因果関係を用いた雑談対話応答におけるリランキングの評価

田中翔平 [1]　吉野幸一郎 [1,2]　須藤克仁 [1]　中村哲 [1]
[1] 奈良先端科学技術大学院大学　[2] 科学技術振興機構 さきがけ
{takana.shohei.tj7, koichiro, sudoh, s-nakamura}@is.naist.jp

1 はじめに

近年，Neural Conversational Model (NCM) [1] を始めとするニューラルネットワークに基づいた対話モデルが盛んに研究されている．しかし，こうした対話モデルから生成される応答はしばしばどのような場合にでも当てはまる単純なものであり，対話の文脈や論理を考慮した応答を生成することがむずかしい．

そこで本研究では，応答候補と対話履歴の因果関係に基づき，より文脈や論理を考慮した多様な応答を選択する手法を提案する．因果関係とは「ストレスが溜まる」→「発散する」など 2 つの事態間に原因と結果の関係が成立する場合を指す．本論文では，原因に相当する事態が発生すると結果に相当する事態が発生する確率が上昇することを指して，因果関係とする．

雑談対話における発話対に関しても因果関係が重要であることは徳久らの研究 [2] により知られており，特に対話を継続する働き（対話継続性）がある間接応答や問い返しにおいて，先行発話との因果関係が多く成立するとされている．この知見に基づき佐藤ら [3] は，因果関係を認定可能な発話対を選択したデータで NCM を学習することで，論理的で対話継続性に優れた応答を生成する手法を提案した．しかしこの手法は，対話モデルの学習に用いるデータを削減してしまうため，元の学習データ量によっては NCM の学習が困難となる．

そこで本研究では，大規模テキスト対話データで学習した NCM によって生成された N-best 応答候補を，因果関係に基づきリランキングする手法を考案した．また，リランキングにより文脈を考慮した多様な応答が選択されるかを評価した．実験の結果，提案する手法により，自然で文脈を考慮した多様な応答が生成されることが示された．

2 因果関係を用いたリランキング

本研究では因果関係辞書を利用して，対話履歴との因果関係を考慮した，より適切な応答選択を実現する．

2.1 ニューラル雑談対話モデル

NCM における処理の流れは次のとおりである．まず入力として対話履歴（ユーザ発話）が与えられ，NCM はこれに対する応答を生成する．ユーザ発話の単語列を $x_1, x_2, \ldots, x_n$ とし，モデルのパラメータ行列を W，バイアスを b とすると，最も単純なエンコーダデコーダの場合，エンコーダにおいては，各単語 x_t ごとに対応する隠れ層 h_t へ $h_t = tanh(W_{xh}x_t + W_{hh}h_{t-1} + b_h)$ によってユーザ発話が記憶される．この時，h_0 はゼロベクトルである．デコーダにおいては，応答発話の単語列 $y_1, y_2, \ldots, y_m$ における各単語 y_t が $y_t = softmax(W_{hy}h_t + b_y)$ によって与えられる．なお，デコーダにおける隠れ層は直前に生成した単語を用い，$h_t = tanh(W_{yh}y_{t-1} + W_{hh}h_{t-1} + b_h)$ によって計算される．対話履歴として直前 1 発話のみでなく，直前 N 発話が与えられる場合もある．また NCM のデコーダは各単語を逐次的に予測するため，ビームサーチやサンプリングなどを用いることで N-best 応答を生成することもできる [4]．

2.2 因果関係を用いたリランキング

図 1 に提案手法の概要を，図 2 にリランキングの例を示す．雑談対話モデルから生成された N-best 応答候補と対話履歴中の発話との間に，因果関係辞書 (2.3 節参照) にマッチするものがある場合，式 (1) に基づきスコアを再計算し，リランキングする．

$$l_{new} = \frac{l}{(\log_2 lift)^{\lambda}} \tag{1}$$

リランキングの際は応答候補と対話履歴のみに着目し，応答候補がどのような雑談対話モデルから生成されたかは考慮しない．ここで $l(<0)$ は対話モデルが算出する対数尤度であり，λ は因果関係がどの程度重視されるかを表すパラメータ，$lift$ (2.3 節参照) は因果関係に含まれる 2 つの事態間の相互情報量である．$\lambda = 0$ のとき，因果関係は全く考慮されない．ある応答候補と対話履歴中の発話との間に複数の因果関係が認められる場合，$lift$ の値が最も大きい因果関係のみを考慮する．ただし $lift$ は値域が広い $(10 < lift < 10,000)$ ため，対数をとった値を使用する．

