



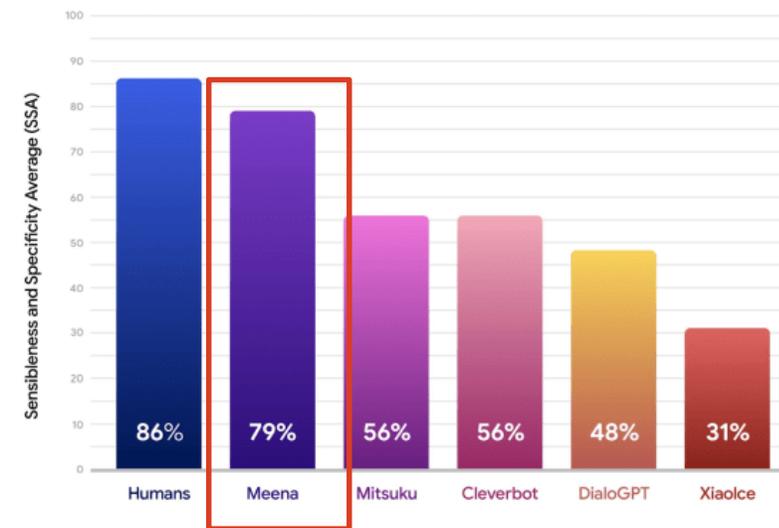
Transformer encoder-decoderモデルによる 趣味雑談システムの構築

○杉山弘晃 成松宏美 水上雅博 有本庸浩 千葉祐弥 目黒豊美 中嶋秀治
NTT コミュニケーション科学基礎研究所

2020/11/30

雑談対話システム研究の情勢

- 昨年まで：ニューラル全盛の機械翻訳等と異なり，雑談対話システム研究はルールベースがニューラルに対しても互角に戦っていた領域
 - ライブコンペ1・2ではルールベースシステムが優勢
 - ConvAI等英語圏コンペでも，ニューラルが圧倒的というほどでもなかった
- **この1年で世界が一変**
= 大規模Transformerを利用したNNベースシステムが極めて自然な応答生成を実現
- Meena (Google: 2020/02)
 - › シンプルなEncoder-decoderモデルをSNSデータで学習したもの
 - › 2.6Bパラメータを341GBのテキストデータで学習（モデル・データ非公開）
 - › 従来，最終的な品質では優位にあったルールベースシステムを，初めて明確に上回ったNNベース雑談対話システム
- BlenderBot (Facebook: 2020/04)
 - › 対話に必要な3種類のスキル（個人性・知識・共感）を専用コーパスで学習し統合するモデル
 - › 2.7B・9.4Bパラメータのモデル2種が公開（データは既存コーパス利用）
 - › 現在の英語雑談システムのState-of-the-art（Meenaに対し，同規模モデルでも有意に改善）



BlenderBot (9.4B) 対話例和訳

ユーザー
生きてる人でも死んでる人でもいいんだけど、
一番夕飯を一緒に食べてみたい人って誰？

BlenderBot
難しいなあ・・・挙げるならステーキジョブスかなあ。あの人の知恵を借りたいよ。

あ，面白いね。ステーキジョブスについて
どんなこと知ってるの？

技術の歴史において最も影響を与えた一人
だね。先見の明がある。

現状

- 言語処理全般において公開データ・モデル規模の両面で大きく出遅れ
 - › 最大のGPT2-Japanese*でも0.27Bパラメータ（モデルサイズ差10倍以上）・公開対話データもごく少量

