知的対話アシスタントにおけるユーザエンゲージメントの予測

佐野 峻平　　鍛治 伸裕　　颯々野 学
ヤフー株式会社
{shsano, nkaji, msassano}@yahoo-corp.jp

1 はじめに

近年, Apple の Siri をはじめ, ユーザとの対話から意図を汲み取りユーザの代わりに様々なタスクをこなしたり, 雑談相手になったりと, 秘書のような役割をこなす対話システム (以下, 知的対話アシスタントと呼ぶ) が普及してきている.

そうした知的対話アシスタントシステムの運用においては, 長期間にわたって継続的に利用してくれるユーザの数を増やすことが重要な課題の一つとなる. そのためには, 利用の継続や中止につながる要因の分析を行い, それに基づいて機能拡張及び改善を行うことが考えられる. また, 利用をやめそうなユーザを事前に検知して, 利用を継続するよう働きかけることも考えられる. しかし一般的に, 知的対話アシスタントは天気情報の取得やウェブ検索, 雑談といった多種多様な機能を備えている. そのため, どの機能がユーザの継続的利用に寄与しているのかは, 自明に分かることではない.

上記の課題を解決するため, 本研究では, 一定期間知的対話アシスタントを継続利用したユーザが, その後も利用を継続するか否かを予測するモデルの構築を行った (このタスクをユーザエンゲージメント予測と呼ぶ). これにより, 近い将来, 利用をやめてしまうユーザを事前に発見することができるようになると考えられる. また, モデルの素性選択や, 学習された素性の重みの調査を通して, 継続利用および利用の中止につながる要因に関する知見が得られることも期待できる.

2 ユーザエンゲージメント予測

本節ではまず知的対話アシスタントの特徴を概観したうえで我々が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 や NTT ドコモのしゃべってコンシェル, Microsoft の Cortana, Yahoo! JAPAN の Yahoo!音声アシストなどがある. これらのシステムはユーザとの対話から意図を汲みとり, ユーザの代わりに様々なタスクをこなす秘書のような身近な存在であり, 下記の特徴を有している.

- 主にモバイル端末で利用される.
- 音声認識, および音声合成を用いることによって, ユーザと対話的にやりとりを行う.
- 天気情報の取得や雑談など, 複数の機能をユーザに提供する.

システム毎に対応している機能に細かな違いはあるが, 大抵のシステムに共通しておりかつユーザに利用されやすい機能として, モバイル端末の操作 (ex. アプリ起動, 音楽再生), コミュニケーション (ex. 通話, メール), 位置 (ex. 目的地への経路表示), カレンダー (ex. 予定表の確認), 天気 (ex. 天気の確認) などがあげられる [4].

本稿での分析や評価には, 知的対話アシスタントの一つである Yahoo!音声アシスト [8] を用いる.

2.2 タスク設定

本項では, 我々が取り組むユーザエンゲージメント予測というタスクの詳細を述べる.

まず, 知的対話アシスタントの継続利用という概念を具体的に定義する. 知的対話アシスタントを継続して利用するユーザには毎週土曜日にスポーツの試合結果を確認するといった習慣的な利用傾向が考えられるため, 毎週一回以上の発話があれば継続して利用しているものとする. 一定期間継続したユーザだけを対象としているのは, 知的対話アシスタントのユーザには, システムの性能や備わっている機能を試すために 1, 2 週ほど使って利用をやめてしまうユーザが多く含まれるからである. これらのユーザは, そもそも継続利用する見込みが少ないことや, 予測に利用できる発話ログが少ないことから, 今回の予測対象からは除外した.

上記の考えに基づき, 本稿ではユーザエンゲージメント予測を「4 週間継続して利用したユーザがその後 4 週間も継続して利用するか否か」を予測する問題とする. このとき, 利用を継続したユーザを継続ユーザ, それ以外のユーザを離脱ユーザと呼ぶ.

