Vol.2017-NL-231 No.18
Vol.2017-SLP-116 No.18
2017/5/16

知的対話アシスタントにおける雑談を目的としたユーザ発話の検出

赤崎　智[†1,a)]　鍜治　伸裕[†2,b)]

概要：ユーザとの対話により様々な仕事や雑談などを行う知的対話アシスタントが注目されている．知的対話アシスタントはタスク指向型対話システムと非タスク指向型対話システムが混ざった，多種多様なドメインでの対話が行われる全く新しい対話システムである．本稿ではそのような対話システムの実現のため，ユーザ発話が雑談を行おうとしているか否かを判定するという新しいタスクを提案する．我々は商用の知的対話アシスタントの一つである Yahoo!音声アシストのデータを対象に，n-gram 等の基本的な特徴量に加え，多様なユーザ発話を扱うためにツイッターとウェブ検索エンジンのクエリを用いた外部特徴量を設計した．実験では 2 値分類器に上記の特徴量を与え，実際にユーザ発話が雑談であるか否かの判定を行った．結果，基本的な特徴量だけを使うと F 値が 86.21%だったものが，外部特徴量を加えることで 87.53%まで改善できることを確認した．

1. はじめに

1.1 発話の雑談意図判定

従来の音声対話システムの研究の大半は，タスク指向型対話システム [1] または非タスク指向型対話システム [2] のどちらかを対象としてきた．前者は特定のタスクを遂行しユーザを支援する対話システムで，後者は雑談などを行いユーザを楽しませることを目的とした対話システムである．タスク指向型，非タスク指向型それぞれの対話システムに備わっている機能は明らかに相互補完の関係であるが，2 つのシステムの組み合わせを考慮した研究はこれまでほとんど行われていない．

この状況は知的対話アシスタントと呼ばれる音声対話システムの出現により変わりつつある．知的対話アシスタントには代表的なものにスマートフォンで動作する Siri[*1]や Alexa[*2]があり，いずれもユーザとの雑談のほか，ウェブ検索，天気の確認，アラームの設定といった様々なタスクも行う．これらの知的対話アシスタントは，タスク指向型，非タスク指向型対話システムの 2 つの要素を組み合わせた全く新しい対話システムである．

タスク指向型，非タスク指向型の両発話が混ざった音声対話システムにおいては，ユーザの発話が与えられた時にそれがシステムと雑談を行おうとしているか否かを判定する必要がある．例えば，ユーザがシステムに「あなたの趣味はなんですか？」と尋ねたなら，恐らくユーザはシステムと雑談を行おうとしている．一方で，「8 時にアラームをセットしてください」と尋ねたなら，ユーザは端末の操作をシステムに依頼している．我々は前者のような発話を雑談意図発話と定義し，2 値分類器によりこれらの検出を行うタスクを提案する（以降，このタスクを**発話の雑談意図判定**とする）．

発話の雑談意図判定は知的対話アシスタントのようなタスク指向型，非タスク指向型の両方を組み合わせた対話システムの実現には重要な要素となるが，既存の研究では十分に検討されていない．タスク指向型または非タスク指向型の音声対話システムのどちらか一方に焦点を当てた研究はこれまでにも多くなされているが，いずれも発話の雑談意図判定は必要ない．これはタスク指向型対話システムのユーザはシステムと雑談をせず，逆に非タスク指向型対話システムのユーザは常にシステムと雑談をするからである．

1.2 意図判定，ドメイン判定，Slot-filling との関連

発話の雑談意図判定は，既存の音声対話システムの分野で研究されてきた発話のドメイン判定や意図判定，対話管理における Slot filling などと関係はあるがその目的は大き

[†1] 東京大学
The University of Tokyo
本研究はヤフー株式会社におけるインターンの成果である．
[†2] ヤフー株式会社
Yahoo Japan Corporation
[a)] akasaki@tkl.iis.u-tokyo.ac.jp
[b)] nkaji@yahoo-corp.jp
[*1] http://www.apple.com/ios/siri
[*2] https://developer.amazon.com/alexa

く異なる [3–7].

意図判定は単一ドメインからなるタスク指向型対話において発話の意図を推定するタスクである [6]．例えば，航空機の予約サービスの対話記録から構築された ATIS データセットでは，飛行便や都市の情報などが発話の意図にあたる [8]．対して，ドメイン判定は複数ドメインからなるタスク指向型対話においてユーザの発話がどのドメインと関連するかを判定するタスクである [4]．Slot filling はタスク指向型対話においてユーザの発話から場所や対象といった重要な情報を抽出し，対話管理のためのスロットに格納していくタスクである [3,7].

雑談意図判定はこれらと異なり，オープンドメインな非タスク指向型対話と複数ドメインからなるタスク指向型対話が混ざったシステムを対象とし，ユーザの発話に対し非タスク指向型の要素があるか，すなわちユーザがシステムと雑談を行おうとしているか否かを判定する．したがって，発話の雑談意図判定は意図判定やドメイン判定とは目的が大きく異なる．

我々は雑談意図判定を 2 値分類問題と定義した．このとき，雑談意図と非雑談意図の他にもより細かいドメインを設定し，多値分類問題として定義することも可能である．しかし一般に対話システムのドメインは逐次的に拡張されていくため，多値分類の際はドメイン毎に学習データを継続的に整理していく必要がある．このような作業はコストがかかるため，我々は雑談意図判定を発話が雑談意図かそうでないかの単純な 2 値分類問題と定義した．2 値分類問題と定義することによって，タスク意図におけるドメイン追加や削除といったシステムの仕様変更に影響されにくくなる．

発話の雑談意図を判定するときの技術的な課題は，タスク指向型，非タスク指向型の両方の要素をもつシステムにおいての発話のオープンドメインな性質，いわゆる多様性への対応である．このような多様性の高い発話は雑談意図やタスク意図についてのあらゆるトピックが含まれるため，発話のラベル付きデータを人手で大量に用意するのに膨大な労力が必要となる．これは特定ドメインの発話にのみ注視すれば良いタスク指向型対話や非タスク指向型対話とは状況が異なる．