課題

- 研究速度の低下
- 英語圏との課題レベルのギャップ

(少なくとも) BlenderBot相当の日本語モデルを作らないことには始まらない！

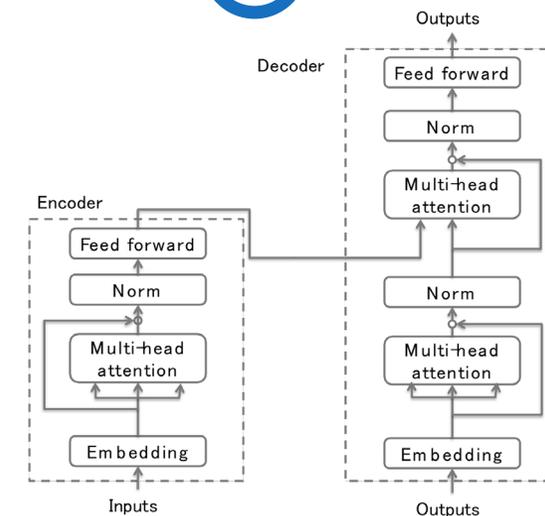
- 日本語でもそれなりに動くと予想されるものの、実際に何ができるか・問題になるかを詳細に知りたい
 - › 日本語はハイコンテキスト言語のため、省略や発話スタイルのバリエーションが多くかなり難しい言語と思われる
- 正直英語だけで研究する方が早いし成果出る。が、自分たちの日常で使えないものはテンション上がらない
 - › しかも作っても自分でよし悪しがいまいちわからない（すこぶるブロークンな対話は良さが理解不能・・・）

BlenderBotのコピーを作る！ & ライブコンペに出してみよう！

構築システムの概要



- ベース：Transformer encoder-decoderモデル
- + フィルタ：過去発話類似フィルタ + 不自然共起フィルタ
- ベース=いわゆるEnd-to-endモデル
 - 構造はオリジナルのBlenderBot Generativeモデル相当（小細工したものより性能↑）
 - 1.6Bパラメータ（2層Encoder, 24層decoder, 各Embedding1920次元）
= BlenderBotの2.7Bよりもやや小さい（V100 16GBにfairseqライブラリで載せられる限界値）
 - デコードはSample-and-rank方式
- 不自然共起フィルタ = 「有川浩の容疑者Xの献身」などの共起関係が不自然な発話へのペナルティ
 1. 発話候補・文脈に含まれる固有名詞（Wikipedia見出し語）を抜き出す
 2. 発話候補内の固有名詞と全固有名詞のいずれかのペア（例：「有川浩」 - 「容疑者Xの献身」）がWikipediaの各ページの概要内で共起していればスルー。なければフィルタ。



