未知語属性獲得のための質問の形式がユーザに与える印象の実験的分析

Empirical Analysis of User Impressions by Question Types to Acquire Knowledge about Unknown Words

駒谷和範 [1*]　中野幹生 [2]
Kazunori Komatani[1]　Mikio Nakano[2]

[1] 大阪大学 産業科学研究所
[2] 株式会社ホンダ・リサーチ・インスティチュート・ジャパン
[1] The Institute of Scientific and Industrial Research (ISIR), Osaka University
[2] Honda Research Institute Japan Co., Ltd.

Abstract: For acquiring knowledge through dialogues, it is crucial for the system to ask questions that do not diminish the user's willingness to talk, i.e., that do not deteriorate the user's impressions. Implicit confirmation is a way for the system to verify whether or not its content is correct by using the dialogue contexts. We report the results of empirical analysis of users' impressions when users receive several question types including the implicit and explicit ones. User impressions were assessed via crowdsourcing from 104 participants. Regression analysis is conducted to investigate the impact on user impressions by each question type with correct or wrong content after individual deviation of the impression scoring and temporal position of the questions are taken into consideration. The results reveal that explicit questions are more annoying when repeated with correct contents than implicit ones. Helpful empirical insight is also provided into creating a strategy that would avoid user impression deterioration during knowledge acquisition.

1　はじめに

対話を通じて対話相手から知識を獲得できることは，対話システムが持つべき重要な能力のひとつである．これが可能になれば，未知語が現れた場合 [1, 2] や，初期の知識ベースが不完全である場合 [3] にも，システムは話すほどに知識を得ることができる．いくつかの研究でこのような問題は取り組まれており [4, 5, 6, 7, 8, 9]，Life-long learning というキーワードも使われ始めている [10, 11]．

対話を通じて知識を獲得するには，システムが質問をすることになるが，この際にシステムが行う質問をユーザが煩わしく感じたり，その結果としてユーザが対話を中断したりしないことが重要である．Amazon Alexa Prizes [12, 13] でも示されたように，対話を継続するということ自体も挑戦的な課題である．あたかもクラウドワーカに対して行うかのように [14]，システムが煩わしい質問を繰り返した場合，ユーザはシステムを使うのを止めてしまう．したがって，知識を獲得するために行う質問は，できるだけユーザが煩わしく感じないものとなるよう設計する必要がある．

唐突な質問を行うことなく知識を獲得するための手法として，我々は暗黙的確認を提案してきた [15, 8]．図 1 に暗黙的確認の例を示す．まず，ユーザ発話中に未知語が存在した場合，システムはまずその属性を推定する [4]（図 1 の (1)）．次に，その推定結果について明示的に質問はせず，推定結果を含めた発話，つまり暗黙的質問[1] を行う（図 1 の (2)）．それに対するユーザの応答を考慮することによって，暗黙的質問に含めた推定結果が正しいかどうかを推定し，正しいとみなせた場合にはその内容をシステム知識として獲得する [8]（図 1 の (3)）．暗黙的質問は明示的に質問するよりもユーザにとって煩わしくないとしてきたが，実験的に検証されてはいなかった．

