

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 -

電気通信大学

稲葉 通将

本発表の趣旨

- TripiaBot アルゴリズムの説明は1分で終わる
 - 極めて単純
- なぜそのような単純な対話システムがライブコンペの予選を通過したか？
 - ⇒ 雑談対話システムの評価の難しさ

雑談対話システムの評価尺度を概観し、
TripiaBotの予選通過の要因と適切な評価方法について考える

TripiaBotのアルゴリズム

- 完全シナリオベース
 - 16個の発話を決められた順番で出力
- ユーザの発話は完全に無視
 - 発話の内容と出力順は一切変化しない

TripiaBotの16個の発話

1. こんにちは！私はシンといいます。あなたのお名前を教えてくださいませんか？
2. では、短い間ですがよろしくおねがいします。実は私、旅行が好きでよく行くのですが、あなたはよく旅行には行きますか？
3. そうなんですね。実は旅行って、健康にも良いいってことご存知ですか？
4. 旅行に行くとストレスが解消されて健康になる、っていうのもあるんですが、旅行に行くと心臓病になりにくいという研究結果があるんですよ。
5. その研究によると、半年に1回以上旅行に行くと、心臓病のリスクが低下するんだそうですよ。旅行に行って、楽しくて、さらに健康になるなんて最高だと思いませんか？
6. さらに、アメリカのアーカンソー大学の研究によると、高齢者になってからも定期的に旅行に行くことで、寿命も伸びるらしいですよ。
7. これは、旅行先でいろいろな刺激を受けて脳が活性化したり、あとは夫婦で旅行に行くことで一層仲が良くなったりするってことが要因のようですよ。
8. ところで、あなたが旅行で一番楽しいと思う時ってどんな時ですか？
9. 私は旅行先で新しい文化に触れるときが一番楽しみなのですが、科学的には旅行で一番幸福度が高くなるのは、実は旅行の前らしいですよ！
10. 旅行中よりも、旅行の前のほうが幸せって結構驚きですよ。
11. 人間は未来のことを考えると幸せになるっていう性質があることが関係しているみたいです。例えば、なにか欲しい物があった時、実際に手に入れた後よりも手に入る前のほうが楽しい感じってなんとなくわかりますよね。
12. なので、旅行は出発の何週間か前にしっかり予定を立てることが楽しい旅行をする秘訣らしいです！
13. そういった意味では、ガイド付きのツアー旅行よりも、自分でいろいろ計画して行くのが良いみたいですね。
14. おっと、残念ですがそろそろ時間ですね。あなたとお話していて、なんだか旅行に行きたくなってきました！
15. それでは、また機会がありましたらお話ししましょう。
16. ありがとうございました！

TripiaBotシナリオ作成のポイント

- システム側が対話の主導権を握る
 - 質問や同意を求める発話を多用
 - ユーザの応答を予測して次の応答を作成
- 話題は多くの人に興味を持つ「旅行」
- ユーザの興味を引き続ける
 - 旅行の雑学を語る

ライブコンペ予選における評価

- TripiaBotは平均4.20で予選を一位タイで通過
 - 作成したシナリオは極めて有効だった
- なぜTripiaBotは高い評価となったか？
 - ⇒ 本コンペで用いられた評価尺度の限界

Amazon Alexa Prizeでも同じ評価尺度では？

- Alexa Prizeの決勝ラウンドではライブコンペと同じ評価尺度
- Alexa Prizeでは、書類選考の段階で技術的に新規性のないものは落ちる
 - TripiaBotはこの段階で間違いなく落ちる

雑談対話システムをどう評価すべきか

- 様々な方法でシステムは評価されている
 - 人手評価
 - 対話ごと，発話ごと，絶対評価，相対評価
 - 自動評価
 - 他タスク用評価指標，機械学習ベース，語彙数
- 評価方法が不適切な場合，システムを正しく評価できない

