雑談対話システムの評価とその問題点

Evaluation of chat-oriented dialogue systems and its problems

東中竜一郎 [1*]
Ryuichiro Higashinaka[1]

[1] 日本電信電話株式会社 NTT メディアインテリジェンス研究所
[1] NTT Media Intelligence Laboratories, NTT Corporation

Abstract: Dialogue systems can be divided into two categories: task-oriented dialogue systems that accomplish certain tasks through dialogue and non task-oriented dialogue systems or chat-oriented dialogue systems that perform casual conversation with users. As for the former, its performance can be measured by task completion measures; however, it is not trivial to perform the evaluation of the latter. This paper introduces current chat-oriented dialogue systems and describes how they have been evaluated, highlighting the difficulties of evaluation.

1 はじめに

対話システムは，ユーザとの対話を通してタスクを遂行するタスク指向型対話システムとコミュニケーション自身を目的とする非タスク指向型対話システム（雑談対話システムとも呼ぶ）に大別される [1]．特に，雑談対話システムは，Apple 社の Siri や NTT ドコモ社のしゃべってコンシェルといったスマートフォン上のパーソナルアシスタントやソフトバンク社の Pepper や Jibo といったパーソナルロボットが身近になるにつれ，システムが日常会話を行う必要が生じてきていることなどから，近年注目を集めている．しかし，タスク指向型対話システムはタスク達成率などの尺度でそのパフォーマンスを計ることができるのに対し [2]，後者は主観的要素も強く，その評価は難しい．このことが，雑談対話システムの改善サイクルを阻んでいると言える．本稿では，雑談対話システムの構成法，評価手法を説明し，現状の評価の問題点について述べる．

2 雑談対話システム

雑談対話システムを構成する手法は主に三つある．一つ目は，ルールベースによる手法で，手作業で入力に対する応答ルールを記述するものである [3]．高い質の応答を実現できる一方で，多くの話題に対応しようとすると，ルール作成にコストがかかる．二つ目は，抽出ベースの手法で，大量のテキストデータ（たとえば，新聞記事や映画のスクリプト，ツイッターのデータ）から，現在の入力の応答として相応しいものを抽出するという方法である [4]．発話の質は低いかもしれないが，比較的低いコストで多くの話題に対応できる．三つ目は，生成ベースの手法で，大量のテキストデータの中でも，会話形式（特に発話ペア）に着目し，機械翻訳で用いられる手法を用い，発話を生成モデルによって生成するというものである [5]．深層学習の進展により人気となっている手法だが，現在のところ発話の質は高くない．そのため，現在，生成ベースの手法は，抽出ベースの手法と組み合わせて用いることが多い [6, 7]．なお，我々は，発話理解部，対話管理部，発話生成部からなる構成の雑談対話システムを構築しており，発話生成部では，ルールベース・抽出ベースの両方の手法を用いている [8]．

3 雑談対話システムにおける評価

音声認識の進展は評価セット（主に，Word Error Rate）についての精度を改善していくという方法論によるところが大きい．このように，評価セットを事前に決めて，計算機に閉じてアルゴリズムを評価することをオフライン評価と言う．一方，対話のようにやり取りの内容がダイナミックに変わるものは評価セットを構築することが難しい．その場合は，実システムを構築し，ユーザにシステムを使ってもらって評価する必要が出てくる．これをオンライン評価と言う．雑談対話システムの基本的な性能（たとえば，一往復のやり取りを行う性能）については，オフライン評価が使われることが多い．しかし，複数回のやり取りの評価はオンライン評価で行うことが多い．

3.1 オフライン評価

雑談対話システムのオフライン評価では，発話選択の精度，および，発話生成の精度が評価されている．すなわち，ある入力発話について，正解であるシステム発話をあらかじめ準備しておき，正解を選択できる精度 [9, 10] や，正解と類似した発話を生成できる精度（BLEU に似た評価尺度）[11]，もしくは，正解に対する予測性能（perplexity）[12] などで評価する．なお，正解は人手

*連絡先：higashinaka.ryuichiro@lab.ntt.co.jp

で作成される場合と世の中にすでに存在する対話データを正解として用いる場合がある．また，雑談では，入力に対するシステム応答のバリエーションが多い．そのため，複数の正解を準備しておく方がよい．我々は，大規模マルチリファレンスを用いる手法を提案している [13].

