Vol.2019-NL-241 No.12
2019/8/29

因果関係と事態分散表現を用いた雑談対話応答のリランキングにおける傾向分析

田中 翔平[1,a)]　吉野 幸一郎[1,2,b)]　須藤 克仁[1,c)]　中村 哲[1,d)]

概要：本論文では，対話履歴に対し一貫した多様な応答を選択する手法を提案する．提案手法では対話履歴に対する一貫性を保つため，対話モデルより生成された応答候補を，対話履歴と応答候補の間に存在する因果関係（ストレスが溜まる → 発散する，など）を用いてリランキングする．この際，因果関係の認定には統計的に獲得された因果関係ペアを用いるが，対話中に存在する全ての因果関係を被覆するような辞書を用意することは難しい．そこで，Role Factored Tensor Model を用いて事態を分散表現に変換することで，因果関係知識のカバレージを向上させ，因果関係知識と対話中の因果関係の頑強なマッチングを実現した．自動評価，人手評価の結果，提案手法は応答の一貫性や対話継続性を向上させることが確認できた．一方で，事態の過汎化に由来する応答の自然性低下が見られる場合もあった．これらの問題についても例示し，解決の方向性について論じる．

1. はじめに

近年，Neural Conversational Model (NCM) [32] 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はしばしばどのような場合にでも当てはまる単純な応答を生成し，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．そこで本論文では，応答候補と対話履歴に存在する因果関係に基づき，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 [29], [30]．因果関係はこれまで質問応答システムなどで利用されており，質問と応答の間に成立する因果関係を考慮することで，質問に対する適切な応答を生成できることが示されている [21], [22], [23]．雑談対話システムにおいても因果関係を考慮することで，多様で文脈に沿った応答を生成できることが示されている [6]．しかし雑談対話システムにとって重要な対話を継続する働き（対話継続性）が向上するかは示されていない．

そこで本論文では，NCM によって生成された N-best 応答候補より，文脈に沿った，対話継続性の高い応答を選択する手法を考案する．この手法は対話履歴との間に因果関係が成立する応答を選択するために，因果関係を考慮したスコアの計算を行い，これに基づいて応答候補から応答を選択する．こうしたリランキングは，質問応答システムや対話システムなどの言語生成タスクにおいて様々な要素を考慮した候補の選択に用いられる [2], [8], [22], [24]．本研究では因果関係の考慮を行うため，大規模コーパスから統計的に獲得された因果関係ペア [29], [30] を用いた．この際，単純にこれらのペアを用いるとカバレージの問題が生じるため，Role Factored Tensor Model [33] を用いた事態の分散表現によって汎化を行った．自動評価及び人手評価の結果，提案する手法は文脈に沿った，論理的かつ多様で対話継続性の高い応答を選択できることが示された．

2. 因果関係を用いた応答のリランキング

図 1 に提案手法の概要を示す．提案手法は大きく分けて 4 つのパートから構成される．まず対話履歴をもとに既存の NCM モデルから N-best 応答候補を生成する（図 1 ①; 2.1 節）．次に対話履歴と応答候補に含まれる事態（述語項構造）を事態パーサーを用いて抽出する（図 1 ②）．この事態パーサーには KNP[*1] [9], [27] を用いる．その後，抽

[1] 奈良先端科学技術大学院大学
NAIST, Takayama-cho 8916-5, Ikoma-shi, Nara 630–0192, Japan
[2] 科学技術振興機構さきがけ
PRESTO, Japan Science and Technology Agency
[a)] tanaka.shohei.tj7@is.naist.jp
[b)] koichiro@is.naist.jp
[c)] sudoh@is.naist.jp
[d)] s-nakamura@is.naist.jp

