知的対話アシスタントにおける意図が曖昧な発話の検出

赤崎 智　　颯々野 学
ヤフー株式会社
sakasaki@yahoo-corp.jp　　msassano@yahoo-corp.jp

概要

対話によってユーザとの雑談とタスクの両方をこなす知的対話アシスタントにおいては「お腹が空いた」,「腰痛がひどい」のような雑談ともタスクとも取れる意図が曖昧なユーザ発話が存在する．そのような発話の意図を決定的に推定すると，結果として要求に沿わない応答をすることがありユーザ体験が損なわれてしまう．そのため，予めユーザの発話の曖昧性を判定し，必要であれば後の応答で対応することが望ましい．本稿では，実際の知的対話アシスタントのログデータにクラウドソーシングによるラベル付けを行い，意図が曖昧な発話にどのような傾向があるかを分析する．得られたラベル付きデータを用いて BERT による教師あり学習モデルを構築し，意図が曖昧な発話を検出することを試みる．

1　はじめに

スマートフォンや AI 搭載型デバイスなどの台頭により，Apple の Siri や Amazon の Alexa，Google の Google Home など，ユーザと対話をすることにより情報検索や端末操作などを行ったり，時には人間同士がするように雑談を行う知的対話アシスタントが普及して久しい [1].

知的対話アシスタントはユーザの要求に対応するために，与えられたユーザの発話の意図を適切に認識し，それに対応した後段のモジュールを呼び出し要求を満たす応答を行う．近年ではパターンマッチや機械学習モデルにより発話の意図を判定することが主流であるが [2, 3, 4, 5]，ユーザの発話には意図が曖昧なものが存在する．例えば,「お腹が空いた」という発話は単なる雑談のきっかけとも飲食店の情報を求めているとも取れるし,「東京駅」という発話は路線検索とも地図検索とも取れる．このような意図が曖昧な発話はタスク型，非タスク型の発話が混在し発話長も短い知的対話アシスタントにおいて顕著に現れ，それらの意図を決定的に推定してしまうと，結果的にユーザの要求にそぐわない不適切な応答を返してしまう可能性が高い．

このような意図が曖昧なユーザ発話に対処するために，聞き返しや意図を明確化する応答の生成 [6, 7, 8] を行う取り組みもあるが，実際の状況ではまずどの発話に対してそのような応答を行うかを判定する必要がある．また，これらの取り組みは基本的にタスク指向型対話システムを対象としており，知的対話アシスタントについてはどんな発話が曖昧な意図になるか明らかとなっていない．

これらを踏まえ，本稿では知的対話アシスタントにおける意図が曖昧なユーザ発話を判定する問題を設定する．我々は実際の知的対話アシスタントである Yahoo!音声アシストの対話ログからユーザ発話とシステム応答からなるペアを収集し，クラウドソーシングを利用してラベル付けを行う．得られたデータセットから，どのような発話が曖昧な意図になるのか傾向を確認したのち，教師あり学習によってそれらを検出するモデルを構築する．

2　関連研究

タスク指向型対話システムにおけるユーザ発話のドメイン・意図判定は古くから数多く取り組まれている [2, 3, 4]．いくつかの取り組みでは，発話についてのシステムのドメイン・意図判定の確信度に閾値を設けることにより意図が曖昧な発話を判定しているが [9, 10, 11, 12]，マルチドメインなシステムでは個別に閾値を設定することが難しい上，ドメインが増える度に対応する必要が生じる．

意図が曖昧な発話に焦点を当て，教師あり学習により発話の曖昧性判定を行う取り組みもある [13, 14, 15]．しかしながら，いずれの取り組みも対象をタスク指向型または非タスク指向型対話システムのどちらかに限定しており，曖昧性判定もそれに特化した問題設定や判定手法となっているため，知的対話アシスタントに対して同様の手法を適用することは難しい．

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

Akasaki [5] は知的対話アシスタントのユーザ発話を対象に，それが非タスク指向型（雑談）かタスク指向型の応答を求めているかを判定するためのデータセットを構築し，教師あり学習による二値分類で判定を行った．しかしながら，意図が曖昧な発話についても決定的にラベルを割り当てているため，それらの発話に対応できないという問題がある．

