択肢を提示するよりも Wh 質問の方が煩わしくない印象を与えることが確認できた．また，提示する選択肢の数が少ない方が，より煩わしくない印象を与えることが確認できた．具体的には，誤り時の効用の妥当性を検証した組 (B) では，10/12(83%) の一致数であった．これは，表 2 で $U_{\mathrm{Wh}}^- \geq U_1^- \geq U_2^- \geq U_3^-$ となるように定めたことが妥当であったことを意味する．

組 (C)-賢さから,「質問の賢さ」の点で考えた場合，表 2 で $U_x^+ \geq 0 \geq U_x^-$ となるように定めたことは概ね妥当でありそうだと言える．具体的には，組 (C)-賢さの一致数が 9/12(75%) であった．この結果は，ユーザが質問の賢さを考える際，その内容が正解したか否かを考慮しているためだと考えられる．

表 3: 効用の順位と被験者の印象の一致数

組	一致数
(A)	11/12 (92%)
(B)	10/12 (83%)
(C)-賢さ	9/12 (75%)
(C)-煩わしさ	5/12 (42%)

表 4: 正解ジャンルを含む質問を生成できた割合

店舗集合	都道府県	正解率
DB 内	愛知県	72.0% (1193/1656)
DB 外	愛知県	73.5% (236/ 321)
	東京都	70.0% (290/ 414)
	その他	69.7% (641/ 919)
	全店舗	70.6% (1167/1654)

一方，組 (C)-煩わしさから,「質問の煩わしさ」の点で考えた場合，表 2 で $U_x^+ \geq 0 \geq U_x^-$ となるように定めたことが妥当でない可能性があることが分かった．具体的には，組 (C)-煩わしさの一致数は 5/12(42%) に留まった．一致しなかった 7 名は，二択 or 三択の中に正解がある質問よりも，Wh 質問の方が煩わしくないという印象を持っていた．この結果は，ユーザが，質問の煩わしさを考える際，その内容が正解したか否かよりも，システムの発話時間に主眼を置いたためだと考えられる．

2 点目の検証のため，DB 外の店舗に関して，質問選択を行い，DB 内の店舗での結果と比較した．DB 内の店舗とは，愛知県内 1656 件の飲食店である．DB 外の店舗とは，DB 内に存在しない，愛知県，東京都とその他都道府県にある 1654 件の飲食店である．その他都道府県は，北海道，神奈川県，石川県，大阪府，京都府の全 5 都道府県である．正解ジャンルは，対象 DB 内のジャンルのフィールドに登録できる 16 種類の中から，最も近いであろうものを，実験前に人手で選択した．質問の正誤は，その内容に正解ジャンルが含まれているか否かで判定する．

表 4 から，オープンデータである DB 外の飲食店に関しても，開発データと同等の正解率が得られることを確認した．また，愛知県のみならず，東京都やその他都道府県でも同等の正解率で正しいジャンルを含む質問生成が可能であり，質問選択の性能が地域に依存しないことも確認した．表 4 は，DB 内 1656 件と，DB 外の飲食店それぞれに関し，正解ジャンルを含む質問が生成できた割合を示したものである．DB 内にある愛知県内の飲食店で 72%(1193/1656) の正解率であったのに対し，DB 外にある愛知県の飲食店では，73.5%(236/321) の正解率であった．また，東京都では 70.0%(290/414)，その他都道府県では 69.7%(641/919)，愛知件，東京都，その他都道府県の全店舗では 70.6%(1167/1654) の正解率であった．

参考文献

[1] 吉村健. しゃべってコンシェルと言語処理. 情報処理学会研究報告, Vol.2012-SLP-93, pp. 1–6, 2012.

[2] T. Otsuka, K. Komatani, S. Sato and M. 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.

[3] 大塚嗣巳, 駒谷和範, 佐藤理史, 中野幹生. 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, 人工知能研資, SIG-SLUD-B402-07, pp. 37–44, 2014.