[*1] http://nlp.ist.i.kyoto-u.ac.jp/?KNP

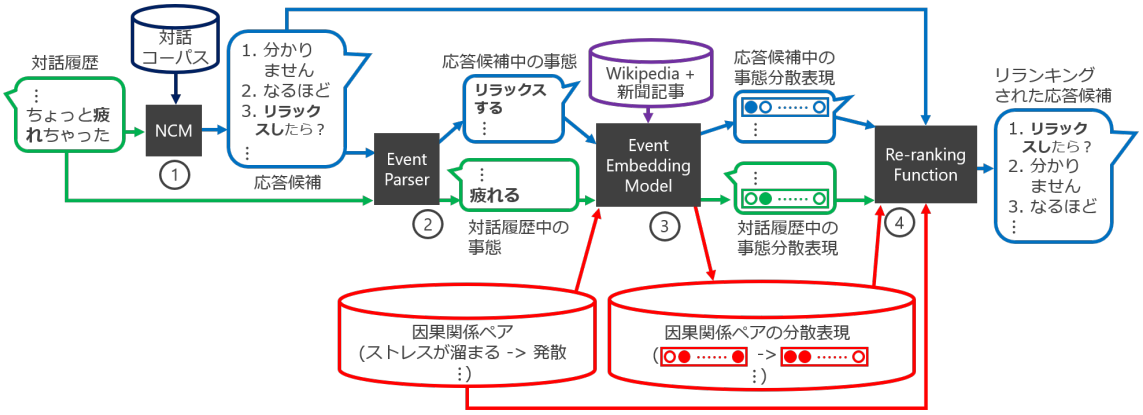

図 1 Neural Conversational Model+ 因果関係を用いたリランキング;「疲れる」→「リラックスする」という因果関係が対話履歴との間に成立する応答がリランキングにより選択される.

表 1 因果関係の一例

述語 1	項 1	述語 2	項 2	$lift$
溜まる	ガ:ストレス	発散	-	10.02

出した事態及び統計的に獲得された因果関係ペア [29], [30] を事態埋め込みモデルを用いて分散表現に変換する（図 1 ③; 2.3 節）. 事態埋め込みモデルとして，RFTM を利用する. 最後に応答候補を因果関係に基づきリランキングする（図 1 ④; 2.2, 2.4 節）.

2.1 Neural Conversational Model (NCM)

. NCM は入力系列と出力系列のマッピングを Recurrent Neural Network (RNN) により学習するものである. 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 [15]. 本研究ではビームサーチにより生成された N-best 候補を用いる.

2.2 因果関係ペア

柴田ら [29], [30] 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因果関係ペアをリランキングに用いる. 16 億文のテキストから約 42 万件の因果関係知識が抽出され，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 $lift$ は，2 つの事態間の因果関係としての結びつきの強さを表す相互情報量である. $lift$ を用い，リランキングのためのスコアの計算を次のように定義する.

$$score = \max_{<\mathrm{e}_h,\mathrm{e}_r>} \frac{\log_2 p}{\left(\log_2 lift(\mathrm{e}_h,\mathrm{e}_r)\right)^{\lambda}}. \qquad (1)$$

p は NCM より与えられる各応答候補の事後確率であり，λ は因果関係の重みを決定するハイパーパラメータである. $lift(\mathrm{e}_h,\mathrm{e}_r)$ は対話履歴中の事態 e_h と応答候補中の事態 e_r との間の $lift$ の値である. この事態ペアが因果関係ペアに含まれない場合，$lift$ の値は 2 である. ただし $lift(\mathrm{e}_h,\mathrm{e}_r)$ は値域が広い $(10 < lift(\mathrm{e}_h,\mathrm{e}_r) < 10,000)$ ため，対数をとった値を使用する. 応答候補と対話履歴との間に複数の因果関係が認められる場合，$lift(\mathrm{e}_h,\mathrm{e}_r)$ の値が最も大きい因果関係のみを考慮する. このモデルを “Re-ranking” と呼ぶ.

2.3 Role Factored Tensor Model (RFTM) に基づく事態分散表現

統計的に獲得された因果関係ペアには網羅性の問題が存在し，これのみを用いて対話履歴と応答候補に存在する全ての因果関係を考慮することは難しい. そこで因果関係ペア，および発話中に含まれる事態を分散表現に変換し，因果関係ペアと対話中に含まれる因果関係との頑強なマッチングを実現する. 各事態をベクトルに変換することで，事態の類似度を計算する.

本論文では，事態を述語，もしくは述語と付随する格要素のペアと定義して用いる. 格要素 a は Skip-gram [16], [17], [18] によりベクトル v_a へと変換される. 述語 p は predicate embedding によりベクトル v_p へと変換される. predicate embedding は Skip-gram をもとにした単語分散表現である. 図 2 に predicate embedding モデルの概要を示す. このモデルは与えられた述語に付随する格要素を予測するよう学習を行う. v_p および v_a より事態の分散表現を得る手法として，Weber ら [33] が提案した RFTM を利用する. RFTM は述語と項を次式により事態分散表現 e へと変換する.