本稿では，知識獲得のために行うシステム質問が，ユーザの印象にどのような影響を与えるかを，そのタイプごとに実験的に調査する．システムが行う質問の

*連絡先： 大阪大学産業科学研究所
大阪府茨木市美穂ヶ丘 8-1
e-mail: komatani@sanken.osaka-u.ac.jp

[1] 文献 [15, 8] では暗黙的確認要求 (implicit confirmation request) と呼んでいたものである．

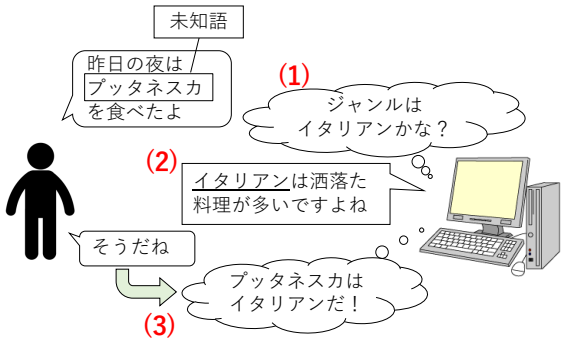

図 1: 暗黙的確認プロセスの例

タイプとして，明示的質問や暗黙的質問を用意し，それらの質問内容が正しい場合と誤っている場合とを含めて 5 種類設計した．これらを通じて，とりわけ,「明示的質問が連続すると煩わしい」かどうかを検証する．

質問タイプごとの印象は，複数回の質問を含むセッション全体に対する印象を尋ね，その結果から分析するというアプローチを採った．これに対して最も単純には，質問を一回行うごとにその印象を入力してもらうことも考えられるが，これは明らかに入力自体が煩わしく，対話の流れを阻害する．このため，セッションごとに付与された印象から，そこに含まれる質問タイプによる影響を回帰モデルにより推定する．またこのモデルを用いることで，同じ質問タイプが連続した場合の印象も分析する．

2 知識を獲得するための 5 つの質問タイプ

対話を通じた知識獲得の枠組みとして，獲得対象は，未知語とその属性値の組とする．つまり，未知語が現れた場合に，その属性を対話を通じて獲得して，新たな知識とすることを目標とする．本研究におけるタスクでは，具体的には料理名とそのジャンルの組とする．まずこのジャンルをその未知語の文字列情報などから推定し [4]，これが正しいかどうかを質問を通じて検証して，正しいと分かれば知識として獲得することを目指す．

システムは，5 種類の質問タイプのうちのいずれかに基づく質問を行うことで，その推定された属性値が正しいかどうかを確かめる．質問タイプの構成要素として，まず明示的（explicit）質問，暗黙的（implicit）質問と，Wh 質問を考える．明示的質問では，その内容が正しいか否かを，Yes/No 質問を通じて，例えば「プッタネスカはイタリアンですか？」のように明示的に尋ねる．これを E で表す．暗黙的質問は I で表す．これは，システムが「イタリアンはデートにぴったりですね。」のように，推定値を含めた発話を行って対話を続け，その後のユーザ発話も考慮することで，その推定値が正しかったかどうかを決定する [8]．Wh 質問は Whq で表し，推定された属性値を使うことなく，単純に，例えば「プッタネスカって何ですか？」のように尋ねる．

EC	プッタネスカはイタリアンですか？
EW	プッタネスカは和食ですか？
IC	イタリアンはデートにぴったりです。
IW	和食は健康にいいです。
Whq	プッタネスカって何ですか？

表 1: 5 つの質問タイプの例

さらに，明示的質問と暗黙的質問それぞれに，その内容が正しい場合と誤りである場合を用意する．推定した内容の正誤がユーザの印象に与える影響を調べるためである．このため，C と W をそれぞれ正しい内容 (correct) と誤った内容 (wrong) を表す記号として，E と I に付加する．Whq は推定値を使わないので，これらの記号は付加されない．簡単のため，明示的質問の選択肢はひとつだけとし，それ以上の数の内容を含んだ明示的質問 [16] はここでは考慮しない．

表 1 に，設計した 5 種類の質問タイプとその例を示す．これは「プッタネスカ」が未知語とした場合，それに対する正しいジャンル推定結果がイタリアンで，誤ったジャンル推定結果が和食であるとした場合の例である．

3 ユーザスタディの設計

前節で準備した 5 種類の質問タイプによる印象をクラウドソーシングにより調査した．クラウドソーシングのプラットフォームにはクラウドワークス社[2]のものを利用した．

