対話システムログから不自然な会話を検出する

橋本 力　　颯々野 学
ヤフー株式会社

1 はじめに

Yahoo!音声アシスト[1]等の対話システムのログから不自然な会話を検出する手法を提案する。自然な会話には「わかったありがとう」等のユーザからの好意的 feedback が、不自然な会話には「どういう意味？」等の非好意的 feedback が後続しやすいが、本手法はこうした feedback を活用する（表 1）。本研究の好意的 feedback とは賞賛、理解、感謝、歓喜等を、非好意的 feedback とは失望、誤解、侮蔑、退屈等を示すユーザ発話である。前者は「やるねぇ」、後者は「もうアンインストール」等を含む。単純な手法はログ中の非好意的 feedback の直前の会話（例: 表 1 下の会話）を不自然として検出するものだが、次の困難に直面する。

ユーザ発話	大阪から東京まで
システム応答	大阪 9 時発の新幹線があります
ユーザ発話	わかったありがとう **（好意的 feedback）**
ユーザ発話	風邪ひいたかも
システム応答	ふふふ
ユーザ発話	どういう意味？ **（非好意的 feedback）**

表 1: 対話システムとユーザの会話の例

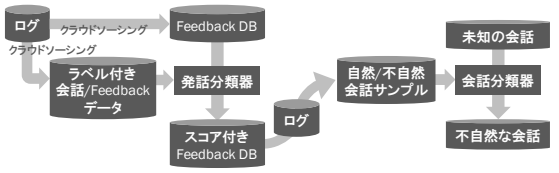

図 1: システム全体像

曖昧性 Feedback の好意的/非好意的評価の対象は会話の自然さ/不自然さだけとは限らない。例えば対話システムの冗談に、たとえ会話が成立していても「ふざけるな」と非好意的に feedback する場合や、たとえ会話が成立していなくても「あなたはそのままで十分だと思います」等の心温まるシステム応答に「ありがとう」と好意的に feedback する場合、音声認識が成功しただけで好意的に feedback する場合もある。

低頻度 Feedback の出現頻度は低く（本研究では全ユーザ発話の約 12%）会話の多くは feedback が無く、上記の単純な手法では不自然な会話の多くを見逃す。

本研究では曖昧性に対処すべく、会話の自然/不自然を曖昧性無く示す傾向が強い feedback （例:「了解！」「意味不明」）とその他を区別する**発話分類器**を学習する。そうした曖昧性の低い好意的/非好意的 feedback 直前の会話を自然/不自然会話の学習データとしてログから大量に抽出し、後続する feedback の有無に関わらずログ中の全会話を自然/不自然に分類できる**会話分類器**を学習する。こうして低頻度問題に対処する。

会話分類器は曖昧性の低い feedback とログがあれば自動で学習でき、また feedback はドメイン（雑談や天気予報等の対話領域）に依存しにくい。つまり本手法の新ドメインへの適応に必要なのは当該ドメインのログだけで、ドメイン依存性が低いと期待できる。

本研究では Yahoo!音声アシストの約 5 年分のログを用いた実験により、発話分類器と会話分類器の有効性、本手法のドメイン依存性の低さ等を定量的に示した。

2 関連研究

関連が深いのは対話システムの自動評価に関する研究であり、システム応答を人手作成した参照応答と比較する手法 [4] や、報酬等の外部信号を前提とする手法 [6] 等がある。参照応答は feedback に比べドメインに強く依存する。外部信号は一般に高コストだが、会話に自然に出現する feedback は低コストで利用できる。

対話システムのユーザ満足度自動評価も関連が深い。[2] はセッション毎、[5] はユーザ毎の満足度評価手法を提案した。本手法は会話毎の評価手法と言えるが、不自然な会話の検出には会話毎の手法が適切である。

Yes や *nope* 等の確認 feedback を利用する手法 [3] もあるが、本手法は「もうアンインストール」等のより広範な feedback を利用する。また、feedback に基づく手法が直面する曖昧性と低頻度の問題への対処法を提案したのは我々の知る限り本研究が初めてである。

3 提案手法

図 1 に本手法の全体像を示す。ログから構築したラベル付き会話/Feedback データ（後述）で発話分類器

[1] https://v-assist.yahoo.co.jp/

ユーザ発話	u	大阪から東京まで
システム応答	r	大阪 9 時発の新幹線があります
自然/不自然ラベル	l	自然
Feedback	f	わかったありがとう **（好意的）**
ユーザ発話	u	風邪ひいたかも
システム応答	r	ふふふ
自然/不自然ラベル	l	不自然
Feedback	f	どういう意味？ **（非好意的）**

