知的対話アシスタントにおける発話の雑談意図の判定

赤崎 智[1]† 鍛治 伸裕‡
†東京大学大学院 情報理工学系研究科 ‡ヤフー株式会社
akasaki@tkl.iis.u-tokyo.ac.jp nkaji@yahoo-corp.jp

1 はじめに

近年，Apple の Siri などをはじめとする，ユーザーと対話的やり取りを行いユーザーの代わりに調べ物をしたり，雑談をしたりする対話システム（以下，知的対話アシスタントとする）が普及してきている．

知的対話アシスタントは対話システムにおけるタスク型･非タスク型両方の側面を持つため，ユーザーの発話に対してその意図を正しく判定することは難しい．特に非タスク型，いわゆる雑談発話については発話の多様性が高く，ある種の定形のパターンで返答が可能なタスク発話と同様に取り扱うことは難しい．そのため，システムはまずユーザーの発話の意図が雑談かそれ以外なのかを判定することが重要である．

そこで本研究ではユーザーの発話の意図について雑談か否かを判定するモデルを構築する．これによりシステムは，発話に対してまず雑談かそうでないかを判定し，そこから判定に応じた応答を生成することが可能となる．特に近年ではニューラル対話モデル [1] などの台頭により，人とシステムとのより自然な雑談対話への期待が高まっているが，そうした技術を知的対話アシスタントに組み込むためには雑談意図の判定は欠かせないものとなる．

実験では，Yahoo! 音声アシスト [2] のユーザー発話ログを用いて教師ありモデルを構築し，実際の発話に対し雑談か否かを判定しその精度を確かめる．

[1]本研究はヤフー株式会社におけるインターンによる成果である．

2 ユーザーの雑談意図の判定

本節ではまずはじめに知的対話アシスタントについて概観し，設定したタスクの詳細を述べる．

2.1 知的対話アシスタント

知的対話アシスタントには Apple の Siri， Microsoft の Cortana，Yahoo! JAPAN の Yahoo! 音声アシストなどがある．いずれのシステムも，音声やテキストを用いてユーザーと対話を行い，ユーザーの目的となる動作を遂行するアシスタントのような存在である．近年では音声認識技術の向上やモバイル端末の普及に伴い，幅広いユーザーに使用されている．知的対話アシスタントの基本的な特徴として以下が挙げられる．

- モバイル端末での利用が主流
- ユーザーの音声発話に対して音声認識を行い，発話に沿った返答を行う
- システムによっては基本的なアシスト動作に加え，雑談などの動作も行う

知的対話アシスタントが有する機能はシステム毎に微細な違いがあるが，大抵のシステムは天気（ex. 明日の天気），位置情報（ex. 新宿駅まで），ウェブ検索（ex. 渋谷　ラーメン），端末操作（ex. アラームセット）などの機能が備わっている．

本研究での分析や評価には，知的対話アシスタントとして Yahoo! 音声アシスト [2] を用いる．

2.2 タスク設定

本項では，我々が取り組む発話の雑談意図判定タスクについての詳細を述べる．

まず，本タスクに置ける発話意図について定義する．知的対話アシスタントとユーザーのやり取りにおいて，ユーザーの発話は前項で述べたような機能のどれかに関する応答を求めていると仮

表 1: 各機能の発話例

機能	発話例
雑談	おはよう，今からお仕事です，名前を教えて
タスク	今日の天気，近くのラーメン屋，アラームセットして

定し，それを本タスクでの発話意図と定義する．我々は発話意図として，知的対話アシスタントに備わっている機能をおおまかに「雑談」と「タスク」の 2 つに分けた. 各機能に該当する発話の例を表 1 に示す．

雑談に該当する発話としては，日常的な挨拶や質問，自己開示など知的対話アシスタントとのコミュニケーションを試みているものが挙げられる．それ以外の天気の確認やウェブ検索，端末の操作などがタスクに該当する発話となる．