→フィルタ結果（いまいち）：

上記のパターンはフィルタ出来た一方、有用な発話も一緒にフィルタされた

- 並列：「好きな作家は？」 → 「恩田陸とか有川浩とか」
- 対比：「フランスとかたまに行きます」 → 「私はドイツに行くことが多いです」

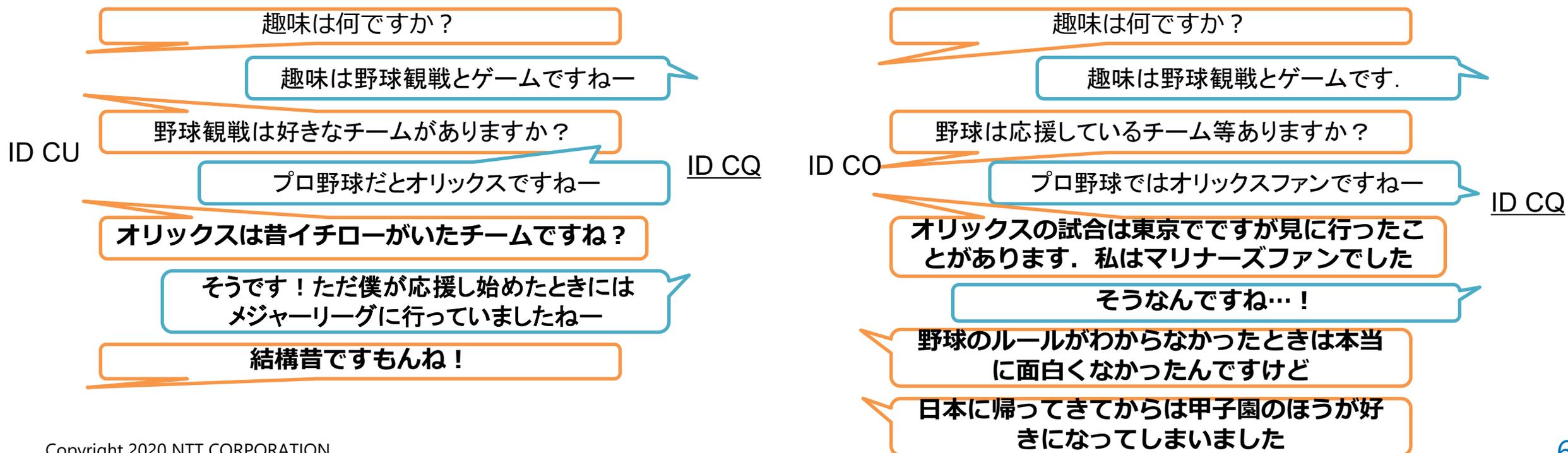
構築システムの学習詳細



- モデルの学習
 - 利用GPUリソース：産総研ABCIクラウド（各ノードV100 16GB x4）
 - › Pre-train: 400GPU（100ノード） x 28時間程度（下記Pre-trainデータを1回し半）
 - › Finetune: 128GPU x 1~3時間程度
 - いずれもvalidation（Twitterリプライ）のppl最小のモデルを利用
- データ
 - Pre-train: クリーニング済みTwitterリプライペア(201601-201803) 512GB程度（21億発話ペア）
 - › クリーニング
 - › 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
 - Finetune: 22万発話ペアの対話データ
 - › NTT内部の雑談コーパス（4700対話ほど）
 - › BlenderBotのスキル学習用コーパスの日本語版（Persona-chat, Wizard-of-Wikipedia, Empathetic Dialogue. それぞれ作成中のため少量・計2500対話ほど）
 - › Blended Skill Talkコーパス（スキル統合用コーパス）の日本語版（は用意できず→初対面趣味雑談コーパス（3000対話ほど）を流用）

初対面趣味雑談コーパス

- 初対面の人同士の2者間テキストチャットを高密度に収集したコーパス
 - › 80名の各参加者が他の全ての参加者と1回ずつ対話（最低60回）
 - › 話題を自身の趣味に限定
 - 個々人で異なり，かつ相手話者によって多少の影響を受ける対話を収集できる（個人性を反映した対話を学習する基礎データ）



ライブコンペ投稿



- 構築したシステムをライブコンペ3 オープントラックへ投稿
 - ライブコンペ2 で上位がシナリオベースのみになったことを受け, レギュレーション変更 (通知直前…)
- 対話ルール：
 - 所定リストから対話者が選んだ2つの話題 (システムには非開示) について対話
- 評価尺度：
 - 自然性：対話が自然かどうか
 - 話題追従：システムはユーザが選択した話題に関して適切に応答できたかどうか
 - 話題提供：システムはユーザが選択した話題に関して新たな情報を提供できたかどうか
- 幸いにして趣味雑談でのFinetuneにうまくフィット
 - 「趣味」を軸に展開しつつ, 個々の話題にそれなりに適応できそう (趣味は基本オープンドメイン)
 - 今回は話者を相手の1発話目の内容に合わせて変更 (似た趣味を持っている人のIDを入力文に埋め込み)