2.3 因果関係辞書

因果関係辞書として，柴田ら [5] が提案した，共起情報と格フレームに基づき自動獲得された辞書を利用する．辞書は約 42 万件の因果関係知識で構成され，表 1 に示されるような情報を含む．各事態は述語項構造により表現され，述語 1 及び項 1 は原因となる事態を，述語 2 及び項 2 は結果となる事態を表す．ここで各事態は述語を必ず含むが，項（ガヲニデ格のいずれか）は含まない場合もある．また $support, confidence, lift$ はそれぞれ 2 つの事態の同時確率，原因となる事態が起

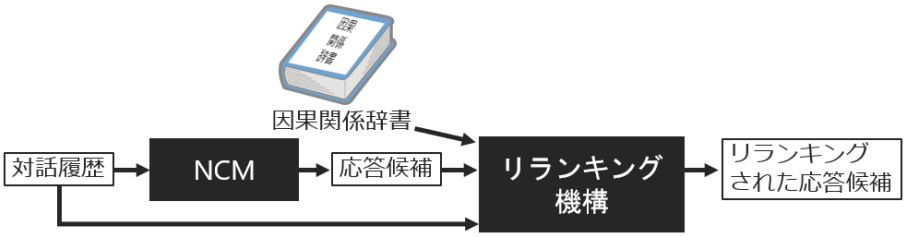

図 1　ニューラル雑談対話モデル + リランキング機構

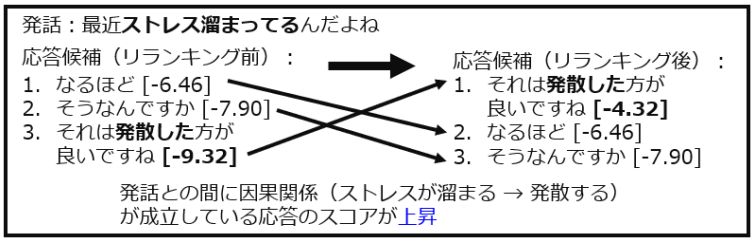

図 2　因果関係を用いたリランキング

こった場合に結果となる事態が生じる条件付き確率，2 つの事態間の相互情報量を表す．

3　実験

ここでは，因果関係を用いた雑談対話応答におけるリランキングの有効性を，定量的指標を用いて従来手法と比較することで評価する．これまでの NCM の研究では，Hierarchical Recurrent Encoder-decoder (HRED)[6] のように，NCM のモデル自体で履歴を考慮する取り組みがなされてきた．これに対し本研究は，NCM の生成結果に対して外部辞書を用いて履歴との一貫性を考慮するものである．HRED のモデルは，単純な Encoder-Decoder などのモデルより既に履歴を考慮した結果を生成する可能性がある一方で，出力結果のバリエーションが対話履歴により制約され，N-best のリランキングには不向きである可能性もある．

そこで実験では，直前 1 発話のみを利用する対話モデルとして Encoder-Decoder + Attention モデル（以下 “EncDec”）[7] を，直前 N 発話を利用する対話モデルとして HRED[6] を利用する．リランキングの際には，どちらのモデルが生成した N-best 応答に対しても直前 1 発話のみ，または直前 N 発話を考慮した場合を比較する．また，佐藤らの手法 [3] を用い，あらかじめ因果関係を認定可能な発話対のみをフィルタリングした学習データで学習した EncDec（以下 “Filtered”）とも比較する．

3.1　データセット

学習及びテストに用いるコーパスとしてマイクロブログ (twitter) から収集した 2,632,114 対話を使用した．平均対話長は 21.99 ターン，平均発話長は 22.08 文字である．絵文字などはあらかじめ発話から除外した．対話コーパスのうち学習データ，バリデーションデータ，テストデータとしてそれぞれ 2,509,836 対話，63,308 対話，58,970 対話を用いた．このうち学習データを佐藤らの手法によってフィルタリングしたところ，学習に利用可能な対話データは 62,700 対話 (2.50%) に減少した．

3.2　モデル設定

NCM の学習設定は次のとおりである．隠れ層 256 次元の GRU[8]2 層，対話履歴数 $n = 5$，バッチサイズ 100，Dropout 確率 0.1，teacher forcing 率 1.0，Optimizer を Adam[9] とし，Gradient Clipping 50，Encoder 及び HRED における Context RNN の学習率 $1e^{-4}$，Decoder の学習率 $5e^{-4}$，目的関数を ITF loss[10] とした．トークン分割には sentencepiece[11] を用い，語彙数 32,000 とした．これらの設定は EncDec，HRED，Filtered 全てにおいて同一である．

応答生成時はビームサーチを用い，ビーム幅は 10，20 の場合をそれぞれ比較した．探索の際 Repetitive Suppression[10] を用い，EncDec，Filtered では更に length normalization[4] を用いた．最後に，リランキングの式 (1) における $\lambda = 1$ とした．

3.3　ビームサーチ結果の多様性

リランキング以前の各モデルにおける N-best 応答内の多様性評価を行う．これは，ビームサーチにより生成される N-best 応答が多様であるほど，リランキングの効果が期待できるためである．そこで，テストデータに対するそれぞれの N-best 応答内の多様性を，dist-1, 2[12] によって評価した．