以上に基づき本研究ではユーザー発話の雑談意図の判定を，「ユーザーから発話があたえられた時，その発話意図が雑談か否かを判定する問題」と定義する．我々はユーザー発話が雑談か否かの 2 値分類に加え，雑談意図を持つ発話をそれらの発話行為に応じて細分化したうえでの多値分類も試みる．

3 データセット

本タスクの実験のため，Yahoo! 音声アシストにおいての全ユーザーの 8 ヶ月分の発話から，セッション先頭の発話のみを 30,000 件ランダムにサンプリングする．この時，単純にサンプリングを行うと「おはよう」等の発話データ内で高頻度な発話が集中するため，そうならないようにデータ内の発話を高，中，低頻度の 3 つのクラスタに分けそこから均等にサンプリングを行う．

教師ありモデルの構築のため, サンプリングされた発話に対しラベルを付与する必要がある．我々はクラウドソーシングにより発話データへのラベル付与をワーカーに依頼した．クラウドソーシングを依頼するにあたり，サンプリングされたデータに含まれる音声認識誤りや個人情報などの，提供するデータとしてふさわしくないものを事前に人手で可能な限り修正及び削除した．これにより，22,000 件の発話データが残った．残ったデータを用いて，Yahoo! クラウドソーシングに「提示された発話に対し，表に示した 4 つのラベルから発話の意図として尤もらしいものを選択する」というタスクを依頼した．この際，選択の多数決によってラベルを決定するため，1 つの発話につき 7 人のワーカーにタスクを依頼した．

表 2: ラベル集計結果

ラベル	発話数
雑談	4,752
音声検索	8,387
端末操作	1,958
わからない	60
計	15,157

表 3: 得票数の内訳

得票数	発話数
(6,7]	9,277
(5,6]	2,929
(4,5]	1,882
(3,4]	950
(2,3]	119
計	15,157

表 4: 雑談ラベルの細分類

ラベル	発話数	発話例
定形挨拶	280	おはよう，おつかれ，メリークリスマス
自己開示	1,432	今日は疲れました，テレビでびっくりしました
命令	846	歌をうたって，俺と結婚しろ，なんか話して
質問	1,402	あなたの母親は？，あなたは人工知能？
罵倒	165	死ね，バカ，アホ，役に立たんわ
意図不明	492	ニャーンニャーン，バルス，コケコッコー
フィラー	135	おお〜，えー，うん，ほおー
計	4,752	

クラウドソーシングの結果を集計し，多数決によりラベルを決定した結果を表 2 に示す．[2] また，得票数毎の頻度を集計したものを表 3 に示す．[3]

表 3 より，概ね 8 割程度の発話についてはワーカーの選択が一致していることがわかる．また，表 2 で雑談ラベルと判定されたものに対し，人手でラベルを細分化し分類した結果を表 4 に示す．

4 提案手法

我々は 2 つの教師ありモデルを用いて雑談意図の判定を行う．本節ではこれらのモデルで使用する素性とモデルの概要について述べる．

4.1 素性

モデルの素性として用いるものを箇条書きで示す．

- N-gram
  我々は多くの分類タスクで用いられる特徴量である単語 n-gram に加え，文字 n-gram も素性として用いる．文字 n-gram は知的対話アシスタントにおける短い発話をモデリングするのに有効だと考えられる．
- 単語分散表現
  近年分類タスクにおいてよく用いられるよう

[2] 同一発話については重複を排除している

[3] 投票の際ワーカーは複数のラベルを選択することが可能であるため，そのような投票に関しては選択数で割ったものを各ラベルの票数に加えている．

になり，良い性能を挙げているものとして単語分散表現がある．我々は Yahoo! 音声アシストにおける 2016 年 1 月から 8 月の全発話データから word2vec [3] によ り分散表現を学習したものを，発話中の各単語のベクトル表現として与える．

