

2 コーパス

2.1 Pre-train

Transformer に基づく言語処理は、大量のテキストで Pre-train することで高い性能を発揮する。文分類タスクや機械翻訳等では、収集の容易さから、構造を持たない平文の denoising タスク（ノイズを付与した文からの復元問題）を用いることが一般的である。一方 BlenderBot では、より対話に特化した Pre-train として、大量の対話形式のデータ（Reddit のデータを加工した入力文脈・ターゲット発話のペア：入力文脈 560 億トークン・ターゲット発話 888 億トークン）を用い、直接入力文脈とターゲット発話の関係を学習している。

本研究では BlenderBot を踏襲し、対話形式のデータである Twitter のリプライペアを大量に収集し Pre-train に用いる。まず、ランダムにサンプリングした日本語ユーザの、2016 年 1 月から 2018 年 3 月までの全ツイートを取得し、以下に合致するものを取り除いた後、全角化、アカウント名等の除去、絵文字の除去のクリーニング処理を行う。

- 同日にコサイン類似度 0.9 以上の別ツイートが存在するツイート（20 文字以下はフィルタしない）
- URL・括弧等を含むツイート
- ユーザが Bot であるツイート
- リツイート
- ひらがな・カタカナの割合が 30% 以下のツイート

次に、クリーニングされたツイートから、リプライ関係にあるツイートを抜き出し、入力文脈・ターゲット発話のペアとする。この際、リプライの連鎖となっているものについては、連鎖のルートツイートから 1 つずつ鎖を伸ばし、最後のツイートをターゲット発話、それまでのツイート列を入力文脈として複数のペアを作成する。例えば、A-B-C-D の鎖となっている場合、A-B, AB-C, ABC-D を利用し、B-C 等は用いない。これにより得られたペアの総数は 21 億ペア（521GB）であった。入力文脈の長さの平均^{*1}は 2.91、文字数の平均は入力文脈で 62.3 文字、ターゲット発話で 20.3 文字であった。学習時は、このデータを Sentencepiece を用いてトークンへ分割して用いる。Sentencepiece のモデルは教えてgoo! のデータからサンプリングした 2000 万文で構築し、構築時のパラメータには語彙数 32000、文字カバー率 99.995% を設定する。

2.2 Fine-tune

BlenderBot の Fine-tune では、話者性・知識・共感という、雑談対話に必須と考えられるスキル

を獲得させるため、各スキルに対応する 3 種類のコーパス（話者性：Persona-chat[Zhang 18]、知識：Wizard-of-wikipedia[Dinan 18]、共感：Empathetic Dialogues[Rashkin 19]）を用いて学習している。また、これらのスキル間の使い分けを学習するため、スキル統合用の Blended Skill Talk コーパス [Smith 20] を加え、まとめて Fine-tune を行っている。

本研究の Fine-tune では、BlenderBot の各スキル学習用コーパスの日本語版コーパスを作成して用いる。Blended Skill Talk コーパスは、上記スキルを学習したモデルの出力を人が編集する形で収集した対話コーパスであり、学習済みのモデルがない場合には構築することができない。本研究ではオリジナルと同じ方法では Blended Skill Talk コーパスを構築できなかったため、その代用として、上記スキルが用いられていると想定される雑談コーパスを 3 種類用意し、スキル統合用コーパスとして用いる。以下に各コーパスの詳細を述べる。

2.2.1 スキル学習用コーパス

■Wizard-of-wikipedia Wizard-of-wikipedia [Dinan 18] は、好奇心旺盛な apprentice 役と、相手に情報を与える wizard 役による、Wikipedia を知識源とする雑談で、22311 対話が収録されている。Wizard 役のみがシステムが提示した Wikipedia ページを参照できる。対話は、どちらかの話者がトピックを選択して相手に話しかけることで開始され、最低 4、5 ターン以上続けた後、いずれかの話者の任意で対話を終了する。話者は Wikipedia の文章をそのままコピーしたような発話はできないが、それ以外の制限はない。