Slot-filling との関連であるが，タスク指向型対話における意図判定やドメイン判定と異なり，雑談意図判定の後に Slot-filling は通常行われない．これは，雑談意図発話に対する応答の生成手法として sequence to sequence モデル [9] や，情報検索ベースのアプローチ [10] などの手法が用いられることが多いからである．意図判定やドメイン判定と同時に Slot-filling を行い判定精度を向上させる研究 [3,7] も行われてきたが，雑談意図発話にはスロットが定義されないため，同様の手法を雑談意図判定に適用することは出来ない．

1.3 本研究の要旨

本研究では，タスク指向型，非タスク指向型の両方の要素をもつシステムにおいて発話の雑談意図を判定するというタスクを提案する．我々は商業用の知的対話アシスタントのログデータから 15,160 件のユーザ発話を収集し，クラウドソーシングを用いワーカーにそれらの発話が雑談意図か否かのラベル付けを依頼した．その後，クラウドソーシングによって得られたラベル付きデータを用いて教師ありの 2 値分類器を構築し，発話の雑談意図の判定を行った．この際，発話の多様性に対応するためにラベルなしの外部資源を用いての実験も行った．具体的には Twitter の投稿とウェブ検索エンジンのクエリを外部資源として用い，それらを特徴量として加えることでどの程度判定精度が向上するかを調査した．実験の結果，基本的な特徴量を使った教師ありの 2 値分類手法でも高い精度で判定ができるが，そこに外部資源を加えることで更に精度が向上することを確認した．外部資源を加えなかった場合と加えた場合では，F 値に 1%以上の差があることが確認できた (86.21%から 87.53%に向上).

本稿では 2 節では本研究の関連研究について述べ，3 節と 4 節ではそれぞれデータセットと発話意図判定の手法について述べる．5 節では実験結果について述べ，6 節と 7 節でそれぞれ今後の予定と本研究のまとめについて述べる．

2. 関連研究

タスク指向型，非タスク指向型対話システムに関する研究は古くから続いているが，これら 2 つの対話システムを組み合わせたものに関する研究はほとんどない．よって，タスク指向型，非タスク指向型の 2 つを組み合わせた対話システムにおける発話の雑談意図判定もこれまでの研究では検討されていない[*3]．Lee ら [11] は用例ベースのタスク指向型，非タスク指向型対話システムを組み合わせるための対話管理モジュールを提案している．しかし，彼らの手法は用例ベースの対話システムを前提としているため，対話システムのモジュールとして統計的な手法を取り入れるのは難しい．また，雑談意図判定のための外部資源についても検討されていない．Yu ら [12] はタスク指向型対話システムにおいて，ユーザの発話に対してタスク応答で対応できない場合に，強化学習の枠組みで適当な非タスク応答をすることによって，タスクの達成率とユーザーエンゲージメントを向上させる手法を提案した．彼らの手法ではユーザーの発話に対してシステムが自発的に雑談意図応答を行うが，本研究のように発話が雑談意図か否かの判定は行っていない．

[*3] 既存の Siri などの商業用知的対話アシスタントは技術的な部分が隠匿されているため，これらと本研究との比較はできないが，本稿では実験で我々の内製の知的対話アシスタントを提案手法と比較し，有効性を確認する．

知的対話アシスタントに関する既存の研究は，知的対話アシスタントに対するユーザ満足度やエンゲージメントの予測などを行っている [13–16]．Jiang ら [13] はユーザが知的対話アシスタントの応答に対する満足度を，ユーザ発話の他にクリックやスワイプ等の様々なデバイス情報を用いて予測した．Sano ら [14] は，ユーザが知的対話アシスタントを継続利用するかの予測をユーザの過去の使用履歴などを用いて予測した．

いくつかの先行研究では，知的対話アシスタントの 1 つである Cortana の発話データセットをドメイン判定のベンチマークとして使用した [3,4,6]．これらの研究はいずれも知的対話アシスタントを単なる複数ドメインのタスク指向型対話システムとみなしており，雑談意図の発話については考慮していない．

対話システムにおける統計的な応答生成手法に関する研究は近年急激に発達している [17,18]．はじめは Ritter ら [17] がフレーズベース機械翻訳の手法を応答生成に適用し，その後 Vinyal と Le [18] が Sutskever ら [9] の sequence to sequence モデルを応答生成に用いて以降，ニューラルネットワークベースの応答生成手法が多く研究されている．これまで，生成される応答の質を様々な観点から向上させる試みはいくつも存在したが [10,19–24]，いずれもユーザは常にシステムと雑談をすると仮定し，それを踏まえ応答を生成している．よって，これらの応答生成手法をタスク指向型と非タスク指向型の要素が混ざった対話システムに適用するには，発話の雑談意図判定が必要である．

対話のモデリングに Twitter 投稿を用いた研究もいくつか存在する [25,26]．これらの研究は Twitter 投稿を応答生成の訓練データとして用いたり，教師なしの発話行為推定へ用いたりしている．この種の外部資源の扱い方は本研究における外部資源の利用法とは異なる．

3. 雑談意図判定のためのデータセット

本節では雑談意図判定に用いるデータセットの構築について説明する．その後，構築したデータセットの内容について分析する．

3.1 データセットの構築手順

我々は，Yahoo! 音声アシスト[*4]という商業用の知的対話アシスタントのログデータ[*5]から内容の重複なしで 15,160 件の発話をランダムサンプリングした．サンプリングの際はデータセットに高頻度で出現する発話と低頻度で出現する発話の両方をバランスよく含めるため，ログデータの発話を高，中，低頻度の 3 つのグループに分割し，それぞれのグループから均等にサンプリングを行った．その後，得られた発話からプライバシーに関わる内容や人名，住所といった情報を人手により排除した．

