連続する事態の一貫性に基づく雑談対話応答のリランキングにおける事例分析

田中 翔平　吉野 幸一郎　須藤 克仁　中村 哲
奈良先端科学技術大学院大学
{tanaka.shohei.tj7, koichiro, sudoh, s-nakamura}@is.naist.jp

1　はじめに

Neural Conversational Model (NCM) [9] を始めとする，ニューラルネットワークで対話のクエリ-応答ペアを学習する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対話の文脈や論理を考慮せず，どのような場合にでも当てはまる単純な応答を生成してしまい，その結果として対話継続性が低下する，対話破綻を起こすといった問題が知られている（dull response 問題）．この問題に対し我々は，文脈や論理を考慮した応答を，対話モデルの生成する応答候補からリランキングにより選択する手法を提案した [4]．

本研究では「ストレスが溜まる」と「発散する」など，関連すると認められる事態ペアが対話履歴と応答候補の間に存在する場合，対話中の事態の一貫性が高いと考える．この事態間関係の一つとして，因果関係がある．因果関係とは 2 つの事態間に原因と結果の関係が成立すること [8, 7] と定義され，この定義に従い，「ストレスが溜まる」が原因,「発散する」が結果，のように認定する．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 [5]．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，文脈に沿った応答を生成できることが示されている [2, 3]．そこで，こうした因果関係に基づくリランキング手法について提案した．

また一貫性推定に関する研究として，Coherence Model [11] がある．このモデルは文書中に出現する単語の品詞情報や文の分散表現をもとに，入力された文書の一貫性を推定する．対話においてもこの一貫性推定は有効であることが知られている [1]．そこで，この Coherence Model に基づくリランキング手法についても提案した．

提案手法は，NCM によって生成された N-best 応答候補より，一貫した，対話継続性の高い応答を選択するものである．この手法では，対話履歴に対し一貫した応答を選択するために，事態の一貫性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．事態の一貫性の考慮を行うため，大規模コーパスから統計的に獲得された因果関係ペア [8, 7] を用いる．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 (RFTM) [10] を用いた事態の分散表現によって汎化を行った．また上述の事態の一貫性のみを考慮したリランキングでは応答全体の一貫性が低下する可能性があるため，異なるリランキング手法として Coherence Model [11] に基づく応答候補の一貫性推定を提案した．自動評価及び人手評価の結果，因果関係ペアを用いたリランキングにより応答の一貫性，対話継続性が最も向上することが示された [4] ものの，これらの手法が具体的にどのような場面で有効かについて検討する必要がある．そこで本稿では実際のリランキング結果に対して事例分析を行うことで，傾向調査を行った．

2　事態の一貫性に基づく応答のリランキング

本実験で我々が提案したリランキングモデル [4] を使用する．図 1 に手法の概要を示す．この手法は大きく分けて 3 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）．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 [6] を用いる．最後に応答候補を事態の一貫性に基づきリランキングする（図 1 ③）．このリランキングのために，2 つの異なる手法を提案した．

1 つ目の手法は事態の一貫性に関する外部知識として，統計的に獲得された因果関係ペア [8, 7] を用いるリランキングである．このリランキングでは，抽出した事態及び因果関係ペアとの表層マッチングにより対話中の因果関係を抽出し，抽出された因果関係に基づき応答候補をリランキングする．この手法を “Reranking (Pairs)” と呼ぶ．大規模テキストから抽出し

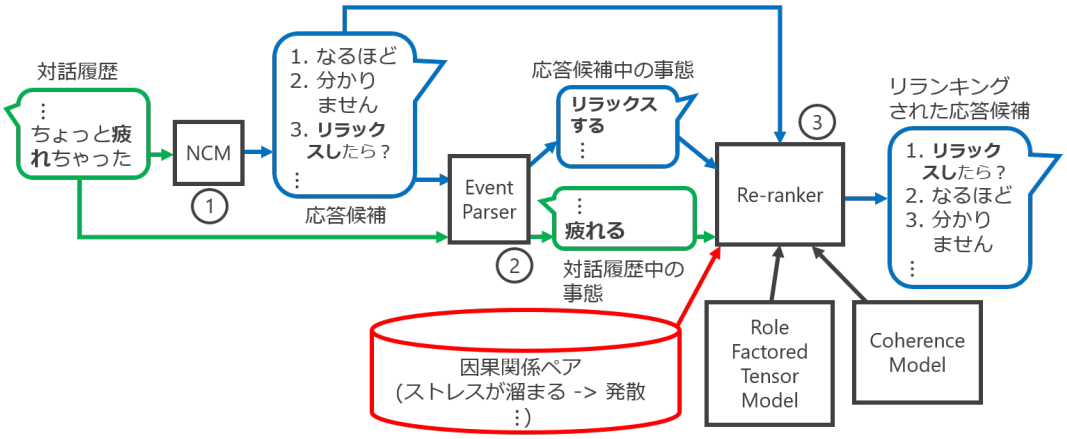

図 1: Neural Conversational Model+ リランキング; 「疲れる」と「リラックスする」が関連した事態であるという知識に基づき応答を選択.