表 2: ラベル付き会話/Feedback データの例

を学習し、同じくログから収集した feedback にスコア付けする。スコアの低い/高い feedback ほど会話の自然/不自然を曖昧性無く示す傾向にあり、それらを用いて自然/不自然な会話をログから抽出できる。それらの会話で学習した会話分類器により、直後に feedback が続かないものも含め全会話を自然/不自然に分類する。

3.1 Feedback DB の構築

Yahoo!クラウドソーシング[2]（作業者 780 名）により発話約 4 万件に好意的/非好意的/中立のラベルを付与した。発話あたり 3 名で判断し、ラベルは多数決で決めた。3 名が同じ判断をした発話は全体の 69.8%、3 名とも違う判断をしたのは 1.5%のみである。著者らが別途収集した 4,066 の好意的、2,434 の非好意的 feedback を追加し、全体を著者らでチェックして 8,988 の好意的、11,316 の非好意的 feedback から成る DB を得た。クラウドソーシング結果から好意的/非好意的 feedback は全発話（異なり）の 5%/7%と推定する。

3.2 ラベル付き会話/Feedback データの構築

ラベル付き会話/Feedback データ（例:表 2）は発話分類器の学習と会話分類器の tuning、評価に用いる。次の通り構築した。まず Yahoo!クラウドソーシング（作業者 1,071 名）でログ中の会話 $\langle u, r\rangle$ に自然/不自然ラベル l を付与し $\langle u, r, l\rangle$ を作る。表 1 下の場合 $\langle u, r, l\rangle$ は $\langle$ 風邪ひいたかも, ふふふ, 不自然 $\rangle$ となる。$\langle u, r, l\rangle$ 群は 3 set に分割する。次に $\langle u, r, l\rangle$ の $\langle u, r\rangle$ に後続する feedback f をログから抽出し $\langle u, r, l, f\rangle$ とする。対象の f は 3.1 節で述べた feedback DB に収録のものである。表 1 下の場合 $\langle u, r, l, f\rangle$ は $\langle$ 風邪ひいたかも, ふふふ, 不自然, どういう意味?$\rangle$ となる。$\langle u, r, l, f\rangle$ 群も $\langle u, r, l\rangle$ 群に対応させる形で 3 set に分割する。

表 3 にデータの内訳を示す。$\langle u, r, l\rangle$ より $\langle u, r, l, f\rangle$ の数が多いのは、1 つの $\langle u, r\rangle$ がログに複数回出現し、複数の異なる f がそれに後続しうるためである。

$\langle u, r, l, f\rangle$ の Set 1 は発話分類器の学習に用いる。$\langle u, r, l, f\rangle$ からは会話の自然/不自然 (l) と feedback f の結びつきが観察でき、発話分類器の学習データとして利用できることに注意されたい。$\langle u, r, l\rangle$ の Set 2、3 は各々会話分類器の tuning と評価に用いる。

[2] https://crowdsourcing.yahoo.co.jp/

		Set 1	Set 2	Set 3
$\langle u, r, l\rangle$	不自然	4,787	1,870	1,827
	自然	7,746	3,056	3,108
$\langle u, r, l, f\rangle$	不自然	77,366	7,907	6,677
	自然	167,826	19,505	19,882

表 3: $\langle u, r, l\rangle$ 群と $\langle u, r, l, f\rangle$ 群の内訳

3.3 発話分類器の学習と Feedback のスコア付け

発話分類器は曖昧性無く会話の自然/不自然を示す feedback に低い/高いスコアを、曖昧なものに中間のスコアを付与する。分類器として fastText[3]を default 設定のまま、Wikipedia から skipgram で学習した単語ベクトルと共に用いる。単語分割には MeCab[4]を用いる。この分類器は入力された feedback の直前に現れる会話の自然/不自然を表すラベルとその尤度を出力するが、自然ラベルに対する尤度に -1 を掛け全尤度を feedback 直前の会話の不自然さを示すようにする。

発話分類器で DB 中の全 feedback にスコア付けし、スコア付き feedback DB を得る。例えば「会話になってません」は 0.79、「頭良い」は -0.78 が付与される。

