対話応答選択による対話応答生成モデルの評価

佐藤 志貴 [1]　赤間 怜奈 [1,2]　大内 啓樹 [2,1]　鈴木 潤 [1,2]　乾 健太郎 [1,2]
[1] 東北大学　[2] 理化学研究所
{shiki.sato,reina.a,jun.suzuki,inui}@ecei.tohoku.ac.jp
hiroki.ouchi@riken.jp

1　はじめに

雑談対話応答生成システムの自動評価は，人手評価に比べ低コストで再現性が高いため，日々のシステム改良が良い方向に効いているか継続的に評価する場合などに重要となる．しかし，応答生成システムの評価において，既存の自動評価指標は人手評価との相関が低いことが報告されている [1]．こうした問題は，ひとつの入力に対し適切な応答が複数存在することがあるという対話の性質に起因する (one-to-many 問題 [2])．

One-to-many 問題の影響を受けにくい評価方法のひとつとして対話応答選択が考えられる．対話応答選択では，与えられた対話履歴に対して適切な応答を応答候補から選ばせることでシステムの性能を評価する．システムに各候補を生成させたとき，最も尤度の高いものをシステムが選択した候補とみなすことで，応答生成システムによる応答選択が可能となる．応答選択による評価にはふたつの利点がある．ひとつめは，限られた候補のなかから応答を選択するという問題設定であるため，適切な応答が複数存在するという問題が発生しないよう設計できる点である．ふたつめは，正解率でシステムを評価するため，容易かつ明快にシステムの性能を比較できる点である．適切な応答を選択できるシステムが適切な応答を生成できるとは限らないものの，適切な応答を生成できるシステムならば適切な応答を選択できることが期待される．そのため，どの生成システムを人手評価にかけるかを選定するための事前評価など，日々のシステム改良の際におこなう評価に役立つ指標となると考えられる．

一般に，応答選択における応答候補は，対話履歴に対する本来の応答 (正例) に加え，誤り候補 (負例) をテストセットから無作為に抽出することで応答候補を構成する [3]．しかし，従来のように無作為抽出により負例を獲得する方法には，少なくともふたつの問題がある．ひとつめは，正例とかけ離れすぎていて容易に不適切と判別できる負例のみで候補が構成される可能性があること

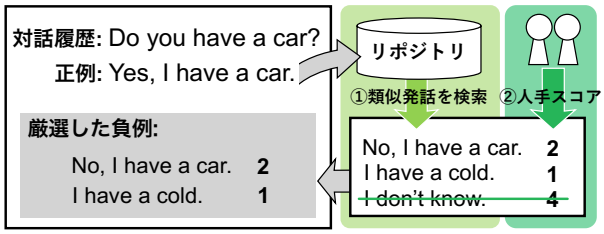

図1: テストセット構築方法の概要．①リポジトリから正例に類似した発話を検索し，②応答として成立してしまう発話を人手評価により除くことで，負例を取得する．

である．たとえば，「好きな果物はなんですか？」という対話履歴に対し「車が故障しました．」という発話が負例として無作為抽出されたとき，対話履歴と発話の間で関連性のある内容語が存在しないため，システムは両者の内容語を比較するだけで負例かどうかを容易に識別できてしまう．これによって，システムが対話の性質を理解していなくとも正解できてしまうような簡単すぎるテストセットが構築されうる．ふたつめは，応答として必ずしも不適切とはいいきれない発話が負例として候補に混入する可能性があることである．たとえば，「わかりません．」という発話は雑談対話において頻出するため，無作為抽出により頻繁に負例として取得されるが，この発話はさまざまな対話履歴に対して応答として成り立ってしまう．これらふたつの問題により，負例を無作為抽出により取得した応答選択テストセットの有効性は低いものとなる．

本研究では，図1に示す方法により負例を厳選することで上で述べたふたつの問題を解消し，負例の識別が人間にとっては容易だがシステムにとっては難しい応答選択テストセットを構築する[*1]．負例を厳選したテストセットを用いた応答選択によるシステムの評価が，人手評価と相関するかどうかを実験により確認する．

[*1] 構築したテストセットは公開予定

2 関連研究

応答生成システムの自動評価指標 対話応答生成システムの自動評価指標には，BLEU[4], METEOR[5], ROUGE[6] などの，機械翻訳システムや自動要約システムの評価を目的として提案された指標が広く用いられている．これらの指標では，システムが生成した応答と，正解となる応答例の類似度を測ることでシステムの評価をおこなう．しかし，対話の性質である one-to-many 問題の影響で，特定の応答例との類似度のみによって，生成された応答が対話履歴に対するものとして適切かどうかを評価することは難しい．