3 データセット

Yahoo!音声アシストのユーザのうち, 以下の二条件を満たすユーザから 50,000 ユーザを無作為に選別した.

- 2015 年 3 月から 8 月の間にインストールした.
- インストール後 4 週間, 毎週一回以上発話した.

これらユーザの, インストール後 4 週間以内の全発話ログをデータセットとし, 前節で述べた継続/離脱のラベルを各ユーザに付与する. 継続ユーザと離脱ユーザの数はそれぞれ 25,093 と 24,907 であった. 表 1 に発話ログの具体例を示す. 一つの発話には, ユーザ ID, ユーザの発話時刻, ユーザの発話, システムの応答, システムの応答タイプ, 誕生日設定の情報, 呼び名設定の情報の七つが含まれている. 応答タイプとは, システムが返す応答を機能ごとに分類したものであり, 天気や検索など, 68 種類のタイプが定義されている.

50,000 ユーザを無作為に 8:1:1 に分け, それぞれ学習データ, 開発データ, 評価データとする.

4 素性

本節では本稿での実験に用いる素性について説明する. 素性のリストを表 2 に示す. これらの素性は大別してユーザの属性 (Attribute, A), ユーザの利用傾向 (Usage, U), システムの応答 (Responce, R), 音声認識・合成に関するもの (Speech, S) の四つになる.

4.1 ユーザの属性に関する素性 (A)

知的対話アシスタントの多くで設定できる, 呼び名や誕生日などユーザの属性に関する情報に関する素性について述べる. それらの情報はユーザの誕生日にシステムがお祝いの応答を返す, ユーザが設定した呼び名をつけた応答を返すといった活用がなされている. 属性情報の設定をしているユーザほどシステムの習熟度が高く, また属性の設定をするという行為は一種のエンゲージメント行動と言えるため, 設定をしていないユーザよりも継続して利用する割合が多いと予想される.

素性としては, インストール後 4 週間以内に一度でも呼び名情報を設定したか (呼び名設定の有無), インストール後 4 週間以内に一度でも誕生日情報を設定したか (誕生日設定の有無), 誕生日情報が設定されている場合には 2016 年 1 月 1 日時点での年代 (10 代以下か, 20 代か, . . . , 50 代か, 60 代以上か, の六つのいずれか一つ) の三つを用いる. いずれも, 設定ありもしくは該当する場合に 1, そうでない場合に 0 の値が付与される.

4.2 ユーザの利用傾向に関する素性 (U)

システムとの雑談を好む, 休日の昼間に利用しやすいなどの, ユーザ毎の利用傾向に関する素性を説明する. まず, エンゲージメントを顕著に表すものとして, 発話数を素性とした. ただし, 発話数はそのままでは大きすぎる値の場合もあるため, 対数をとった. さらに週末の昼間に利用しやすい, 毎朝欠かさず利用するといった周期的な利用傾向を見るため, 次のような素性を設計した. ユーザの発話時間帯を各曜日 0-6 時, 6-12 時, 12-18 時, 18-24 時の 28 個に区切り, 全発話数に対する, 各区切り内における発話数の割合を素性とした. (週内発話分布). 例えば月曜 0-6 時に 4 発話, 金曜 0-6 時に 2 発話あるユーザの場合, 月曜 0-6 時が 0.8, 金曜 0-6 時が 0.2, その他の時間区切りが 0 となる. またインストール後の週毎のシステム利用率変化を見るため, 週毎の発話日時の分布も同様の発話割合を週毎に求めて, それらを素性とする (週間発話分布).

4.3 システムの応答に関する素性 (R)

検索や雑談など, 知的対話アシスタントがユーザの発話意図を解釈して返す応答タイプに関する素性について述べる. 応答タイプの傾向を見ることで, 継続ユーザに好んで使われやすい機能や, 意図解釈誤りにより離脱につながりやすい機能などが見えてくる.