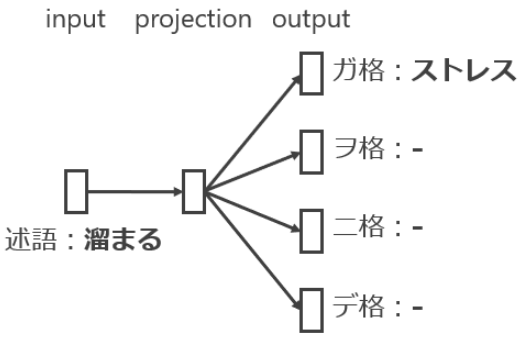

図 2 Predicate Embedding

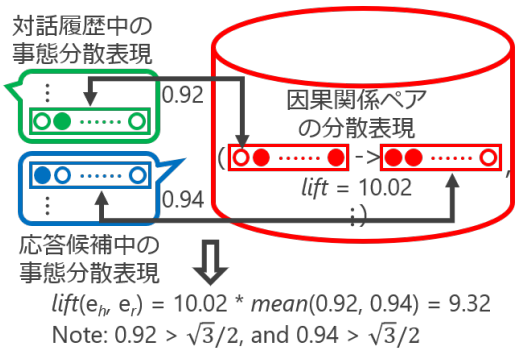

図 3 因果関係のマッチング;「疲れる」→「リラックスする」という因果関係の $lift$ は最もコサイン類似度が高い因果関係である「ストレスが溜まる」→「発散」の $lift$ から計算される.

$$\mathrm{e} = \sum_a W_a T(\mathrm{v}_p, \mathrm{v}_a). \tag{2}$$

述語と付随する格要素の関係は 3 次元パラメータテンソル T, パラメータ行列 W_a により計算される. 述語が格要素を持たない場合, e は v_p により代替される. RFTM は学習の目標として連続して起こる事態を予測する. これは分布仮説同様, 似た事態が似た文脈を持つことを仮定するものである. これにより, 文脈に沿った事態の意味を捉えることが可能である.

2.4 事態分散表現を用いた因果関係のマッチング

図 3 に事態分散表現による事態のマッチングの手続きを示す. 提案手法は事態分散表現に基づき, 応答候補と対話履歴中の発話との間の事態ペアに対し, 最も高いコサイン類似度を持つ因果関係を因果関係ペアより選択する. ここで 2 つの事態間の $lift_{emb}$ は次の式のように定義される.

$$lift_{emb}(\mathrm{e}_h, \mathrm{e}_r) = lift(\mathrm{e}_c, \mathrm{e}_e) * mean\big(sim(\mathrm{e}_h, \mathrm{e}_c), sim(\mathrm{e}_r, \mathrm{e}_e)\big). \tag{3}$$

e_h は対話履歴中の事態, e_r は応答候補中の事態であり, e_c と e_e はそれぞれ因果関係ペア中の原因となる事態, 結果となる事態である. 提案手法では対話履歴中の事態が結果, 応答候補中の事態が原因となる場合も考慮する. ただし事態を過剰に汎化してしまうことを避けるために, 各

表 2 N-best 応答候補内の多様性

	Ave.dist-1	Ave.dist-2
EncDec	**0.44**	**0.56**
HRED	0.33	0.42

sim はしきい値 ($=\sqrt{3}/2$) を持つ. 式 (1) の $lift(\mathrm{e}_h, \mathrm{e}_r)$ を $lift_{emb}(\mathrm{e}_h, \mathrm{e}_r)$ で更新することで, 事態分散表現を用いたリランキングスコアは次のように定義される.

$$score = \max_{<\mathrm{e}_h, \mathrm{e}_r>} \frac{\log_2 p}{\big(\log_2 lift_{emb}(\mathrm{e}_h, \mathrm{e}_r)\big)^{\lambda}}. \tag{4}$$

このモデルを “Re-ranking (emb)” と呼ぶ.

3. 実験

ここでは, 自動評価, 人手評価によりリランキングを行わない場合と行う場合を比較し, 因果関係を用いたリランキングの有効性を検証する. 実験では NCM として Encoder-Decoder with Attention (EncDec) [1], [14] と Hierarchical Recurrent Encoder-Decoder (HRED) [28], [31] を用いる. HRED のモデルは, 単純な Encoder-Decoder などのモデルより既に履歴との関係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 応答候補のリランキングには不向きである可能性もある.