表 2 に結果を示す．ここで Ave.dist は各 N-best 応答内で計算された dist の平均を表す．beam はビーム幅を示し，10, 20 どちらの場合においても EncDec の方が HRED よりも多様性が高いものの，大きな差は見られないことが分かる．また EncDec と Filtered を比較すると，学習データが減少することにより N-best 応答内の多様性も減少することが確認できる．

表 1　因果関係辞書に含まれる因果関係知識の一例

述語 1	項 1	述語 2	項 2	$support$	$confidence$	$lift$
溜まる	ガ:ストレス	発散	-	$2e^{-7}$	0.0099	10.02

表 2　N-best 応答内の多様性

	Ave.dist-1	Ave.dist-2
EncDec（beam 10）	**0.49**	**0.61**
HRED（beam 10）	0.42	0.52
Filtered（beam 10）	0.41	0.53
EncDec（beam 20）	**0.38**	**0.50**
HRED（beam 20）	0.35	0.46
Filtered（beam 20）	0.28	0.40

3.4　リランキングされた割合

表 3 に，提案手法によりリランキングされた応答候補の割合を示す．ここで Reference は，テストデータ中の実際の応答と対話履歴との間で因果関係が認められるものの割合を表し，history は直前何発話までとの因果関係を考慮するかを表す．All はいずれかの手法でリランキングされた応答の和集合である．今回提案した手法では，リランキングの際，因果関係の順序は考慮していない．また発話応答中に含まれる述語項構造の抽出には KNP[13] を用いた．

ビーム幅 10, 20 どちらの場合でも，HRED (history N) において最も多くの応答がリランキングされている．これは HRED は生成時においても直前 N 発話を考慮しているため，N-best 応答候補内に先行発話との因果関係が成立するものが含まれやすいのだと考えられる．

また Reference と比較すると，EncDec (history N, beam 10) のみ Reference 未満の割合の応答がリランキングされているものの，その他の場合は Reference における値以上の割合の応答がリランキングされていることが分かる．ただしリランキングされた応答の割合は最大でも 12% 程度であり，応答の大多数はリランキングされていない．

表 3　リランキングされた応答の割合

	Reranked (%)
Reference（history 1）	1,566 (2.66)
Reference（history N）	2,681 (4.55)
EncDec（history 1, beam 10）	2,258 (3.91)
EncDec（history N, beam 10）	2,518 (4.36)
HRED（history 1, beam 10）	3,716 (6.44)
HRED（history N, beam 10）	**4,465(7.74)**
All（beam 10）	7,114 (12.32)
EncDec（history 1, beam 20）	3,231 (5.60)
EncDec（history N, beam 20）	3,921 (6.79)
HRED（history 1, beam 20）	5,401 (9.36)
HRED（history N, beam 20）	**6,936(12.02)**
All（beam 20）	10,433 (18.07)

3.5　リランキング前後の比較

表 4 にリランキング前後及び Filtered との比較を示す．表では 3.4 節における All に含まれる応答についてのみの評価を行っている．評価には Reference に対する BLEU と dist，リランキングされた応答の平均応答長を用いた．手法名は左から順に，用いた NCM，リランキングに考慮した履歴の範囲，N-best 候補生成時に用いたビーム幅を示している．Reference はテストデータ中の実際の応答から算出した値である．

ビーム幅 10 の場合 BLEU の値は HRED (history N) が最も高く，リランキング前と比較して 0.41 上昇している．BLEU は実際の応答とどの程度 N-gram が類似しているかを示す値であり，実際の応答は対話の文脈や論理を考慮していると仮定できることから，BLEU の値が高いことはその応答が文脈や論理を考慮しているかどうかをある程度示すと考えられる．また dist-1, 2 は EncDec (history N) が最も高いものの，リランキング前後で大きな変化は見られない．しかし HRED の場合リランキング前後で dist の値はある程度上昇している．

ビーム幅 20 の場合 BLEU, dist-1, 2 は EncDec (history N) が最も高く，BLEU はリランキング前と比較して 0.23 上昇している．ただしビーム幅 10 の場合と同様に dist の値は大きく変化していない．HRED の場合リランキング前後で dist の値がある程度上昇する点も同様である．

平均応答長に関してはリランキング前後で大幅な差は見られないものの，EncDec ではわずかに減少し，HRED ではわずかに上昇することが確認できた．応答長が長いほどその応答は単純ではない場合が多いと考えられる．ビーム幅 10 の場合 HRED（history N）が，20 の場合 Filtered が最も長い平均応答長を持つ．しかし Filtered と EncDec を比較するとビーム幅 10, 20 どちらの場合においても BLEU, dist の値が減少しており，特に dist の値が大きく減少している．これは応答長は長いもののどんな発話に対しても同じような，単純で汎用的な応答を生成していることを示している．