意図が曖昧なユーザ発話について，その意図を明確化するための質問を生成するという取り組みも行われている [6, 7, 16, 8, 12]．生成した質問をシステム応答として出力することで意図の曖昧性判定ができるが，ほとんどの研究はあくまで生成の部分に焦点を当てており，いつどのユーザ発話に対し生成を行うかという部分については無視している．

本稿では，タスク指向型と非タスク指向型対話システムの複合である知的対話アシスタントのユーザ発話の意図の曖昧性を分類により判定する．

3 意図が曖昧な発話の検出

本節では本稿で扱う知的対話アシスタントと，そこから意図が曖昧なユーザ発話を判定する課題について説明する．

3.1 知的対話アシスタント

知的対話アシスタントの例として，Apple の Siri や Amazon の Alexa，Yahoo! JAPAN の Yahoo! 音声アシストなどが挙げられる．いずれのシステムもユーザの要求を遂行するために音声またはテキストを用いユーザと対話を行う．近年ではスマートフォンやスマートスピーカーの普及，音声認識技術の向上により，従来よりも日常的に使用されるようになっている．知的対話アシスタントはシステム毎に軽度な違いはあるが，ほとんどはウェブを介した情報検索（例: 天気予報，交通情報，ウェブ検索），端末操作（例: 電話，時計，音楽再生）などのいわゆるマルチドメインのタスク指向型対話システムが有する機能に加え，人間同士がするようなあいさつや世間話といった雑談，すなわちオープンドメインな非タスク指向型対話システムの能力も併せ持つ．そのため，どちらか一方の機能を持つ対話システムよりも，幅広い要求に対応する必要がある．本稿では Yahoo! 音声アシスト[1)]を知的対話アシスタントとして用い，それらの実際のユーザとシステムとの対話のログを収集しデータセットを構築する．

1） https://v-assist.yahoo.co.jp/

ラベル	発話数
雑談	5,610
タスク	6,994
曖昧	7,396
計	20,000

表 1 ラベル別集計

得票数	発話数
10,9,8	1,260
7,6	2,675
5,4,3	5,851
2,1	6,402
0	3,812
計	20,000

表 2 「曖昧」ラベルへの得票数別集計

3.2 問題設定

本稿では知的対話アシスタントにおける意図が曖昧な発話を検出することを目的としているため，既存のドメイン・意図判定の取り組みを参考に問題を設定した．

既存の取り組み [2, 3, 4] では，タスク指向型対話システムにおいてユーザの発話がどのドメインに属するか，またドメイン内のどの意図であるかを分類するのが一般的であるが，知的対話アシスタントは前述した通り，マルチドメインのタスク指向型とオープンドメインな非タスク指向型の対話システムの複合であるという性質であるため，その両方の範囲の発話を扱う必要がある．また，一般の対話システムでは細かいドメイン意図を分類することが多いが，ドメインは固定ではなく拡充されていくことを考慮すると，ドメイン毎に学習データを整理し都度モデルを更新していくこともコストがかかる．

Akasaki [5] はこれらを踏まえ，知的対話アシスタントにおいてまずユーザの発話がタスク指向型の意図の「タスク」であるか，または非タスク指向型の意図の「雑談」であるかを判定する問題を設定した．これにより，ドメインなどの仕様変更の影響を少なくでき，かつ必要であれば後に判定毎の細かい意図分類などを行うことが可能となっている．本稿でもこの設定に則り，発話の意図が曖昧という「曖昧」のラベルを加えた三値分類として問題を設計する．これにより，システムはユーザの発話が「曖昧」だと検出したとき，それに応じた聞き返しや意図の明確化を行う応答ができる．