我々が先行研究の収集システムを用いて対話したところ、日本人にはなじみのない話題が多く、提示される Wikipedia の記事も発話の参考として使いにくいものが多かった。そのため本研究では、先行研究の英語対話データを基に、日本人同士の対話として自然な形になるよう翻訳および編集し、学習に用いる。

Wizard-of-wikipedia の対話データのうち、日本人の会話に出てこない話題や、略称を含む対話、複雑な対話になりやすい数値情報を含む対話を除外した約 2 万対話を対象に、シナリオライターが作例を行った。本研究ではその一部にあたる、1886 対話（17007 発話）を用いる。

■Persona-chat Persona-chat[Zhang 18] は、話者の特徴を規定するプロフィール文を 5 文程度のセットで設定し、各話者が与えられたプロフィール文セットに従って対話を行うことで、擬似的に様々な話者の対話を収集したコーパスである。収録対話数は 10907 対話である。本研究では、Persona-chat の日本語版コーパスを構築するため、プロフィール文セットの日本語化と、日本語話者による対話の収集を行う。

プロフィール文セットの日本語化は、先行研究の英語

*1 全体の 0.12% のデータを利用して計算

プロフィール文を日本語に翻訳し、出身地や学校名、商品名などの名詞を日本に関連する内容に修正したものを、5 文程度組み合わせることで実現している。

日本語話者による対話の収集では、日本語版プロフィール文セットから 399 セットを抜き出し、クラウドワークに割り当て対話を収集する。各ワークは自身に与えられたプロフィール文セットの内容に触れるように雑談を行う。発話はそれぞれが交互に 1 発話ずつ行うこととし、1 発話は最大で 100 文字、全体で最低 12 発話、最高 15 発話となるように対話を収集した。本研究では、7 名の日本語母語話者が対話した 200 対話を学習に用いる。

■**Empathetic Dialogues** Empathetic Dialogues は、感情的な状況について発話する「話し手」と、それに応答する「聞き手」の 2 者による、共感的な対話を収集した英語対話コーパスである [Rashkin 19]。先行研究では、クラウドソーシングを用いて、32 種類の感情を示す単語（英語）について、話し手がその感情を抱く状況の説明文と対話を 24,850 対話収集している。本研究では、Empathetic Dialogues の日本語版コーパスを作るため、感情を示す 32 種類の単語（英語）を日本語に翻訳し、日本語話者が対話を作例する。並行して、先行研究の英語版コーパスを日本語に翻訳する作業も行った。

日本語話者による作例作業には、クラウドソーシングを用いる。収集コストを抑える観点から、1 つの対話は 2 人の作業者による実際の対話ではなく、1 人の作業者による作例としている。作業者は、翻訳された感情のリストを参照し、その感情に沿った 1-3 文の状況文と、その状況化で対話する「話し手」と「聞き手」の 2 者による 4-8 発話のテキスト対話を作成する。本研究では作例 200 対話、翻訳 200 対話の計 400 対話を学習に用いる。

2.2.2 スキル統合用雑談コーパス

■**聞き役対話コーパス** 対話相手の話を積極的に聞き、話し手に満足を与えることを目的とする、聞き役対話を収集したコーパスである [Meguro 10]。対話のフォーマットが固定的な通常のカウンセリングや傾聴よりも自然な雑談に近い特徴を持つ。10 人の聞き役と 37 人の話し役が収集に参加し、聞き役と話し役が 1 人ずつペアで対話する形式で収集されている。対話数は 1260 対話、対話ごとの平均発話数は 28.2 発話である。

■**初対面雑談コーパス** 初対面の話者が最大 16 回繰り返し対話することによる、対話の変化を調べる目的で収集されたコーパスである [Higashinaka 14]。対話数は 3600 対話、対話ごとの平均発話数は 36.9 発話である。

■**趣味雑談コーパス** 多様な人の、趣味についての雑談を高密度に収集したコーパスである。話者の組み合わせによる対話への影響を調べるため、80 名の各実験参加者が、それぞれ他の 60 名以上の参加者と対話するという

