対話システムライブコンペティション2
The Dialogue System Live Competition 2

東中竜一郎[1,9*]　船越孝太郎[2]　稲葉通将[3]　角森唯子[4]　高橋哲朗[5]
赤間怜奈[6,7]　宇佐美まゆみ[8]　川端良子[8]　水上雅博[9]
Ryuichiro Higashinaka[1,9]　Kotaro Funakoshi[2]　Michimasa Inaba[3]
Yuiko Tsunomori[4]　Tetsuro Takahashi[5]　Reina Akama[6,7]
Mayumi Usami[8]　Yoshiko Kawabata[8]　Masahiro Mizukami[9]

[1] NTT メディアインテリジェンス研究所　NTT Media Intelligence Laboratories
[2] 京都大学　Kyoto University
[3] 電気通信大学　The University of Electro-Communications
[4] NTT ドコモ　NTT DOCOMO INC.
[5] (株) 富士通研究所　Fujitsu Laboratories, LTD.
[6] 東北大学　Tohoku University
[7] 理化学研究所　RIKEN Center for Advanced Intelligence Project
[8] 国立国語研究所　National Institute for Japanese Language and Linguistics
[9] NTT コミュニケーション科学基礎研究所　NTT Communication Science Laboratories

Abstract: Following the success of the first dialogue system live competition, whose aim was for the research community to understand the difficulty and limitations of human-computer dialogue in a live event, we organized "the dialogue system live competition 2". In this edition, we had two tracks; one is the open track and the other the situation track. The former aims at developping an open-domain chat-oriented dialogue system and the latter a human-like chat-oriented dialogue system in a given situation. In the preliminary round of the competition, we had nine and seven entries to the open track and the situtation track, respectively. This paper describes the design, background and procedure of the event as well as the results of the preliminary round of the competition. The final round will be held as a live event in the 10th dialogue system symposium.

1　はじめに

スマートフォン上の音声エージェントやパーソナルロボットなど，対話システムが身近になってきてはいるものの，それらの対話能力は高いとは言えない．たとえば，文脈を踏まえた発話ができなかったり，個性が一貫していなかったり，質問に答えられなかったりする．対話システムが人間の役に立つためにはまだまだ改善すべきことは多い．

対話システムの性能向上を図るために，これまでコンペティション形式の評価型ワークショップが行われてきた．コンペティションには大きく分けて二つの形式があり，一つは固定的なデータセット上での評価値を競うものである．たとえば，Dialogue State Tracking Challenge [1] や対話破綻検出チャレンジ [2] である．もう一方は，ユーザとの対話によって得られる評価値を競うものである．たとえば，Spoken Dialogue Challenge [3] や Conversational Intelligence Challenge [4] である．本稿で説明する対話システムライブコンペティション2 は後者に該当する．本イベントは，昨年度から開始した対話システムライブコンペティション（ライブコンペ）[5, 6] の第二回である．ライブコンペの特徴は，ユーザとシステムの対話を大勢のオーディエンスが一斉にライブで鑑賞・評価するところにある．これにより，対話システム関係者が現状の対話システムの問題点を共有することができ，対話システム研究が加速される．

ライブコンペ 1 では，また話したくなる雑談対話システムを作ることを目指したが，ライブコンペ 2 では，これに加え特定のシチュエーションで人間らしく話す雑談対話システムを目指すトラックも用意した．このトラックをシチュエーショントラックと呼ぶ．これに合わせて，ライブコンペ 1 のトラックはオープントラックと呼ぶ．これから対話システムがより身近になるにつれて，対話システムも人間同士のように相手との関係性やその場の状況を理解して会話することが必要となってくるだろう．そのような場合に何が課題となるかをシチュエーショントラックを通して見極めたいと考えている．

*連絡先：〒 239-0847 神奈川県横須賀市光の丘 1-1 Email: ryuichiro.higashinaka.tp@hco.ntt.co.jp

本稿では，ライブコンペ 2 のトラックのそれぞれについて，仕様，評価尺度，システムの要件を説明する．そして，ライブコンペ 2 におけるエントリ状況について述べる．今回，オーガナイザのチームを入れて，オープントラックには 9 チームのエントリがあった．シチュエーショントラックには 7 チームのエントリがあった．本稿では，これらのチームによるシステムの予選の結果について述べる．予選を通過した上位チームのシステムは，第 10 回対話システムシンポジウム内のライブイベントにおいて鑑賞・評価される．