次に，得られた 15,160 件の発話へのラベル付けを Yahoo! クラウドソーシング[*6]へと依頼した．我々は，ワーカーに発話を提示し，その発話がシステムと雑談を行おうとしているものなら雑談意図ラベル，情報の検索（ウェブ検索や天気，路線情報などの確認）やスマートフォン端末の操作（アラームの設定や音量調整）を行おうとしているなら非雑談意図ラベルをつけるように指示した[*7]．この際，それぞれの発話につき 7 人のワーカーに上記タスクを依頼し，得られた票の多数決によりラベルを決定した．これにより得られたラベルと発話の例を表 1 に示す．票数の列が示す数字は，発話に対し多数派の票，つまり決定されたラベルに投票された数である．例えば，はじめの行の発話「なにかお話しよう」は 5 人のワーカーが雑談意図ラベルへ投票し，2 人のワーカーが非雑談意図ラベルへ投票している．

3.2 データ分析

構築されたデータセットは 4,833 件の雑談意図ラベルの発話，10,327 件の非雑談意図ラベルの発話からなる．

ここで我々は，ワーカー間の発話に対する投票の一致率について調査した．表 1 が，多数派の票数別の発話数であり，7 人のワーカー全てが一致した発話は 5,811 件 (38%)，6 人のワーカーが一致した発話は 4,978 件 (32%) で，計 10,789 件 (71%) の発話についてはアノテータの投票が一致していることになる．よって，付与されたラベルの信頼性は比較的高いと言える．

投票が一致していない発話について分析すると，意図が曖昧な発話が多数確認された．例えば，「お腹が空きました」はユーザがシステムと雑談を試みている発話とも解釈できるが，近所の飲食店を探している発話とも解釈できる．「肌荒れケアはどうする？」や「腰が痛いです」なども同様の例である．このような暗黙的な情報欲求と解釈されうる発話に対しては，ユーザに適度に聞き返しや発話意図の詳細化をする応答をすることが必要となる [27]．このような処理の実現は今後の課題とする．

我々は雑談意図ラベルが付与された発話に対し，ユーザがシステムとどういう雑談をしているのか調査をするための詳細な発話行為ラベルを人手で付与した．表 2 に付与した発話行為と発話を示す．ここでの発話行為は目黒ら [28] の研究を参考にし設計した．表 2 より，いくつかの発話は挨拶（例:「おはよう」）やお礼（例:「ありがとう」）といった典型的な表現で，これらについてはあまり多様性が観察

*4 https://v-assist.yahoo.co.jp/
*5 このログデータは 2016 年の 1 月から 8 月にかけて実際にユーザが発話し，それを音声認識することによって得られたものである．
*6 https://crowdsourcing.yahoo.co.jp/
*7 今回用いた知的対話アシスタントはスマートフォン上で動作するため，非雑談意図ラベルの発話はアラームの設定等のスマートフォン端末の操作を指示する発話が多く存在することに注意されたい．

表 1　発話の例と多数派の票数．非雑談意図ラベルは音声検索（上段）と端末操作（下段）の 2 つに著者が分割した．

ラベル	発話	票数
雑談意図	お話しよう	5
	趣味はなんですか	7
	今月は休みがありません	5
	散歩しにいきます	6
	猫は好き?	5
	あなたけっこうオタクだね	7
非雑談意図	富士山の写真を見せて	6
	世界で一番高い建物はなんですか？	5
	近くのおいしいレストラン	7
	9:10 に起こして	7
	画面を明るくして	6
	アラーム解除	7

できなかった．一方で，残りの発話については多様性が高く，ユーザの個人的な出来事（例:「今日は私の結婚記念日なの」）や，システムへの質問（例:「怒ってますか？」）等の幅広いトピックが確認できた．他にも，「フォースとともにあれ」などの有名な映画の台詞の引用や「コケコッコー」などの動物の鳴き声を真似たものも多数確認でき，それらには「その他」のラベルを付与した．

また，システムが正しい応答をすることができなかったためにユーザがシステムを罵倒したりしている発話も存在しているのは興味深い（発話行為の罵倒，怒り，呆れの行を参照）．このような発話は，既存の対話システムに関する研究で構築されたような，あらかじめ被験者に指示を与え会話をさせることにより構築したデータセットでは確認できない．すなわち，本研究で作成したデータセットは，対話システムにおけるユーザの実際の振る舞いをより反映しているデータセットであるともいえる．以上の分析より，知的対話アシスタントにおける発話は多様性が高いことが伺える．

4. 判定手法

我々は発話の雑談意図判定を 2 値分類問題と定義し，教師ありの分類器を構築し判定を行う．本節でははじめに，本研究で用いる 2 種類の分類器と基本特徴量について説明し，次にこれらの分類器の性能を向上させるための外部資源について説明する．

4.1 分類器

一つ目の分類器として Support Vector Machine (SVM) を用いる．基本特徴量としては，発話の文字と単語両方の 1-gram，2-gram と単語埋め込み表現 [29] を用いる．単語埋め込み表現は Mikolov ら [30] の Skip-gram を Yahoo! 音声アシストの発話ログから学習し，以下のように発話の t 番目の単語の埋め込み表現 $\mathbf{x}_t$ とし，単語数 n で平均したものを素性ベクトルの追加の要素として与える．

表 2　発話行為の分布と発話例．

発話行為 (#発話数)	発話例
挨拶 (206)	こんにちは
	メリークリスマス
自己開示 (1,164)	私はしし座です
	のどが痛いです
命令 (716)	励まして
	歌をうたって
質問 (1,551)	あなたは感情ある?
	怒ってる?
勧誘 (130)	一緒に遊ぼう
	今度カラオケいこうね
情報提供 (214)	猫の行動が変です
	雪降ってるよ
感謝，承認 (126)	ありがとう
	あなたクールだね！
罵倒，怒り，呆れ (172)	あんたバカだね
	お前は役に立たん
謝罪 (9)	ごめんなさい
	間違えましたごめんね
投詞，フィラー (151)	うおー
	ハハハ
その他 (394)	フォースとともにあれ
	コケコッコー

$$\frac{1}{n}\sum_{t=1}^{n}\mathbf{x}_t$$

二つ目の分類器は，近年文書分類タスクで良い性能を挙げている畳み込みニューラルネットワーク (CNN) を用いる [31–33]．我々は Kim ら [31] の単一の畳み込み層，max-pooling 層，ソフトマックス層からなるモデルを使用する．活性化関数としては rectified linear unit (ReLU) を用い，発話の初期単語ベクトル表現として前述の単語埋め込み表現を用いる．