RFTM が利用する Skip-gram, predicate embedding の学習には日本語 Wikipedia ダンプデータを用い, RFTM の学習には毎日新聞 2017 データ集 [*2] を用いた. 対話モデルの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. 語彙サイズを削減しモデルの学習を促進するために, 絵文字などはあらかじめ発話から除外した. 対話コーパスを学習データ, 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 対話に分割した.

3.1 モデル設定

NCM の学習設定は次のとおりである. Skip-gram [16], [17], [18], predicate embedding, RFTM [33] の隠れ層は全て 100 次元とした. 隠れ層 256 次元の GRU [3], [4] 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 [10] 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 [19] 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 どちらにおいても同一である.

[*2] http://www.nichigai.co.jp/sales/mainichi/mainichi-data.html

表 3　自動評価のスコア

Method			Evaluation							
NCM	history	re-ranking	re-ranked (%)	BLEU	NIST	extrema	dist-1	dist-2	PMI	length
reference	-	-	-	-	-	-	0.06	0.40	1.86	21.43
EncDec	-	1-best	-	1.12	1.19	**0.42**	0.06	0.18	1.77	15.55
EncDec	1	Re-ranking	4,016 (7.90)	1.10	1.18	**0.42**	0.06	0.19	1.78	15.52
EncDec	1	Re-ranking (emb)	29,343 (57.71)	1.02	1.07	0.40	0.06	0.20	1.77	15.64
EncDec	5	Re-ranking	6,469 (12.72)	1.09	1.17	**0.42**	0.06	0.19	1.78	15.50
EncDec	5	Re-ranking (emb)	35,284 (69.39)	1.00	1.04	0.39	**0.07**	**0.21**	1.77	15.66
HRED	-	1-best	-	**1.34**	**2.74**	**0.42**	**0.07**	0.20	1.84	35.05
HRED	1	Re-ranking	3,671 (7.22)	1.33	**2.74**	**0.42**	0.06	0.20	1.84	35.20
HRED	1	Re-ranking (emb)	30,992 (60.95)	1.28	**2.74**	0.41	0.06	0.20	**1.86**	34.80
HRED	5	Re-ranking	6,231 (12.25)	1.33	2.73	**0.42**	0.06	0.20	1.84	**35.30**
HRED	5	Re-ranking (emb)	**36, 373(71.53)**	1.28	**2.74**	0.41	0.06	0.20	**1.86**	34.60

探索の際は Repetitive Suppression [19], length normalization [15] を用いた．最後に，リランキングの式 (1)，式 (4) における $\lambda = 1.0$, コサイン類似度のしきい値は $\sqrt{3}/2$ とした．

3.2　ビームサーチの多様性

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．ビーム幅は 20 とし，これは続く実験でも同一である．

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．表を見ると EncDec の方が HRED よりも多様性が高いことが分かる．

3.3　自動評価による比較

表 3 にリランキング前後の客観評価による比較を示す．評価には提案手法によりリランキングされた応答候補の割合 (“re-ranked”)，reference に対する BLEU [25], NIST [5], vector extrema [7] (“extrema”) を用いた．NIST は BLEU に類似した評価指標だが，スコア計算の際に低頻度語を重視する．Vector extrema は参照応答と生成応答の文ベクトルのコサイン類似度を測るものである．各文ベクトル e_s は Skip-gram 単語ベクトル e_w の各次元 d における極値を次式のように取ることで計算される．

$$e_{sd} = \begin{cases} \max_{w \in s} e_{wd} & \text{if } \ e_{wd} > |\min_{w' \in s} e_{w'd}| \\ \min_{w \in s} e_{wd} & \text{otherwise} \end{cases}. \quad (5)$$

ここで e_{sd}, e_{wd} はそれぞれ e_s, e_w の d 次元目の要素を表す．また評価指標として，dist [12], Pointwise Mutual Information (PMI) [20]，リランキングされた応答の平均応答長 (“length”) も用いた．Dists, PMI はそれぞれ応答の多様性，一貫性を測るために用いた．応答と対話履歴の PMI は次式のように計算される．

$$\mathrm{PMI} = \frac{1}{|response|} \sum_{wr}^{|response|} \max_{wh} \mathrm{PMI}(wr, wh). \quad (6)$$

wr, wh はそれぞれ応答中，対話履歴中の単語を表す．表 3 の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，用いたリランキング手法を示している．“1-best” はリランキングを行わない，ベースラインの NCM を表す．“Re-ranking”, “Re-ranking (emb)” はそれぞれ分散表現を用いないリランキング，分散表現を用いるリランキングを表す．