2 ライブコンペの仕様

2.1 オープントラック

オープントラックの仕様については，ライブコンペ 1 に準じており，具体的には以下に述べるとおりである．

2.1.1 対象とする対話システム

オープントラックが対象とする対話システムは，対話内容の想定・制約がない「非タスク指向型対話システム（雑談対話システム）」である．タスク指向型対話システムを対象にしてタスクを固定すると，タスクによって組織間で参加しやすさに違いがでてしまうこと，まだ基礎研究段階にある雑談対話のほうが企業・大学の垣根を越えて参加しやすいこと，そして，タスク指向であっても雑談要素が重要であることから，非タスク指向型対話システムを選んだ．

2.1.2 評価尺度

雑談対話システムの目的は複合的であるが，中でも，会話を続けること・社会的関係を築くことは重要な目的である．そこで，評価尺度として,「どれくらいまた話したいと思うか」という 1 つの評価軸を用いる．これは，対話が面白かったか，役に立ったか，自然だったか，などの様々な観点を包含したものとして設定した．評価は 5 段階のリッカート尺度で行う．

なお，評価においては，対話の相手がシステムであることはあらかじめ通知されることとする．つまり，ローブナー賞に代表されるチューリングテストのように，相手が人間と見分けが付かないかという観点は用いない．あくまでも，雑談対話システムとしての良さを評価する．

2.1.3 対話システムの要件

複数のシステムを同じ条件で評価するために，オープントラックにおける対話システムの要件は以下の通りとした．

1. システムの 1 ターンあたり 1 発話（一つの吹き出し）とする．
2. 発話内容はテキストのみとし，絵文字・顔文字・スタンプは使用しない．
3. 発話内には改行を含めない．
4. アイコン画像やプロフィールは使用しない．
5. システムは 16 発話以上継続できるものとする．
6. ユーザからの 15 発話目を受け取りシステムが 16 発話目を発話した後に，評価のための識別子としてタイムスタンプやボットの名前から成る文字列を出力して対話を終了する．

要件 1 は，不特定多数の評価者が評価を行なう際に，評価者の発話を入れるタイミングを固定し，安定した評価を行なうためである．また，要件 2–4 は，対話以外の要素に評価が左右されることを防ぐためである．なお，対話システムの入力となるユーザ発話についても，評価を行うクラウドワーカーには上記 1–3 に従って対話するよう指示し，ユーザからの発話もテキストのみが渡される前提とした．

対話はユーザからの`/start`というメッセージにより開始されることとする．このメッセージを受けた後にシステムは最初のメッセージをユーザに送信する．

2.2 シチュエーショントラック

シチュエーショントラックは今回新設されたトラックで，特定のシチュエーションで人間らしく話す雑談対話システムを目指すものである．目的の違いから，オープントラックとは評価尺度や要件にいくつかの違いがある．これらの違いを中心に，以下にシチュエーショントラックの仕様について述べる．

2.2.1 対象とする対話システム

シチュエーショントラックが対象とする対話システムは，オープントラックと同様「非タスク指向型対話システム」である．オープントラックと異なる点は，話者の属性，話者間の関係，そして対話のきっかけなどのシチュエーションがあらかじめ設定されている点である．今回設定したシチュエーションは図 1 のとおりである．システムは「田中」として「ところで，これまで行ったところで一番印象に残った場所ってどこ？」から対話を開始し,「鈴木」とのやり取りを行う．

今回設定したシチュエーションにおいては，話者は男性同士または女性同士としており，ライブコンペの参加者にはあらかじめ性別を決めてもらった．またそれに応じて，評価するクラウドワーカーにもシステムと同じ性別の設定として対話するよう指示を出した．このようなシチュエーションを設定することで，一般的な雑談対話システムにおいてあまり検討しないような項目の考慮が要求される．たとえば，敬語を使って話すことはこのシチュエーションではあまりないだろうし，会社員だからこその話題というものもあると考えられる．なお，ここに書かれていないシチュエーションについては任意に想定してよいとした．

システム	名前:田中アイ (女)/アキラ (男), 年齢:20〜30 代, 職業:会社員
ユーザ	名前:鈴木ユウコ (女)/ユウキ (男), 年齢:20〜30 代, 職業:会社員
話者の関係	同性同士, 学生時代の友人関係
場所・時間	自宅, 暇な時間
話題	一番印象に残った旅行・場所

