ニューラル対話モデルの自動評価に向けた対照応答対評価セットの試作

岡野功士朗 [1]　川村真也 [2]　鈴木優 [1]　加藤恒夫 [2]　田村晃裕 [2]　呉剣明 [3]

[1] 同志社大学大学院理工学研究科　[2] 同志社大学理工学部　[3]KDDI 総合研究所

ctwf0136@mail4.doshisha.ac.jp

1　はじめに

非タスク指向型対話における応答生成は正解が一つしかないわけではないが，文脈を適切に踏まえたものであることが望ましい．注意機構付き Encoder-Decoder RNN[1] や Transformer[2] によって比較的流暢な応答が生成されている．しかしながら，非タスク指向型対話における応答品質の自動評価手法は確立されていない．応答の仕方は多様であるため，コーパスから抽出した応答文を参照する BLEU のような指標は主観評価値との相関が低いことが報告されている [3]．この課題に対して，対話相手からの問い掛けとの関係も評価する指標 [4]，主観評価と統計的な評価を統合した指標 [5]，さらには，対話応答の品質は対話のやりとりからしか評価できないとしてインタラクティブに評価する手法 [6] などが提案されている．

一方，文レベルの翻訳品質が向上した機械翻訳では，一連の文章を翻訳する場合に前後の翻訳文間の意味的な整合性が新たな課題となっている．この場合も，文脈に関わる意味的な整合性を BLEU で測るのは難しく，Sennrich らは，機械翻訳システムが生成する翻訳文の言語的な傾向を分析するために自身が提案した対照翻訳対 [7] を発展させ，英仏翻訳において前方照応代名詞の翻訳と語彙選択の一貫性を評価する評価セットを提案している [8]．さらに，文脈を考慮した英露翻訳において，頻出する典型的な誤りを分析し，その誤りに関する翻訳モデルの性能を評価するための対照翻訳対を設計している [9]．英日翻訳では，永田らが共参照と一貫性を評価する対照翻訳対を設計している [10]．これらの対照翻訳対は，対訳コーパスから抽出した正解翻訳文に必要最小限の誤りを加えて誤り文とし，正解翻訳文と誤り文に対して翻訳モデルが算出する生成確率の比較において正解翻訳文の方が高い割合を測ることでモデル性能を定量化する．

本研究では，非タスク指向型対話システムにおける応答品質の自動評価に向けて，文脈を考慮した機械翻訳の評価のための対照翻訳対 [9] に倣い，対照応答対を提案する．大規模な雑談対話コーパスから学習したニューラル対話モデルが生成する応答文の誤り傾向を分析し，頻度の高い 3 種類の誤りを加えた対照応答対を試作した．作成した対照応答対評価セットを用いて性能差がある 3 種類のニューラル対話モデルを評価し，主観評価で測った性能差が対照応答対評価セットによる自動評価値と整合するか検証した．

2　対照応答対の作成手順

文脈を考慮した機械翻訳を評価するための対照翻訳対の設計手順に倣い，まずニューラル対話モデルが生成する応答文の誤り分析を行い，その結果を踏まえて頻度の高い誤りを含む対照応答対を作成し，評価する．具体的には以下の手順で進める．

1. 誤り分析：ニューラル対話モデルが生成した多数の応答文に対して，自然であるか否かを人手で 2 値分類する．自然でないと判定された応答文を誤りの種別に分類し，各誤り種別の頻度をカウントする．
2. 対照応答対作成：対話コーパスに含まれる正しい文に，頻度の高い誤りを加えて誤り文とし，正しい文と組み合わせて対照応答対（ペア）とする．このときの誤り文は，正しい文に含まれる形態素の最小限の置換によって作成する．
3. モデル評価：評価対象のニューラル対話モデルで，対照応答対の正しい文と誤り文の生成確率