設定で収集した。学習に用いた対話数は 3483 対話、対話ごとの平均発話数は 60.9 発話である。

3 手法

3.1 モデルアーキテクチャ

BlenderBot の構築において複数のモデルアーキテクチャを比較したところ、シンプルな Transformer encoder-decoder モデルが最もよい性能であったことが報告されている [Roller 20]。これを踏まえ、本研究でも同一のアーキテクチャを用いる。

3.2 候補文のフィルタリング

上記モデルを後述 (3.3 節) の方法で学習し生成された文を予備的に分析したところ、過去発話の繰り返し、誤ったキーワードの組み合わせ、過去発話との矛盾の、3 種類の不自然な文が多数生成されていた。抑制が難しい過去発話との矛盾以外について、不自然な文をフィルタするため、以下のフィルタルールを定義し、発話候補 u の生成スコア (平均 Perplexity) s_u にペナルティスコアを付加する形でフィルタリングを行った。

■**過去発話繰り返し** 文脈内の発話と同一内容の発話を抑制するためのフィルタである。文脈中の発話およびそれを文に分割したものと、発話候補 u の文間類似度が閾値 σ_r を上回った場合、ペナルティスコア c_r を付加する。

■**誤ったキーワードの組み合わせ** これは、「東野圭吾の『ハリーポッター』が好きです。」のように、誤ったキーワードの組み合わせが含まれる発話を抑制するためのフィルタである。本研究では、キーワードペアの共起の妥当性を調べるため、Wikipedia の概要部分を利用する。具体的には、一方を発話候補 u に含まれるキーワード、他方を直近の文脈に含まれるキーワードとする全てのペアについて、片方のキーワードの概要部分に他方のキーワードが含まれている場合は、正しいキーワードペアであると考え、確認済みリストにペア双方のキーワードを追加する。全てのキーワードペアをチェックし、確認済みリストに含まれなかった発話候補 u 内のキーワードの個数に応じてペナルティスコア c_{kw} を付加する。一方、確認済みリスト内のキーワードが多いほど、より話題を広げる発話であることが期待されるため、リワードスコア c_{kc} を付加した。

3.3 モデル学習の詳細

モデルの実装・学習には、Pytorch ベースのニューラル系列モデリングライブラリである、Fairseq^{*2}を利用した。学習を行う GPU クラスタとして、産総研 ABCI クラウド^{*3}を利用した。また、パラメータ探索には、Weight

^{*2} <https://github.com/pytorch/fairseq>

^{*3} <https://abci.ai/ja/>

and Biases を利用した [Biewald 20].

3.3.1 Pre-train

Pre-train では、2.1節で構築した Twitter リプライベアコーパスを利用する。Encoder への入力形式は、各ペアの入力文脈中の発話を [SEP] トークンで繋げる形式とし、それ以外の情報は付加しない。

モデルパラメータは Meena, BlenderBot の 27 億パラメータを参考に、ABCI クラウドの GPU (V100 16GB) でメモリエラーを起こさないよう隠れ層の次元数を調整し、Encoder 2 層, Decoder 24 層, 各隠れ層 1920 次元, Attention head 数 32 とした。このときのモデルサイズは 16 億パラメータであり、GPT-2[Radford 19] とほぼ同程度であった。また、フィードフォワード層および Attention の dropout は実験的に 0.1 とした。学習の Optimizer にはメモリ効率が良い Adafactor[Shazeer 18] を利用し、学習率は $1e-3$, warmup step は 10000, 1 ステップあたりの処理トークン数は 4.8M トークンとした。目的関数には初期の学習を促すため label smoothed cross entropy を設定した。使用した計算資源は V100 16GB 400 枚で、validation loss がおおよそ平坦になる 30000 ステップまで、おおよそ 28 時間学習を行った。