印象評定はセッションごとに行った．一つのセッションは，3 セットのやりとりと印象評定から構成される．1 セットのやりとりは，ワーカの入力 2 ターンとシステム発話 2 ターンの，4 ターンからなる．この 4 ターンは具体的には以下のとおりである．

1. ワーカが，指定された単語を含む発話を入力する．この単語は実験者側で用意し，例えば「フリカッセについて何か入力してください」のように表示することでワーカに伝えた．

2. システムが，入力された単語に関する 5 つの質問タイプのいずれかにを用いて質問する．質問タイプはランダムに選択する．ここで用いる誤った内

[2] https://crowdworks.co.jp/

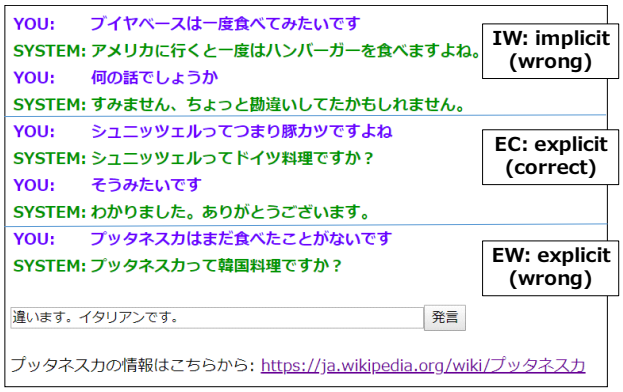

図 2: 3 セット目のやりとり途中でのスクリーンショット．なお右側の質問タイプを含む四角は説明用であり実験時には表示されない．

容（料理ジャンルの推定結果）や暗黙的質問の表現は，それぞれの単語について人手で用意した．

3. ワーカがその質問に対する応答を入力する．

4. システム応答が表示される．このシステム応答は，各質問タイプごとに固定とし，人手で用意した．例えば，誤った暗黙的質問（IW）に対する応答は「すみません，ちょっと勘違いしていたかもしれません。」である．

4 ターンからなるやりとりが終われば，次のセットのやりとりにおける指定単語が示される．3 セットのやりとりを終えた後，ワーカは，同じ画面上に表示されるインタフェース（図 3）により，その印象を入力した．アンケートは 7 段階で，「システムの発話を煩わしいと感じましたか？」と「システムは賢いと感じましたか？」の 2 項目とした．以降，個々で得られた印象スコアを，それぞれ「煩わしさ」「賢さ」と表記する．

各ワーカは，このセッションを 10 回繰り返した．未知語とする単語は 1 セッションあたり 3 個ずつ，合計 30 個を用意した．

図 2 に対話例を示す．YOU と SYSTEM から始まる行は，それぞれワーカの入力とシステムの応答を表す．各やりとり冒頭の，指定単語を指示する部分については，表示されていない．ワーカが指定単語（料理名）を知らない場合のために，最下部にあるリンクから Wikipedia のページをチェックできるようにした．

この結果，104 名のワーカから，1,183 セッション分のデータを得た[3]．ここでは 10 セッション全てを終えなかったワーカによるデータなどは取り除いた．これにより，質問タイプのいずれかを 3 回行ったセッション（3,549 発話）に対応して，「煩わしさ」「賢さ」に関する 1183 個の印象スコアを得た．

[3] システムの不具合により，一部のワーカは 10 を超えるセッションで対話を行った．

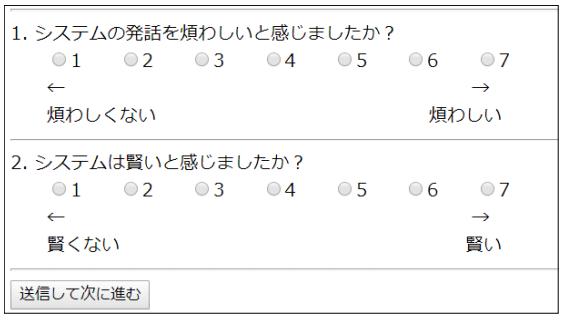

図 3: セッション終了後に表示されるアンケート画面

4 線形回帰による分析

収集した印象スコアを予測する線形回帰モデルの係数を求め，得られた係数を用いてそれぞれの質問タイプによる影響を分析する．まず，基本とする回帰モデルについて述べ，より高い決定係数を得るための改善について述べる．続いて，このモデルを用いて，それぞれの質問タイプがユーザの印象に与える影響や，同じタイプの質問が繰り返された場合の分析結果を示す．