この結果から，リランキングにより参照応答との類似度を測る評価値はすべて減少してしまうことがわかる．これは通常の Encoder-Decoder のようにクロスエントロピーを目的関数として用いると，モデルが生成した N-best 応答のうち一位の応答が最も参照応答と類似度が高くなると考えられるからである．これに対し，dist-2 および PMI はリランキングにより上昇しており，語彙の組み合わせが多様かつ対話履歴と関連したものになっていることが分かる．しかし，リランキングされる応答候補の割合は多くとも 10% 前後に留まり，リランキングの効果が限定的となる．これに対して，RFTM による分散表現で汎化を行ったモデルでは，リランキングの効果が 50-70% の発話とかなり広範囲に広がることが見てとれる (Re-ranking vs. Re-ranking (emb))．dist-1, dist-2, PMI は分散表現を用いたモデルが最大となっている．そこで，実際にどの程度良い候補が選択されるようになったかを確認するため，次の節でベースラインモデルと提案モデルの人手評価による比較を行う．HRED は EncDec と比較してどのリランキング手法の場合でも BLEU, NIST, PMI に関して高いスコアを持つため，人手評価において用いる NCM は HRED とした．

3.4　人手評価による比較

自動評価のみで対話システムの性能を評価することは困難である [13]．そこで，ベースラインモデルと提案モデル

表 4 1-best v.s. Re-ranking; 対話数 100.

	word coherency	dialogue continuity
1-best	28.62	**40.84**
Re-ranking	**33.91**	38.53
neither	37.47	20.62

表 5 1-best v.s. Re-ranking (emb); 対話数 100.

	word coherency	dialogue continuity
1-best	**30.10**	35.50
Re-ranking (emb)	25.40	**38.20**
neither	44.50	26.30

表 6 Re-ranking v.s. Re-ranking (emb); 対話数 100.

	word coherency	dialogue continuity
Re-ranking	**23.70**	35.53
Re-ranking (emb)	22.91	**35.65**
neither	55.39	28.83

を人手評価により比較することで，提案モデルにより選択された応答の一貫性，対話継続性を測った．ベースラインとして HRED を用い，提案モデルのリランキング手法として分散表現を用いない場合，用いる場合を選択した．評価者の負担を軽減するため，対話履歴は最大で過去 2 ユーザ発話まで表示し，内容を理解するために外部知識を必要する対話は評価対象から取り除いた．人手評価にはクラウドソーシングを用い，十人のクラウドワーカーに 2 つのシステムの応答を比較し，次に挙げる 2 点の指標をより満たすものを選択してもらった．1 番目の指標は「どちらの応答に含まれる単語がより対話履歴に関連しているか (word coherency)」であり，これはシステム応答が一貫しているかを計測するために用いる．2 番目の指標は「どちらの応答により返答したいと思うか (dialogue continuity)」であり，これはシステム応答の対話継続性が高いかを計測するために用いる．これらの指標は Alexa Prize [26] を参考に決定した．

人手評価の結果を表 4, 5, 6 に示す．単語の一貫性は分散表現を用いないモデルで上昇している一方，分散表現を用いるモデルでは減少している．これは因果関係ペアでもともと因果関係と認められているものは一貫性の改善に役立つものの，汎化された因果関係には因果関係と認めにくいものが多く含まれてしまうからだと考えられる．しかしながら，対話継続性は分散表現を用いるリランキングにおいて向上しており，単純でつまらない応答の割合が減少していることがわかる．よって今後の課題として，一貫性を向上させつつ文全体の自然性を保つ適切なしきい値を探索する必要がある．

表 7 リランキング結果の分類 (Re-ranking)

Re-ranking / Causality	Good	Bad (pairs)	Sum
Good	14	12	26
Bad	14	10	24
Both Good	12	7	19
Both Bad	11	20	31
Sum	51	49	100

表 8 リランキング結果の分類 (Re-ranking (emb))

Re-ranking / Causality	Good	Bad (pairs)	Bad (over-generalization)	Sum
Good	2	0	6	8
Bad	2	2	5	9
Both Good	0	0	14	14
Both Bad	0	4	65	69
Sum	4	6	90	100