3.3.2 Fine-tune

Fine-tune では、2.2節で説明した各コーパスをすべて混合して学習した。このときのペア数は 220881 であった。Encoder への入力形式は、各コーパスの違いや趣味雑談の話者性、対話中のターンの進行を表現するため、表 1 のテンプレートに沿って情報を埋め込んだ。実際の入力例 (トークナイズ後) を表 2 に示す。発話履歴は最大 128 文字となるまで、最大 4 発話入れ込む。対話種別は、コーパスに応じて趣味雑談・初対面雑談 (初対面雑談 + Persona-chat) ・聞き役対話・共感雑談 (Empathetic Dialogues) ・知識対話 (Wizard-of-wikipedia) と、終了の挨拶を行うための終了処理の 6 種類とした。

対話種別: [SEP] 話者 ID [SEP] [SPK1] システム発話 _{$t-2$} [SEP] [SPK2] ユーザ発話 _{$t-1$} [SEP] 経過ターン数

表1 Encoder への入力テンプレート

趣味 雑談: [SEP] id au [SEP] [SPK2] カフェ も 好きです [SEP] [SPK1] どういった ものを見られますか? [SEP] [SPK2] ミュージカルとかお芝居やお笑いです。劇団四季や宝塚やよしもと漫才劇場もいきます [SEP] ターン 04

表2 実際の入力例 (トークナイズ後)

学習の Optimizer には Pre-train 同様 Adafactor を利用し、学習率は $1e-4$, warmup step は 3000, 1 ステップあたりの処理ペア数は 2048 ペア (おおよそ 200K トークン) とする。この設定で、V100 16GB 128 枚でおおよそ 1 時間ほど (300 ステップ程度) 訓練し、validation set の perplexity が最小となるモデルを利用する。本研究では、validation set には、training set 中のコーパスではなく、Pre-train に用いた Twitter リプライベアを

利用する。検討当初は趣味雑談コーパスや training set と同様に混合したデータを validation set に利用していたが、読書の話題で常に同じ作者を出すなど、話題ごとに選ばれる内容が固定されやすく、ドメインへの過学習が疑われたため、より話題の幅が広いと考えられる Twitter リプライベアを validation set に用いている。趣味雑談を validation set としたときは上記 300 ステップでおおよそ収束が見られた一方、Twitter リプライベアを validation set としたときは 85 ステップで収束が止まり、その後 validation loss が増加していた。

3.4 デコード設定

モデルからの発話のデコードについて、本研究では Meena[Adiwardana 20] 同様に、Sample-and-rank 形式を採用する。これは、サンプリングによって独立に生成された N 個の発話候補のうち、最も低い perplexity の候補を最終出力とする方法である。当初の検討では、BlenderBot 同様の diverse beam search を利用していたが、やや無難な (つまらない) 応答が出力されるケースが多く、Sample-and-rank 形式のほうが話題を広げる上で有利と判断した。本研究では、softmax を計算する際に温度 T でトークンの出力確率 z_i を除算する、Hinton らの方法 [Hinton 15] を採用する。温度 $T = 1$ で通常のサンプリングになり、温度 T を大きくするほど、より文脈的に珍しいトークン (固有名詞等) が生成されるようになるが、同時に文脈的に誤ったトークンも生成されやすくなる。逆に温度 T が小さくなるほど、無難で一般的な単語が選ばれやすくなる。

3.5 対話システムライブコンペティション 3

本研究で構築したモデルを、ライブコンペ 3 のオープントラックへ投稿した。詳細なパラメータを以下に示す。

デコード (3.4節) 温度 $T = 1.2$, 発話候補数 $N = 50$.
過去発話繰り返し (3.2節) 閾値 $\sigma_r = 0.65$, ペナルティスコア $c_r = 0.25$.

誤ったキーワードの組み合わせ (3.2節) ペナルティスコア $c_{kw} = 0.3$, リワードスコア $c_{kc} = 0.3$.