ゲームは, 知的対話アシスタントへのエンゲージメントを向上させる効果があることが [5] により確認されている. 逆に, 発話意図を解釈できなかった場合に返されるエラー応答はエンゲージメントを低下させると考えられる. 音声アシストのゲーム機能であるしりとりと連想ゲームの全応答に占める割合をゲーム応答率, エラー応答の割合をエラー応答率とし, これらを素性として用いる. またユーザ毎にシステムの応答タイプを時系列順に並べたときの, 学習データで 10 回以上登場した応答タイプ n-gram を素性とする. この際, ユーザの次の発話までに 20 分以上の間隔が空いていたらシステム利用の区切りを表す EOS を入れている. 例えば表 1 の User A の応答タイプ列は, ”雑談”, ”天気”, EOS, ”検索”, EOS となる.

4.4 音声認識・合成に関する素性 (S)

エンゲージメントを低下させる要因の一つである, 音声認識・合成誤りに関する素性の説明をする. 音声認識の性能は, ユーザ体験の質に直結するため離脱するかに影響する要素である. 発話の音声認識誤りがあった際の分析を行った [3] では, ユーザは似た発話を繰り返す傾向にあることや, 長い発話ほど音声認識誤りが起こりやすいことがわかっている. また応答文が長いと音声合成を誤る確率が上がる. そのうえ読み上げに時間がかかりユーザが次の発話を行うまでに時間がかかるため, 応答文の長さも離脱に影響する要素と言える.

素性は発話文字数の平均 (平均発話長), 応答文字数の平均 (平均応答長), 直後の発話との編集距離が閾値以下の発話の割合 (言い直し発話率) とする. 編集距離には値が 0 から 1 の範囲を取り 0 に近づくほど発話

表 1: 音声アシストの発話ログの例

ユーザ ID	発話時刻	発話	応答	応答タイプ	誕生日設定	呼び名設定
User A	2015-3-1 07:00	おはよう	おはよう	雑談	2000-1-1	太郎
User A	2015-3-1 07:01	今日の天気	晴れです	天気	2000-1-1	太郎
User A	2015-3-2 12:34	xxx を検索	(検索結果を表示)	検索	2000-1-1	太郎

表 2: 素性リスト

素性	素性 ID
呼び名設定の有無	A_1 (Attribute)
誕生日設定の有無	A_2
年代	A_3
発話数	U_1 (Usage)
週内発話分布	U_2
週間発話分布	U_3
応答タイプ unigram	R_1 (Responce)
応答タイプ bigram	R_2
ゲーム応答率	R_3
エラー応答率	R_4
平均発話長	S_1 (Speech)
平均応答長	S_2
言い直し発話率	S_3

が似ていることを表す正規化レーベンシュタイン距離 [9] を用い, 言い直し発話とする閾値は経験的に 0.5 とした.

5 評価実験

本節ではユーザエンゲージメント予測実験を行い結果を考察する. それに先立って, エンゲージメントに影響を与える要因の分析を行いユーザエンゲージメント予測での素性選択の参考にした.

5.1 エンゲージメントへの影響要因分析

学習データの 40,000 ユーザを用い表 2 の各素性の値と 5 週目以降の継続週数 (毎週一回以上発話があれば継続) との Spearman の順位相関係数を調べ, システムへのエンゲージメントに影響する要素の分析を行う. 相関値の絶対値が大きいほど継続利用に寄与しやすく, エンゲージメントにより影響しやすい素性といえる. 5 週目以降の継続週数は最小 0, 最大 4 (4 以上は 4 に丸めた) で該当ユーザ数は順に 7,546, 5,360, 3,869, 3,107, 20,118 となっている.

各素性で Spearman の順位相関係数の絶対値が最大の要素とその値を, 表 3 に示す. 最も相関の高い素性は週間発話分布で, インストール後 4 週目の発話率が高いほどエンゲージメントが高い結果となった. これは, インストール後の発話の減少率が低いユーザほど継続利用しやすいという直感に沿った結果で, 初週の発話率との相関が-0.20 と負の相関であることからもそのことが裏付けられる. 次いで相関が高いのが応答タイプ n-gram 素性で, 占い機能に関する要素が相関最大となっている. 占い以外ではおみくじやニュース, 天気, アラームなどの習慣的に利用される機能に関する素性が相関が高く, 逆に検索や雑談といった習慣的に利用されにくい機能の素性は負の相関が見られた.