ライブコンペ2：
話題指定なし・完全に自由

ライブコンペ2：
「この対話システムとどのくらい
(また) 話したいか」のみ

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

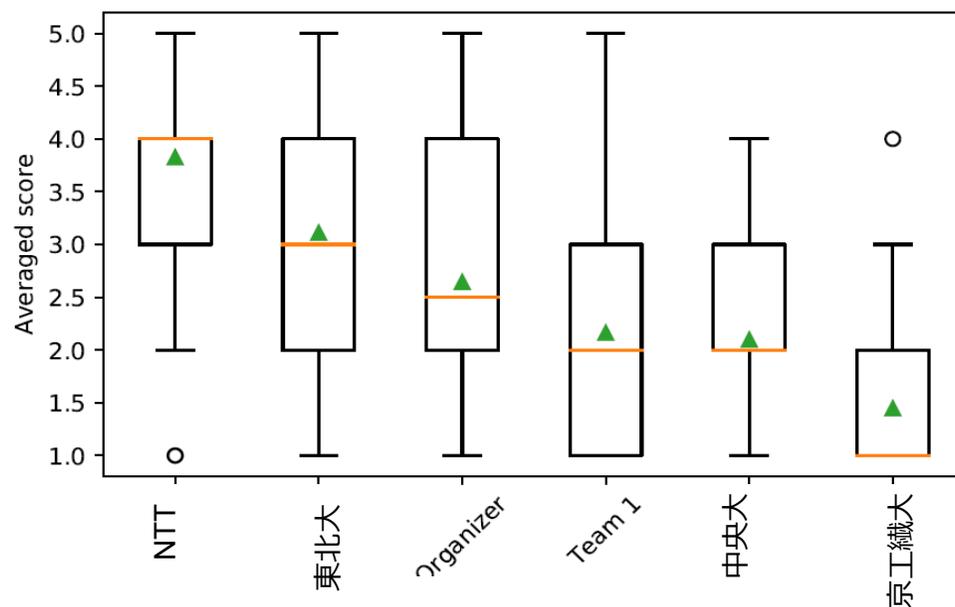
対話例

System: こんにちは。よろしくお願いします。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: 安田さんがリバーズ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、最近はどこに行かれましたか？
User: 最近では東京に行きました。
System: 東京ですか!いいですね!行ってみたいです。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: そうなんですか! 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: そうなんですか! 最近は何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ススメです。
System: 書店ガール、タイトルだけ聞いたことがあります! 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! 気に入りますね! 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! 是非見て見ます! 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。

ライブコンペ3 予選結果



- 有意差ありの1位を獲得（2位にも有意差はライブコンペ初（のはず））
 - 自然性 4.03, 話題追隨 3.87, 話題展開 3.58
 - 話題追隨・展開がやや弱い（人はどのくらいのスコア？）
- 2位の東北大も同様のアプローチを採用
 - 大規模Transformer encoder-decoderモデル+ Wikipediaによるテンプレートベース知識伝達発話
 - モデル4.8億パラメータ・Pre-train3億発話ペア・Finetune15万発話ペア（データはいずれもTwitterペアから作成）
 - モデルサイズが3倍以上・Pre-trainデータ量が7倍以上異なる点と、Finetuneコーパスの質の差が結果の差の要因と予想



対話の分析



- 対話破綻につながる発話の分類（アノテータ1名が付与）
 - 矛盾・話題の飛びが1対話に1,2回発生している
 - 事実誤認もフィルタし切れていない（精度&ペナルティコストの問題？）
- 評価値への影響
 - 自然性・話題追従：事実誤認・話題の飛びで低下（話題を認識できていないように見える）
 - 話題展開：キーワード関係誤り・概念誤りで低下（新規話題導入や話題の深掘りに失敗）

破綻度合
B1：かすかな違和感
B2：明らかな違和感だが継続可
B3：継続困難な破綻

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：有田焼はどんなところ？）	5	2	5	4.4	4.1	3.0
事実誤認	事実と異なる発話（例：有川浩の容疑者Xの献身）	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、相手話者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			