田中と鈴木は，学生時代，仲の良い友人同士であった．2 人とも大学を卒業して会社員になってからはときどき食事に行ったりしていたが， ここ 2，3 年は会う機会も連絡をとることもなくなっていた．ある日，田中が自宅でのんびり過ごしていると，鈴木からテキストメッセージが送られて来た．鈴木も家で暇にしていたらしく，ふと気になって連絡をくれたらしい． 久しぶりにお互いの近況報告をする中で，最近出かけた場所などが話題になった．田中「ところで，これまで行ったところで一番印象に残った場所ってどこ？」

図 1: シチュエーショントラックにおいて設定したシチュエーション

2.2.2 評価尺度

シチュエーショントラックにおいて対話システムは,「どれくらい (シチュエーションに適した) 人らしい会話か」という 1 つの評価軸を用いる．評価者は対話が自然だったか／システムが対話相手として疲れないか，などの様々な観点を考慮して評価することを想定している．評価は 5 段階のリッカート尺度で行う．なお，オープントラックと同様，対話の相手がシステムであることはあらかじめ通知されることとする．

2.2.3 対話システムの要件

シチュエーショントラックにおける対話システムの要件は 2.1.3 節に書かれた要件を基準としながら，要件 2 についてのみ以下の変更を加えた．

2. 発話内容はテキストに加え，Telegram で利用可能な絵文字・顔文字とする．また，テキストを含まない画像および STICKERS（スタンプ）の利用は不可とする．

変更の理由は，人間らしい対話ということを考えた場合，絵文字や顔文字による表現が果たす役割も大きいと考えたからである．

2.3 トラック共通の仕様

ここでは，オープントラックとシチュエーショントラックに共通した仕様について述べる．これらについては昨年度同様であるので簡単に述べる．詳しくは昨年度の予稿 [5] を参照されたい．

2.3.1 プラットフォーム

ライブコンペに参加する対話システムは，インスタントメッセージシステム Telegram[1]でボットとして動作する必要がある．ライブコンペ参加者は各自のサーバで対話システムを立ち上げ，対話システムは，Telegram プラットフォーム上でユーザと対話を行う．

2.3.2 予選と本選

ライブコンペ 1 と同様，ライブコンペ 2 においても予選と本選がある．予選において好成績を収めた対話システムがライブイベント（本選）に進出する．予選と本選は以下の通り実施する．

予選 予選は，クラウドソーシングを用いて評価を行う．クラウドワーカーに Telegram 上で対話システムと対話をしてもらい，トラックごとに決められた評価尺度によって評価する．信頼性のある評価を行うため，今回は最大 50 人のクラウドワーカーにより評価する．なお，クラウドソーシングによる評価の前に，疎通に問題ないか，最低限の対話ができるかなどを確認するためのスクリーニングを，オーガナイザと数名のクラウドワーカーにより実施する．本スクリーニングを通過しなかったシステムは，その時点で評価の対象外とする．

ライブイベント（本選） 予選で好成績を収めたシステムが，ライブイベントに参加できる．ライブイベントでは，リアルタイムでシステムとユーザが対話し，その状況を対話システムシンポジウムの参加者全員で鑑賞・評価する．評価の基準は予選と同じとする．その後，開発者が対話システムについて説明し，システムの挙動について質疑応答を行う．

2.3.3 スケジュール

本ライブコンペのスケジュールは以下のとおりであった．なお，システムトラブルやオーガナイザと参加者間の調整などで，スケジュールが後ろにずれ込んだケースもあったが概ねスケジュール通り進行した．

- 2019 年 7 月：アナウンス
- 2019 年 10 月 1 日：エントリ締切
- 2019 年 10 月 5–11 日：スクリーニングと予選
- 2019 年 10 月 14 日：予選の結果通知
- 2019 年 10 月 18 日：発表申し込み締切
- 2019 年 11 月 1 日：原稿締切
- 2019 年 12 月 2 日：ライブイベント（本選）

[1] https://telegram.org/

2.3.4　情報公開について
予選を通過した参加者（チーム）については，所属組織を公表する．その他の組織については，原則非公開とし，公開を希望した場合のみ公開する．対話ログについては原則非公開とする．ただし，本選における対話ログは公開される．なお，公開を希望したチームの対話ログについては一般公開する予定である．

3　予選

予選の結果を説明する前に，まずオーガナイザが準備したベースラインシステムについて述べ，それからオープントラック，シチュエーショントラックのそれぞれについてエントリ状況および予選の結果について述べる．そして，予選の結果から，現在の到達点および有効な手法について考察する．