4.1 回帰モデル

ユーザの印象スコアを予測するのに線形回帰モデルを用いる．説明変数はセッション内での 5 つのシステムの質問タイプの使用回数，目的変数はそのセッションに対する印象スコアのいずれか，つまり「煩わしさ」もしくは「賢さ」とした．i 番目のセッションに対して，基本とする回帰モデルの式は以下である．

$$\text{score}_i = w_0 + \sum_{c \in \{EC, EW, IC, IW, Whq\}} w_c n_i(c)$$

ここで $n_i()$ は i 番目のセッションで使われたそれぞれの質問タイプの回数であり，この基本とするモデルでは $0, 1, 2, 3$ のいずれかである．以降，賢さと煩わしさは概ね逆傾向を示していたため，記述の重複を防ぐため「賢さ」のスコアの結果のみを示す．

ここで，決定係数を高めるために，2 つの改善を実施した．まず，平均 0 分散 1 になるように各ワーカごとに印象スコアを正規化し，正規化された値を説明変数とした．これはワーカごとに印象スコアのレンジが異なるため必要である．つまり，7 段階評価の高い部分にスコアが分布するワーカや，低い部分に分布するワーカが存在するためである．ここで知りたいのは，質問タイプによる影響であるため，個々のワーカ間の違いを正規化によりキャンセルした．つまりワーカ内での相対評価を用いることに相当する．

次に，1 セッションの中での質問タイプの位置を考慮した．つまり，5 つの質問タイプそれぞれについて，

基本となる回帰モデル	0.368
＋ ワーカごとにスコアの正規化	0.493
＋ セッション内の位置を考慮	0.540

表 2: 回帰モデルの決定係数 R^2

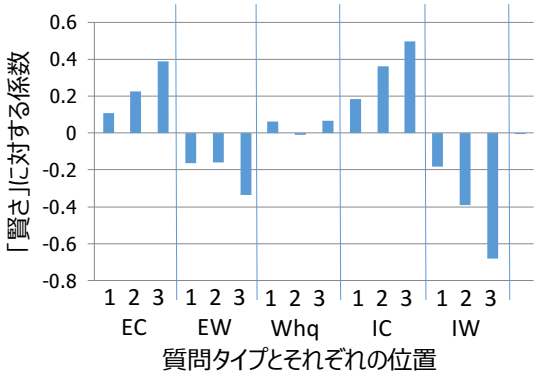

図 4: 「賢さ」に対する回帰モデルの係数

セッションの中で用いられた位置（1 番目，2 番目，3 番目）を考慮して 15 個の説明変数とした．決定係数の変化を表 2 に示す．以降，正規化を施し位置を考慮したモデルにおける 15 個の説明変数の係数を用いて分析を進める．

4.2 得られた係数の分析

「賢さ」に対する回帰モデルの 15 個の係数の値を，図 4 に示す．値が正であり大きいほど，その質問によりユーザに賢いと感じさせたことを示す．負の値はその逆を示す．この図において，3 つの位置に関する値を平均すると，係数は以下の順序を持つことがわかる．

$$\mathrm{IC} > \mathrm{EC} > \mathrm{Whq} > \mathrm{EW} > \mathrm{IW}$$

正しい内容を持つ IC と EC の係数は正で，誤った内容を持つ EW と IW の係数は負であった．これは，システムが誤った内容の質問を行ったとき，ユーザがシステムのことを賢くないと感じるという事実に合致する．具体的な内容を持たない Wh 質問（Whq）はこれらの中間であった．