対話に用いる話者 ID は、趣味雑談コーパス中でユーザと最も趣味が類似していると予想される話者の ID とする。話者間の趣味の類似度計算には、趣味雑談コーパスで収集した各話者の自己申告趣味の文字列と、ユーザの 1 発話目との文間類似度を用いる。

各対話の最初のシステム発話は「こんにちは。よろしくお願ひします。早速ですけど、何か趣味ってありますか?」で固定し、終了ターンの 1 ターン前のシステム発話に「あ、すみません。そろそろ時間ですね。今日はありがとうございました。」を付加するとともに、対話種別を終了処理に設定した。

System こんにちは。よろしくお願ひします。早速ですけど、何か趣味ってありますか？

User こんにちは。水曜どうでしょうを見るのが趣味です。

System 水曜どうでしょう！大好きです。どのシリーズが好きですか？

User 対決列島が好きです。

System 対決列島は名作ですね！僕はヨーロッパリベンジが好きです。好きな回とかありますか？

User 安田さんがリバースする回は何度見ても爆笑してしまいます。

System 安田さん面白いですよ！大泉洋とのコンビも好きです。

User そうですね。あの2人のトークは最高です。

System 大泉洋の笑い声につられて僕も笑っちゃいます

User ですよ。大泉洋さんはとても面白いですね。

System 趣味は旅行とありますが、最近はこちらに行かれましたか？

User 最近東京に行きました。

System 東京ですか！いいですね！行ってみたいですが東京は何が有名なんですか？

User 東京は浅草が有名ですね。ところで女優の渡辺麻友さんを知っていますか？

System そうなんですね！渡辺麻友さんは存じ上げないです…

User AKB出身の元アイドルでとても可愛いんですよ。

System akbの人だったんですね！初耳です！

User 話題作のドラマにもたくさん出ていましたが最近引退されてしまいました…

System そうなんですね！最近どのドラマに出てたんですか？

User 朝ドラのなつぞらに出られていました。

System 朝ドラにも出てたんですね！知らなかったです！

User 是非見てみてください。きっと気に入りますよ。

System なつぞらですね！今度見てみます！

User なつぞらもいいですし、戦う！書店ガールも面白いのでおすすめです。

System 書店ガール、タイトルだけ聞いたことあります！どんな内容なんですか？

User 書店員が奮闘するドラマですね。

System なるほど！気になりますね！時間があるときに見てみます！

User 是非見てみてください。

System はい！是非見てみます！どんなジャンルの漫画をよく読まれるんですか？あ、すみません。そろそろ時間ですね。今日はありがとうございました。

User ギャグ漫画をよく読みます。ありがとうございました。

System いえいえ、こちらこそありがとうございました。ではまた明日。

表3 評価の高い対話（話題：水曜どうでしょう・渡辺麻友。自然性5，話題追隨5，話題提供4）

System こんにちは。よろしくお願ひします。早速ですけど、何か趣味ってありますか？

User 山歩きかなー山の日にも山行っただし

System 山歩きですか。いいですね。どんな山に行かれたんですか？

User 地元の山だよ。そっちは山の日今年は何したの？

System いいですね。私の趣味は読書とウォーキングです。

User 山の日にもやっぱり読書かウォーキングを？

System 読書はミステリー小説が好きで、主に東野圭吾です。ウォーキングは山の近くまで電車で行って、山頂から歩くのが好きです。

User いいねー山の日に関する本とか何か知らない？

System 最近ミステリー小説は読まないのですが、去年は東野圭吾の容疑者xの献身という本にハマってました。

User それは山の日と関係が？

System 「本」でお探ひですか？「東野圭吾」の「容疑者xの献身」ですか？

User 違うけど…まあいいや、山行っただし動物見かけることってあるよね

System そうなんですね。容疑者xの献身はミステリー小説です。