3.1　ベースライン

オープントラックとシチュエーショントラックのそれぞれでオーガナイザが一つずつベースラインシステムを準備した．

3.1.1　オープントラック
オープントラックのベースラインシステムは昨年度と同じ IRS である．IRS は，IR-STATUS[7] に準拠した用例ベースの雑談対話システムであり，ユーザ発話に対し，最も類似した入力部を持つ用例を検索し，その応答部を用いて発話する．用例間の類似度の計算には，全文検索ライブラリ Lucene[2]をデフォルト設定で用いている．形態素解析器には Lucene に付属している JapaneseAnalyzer を用いており，使用した用例は人間同士のテキストチャットデータから抽出した 26,972 用例ペアである．

3.1.2　シチュエーショントラック
シチュエーショントラックのベースラインシステムは Repl-AI[3]を用いて構築した，ルールベース型の雑談対話システムである．Repl-AI は直感的なインタフェースで対話シナリオを構築できるブラウザベースのツールである．話者は男性同士を想定している．ルールはすべて任意の入力を表すワイルドカード（*）に対して作成しており，ユーザ発話に依存せず予め準備された次の発話を行う．ルールに使用したシステム発話はすべて人手で作成し，合計で 16 発話用意した．このベースラインシステムの対話例を図 2 に示す．

3.2　オープントラック

3.2.1　エントリ
本トラックには，ベースラインを含めると，9 チームのエントリがあった．エントリしたチームとその手法については表 1 と表 2 を参照されたい．

システム:　なるほどな．僕はサッカー部だったけど，鈴木は何部だっけ？
ユーザ:　全然帰宅部 …
システム:　そうだっけ？記憶が…鈴木は体重とか気にしてる？僕は最近青汁とか飲んでるわ…
ユーザ:　おじさん体形にならないように気を付けてるよ。
システム:　そかそか．てかさー，またみんなで集まりたいよな．同級生の誰かにあったりしてる？
ユーザ:　最近は会ってないなぁ。

図 2: シチュエーショントラックのベースラインシステムの対話例

3.2.2　クラウドソーシングによる評価
エントリしたすべてのチームが疎通確認をクリアし，クラウドソーシングによる評価に進んだ．

クラウドソーシングサービスには CrowdWorks[4]を利用した．クラウドワーカーは，事前の選別や個別の依頼等はせず，不特定多数から募集した．ただし，タスク承認率（過去に取り組んだタスクの中で，タスク依頼者から成果を承認され支払いを受けた割合のこと）が 95% を超える作業者に限定した．作業単価は，Telegram のインストールなどもあることから，1 件 300 円（税別）とした．

対話システムは，Telegram アプリのボットアカウントとして稼働する．そこで，ボットアカウント毎にクラウドソーシングのタスクを作成して作業者を募り，対話と評価を依頼した．各作業者は，1 つのボットアカウントの評価は一度しか行えないが，複数のボットアカウントに渡って評価を行うことはできた．ただし，その際も他のボットとの相対評価ではなく，「当該チャットボットとしか話したことがない」という認識のもと絶対評価を行うように指示をしている．

前述の通り，評価の観点は「このチャットボットとまた話したいと思いますか？」の 1 項目のみで，回答は 1.「まったくそう思わない」2.「そう思わない」，3.「どちらとも言えない」，4.「そう思う」，5.「とてもそう思う」の 5 つから選択させた．この 1 から 5 の選択肢の番号を評定値として利用した．従って，優れたシステムほど評定値の平均値が大きくなる．他に，自由回答でタスクあるいはチャットボットについての感想を求めた．対話ログは，Telegram のログを export する機能を用いて保存してもらい，ファイル添付により提出してもらった．

予選では，50 名を上限として作業者を募集した．無効な評価結果を除外したあと，有効評価者数はチームによって 34 名から 50 名まで開きがあったが，1 チームを除いて 40 名以上の評価者数を得ており，最小の場合

[2] http://lucene.apache.org/
[3] https://repl-ai.jp/
[4] https://crowdworks.jp/

表 1: オープントラックの予選結果．スコアは最大 50 人のクラウド評価者の平均値．

順位	スコア	組織名	チーム名	Telegram ボット名
1T	4.20	東京工業大学	TokoChanTeam	Toko
1T	4.20	電気通信大学	UEC	Tripia
3	3.84	NTT コミュニケーション科学基礎研究所	NTTCS	tripfreak
4T	3.15	電気通信大学	UEC	NoneJupiter
4T	3.15	ライブコンペオーガナイザ	IRS	IRS
6	2.96	Anonymous	TEAM1	Anonymous
7	2.33	NAIST AHC-Lab.	KyoshiroS	KyoshiroS
8T	2.00	京都工芸繊維大学インタラクティブ知能研究室	MMI2019	Newsbot
8T	2.00	Anonymous	TEAM2	Anonymous