次に，明示的質問と暗黙的質問との間の関係に注目する．内容が正しいとき，暗黙的質問（IC）の係数は，明示的質問（EC）の係数よりも大きい．これは，暗黙的質問が，明示的質問よりも賢いと感じられることを表している．ログを見たところ，対象とした料理名がより珍しいものである場合ほど，「賢さ」を示す印象スコアが高くなる傾向にあった．つまり，システムが対象とする単語について知識を持っているように感じたと思われる．一方で，誤った内容について質問を行った場合，明示的質問（EW）の係数は，暗黙的質問（IW）の係数よりも，絶対値が小さい．つまり，内容が誤っていた場合には，明示的質問の方が相対的にダメージが小さい．これは，誤った内容の暗黙的質問をシステムが行った場合，ユーザは自分の直前の発話に答えることなくシステムが自分勝手に新しい話題を開始したように感じたためと考えられる．

またこの図から，3 回の質問の中での位置に関する傾向も読み取れる．係数が正である場合も負である場合も，5 つ全ての質問タイプにおいて，3 番目の位置の係数が最も大きい．つまり，アンケートを記入する直前の質問が，印象スコアに対して大きな影響を持っていたことが示唆されている．

4.3 同じタイプの質問が連続した場合の印象

明示的質問が繰り返された場合が，暗黙的質問が繰り返された場合よりも煩わしいかどうかを検証する．この節では，印象スコアとして「煩わしさ」を使用した場合の結果を示す．

ここでは，同じ質問タイプが繰り返された場合について，以下の 2 つのユーザの印象スコアを比較する．

- 回帰モデルによる予測値
- 実際に同じタイプの質問が 3 回繰り返された場合の平均値

まず，既に述べた回帰モデルを用いて，それぞれのタイプの質問が繰り返された場合の印象スコアの予測値を得ることができる．この回帰モデルの係数は，それぞれの質問タイプがランダムに選ばれていたことから，前後の質問タイプを考慮していないスコアであるとみなせる．つまり，これらの係数は，同じタイプの質問が繰り返されたかどうかを考慮することなく求められたものである．

一方で，収集したデータの中には，ランダムに質問タイプを選択した結果として，実際に同じタイプの質問が 3 回連続で選ばれていたケースが含まれる．このようなケースは，1 つの質問タイプについて，平均 10.4 回存在した．この 2 つの値を比較することで，前後の質問タイプに関係なく，それぞれの質問タイプ自身が持つ「煩わしさ」が 3 回繰り返された場合の予測値と比較して，実際に同じ質問タイプが繰り返された場合の「煩わしさ」がどの程度違うかを分析できる．

図 5 にその結果を示し，その具体的に値を表 3 に示す．これらより，「煩わしさ」は，全ての質問タイプについて，同じタイプの質問が実際に繰り返された場合の値の平均は，回帰モデルにより予測された値よりも「煩わしさ」が増加していた．質問タイプごとにより詳しく見ると，「煩わしさ」の増加は，正しい内容の質問

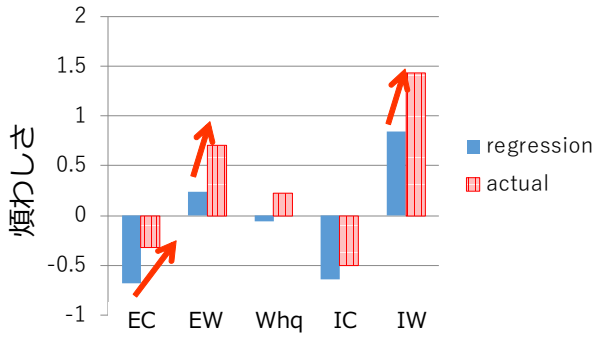

図 5: 「煩わしさ」の予測値と実際値の平均の比較

	予測値	実際値の平均	差分
EC	−0.679	−0.324	+0.355
EW	0.239	0.707	+0.468
Whq	−0.058	0.224	+0.282
IC	−0.639	−0.502	+0.137
IW	0.842	1.429	+0.587