た大規模因果関係ペアデータセットであっても，あらゆる因果関係ペアを網羅できるわけではないため，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 RFTM を用いて分散表現に変換し，ベクトル空間中で因果関係知識と対話中に出現した因果関係との類似度に基づくマッチングを行うことで，表層の一致しない因果関係に対するマッチングを実現する．この手法を “Re-ranking (RFTM)” と呼ぶ.

2 つ目の手法は事態間の関係のみでなく，Coherence Model によって対話全体の一貫性も評価するリランキングである．因果関係の定義の難しさ [8, 7] や，事態分散表現が事態を過汎化する可能性があることから，“Re-ranking (RFTM)” で用いられる因果関係は必ずしも正確ではない．また因果関係ペアを用いたリランキングは応答候補中に出現する事態ペアの一貫性のみに着目しているため，選択された応答候補全体が持つ意味が対話履歴に対して一貫していないことも考えられる．そこで Coherence Model を用いることで事態ペアのみでなく，応答全体の一貫性も評価するリランキングを実現する．この手法を “Re-ranking (Coherence)” と呼ぶ.

3 事例分析

本稿では提案手法による応答のリランキングを詳細に分析するため，述語項構造解析結果，リランキング結果の分類を行った．具体的には，まず本研究で用いた KNP による述語項構造解析結果の分析を行い，次にリランキング結果個別の分析を行う．評価対象は人手評価実験 [4] で用いた各 100 対話，合計 300 対話とした.

表 1: 述語項構造解析結果の分類

	Correct	Wrong	Sum
Each	424	176	600
Both	170	130	300

3.1 述語項構造解析結果の分類

リランキングに用いられた事態の述語項構造解析が適切に行われている割合を調査した．分類結果を表 1 に示す．ここで横軸の “Correct” はリランキングに用いられた事態に述語項構造解析の誤りがなかった場合であり，“Wrong” は何らかの誤りが含まれていた場合を指す．これは例えば,「おはようさぎ」という発話文から「詐欺」という誤った述語（判定詞）を抽出した場合や,「栄行こうか迷う」という発話文に対し「栄が行く」のように格解析を誤った場合などがある．“Each” は事態ペアに含まれる 2 つの事態について別々に正誤を判定した場合であり，“Both” は 2 つの事態をまとめて正誤を判定した場合である．つまり，事態ペアに含まれる 2 つの事態のいずれも述語項構造解析が適切に行われていた場合のみ，“Both” が “Correct” となる.

述語項構造解析によって事態が完全に解析されている割合は “Each” で 70% 前後，“Both” で 60% 前後であり，十分高いとは言えないが，特に後者は複数の述語項構造関係の抽出結果に対する評価という点に留意する必要がある．また，提案した事態の埋め込み表

表 2: リランキング結果の分類 (Re-ranking (Pairs))

Re-ranking / Events	Good	Bad (Pairs)	Sum
Good	20	6	26
Bad	4	8	12
Both Good	11	6	17
Both Bad	26	19	45
Sum	61	39	100

表 3: リランキング結果の分類 (Re-ranking (RFTM))

Re-ranking / Events	Good	Bad (Pairs)	Bad (過汎化)	Sum
Good	1	5	12	18
Bad	0	0	5	5
Both Good	2	2	17	21
Both Bad	2	5	49	56
Sum	5	12	83	100

表 4: リランキング結果の分類 (Re-ranking (Coherence))

Re-ranking / Events	Good	Bad (Sequence)	Sum
Good	18	8	26
Bad	3	5	8
Both Good	17	22	39
Both Bad	14	13	27
Sum	52	48	100

現が，格要素の解析誤りなどの問題を汎化している可能性がある．

3.2 リランキング結果の分類

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事態ペアの妥当性を分類，分析した．分類結果を表 2-4 に示す．表 2 は “Re-ranking (Pairs)”，表 3 は “Re-ranking (RFTM)”，表 4 は “Re-ranking (Coherence)” に関する分類である．ここで各行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” は悪化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．各列はリランキングに用いられた事態ペアの一貫性を表し，“Good” は一貫した事態ペアが用いられたことを示す．“Bad (Pairs)” は事態の汎化を行わず項構造解析結果と因果関係ペアを利用した場合に，当該対話コンテキストで用いることが適当ではないと考えられる因果関係が用いられたことを意味する．また “Bad (過汎化)” は，RFTM を用いて事態を分散表現化した場合に，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．“Bad (Sequence)” は，Coherence Model を利用した時に連続しているが一貫性が低い事態ペアが用いられたことを意味する．以上は著者らが各サンプルを確認して人手で分類した．