雑談対話システムの評価尺度：人手評価

• 人手評価

- 雑談対話システムの評価は人手によって行うのが妥当とされている
- ライブコンペも人手評価
- 以下の2つに大きく分けられる
 - 対話ごとの評価
 - 発話ごとの評価



人手評価：対話ごとの評価

- システムと人が実際に対話し，対話後にアンケート
 - 評価尺度
 - もう一度話したいか (ライブコンペ, Alexa Prize)
 - 流暢性，興味を引いたか，一貫性 (ConvAI)
- 利点
 - 実際に対話システムを運用する場合に近い評価が可能
- 欠点
 - 評価の分散が大きく，システム間性能差の検出力小
 - 1人の評価者と(多くても)数回しか対話しないため，多様な対話が可能かを評価できない

人手評価：発話ごとの評価

- 文脈とそれに対するシステム発話が与えられ、その発話を人手で評価
- 評価基準
 - 文脈との関係の一貫性 [Niu+ 18]
 - 論理性 [Qian+ 18]
 - 対話破綻ラベル [東中+ 16] など
- 絶対評価と相対評価が存在

人手評価：発話ごとの絶対評価

- 絶対評価により文脈に対する発話を評価

- 特徴

- システム間の性能差の検出力がやや大

- 対話ごとの評価よりも評価基準が明確

- システムの実運用に近い評価が可能

- 人とシステムの対話ログを用いて評価できる

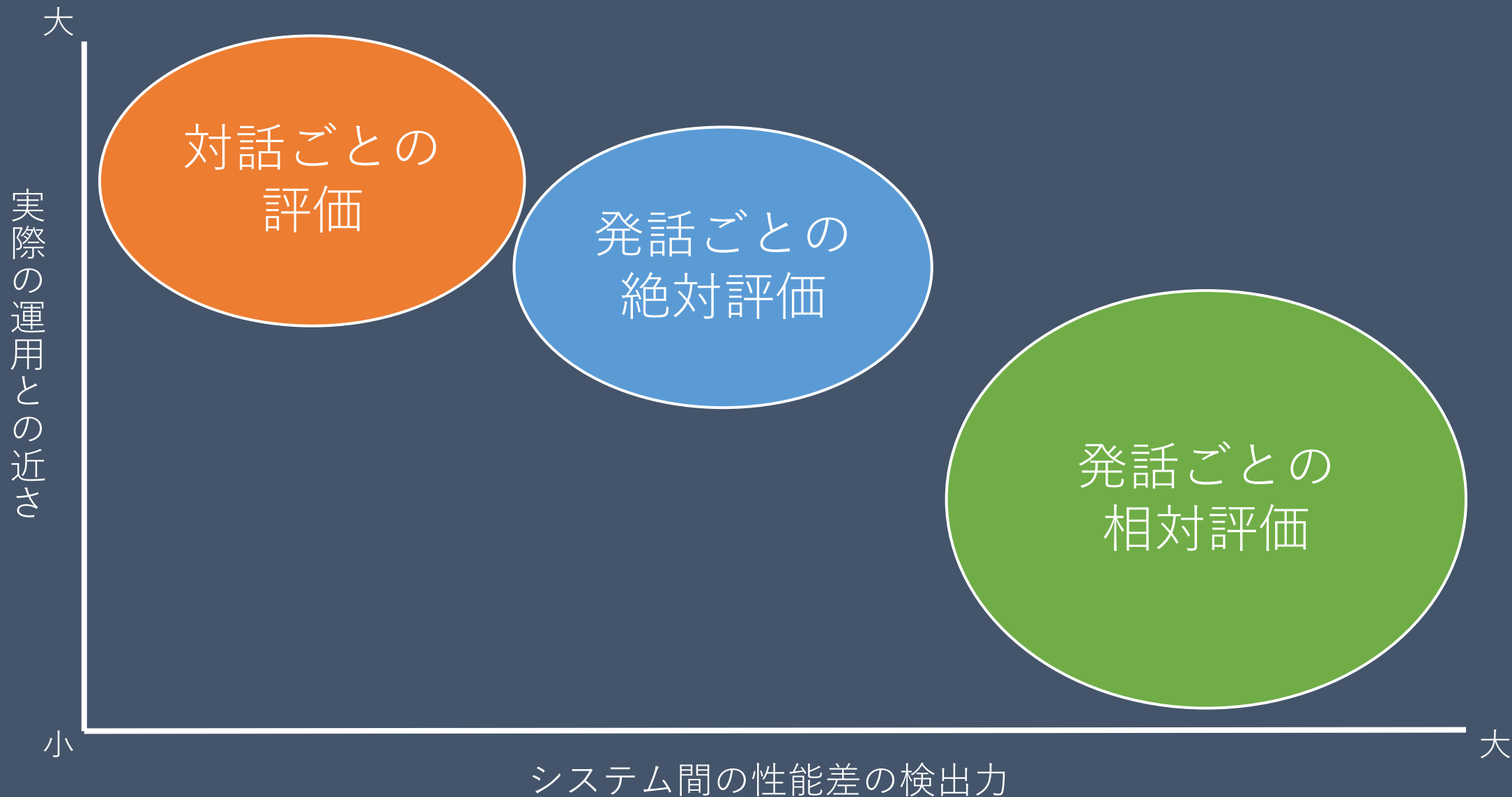
- 発話単位では問題がないが，対話単位では問題のあるシステムを高く評価する可能性