4.2 外部資源の利用

ここでは雑談意図判定のための外部資源を用いた追加の特徴量について述べる．近年の急速なウェブの発展のため，雑談用途と非雑談用途の様々な種類のテキストデータが利用できるようになっており，これらのデータは今回の雑談意図判定における分類器の精度を改善するのに効果的であると考えられる．我々は雑談用途として Twitter の投稿データ（ツイート），非雑談用途としてウェブ検索のクエリデータ（クエリ）をそれぞれ用いる．

我々はツイートとクエリのテキストデータを用いて文字ベースの言語モデルをそれぞれ訓練し，各言語モデルでの発話スコア（発話の各言語モデルでの確率を文字数で正規化したもの）を 2 つの追加の特徴量として与える[*8]．$u = c_1, c_2, \ldots, c_m$ を m 個の文字からなる発話だとすると，

[*8] 単語ベースの言語モデルについても実験したが，文字ベースの言語モデルのほうが性能が良かった．

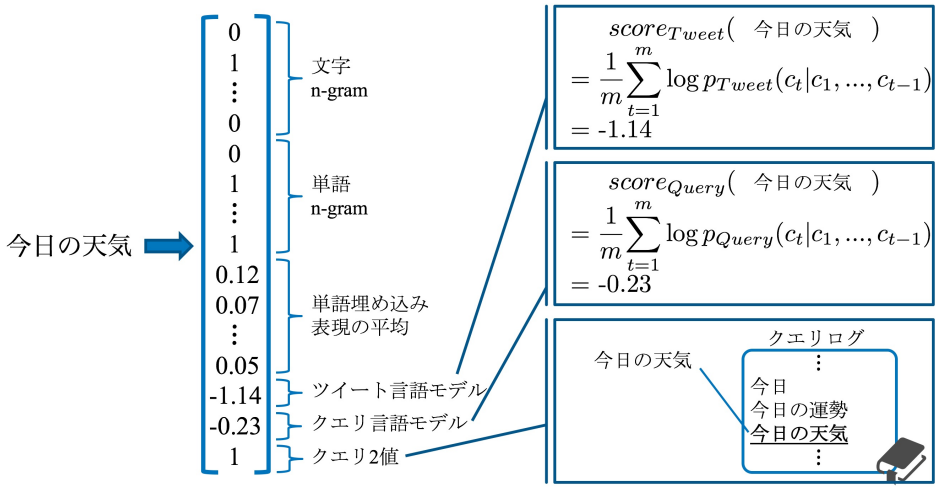

図 1 発話「今日の天気」に対し分類器として SVM を使用するときの素性ベクトル.

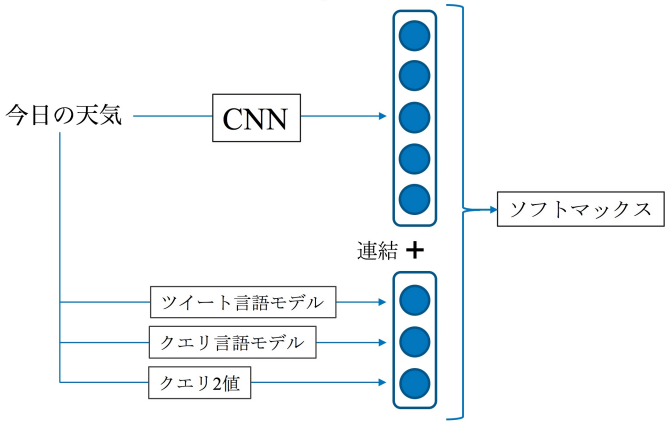

図 2 発話「今日の天気」に対し分類器として CNN を使用するときのモデル概略図.

外部資源 $r \in \{tweet, query\}$ で学習した言語モデルにおいての発話 u のスコア $score_r(u)$ は以下のようになる.

$$score_r(u) = \frac{1}{m}\sum_{t=1}^{m} \log p_{type}(c_t \mid c_1, \ldots, c_{t-1}).$$

言語モデルとしては文書モデリングにおいて良い性能を挙げている GRU 言語モデルを用いる [34, 35]. $\mathbf{x}_t$ を t 番目の文字の埋め込み表現とし, $\mathbf{h}_t$ を t 番目の文字を読み込んだ時の隠れ層だとすると, GRU は隠れ層を以下のように計算する.

$$\mathbf{h}_t = (1 - \mathbf{z}_t) \odot \mathbf{h}_{t-1} + \mathbf{z}_t \odot \tilde{\mathbf{h}}_t$$
$$\mathbf{z}_t = \sigma(\mathbf{W}^{(z)}\mathbf{z}_t + \mathbf{U}^{(z)}\mathbf{h}_{t-1})$$
$$\tilde{\mathbf{h}}_t = \tanh(\mathbf{W}^{(h)}\mathbf{x}_t + \mathbf{U}^{(h)}(\mathbf{r}_t \odot \mathbf{h}_{t-1}))$$
$$\mathbf{r}_t = \sigma(\mathbf{W}^{(r)}\mathbf{x}_t + \mathbf{U}^{(r)}\mathbf{h}_{t-1})$$

$\odot$ は要素積であり, σ はシグモイド関数, tanh はハイパボリックタンジェント関数である. $\mathbf{W}^{(z)}$, $\mathbf{U}^{(z)}$, $\mathbf{W}^{(h)}$, $\mathbf{U}^{(h)}$, $\mathbf{W}^{(r)}$, $\mathbf{U}^{(r)}$ は各ゲートの重み行列である. 隠れ層は次の文字を予測するためにソフトマックス層に入力される.

