

対話システムライブコンペティション

The Dialogue System Live Competition

東中竜一郎^{1*} 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
Ryuichiro Higashinaka¹ Kotaro Funakoshi² Michimasa Inaba³
Yuiko Tsunomori⁴ Tetsuro Takahashi⁵ Reina Akama⁶

¹ NTT メディアインテリジェンス研究所 NTT Media Intelligence Laboratories

² (株) ホンダ・リサーチ・インスティテュート・ジャパン Honda Research Institute Japan Co., Ltd.

³ 広島市立大学 Hiroshima City University

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

Abstract: We organized an evaluation workshop entitled “the dialogue system live competition” in which developers and users can share the difficulty and interestingness of human-computer dialogue in a live event. In the preliminary round of the competition, eleven systems were submitted and evaluated. 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 9th 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 AI スピーカーが一般的に用いられるようになってきた。タスク指向型対話のみならず、非タスク指向型対話（いわゆる雑談対話）についても、現在、活発に研究が進められている。

要素技術を改善するために、従来から、評価型ワークショップの枠組みが用いられてきた。所定のデータをオーガナイザが提供し、それらについて参加者がアルゴリズムの良さを競い合うことで、分野の活性化および技術の発展を目指すものである。タスク指向型対話システムについて言えば、Dialog State Tracking Challenge (DSTC) [1] が有名である。DSTC では、対話ログをもとに、システムが理解すべきスロットを推定するタスクが取り組まれた。非タスク指向型対話システムについて言えば、対話破綻検出チャレンジ [2, 3] や NTCIR Short Text Conversation (STC) [4] が近年行われている。前者は、対話ログについて、対話破綻につながるシステム発話を特定するタスクであり、後者は、入力発話に対するシステムの応答発話を、大規模な発話集合から抽出・生成するタスクである。

これらの評価型ワークショップは、基本的に固定の対話コーパスに対してアルゴリズムを競う形式である。そのため、この営みによって、対話システム全体が良くなっているかが不明であるという問題点がある。固

定的なコーパスで対話システムの特定のモジュールの性能が改善されても、必ずしも「対話システム全体」の性能が改善されるとは限らない。

一方で、対話システム全体を対象とするチャレンジもなされてきた。タスク指向型対話システムについては、DiaLeague [5] や Spoken Dialogue Challenge [6]、非タスク指向型対話システムについては、ローブナー賞、Amazon による Alexa Prize¹ や Conversational Intelligence Challenge² などが開催されてきた。

本稿で説明する「対話システムライブコンペティション（以降ライブコンペ）」は、対話システム全体を評価する評価型ワークショップに分類されるが、従来のものとは以下の点において異なる：

- 学会内において、対話システムとユーザの対話をイベント参加者がライブで鑑賞し、評価する点
- 開発者がイベント内で、対話ログを参照しながら対話システムの挙動について説明し、質疑応答を行う点

このような設計にしたのは、現在の対話システムの抱える問題点をコミュニティ全体で共有するためである。実際に対話システムの一挙手一投足を参加者全員で鑑賞し、現在のシステムは何ができて、何ができないのかといったことを共有することによって、対話システムの問題点に対する理解が深まり、本分野の研究が加速すると考えた。また、対話システムではライブ性が

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

¹<https://developer.amazon.com/alexaprize>

²<http://convai.io/>

極めて重要である。そのライブ性を開発者だけでなく、コミュニティ全体で体験することが、本ライブコンペの趣旨である。

本稿では、ライブコンペの仕様や評価手順などを述べるとともに、参加したチームの、クラウドソーシングを用いて行った予選の結果について述べる。予選を通過したチームについては、研究会内のライブイベント（本選）にて評価が行われる。本稿は本選の前に執筆していることから、本選については触れられないが、結果や本選の対話ログについてはライブコンペのホームページ³にて公開する予定である。

2 ライブコンペの仕様

ライブコンペの目的はライブイベントにおいて、対話システムとユーザの対話を鑑賞・評価し、今何が難しいのかという問題意識を共有することであるが、そのためには、現状の対話システムの中でも高性能なものを用いる必要がある。よって、複数の対話システムを集め、その中でよいものを選定する手続きが要る。また、複数の対話システムを適切に評価するための枠組みも必要となる。ここでは、ライブコンペの仕様（全体設計）について述べる。

2.1 対象とする対話システム

ライブコンペの対象とする対話システムは「非タスク指向型対話システム（雑談対話システム）」とした。タスク指向型対話システムを対象にすると、そのタスクに関わっている組織は参加しやすいが、そうではない組織にとっては参加がしづらくなってしまうこと、雑談対話はまだ基礎研究段階であり、企業、大学の垣根を越えて参加しやすいこと、そして、タスク指向型対話システムであっても社会的な関係を持続させるために雑談の要素が重要であることなどから決定した。