3.4 会話分類器の学習と不自然会話の検出

スコア付き feedback を用いて会話分類器の学習データとなる自然/不自然会話サンプルをログから次の通り抽出する。まず、スコア付き feedback が後続する会話をログから抽出し、その会話に当該 feedback のスコアを付与する。表 1 下の場合、feedback「どういう意味?」に 0.9 が付与されていれば、「風邪ひいたかも」「ふふふ」という会話に 0.9 が付与される。[5] 次に、スコア上位 K 会話を不自然な会話の、スコア下位 K 会話を自然な会話のサンプルとして抽出する。K の値として $\langle u, r, l\rangle$ の Set 2 に対する適合率-再現率曲線の AUC を最大化する $50,000$ を選択した。[6]

発話分類器と同様、会話分類器として fastText を Wikipedia から得た単語ベクトルと共に default 設定のまま用いる。会話分類器への入力は、会話（ユーザ発話とシステム応答）を一文字列に連結し MeCab で単語分割したものである。この分類器は入力された会話の自然/不自然を表すラベルとそのラベルの尤度を出力するが、自然ラベルに対する尤度に -1 を掛けることで全尤度を会話の不自然さを示すようにする。

[3] https://github.com/facebookresearch/fastText

[4] http://taku910.github.io/mecab/

[5] 複数の異なる feedback が後続する会話にはそれら feedback のスコアのうち最大のものを付与する。

[6] 他、10,000、100,000、150,000、200,000、250,000、300,000、350,000、400,000、450,000、500,000 を試した。

PROPOSED: 提案手法。

PROPOSED-UC: 発話分類器無しの PROPOSED。自然/不自然会話サンプル抽出時、発話分類器スコアの代わりに $|FB_{unf}|-|FB_{fav}|$ を抽出対象の会話に付与する。$|FB_{unf}|$ と $|FB_{fav}|$ は各々、当該会話に後続する非好意的、好意的 feedback の数で、大きい値は会話の不自然さを示す。

PROPOSED-CC: 会話分類器無しの PROPOSED。代わりに $|FB_{large}(s)| - |FB_{small}(-s)|$ の値で会話を分類する。$|FB_{large}(s)|$ と $|FB_{small}(-s)|$ は各々、当該会話に後続する、発話分類器スコアが s より大きい feedback 数と同スコア $-s$ 未満の feedback 数であり、大きい値は会話の不自然さを示す。Feedback が後続しない会話には（ログの大半が自然な会話のため）小さい値 $(-9,999)$ を付与する。上記 s と自然/不自然分類のための閾値 t は $\langle u,r,l\rangle$ の Set 2 に基づき各々0.5、−190 とした。s として 0.5、0.6、0.7、0.8 を、t として −200、−190 ... 190、200 を試した。

MAJORITYVOTE: いずれの分類器も使わず直後の feedback に頼るのみの手法。PROPOSED-UC の $|FB_{unf}|-|FB_{fav}|$ の値で会話を分類する。Feedback が後続しない会話には小さい値 $(-9,999)$ を付与する。自然/不自然分類用の閾値 t は PROPOSED-CC と同様の手続きで −80 とした。

SUPERVISED: 単純な教師あり学習による手法。Feedback も発話分類器も使わず、人手作成したデータ（$\langle u,r,l\rangle$ の Set 1）で会話分類器を学習する。分類器は PROPOSED 同様 fastText を用いる。SUPERVISED と次の PROPOSED+SUP は新ドメインへの適応の都度学習データを作る必要がある。

PROPOSED+SUP: PROPOSED の会話分類器学習の際、SUPERVISED で使用した人手作成したデータも併せて使う。

表 4: 提案手法と比較手法

最後にログ中の会話を会話分類器でスコア付けする。評価実験では $\langle u,r,l\rangle$ の Set 3 にスコア付けした。分類の失敗例として諺に関する会話がある。「ことわざ」「犬も歩けば棒に当たる」は自然な会話だが、「どういう意味？」というユーザ発話が続く場合がある。これは feedback として意図された発話ではないが表層上は非好意的 feedback に見え、会話分類器は誤って不自然と判定した。一方「リオデジャネイロの天気」「今日の仙台の天気は、雪後晴でしょう」は不自然だが、この不自然さの理解には会話分類器には備わっていない場所に関する常識が必要であり、また場所以外については自然な会話であるため誤って自然と判定した。

4 評価実験

Yahoo!音声アシストの約 5 年分のログで、次の点を定量的に示す。(1) 発話分類器、会話分類器は性能に寄与する。(2) 直後の feedback のみに頼る単純な手法は性能が低い。(3) 本手法はドメインに依存しにくい。

4.1 本手法と比較手法の性能評価

表 4 の手法を比較する。表 5 に各手法の正解率、適合率、再現率、F1 を挙げる。PROPOSED との差は全て統計的に有意 (McNemar's test: $p < 0.01$) である。