表 3: 5 週目以降の継続週数との順位相関

素性 ID	素性	相関値	最大要素
A_1	呼び名設定の有無	0.09	
A_2	誕生日設定の有無	0.08	
A_3	年代	0.08	50 代
U_1	発話数	0.19	
U_2	週内発話分布	0.20	火曜 6-12 時
U_3	週間発話分布	0.25	4 週目
R_1	応答タイプ unigram	0.21	占い (通知)
R_2	応答タイプ bigram	0.21	占い (通知)_EOS
R_3	ゲーム応答率	0.03	
R_4	エラー応答率	-0.07	
S_1	平均発話長	-0.08	
S_2	平均応答長	0.19	
S_3	言い直し発話率	-0.11	

表 4: 評価データでの分類精度

手法	ベースライン (発話数のみ)	提案手法
精度 (%)	58.2	71.4

週内発話分布は, 平日の 6-12 時の利用率が全て相関 0.20 と他に比して高い. ここからも習慣的に知的対話アシスタントを使用しているユーザが継続利用しやすい傾向が見て取れる. 平均応答長は長いほど音声読み上げに時間がかかりエンゲージメントが低下する, という分析前の予想に反し, 長いほどエンゲージメントを高める結果となった. これは, 占いやニュースなど継続ユーザに利用されやすい機能の応答は長く, 雑談や検索は短いことが影響したと考えられる.

5.2 ユーザエンゲージメント予測

本項では, ユーザエンゲージメント予測と予測結果の考察を行う. 学習データで線形 SVM[1] のモデル作成, 開発データでモデルのパラメータ調整を行い最終的な評価には評価データを用いる. 素性は前項の結果を踏まえ, 表 3 での相関の絶対値 0.10 以上の素性のみを使用する. 尚, 各素性は事前に学習データ中の最小値, 最大値がそれぞれ 0, 1 となるように正規化した.

ベースラインとして発話数を用いた場合との評価データでの精度比較結果を表 4 に示す. 提案手法はベースラインの性能を大きく上回っている. 5 週目以降の継続利用週数毎に離脱を継続と誤ったユーザの内訳を見ると, 0 週が 19.7%, 1 週が 27.4%, 2 週が 37.7%, 3 週が 40.5%となっており, 早々に離脱したユーザに限定すると 80%を上回る精度で分類できている.

次に提案手法で予測に効いていた素性を分析する. SVM モデルの重みの絶対値が大きい素性のリストを表 5 に示す. アラームや占いなどの習慣的な機能は継続につながりやすく, 習慣的に使われにくい雑談やしりとりが離脱につながりやすい傾向がみてとれる.

表 5: SVM モデルの重みが大きい素性

重み	素性	重み	素性
3.51	U_1 発話数	-0.82	定数項
0.42	R_1 EOS	-0.67	U_3 1 週目発話割合
0.40	U_3 4 週目発話割合	-0.47	R_1 雑談
0.40	R_2 EOS_アラーム	-0.41	U_3 2 週目発話割合
0.33	R_1 アラーム	-0.39	S_2 平均応答長
0.29	R_1 占い (通知)	-0.35	R_2 雑談_雑談
0.24	R_1 雑談 (通知)	-0.30	R_1 端末操作
0.24	R_2 EOS_占い (通知)	-0.29	R_1 検索
0.23	R_2 アラーム_EOS	-0.27	R_1 しりとり
0.22	R_2 アラーム_天気	-0.27	R_1 しりとり_しりとり