これに加え, 我々は発話がウェブ検索のクエリデータに出現するか否かという 2 値の特徴量も用いる. 3 節のデータセット構築の際, いくつかの非雑談意図ラベルの発話は場所名や商品名などの単一のエンティティやその組み合わせ (「iPhone7」,「東京　観光地」など) で構成されることを確認した. このような発話はこれらのエンティティについての情報を求めている, すなわち非雑談意図の発話とみなせるため, 我々はこれと対応するクエリデータを辞書として用い, 発話そのものがこの辞書に含まれているか否かという 2 値の特徴量を分類器に与える.

以上の 3 つの特徴量を分類器の追加特徴量として与える. 各分類器において以上で説明した特徴量の入力方法について, 発話「今日の天気」を与えた時の概略図を図 1 と図 2 に示す. SVM では 4.1 節で述べた基本特徴量と追加特徴量を併せ, 入力の素性ベクトルとして用いる (図 1). CNN においては, 追加特徴量を全結合層の出力と連結することでソフトマックス層への入力として与える (図 2).

5. 評価実験

我々は 3 節で構築したデータセットを用いて, 実際に雑談意図判定を行い提案手法を評価した. 最初に実験設定について述べ, その後実験結果について説明する.

5.1 実験設定

我々は 3 節で得られた 15,160 件の発話データを 10 分割交差検証により評価した. 分割したそれぞれのセットの 80%を訓練データ, 10%を開発データ, 10%を評価データとしてそれぞれ用いた.

4.1 節における特徴量の単語埋め込み表現の学習には word2vec[*9] を用い, Skip-gram で 300 次元の単語埋め込み表現を学習した. SVM には素性ベクトルの 300 次元分の追加要素として与えられ, CNN には発話の初期単語ベク

[*9] https://code.google.com/archive/p/word2vec

トル表現として与えられる．

4.2 節で述べた GRU 言語モデルの学習には faster-rnn ツールキット[*10]を用いる．単語埋め込み層と隠れ層の次元は 256 に設定し，ソフトマックス層の学習にノイズサンプル数 50 の Noise contrastive estimation [36] を使用した．また，精度向上のため 4-gram の最大エントロピー言語モデルを GRU 言語モデルと同時に訓練し，2 つのモデルの出力を組み合わせた [37]．

各言語モデルはツイートについては 2016 年 4 月から 7 月の期間で収集した 1 億件のツイートとリプライの対 [17] で，クエリについては 2016 年 3 月から 6 月の間に発行された 1 億件のウェブ検索クエリで学習を行った．なお，クエリについては上記と同様のデータを 4.2 節で述べたようにエンティティ辞書として用い 2 値の特徴量を与えた．

分類器の SVM については liblinear[*11]を用い，L2-regularized L2-loss SVM を学習する．C パラメータについては開発データにおいて $C = \{2^{-10}, 2^{-9}, \ldots, 2^{10}\}$ の刻み幅で雑談意図ラベルの F 値を最大化するものを用いる．

CNN については Chainer[*12]を用いて実装した．パラメータについては，開発データにおいて各フィルタごとの個数が $\{100, 150\}$，フィルタ幅が $\{\{2\}, \{3\}, \{1,2\}, \{2,3\}, \{3,4\}, \{1,2,3\}, \{2,3,4\}\}$ の組み合わせで雑談意図ラベルの F 値を最大化するものを用いる．またミニバッチ学習のバッチサイズを 32 に設定し，層数 3 の全結合層にレート 0.5 の Dropout を適用する．最適化のため Adam ($\alpha = 0.001, \beta 1 = 0.9, \beta 2 = 0.999, \epsilon = 10^{-8}$) を用いて確率的勾配降下法を行う [38]．

5.2 ベースライン手法

以下のようなベースライン手法を比較のために用いる．

マジョリティ　全ての発話に対し，構築されたデータセット中で最も数が多い非雑談意図ラベルと判定する．

ツイート言語モデル　5.1 節で構築したツイートの GRU 言語モデルにおいての発話スコアが特定の閾値を越えた時に雑談意図ラベルと判定する．閾値は開発データにおいて雑談ラベルの F 値を最大化するものを用いる．

内製の意図推定システム　Yahoo! 音声アシストの内製の発話意図推定システムを用いる．このシステムはルールベースと用例ベースの手法を組み合わせたものである．このシステムの技術的な詳細は公開できないため，参考結果として載せる．

5.3 実験結果

表 3 に雑談意図ラベルについての適合率，再現率，F 値と全体の分類精度を示した．ベースラインのマジョリティについては分類精度のみを示し，＋ **単語埋め込み表現**は 5.1 節で述べた SVM においての単語埋め込み特徴量，＋ **事前学習**は CNN においての初期単語ベクトル表現である．＋ **外部特徴量**は 5.1 節で述べたツイートとクエリから得られた特徴量を分類器に追加したものである．最も性能の良かったものは **SVM ＋ 単語埋め込み表現 ＋ 外部特徴量**で，92%の分類精度と 87%の F 値を達成しベースラインの手法を上回った．CNN については SVM の結果より数値が低くなり，Kim ら [31] の報告と異なる結果となった．これは今回用いた CNN のモデルが比較的単純なものであるからと考えられ，より複雑な CNN ベースの分類器を用いることで性能が改善すると思われる[*13]．また，表 3 は分類器の追加特徴量として用いた外部資源が有効であることも示しており，SVM と CNN の両方で F 値が 1%以上向上している．

表 4 に発話の例とその発話の各言語モデルにおけるスコアを示す．これより，ツイート及びクエリの言語モデルが雑談意図ラベルの発話に対しそれぞれ高いスコアと低いスコアを与えていることが確認できる．

