負例を厳選した対話応答選択テストセット構築の試みと分析

The first step to create a better response selection test set with carefully chosen false candidates

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]
Shiki Sato[1]　Reina Akama[1,2]　Hiroki Ouchi [2,1]　Jun Suzuki [1,2]　Kentaro Inui [1,2]

[1] 東北大学　Tohoku University
[2] 理化学研究所　RIKEN Center for Advanced Intelligence Project

Abstract: False candidates for dialogue response selection test sets are usually randomly sampled from other dialogue data. However, randomly sampled candidates may be too far from proper responses. In addition, there is no guarantee that randomly selected candidates will be improper responses. This paper describes the first step to create a better dialogue response selection test set with carefully chosen false candidates. We retrieved candidates which are similar to the gold responses, and then scored them using crowdsourcing to remove candidates which might still be considered proper responses.

1　はじめに

対話システムの評価方法のひとつに対話応答選択がある．与えられた対話履歴に対する適切な応答を，選択肢の中から選ばせることによってシステムの性能を評価する．例えば，“How are you?”という発話とそれに対する応答の選択肢 “1. I’m fine.”と “2. This is mine.” が与えられ，システムが選択肢 1 を選べれば正解となる．応答選択は，適切な選択肢を選択できるかどうかを評価することにより，評価対象の対話システムが多様な入力文に対して適切な応答ができるかを間接的に評価することができる．また，評価指標が正解率なので，システム間の比較が容易かつ明快であり，対話システム開発の際に，日々のシステム改良が良い方向に効いているかを継続的に評価する場合などには最適な評価方法の一つと言える．応答選択に使用する選択肢は自動で作られるのが一般的である．具体的には，ある対話履歴に対する本来の応答 (正例) に加え，誤りとなる選択肢 (負例) をテストセットから無作為にサンプリングすることで選択肢を構成する [1, 2].

しかし，従来のように無作為サンプリング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選択肢が構成されてしまう可能性があることである．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テストセットが構築されうる．実際に負例を無作為サンプリングして BERT[3] に応答選択を解かせたところ，表 3 のようにほぼ全ての問題で正解した．ふたつめは，応答として必ずしも不適切とは言いきれない発話が負例として選択肢に混入する可能性があることである．これによって，選択肢のなかに正解がひとつだけ含まれているという前提が崩れ，算出される精度の信頼性が低いテストセットが構築されうる．本稿では，上記のふたつの問題を解消し，より信頼性の高い対話応答選択テストセットを構築する．具体的には，次のふたつの段階を踏んで負例を厳選し，選択肢を構成する．(i) 正例をクエリとした検索をおこない，正例に類似した発話を収集する．(ii) 収集した各発話に「応答としての良さ」を示すスコアをクラウドソーシングを通して付け，応答として不適切とは言い切れない発話を負例から除外する．以上の手法で構築したテストセットを用いて，既存の対話システムを評価および分析し，既存システムの抱える課題について考察する．

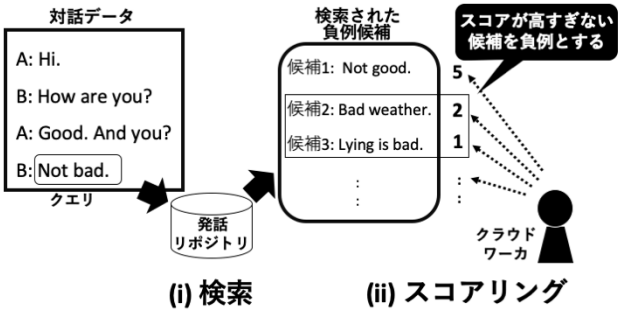

図 1: 負例候補の取得方法

*連絡先：東北大学大学院情報科学研究科システム情報科学専攻
〒 980-0845 仙台市青葉区荒巻青葉 6-3
E-mail：shiki.sato@ecei.tohoku.ac.jp

表 1: 構築したテストセットの概要

総対話数	103
各問題ターン数	4
各問題候補数	11
Fleiss' kappa(6 クラス)	0.23

2 選択肢の構成方法

テストセットの選択肢の構成にあたって，(i) 各問題の選択肢の収集，(ii) 各選択肢のスコアリング，のふたつの段階を踏んだ．以下，(i) と (ii) について詳述していく．なお，テストセットの問題は DailyDialog[4] で作成した．[1]