	正解率	適合率	再現率	F1
PROPOSED	0.784	0.7158	**0.6907**	0.7031
PROPOSED+SUP	**0.7998**	**0.751**	0.6869	**0.7176**
PROPOSED-UC	0.7106	0.6041	0.6338	0.6186
PROPOSED-CC	0.5277	0.3683	0.3859	0.3769
SUPERVISED	0.7435	0.6735	0.5961	0.6324
MAJORITYVOTE	0.5293	0.3685	0.3804	0.3744

表 5: 提案手法と比較手法の性能

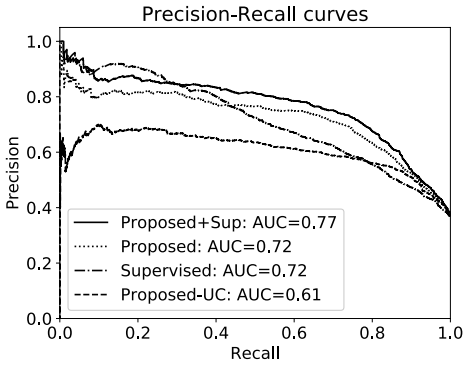

図 2: F1 上位 4 手法の適合率-再現率曲線

図 2 に F1 上位 4 手法の適合率-再現率曲線を示す。PROPOSED、PROPOSED-UC、PROPOSED-CC を比較すると発話分類器、会話分類器がいずれも PROPOSED の性能に寄与しているのがわかる。MAJORITYVOTE の性能の低さは、直後の feedback に頼るのみでは不自然な会話をうまく検出できないことを示している。PROPOSED は SUPERVISED より高性能だが、これは PROPOSED が自動抽出した会話分類器の学習データ（自然/不自然会話サンプル各々5 万件、計 10 万件）が SUPERVISED の学習データ（$\langle u,r,l\rangle$ の Set 1 約 1 万件）より膨大なためと考えられる。PROPOSED+SUP の結果は PROPOSED に人手作成したデータを追加することでさらに性能が向上することを示している。

4.2 ドメイン非依存性の評価

Yahoo!音声アシストで最も頻繁に用いられる雑談、web 検索、天気予報の 3 ドメインを対象に、本手法の out-of-domain 版と in-domain 版を比較することでドメイン非依存性を評価した。Out-of-domain 版は次の通り構築する。まず、評価対象のドメインに属する $\langle u,r,l,f\rangle$ をその Set 1 から除外した上で発話分類器を学習する。Yahoo!音声アシストは各ユーザ発話 u をドメインに分類するが、$\langle u,r,l,f\rangle$ のドメインはその u に割り当てられたドメインとする。次に、対象ドメインの feedback を除外した上でスコア付き feedback DB を構築する。Feedback のドメインはその直前の会話のユーザ発話のドメインとする。それ以降の手続きは

	雑談	Web 検索	天気予報
$\langle u, r, l, f \rangle$	55,573 (9.2)	223,457 (33.5)	243,313 (31.7)
Feedback	10,664	19,025	19,998
評価 (対象)	2,274 (66.9)	1,084 (12.0)	827 (9.7)
評価 (全部)	4,935 (37.0)	4,935 (37.0)	4,935 (37.0)

表 6: ドメイン非依存性評価実験条件

3.4 節で述べたものと同じである。つまり対象ドメインについてはログは蓄積されているが、対象ドメインの $\langle u, r, l, f \rangle$ と feedback は整備されていない、という想定の状況のもと実験する。なお、本手法で手作業を要するのはラベル付き会話/Feedback データ $\langle u, r, l, f \rangle$ と feedback の収集のみで、これらを skip できればドメイン適応を自動化できることに注意されたい。

In-domain 版は 3 節で述べた通りに構築する。但しドメイン以外の条件を同じにすべく、$\langle u, r, l, f \rangle$ の数とスコア付き feedback DB の feedback 数は、random sampling により out-of-domain 版と同じ数に減らす。

評価データは 2 種類用意した。1 つは $\langle u, r, l \rangle$ の Set 3 で、Yahoo!音声アシストの全ドメインが含まれる（以降このデータを「評価 (全部)」と表記する）。もう 1 つは $\langle u, r, l \rangle$ の Set 3 から対象ドメインの会話のみを抽出したものである（「評価 (対象)」と表記する）。つまり 3 ドメイン × 評価データ 2 つの 6 条件で評価する。