2.2 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。これは抽象的な評価軸であり、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。各参加者が雑談対話システムを構築する目的はさまざまであるため、特定の目的に偏り過ぎないようにした。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つ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。

2.3 対話システムの要件

同じ条件で対話システムを評価するために、ライブコンペで用い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。

2.4 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 Telegram⁴でボットとして動作する必要がある。

TelegramはTelegram Messenger LLPが開発・提供するインスタントメッセージシステムでありConversational Intelligence Challengeでも用いられているプラットフォームである。利用者間でのチャットができるほか、ボットを接続するためのAPIを備えており、このAPIを用いることで独自で開発した対話システムをTelegram上で動作させることができる。ユーザは、ボットのusernameを指定することで、任意のTelegramクライアントからそのボットに接続し、チャットが可能である。

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。対話システムとTelegram

³<https://dialog-system-live-competition.github.io/dslc1/index.html>

⁴<https://telegram.org/>

サーバの間の通信のためには Python, Java, Ruby など様々な言語で実装されたライブラリが GitHub で公開されており、これらを使って対話システムを容易に開発することができる。本ライブコンペでは、サンプルプログラムとして現在の対話ログからランダムに発言を選択し送信するボットを用意し参加者に提供した。

2.5 予選と本選

ライブコンペにエントリーした対話システムについて、ライブイベントで用いるシステムを選定するために予選を行う。予選において好成績を取めた対話システムがライブイベント（本選）に進出する。以下、予選と本選について説明する。

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してもらい、所定の評価尺度（どれくらいまた話したいと思うか）によって評価する。雑談対話システムの評価は主観的要素が強いため、少数のワーカーで評価することは適切ではない。そこで、本ライブコンペでは、各対話システムについて多数のクラウドワーカーを用いる（今回は最大で 30 名を用いた）。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） 予選で好成績を取めたシステムが、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。

2.6 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る。

- 2018 年 6 月：ライブコンペのアナウンス
- 2018 年 9 月 30 日：エントリー締切
- 2018 年 10 月 1–10 日：スクリーニングと予選
- 2018 年 10 月 12 日：予選の結果通知
- 2018 年 10 月 26 日：予稿締切
- 2018 年 11 月 21 日：ライブイベント（本選）

2.7 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（opt-in policy）。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、本ライブコンペのオーガナイザによって、これらに対話破綻アノテーションを行った上で、一般公開する予定である。

3 予選

ここでは予選とその結果について述べる。また、予選の結果から、現在の雑談対話システムの到達点および有効な手法について考察する。

3.1 エントリー

本ライブコンペには 12 チームのエントリーがあった。エントリーしたチームは企業、大学など様々である。これらのチームおよびこれらのチームが用いた手法については、後段の表 2 にまとめてあるので参照されたい。

なお、エントリーしたシステムのうち一つはオーガナイザによって準備した IRS と呼ばれるものである。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、ユーザ発言に対し、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言する。用例間の類似度の計算には、全文検索ライブラリ Lucene⁵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。

3.2 クラウドソーシングによる評価

エントリーした 12 チームのうち、1 チームについては、疎通確認において、システムの動作確認ができなかったため、最終的に 11 チームに対し予選を実施した。

クラウドソーシングサービスには CrowdWorks⁶を利用した。評価者（作業員）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% を超える作業員に限定した。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員を募り、対話と評価を依頼した。各作業員は、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

⁵<http://lucene.apache.org/>

⁶<https://crowdworks.jp/>

表 1: 予選の結果. スコアは最大 30 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	1.97	株式会社 NTT ドコモ	NTTdocomo	MarikoZatsudanBot
2	2.10	NTT コミュニケーション科学基礎研究所	NTTCS	tripfreak
3	2.27	東北大学 大学院情報科学研究科	teamzunko	zunko
4	2.57	ライブコンペオーガナイザ	IRS	IRS
5	2.72	Anonymous	TEAM1	Anonymous
6	2.73	Anonymous	TEAM2	Anonymous
8	2.75	Anonymous	TEAM3	Anonymous
7	2.76	Anonymous	TEAM4	Anonymous
9	3.31	早稲田大学	RSL	momokoBot
10	4.10	Anonymous	TEAM5	Anonymous
11	4.77	Anonymous	TEAM6	Anonymous

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せんか？」の 1 項目のみで、回答は 1.「とてもそう思う」、2.「そう思う」、3.「どちらとも言えない」、4.「そう思わない」、5.「まったくそう思わない」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小さくなる。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、20 名以上の評価者の利用を想定していたが、実際には規定に従わない作業、例えば、システムの応答を待たずに 2 回連続で発話をする、不正に編集したログを提出するなど、を行う者がいる可能性があったため、余裕を見て対話システム毎に 30 名の作業者を募集した。このように 11 チームに対して各 30 名、延べ 330 名から評価を集めたが、対話ログを確認して問題があると判断したものを除外した結果、有効な評価者数は延べ 318 名（異なりでは 55 名）であった。チーム毎の有効評価者数は、一番少ないチームで 26 名（1 チーム）、一番多いチームで 30 名（4 チーム）であった。