4 データセット

本節では意図が曖昧な発話の検出に用いるデータセットの構築手順について説明し，得られた会話データについて分析を行う．

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

ラベル	数	会話例
音声認識誤り	341	u_0: 今日の東京の　r_{-1}: 今，このような言葉が話題になっています．...　u_{-1}: 最新の話題 u_0: もっと他て　r_{-1}: 一生懸命歌います．でんでんむしむしかたつむり...　u_{-1}: 歌って
名詞	319	u_0: ポッカレモン　r_{-1}: ご指定のアプリはインストールされていないようです．　u_{-1}: モバゲーを開いて u_0: 会議場　r_{-1}: そ，そんなこと急に言われても困ります．　u_{-1}: 抱きしめる
質問	110	u_0: 今日は雨が降りますかね　r_{-1}: 私にクーラー機能が付いているといいんですが．　u_{-1}: 今日は暑いですね u_0: 何か食べたいですか　r_{-1}:11 時 9 分です．　u_{-1}: 今の時間じゃなくて
自己開示	82	u_0: お腹が出てきた　r_{-1}: こちらで進捗状況がわかります．　u_{-1}: みせて u_0: 目がおかしい　r_{-1}: どういたしまして．　u_{-1}: いえい
要望・命令	104	u_0: 自宅に帰りたいんだけど　r_{-1}: お話，つづけましょうよ．　u_{-1}: 何を話しましょう u_0: ちょっと寝たい　r_{-1}: 音声アシストと申します．　u_{-1}: 名前教えてくれ
指摘	16	u_0:17 時 5 分　r_{-1}: よく眠れましたか？ 今日の東京の天気は，曇後雨のようです．...　u_{-1}: 今何時 u_0: 会話が続きません　r_{-1}: もちろんです．　u_{-1}: 何を考えてるの
その他	28	u_0: ヒッヒッヒッ　r_{-1}: ウェブで見つかった「パズドラを終了」の情報です．　u_{-1}: パズドラを終了 u_0: (ゞノ・∀・｀)　r_{-1}:Yahoo!ロコで，さまざまなお店，施設が検索できます．　u_{-1}: プリズンスクール

表 3　構築したデータセットの対話例

4.1　構築手順

Yahoo! 音声アシストで 2014 年から 2022 年にかけてユーザとシステムが実際に行った対話から，10 回以上出現したユーザ発話 u_0 についてその前のシステム応答 r_{-1} およびユーザ発話 u_{-1} からなる会話 $< u_0, r_{-1}, u_{-1} >$ をランダムに 20,000 件収集する．この時，同一の発話は最大 5 件までに数を制限する．

次に得られた会話を，Yahoo! クラウドソーシング[2]のワーカに提示し「表示された会話の発話 u_0 に対しその意図を前述の 3 つのラベルから選択する」というタスクを依頼した．この時，一つの会話につき 10 人のワーカを割り当てた．

我々はここから，各会話に対し過半数である 6 以上の票を得たラベルを割り振った．また，いずれのラベルも得票が 5 以下のもの，すなわち票が割れている例を多数確認した．このような発話も意図が曖昧な所為で票が割れていると考えられるため，「曖昧」ラベルを割り振った．最終的な集計結果を表 1 に示している．また，表 2 に各会話の「曖昧」ラベルへの得票数の内訳を示している．これより，実際のユーザ発話には人間から見て意図が曖昧だと感じる例が多数存在することがわかる．

4.2　意図が曖昧な発話の分析

「曖昧」ラベルが割り当てられた会話からランダムに 1000 件抽出し，それらがなぜ曖昧であるか第一著者が分類を行った結果を表 3 に示す．

表 3 より，まず音声認識誤りと名詞の割合が高いことがわかる．音声認識誤りについては，かな漢字変換の誤りや語の脱落などが含まれており，結果的に意図が取れず曖昧となっている例が多かった．名詞や要望・命令，質問は多くは一般的には情報検索や端末操作の意図で用いられるが，例のように必ずしもそれらの意図と取れないものがあった．情報開示は雑談の文脈でよく使われるが，例のように暗黙的にタスクの要求をしているとも取れるものがあった．これらから，知的対話アシスタント特有の発話意図の曖昧性判定が必要であることがわかる．

5　実験

本節では構築したデータセットを用いて教師あり分類器を構築し，三値分類を行うことで意図が曖昧な発話の検出を試みる．

5.1　比較手法

以下の手法を用いて発話意図の分類を行い，性能を比較する．

AllAmbig: 全てのユーザ発話に対し「曖昧」ラベルを出力する手法．

Threshold: 多くのタスク指向型対話システムで用いられている，意図判定の閾値により曖昧なユーザ発話の判定を行う手法．我々は Akasaki [5] が構築した知的対話アシスタントの発話が雑談かタスク意図かを分類する意図判定データセット[3]を用い，BERT [17] モデルを fine-tuning し評価データを分類する．この時，モデルの予測に対する softmax のスコアが最大のラベルについて，そのスコア s が $0.5 \leqq s \leqq 0.8$ の時に「曖昧」ラベルを出力する．