対話の分析



- 対話破綻につながる発話の分類（アノテータ1名が付与）
 - 矛盾・話題の飛びが1対話に1,2回発生している
 - 事実誤認もフィルタし切れていない（精度&ペナルティコストの問題？）
- 評価値への影響
 - 自然性・話題追従：事実誤認・話題の飛びで低下（話題を認識できていないように見える）
 - 話題展開：キーワード関係誤り・概念誤りで低下（新規話題導入や話題の深掘りに失敗）

破綻度合
 B1：かすかな違和感
 B2：明らかな違和感だが継続可
 B3：継続困難な破綻

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：有田焼はどんなところ？）	5	2	5	4.4	4.1	3.0
事実誤認	事実と異なる発話（例：有川浩の容疑者Xの献身）	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、相手話者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			

対話の分析



- 対話破綻につながる発話の分類（アノテータ1名が付与）
 - 矛盾・話題の飛びが1対話に1,2回発生している
 - 事実誤認もフィルタし切れていない（精度&ペナルティコストの問題？）
- 評価値への影響
 - 自然性・話題追従：事実誤認・話題の飛びで低下（話題を認識できていないように見える）
 - 話題展開：キーワード関係誤り・概念誤りで低下（新規話題導入や話題の深掘りに失敗）

破綻度合
 B1：かすかな違和感
 B2：明らかな違和感だが継続可
 B3：継続困難な破綻

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：有田焼はどんなところ？）	5	2	5	4.4	4.1	3.0
事実誤認	事実と異なる発話（例：有川浩の容疑者Xの献身）	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、相手話者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			

対話の分析



- 対話破綻につながる発話の分類（アノテータ1名が付与）
 - 矛盾・話題の飛びが1対話に1,2回発生している
 - 事実誤認もフィルタし切れていない（精度&ペナルティコストの問題？）
- 評価値への影響
 - 自然性・話題追従：事実誤認・話題の飛びで低下（話題を認識できていないように見える）
 - 話題展開：キーワード関係誤り・概念誤りで低下（新規話題導入や話題の深掘りに失敗）

破綻度合
 B1：かすかな違和感
 B2：明らかな違和感だが継続可
 B3：継続困難な破綻

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：有田焼はどんなところ？）	5	2	5	4.4	4.1	3.0
事実誤認	事実と異なる発話（例：有川浩の容疑者Xの献身）	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、相手話者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			

対話の分析



- 対話破綻につながる発話の分類（アノテータ1名が付与）
 - 矛盾・話題の飛びが1対話に1,2回発生している
 - 事実誤認もフィルタし切れていない（精度&ペナルティコストの問題？）
- 評価値への影響
 - 自然性・話題追従：事実誤認・話題の飛びで低下（話題を認識できていないように見える）
 - 話題展開：キーワード関係誤り・概念誤りで低下（新規話題導入や話題の深掘りに失敗）

破綻度合
 B1：かすかな違和感
 B2：明らかな違和感だが継続可
 B3：継続困難な破綻

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：有田焼はどんなところ？）	5	2	5	4.4	4.1	3.0
事実誤認	事実と異なる発話（例：有川浩の容疑者Xの献身）	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、相手話者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			

まとめ

- 日本語最大規模のNNベースの雑談対話システムを構築
- ライブコンペ3 オープントラック予選で1位を獲得

展望

- モデルサイズ変動による性能評価 (ppl・主観評価)
- 破綻発話のフィルタ
- 外部知識や文脈の利用方法改善
 - 現状はまだまだ、よくできた反射的応答モジュールに過ぎない
 - これを使いこなす制御系が必要 (記憶や対話状態, 利用する外部知識の制御)