- 例：高頻度で質問し，詰問されているような印象を与えるシステム

人手評価：発話ごとの相対評価

- 文脈に対する複数のシステムによる発話を評価基準に基づき順位付け
 - 特徴
 - システム間の性能差の検出力大
 - 実運用における対話と乖離する可能性
 - 人間同士の対話ログの一部を文脈として用い、応答を生成して評価する場合が多い
 - 複数システムに同一の文脈を与える必要があるため

人手による評価の比較



自動評価

- 人手による評価はコストが高い
- 深層学習を用いた対話システムの研究ではほとんどの論文で自動評価を実施
- 自動評価の分類
 - 他タスク用の自動評価尺度による評価
 - 機械学習に基づく評価
 - 発話中の語彙による評価



自動評価：他タスク用の自動評価尺度

- 自然言語処理の他タスクにおける自動評価尺度を用いて評価
 - 文脈から発話を生成し，それが実際のログ中の発話をどの程度再現できたかにより評価
 - BLEU(翻訳), ROUGE(要約), Perplexity(言語モデル)など
- 人手評価との相関が低い

自動評価：機械学習に基づく評価

- 機械学習によりシステムの発話を評価
 - 隠れマルコフモデル[磯村+ 09], 回帰モデル[Sugiyama+ 19], 深層学習[Lowe+ 17] など
- 敵対的サンプルに対する頑健性の欠如が指摘 [Sai+ 19]

自動評価：発話中の語彙による評価

- 応答の多様性を評価
 - ニューラル対話モデルでは簡潔な応答が頻出する傾向があり，それを評価
- n-gramのType Token Ratio(distinct-nとも)
 - 異なりn-gram数を総語数で正規化した値

人手評価は応答の多様性の評価には不向きであり
客観的な自動評価が必要

考察

- 人手評価は対話の質と内容の評価に適する
一方、多様性の評価には不向き
 - TripiaBotが高評価となった要因
- 自動評価は人手評価の代替としては課題が残るが、多様性の評価には適する

まとめ

- 雑談対話システムの適切な評価のために
 - 人手評価は必須
 - 対話ごとの評価と発話ごとの絶対評価が望ましい
 - 実運用に近い評価が可能
 - 発話ごとの相対評価は比較システムとの性能差が小さい場合のみ行うべき
 - 自動評価については多様性の評価を行うべき
 - 人手評価の欠点を補完