一方で本研究では，応答選択によるシステムの評価をおこなうことで，one-to-many 問題への対処を図っている．

人手ラベル付き応答選択テストセット 広く用いられている対話応答選択データセットのひとつに Douban Conversation Corpus[7] がある．Douban Conversation Corpus は，テストセットの各応答候補に対し，与えられた対話履歴に対する応答として適切かどうかを示すラベルが人手で付与されている中国語のデータセットである．このテストセットは本研究で構築するテストセットと類似しているものの，その目的と構築方法が異なる．Douban Conversation Corpus は検索型対話システムの性能評価を目的として構築されている．そのため，テストセットの応答候補は検索型対話システムと同様に対話履歴をクエリとして収集されている．

一方で，本研究で構築するテストセットは応答生成システムの性能評価を目的とする．各問の負例は，正例をクエリとして収集することで，適切な応答からかけ離れた発話が候補に含まれないようにする．

3 テストセットの構築

3.1 構築方法

対話履歴 c と正例 r^{true} に対し，大量の発話からなるリポジトリ $\mathcal{U}$ から負例となる候補 $r^{\text{false}} \in \mathcal{R}^{\text{false}}$ を収集する．このとき，1章で示したような，適切な応答からかけ離れた発話および応答として成立してしまう発話が $\mathcal{R}^{\text{false}}$ に含まれないようにする．本研究では，次のふたつの段階を踏み $\mathcal{R}^{\text{false}}$ を構築する:

1. 正例に類似した M 個の発話 $\{u_1, ..., u_M\}$ を $\mathcal{U}$ から検索する．
2. $\{u_1, ..., u_M\}$ のなかから，応答として成立してしまう発話を人手評価により除く．

1. 正例に類似した発話の検索 最初に，適切な応答からかけ離れておらず，対話履歴と内容語を比較するだけでは正例か負例かを判別できないような発話を収集する．正例に類似した発話のなかには，適切な応答からかけ離れた発話が含まれにくいと考えられるため，これらの発話を収集する．正例に類似した発話は，正例の内容語に類似した内容語を含むと考えられる．そこで，正例との内容語の類似度によって $\{u_1, ..., u_M\}$ を検索する．

2. 応答として成立する発話の除去 次に，検索された発話のうち，応答として成立してしまうものが $\mathcal{R}^{\text{false}}$ に含まれないようにする．応答として成立してしまうかどうかの判断には人手評価を用いる．具体的には，各発話 5 人のアノテータにより，その発話が与えられた対話履歴に対する応答として適切かどうかを示すスコアを付与するよう指示する．スコアは，1 から 5 までの 5 段階とする．スコア 5 は，発話が与えられた対話履歴に対する応答として適切であることを示し，スコア 1 は，応答として不適切であることを示す．また，明らかな文法誤りを含むような発話に対しては，スコア 0 を付与させる．

5 人中 3 人以上のアノテータにより 3 以上のスコアが付与された発話は，応答として成立している可能性があるとして $\mathcal{R}^{\text{false}}$ に含まれないようにする．加えて，3 人以上のアノテータによりスコア 0 が付与された発話は，文法誤りを含む可能性があるとして $\mathcal{R}^{\text{false}}$ に含まれないようにする．

3.2 構築したテストセットの概要

テストセットの統計情報 前節の方法により，各問正例に負例 3 発話を加えた 4 発話を応答候補とする 1,019 問のテストセットを構築した．テストセットの対話履歴および正例は DailyDialog[8] より取得した．リポジトリは OpenSubtitles2018[9] に含まれる発話を収集することで構築した．表1に構築したテストセットの基本的な統計量を示す．また，スコアリングの信頼度を確認するために計算したアノテータ間でのスコアリングの Fleiss' kappa を同表に示す．アノテータによるスコアリングを 0 から 5 までの 6 クラス分類と考えたとき，Fleiss' kappa は 0.22 となった．また，スコアリングを Douban Conversation Corpus と同様に 2 クラス分類と考えると[*2]，Fleiss' kappa は 0.63 であり，これは Doucan Conversation Corpus の 0.41 を上回った．

[*2] 4 以上のスコアを適切な応答とみなし，3 以下のスコアを適切ではない応答とみなした．

表1: 構築したテストセットの基本統計量およびアノテータ間でのスコアリングの一致率

問題数	1,019
問題あたりの応答候補数	4
問題あたりの対話履歴ターン数	3
Fleiss' kappa (6 クラス)	0.22
Fleiss' kappa (2 クラス)	0.63

テストセットの例　表2に，テストセットに含まれる問題例の 1 つを示す．表中の全負例は対話履歴に含まれる "camera" に関連する内容語 "focus" を，正例と同様に含んでいる．