表 1　定義した誤りラベルとその概要

ラベル名	誤り種別	例
ICW	文脈上，不適切な内容語を含んでいる	梅酒の話題の際に“ライチって美味しいですよね”
RUDE	相手に対して失礼な発話	相手に対して“メイドっぽいですね”
FNC	適切でない助詞を選択している	食べたいものについて尋ねられた際に“パエリアとか食べたいです”
ESE	文末表現を誤っている	通勤中の読書の話題に対して“寝過ごしちゃったりしますか？”
SC	自身の過去の発話と矛盾している	定時退社するという発言後に“フレックスタイムの仕事です”
RL	自身の過去の発話を繰り返している	朝活の話題の際に“朝活ってやつですね”
IA	相手の質問に回答していない	いつ結婚したかと尋ねられた際に“7 歳になりましたよ”
DIS	文法的に適切でない応答文	“疲れると疲れる”，“めったに行きます”などの発話
COL	応答内の修飾語が過去の発話と矛盾	都内の温泉について尋ねられた際に“山形県の温泉が多いです”
UNK	上記のいずれにも当てはまらない誤り	海で泳ぐかという質問に対して“バタバタバタフライ”

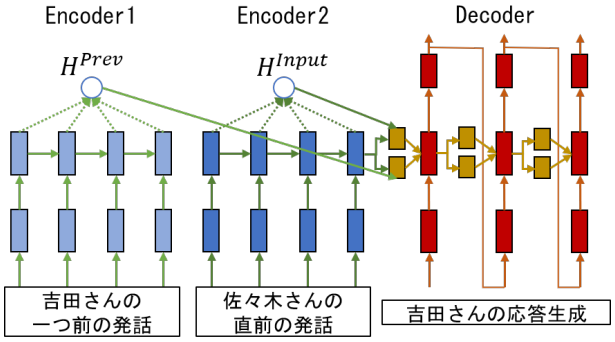

図 1　Double Attention モデル

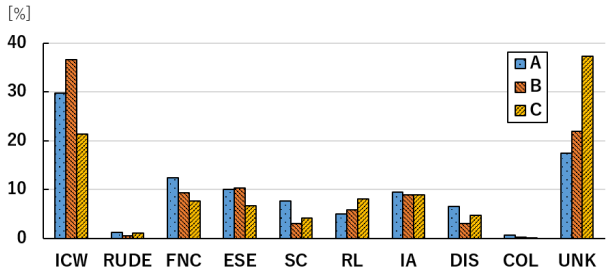

図 2　各評価者の誤りラベル使用割合

を算出し，正しい文の生成確率の方が誤り文の生成確率よりも高いペアの割合を，モデルの正解率とする．

以降，手順 1 の誤り分析を第 3 節に，手順 2 の対照応答対作成を第 4 節に，手順 3 のモデル評価を第 5 節に記述する．

3　ニューラル対話モデルによる応答文の誤り分析

非タスク指向型対話は女性同士の雑談を対象にした．大規模な雑談対話コーパスとして，クラウドソーシングを用いて作成した架空の 20 代女性 2 名「吉田さん」と「佐々木さん」の間の仮想雑談コーパスを用いた．二人のペルソナを細かく規定し，クラウドワーカーの間で共有することで，一貫性のある雑談を約 168 万発話分集めている．吉田さんと佐々木さんの雑談は交互に最大 10 ターン続くので，比較的長い文脈に対する応答の整合性を評価することも可能である．今回は「吉田さん」の対話モデル用に対照応答対評価セットの試作と評価を行うため，同コーパスより学習データ約 110 万文，検証データ約 6.4 万文，評価データ約 6.4 万文を抽出した．

「吉田さん」の対話モデルとして，GRU ベースの注意機構付き Encoder-Decoder モデルとその変形版の 2 種類を用いた．前者のモデル構造を図 1 に示す．1 つ前の吉田さん自身の発話と直前の佐々木さんの発話を別の Encoder に入力し，Decoder は 2 つの Encoder の隠れ層に対してそれぞれ注意機構を用いるので Double Attention モデルと呼ぶ．変形版は図に示していないが，2 つの注意機構のいずれにより強い注意を向けるかを考慮する機構を加えたものである．モデル学習はクロスエントロピ誤差を損失関数とする教師あり学習で行った．

検証データから無作為抽出した 1500 種類の対話文脈に対して，上記の 2 種類のニューラル対話モデルで応答文を生成した．応答生成は相互情報量最大化を基準にして行っている [11]．計 3000 種類の応答文に対して，著者ら 3 名が誤り分析を行った．最初に 3 名それぞれが，応答が文脈に照らして自然であるか否かを基準に 2 値分類した．次に，自然でないと考えた応答文を対象に各評価者がその原因を分析し，その結果を評価者間で擦り合わせることで最終的に表 1 に示す 10 種類の誤りラベルを定義した．