BERT: 4 節で構築したデータセットを用いて BERT モデルを fine-tuning し，評価データの三値分類を行う手法．会話データの各発話，応答は [SEP] タグで繋いだ上でエンコーダに入力する．

2)　https://crowdsourcing.yahoo.co.jp/

3)　このデータセットではユーザの 1 発話のみが入力として与えられているため，モデルも u_0 のみを入力として用いる

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

BERT+文埋め込み: Akasaki [5] の研究を参考に，ツイートおよびウェブ検索ログで事前学習した言語モデルを用いてユーザ発話 u_0 の文埋め込みを獲得し，**BERT** の fine-tuning の際に分類層に追加の特徴量として与える．

5.2 設定

各手法で fine-tuning を行う BERT モデルについては bert-base を 2021 年 2 月の日本語 Wikipedia 約 1800 万文で事前学習したものを用いた．BERT モデルは Tensorflow2 で実装し，最大文長は 64，バッチ数は 16 に設定した．トークナイズは sentencepiece[4] を用い，最適化に Adam を用いた．学習と評価は，データセット 20,000 件を用いて五分割交差検証で行った．各手法で 3 エポック学習し，開発データで最も f 値が高かったモデルを評価データに適用した．

BERT+文埋め込みで用いる文埋め込み用の BERT については，2021 年 7 月から 2022 年 7 月の期間のウェブ検索の頻度上位クエリ 5000 万件と，同様の期間のランダムにサンプルしたツイート 5000 万件を sentencepiece でトークナイズしたものをそれぞれ用い，bert-base を 40 エポック事前学習した．構築した各 BERT を文埋め込みの出力で用いるため，対照学習により文埋め込みを獲得する手法である教師なし SimCSE [18] を各データで標準のパラメータで 1 エポック学習した．これらのモデルを用いて，各ユーザ発話 u_0 をエンコードし，分類用 BERT モデルのエンコード結果と連結し分類層に与えた．

5.3 結果

表 4 に評価データに対し三値分類を行った結果を示す．表 4 より，3 節で構築したデータセットで学習を行った **BERT** および **BERT+文埋め込み**が最も高い性能となっており，ラベル付きデータの有効性が示されている．

表 5 は各ラベルの F 値を示している．「曖昧」ラベルにおいて，**BERT** および **BERT+文埋め込み**は発話に対し全て「曖昧」ラベルと予測する **AllAmbig** の性能を上回っているため，意図が曖昧な発話の傾向を学習できているといえる．「雑談」と「タスク」ラベルの予測の閾値で曖昧性を判定する **Threshold** は「曖昧」ラベルに対する性能が極めて低かった．これについては Recall 等を確認したところ，ほとんど「曖昧」ラベルを出力できておらず，モデルのスコアの閾値などで曖昧性判定を行うことの難しさを示している．また，**BERT** と **BERT+文埋め込み**の間では有意な性能差は確認できなかった．

	Acc.	Prec.	Rec.	F1
AllAmbig	36.95	12.32	33.33	17.99
Threshold	61.70	62.74	64.61	54.87
BERT	78.35	79.02	78.35	78.62
BERT+文埋め込み	78.30	79.08	78.32	78.64

表 4 三値分類の各手法の分類結果

	雑談	タスク	曖昧
AllAmbig	0.00	0.00	53.96
Threshold	73.58	73.86	17.16
BERT	80.30	83.10	72.47
BERT+文埋め込み	80.86	82.49	72.57

表 5 三値分類の各手法のラベル別 F 値

BERT の出力を確認してみたところ，表 3 でのラベル「名詞」にあたるユーザ発話の判定誤りが多かった．これらの例は多くの場合名詞（句）一つのみの発話であり長さが短く，豊富な言語資源を生かしたモデルでも扱うのが難しい．また，名詞だからといってそれら全ての意図が曖昧というわけでもなく，例えば「熱海駅」は路線検索や地図検索を求める曖昧なタスク要求である可能性が高いが，一方で「カローラクロス」は該当するタスク要求が情報検索くらいしか存在せず，曖昧でない可能性が高い．これらを区別するためには，発話に対する対話システムの各モジュールの確信度などの情報も特徴として加える必要がある．