4. リランキング結果の事例分析

リランキングが適切である割合を測るために，リランキング結果およびリランキングに用いられた因果関係の妥当性を分類，分析した．事例の分類は一人のアノテータが行い，評価対象は表 4, 5 の実験で用いられた各 100 対話とした．分類結果を表 7, 8 に示す．表 7 は事態分散表現を用いないリランキング，表 8 は事態分散表現を用いるリランキングに関する分類である．ここで縦軸はリランキングの妥当性を表し，“Good” はリランキング後の応答の対話履歴に対する一貫性がリランキング前の応答と比較して向上していることを，“Bad” は低下していることを意味する．また “Both Good” はリランキング前後どちらの応答も一貫性が高いことを，“Both Bad” はどちらの応答も一貫性が低いことを意味する．横軸はリランキングに用いられた因果関係の妥当性を表し，“Good” は妥当な因果関係が用いられたことを，“Bad (pairs)” は因果関係ペアに含まれる，妥当でない因果関係が用いられたことを意味する．また “Bad (over-generalization)” は事態分散表現により過汎化された因果関係が用いられたことを意味する．

“Good” と “Both Good” を合算したものをリランキングが妥当な場合と定義すると，その割合は分散表現を用いない場合 50%前後，分散表現を用いる場合 20%前後である．妥当な因果関係が用いられた割合は，分散表現を用いない場合 50%前後，分散表現を用いる場合 5%前後である．妥当な因果関係により，適切なリランキングが行われている例を下に示す．“[]” はリランキングに使用された因果関係を，“<>” はリランキング前の応答を示す．

対話 1:

ユーザ 1： もう不安な事が多すぎて**ストレスが溜ま**ってく.
ユーザ 2（システム）： 大丈夫ですか**無理し**ないでくださいね
[無理をする → ストレスが溜まる] < 大丈夫ですか >

対話 2:

ユーザ 1： だれか，明日，弥生が丘周辺で昼ごはん食べに**行こ**ー誰でもいいけんお願い
ユーザ 2： いきますかー
ユーザ 1： どこに食い行く
ユーザ 2（システム）： 新宿駅の**近く**
[近く → 行ける]
< どこいんの >

上記の例では，因果関係を用いた妥当なリランキングにより，論理的で一貫した応答が選択されている．しかし表 7, 8 からわかるように，このような例は多数派ではない．次に示す対話 3 のように妥当でない因果関係が用いられている場合や，対話 4, 5, 6 のように，妥当な因果関係が使用されていても応答が不自然である場合が多く見られた.

対話 3:

ユーザ 1： **お酒を飲め**ない社長がビールをくれる ...
ユーザ 2（システム）： **お酒飲む**
[お酒を飲む → お酒を飲める]
< わかる >

対話 4:

ユーザ 1： 変な**咳が止まら**ない ... 涙
ユーザ 2（システム）： **風邪ひか**ないでね
[風邪をひく → 咳が止まる]
< 大丈夫か >

対話 5:

ユーザ 1： **朝**なのに，救急車とパトカーの音がすごーい ... どうしたんだろ
ユーザ 2（システム）： **おはよう**ございます
[朝 → おはよう]
< 大丈夫ですか >

対話 6:

ユーザ 1： お，起き上がれない ... 会社**休む** ... 熱 ...
ユーザ 2（システム）： お**疲れ**様です
[疲れ → 休む]
< お大事に >

人手評価および分類結果を鑑みると，提案手法には妥当なリランキングを行うために解決すべき 2 つの問題があると考えられる．1 つ目の問題は対話 3 のような事態の過汎化である．対話 3 で使用されている「お酒を飲む → お酒を飲める」という因果関係は因果関係ペア中の「レストランに入る → ビールを頼む」という因果関係を汎化することで得られている．過汎化の割合は表 8 が示すとおり 90%前後であり，事態分散表現を改善することでこのような過汎化を防ぐ必要がある．2 つ目の問題は対話 4, 5, 6 で見られるように，提案手法は単語の一貫性のみに着目し，応答の自然性を考慮していない点である．この問題を解決するためには，リランキングの際に応答の自然性も考慮するなどの対策を講じ，応答の自然性を保ちつつ，一貫した単語の選択を行う必要がある.