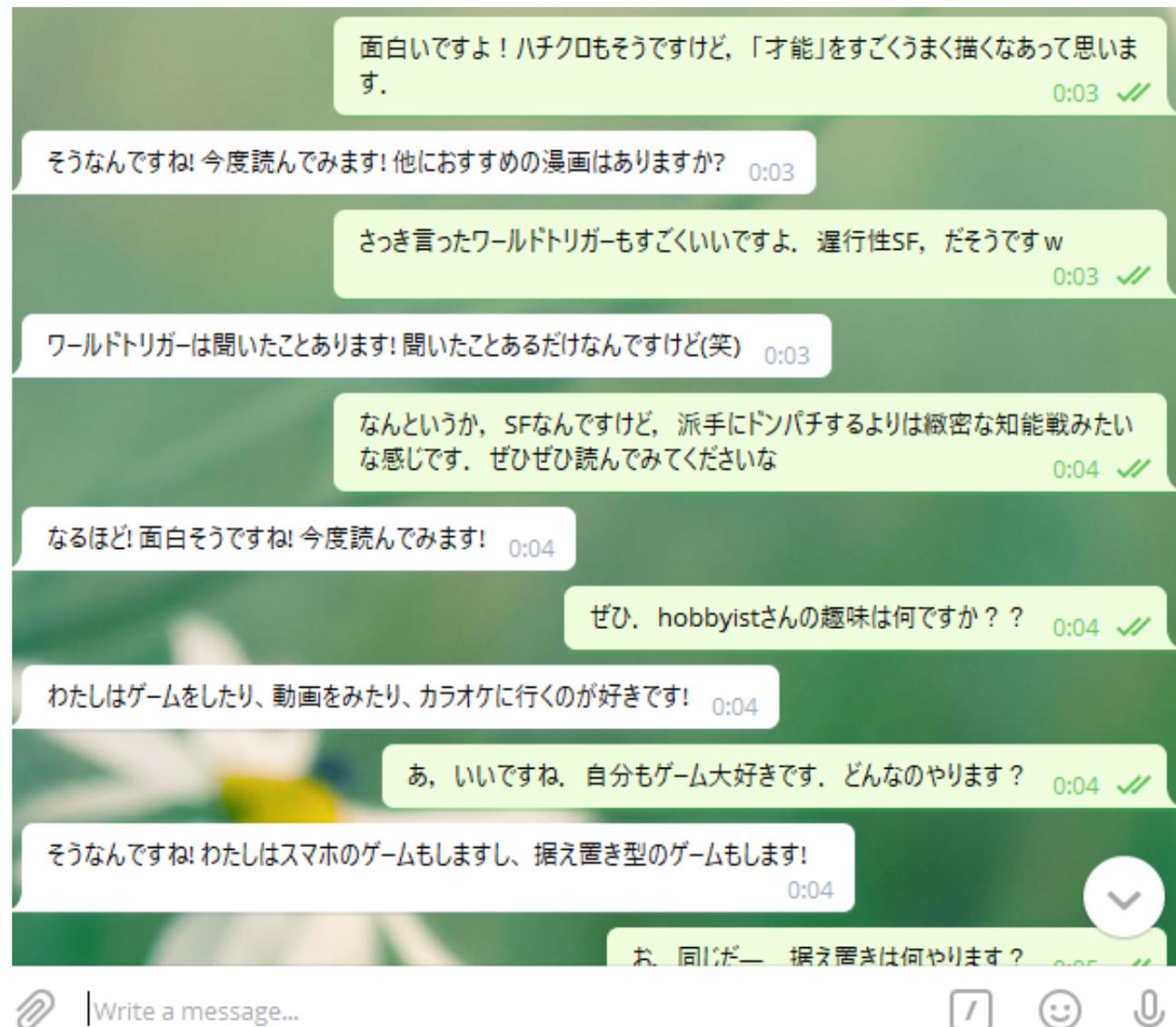
まとめ

- 日本語最大規模のNNベースの雑談対話システムを構築
- ライブコンペ3 オープントラック予選で1位を獲得

展望

- モデルサイズ変動による性能評価 (ppl・主観評価)
- 破綻発話のフィルタ
- 外部知識や文脈の利用方法改善
 - 現状はまだまだ、よくできた反射的応答モジュールに過ぎない
 - これを使いこなす制御系が必要 (記憶や対話状態, 利用する外部知識の制御)
- **本モデル・Finetuneデータの一部を公開予定 (ちょっとシュリンクするかも?)**