雑談対話システムに関する評価型ワークショップの対話破綻検出チャレンジ [14] では，ユーザとシステムの雑談ログにおいて，対話の破綻につながるシステムの不適切な発話を検出するタスクに取り組んでいる．ここでは，対話破綻検出の精度が評価尺度となっている．

3.2 オンライン評価

雑談対話システムのオンライン評価では，実システムをユーザに使ってもらい，その主観評価を行う．主観評価はアンケートによって行うことが多い．たとえば，質問項目として,「システムとの対話は自然でしたか」や「システムの発話には多様性がありましたか」,「システムとまた話をしたいですか」などがある [8]．タスク指向型対話では，SASSI[15] のような，評価の観点を網羅したアンケートが提案されているが，雑談対話システムにおいては，標準的なものは存在しない．

なお，近年では，クラウドソーシングを用いて対話システムを評価することも多くなってきた．ユーザ発話に対するシステムの出力をクラウド上のユーザに速く・安価に主観評価してもらうといった方法である．抽出ベースの雑談対話システムに関する評価型ワークショップ NTCIR Short Text Conversation (STC) [16] では，この手法が取られている．システムが出力した発話について，入力発話に対する関連度（relevance）を複数名が評価し，情報検索の尺度である normalized gain などを算出している．

4 評価における問題点

雑談対話システムの評価における問題の一つはその主観性の高さである．あるユーザがよい思ったシステム発話が，他のユーザは低い評価を付けることがある．評価型ワークショップを運営し，その際に得られた主観評価値を分析したところ，おおよそ，ユーザの主観評価の一致率は 0.2 から 0.4 の間であり [17]，低い一致となっている．人間同士でも評価が揺れるような課題は工学的な問題としてはふさわしくないかもしれない．現在は，人間同士の評価値の高い一致を求めることはあきらめて，入力発話に対する多数のアノテータの評価値の分布そのものを正解とするのがよいのではないかと考えているが，この妥当性も不明である．

もう一つの大きな問題は，そもそも，よい雑談とは何かが分かっていないことである．これまでの評価は非常に局所的な文脈を切り取って，一番よさそうな発話を選択したり，生成したりしてその精度を測っており，どちらかと言えば，破綻の無いような自然な流れが実現できればよいとしている評価となっている．しかし，われわれは雑談によって多くのことを行っている．社会的な関係の構築 [18]，相手の属性・状態・状況の理解 [19]，思考の喚起・整理 [20]，承認欲の充足 [21] などが雑談の機能として挙げられる．この中には長期的な評価が必要なものもある．雑談が実現する機能を観察し，これらの機能についても個別に評価していく必要があるだろう．

5 おわりに

本稿では，雑談対話システムの手法を紹介するとともに，現状の評価手法およびそれらの問題点について述べた．自然性を担保するための評価を行いつつも，雑談を工学的に扱うために，雑談というものの深い理解が必要である．人間同士の対話を観察する以外にも，実際に雑談対話システムをユーザに使ってもらい，どのような機能が必要とされているのかについても分析を進めていく必要がある．

参考文献

[1] 中野幹生, 駒谷和範, 船越孝太郎, 中野有紀子, 奥村学（監修）. 対話システム. コロナ社, 2015.

[2] Marilyn A Walker, Diane J Litman, Candace A Kamm, and Alicia Abella. PARADISE: A framework for evaluating spoken dialogue agents. In *Proc. ACL*, pp. 271–280, 1997.

[3] Richard S Wallace. The anatomy of ALICE. In *Parsing the Turing Test*, pp. 181–210. Springer, 2009.