User 自分はこないだリスをたくさん見かけてさ

System 「天空の城ラピュタ」も映画化されてますね。

User そういえばラピュタにもリスような動物出てきたよね、ロボットと仲良しの。あんな感じ

System 「天空の城ラピュタ」の続編は「ルパン三世」でした。

User ルパンで動物はあまり馴染みないかな…好きな動物とかいるの

System 一番好きなのは「トトロ」です。ジブリ作品は「もののけ姫」が好きです。

User トトロは動物か微妙だけど、まあ動物っぽいよねー。もののけ姫も動物いっぱい出てくるしね

System 「千と千尋の神隠し」は見ましたか？

User 何か動物は出てくるの？

System 出てきますよ。千と千尋の神隠しに出てくる「アシタカ」が大好きなんです。

User アシタカってもののけ姫のなんかシカっぽい動物に乗ってる人じゃなかったっけ？

System それです！そのアシタカです！

User でもそれじゃ千と千尋の神隠しに出てくる動物って何なの？

System アシタカは千と千尋の神隠しの登場人物です。

User さっき違っって言ったじゃない。動物は出てこないの？

System え？さっきのなんでしたっけ？ジブリで1番好きなのはハウルの動く城です。あ、すみません。そろそろ時間ですね。今日はありがとうございました。

User 動物にはあんまり注目してないんだね…まあこちらこそありがとうございました

System そうですね。またお話ししましょう。お疲れさまでした。

表4 評価の低い対話（話題：山の日・動物。自然性2，話題追隨2，話題提供4）

4 結果

4.1 対話システムライブコンペティション3

ライブコンペ3のオープントラックでは、ユーザはオーガナイザが指定した名詞リストから話題となる語を2つ選択し、それについて15ターン対話を行うよう指示された。評価は自然性・話題追隨・話題提供の3つが用いられ、5段階のLikert尺度で付与された。本システムは自然性4.03、話題追隨3.87、話題展開3.58で、オープントラック予選を1位で通過した。

ライブコンペ3でのユーザとの対話のうち、高評価・低評価の2対話を表3、表4に示す。評価の高い対話では、ユーザの提示した話題に適切に追隨し、可能な範囲でシステム自身の知識に基づいて話題を展開させている。一方評価の低い対話では、そもそもユーザの提示した話題を適切に捉えることができず、また適切に話題を転換できずに文脈を引きずりすぎてしまい、応答に失敗していた。検証として、本話題を文脈に入れない状態で「山行っただし動物見かけることってあるよね」と発話すると、システムは「そうですね。へびとかトカゲとかはよく見ますね。」のように適切に応答できていたことから、文脈の管理が課題の一つであることがわかる。

対話へのコメントについて分析すると、コメントを記述した16名のうち15名が、「ほとんど違和感が感じら

れず、驚きました。中の人と話してるとしか思えなかったです。」のような評価の高いコメントをしていた。一方、5名が「少し矛盾があった」や「1回だけ変な応答があった」というコメントをしており、全体としては自然なもの、不自然な発話が含まれていたことがわかる。

4.2 自己対話例

表5、表6に、構築したモデル同士（学習データに含まれない組）の対話の例を示す。モデル同士でも自然に対話を展開することができており、コンピュータ囲碁等のように、対話でも自己学習を行える可能性がある。

また、話者IDを切り替えることで、対話の展開が変化することが示された。しかしながら、どちらも旅行の話題が選ばれており、本来の話者間の対話と比較するとバリエーションに乏しい。単純な話者IDの指定のみでは、話者性の表現には不十分であることが示唆される。

4.3 誤りタイプの分析

表7に、対話中の不自然な発話について、誤りタイプの分類と以下の3段階の破綻度を示す。B1：わずかに違和感があるが、ほぼ気にならない程度。B2：明らかに違和感を感じ、発話内容・意図を疑問に思う程度（焼き物好きで有田焼知らないなど）。B3：強く違和感を感じ、対話継続に支障が出る程度（「佐賀は長崎です」など）。また、破綻度がB2以上の対話の評価値も合わせて示す。