評価者 3 名が，自然でないと考えた応答文は全体の 41.9%であった．それらに対して各評価者が選択した誤りラベルの割合を図 2 に示す．頻度の高いものから，文脈上不適切な内容語を含む誤り (ICW, 28.9%)，助詞の誤り (FNC, 9.8%)，文末表現の誤り (ESE, 8.9%) であった．そこで，頻度上位 3 種類の誤りに関わるモデル性能評価のための対照応答対を作成することにした．なお，誤り分析の詳細については文献 [12] を参照されたい．

表 2　内容語を置換した対照応答対の例

吉田	やっぱり日本人なので和食が 1 番体に合うと思います
佐々木	毎日食べても飽きないのは 和食ですね
吉田 (正しい文)	お**味噌汁**の具って何が好きですか?
吉田 (誤り文)	お**休み**の具って何が好きですか?

表 3　文末表現の誤りの詳細な分類

誤り種別	割合 [%]
平叙文と疑問文の反転	33.3
肯定文と否定文の反転	11.1
省略された主語の変化	11.1
共感表現の欠如	8.9
時制の誤り	4.4
動詞の誤り	4.4
願望表現の欠如	4.4
その他の誤り	22.2

4　対照応答対評価セットの試作

4.1　内容語を置換した対照応答対

文脈上不適切な内容語を含む誤り (ICW) に対応する対照応答対は，正しい文に含まれる文脈上重要な名詞を他の名詞に置換して誤り文とした．置換後の語として単に言語確率の低い語を選択してしまうことを避けるため，同じコーパスから bi-gram 言語モデルを学習し，置換後の語を 1)bi-gram 言語確率が置換前の名詞と同等になる名詞（Equally-likely, EL), 2)bi-gram 言語確率が最大の名詞 (Most-likely, ML), の 2 種類の基準で選択した．具体的には，正しい文 $W = \{w_1, ..., w_n\}$ 中の単語 w_i $(1 \leq i \leq n)$ を置換する場合，EL の置換後の語は式 (1)，ML の置換後の語は式 (2) によって選択する．

$$\hat{w_i} = \arg\min \left[\left\{ \log \frac{P(v \mid w_{i-1})}{P(w_i \mid w_{i-1})} \right\}^2 + \left\{ \log \frac{P(w_{i+1} \mid v)}{P(w_{i+1} \mid w_i)} \right\}^2 \right] \quad (1)$$

$$\hat{w_i} = \arg\max \{ \log P(v \mid w_{i-1}) + \log P(w_{i+1} \mid v) \} \quad (2)$$

ただし，置換後の語は，コーパス中に 2 回以上出現する名詞を候補として，Encoder に入力される語は除外して選択する．表 2 に内容語を置換した対照応答対 (ML) の例を示す．

4.2　文末表現を置換した対照応答対

図 2 で 8.9%を占めた文末表現の誤り (ESE) に対応する対照応答対を作成するため，誤りを細かく分類した結果を表 3 に示す．「平叙文と疑問文の反転」とは，コーパス中の正しい文が平叙文であるのに対してモデル応答が疑問文の場合，またはその逆の誤りを指す．「肯定文と否定文の反転」も同様である．「省略された主語の変化」とは，例えば対話相手の彼氏がいない発言に対して“彼氏募集中ですよ”という応答を返す誤りを指す．彼氏を募集するのは対話相手であるため，“彼氏募集中ですね”という応答が正しい．このように文末表現の違いにより，省略された主語が変化する場合にこの誤りラベルを用いる．「共感表現の欠如」とは，「ですよね」のような共感を含む応答が期待される場合に「ですよ」のような共感が含まれない表現が現れる場合を指す．「願望表現の欠如」も同様である．「時制の誤り」は時制，「動詞の誤り」は生成された動詞が文脈的に誤っている場合を指す．

頻度上位の 2 種類「平叙文と疑問文の反転」と「肯定文と否定文の反転」それぞれの 2 方向の誤り，すなわち「平叙 → 疑問」，「疑問 → 平叙」，「肯定 → 否定」，「否定 → 肯定」の計 4 種類の誤りを人手で付与した．表 4 に文末表現を置換した対照応答対の一例を示す．