表 5 は外部資源を用いた 3 つの特徴量のそれぞれを **SVM ＋ 単語埋め込み表現**に与えたときの結果である．3 つとも **SVM ＋ 単語埋め込み表現 ＋ 外部特徴量**より悪い結果となっているため，良い性能を達成するには 3 つの特徴量を全て組み合わせて用いることが重要であることがわかる．

表 6 は **SVM ＋ 単語埋め込み表現 ＋ 外部特徴量**の特徴量の正負のそれぞれの重みの上位 15 件を抜き出したものである．正と負でもっとも重みがかかっているものとしてツイート言語モデル，クエリ言語モデルの特徴量があるため，言語モデル特徴量が有効であることを示している．他には，単語埋め込み表現や雑談意図，非雑談意図で特徴的な文字や単語（例: 雑談意図なら発話のモダリティを表すような文末表現，非雑談意図なら電話の「電」や時間の「時」などの非雑談意図発話に特徴的な文字）が確認できる．

表 7 は **SVM ＋ 単語埋め込み表現 ＋ 外部特徴量**を用いた時のワーカーの投票数別（多数決での多数票別）の結果である．これからわかるように，ワーカーの投票の一致度合いが高くなればなるほど結果の数値も高くなっている．7 人全員が一致した発話については 98%の分類精度を達成しており，人間が容易に判定できるものは分類器による判定も容易であることがわかる．逆に 4 人しか一致していないような，すなわち投票が割れているものについては判断が難しい発話や 3.2 節で述べたような意図が曖昧な発話が多く，分類器による判定結果も著しく悪化している．

[*10] https://github.com/yandex/faster-rnnlm
[*11] https://www.csie.ntu.edu.tw/cjlin/liblinear
[*12] http://chainer.org
[*13] 複雑になればなるほどモデルの訓練に時間がかかる．

表 3　各モデルにおける判定結果．各指標で最も数値が高いものを太字で示している．

モデル	分類精度	適合率	再現率	F 値
マジョリティ	68.12	N/A	N/A	N/A
ツイート言語モデル	72.07	54.54	74.40	62.94
内製の意図推定システム	78.31	62.57	79.51	70.03
SVM	90.51	86.42	83.45	84.91
SVM + 単語埋め込み表現	91.35	87.62	84.88	86.21
SVM + 単語埋め込み表現 + 外部特徴量	**92.15**	**88.61**	**86.50**	**87.53**
CNN	85.16	83.40	68.12	74.41
CNN + 事前学習	90.84	87.03	83.80	85.36
CNN + 事前学習 + 外部特徴量	91.48	87.78	85.18	86.56

表 4　言語モデルスコアと発話例．最初の 2 列はツイートとクエリの各言語モデルが発話に対し与えたスコアで，3 列目と 4 列目はそれぞれラベルと発話例．

スコア（ツイート/クエリ）		ラベル	発話
−1.157	−2.137	雑談意図	もっと感情込めて喋って
−1.760	−2.521	雑談意図	君頭悪いね
−0.627	−1.166	雑談意図	はいおやすみなさい
−1.139	−0.682	非雑談意図	facebook を開く
−2.469	−1.090	非雑談意図	自宅に電話
−1.796	−0.422	非雑談意図	ポケモン GO 遊び方

表 5　各外部特徴量の効果．各指標で最も数値が高いものを太字で示している．

特徴量	分類精度	適合率	再現率	F 値
ツイート言語モデル	**91.53**	**87.62**	**85.49**	**86.53**
クエリ言語モデル	91.38	87.55	85.06	86.28
クエリ 2 値	91.42	87.56	85.21	86.36

表 6　SVM + 単語埋め込み表現 + 外部特徴量の特徴量の重み．左列が正値の絶対値上位 15 件で右列が負値の絶対値上位 15 件．$e(m)$ は単語埋め込み表現の m 次元目の要素を示しており，$\{w_1, ..., w_n\}$ は n-gram を示している．

特徴量	重み	特徴量	重み
ツイート言語モデル	1.127	クエリ言語モデル	-0.770
{ た,</s>}	0.265	{ 電 }	-0.323
{ の,</s>}	0.254	$e(208)$	-0.278
$e(266)$	0.250	{ 時,?}	-0.258
{ よ,</s>}	0.246	$e(287)$	-0.253
{ う,</s>}	0.243	$e(26)$	-0.250
$e(29)$	0.243	{ 画 }	-0.240
{ 何, の }	0.237	{ 雨, は }	-0.240
$e(28)$	0.233	$e(277)$	-0.235
$e(121)$	0.231	{ ん, に }	-0.235
$e(154)$	0.220	{ 何,?}	-0.223
{ 俺 }	0.215	$e(136)$	-0.220
{ 何, か }	0.214	{ に, 電話 }	-0.217
{ い,</s>}	0.213	{ を, 見せ }	-0.216
{ 好 }	0.210	{ 起, こ }	-0.209

5.4　訓練データ量

我々は訓練データの量が分類精度に与える影響を調査した．図 3 は **SVM + 単語埋め込み表現**と **SVM + 単語埋め込み表現 + 外部特徴量**の学習曲線で，訓練データの量が増えると分類精度も単調に増加していくことを示している．また，本研究で用いたデータの量は決して小さいものではないが，データの量を更に増やすことでより分類精度が向上し得ることを学習曲線の形が示している．これらから，タスク指向型と非タスク指向型が混ざった対話システムにおける発話の多様性をカバーするためには，膨大な量の訓練データが必要であることがわかる．図 3 は同時に外部資源の有効性を示しており，**SVM + 単語埋め込み表現 + 外部特徴量**を用いて 25%の訓練データで学習したときの精度と，**SVM + 単語埋め込み表現**を用いて全ての訓練データで学習したときの精度が同等であることがわかる．よって，外部資源を用いることでアノテーション済みのデータが不足していても性能を補うことが可能となっている．