表 3: 同じタイプの質問が 3 回繰り返された場合の「煩わしさ」の予測値と実際値の平均

（EC や IC）の場合よりも，誤った内容の質問（EW や IW）の場合の方が大きい．これは，誤った内容の質問を繰り返した場合は，正しい内容の質問を繰り返した場合よりもより煩わしいという事実に相当する．

この結果を用いて，「明示的質問が連続すると煩わしいか」という問題に対して回答することができる．図 5 や表 3 において，予測値と実際値の平均を比較した場合，正しい明示的質問（EC）における「煩わしさ」スコアの増加は，正しい暗黙的質問（IC）における増加よりも大きい．具体的には，これらの値はいずれも負であることから，「煩わしくない」とする度合が減少していることになる．すなわち，同じ質問タイプを繰り返した場合の印象の劣化は，正しい暗黙的質問（IC）よりも正しい明示的質問（EC）の場合の方が大きい．つまり，内容が正しい場合でも，明示的質問（EC）を繰り返すことは，暗黙的質問（IC）を繰り返すことよりも煩わしいという結果が示されている．

この結果の原因のひとつとして，明示的質問は全て単純で同じ形式（例えば「プッタネスカはイタリア料理ですか？」）であることから，システムが何も考えることなく質問を繰り返しているようにユーザに感じられるという点が挙げられる．一方で暗黙的質問は，未知語の料理ジャンルが正しく推定されていた場合には，ユーザの発話を受け入れた回答（質問）となることから，対話の流れを乱しておらず，煩わしさが低いと考えられる．

5 おわりに

5 種類の質問タイプがユーザの印象に与える影響を実験的に調査した．同じタイプの質問を繰り返すと，その内容が正しい場合でも，ユーザは煩わしいと感じ，ユーザの印象は悪化する．さらに，その内容が正しい場合，明示的質問が繰り返された場合は，暗黙的質問を繰り返した場合よりもユーザの印象はより悪化する．これらの結果は，知識獲得のための質問を設計するにあたって重要な知見である．

質問を繰り返す場合，明示的質問と比較して，質問内容が正しい場合には暗黙的質問はより良い印象を与えることができる．一方で質問内容が誤っている場合には，暗黙的質問の印象は明示的質問よりも悪い．これにより，ユーザの印象が悪化するリスクを低減するために，料理ジャンル推定の確信度を用いた戦略が考えられる．具体的には，確信度が高い場合には暗黙的質問を用い，一方で確信度が低い場合には推定が誤っていた場合の印象劣化の大きい暗黙的質問を避け，明示的質問や Wh 質問を用いるという戦略が合理的となる．

本稿における実験では，暗黙的質問の具体的な表現は，それぞれの入力単語ごとに人手で用意した．今後の課題として，暗黙的質問の表層表現を自動的に得る手法が必要である．次に，今回の印象スコアには，実験の仕様上，設計した 5 種類の質問タイプの質問だけでなく，ユーザの回答の後のシステム発話（3 節におけるやりとり内の 4 ターン目）による印象も含まれる可能性がある．このため，本来はここにも応答表現を複数用意し，ランダム化するなどして影響を抑える必要もある．また，最適な質問戦略を考える際には，それぞれの質問の知識獲得に関する効用 [16] も考慮されるべき要素である．

参考文献

[1] Raveesh Meena, Gabriel Skantze, and Joakim Gustafson. A data-driven approach to understanding spoken route directions in human-robot dialogue. In *Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 226–229, 2012.