“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 “Re-ranking (Pairs)”，“Re-ranking (RFTM)” の場合 40% 前後，“Re-ranking (Coherence)” の場合 60% 前後である．一貫した事態ペアが用いられた割合は，“Re-ranking (Pairs)” の場合 60% 前後，“Re-ranking (RFTM)” の場合 5% 前後，“Re-ranking (Coherence)” の場合 50% 前後である．一貫した事態ペアにより，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事態ペアを示す．“→” で結ばれた事態ペアは因果関係を表す．

対話 1:

> **ユーザ：** 新年早々**体調崩す**とは．辛い笑明日，明後日までに完全に治さな．
> **システム (1-best)：** お大事に
> **システム (Re-ranked (Pairs))：** 大丈夫ですか**無理**しないでくださいね
> [無理をする → 体調を崩す]

対話 2:

> **ユーザ：** おはようございます☆
> **システム：** おはようございますヒラメ釣り行ってましたが撃沈でした TT
> **ユーザ：** おはようございます☆**冷え**ますね...
> **システム (1-best)：** おはようございます
> **システム (Re-ranked (Pairs))：** おはよーございます朝晩**涼しく**なりました
> [涼しくなる → 冷える]

対話 3:

> **ユーザ：** **仕事**おおお行かなきゃ
> **システム (1-best)：** いってらっしゃい
> **システム (Re-ranked (RFTM))：** お**仕事頑張って**ください
> [仕事 → 仕事を頑張って下さる]

対話 4:

ユーザ： みんな美味しいセブンドーナツ食べよう
システム (1-best)： おいしいよ
システム (Re-ranked (Coherence)): チョコチップ食べたい
[ドーナツが食べる, チップが食べたい]

上記の例では，事態ペア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 2-4 からわかるように，そうではない例も多数存在する．次に示す対話 5 のように一貫していない事態ペアが用いられている場合や，対話 6 のように，一貫した事態ペア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 5:

システム： 今日はたのしかったね
ユーザ： 電話も楽しかったよ
システム (1-best)： ほんとね
システム (Re-ranked (RFTM))： 今度行こうよ
[電話が楽しい → 行く]

対話 6:

システム： はよ風邪なおそなおったら行こう
ユーザ： かれこれ 3 週間くらい風邪引いてる
システム (1-best)： お大事に
システム (Re-ranked (Coherence))： 風邪治ってきたから
[風邪を引く, 風邪が治る]

この結果から今後の課題として，事態分散表現や Coherence Model を改善することで，事態ペアや応答の一貫性をより正確に考慮する必要があると考えられる．

4　おわりに

本論文では著者らが以前に提案した，ニューラル雑談対話モデル (NCM) により生成された N-best 応答を連続する事態の一貫性に基づきリランキングする手法の，事例分析を行った．事例分析の結果，一貫していない事態ペアがリランキングに用いられた場合や，リランキング後の応答が不自然である場合が多く存在することがわかった．今後は一貫した対話中の事態を生成した上で応答生成を行う生成的アプローチについて検討していく．

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禎夫教授，柴田知秀博士に感謝いたします．

本研究は JST さきがけ (JPMJPR165B) の支援を受けた．

参考文献

[1] Alessandra Cervone, Evgeny Stepanov, and Giuseppe Riccardi. Coherence Models for Dialogue. In *Proceedings of INTERSPEECH 2018 (INTERSPEECH)*, 2018.

[2] Motoyasu Fujita, Rafal Rzepka, and Kenji Araki. Evaluation of Utterances Based on Causal Knowledge Retrieved from Blogs. In *Proceedings of the 14th IASTED International Conference Artificial Intelligence and Soft Computing (ASC)*, pp. 294–299, 2011.

[3] 佐藤祥多, 乾健太郎. 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 (ANLP), pp. 1219–1222, 2018.

[4] 田中翔平, 吉野幸一郎, 須藤克仁, 中村哲. 事態の一貫性推定に基づく雑談対話応答選択モデル. 人工知能学会 第 87 回言語・音声理解と対話処理研究会 (SIG-SLUD), 2019.

[5] Jong-Hoon Oh, Kentaro Torisawa, Chikara Hashimoto, Motoki Sano, Stijn De Saeger, and Kiyonori Ohtake. Why-Question Answering Using Intra- and Inter-Sentential Causal Relations. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 1733–1743, 2013.

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

[7] Tomohide Shibata, Shotaro Kohama, and Sadao Kurohashi. A Large Scale Database of Strongly-Related Events in Japanese. In *Proceedings of the 9th International Conference on Language Resources and Evalu ation (LREC)*, 2014.

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

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

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

[11] Peng Xu, Hamidreza Saghir, Jin Sung Kang, Teng Long, Avishek Joey Bose, Yanshuai Cao, and Jackie Chi Kit Cheung. A Cross-Domain Transferable Neural Coherence Model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 678–687, 2019.