予備実験　構築したテストセットの負例が，内容語のみを考慮するシステムによって識別されるかを確認した．実験には，対話履歴の内容語と各候補の内容語を比較する TF-IDF モデル [3] を用いた[*3]．構築したテストセットにおいて，TF-IDF モデルの応答選択の正解率は 0.461 となった．比較のために，構築したテストセットの負例を $\mathcal{U}$ から無作為抽出した発話に置き換えたところ，正解率は 0.671 となった．このことから，構築したテストセットの負例は，内容語を比較するだけでは識別することが難しいということがいえる．

4　実験

構築したテストセットを用いた応答選択による応答生成システムの自動評価結果が，人間によるシステムの評価結果と相関するかどうかを確認した．

4.1　実験設定

OpenSubtitles2018 を学習データとして，20 個の対話応答生成システム[*4]を学習した．その後，20 個のシステムを人手評価と自動評価それぞれによりランク付けした．人手評価により作成したシステムのランキングと自動評価により作成したシステムのランキングを比較することで，自動評価結果の人手評価結果との相関を確かめた．

人手評価によるシステムのランキング　学習した 20 個のシステムにより，与えられた対話履歴 $c \in \mathcal{C}$ に対する応答 r^{gen} をそれぞれ生成させた[*5]．その後，r^{gen} それぞれに対し 5 人のアノテータに，1 から 5 までの 5 段階のスコアを付与させた．各 r^{gen} に対し，5 つのスコア $s_1, s_2, ..., s_5$ が得られるため，この平均値 $s^{\text{mean}} = mean(s_1, s_2, ..., s_5)$ を求めた．さらに，テストセットの各問ごとに s^{mean} が得られるため，テストセット全体での s^{mean} の平均値 s^{final} を，各システムの最終的な評価スコアとした．s^{final} をもとに作成したシステムのランキングを，正解ランキングとした．

表2: 構築したテストセットの例

対話履歴:

A: Excuse me. Could you please take a picture of us with this ***camera***?

B: Sure. Which button do I press to shoot?

A: This one.

応答候補:

1. Could he not ***focus*** on that?
2. But I do have ninja ***focus***.
3. Do not lose your ***focus***!
4. Do I have to ***focus*** it? [正例]

応答選択正解率によるシステムのランキング　構築したテストセットを用いた応答選択の正解率を 20 個のシステムそれぞれについて計算し，正解率をもとに 20 個のシステムのランキング (`Well-Chosen`) を作成した．具体的には，学習した応答生成システムにより，各応答候補 r が生成されるときの cross-entropy loss ℓ_r を計算し，最もロスの低い候補 $\hat{r} = \underset{r \in \mathcal{R}}{\text{argmin}}\, \ell_r$ をシステムが選択した候補とみなすことで応答選択の正解率を計算した．比較のために，テストセットの負例をリポジトリから無作為抽出によって取得した発話に置き換えたうえで，同様にシステムのランキング (`Random`) を作成した．

その他の評価指標によるシステムのランキング　比較のために，既存の評価指標である BLEU, METEOR, ROUGE-L によるシステムのランキングを作成した[*6]．BLEU については，brevity penalty を適用しない場合についてもランキング ($\text{BLEU}^{\text{w/o BP}}$) を作成した．

4.2　実験結果

自動評価により作成したランキングがどれだけ正解ランキングに近いかを測るために，両ランキングの Spearman の順位相関係数を計算した．結果を表3に示す．

最初に，ランキングを比較する際の順位相関係数の upper bound を求めた．異なるアノテータが付与したス

[*3] 学習データには OpenSubtitles2018 を用いた．

[*4] GRU，LSTM，ConvS2S[10]，Transformer[11]，DialoGPT[12] の 5 種類のアーキテクチャから，異なるハイパーパラメータの計 20 個のシステムを用いた．

[*5] テストセット中の対話履歴のうち 56 個を取り出し，応答生成システムに入力したときに生成される応答を r^{gen} とした．

[*6] DailyDialog の対話データ 7,393 個の冒頭 3 発話を対話履歴，4 発話目を正解例としてそれぞれ取り出し，評価をおこなった．

表3: 正解ランキングと自動評価により作成したランキングの Spearman の順位相関係数および p 値

評価指標	順位相関係数	p 値
BLEU-1	−0.45	0.046
BLEU-2	−0.067	0.78
BLEU-1$^{\text{w/o BP}}$	0.68	<0.001
BLEU-2$^{\text{w/o BP}}$	0.58	0.0072
METEOR	0.16	0.51
ROUGE-L	0.50	0.026
`Random`	0.68	<0.001
`Well-Chosen`	**0.76**	<**0.001**
upper bound	0.94	<0.001