6 まとめ

本稿では，知的対話アシスタントで意図が曖昧なユーザ発話が存在することを指摘し，実際の知的対話アシスタントである Yahoo! 音声アシストのログデータにクラウドソーシングを利用してラベル付けを行うことで意図が曖昧な発話にどのような傾向があるかを分析した．実験では教師あり学習により「曖昧」ラベルを分類先に含んだ三値分類器を構築し，ラベル付きデータでの学習が有効であることを示した．

今後の予定としては，意図が曖昧な発話を判定するのに有効な特徴量や手法を設計し適用することである．また，本稿で構築したデータセットについては公開を予定している．

4) https://github.com/google/sentencepiece

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

参考文献

[1] Graeme McLean and Kofi Osei-Frimpong. Hey alexa… examine the variables influencing the use of artificial intelligent in-home voice assistants. **Computers in Human Behavior**, Vol. 99, pp. 28–37, 2019.

[2] Daniel (Zhaohan) Guo, Gokhan Tur, Scott Wen tau Yih, and Geoffrey Zweig. Joint semantic utterance classification and slot filling with recursive neural networks. In **Proceedings of IEEE SLT Workshop**, 2014.

[3] Puyang Xu and Ruhi Sarikaya. Contextual domain classification in spoken language understanding systems using recurrent neural network. In **Proceedings of ICASSP**, pp. 136–140, 2014.

[4] Joo-Kyung Kim, Gokhan Tur, Asli Celikyilmaz, Bin Cao, and Ye-Yi Wang. Intent detection using semantically enriched word embeddings. In **Proceedings of IEEE SLT Workshop**, 2016.

[5] Satoshi Akasaki and Nobuhiro Kaji. Chat detection in an intelligent assistant: Combining task-oriented and non-task-oriented spoken dialogue systems. In **Proceedings of ACL**, 2017.

[6] Johannes Kiesel, Arefeh Bahrami, Benno Stein, Avishek Anand, and Matthias Hagen. Toward voice query clarification. In **Proceedings of SIGIR**, pp. 1257–1260, 2018.

[7] Mohammad Aliannejadi, Hamed Zamani, Fabio Crestani, and W Bruce Croft. Asking clarifying questions in open-domain information-seeking conversations. In **Proceedings of SIGIR**, pp. 475–484, 2019.

[8] Hamed Zamani, Susan Dumais, Nick Craswell, Paul Bennett, and Gord Lueck. Generating clarifying questions for information retrieval. In **Proceedings of WWW**, pp. 418–428, 2020.

[9] Kazunori Komatani and Tatsuya Kawahara. Flexible mixed-initiative dialogue management using concept-level confidence measures of speech recognizer output. In **Proceedings of COLING**, pp. 467–473, 2000.

[10] Ingrid Zukerman, Su Nam Kim, Thomas Kleinbauer, and Masud Moshtaghi. Employing distance-based semantics to interpret spoken referring expressions. **Computer Speech and Language**, Vol. 34, pp. 154–185, 2015.

[11] Svetlana Stoyanchev and Michael Johnston. Localized error detection for targeted clarification in a virtual assistant. In **Proceedings of ICASSP**, pp. 5241–5245, 2015.

[12] Kaustubh D Dhole. Resolving intent ambiguities by retrieving discriminative clarifying questions. **arXiv preprint arXiv:2008.07559**, 2020.

[13] 大原康平, 佐藤翔悦, 吉永直樹, 豊田正史. 対話によって曖昧性解消を行う質問応答. 第 9 回データ工学と情報マネジメントに関するフォーラム, 2017.

[14] Joo-Kyung Kim, Guoyin Wang, Sungjin Lee, and Young-Bum Kim. Deciding whether to ask clarifying questions in large-scale spoken language understanding. In **Proceedings of ASRU**, pp. 869–876. IEEE, 2021.

[15] Shohei Tanaka, Koichiro Yoshino, Katsuhito Sudoh, and Satoshi Nakamura. Reflective action selection based on positive-unlabeled learning and causality detection model. **Computer Speech and Language**, 2023.

[16] 中野佑哉, 河野誠也, 吉野幸一郎, 中村哲ほか. 対話によって曖昧性解消を行う質問応答. 第 244 回自然言語処理研究会, 2020.

[17] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In **Proceedings of NAACL-HLT**, pp. 4171–4186, 2019.

[18] Tianyu Gao, Xingcheng Yao, and Danqi Chen. SimCSE: Simple contrastive learning of sentence embeddings. In **Proceedings of EMNLP**, 2021.

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).