表 7　ワーカーの投票数別（多数決での多数票別）の分類精度．

#票数	#発話数	分類精度	適合率	再現率	F 値
4	1,701	66.67	55.41	59.81	57.53
5	2,670	87.72	80.46	83.01	81.72
6	4,978	96.02	92.73	93.87	93.30
7	5,811	98.33	96.73	97.68	97.20

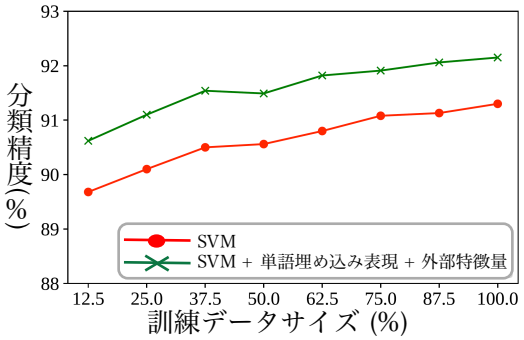

図 3　提案手法の学習曲線．横軸が交差検証の各セットで用いられる訓練データの割合．縦軸が分類精度．

5.5　発話長ごとの精度

我々は最後に，発話の文字数が分類精度に与える影響について調査した．図 4 は **SVM + 単語埋め込み表現**と **SVM + 単語埋め込み表現 + 外部特徴量**の各文字数別の

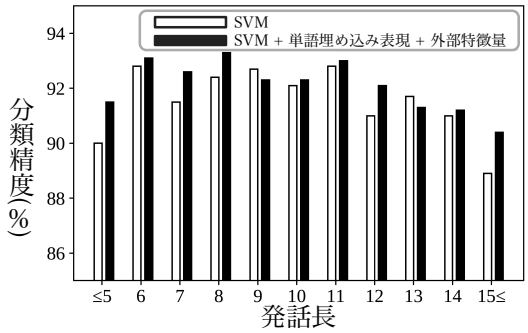

図 4 発話の文字数ごとの分類精度.

分類精度である．図 4 より，外部資源を用いることで文字数が少ない発話 (≤ 5) に効果があることが確認できる．このような短い発話は短いがゆえに判定に必要な情報を十分に含んでいないことが多いため，外部資源が有効であると考えられる[*14]．

また，文字数が多い発話 ($15 \leq$) に対しても外部資源が有効であることが確認できる．このような長い発話は発話に不必要な情報を含んでいることが多いため，そのような情報が n-gram や単語埋め込み表現の特徴量のノイズとなっていると考えられる．一方で，言語モデルのスコアなどは発話の文字数に関わらず発話の特徴を捉えることができるため，ここでも外部資源の有効性が確認できている．

6. 今後の予定

3.2 節で述べたように，「お腹が空きました」などの発話は本質的に意図が曖昧なため，現在の枠組みでこれの意図を判定するのは難しい．このような発話に対してはより洗練された対話管理モジュールを設計することが必要で，例えば発話の意図を詳細化する質問などを行うことが考えられる [27].

また 3.2 節では雑談意図ラベルの発話に対して発話行為の付与を人手で行ったが，これを対話中に自動的に行うことで，システムが適切な雑談応答を返すことを支援できる．これについてはいくつかの関連研究が既に存在している [28].

本研究ではユーザ発話の音声認識結果のみを用いて雑談意図を判定したが，他にも以前の発話などのコンテキスト情報 [4] を用いたり，発話の音声情報 [13] やユーザのプロフィール [14] を用いることができる．特に音声などのテキスト以外の情報を組み合わせることは大変興味深く，ニューラルネットワークをベースとする手法などでそのような異種情報を組み合わせることができると考えられる．

音声認識誤りは音声対話システムにおいて主要な問題であり，既存の研究ではリランキング [39] や POMDP [1] などの手法を用いて音声認識誤りに対処している．これらの手法を我々の手法に組み込むことも今後の予定として重要である．

[*14] 日本語は表意文字を含むため，5 文字程度の長さでも単純な文として機能することに注意されたい．

7. おわりに

本稿では知的対話アシスタントと呼ばれるタスク指向型，非タスク指向型の両方の性質を持つ音声対話システムにおける発話の雑談意図判定の重要さについて指摘し，2 値分類器を用いた実験を行った．知的対話アシスタントとして Yahoo! 音声アシストを用いて，実際のユーザ発話を収集し発話の雑談意図判定のためのデータセットを構築した．我々は発話の多様性に対応するため外部資源に着目し，Twitter 投稿とウェブ検索クエリをそれぞれ用いて分類器の性能向上を試みた．実験では基本特徴量を用いた 2 値分類器に外部資源から得られた特徴量を与えることでより判定性能が向上し，ベースラインの手法の性能を大きく上回ることを確認した．我々は対話システムにおけるタスク指向型，非タスク指向型の間の障壁が本研究を起点に取り除かれていくことを望んでいる．

8. 謝辞

本研究を遂行するにあたり，活発な議論にお付き合い頂いたヤフー株式会社の颯々野 学さん，橋本 力さん，東京大学の豊田 正史さん，吉永 直樹さんに感謝いたします．本論文は，The 55th Annual Meeting of the Association for Computational Linguistics に投稿した内容をもとに，議論を追加して再構成したものとなります [40].

参考文献

[1] Jason D. Williams and Steve Young. Partially observable markov decision processes for spoken dialog systems. *Computer Speech & Language*, Vol. 21, No. 2, pp. 393–422, 2007.

[2] Richard S. Wallace. *The Anatomy of A.L.I.C.E.*, pp. 181–210. Springer, 2009.

[3] 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.

[4] 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.

[5] Suman Ravuri and Andreas Stolcke. A comparative study of neural network models for lexical intent classification. In *In Proceedings of ASRU*, pp. 368–374, 2015.

[6] 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.

[7] Xiaodong Zhang and Houfeng Wang. A joint model of intent determination and slot filling for spoken language understanding. In *Proceedings of IJCAI*, pp. 2993–2999, 2016.