[4] Rafael E Banchs and Haizhou Li. IRIS: a chat-oriented dialogue system based on the vector space model. In *Proc. ACL (System Demonstrations)*, pp. 37–42, 2012.

[5] Oriol Vinyals and Quoc Le. A neural conversational model. In *Proc. ICML Deep Learning Workshop*, 2015.

[6] 呉先超, 伊藤和重, 飯田勝也, 坪井一菜, クライアン桃. りんな:女子高生人工知能. 言語処理学会 第 21 回年次大会発表論文集, pp. 306–309, 2015.

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

[8] Ryuichiro Higashinaka, Kenji Imamura, Toyomi Meguro, Chiaki Miyazaki, Nozomi Kobayashi,

Hiroaki Sugiyama, Toru Hirano, Toshiro Makino, and Yoshihiro Matsuo. Towards an open-domain conversational system fully based on natural language processing. In *Proc. COLING*, pp. 928–939, 2014.

[9] David DeVault, Anton Leuski, and Kenji Sagae. Toward learning and evaluation of dialogue policies with text examples. In *Proc. SIGDIAL*, pp. 39–48, 2011.

[10] Atsushi Otsuka, Toru Hirano, Chiaki Miyazaki, Ryuichiro Higashinaka, Toshiro Makino, and Yoshihiro Matsuo. Utterance selection using discourse relation filter for chat-oriented dialogue systems. In *Proc. IWSDS*, 2016.

[11] Michel Galley, Chris Brockett, Alessandro Sordoni, Yangfeng Ji, Michael Auli, Chris Quirk, Margaret Mitchell, Jianfeng Gao, and Bill Dolan. deltableu: A discriminative metric for generation tasks with intrinsically diverse targets. In *Proc. ACL*, pp. 445—450, 2015.

[12] Iulian V Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Proc. AAAI*, 2016.

[13] 杉山弘晃, 目黒豊美, 東中竜一郎. 大規模マルチリファレンスに基づく雑談対話システムの自動評価に向けた実験的検討. 人工知能学会研究会資料 SIG-SLUD-B401-01, Vol. 71, pp. 1–6, 2014.

[14] 東中竜一郎, 船越孝太郎, 小林優佳, 稲葉通将. 対話破綻検出チャレンジ. 人工知能学会研究会資料 SIG-SLUD-075-07, pp. 27–32, 2015.

[15] Kate S Hone and Robert Graham. Towards a tool for the subjective assessment of speech system interfaces (SASSI). *Natural Language Engineering*, Vol. 6, No. 3&4, pp. 287–303, 2000.

[16] Lifeng Shang, Tetsuya Sakai, Zhengdong Lu, Hang Li, Ryuichiro Higashinaka, and Yusuke Miyao. Overview of the NTCIR-12 short text conversation task. *Proc. NTCIR*, 2016.

[17] Ryuichiro Higashinaka, Kotaro Funakoshi, Masahiro Araki, Hiroshi Tsukahara, Yuka Kobayashi, and Masahiro Mizukami. Towards taxonomy of errors in chat-oriented dialogue systems. In *Proc. SIGDIAL*, pp. 87–95, 2015.

[18] Timothy Bickmore and Justine Cassell. Relational agents: a model and implementation of building user trust. In *Proc. CHI*, pp. 396–403, 2001.

[19] 平野徹, 小林のぞみ, 東中竜一郎, 牧野俊朗, 松尾義博. パーソナライズ可能な対話システムのためのユーザ情報抽出. 人工知能学会論文誌, Vol. 31, No. 1, pp. DSF–B_1–10, 2016.

[20] 前田英作, 南泰浩, 堂坂浩二. 人ロボット共生におけるコミュニケーション戦略の生成. 日本ロボット学会誌, Vol. 29, No. 10, pp. 887–890, 2011.

[21] 目黒豊美, 東中竜一郎, 堂坂浩二, 南泰浩. 聞き役対話の分析および分析に基づいた対話制御部の構築. 情報処理学会論文誌, Vol. 53, No. 12, pp. 2787–2801, 2012.