3.3 結果

表 1 に予選の結果を示す。3 位と 4 位の間のスコアに開きがあることから、3 位までを予選通過者とした。表中では、予選通過チーム、および、組織名の公表を希望したチームのみ、その情報を掲載している。組織名の公開を希望していないチームについては、スコア順で TEAM1-6 を割り振っている。各チームのスコアを箱ひげ図にしたものを図 1 に示す。

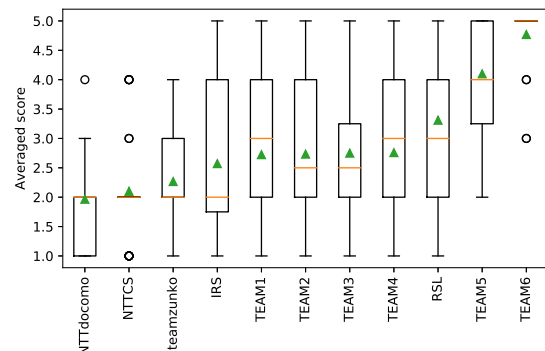


図 1: 各チームのスコア (箱ひげ図)

また、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。

```

NTTdocomo > TEAM3+, RSL**, TEAM5**,
TEAM6**
NTTCS > RSL*, TEAM5**, TEAM6**
teamzunko > RSL*, TEAM5**, TEAM6**
IRS > TEAM5**, TEAM6**
TEAM1 > TEAM5**, TEAM6**
TEAM2 > TEAM5**, TEAM6**
TEAM3 > TEAM5**, TEAM6**
TEAM4 > TEAM5**, TEAM6**
RSL > TEAM6**
TEAM5 > TEAM6+
    
```

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。

RSL に対し、統計的な差が見られたのは予選を通過した上位 3 チームのみである。また、NTTdocomo に

ついでのみ、TEAM3に対して有意傾向が見られた。上位チーム（1から3位）と下位チーム（9から11位）には大きな差が見られるが、今回の評価においては、上位チームと中位チームの間には有意差は見られなかった。

各チームが利用している手法について、その特徴をまとめたものを表2に示す。また、予選通過チームのシステム概要は以下のとおりである。

NTTdocomo 対話履歴から焦点と対話行為を推定し、大規模発話データベースとルールを併用して発話するオープンドメイン雑談対話システム。

NTTCS 対話の流れを人手で定義したフローに基づいて制御しつつ、半自動的に構築した発話知識を対話の内容として組み合わせることで、文脈に沿った発話理解・生成を行う。

teamzunko 3種類の知識（Wikipediaから抽出されたKB、早押しQAデータセット、Webからのスクレイピングによるダジャレデータ）を用いたルールベース発話と、「東北ずん子」という既存のキャラクターの人格に沿って発話を行う人格依存のルールベース発話をメインの応答に用い、ルールベースで対応できない部分は直前の発話を考慮した seq2seq 応答を返す。

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。

3.4 考察

今回の評価におけるランキングと各システムの方式および知識源を比較すると、方式については、チームをランキング上の上位、中位、下位に分類した際に、上位のものはルールと生成の組合せ、中位のものはルールと抽出の組合せ、下位のものはルールまたは生成の単独利用が特徴のようである。ルールベースの手法を用いつつも、それ以外の知識源をうまく組み合わせることが重要であると言える。また、上位にナレッジベースを用いたシステムがランクインしていることからナレッジベースが有効に働いていることが分かる。学習手法としては、どの手法が効果的であるかは明確ではないが、必ずしもニューラルベースの手法がよいとは言えない結果である。

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る。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 MeCab⁷（バージョン 0.996）を用いた。システムの発話数はシステムによって大きく異なるが、一般に長い方が上位のようである。ただし、異なり語が多ければよいわけでもない。画一的な（パターンの）応答であっても、適切に利用することで満足度の高い雑談が可能であることが分かる。

なお、上位チームのいくつかは、旅行や料理など、ある程度ドメインを限定している点も特徴的である。これにより、無理な話題の飛躍を避けることができると考えられる。その反面、中位以降のチームでは、ドメインを限定しない難しさからか、文脈を無視したようなシステム発話も多く見受けられた。

4 本選

予選を通過した3チームについては、ライブイベント（本選）において、対話システムシンポジウムの参加者の前で評価が行われる。具体的には、以下の流れでライブイベントは実施される。