表 2: 各システムの手法（オープントラック）．方式はルールベース，抽出ベース，生成ベースの中から，知識源は，大規模テキストデータの利用，知識ベースの利用，対話データの利用の中から該当するものを参加チームが選択した．学習手法については，機械学習を用いていない場合は N/A としている．

順位	チーム名	方式			知識源			学習手法
		ルール	抽出	生成	テキスト	KB	対話	
1T	TokoChanTeam	✓				✓		N/A
1T	UEC (Tripia)	✓						N/A
3	NTTCS	✓				✓		SVM
4T	UEC (NoneJupiter)		✓		✓			BERT
4T	IRS		✓				✓	N/A
6	TEAM1	✓	✓					N/A
7	KyoshiroS			✓				RNN
8T	MMI2019	✓	✓	✓			✓	RNN
8T	TEAM2		✓	✓			✓	RNN,SVM

でも，昨年度の最大評価者数 30 名よりも多く確保した．

3.2.3 結果

オープントラックの予選評価結果は表 1 に示す通りである．各チームの評価スコアは，評価者の評定値の平均値である．順位付けは小数点 2 位までで実施した．各チームのスコアを箱ひげ図にしたものを図 3 に示す．チーム間のスコアの開きも勘案して，上位 3 チームを予選通過とした．今回昨年度 1 位であった tripfreak が 3 位となる結果であった．今年度 1 位タイの Tripia は，完全にルールベースで知識源も用いておらず，ユーザの話も全く聞かないという特異な設定であったにも関わらず，高いスコアを獲得した．

ノンパラメトリックな多重比較の手法である Steel-Dwass の手法によってシステム間の評定値を比較した結果は以下の通りであった．

```
TokoChanTeam > UEC (NoneJupiter)**, IRS**,
    TEAM1**, KyoshiroS**, MMI2019**,
    TEAM2**
UEC (Tripia) > UEC (NoneJupiter)**, IRS**,
    TEAM1**, KyoshiroS**, MMI2019**, TEAM2**
NTTCS > IRS+, TEAM1*, KyoshiroS**,
    MMI2019**, TEAM2**
IRS > KyoshiroS+, MMI2019**, TEAM2**
UEC (NoneJupiter) > KyoshiroS+, MMI2019**,
    TEAM2**
TEAM1 > MMI2019*, TEAM2**
```

ここで，> の左側のチームのシステムは右側のチームのシステムよりも統計的に優れていることを示している．**，*，+はそれぞれ，p<0.01，p<0.05，p<0.1（有意傾向）を示す．

上位 3 チームが以降のチームよりも良いことが分かるが，IRS と有意差が見られるチームは UEC (Tripia) と TokoChanTeam の二つであり，NTTCS よりも若干優れていると評価された．

各システムの挙動を概観するために，各チームの総対話数，システム発話およびユーザ発話について，1 発話あたりの平均文字数および単語数，ならびに，同一発話者による発話全体での異なり語の割合を算出した．算出した発話に関する基本統計量は表 3 の通りである．なお，総対話数とは各チームのシステムがユーザ（評価者）と一連の対話をおこなった回数であり，有効評価者数に等しい．また，単語数の算出に際しては形態素解析器 MeCab[5]（バージョン 0.996）を用いた．

予選通過チームのシステム概要（著者による説明に基づく）は以下のとおりである．

TokoChanTeam: 処理を行ったユーザー入力と，スクレイピングによって収集したインターネット上

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

表 3: システム発話およびユーザ発話の統計量（オープントラック）．文字数，単語数は発話ごとの平均値．

順位	チーム名	総対話数	システム発話			ユーザ発話		
			文字数	単語数	異なり語の割合	文字数	単語数	異なり語の割合
1T	TokoChanTeam	45	35.76	20.66	2.58 %	16.59	9.35	14.92 %
1T	UEC (Tripia)	49	49.63	30.13	0.78 %	18.38	10.83	11.46 %
3	NTTCS	44	46.92	28.19	2.73 %	17.60	10.42	13.69 %
4T	UEC (NoneJupiter)	48	28.38	16.66	16.34 %	15.44	9.23	20.26 %
4T	IRS	47	29.51	17.68	14.77 %	15.14	9.03	18.74 %
6	TEAM1	50	20.40	11.87	12.91 %	15.55	9.11	17.27 %
7	KyoshiroS	49	8.50	4.12	21.65 %	11.85	6.82	17.30 %
8T	MMI2019	34	13.37	7.05	10.59 %	13.69	8.08	17.64 %
8T	TEAM2	48	18.65	11.17	10.75 %	13.20	8.00	16.48 %

