雑談対話における会話への関心度と継続可能性を考慮した自動評価手法

蔦侑磨（東京大学）　吉永直樹，佐藤翔悦, 豊田正史（東京大学生産技術研究所）

概要

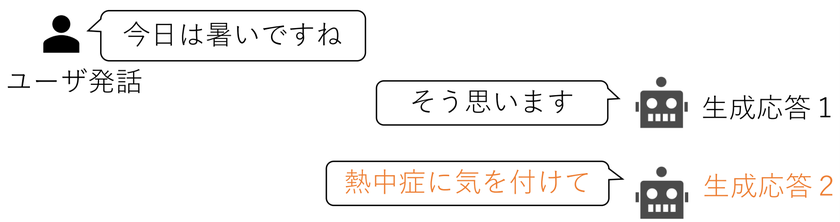
研究目的：対話システムを以下の観点から評価する
- ユーザの興味を引けるか
- ユーザの応答が期待できるか

提案手法：Twitter上の会話での「いいね」や返信数を間接的な評価指標に利用した自動評価手法

研究課題：評価に関する以下の二点を確認
- 上記の評価尺度による評価が可能かどうか
  - 人手評価間の一致度による確認
- 間接的な評価指標が教師データとして適切か
  - 提案手法による自動評価手法の性能の確認

背景：ユーザが対話システムに期待する雑談

対話システムの応答は妥当であるだけではなく、
ユーザの興味を引き、会話が継続することが重要



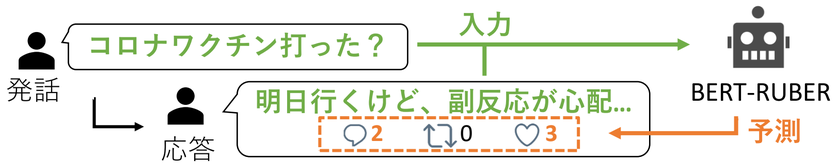
応答は両方とも妥当だが、応答 2 の方が興味を引き、
会話が継続しやすいと考えられる

関連研究・課題：評価モデルの学習データ

- 会話データには、応答への関心度や会話の継続性に関する評価が存在しない
- 人手によるアノテーション[Ghazarian+, AAAI2020]はコストが高い

提案手法：間接的な評価指標の利用

関心度や継続可能性の推定のため、自動収集可能な指標を利用した自動評価手法 [Ghazarian+, AAAI2020]の学習を行う
- 入力：Twitterでのリプライを応答とした会話
- 出力：応答文に付与される**「いいね」または返信の数**
  - 以降は返信数に関してのみ記載



出力値の補正

- **「いいね」や返信がある場合**

$$\frac{\log(\#of\ replies\ \text{or}\ likes)}{\log(\#\ of\ followeres)}$$

最大値がフォロワー数に比例するため

- **「いいね」や返信がない場合**

$$\frac{-\log(\#\ of\ followeres)}{\max(\log(\#\ of\ followeres))}$$

「いいね」や返信が付かない偶発性を考慮

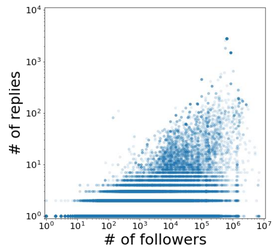


図. ツイートのフォロワ数に対する返信数

上記により補正した出力値は(−∞, ∞) の値を持つため、
それぞれの平均値・分散を利用して[0,1] に正規化する

分析：内的評価（返信数の有無に関する分類）

		教師ラベル	
		Pos	Neg
予測	Pos	102,851	31,411
	Neg	4,325	9,772

Accuracy: 0.76　Precision:0.77
Recall: 0.96　F1: 0.85
(Majority Class: 0.727)

ラベル
- Pos：出力値 > 0.5
- Neg：出力値 ≤ 0.5

結果
検証用データでの性能が低く評価指標が不適切である可能性が高い

「いいね」の数で学習したモデルも同様のため省略

分析：評価指標と人手評価に関する解析

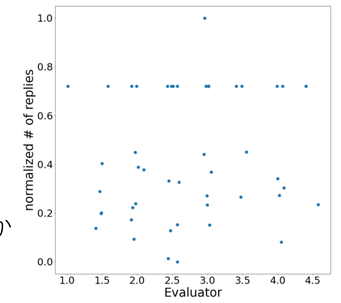
人手評価と評価指標の相関（下図）

評価指標が人手評価の間接的な指標として適切かを確認する

データ：約50件のTwitter上の会話

人手評価 (x軸)：
- 応答に返答（リプライ）を行うか
- [0,5]の6段階で評価
- アノテータ二人による平均値

評価指標 (y軸) ：正規化した返信数



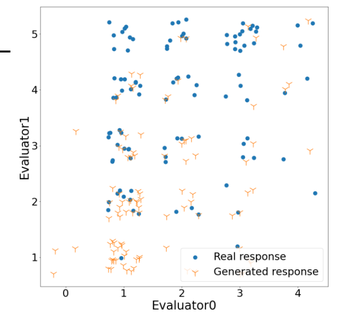
結果：人手評価と評価指標での相関は確認されない

人手評価間の一致度（右図）

二人のアノテータ間での人手評価の相関値を計測

データ：約50件の生成応答・実応答
- 同じ発話に対する応答を利用

人手評価：上記と同様



結果：個人によるばらつきが大きいことが確認された（r=0.363）

結論・今後の予定

- 個人ごとの評価のばらつきが大きいため、文レベルでのメタ評価は安定しないと予想される
  - 人手評価と評価指標での相関が確認されない原因の可能性としても考えられる
- 今後の予定：システムレベルでの評価方法を検討する