Sys-CC こんにちは。よろしくお願ひします。早速ですけど、何か趣味ってありますか？
Sys-CD 趣味は、旅行、食べ歩き、観劇です。ご趣味は何ですか？
Sys-CC 旅行は国内ですか？海外ですか？私も旅行好きです。
Sys-CD 最近国内が多いですね海外は行ったことないですどちらに行かれましたか？
Sys-CC 最近、国内旅行は、どちらに行かれましたか？私は、沖縄と北海道にいきました。
Sys-CD 北海道いいですね！私も沖縄は行ったことあります北海道は食べ物おいしいですよ
Sys-CC そうなんです！沖縄は、修学旅行で行きました。北海道は、食べ物美味しいですね。

表5 自己対話（話者 CC-CD）

Sys-CG こんにちは。よろしくお願ひします。早速ですけど、何か趣味ってありますか？
Sys-CH 趣味は旅行と食べ歩きです。ご趣味は何ですか？
Sys-CG 旅行は国内ですか？海外ですか？食べ歩きは私も好きです。
Sys-CH 国内も海外も行きます食べ歩きはどのあたりに行かれますか？
Sys-CG 食べ歩きは、京都と奈良です。海外は、どこに行かれましたか？
Sys-CH 私も京都と奈良です！海外はオーストラリアとニュージーランドに行きました
Sys-CG オーストラリアとニュージーランドは、どちらに行かれたんですか？

表6 自己対話（話者 CG-CH）

誤りタイプ	定義・例	B1	B2	B3	自然	追従	展開
既知	対話内に含まれている内容の質問	2	7	2	4.4	4.1	3.3
矛盾	既知質問以外の矛盾	12	23	21	3.8	3.6	3.3
話題の飛び	話題や話が不連続に遷移	11	16	21	3.4	3.3	3.7
キーワード関係誤り	キーワードの関係が誤り。例：佐賀は長崎ですか？	3	5	8	3.9	4.0	2.9
概念誤り	キーワードではないが関係が誤り。例：有田焼はどこなところ？	5	2	5	4.4	4.1	3.0
事実誤認	事実と異なる発話。例：邦画のハリーポッター	0	10	4	3.6	3.4	3.7
対象違い	システム自身の発話に対して、相手が発話したかのように話す	0	2	0	4.0	4.0	4.5
勝手	勝手に対話を終わらせようとする	2	1	0	3.0	4.0	3.0
その他	上記に含まれないが不自然な発話	17	5	0	4.0	4.0	3.8
合計		51	70	58			

表7 誤りタイプと定義、全52対話中の各誤りタイプの破綻度ごとの出現数、破綻度B2以上の平均評価値

表7より、不自然な発話は1対話（16システム発話）あたり平均3.44発話（破綻度B2以上で2.46発話）出現していた。誤りタイプ分類では、矛盾、話題の飛び、キーワード関係誤り、事実誤認が出現数として多いことがわかる。キーワード関係誤り・事実誤認はペナルティをかけたにも関わらず相当数が出現しており、単純なルールでは抑制が不十分であったと考えられる。

評価値を調べると、事実誤認や話題の飛びが起こった場合に、自然性と話題追従が大きく下がっていた。話題の飛びは話題追従と直接対応するものであり、事実誤認は話題を正しく理解できていないことが表出されたものであるため、話題追従が低下したと考えられる。なお、自然性と話題追従の spearman の相関係数は $r = 0.66$ と他の組み合わせ ($r < 0.1$) より高かった。適切に話題を捉え応答すること自体が、自然さと強く関連していることがわかる。一方、話題展開を調べると、キーワード関係誤り、概念誤りが含まれる場合に評価が低下していた。これらの誤りタイプは、新たな話題の導入や話題を深掘る質問に関する誤りであるため、話題展開ができていないと判断されたと考えられる。

5 結論

本研究では、大規模な Transformer encoder-decoder モデルを利用した、日本語雑談対話システムの構築につ

いて詳細に報告し、対話システムライブコンペティション3での結果を分析した。構築したシステムは、ライブコンペ3 オープントラックの予選を1位通過するなど、非常に自然な対話が行えていた。一方、過去発話との矛盾や、関連の薄いキーワードの共起などの問題が明らかになった。今後はこれらの問題への対処を進めるとともに、効率的なモデルの評価方法についても検討していく。