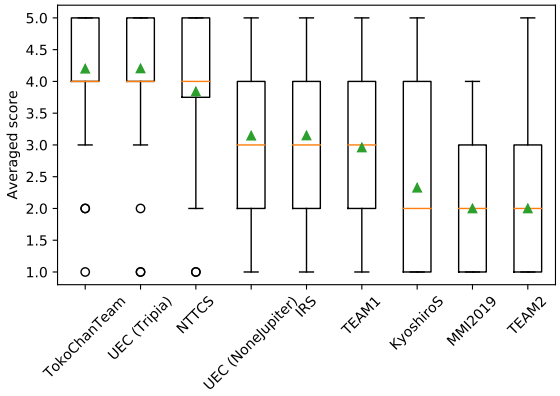

図 3: 各チームのスコア（オープントラック）

の情報を用いて，FSM に基づいたルールベースシステムにより，手書きのテンプレートから出力を生成する．

UEC (Tripia): 人手で作成したスクリプト中の発話を決められた順番で出力する．ユーザの発話は一切考慮しない．

NTTCS: 対話が自然な流れとなるよう対話フローを Agenda 的に定義・管理し，併せて発話理解・生成に必要な外部知識を導入したシステム．昨年度のバージョンから，簡易な質問応答機能のみ追加している．

これらの手法の詳細についてはそれぞれの文献 [8, 9, 10] を参照されたい．

3.3 シチュエーショントラック

3.3.1 エントリ

本トラックにはオーガナイザによるベースラインを含めて，7 チームのエントリがあった．ただ，これらのチームのうち，一つのシステムについては疎通確認ができなかったため，6 チームがクラウドソーシングによる評価に進んだ．各チームが用いた手法については，後段の表 5 にまとめている．

3.3.2 クラウドソーシングによる評価

シチュエーショントラックにおいては，評価の観点として「どれくらい (シチュエーションに適した) 人らしい会話でしたか？」という観点を用いた．その他の評価方法は 3.2.2 節で述べたオープントラックの評価方法と同じ方法を用いた．

予選では，50 名を上限として作業者を募集した．参加チームのうち 1 チームのシステムは予選評価中に応答ができなくなったため，結果的に 18 名による評価となった．評価者数が大きく異なるため他チームとの比較においては注意が必要である．他の 5 チームにおける有効評価者数は 44 名から 50 名まで開きがあったが，いずれも 40 名以上の評価者数を得ていた．

3.3.3 結果

シチュエーショントラックの予選評価結果は表 4 に示す通りである．各チームの評価スコアは，評価者の評定値の平均値である．各チームのスコアを箱ひげ図にしたものを図 4 に示す．チーム間のスコアの開きを勘案して，上位 3 チームを予選通過とした．

特定のシチュエーションが設定されていることもあり，どのチームもルールが用いられていた．

Steel-Dwass の手法によってシステム間の評定値を比較した結果は以下の通りであった．

```
OUHRI > nksw*
```

有意差が見られたのは，OUHRI と nksw の間のみであった．オープントラックと比べ，システム間に差がほとんど見られなかった．全体的にスコアが 3.5 を超えており，オープントラックと比べて下限が高い．詳細な分析が必要であるが，特定のシチュエーションを設定することで発話の解釈性が増した可能性がある．

3.3.4 予選通過システムの概要

予選通過チームのシステム概要（著者の説明に基づく）は以下のとおりである．

表 4: シチュエーショントラックの予選結果．スコアは最大 50 人のクラウド評価者の平均値．

順位	スコア	組織名	チーム名	Telegram ボット名
1	4.10	大阪大学駒谷研究室	OUHRI	OHBot
2	4.02	ライブコンペオーガナイザ	LiveCompetition2019	LiveCompetition2019
3	3.96	NTT コミュニケーション科学基礎研究所	NTTCS	ArmeriaVulgaris
4	3.66	東京工芸大学大学院	HSSLAB	RiskyPolitenessBot
5	3.55	筑波大学	110	meshimeshi
6	3.52	明海大学	nskw	nskw