1. 予選を通過した各対話システムと審査員が Telegram を用いて対話する。審査員はそれぞれの対話システムと1度ずつ話し、その後、これをもう一度繰り返す。つまり、各対話システムは2回ずつ審査員と対話する。対話の順番は事前にくじ引きにより決定する。審査員はオーガナイザが利害関係のない者から選定する。対話の様子は予選と同一とする。システムは、対話システムシンポジウムの参加者全員によって評価される。評価の基準は予選と同様とする。
2. 予選通過チームは、自身のシステムについてプレゼンテーションを行う。そして、対話ログについて、なぜそのようにシステムが応答したかについて説明し、参加者との質疑応答を行う。
3. 対話システムシンポジウムの参加者全員による評価の結果により、最終的な順位を決定する。

5 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について述べた。12チームのエントリがあり、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。ぜひ第二回も行いたい。

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。

⁷<http://taku910.github.io/mecab/>

表 2: 各システムの手法. 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	NTTdocomo	✓		✓	✓			Other
2	NTTCS	✓			✓	✓	✓	CRF, SVM
3	teamzunko	✓		✓	✓	✓	✓	RNN
4	IRS		✓				✓	N/A
5	TEAM1	✓	✓				✓	N/A
6	TEAM2	✓		✓	✓			CNN, RNN
7	TEAM3	✓	✓		✓		✓	CNN
8	TEAM4	✓	✓				✓	RNN
9	RSL	✓						N/A
10	TEAM5			✓			✓	RNN
11	TEAM6			✓			✓	Transformer

表 3: システム発話およびユーザ発話の統計量. 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	NTTdocomo	29	22.99	13.46	18.96 %	15.51	9.40	23.86 %
2	NTTCS	30	45.34	27.16	3.39 %	15.84	9.41	16.81 %
3	teamzunko	30	40.78	25.05	10.43 %	13.09	7.94	21.25 %
4	IRS	28	30.30	18.08	19.09 %	16.97	10.06	22.56 %
5	TEAM1	29	32.27	18.65	14.89 %	13.59	8.05	21.77 %
6	TEAM2	30	17.62	10.09	6.98 %	14.78	8.70	19.76 %
7	TEAM3	28	15.23	8.03	21.46 %	15.34	8.97	23.18 %
8	TEAM4	29	13.96	8.25	21.07 %	16.89	10.13	20.63 %
9	RSL	29	57.44	30.70	3.42 %	11.59	6.31	18.71 %
10	TEAM5	30	9.82	6.23	1.46 %	11.50	6.68	20.81 %
11	TEAM6	26	4.36	1.93	1.24 %	11.59	7.12	1.94 %

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] 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. 第 75 回言語・音声理解と対話処理研究会 (第 6 回対話システムシンポジウム), 人工知能学会研究会資料 SIG-SLUD-75-B502, pp. 27–32, 2015.
- [3] 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.
- [4] Lifeng Shang, Tetsuya Sakai, Zhengdong Lu, Hang Li, Ryuichiro Higashinaka, and Yusuke Miyao. Overview of the NTCIR-12 short text conversation task. In *Proc. NTCIR*, 2016.
- [5] 橋田浩一, 伝康晴, 長尾確, 柏岡秀紀, 酒井桂一, 島津明, 中野幹生. DiaLeague: 自然言語処理システムの総合評価. 人工知能学会誌, Vol. 12, No. 3, pp. 390–399, 1997.
- [6] Alan W Black and Maxine Eskenazi. The spoken dialogue challenge. In *Proc. SIGDIAL*, pp. 337–340, 2009.
- [7] Alan Ritter, Colin Cherry, and William B. Dolan. Data-driven response generation in social media. In *Proc. EMNLP*, pp. 583–593, 2011.
- [8] 角森唯子, 大西可奈子, 藤本拓, 角野公亮, 吉村健, 磯田佳徳. カスタマイズ可能なオープンドメイン雑談対話エンジンの開発. 人工知能学会 言語・音声理解と対話処理研究会第 84 回 (第 9 回対話システムシンポジウム), 2018.
- [9] 杉山弘晃, 成松宏美, 水上雅博, 有本庸浩. 文脈に沿った発話理解・生成を行うドメイン特化型雑談対話システムの実験的検討. 人工知能学会 言語・音声理解と対話処理研究会第 84 回 (第 9 回対話システムシンポジウム), 2018.
- [10] 阿部香央莉, 佐藤志貴, 佐藤拓海, 藤井諒, 松田耕史, 鈴木正敏, 山口健史, 赤間怜奈, 大内啓樹, 鈴木潤, 乾健太郎. Zunkobot: 複数の知識モジュールを統合した雑談対話システム. 人工知能学会 言語・音声理解と対話処理研究会第 84 回 (第 9 回対話システムシンポジウム), 2018.