- GRU 言語モデルにおける発話確率
  ツイートおよびウェブ検索クエリログから GRU 言語モデルを学習し，それぞれの言語モデルを用いて求めた発話の生起確率の負の対数値を特徴量として用いる．このような特徴量を設計したのは，雑談 / タスクを意図した発話はツイート / 検索クエリと類似していて，生起確率が高くなると考えたためである．ツイートデータは，2016 年の 1 月から 12 月の間に収集したツイートの中から，別ユーザーと対話を行っているものだけを抽出した 1,000 万ツイートを用いた．一方，検索クエリログは，2016 年 4 月から 8 月に検索された 1 億クエリからランダムに抽出したクエリを用いた．

4.2 モデル

本タスクで用いる 2 つの教師あり学習手法について概要を述べる．

- 線形 SVM
  文書分類タスクで比較的良く用いられる線形 SVM を用いて学習，判定を行う．素性として発話の単語 n-gram，文字 n-gram，単語分散表現の平均を基本素性として用いる．それに加え，各 GRU 言語モデルにおいての発話確率を加えたものについても実験する．
- Convolutional Neural Network
  Kim ら [4] の，文書中の各単語のベクトル表現を入力とする Convolutional Neural Network (CNN) を用いて学習，分類を行う．入力単語のベクトル表現として， word2vec を用いて学習したモデルから得られる発話中の各単語のベクトル表現を用いて実験を行う．それに加え，ソフトマックス層に各 GRU 言語モデルにおいての発話確率を中間層の出力とともに入力するモデルについても実験する．

表 5: 2 値分類の実験結果

手法	Accuracy (%)
LM	67.8
SVM	90.9
SVM+LM	91.9
CNN	91.2
CNN+LM	91.6

表 6: 判定を誤った発話例

発話	ラベル	判定
人類滅亡まで後何日？	雑談	タスク
免許証がなくても運転できる？	タスク	雑談
にっこにっこにー	雑談	タスク
みなとくん	タスク	雑談

5 評価実験

3 節で作成したデータセットを用い，線形 SVM と CNN を用いたモデルで実際に判定を行った．ラベル付与された発話データから「わからない」ラベルの発話を除いた 15,097 発話のうち，「雑談」ラベルを雑談意図，「音声検索」，「端末操作」ラベルをタスク意図とし 2 値分類を行った．評価は 10 分割交差検定を用いて行い，両手法ともに開発データでパラメータチューニングを行った．

ベースラインとして，4.1 節で説明した 2 つの言語モデルにおいての発話の確率の大小で 2 値分類を行う手法との比較を行う．提案手法の実装として線形 SVM は liblinear [5] ， CNN は Chainer [6] を用いた．また，各 GRU 言語モデルの構築に faster-RNNLM [7] を用いた．

各手法による 2 値分類の結果は表 5 に示すとおりとなっている．LM は 2 つのベースラインの手法で，SVM ，CNN は 4.1 節の GRU 言語モデルの発話確率を用いない手法で，SVM+LM ，CNN+LM は 発話確率を用いた手法である．3 つの手法はいずれもベースラインの性能を大きく上回っていることが確認できる．また，いずれの手法も言語モデルの特徴量を加えることで精度が向上していることが確認できる．

判定を誤った発話についていくつか例を表 6 に示した．上半分の例は内容ははっきりしているがどちらの意図か曖昧な発話で，人間でもどのラベルとするか迷うような発話である．このような例を正確に判定するのは難しいと考えられる．下半分については内容も意図もはっきりとしない発話である．このような短い発話は，ユーザーが知的対話アシスタントの反応を伺うために発せられていることが多く，特定のドメインに関する知識等

表 7: CNN+LM による雑談ラベル発話の多値分類の実験結果 (%)

	定形挨拶	自己開示	命令	質問	罵倒	意図不明	フィラー
Precision	0.81	0.71	0.74	0.74	0.69	0.32	0.58
Recall	0.64	0.78	0.74	0.72	0.50	0.27	0.50
F	0.71	0.74	0.74	0.73	0.56	0.29	0.52