表 5: 各システムの手法（シチュエーショントラック）．方式はルールベース，抽出ベース，生成ベースの中から，知識源は，大規模テキストデータの利用，知識ベースの利用，対話データの利用の中から該当するものを参加チームが選択した．学習手法については，機械学習を用いていない場合は N/A としている．

順位	チーム名	方式			知識源			学習手法
		ルール	抽出	生成	テキスト	KB	対話	
1	OUHRI	✓						CRF, ロジスティック回帰
2	LiveCompetition2019	✓						N/A
3	NTTCS	✓				✓		Transformer
4	HSSLAB	✓						N/A
5	110	✓						N/A
6	nskw	✓						N/A

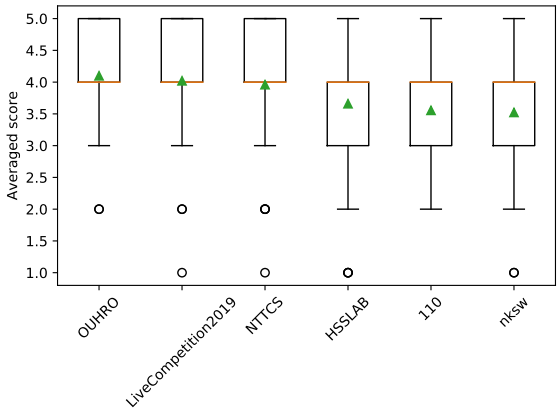

図 4: 各チームのスコア（シチュエーショントラック）

OUHRI: ネットワークモデルに基づく対話制御を行う．対話のフローを状態とその間の遷移として人手で記述し，ユーザ発話の言語理解結果に応じて条件分岐を行うシステム．

LiveCompetition2019: ユーザ発話に依存せず予め準備された次の発話を行うルールベースのシステム（3.1.2 節参照）．

NTTCS: ユーザの発話から 5W1H に該当するフレーズをユーザの体験情報として抽出し，ユーザの体験に類似するシステムの知識や体験を用いて，共感や質問，自己開示を行う．

これらの手法の詳細についてはそれぞれの文献 [11, 12] を参照されたい．

3.4 考察

オープントラックは，昨年度好成績を収めた NTTCS よりもスコアが高かったシステムが 2 つあり，それぞれルールと知識ベース（KB）を用いているものであった．中位のシステムは抽出ベースのもの，低位のシステムは生成モデルを用いているシステムが中心となっている．このことは，雑談対話システムにおいては，いまだルールによって実装される人間の知見が重要であることを示している．一問一答では深層学習のモデルの対話適用が広くなされているが，文脈を踏まえた長い対話となってくるとまだ対応が困難であると考えられる．

発話の統計情報を見ると分かる通り，上位 3 チームのシステムの異なり語の割合が極めて低い．いずれも 3%以下である．また，これらのシステムの発話は他のシステムに比べて長い．発話内容のバリエーションが少なくとも，意味のある発話を対話の流れに応じて行うことが重要ということが見て取れる．

シチュエーショントラックは，すべてのシステムがルールに基づくものであった．OUHRI はルールに加えて機械学習も併用しており，それが有効に働いたようである．対話の統計量について言えば，ルールに基づいていることもあり，異なり語の割合はいずれも低くなっている．また，一発話の長さは比較的長い傾向にある．ルールに基づくシステムでは，対話を円滑にするための知見がルール中に埋め込まれている．これら

表 6: システム発話およびユーザ発話の統計量（シチュエーショントラック）．文字数，単語数は発話ごとの平均値．

順位	チーム名	総対話数	システム発話			ユーザ発話		
			文字数	単語数	異なり語の割合	文字数	単語数	異なり語の割合
1	OUHRI	50	41.57	24.69	1.38 %	16.15	9.61	17.17 %
2	LiveCompetition2019	48	36.88	23.31	0.92 %	13.76	8.54	16.62 %
3	NTTCS	50	36.60	23.10	4.36 %	15.65	9.47	17.35 %
4	HSSLAB	50	63.55	38.80	0.89 %	16.14	9.95	14.70 %
5	110	18	40.36	25.08	2.46 %	16.78	10.38	23.20 %
6	nskw	44	25.02	15.47	2.01 %	12.79	7.87	16.07 %

のルールを分析していくことで，特定のシチュエーションにおいて有効な知見が得られることが期待できる．

オープントラックとシチュエーショントラックを通して，ルールに基づくシステムが上位を占めていることは示唆的である．15 ターンの対話を円滑に進めるためには，人間による対話の設計が重要ということであろう．しかし，深層学習に代表される End2End の手法はデータがあればシステムが構築できるという利点がある [13]．今後は，ルールに含まれる人間の知見を機械学習に取り込んでいくことが重要だと考えられる．