[8] Gokhan Tur, Dilek Hakkani-Tür, and Larry Heck. What is left to be understood in atis? In *Proceedings of IEEE*

SLT Workshop, pp. 19–24, 2010.

[9] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in NIPS*, pp. 3104–3112. 2014.

[10] Zhao Yan, Nan Duan, Junwei Bao, Peng Chen, Ming Zhou, Zhoujun Li, and Jianshe Zhou. Docchat: An information retrieval approach for chatbot engines using unstructured documents. In *Proceedings of ACL*, pp. 516–525, 2016.

[11] Cheongjae Lee, Sangkeun Jung, Seokhwan Kim, and Gary Geunbae Lee. Example-based dialog modeling for practical multi-domain dialog system. *Speech Communication*, Vol. 51, No. 5, pp. 466–484, 2007.

[12] Zhou Yu, Alan W Black, and Alexander I. Rudnicky. Learning conversational systems that interleave task and non-task content. arXiv:1703.00099, 2017.

[13] Jiepu Jiang, Ahmed Hassan Awadallah, Rosie Jones, Umut Ozertem, Imed Zitouni, Ranjitha Gurunath Kulkarni, and Omar Zia Khan. Automatic online evaluation of intelligent assistants. In *Proceedings of WWW*, pp. 506–516, 2015.

[14] Shumpei Sano, Nobuhiro Kaji, and Manabu Sassano. Prediction of prospective user engagement with intelligent assistants. In *Proceedings of ACL*, pp. 1203–1212, 2016.

[15] Julia Kiseleva, Kyle Williams, Ahmed Hassan Awadallah, Aidan C. Crook, Imed Zitouni, and Tasos Anastasakos. Understanding user satisfaction with intelligent assistants. In *Proceedings of SIGCHIIR*, pp. 121–130, 2016.

[16] Julia Kiseleva, Kyle Williams, Ahmed Hassan Awadallah, Aidan Crook, Imed Zitouni, and Tasos Anastasakos. Predicting user satisfaction with intelligent assistants. In *Proceedings of SIGIR*, pp. 45–54, 2016.

[17] Alan Ritter, Colin Cherry, and William B. Dolan. Data-driven response generation in social media. In *Proceedings of EMNLP*, pp. 583–593, 2011.

[18] Oriol Vinyals and Quoc Le. A neural conversational model. In *Proceedings of Deep Learning Workshop*, 2015.

[19] Takayuki Hasegawa, Nobuhiro Kaji, Naoki Yoshinaga, and Masashi Toyoda. Predicting and eliciting addressee's emotion in online dialogue. In *Proceedings of ACL*, pp. 964–972, 2013.

[20] Lifeng Shang, Zhengdong Lu, and Hang Li. Neural responding machine for short-text conversation. In *Proceedings of ACL*, pp. 1577–1586, 2015.

[21] Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. A neural network approach to context-sensitive generation of conversational responses. In *Proceedings of NAACL*, pp. 196–205, 2015.

[22] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-promoting objective function for neural conversation models. In *Proceedings of NAACL*, pp. 110–119, 2016.

[23] Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. A persona-based neural conversation model. In *Proceedings of ACL*, pp. 994–1003, 2016.

[24] Jiatao Gu, Zhengdong Lu, Hang Li, and Victor O.K. Li. Incorporating copying mechanism in sequence-to-sequence learning. In *Proceedings of ACL*, pp. 1631–1640, 2016.

[25] Alan Ritter, Colin Cherry, and Bill Dolan. Unsupervised modeling of twitter conversations. In *In Proceedings of NAACL*, pp. 172–180, 2010.

[26] Ryuichiro Higashinaka, Noriaki Kawamae, Kugatsu Sadamitsu, Yasuhiro Minami, Toyomi Meguro, Kohji Dohsaka, and Hirohito Inagaki. Building a conversational model from two-tweets. In *Proceedings of ASRU*, pp. 330–335, 2011.

[27] Julian J. Schlöder and Raquel Fernandez. Clarifying intentions in dialogue: A corpus study. In *Proceedings of the 11th International Conference on Computational Semantics*, pp. 46–51, 2015.

[28] Toyomi Meguro, Ryuichiro Higashinaka, Yasuhiro Minami, and Kohji Dohsaka. Controlling listening-oriented dialogue using partially observable markov decision processes. In *Proceedings of Coling*, pp. 761–769, 2010.

[29] Joseph Turian, Lev-Arie Ratinov, and Yoshua Bengio. Word representations: A simple and general method for semi-supervised learning. In *Proceedings of ACL*, pp. 384–394, 2010.

[30] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *Advances in NIPS*, pp. 3111–3119, 2013.

[31] Yoon Kim. Convolutional neural networks for sentence classification. In *Proceedings of EMNLP*, pp. 1746–1751, 2014.

[32] Rie Johnson and Tong Zhang. Effective use of word order for text categorization with convolutional neural networks. In *Proceedings of NAACL*, pp. 103–112, 2015.

[33] Rie Johnson and Tong Zhang. Semi-supervised convolutional neural networks for text categorization via region embedding. In *Advances in NIPS*, pp. 919–927. 2015.

[34] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of EMNLP*, pp. 1724–1734, 2014.

[35] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv:1412.3555, 2014.

[36] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of AISTATS*, pp. 297–304, 2010.

[37] Tomas Mikolov, Anoop Deoras, Daniel Povey, Lukas Burget, and Jan Cernocky. Strategies for training large scale neural network language models. In *Proceedings of IEEE ASRU Workshop*, pp. 196–201, 2011.

[38] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of ICLR*, 2015.

[39] Fabrizio Morbini, Kartik Audhkhasi, Ron Artstein, Maarten Van Segbroeck, Kenji Sagae, Panayiotis Georgiou, David R. Traum, and Shri Narayanan. A reranking approach for recognition and classification of speech input in conversational dialogue systems. In *Proceedings of SLT*, pp. 49–54, 2012.

[40] 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 (to appear).