[5] でしりとりのようなゲーム機能はシステムへのエンゲージメントを高める効果があると確認されているが, 継続して使ううちに徐々にゲームは遊ばれなくなるのだと考えられる. 平均応答長は, 前項で継続週数との正の相関がみられたのに反して離脱に影響のある結果となったが, 占いなどの継続につながりやすい機能が平均応答長が長いため正の相関が出ており, それらの影響を排した平均応答長自体で言えば長いほど離脱につながりやすいと言える. また応答タイプ unigram 中の EOS の割合が多いほど, すなわち知的対話アシスタントとの一回のやりとりの回数が少ないほど継続しやすいこともわかる. ここから, 知的対話アシスタントを継続して利用するユーザは, 特定の機能を用いる際に起動し目的を達成したら終了する, という使い方を繰り返しやすいと考えられる.

6 関連研究

知的対話アシスタントは Web 検索や雑談, 天気情報の取得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 [6] とは一線を画する. 個々の機能に着目すれば, Web 検索 [2] や雑談 [7] などの評価は精力的に研究が行われてきた分野である. これらの研究は知的対話アシスタントの分析や評価を行う上で参考になるが, エンゲージメントに関しては論じていない.

知的対話アシスタントに関する研究は, 評価指標の提案を行った [4] が先駆けとなっている. [4] は知的対話アシスタントとのやりとりを通じて被験者に特定のタスク (指定した人に電話をかける, 今日の天気を調べるなど) を行わせ, その際のシステムに対するユーザ満足度を予測するモデルを提案している. [4] では特定のタスクの達成という短期的な利用での知的対話アシスタントの評価を目指したものだが, 本研究はより長期間にわたる利用に対する評価を目指しているため相補的な関係にある.

7 おわりに

本稿では, 知的対話アシスタントを継続利用したユーザが, その後も利用を継続するか否かを予測する, ユーザエンゲージメント予測のモデル構築を行った. 予測精度は 71.4%, 早期にシステムの利用をやめてしまうユーザに関しては 80%強であった. また, モデルの素性選択や学習された素性の重みの調査を通して, 平日の毎朝利用する, アラームや占いなど特定の機能を好んで使うなど, 習慣的な使い方をしているユーザは継続利用しやすく, 雑談やウェブ検索など習慣的に利用されにくい機能を多く使うユーザは離脱しやすいことがわかった.

今後は, 構築したユーザエンゲージメント予測のモデルを用いて近い将来利用をやめてしまいそうなユーザを見つけ, 利用を継続するよう働きかけることで, 実際にエンゲージメントの改善に繋がるかを調査したい. また, ユーザエンゲージメント予測の更なる精度向上にも取り組む予定である.

参考文献

[1] Fan et al. LIBLINEAR: A library for large linear classification. *The Journal of Machine Learning Research*, Vol. 9, pp. 1871–1874, 2008.

[2] Hassan et al. Beyond DCG: user behavior as a predictor of a successful search. In *Proc. of WSDM*, pp. 221–230, 2010.

[3] Jiang et al. How do users respond to voice input errors?: Lexical and phonetic query reformulation in voice search. In *Proc. of SIGIR*, pp. 143–152, 2013.

[4] Jiang et al. Automatic online evaluation of intelligent assistants. In *Proc. of WWW*, pp. 506–516, 2015.

[5] Kobayashi et al. Effects of game on user engagement with spoken dialogue system. In *Proc. of SIGDIAL*, pp. 422–426, 2015.

[6] Walker et al. Paradise: A framework for evaluating spoken dialogue agents. In *Proc. of EACL*, pp. 271–280, 1997.

[7] Young et al. POMDP-based statistical spoken dialog systems: A review. *Proc. of IEEE*, Vol. 101, No. 5, pp. 1160–1179, 2013.

[8] Yahoo! JAPAN. Yahoo!音声アシスト, 2015. `http://v-assist.yahoo.co.jp/`.

[9] Yujian and Bo. A normalized levenshtein distance metric. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, Vol. 29, No. 6, pp. 1091–1095, 2007.