謝辞

本研究は科研費「人と社会的に共生する対話システムのための行動決定モデル基盤技術の確立」（課題番号19H05693）の支援を受けた。

参考文献

- [Adiwardana 20] Adiwardana, D., Luong, M.-T., So, D. R., Hall, J., Fiedel, N., Thoppilan, R., Yang, Z., Kulshreshtha, A., Nemade, G., Lu, Y., and Le, Q. V.: Towards a Human-like Open-Domain Chatbot (2020)
- [Biewald 20] Biewald, L.: Experiment Tracking with Weights and Biases (2020)
- [Dinan 18] Dinan, E., Roller, S., Shuster, K., Fan, A., Auli, M., and Weston, J.: Wizard of Wikipedia: Knowledge-Powered Conversational agents, in *Proceedings of Seventh International Conference on Learning Representations* (2018)
- [Higashinaka 14] Higashinaka, R., Imamura, K., Meguro, T., Miyazaki, C., Kobayashi, N., Sugiyama, H., Hirano, T., Makino, T., and Matsuo, Y.: Towards an open-domain conversational system fully based on natural language processing, in *Proceedings of the 25th International Conference on Computational Linguistics*, pp. 928–939 (2014)
- [Hinton 15] Hinton, G., Vinyals, O., and Dean, J.: Distilling the Knowledge in a Neural Network, *arXiv preprint arXiv:1503.02531* (2015)
- [Meguro 10] Meguro, T., Higashinaka, R., Minami, Y., and Dohsaka, K.: Controlling Listening-oriented Dialogue using Partially Observable Markov Decision Processes, in *Proceedings of the 23rd International Conference on Computational Linguistics*, pp. 761–769 (2010)
- [Radford 19] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I.: Language Models are Unsupervised Multitask Learners (2019)
- [Rashkin 19] Rashkin, H., Smith, E. M., Li, M., and Boureau, Y.-L.: Towards Empathetic Open-domain Conversation Models: A New Benchmark and Dataset, in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 5370–5381 (2019)
- [Roller 20] Roller, S., Dinan, E., Goyal, N., Ju, D., Williamson, M., Liu, Y., Xu, J., Ott, M., Shuster, K., Smith, E. M., Boureau, Y.-L., and Weston, J.: Recipes for building an open-domain chatbot, *Arxiv preprint arXiv:2004.13637* (2020)
- [Shazeer 18] Shazeer, N. and Stern, M.: Adafactor: Adaptive Learning Rates with Sublinear Memory Cost, *arXiv preprint arXiv:1804.04235* (2018)
- [Smith 20] Smith, E. M., Williamson, M., Shuster, K., Weston, J., and Boureau, Y.-L.: Can You Put it All Together: Evaluating Conversational Agents' Ability to Blend Skills (2020)
- [Vinyals 15] Vinyals, O. and Le, Q.: A Neural Conversational Model, in *Proceedings of ICML Deep Learning Workshop* (2015)
- [Wang 18] Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. R.: GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding, in *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 353–355 (2018)
- [Wang 19] Wang, Q., Li, B., Xiao, T., Zhu, J., Li, C., Wong, D. F., and Chao, L. S.: Learning Deep Transformer Models for Machine Translation, in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1810–1822 (2019)
- [Zhang 18] Zhang, S., Dinan, E., Urbanek, J., Szlam, A., Kiela, D., and Weston, J.: Personalizing Dialogue Agents: I have a dog, do you have pets too?, in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pp. 2204–2213 (2018)
- [杉山 20] 杉山弘晃, 成松宏美, 菊井玄一郎, 東中竜一郎, 堂坂浩二, 平博順, 南泰浩, 大和淳司: センター試験を対象とした高性能な英語ソルバーの実現, 言語処理学会第26回年次大会, pp. 1–3 (2020)