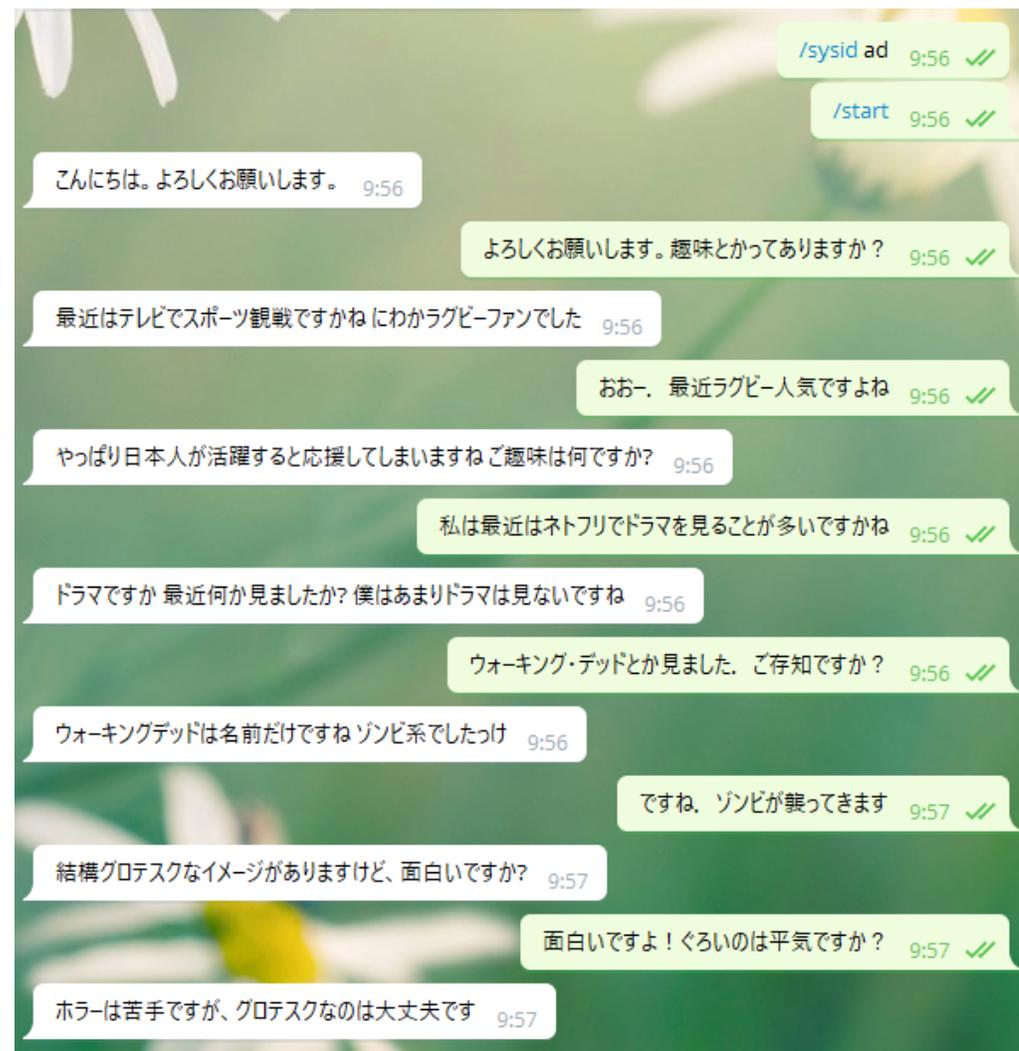
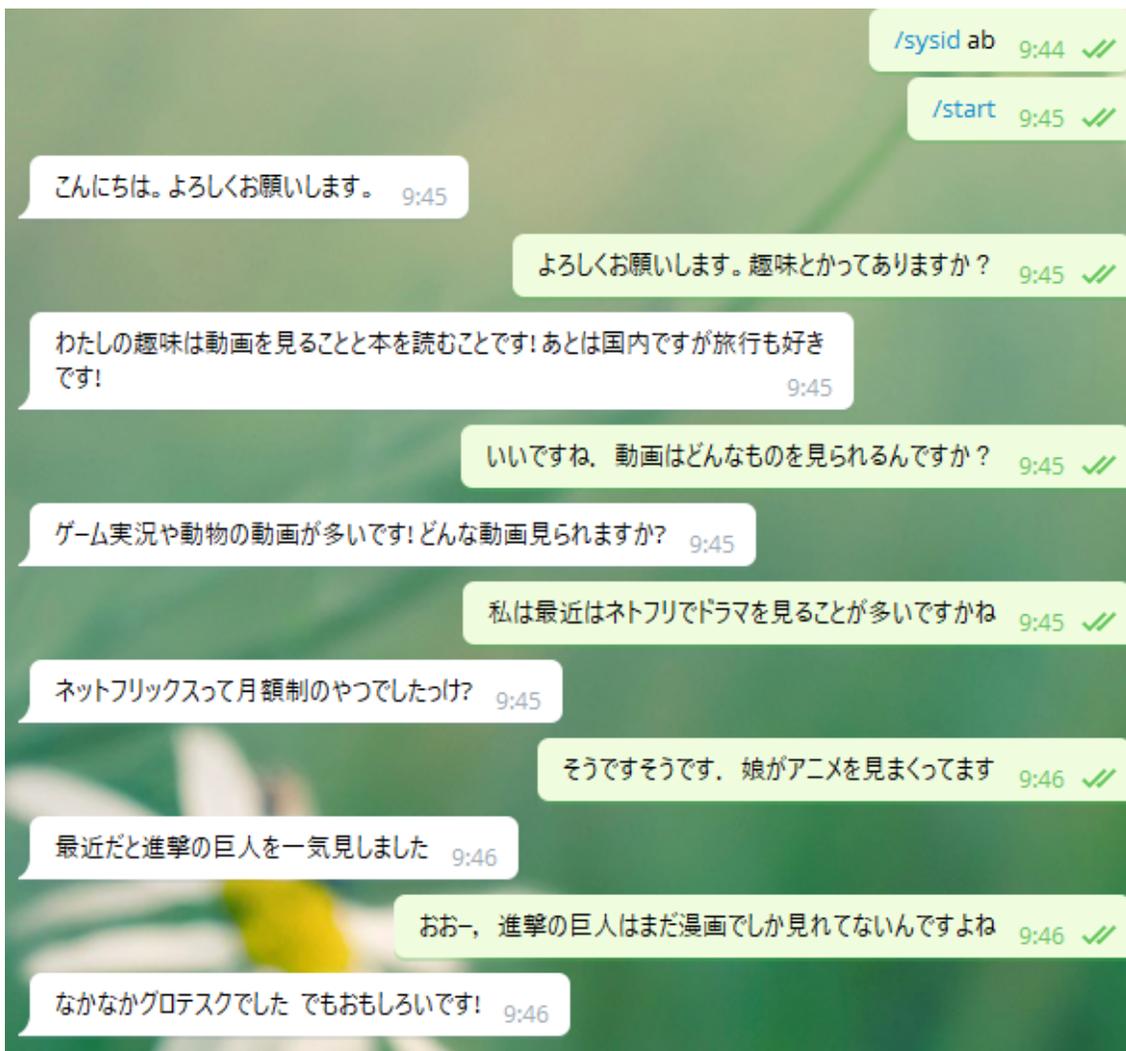
長めの対話例



話者の切り替え

システムID : AB

システムID : AD



推薦・常識理解の例