表 6 に発話分類器学習データとして用いる $\langle u, r, l, f \rangle$ の件数、DB 中の feedback 数、2 つの評価データの件数を対象ドメイン毎に示す。[7] 括弧内の数字は不自然ラベルが付与されたものの割合である。雑談ドメインの $\langle u, r, l, f \rangle$ 数 (55,573) が他と比べて少ないが、雑談が最頻出ドメインであり、雑談に属する $\langle u, r, l, f \rangle$ を除外すると残りが僅かになるからである。雑談の $\langle u, r, l, f \rangle$ の不自然ラベルの割合 (9.2) が他と比べて少ないが、雑談は open-ended で対話システムにとって難しいためその多くは不自然であり、雑談に属する $\langle u, r, l, f \rangle$ を除外すると残りの多くは自然ラベルの $\langle u, r, l, f \rangle$ となるためである。雑談ドメインの feedback 数 (10,664) が他と比べて少なく、評価 (対象) の件数 (2,274) が他と比べて多いのも雑談が最頻出のドメインだからである。雑談の評価 (対象) の不自然ラベルの割合 (66.9) が他と比べて多いのは、雑談が web 検索や天気予報の会話に比べ対話システムにとって困難だからである。一方、評価データ中の不自然会話の割合が高いため、不自然会話検出は雑談ドメインが最も容易なことに注意されたい。同様の理由で、web 検索と天気予報ドメインでは不自然会話検出は難易度の高いタスクとなる。

表 7 に実験結果を示す。In は In-domain 版を、Out は Out-of-domain 版を表す。6 条件全てで両者の差は僅かである。天気予報ドメインの評価 (全部) においてのみ統計的に有意な差 (McNemar's test: $p < 0.01$) が確認されたが、両者の適合率-再現率曲線（図 3）にも示されている通り、大きな差ではない。以上から、提案手法はドメインに依存しくい手法と考えられる。

[7] $\langle u, r, l, f \rangle$ と DB 中の feedback 数は、対象ドメインに属するものを全ドメインのものから除外した残りの数であることに注意。

評価 (対象)		正解率	適合率	再現率	F1
雑談	In	0.6988	0.7887	0.7508	0.7693
	Out	0.6812	0.7671	0.7515	0.7592
Web 検索	In	0.7934	0.2169	0.2769	0.2432
	Out	0.7887	0.2171	0.2923	0.2492
天気予報	In	0.9069	1.0	0.0375	0.0723
	Out	0.9069	1.0	0.0375	0.0723
評価 (全部)		正解率	適合率	再現率	F1
雑談	In	0.7627	0.682	0.6727	0.6773
	Out	0.7611	0.6884	0.6481	0.6676
Web 検索	In	0.7803	0.7027	0.705	0.7038
	Out	0.7868	0.7246	0.6842	0.7038
天気予報	In	0.7919	0.7307	0.6935	0.7116
	Out	0.7801	0.7032	0.7028	0.703

表 7: ドメイン非依存性実験結果

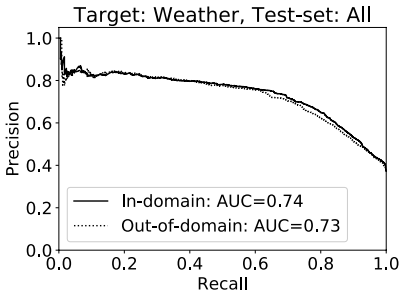

図 3: 天気予報、評価 (全部) の適合率-再現率曲線

5 おわりに

Feedback に基づく不自然な会話の検出法を提案し、実システムの大規模ログで有効性を定量的に示した。[1] でさらに詳述する。データは公開を予定している。

参考文献

[1] Chikara Hashimoto and Manabu Sassano. Detecting absurd conversations from intelligent assistant logs by exploiting user feedback utterances. In *WWW, Research track, full*, 2018.

[2] 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 *WWW*, 2015.

[3] Emiel Krahmer, Marc Swerts, Mariet Theune, and Mieke Weegels. Error detection in spoken human-machine interaction. *International Journal of Speech Technology*, 2001.

[4] Ryan Lowe, Michael Noseworthy, Iulian V. Serban, Nicolas Angelard-Gontier, Yoshua Bengio, and Joelle Pineau. Towards an automatic turing test: Learning to evaluate dialogue responses. In *ACL*, 2017.

[5] Shumpei Sano, Nobuhiro Kaji, and Manabu Sassano. Prediction of prospective user engagement with intelligent assistants. In *ACL*, 2016.

[6] Jason Weston. Dialog-based language learning. In *NIPS*, 2016.