5. 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．提案手法は因果関係を構成する事態を分散表現に変換するため，雑談対話における多くの場面でリランキングを適用可能である．実験の結果，因果関係を用いたリランキングにより，一貫した多様な応答が選択できることを確認した．提案手法は述語項構造で表現された事態に基づいているため，構文解析器を持つあらゆる言語に適用可能である．一方で過汎化された因果関係を用いたために，不自然な応答が選択される場面が存在することも判明した．今後は過汎化を防ぐよう事態分散表現を改善し，応答の自然性を保つリランキング手法を考案していく.

謝辞

本研究で使用した因果関係ペアをご提供頂いた京都大学黒橋研究室の黒橋教授，柴田助教に感謝いたします.

本研究は JST さきがけ (JPMJPR165B) の支援を受けた.

参考文献

[1] Bahdanau, D., Cho, K. and Bengio, Y.: Neural Machine Translation by Jointly Learning to Align and Translate, *Proceedings of the 3rd International Conference on Learning Representations (ICLR)* (2015).

[2] Bogdanova, D. and Foster, J.: This is how we do it: Answer Reranking for Open-Domain How Questions with Paragraph Vectors and Minimal Feature Engineering, *Proceedings of the 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*

(NAACL-HLT), pp. 1290–1295 (2016).

[3] Cho, K., Merrienboer, B. v., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H. and Bengio, Y.: Learning Phrase Representations Using RNN Encoder-Decoder for Statistical Machine Translation, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (2014).

[4] Chung, J., Gulcehre, C., Cho, K. and Bengio, Y.: Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling, *Proceedings of the 28th Conference Neural Information Processing Systems, Deep Learning and Representation Learning Workshop (NIPS)* (2014).

[5] Doddington, G.: Automatic Evaluation of Machine Translation Quality Using N-gram Co-occurrence Statistics, *Proceedings of the 2nd International Conference on Human Language Technology Research (HLT)*, pp. 138–145 (2002).

[6] Fujita, M., Rzepka, R. and Araki, K.: Evaluation of Utterances Based on Causal Knowledge Retrieved from Blogs, *Proceedings of the 14th IASTED International Conference Artificial Intelligence and Soft Computing (ASC)*, pp. 294–299 (2011).

[7] Gabriel, F., Pineau, J., Larchevêque, J.-M. and Tremblay, R.: Bootstrapping Dialog Systems with Word Embeddings (2014).

[8] Jansen, P., Surdeanu, M. and Clark, P.: Discourse Complements Lexical Semantics for Non-factoid Answer Reranking, *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 977–986 (2014).

[9] Kawahara, D. and Kurohashi, S.: A Fully-Lexicalized Probabilistic Model for Japanese Syntactic and Case Structure Analysis, *Proceedings of Human Language Technology Conference - North American Chapter of the Association for Computational Linguistics Annual Meeting (HLT-NAACL)*, pp. 176–183 (2006).

[10] Kingma, D. P. and Ba, J.: Adam: A Method for Stochastic Optimization, *Proceedings of the 3rd International Conference on Learning Representations (ICLR)* (2015).

[11] Kudo, T. and Richardson, J.: SentencePiece: A Simple and Language Independent Subword Tokenizer and Detokenizer for Neural Text Processing, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (2018).

[12] Li, J., Galley, M., Brockett, C., Gao, J. and Dolan, B.: A Diversity-Promoting Objective Function for Neural Conversation Models, *Proceedings of the 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pp. 110–119 (2016).

[13] Liu, C.-W., Lowe, R., Serban, I. V., Noseworthy, M., Charlin, L. and Pineau, J.: How NOT to Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (2016).

[14] Luong, M.-T., Pham, H. and Manning, C. D.: Effective Approaches to Attention-Based Neural Machine Translation, *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (2015).

[15] Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, Ł., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G., Hughes, M. and Dean, J.: Google's Neural Machine Translation System: Bridging the Gap Between Human and Machine Translation, *arXiv:1609.08144* (2016).

[16] Mikolov, T., Chen, K., Corrado, G. and Deany, J.: Efficient Estimation of Word Representations in Vector Space, *Proceedings of the 1st International Conference on Learning Representations (ICLR)* (2013).

[17] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. and Dean, J.: Distributed Representations of Words and Phrases and Their Compositionality, *Proceedings of the 26th International Conference on Neural Information Processing Systems (NIPS)*, Vol. 2, pp. 3111–3119 (2013).