(i) 負例候補の収集　選択肢の収集時には，正例とかけ離れた発話が選択肢に含まれないように，各対話の正例をクエリとし，大量に用意した発話 (リポジトリ) から正例に類似した発話を検索した．リポジトリ構築のために，OpenSubtitles2018[5] から約 8500 万発話を収集した．構築したリポジトリから，正例をクエリとして (a) 表層的近さによる検索，(b) 意味的近さによる検索をおこなって選択肢に入れる発話を取得した．(a) は，Apache Lucene[2]を用いておこなった．(b) は，文ベクトル同士の cos 類似度を計算することでおこなった．文ベクトルは，ELMo[6] により求めた文中の単語の単語ベクトルの SIF[7] による加重平均とした．(a), (b) の検索法によりそれぞれ得られた検索結果上位 5 件をあわせた合計 10 発話に，正例を加えた 11 発話で選択肢を構成した．

(ii) 負例候補のスコアリング　構成した選択肢中の発話に対して，5 人のクラウドワーカ[3] に「応答としての良さ」を示すスコアを付与するよう指示した．具体的には，対話履歴に対して適切な応答と思われる発話には 5 に近いスコアを，不適切な応答と思われる発話には 1 に近いスコアを付与するような 5 段階評価をさせた．なお，文法誤りを含むような発話についてはスコア 0 を付与するよう指示した．スコアが付けられた発話のうち，スコア 4 以上を付与したワーカが 2 人以下の発話のみ負例として用いることとした．

3 構築したテストセットの概要

2 章で示した方法により選択肢を構成し構築したテストセットの統計を表 1 に示す．クラウドワーカによるスコアリングを 6 クラス分類と考えたとき，Fleiss' kappa[8] は 0.23 であり，Landis らの基準 [9] によれば，スコアリングはまずまずの一致 (Fair agreement) となった．

実際に構築したテストセットの候補例を表 2 に示す．表 2 の例 1 のひとつめの候補は，ワーカによる平均スコアが低かった例である．スコアが低い候補としては，応答として明らかに不適切であるものの，同例の太字部のように対話履歴や正例と同じまたは関連した内容語を含む発話が多く見られた．表 2 の例 1 のふたつめの候補は，スコア 4 以上を付与したワーカが 2 人以下である負例のうち，ワーカによる平均スコアが高かった例である．スコアが高い候補としては，正例に類似した発話であるものの，同例の下線部のように主語などが部分的に不適切である発話が多く見られた．

以上から，対話履歴と候補の内容語の一致などを考慮するだけでは正解することが難しいテストセットが構築できたと考えられる．

4 構築したテストセットによる評価

構築したテストセットで既存システムがどのように評価されるのかを分析するために，ふたつの実験をおこなった．ひとつめの応答選択実験では，テストセットの負例を厳選した場合と，無作為サンプリングした場合の評価結果を比較した．ふたつめの応答ランキング実験では，システムがどのような候補を適切，不適切な応答と予測するかを詳細に分析するために，付与されたワーカの平均スコア順に候補を並び替えさせた．なお，両実験ともテストセットのうち 85 問を用いた．既存システムは，著者らにより公開されているパラメータ[4]を，The Self-dialogue Corpus[10] により fine-tuning[5]した BERT を用いた．

応答選択実験の設定　負例を厳選した場合と，無作為サンプリングした場合それぞれについて，システムに 3 発話を与えたとき正例を選ぶ精度を求めた．前者の場合は次の 3 発話を候補としてシステムに与えた．

1. 正例
2. 最もワーカの平均スコアが高い負例
3. 最もワーカの平均スコアが低い負例 (非文除く)

後者の場合は正例に加え，OpenSubtitles2018 から無作為サンプリングした 2 発話を候補とした．

応答選択実験の結果　結果を表 3 に示す．厳選した負例を候補に用いたところ，応答選択精度は 0.76 となった．無作為サンプリングした負例を候補に用いたところ，応答選択精度は 0.99 となった．以上の結果から，無作為サンプリングと比べ，システムにとって識別の難しい候補が構築したテストセットに含まれていることがわかった．

実際に実験において出現した負例候補の例を表 4 に

[1] 各対話の冒頭 4 発話 (ABAB) を抜き出し，先頭 3 発話 (ABA) を対話履歴，4 発話目 (B) を正例とした．
[2] https://lucene.apache.org
[3] Amazon Mechanical Turk を利用．https://www.mturk.com
[4] Base(Uncased)．https://github.com/google-research/bert
[5] 連続した 4 発話を抜き出した対話を正例，4 発話目をコーパス中の別の発話で置き換えた対話を負例とした 2 値分類問題で fine-tuning をおこなった．