表 4　文末表現を置換した対照応答対の例

吉田	私はカレーは甘口なんです
佐々木	カレーは辛いのは苦手ですか?
吉田 (正しい文)	そうなんです**よね〜**
吉田 (誤り文)	そうなんです**か?**

表 5　助詞を置換した対照応答対の例

吉田	一人暮らしすると料理が もっと楽しくなったりして
佐々木	確かにその可能性もありますね
吉田 (正しい文)	料理**が**できる女子ってモテますよね
吉田 (誤り文)	料理**に**できる女子ってモテますよね

4.3　助詞を置換した対照応答対

助詞を置換した対照応答対も，文末表現を置換した対照応答対と同様に人手で作成した．内容語の置換と同様に bi-gram 言語モデルを用いた自動作成も試みたが，置換後の語として文脈・文法的に適切な助詞しか候補として挙がらなかったために人手で作成した．作成時の注意点として，置換後も同じ深層格（動作主格・対象格など）になる助詞を選択しないようにルールを設けた．例えば，「私は学生です」という正しい文の主語「私」に続く助詞「は」を，「が」に置換した「私が学生です」は誤り文にならない．そこで，明らかに誤りとなる助詞に置換するようにした．表 5 に助詞を置換した対照応答対の一例を示す．

表 6　3 種類のモデルが生成した応答文の適切性に関する主観評価値の割合

	1:不適切	2:どちらでもない	3:適切
No Attention	27.4%	20.6%	52.0%
Single Attention	26.6%	20.5%	53.0%
Double Attention	23.3%	22.2%	54.5%

5　対照応答対評価セットの評価実験

5.1　比較対象のニューラル対話モデルと主観評価による応答文の品質

内容語を置換した対照応答対を EL, ML で 350 ペアずつ，文末表現を置換した対照応答対を 231 ペア（内訳：平叙→疑問 89，疑問→平叙 51，肯定→否定 66，否定→肯定 25），助詞を置換した対照応答対を 100 ペアの計 1031 ペアを作成した．

この対照応答対評価セットが，対話モデルの応答性能を正しく反映できるか評価するために，3 種類のニューラル対話モデルを用意した．

- Double Attention: 第 3 節の誤り分析に用いた 2 つの Encoder とそれぞれに対する注意機構をもつモデル．
- Single Attention: 佐々木さんの直前の発話用の Encoder 1 つと注意機構をもつモデル．吉田さんの一つ前の発話は考慮できない．
- No Attention: 佐々木さんの直前の発話用の Encoder を持つが，注意機構は持たないモデル．

Single Attention モデル, No Attention モデルは, Double Attention モデルをベースに性能を劣化させたモデルなので，生成される応答文の品質は順に低くなると予想されるが，それを確認するために主観評価を実施した．

主観評価には，評価データから無作為抽出した 1200 セットについて，各モデルが生成した応答文を用いた．まず応答の適切性について，1:応答が不適切，2:どちらともいえない，3:応答が適切の 3 段階で評価してもらった．次に，1:の不適切な応答について，1）文脈上不適切な内容語を含む誤り (ICW)，2）文末表現の誤り (ESE)，3) 助詞の誤り (FNC)，に該当するかチェックしてもらった．各モデル 1200 個の応答文について，1 応答文あたり 5 名の評価者に評価してもらったため，各モデル 6000 の主観評価値を得た． 応答文の適切性に関する 3 段階評価の割合を表 6 に示す．No Attention, Single Attention, Double Attention の順に，適切な応答は増加し，不適切な応答は減少した．ただし，適切な応答の差分は 2.5%に収まった．評価者が不適切と判定した応答文に対して該当するとされた 3 種類の誤りラベルの割合を表 7 に示す．No Attention, Single Attention, Double Attention の順に，文脈上不適切な内容語を含む誤り (ICW) は減少したが，助詞の誤り (FNC) は逆に増加した．文末表現の誤り (ESE) は 3 種類のモデルで僅かな差であった．

表 7　不適切と判定された応答文に対して評価者が該当するとした 3 種類の誤りラベルの割合

	ICW	ESE	FNC
No Attention	22.5%	5.2%	2.9%
Single Attention	22.0%	5.0%	3.3%
Double Attention	19.5%	4.9%	4.4%