[18] Mikolov, T., Yih, W.-t. and Zweig, G.: Linguistic Regularities in Continuous Space Word Representations, *Proceedings of the 12th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pp. 746–751 (2013).

[19] Nakamura, R., Sudoh, K., Yoshino, K. and Nakamura, S.: Another Diversity-Promoting Objective Function for Neural Dialogue Generation, *Proceedings of the 33rd Association for the Advancement of Artificial Intelligence Conference on Artificial Intelligence, Workshop on Reasoning and Learning for Human-Machine Dialogues (DEEP-DIAL 2019) (AAAI)* (2019).

[20] Newman, D., Lau, J. H., Grieser, K. and Baldwin, T.: Automatic Evaluation of Topic Coherence, *Proceedings of the 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pp. 100–108 (2010).

[21] Oh, J.-H., Torisawa, K., Hashimoto, C., Iida, R., Tanaka, M. and Kloetzer, J.: A Semi-supervised Learning Approach to Why-Question Answering, *Proceedings of the 30th Association for the Advancement of Artificial Intelligence Conference on Artificial Intelligence (AAAI)*, pp. 3022–3029 (2016).

[22] Oh, J.-H., Torisawa, K., Hashimoto, C., Sano, M., Saeger, S. D. and Ohtake, K.: Why-Question Answering Using Intra- and Inter-Sentential Causal Relations, *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 1733–1743 (2013).

[23] Oh, J.-H., Torisawa, K., Kruengkrai, C., Iida, R. and Kloetzer, J.: Multi-Column Convolutional Neural Networks with Causality-Attention for Why-Question Answering, *Proceedings of the 10th Association for Computing Machinery International Conference on Web Search and Data Mining (WSDM)*, pp. 415–424 (2017).

[24] Ohmura, J. and Eskenazi, M.: Context-Aware Dialog Re-ranking for Task-Oriented Dialog Systems, *Proceedings of IEEE Spoken Language Technology Workshop (SLT)* (2018).

[25] Papineni, K., Roukos, S., Ward, T. and Zhu, W.-J.: BLEU: a Method for Automatic Evaluation of Machine Translation, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 311–318 (2002).

[26] Ram, A., Prasad, R., Khatri, C., Venkatesh, A., Gabriel, R., Liu, Q., Nunn, J., Hedayatnia, B., Cheng, M., Nagar, A., King, E., Bland, K., Wartick, A., Pan, Y., Song, H., Jayadevan, S., Hwang, G. and Pettigrue, A.: Conversational AI: The Science Behind the Alexa Prize,

arXiv:1801.03604 (2018).

[27] Sasano, R. and Kurohashi, S.: A Discriminative Approach to Japanese Zero Anaphora Resolution with Large-Scale Lexicalized Case Frames, *Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP)*, pp. 758–766 (2011).

[28] Serban, I. V., Sordoni, A., Bengio, Y., Courville, A. and Pineau, J.: Building End-To-End Dialogue Systems Using Generative Hierarchical Neural Network Models, *Proceedings of the 30th Association for the Advancement of Artificial Intelligence Conference on Artificial Intelligence (AAAI)* (2016).

[29] Shibata, T., Kohama, S. and Kurohashi, S.: A Large Scale Database of Strongly-Related Events in Japanese, *Proceedings of the 9th International Conference on Language Resources and Evalu ation (LREC)* (2014).

[30] Shibata, T. and Kurohashi, S.: Acquiring Strongly-Related Events Using Predicate-Argument Co-occurring Statist ics and Case Frames, *Proceedings of the 5th International Joint Conference on Natural Language Proce ssing (IJCNLP)*, pp. 1028–1036 (2011).

[31] Sordoni, A., Bengio, Y., Vahabi, H., Lioma, C., Simonsen, J. G. and Nie, J.-Y.: A Hierarchical Recurrent Encoder-Decoder For Generative Context-Aware Query Suggestion, *Proceedings of the 24th Association for Computing Machinery International Conference on Information Knowledge and Management (ACM)* (2015).

[32] Vinyals, O. and Le, Q. V.: A Neural Conversational Model, *Proceedings of the 32nd International Conference on Machine Learning, Deep Learning Workshop (ICML)* (2015).

[33] Weber, N., Balasubramanian, N. and Chambers, N.: Event Representations with Tensor-Based Compositions, *Proceedings of the 32nd Association for the Advancement of Artificial Intelligence Conference on Artificial Intelligence (AAAI)* (2018).