4 おわりに

本稿では，対話システムライブコンペティション 2 について述べた．オープントラックとシチュエーショントラックを実施し，オープントラックには 9 チーム，シチュエーショントラックには 7 チームのエントリがあった．それぞれのトラックで，上位 3 チームが予選を通過し，本選のライブイベントに進んだ．本イベントによって，現状の対話システムの問題点が共有され，今後の対話システム研究の進展に寄与できればと考えている．ぜひ第三回も行いたい．

謝辞

本イベントの実施にあたっては人工知能学会より特別補助をいただきました．また，タイトなスケジュールにもかかわらずエントリいただいた参加チームの皆様にも感謝いたします．シチュエーショントラックについては，参加を促すために講習会を実施しました．講習会においてご講演いただいた飯尾尊優氏に感謝いたします．また，講習会に参加し，シチュエーショントラックへの参加を検討いただいた諸氏にも感謝いたします．

参考文献

[1] Jason Williams, Antoine Raux, Deepak Ramachandran, and Alan Black. The dialog state tracking challenge. In *Proc. SIGDIAL*, pp. 404–413, 2013.

[2] Ryuichiro Higashinaka, Kotaro Funakoshi, Yuka Kobayashi, and Michimasa Inaba. The dialogue breakdown detection challenge: Task description, datasets, and evaluation metrics. In *Proc. LREC*, pp. 3146–3150, 2016.

[3] Alan W. Black, Susanne Burger, Alistair Conkie, Helen Hastie, Simon Keizer, Oliver Lemon, Nicolas Merigaud, Gabriel Parent, Gabriel Schubiner, Blaise Thomson, Jason D. Williams, Kai Yu, Steve Young, and Maxine Eskenazi. Spoken dialog challenge 2010: Comparison of live and control test results. In *Proc. SIGDIAL*, pp. 2–7, 2011.

[4] Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan Lowe, Shrimai Prabhumoye, Alan W Black, Alexander Rudnicky, Jason Williams, Joelle Pineau, Mikhail Burtsev, and Jason Weston. The second conversational intelligence challenge (ConvAI2). *arXiv preprint arXiv:1902.00098*, 2019.

[5] 東中竜一郎, 船越孝太郎, 稲葉通将, 角森唯子, 高橋哲朗, 赤間怜奈. 対話システムライブコンペティション. 第 84 回人工知能学会言語・音声理解と対話処理研究会 (第 9 回対話システムシンポジウム), pp. 106–111, 2018.

[6] Ryuichiro Higashinaka, Kotaro Funakoshi, Michimasa Inaba, Yuiko Tsunomori, Tetsuro Takahashi, and Reina Akama. Dialogue system live competition: identifying problems with dialogue systems through live event. In *Proc. IWSDS*, 2019.

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

[8] Dolça Tellols, 小河晴菜. ドメイン知識を用いてユーザを楽しませるルールベース雑談対話システム TokoChanBot. 第 87 回人工知能学会言語・音声理解と対話処理研究会 (第 10 回対話システムシンポジウム), 2019.

[9] 稲葉通将. 雑談対話システムをどう評価すべきか—TripiaBot のライブコンペ予選通過から考える—. 第 87 回人工知能学会言語・音声理解と対話処理研究会 (第 10 回対話システムシンポジウム), 2019.

[10] 杉山弘晃, 成松宏美, 水上雅博, 有本庸浩. 自然な流れに沿って対話を進めるアジェンダベース雑談対話システム. 第 87 回人工知能学会言語・音声理解と対話処理研究会 (第 10 回対話システムシンポジウム), 2019.

[11] 中島圭祐, 駒谷和範, 中野幹生. 雑談対話システム構築フレームワーク pychat に基づく特定シチュエーション向け対話システム. 第 87 回人工知能学会言語・音声理解と対話処理研究会 (第 10 回対話システムシンポジウム), 2019.

[12] 成松宏美, 杉山弘晃, 水上雅博, 有本庸浩. 自らの体験に基づき雑談する対話システム. 第 87 回人工知能学会言語・音声理解と対話処理研究会 (第 10 回対話システムシンポジウム), 2019.

[13] 東中竜一郎. 最近の対話システム事情：深層学習・データセット・コンペティションの観点から. 映像情報メディア学会誌, Vol. 73, No. 2, pp. 271–276, 2019.