以上のことから，提案手法により文脈や論理を考慮した多様な応答が選択できることが確認できた．また学習データが減少すると対話モデルの学習が困難になり，文脈や論理を考慮した多様な応答の生成が難しくなることが分かった．最後に，提案手法によりリランキングされた応答の例を表 5 に示す．対話モデルは EncDec（history N, beam 20）であり，リランキング後応答の括弧はリランキング前の順位を表す．

表 4　リランキング前，リランキング後，フィルタリングの客観評価における比較

	BLEU	dist-1	dist-2	平均応答長
Reference（beam 10）	-	0.22	0.72	22.17
EncDec（no reranking, beam 10）	0.86	0.22	0.37	11.49
EncDec（history 1, beam 10）	0.88	**0.23**	0.39	11.46
EncDec（history N, beam 10）	0.95	**0.23**	**0.40**	11.39
HRED（no reranking, beam 10）	0.69	0.09	0.13	17.39
HRED（history 1, beam 10）	0.91	0.12	0.19	18.24
HRED（history N, beam 10）	**1.10**	0.15	0.23	**19.13**
Filtered（beam 10）	0.81	0.07	0.21	13.17
Reference（beam 20）	-	0.18	0.70	21.90
EncDec（no reranking, beam 20）	1.46	0.14	0.33	15.78
EncDec（history 1, beam 20）	1.67	**0.15**	0.35	15.59
EncDec（history N, beam 20）	**1.69**	**0.15**	**0.36**	15.33
HRED（no reranking, beam 20）	0.84	0.06	0.09	16.69
HRED（history 1, beam 20）	0.98	0.10	0.15	17.38
HRED（history N, beam 20）	1.05	0.13	0.20	17.69
Filtered（beam 20）	1.14	0.04	0.16	**17.98**

表 5　リランキング結果の例

発話	リランキング前の応答	リランキング後の応答	因果関係
リツイートありがとうフォローしますね	いえいえありがとうございます	フォローありがとうございますフォロバしました (2)	フォローする → フォロバする
徹頭徹尾夜な夜なドライブ	はい	明日休みだから (13)	休み → ドライブする

4　おわりに

本論文では，ニューラル雑談対話モデル (NCM) により生成された N-best 応答を，因果関係を用いてリランキングする手法を提案した．Reference との比較による評価の結果，因果関係を用いたリランキングにより，文脈や論理を考慮した多様な応答が選択できることを確認した．しかし本手法によりリランキング可能な応答の割合はやや低かった．今後の課題として因果関係辞書を汎化し，リランキングできる応答の割合を高めることが挙げられる．また，今後大規模な被験者実験により，実際にユーザが良いと感じる因果関係の考慮ができているかを確認する必要がある．

謝辞

本研究で使用した因果関係辞書をご提供頂いた京都大学黒橋研究室の黒橋先生，柴田先生に感謝いたします．

本研究は JST さきがけ（JPMJPR165B）の支援を受けた．

参考文献

[1] Vinyals et al. A Neural Conversational Model. In *Proc. ICML-DLW*, 2015.
[2] 徳久ら．非課題遂行対話における発話の特徴とその分析．人工知能学会論文誌, Vol. 4, pp. 425–435, 2007.
[3] 佐藤ら．因果関係に基づくデータサンプリングを利用した雑談応答学習．言語処理学会 第 24 回年次大会 発表論文集, pp. 1219–1222, 2018.
[4] Wu et al. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. In *arXiv:1609.08144*, 2016.
[5] 柴田ら．述語項構造の共起情報と格フレームを用いた事態間知識の自動獲得．情報処理学会 自然言語処理研究会, Vol. 2011-NL-203, No. 2, pp. 1–8, 2011.
[6] Serban et al. Building End-To-End Dialogue Systems Using Generative Hierarchical Neural Network Models. In *Proc. AAAI*, 2016.
[7] Luong et al. Effective Approaches to Attention-based Neural Machine Translation. In *Proc. EMNLP*, 2015.
[8] Cho et al. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. In *Proc. EMNLP*, 2014.
[9] Kingma et al. Adam: A Method for Stochastic Optimization. In *Proc. ICLR*, 2015.
[10] Nakamura et al. Another Diversity-Promoting Objective Function for Neural Dialogue Generation. In *arXiv:1811.08100*, 2019.
[11] Kudo et al. SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. In *Proc. EMNLP*, 2018.
[12] Li et al. A Diversity-Promoting Objective Function for Neural Conversation Models. In *Proc. NAACL*, 2016.
[13] Kawahara et al. A fully-lexicalized probabilistic model for japanese syntactic and case structure analysis. *Proc. HLT-NAACL*, pp. 176–183, 2006.