が必要であるため難しい．

また，2 値分類器で雑談と判定されたデータに対し，表 4 のデータのみを学習した CNN+LM を用いての多値分類を行った結果を表 7 に示した．「罵倒」,「その他」,「フィラー」ラベルについてはデータ数が少ないため参考結果であるが，それ以外のラベルのものは概ね 70% 程度の精度で検出ができていることがわかる．

6 関連研究

知的対話アシスタントは Web 検索や端末の操作，雑談など様々な機能を備えた対話システムであり，これまでの研究で対象とされてきた音声対話システムとは性質が異なる．

本研究に最も近いのはタスク型の対話システムで行われている意図判定の研究 [8] である．それと比べた本研究の独自な点は，オープンドメインな雑談発話を扱っていることや，それに伴いツイッターやクエリを用いた言語モデルが有効に働くことを示したことである．

知的対話アシスタントに関する研究は，2014 年に評価指標の提案とユーザー満足度を行った Jiang [9] の研究以降，徐々に数を増やしつつある．[9] は発話ログからユーザーがどの程度利用に満足したかを予測するモデルを構築した．これに対し，Sano [10] はユーザーが今後利用を継続するかしないかの予測と，どの位の期間利用するかのエンゲージメント予測を行った．いずれの研究も知的対話アシスタントのユーザーとの一連のやり取りにおけるユーザーの満足度や利用頻度の予測をすることに注力しており，アシスタントの応答の質やユーザーの発話の分析といった個々の機能への改善につながる部分には未着手である．

7 おわりに

本研究では，知的対話アシスタントにおいてのユーザーの雑談意図を判定するモデルを構築し，高精度に雑談意図を判定できることがわかった．

今後は，本質的に曖昧な発話に対し，どのようにして意図を推定するかが課題である．例えば，「今日は曇っているね」のように，雑談意図なのかタスク意図（天気予報を知りたがっているのか）なのかが本質的に曖昧な曖昧な発話に対し，「今日の天気予報をお調べしましょうか？」という風に適当な聞き返しを行うことで，意図を推測していくような手法を検討している．

参考文献

[1] Orioi Vinyals and Quoc V. Le. A Neural Conversational Model. *Proceedings of the Workshop on Deep Learning in ICML*, Vol. 37, , 2015.

[2] Yahoo! JAPAN. Yahoo!音声アシスト. http://v-assist.yahoo.co.jp/., 2016.

[3] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed Representations of Words and Phrases and their Compositionality. *Proceedings of NIPS*, pp. 1–9, 2013.

[4] Yoon Kim. Convolutional Neural Networks for Sentence Classification. *Proceedings of EMNLP*, pp. 1746–1751, 2014.

[5] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. LIBLINEAR: A Library for Large Linear Classification. *Journal of Machine Learning Research*, Vol. 9, No. 2008, pp. 1871–1874, 2008.

[6] Seiya Tokui, Kenta Oono, Shohei Hido, and Justin Clayton. Chainer: a Next-Generation Open Source Framework for Deep Learning. *Proceedings of the Workshop on Machine Learning Systems in NIPS*, pp. 1–6, 2015.

[7] Anton Bakhtin and Ilya Edrenkin. Faster RNNLM. https://github.com/yandex/faster-rnnlm., 2015.

[8] D Michael, A Mladen, C James, North Carolina, and North Carolina. Dialogue Act Modeling in a Complex Task-Oriented Domain. *Proceedings of SIGDIAL*, pp. 297–305, 2010.

[9] Jiepu Jiang, Ahmed Hassan Awadallah, Rosie Jones, Umut Ozertem, Imed Zitouni, Ranjitha Gurunath Kulkarni, and Omar Zia Khan. Automatic Online Evaluation of Intelligent Assistants. *Proceedings of WWW*, 2015.

[10] Shumpei Sano, Nobuhiro Kaji, and Manabu Sassano. Prediction of Prospective User Engagement with Intelligent Assistants. *Proceedings of ACL*, pp. 1203–1212, 2016.