[2] Ming Sun, Yun-Nung Chen, and Alexander I. Rudnicky. Learning OOV through semantic relatedness in spoken dialog systems. In *Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 1453–1457, 2015.

[3] Robert West, Evgeniy Gabrilovich, Kevin Murphy, Shaohua Sun, Rahul Gupta, and Dekang Lin. Knowledge base completion via search-based question answering. In *Proc. International Conference on World Wide Web (WWW)*, pp. 515–526, 2014.

[4] Tsugumi Otsuka, Kazunori Komatani, Satoshi Sato, and Mikio Nakano. Generating more specific questions for acquiring attributes of unknown concepts

from users. In *Proc. Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pp. 70–77, August 2013.

[5] Aasish Pappu and Alexander Rudnicky. Knowledge acquisition strategies for goal-oriented dialog systems. In *Proc. Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pp. 194–198, June 2014.

[6] Ben Hixon, Peter Clark, and Hannaneh Hajishirzi. Learning knowledge graphs for question answering through conversational dialog. In *Proc. North American Chapter of Association for Computational Linguistics (NAACL)*, pp. 851–861, 2015.

[7] Jason Weston. Dialog-based language learning. In *Proc. International Conference on Neural Information Processing Systems (NIPS)*, pp. 829–837, 2016.

[8] Kohei Ono, Ryu Takeda, Eric Nichols, Mikio Nakano, and Kazunori Komatani. Lexical acquisition through implicit confirmations over multiple dialogues. In *Proc. Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pp. 50–59, 2017.

[9] Jiwei Li, Alexander H. Miller, Sumit Chopra, Marc'Aurelio Ranzato, and Jason Weston. Learning through dialogue interactions by asking questions. In *Proc. International Conference on Learning Representations (ICLR)*, 2017.

[10] Sahisnu Mazumder, Nianzu Ma, and Bing Liu. Towards a continuous knowledge learning engine for chatbots. *Computing Research Repository*, Vol. arXiv:1802.06024, , 2018. version 2.

[11] Sahisnu Mazumder, Bing Liu, Shuai Wang, and Nianzu Ma. Lifelong and interactive learning of factual knowledge in dialogues. In *Proc. Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pp. 21–31, 2019.

[12] Hao Fang, Hao Cheng, Maarten Sap, Elizabeth Clark, Ari Holtzman, Yejin Choi, Noah A. Smith, and Mari Ostendorf. Sounding board: A user-centric and content-driven social chatbot. In *Proc. North American Chapter of Association for Computational Linguistics (NAACL)*, pp. 96–100, June 2018.

[13] Chun-Yen Chen, Dian Yu, Weiming Wen, Yi Mang Yang, Jiaping Zhang, Mingyang Zhou, Kevin Jesse, Austin Chau, Antara Bhowmick, Shreenath Iyer, Giritheja Sreenivasulu, Runxiang Cheng, Ashwin Bhandare, and Zhou Yu. Gunrock: Building a human-like social bot by leveraging large scale real user data. In *2nd Proceedings of Alexa Prize (Alexa Prize 2018)*, 2018.

[14] Saleema Amershi, Maya Cakmak, William Bradley Knox, and Todd Kulesza. Power to the people: The role of humans in interactive machine learning. *AI Magazine*, Vol. 35, No. 4, December 2014.

[15] Kohei Ono, Ryu Takeda, Eric Nichols, Mikio Nakano, and Kazunori Komatani. Toward lexical acquisition during dialogues through implicit confirmation for closed-domain chatbots. In *Proc. of Second Workshop on Chatbots and Conversational Agent Technologies (WOCHAT)*, 2016.

[16] Kazunori Komatani, Tsugumi Otsuka, Satoshi Sato, and Mikio Nakano. Question selection based on expected utility to acquire information through dialogue. In *Proc. International Workshop on Spoken Dialogue System Technology (IWSDS)*, pp. 27–38, 2016.