表 8　対照応答対評価セットならびにサブセットに対する 3 種類のモデルの正解率

	ALL	ICW(EL)	ICW(ML)	ESE	FNC
No Attention	88.0%	94.9%	80.0%	88.3%	91.0%
Single Attention	88.9%	96.3%	81.1%	89.2%	90.0%
Double Attention	89.6%	94.6%	82.0%	92.2%	93.0%

5.2　対照応答対による評価結果比較

対照応答対評価セット全体と各サブセットに対する正解率を表 8 に示す．評価セット全体 (ALL) に対する正解率は，No Attention, Single Attention, Double Attention の順に上昇したが，サブセットに対する正解率は表 7 に示した主観評価による誤りラベルの割合と整合するものと整合しないものがあった．内容語を置換した対照応答対について，EL では正解率が高すぎて，モデルの性能を測るよい指標になっていないと考えられる．これに対して，ML では正解率が 80%前後となり，モデルの性能を測るのに適当な値を示している．文末表現を置換した対照応答対については，主観評価ではモデル間の差が小さかったが，正解率は Double Attention モデルが高かった．助詞を置換した対照応答対については，主観評価では Double Attention モデルの誤りが僅かに多かったが，正解率は Double Attention モデルが高かった．

6　おわりに

本稿では，ニューラル対話モデルの自動評価に向けた対照応答対評価セットを試作した．ニューラル対話モデルの性能を粗く評価できる感触は得られたが，評価セットのサイズを大きくするとともに，より多くのモデルを用いて検証を進める必要があると考えている．さらに，非タスク指向型対話モデルといってもドメインは様々に異なるため，対照応答対の自動生成を検討する予定である．

参考文献

[1]A. Sordoni, M. Galley, M. Auli, C. Brockett, Y. Ji, M. Mitchell, J. Nie, J. Gao and B. Dolan, “A Neural Network Approach to Context-Sensitive Generation of Conversational Responses”, in Proc. NAACL-HLT 2015, pp. 196-205, 2015.

[2]A. Vaswani, N. Shazeer, N. Parmer, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser and I. Polosukhin, “Attention Is All You Need”, in NIPS 2017, pp. 5998-6008, 2017.

[3]C. Liu, R. Lowe, I. Serban, M. Noseworthy, L. Charlin and J. Pineau, “How NOT to Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation”, in Proc. EMNLP 2016, pp. 2122-2132, 2016.

[4]C. Tao, L. Mou, D. Zhao and R. Yan, “RUBER: An Unsupervised Method for Automatic Evaluation of Open-Domain Dialog Systems” arXiv:1701.03079, 2017.

[5]T. Hashimoto, H. Zhang and P. Liang, “Unifying Human and Statistical Evaluation for Natural Language Generation”, in Proc. NAACL 2019, pp. 1689-1701, 2019.

[6]A.Ghandeharioun, J. Shen, N. Jaques, C. Ferguson, N. Jones, A. Lapedriza and R. Picard, “Approximating Interactive Human Evaluation with Self-Play for Open-Domain Dialog Systems”, in NIPS 2019, pp.13658-13669, 2019.

[7]R Sennrich, “How Grammatical is Character-level Neural Machine Translation? Assessing MT Quality with Contrastive Translation Pairs”, in Proc. EACL, pp.376-382, 2017.

[8]R. Bawden, R. Sennrich, A. Birch and B. Haddow, “Evaluating Discourse Phenomena in Neural Machine Translation”, in Proc. NAACL-HLT 2018, pp. 1304-1313, 2018.

[9]E. Voita, R. Sennrich and I. Timov, “When a Good Translation is Wrong in Context: Context-Aware Machine Translation Improves on Deixis, Ellipsis, and Lexical Cohesion”, in Proc. ACL 2019, pp.1198-1212, 2019.

[10]永田, 森下, “日本語から英語への文脈翻訳テストの提案”, 言語処理学会第 25 回年次大会, 2019.

[11]J. Li, M. Galley, C. Brockett, J. Gao and B. Dolan, “A Diversity-Prompting Objective Function for Neural Conversation Models, in Proc. NAACL-HLT 2016, pp. 110-119, 2016.

[12]鈴木, 加藤, 田村, 呉, 楊, 服部, “ニューラル対話モデルの自動評価に向けた応答文の誤り分析”, 電子情報通信学会総合大会 2021, 発表予定.