表 2: システムによる応答ランキングの例
「ワーカスコアによる順位」は，() 内に示したワーカの平均スコアによって候補を並び替えたもの．

	対話履歴	応答候補	ワーカスコアによる順位	システムの予測した順位
例 1	A: Excuse me. Could you please take a picture of us with this **camera**?	Why don't you just **focus** on the work?	3 位 (1.2)	1 位
	B: Sure. Which button do I press to shoot?	Alright, we have to focus, okay?	2 位 (3.6)	2 位
	A: This one.	Do I have to focus it? (正例)	1 位 (4.8)	3 位
例 2	A: Good morning, Madam. What can I do for you?	I've put your **name** on my boat.	3 位 (0.8)	1 位
	B: I'd like to withdraw 35,000 RIB from my	Um, Mrs. Bowman, my name is Jennifer MacMahon.	2 位 (2.6)	3 位
	A: Do you have an **appointment**?	Yes, my name is Ms Jane Reeve, R-E-E-V-E. (正例)	1 位 (4.2)	2 位

表 3: 構築したテストセット (85 問) による評価結果

システム	応答選択精度 (負例無作為取得)	応答選択精度 (負例厳選)	ランキング精度
BERT	0.99 (84/85)	0.76 (65/85)	0.46 (39/85)
Chance rate	0.33 (28/85)	0.33 (28/85)	0.17 (14/85)

表 4: 厳選した負例と無作為サンプリングした負例

対話履歴	
A: And anything to drink?	
B: Yes, a red wine and a cup of coffee.	
A: How do you like your **coffee**?	
厳選された負例	無作為サンプリングされた負例
But this is **sugar**?	You just let her win!

示す．表 4 において，厳選された負例は太字部で示したように対話履歴の関連語が含まれており，負例であると判別するためには対話の内容を理解する必要があると考えられる．一方で，無作為サンプリングされた負例の内容語は対話履歴と関連性のないものであり，容易に負例であると判別できてしまう可能性がある．

応答ランキング実験の設定　応答選択実験で取り出した 1，2，3 の候補をそれぞれ 1 位，2 位，3 位として，システムに正しい順で並び替えさせた．評価指標として，3 候補全てを正しく並び替える精度を求めた．

応答ランキング実験の結果　結果を表 3 に示す．与えた 3 候補を正しく並び替える精度は 0.46 となった．

システムによる実際のランキング結果の例を表 2 に示す．例 1 に示す問題では，ワーカによる平均スコアが非常に低い候補を 1 位と，正例を 3 位とシステムが予測した．1 位と予測した発話は応答としては不適切であるが，太字部のように対話履歴の単語の関連語を含んでいる．例 2 に示す問題でも，ワーカによる平均スコアが非常に低い候補を 1 位とシステムが予測した．1 位と予測した発話も応答としては不適切であるが，太字部のように対話履歴の単語の関連語を含んでいる．このような例から，用いたシステムが内容語の出現のみを見て 1 位と予測した可能性がある．

5　まとめ

本稿では，従来の対話応答選択テストセットにおいて問題となるような選択肢を除外したテストセットの構築の試みを報告した．構築したテストセットを既存システムに解かせ，既存システムがどのような候補に対して正しい予測をおこなえていないかを分析した．

今後の取り組みとして，主にふたつ挙げられる．ひとつは，妥当なテストセットとするために問題数を 1,000 程度までスケールアップしたうえで，対話システム研究のさらなる発展を目的として，構築したテストセットを公開資源とすることである，もうひとつは，用意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してやることで，システムがどのような負例を誤って選択するかを分析できるテストセットを構築することである．

謝辞

本研究の一部は JST 未来社会創造事業（JPMJMI17C7）の支援を受けておこなった．

参考文献

[1] 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.
[2] Matthew Henderson et al. "A Repository of Conversational Datasets". In: *NLP for Conversational AI*. 2019.
[3] Jacob Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *NAACL-HLT*. 2019, pp. 4171–4186.
[4] Yanran Li et al. "DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset". In: *Proceedings of the Eighth International Joint Conference on Natural Language Processing*. 2017, pp. 986–995.
[5] Pierre Lison, Jörg Tiedemann, and Milen Kouylekov. "OpenSubtitles2018: Statistical Rescoring of Sentence Alignments in Large, Noisy Parallel Corpora". In: *IJCNLP*. 2018.
[6] Matthew Peters et al. "Deep Contextualized Word Representations". In: *NAACL-HLT*. 2018, pp. 2227–2237.
[7] Sanjeev Arora, Yingyu Liang, and Tengyu Ma. "A Simple but Tough-to-Beat Baseline for Sentence Embeddings". In: *ICLR*. 2017.
[8] Joseph L Fleiss. "Measuring nominal scale agreement among many raters". In: *Psychological bulletin 76(5)* (1971), pp. 378–82.
[9] J. Richard Landis and Gary G. Koch. "Measuring nominal scale agreement among many raters". In: *Biometrics 33(1)* (1977), pp. 159–174.
[10] Joachim Fainberg et al. "Talking to myself: self-dialogues as data for conversational agents". In: *arXiv preprint arXiv:1809.06641* (2018).