コア[*7]からシステムのランキングを 2 つ作成したうえで両ランキングの順位相関係数を upper bound とした．

`Well-Chosen` の正解ランキングとの順位相関係数は，既存の自動評価指標および `Random` に比べて高く，強い相関が認められる 0.7 を上回った．この結果から，負例を厳選したテストセットを用いた応答選択による応答生成システムの評価結果は，本実験で比較に用いた既存の評価指標に比べて，人手評価の結果と強く相関するということがいえる．

4.3 議論：エラー分析における解釈可能性

負例を厳選したテストセットを用いて自動評価をおこなうことの利点のひとつとして，エラー分析の解釈可能性が挙げられる．表4に，構築したテストセットの例を示す．厳選した負例 (厳選) は正例に類似しているが，対話履歴を考慮すると，“You” は主語としては明らかに不適切である．そのため，評価する応答生成システムが厳選した負例を選択した場合，システムが入出力文の主語を適切に考慮できていない可能性があるということが推測できる．このように，厳選した負例は正例に類似した発話であることから，正例と比較してどこが不適切かを明確に指摘することができる場合があるため，システムの予測について詳細に分析することが可能となる．

実際に，その負例がなぜ応答として不適切かを示すラベルセットを設計し[*8]，構築したテストセットから無作為に取り出した負例 50 発話に対しラベルを付与することを試みた．取り出した 50 発話のうち，22 発話にラベルセット中のラベルを付与することができた．

[*7] 5 人のアノテータにより付与された各発話 5 つのスコアを無作為に 2 グループに分割した．

[*8] 1. 対話履歴と矛盾，2. 情報量が不足，3. 主語が不適切，4. 時制が不適切，の 4 つからなるラベルセットを設計した．複数のラベルが該当する候補や，適切な応答からかけ離れているために具体的な誤りが指摘できない候補は，ラベルを付与できない候補とした．

表4: 無作為抽出した負例（従来）と厳選した負例（提案）の例

対話履歴:

A: Peter, enough with your computer games. Go do your homework now.

B: Can’t I play more?

A: No! Stop playing computer games!

応答候補:

正例: Mom, ***I***’ll be finished soon.

無作為: Thats the problem with small towns.

厳選: ***You*** are to be finished very soon.

5 おわりに

本研究では，人手評価により負例を厳選した対話応答選択テストセットを構築した．構築したテストセットを用いた応答選択による応答生成システムの評価結果が，人手評価の結果と相関することがわかった．今後の取り組みとして，厳選した各負例に対して「この候補はなぜ不適切な応答か」を示すラベルを付与することで，本テストセットを，評価するシステムがどのような負例を誤って選択する傾向にあるかを分析できるテストセットへと拡張することを考えている．

謝辞 本研究の一部は JST 未来社会創造事業（JPMJMI17C7）および JSPS 科研費 JP19H04162 の支援を受けて行った．

参考文献

[1] Chia-Wei Liu et al. “How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation”. In: *EMNLP*. 2016, pp. 2122–2132.

[2] Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. “Learning Discourse-level Diversity for Neural Dialog Models using Conditional Variational Autoencoders”. In: *ACL*. 2017, pp. 654–664.

[3] Ryan Lowe et al. “The Ubuntu Dialogue Corpus: A Large Dataset for Research in Unstructured Multi-Turn Dialogue Systems”. In: *SIGDIAL*. 2015, pp. 285–294.

[4] Kishore Papineni et al. “Bleu: a Method for Automatic Evaluation of Machine Translation”. In: *ACL*. 2002, pp. 311–318.

[5] Satanjeev Banerjee and Alon Lavie. “METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments”. In: *ACL*. 2005, pp. 65–72.

[6] Chin-Yew Lin. “ROUGE: A Package for Automatic Evaluation of Summaries”. In: *Text Summarization Branches Out*. 2004, pp. 74–81.

[7] Yu Wu et al. “Sequential Matching Network: A New Architecture for Multi-turn Response Selection in Retrieval-Based Chatbots”. In: *ACL*. 2017, pp. 496–505.

[8] Yanran Li et al. “DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset”. In: *IJCNLP*. 2017, pp. 986–995.

[9] Pierre Lison, Jörg Tiedemann, and Milen Kouylekov. “OpenSubtitles2018: Statistical Rescoring of Sentence Alignments in Large, Noisy Parallel Corpora”. In: *LREC*. 2018.

[10] Jonas Gehring et al. “Convolutional Sequence to Sequence Learning”. In: *ICML*. 2017.

[11] Ashish Vaswani et al. “Attention is All you Need”. In: *NIPS*. 2017, pp. 5998–6008.

[12] Yizhe Zhang et al. “DialoGPT: Large-Scale Generative Pre-training for Conversational Response Generation”. In: *ArXiv* abs/1911.00536 (2019).