



因果関係を用いた雑談対話応答 におけるランキングの評価

2019/03/15

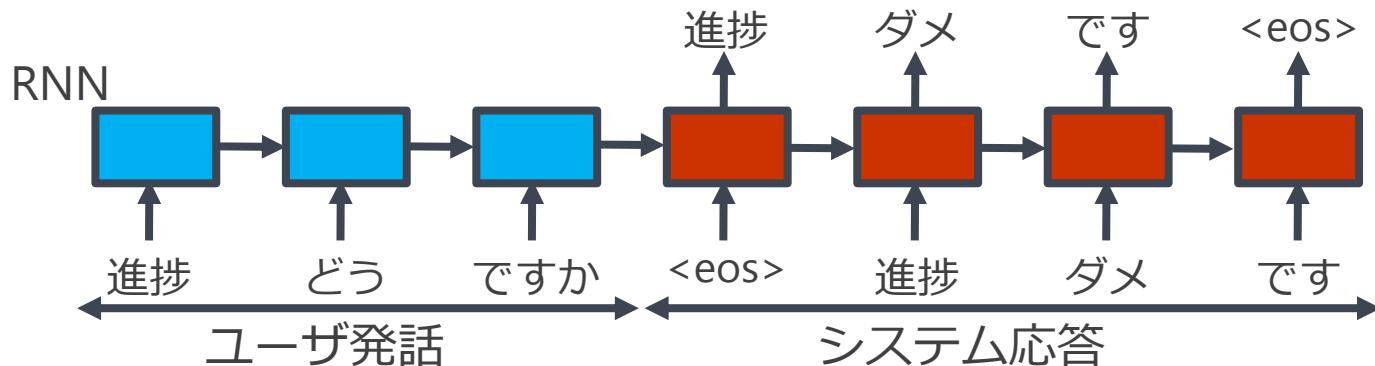
田中翔平¹ 吉野幸一郎^{1,2} 須藤克仁¹ 中村哲¹

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

はじめに

ニューラル雑談対話モデルの現状

Neural Conversational Model (NCM) [Vinyals et al., 2015]
などのニューラルネットワークに基づいた雑談対話モデル
の研究が主流



ルールベースや用例ベースよりも柔軟に応答を生成可能

↓
単純でつまらない応答が生成される傾向

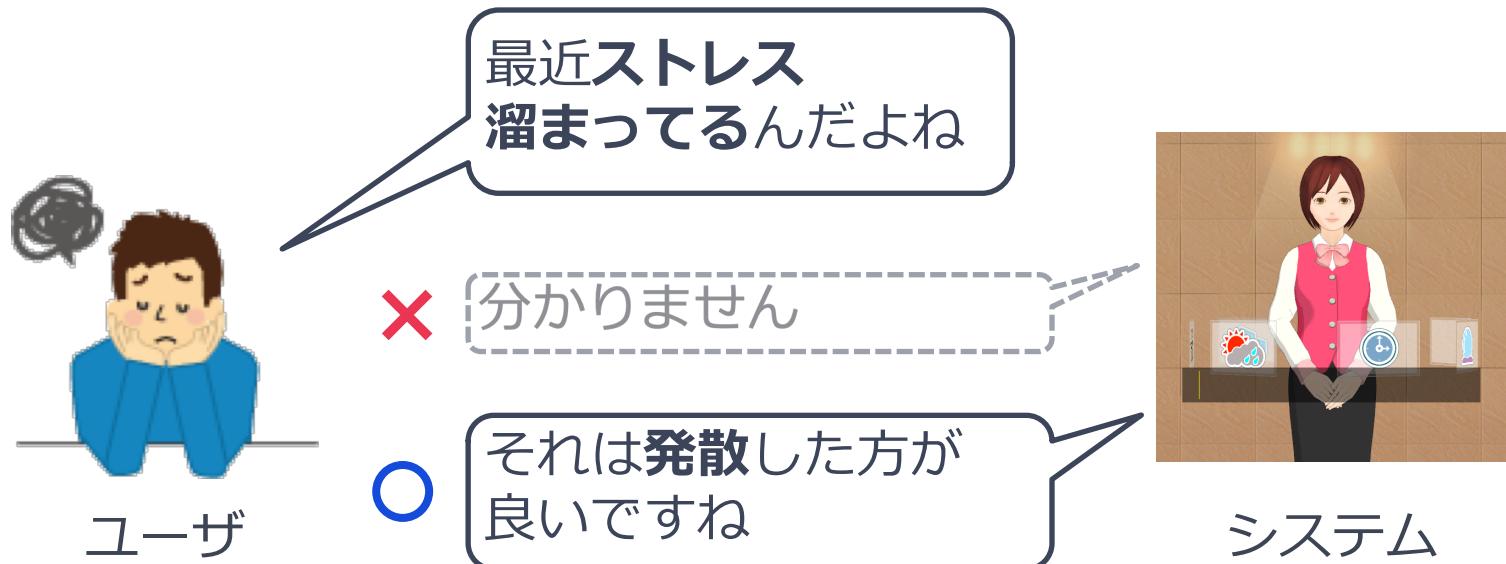
なるほど

分かりません

文脈や論理を考慮した、多様な応答の生成が困難

因果関係を用いた応答選択

因果関係を考慮して雑談対話モデルから生成された N -best 応答候補をリランキング



「ストレスが溜まる」 → 「発散」という因果関係
が成立している応答を選択

因果関係

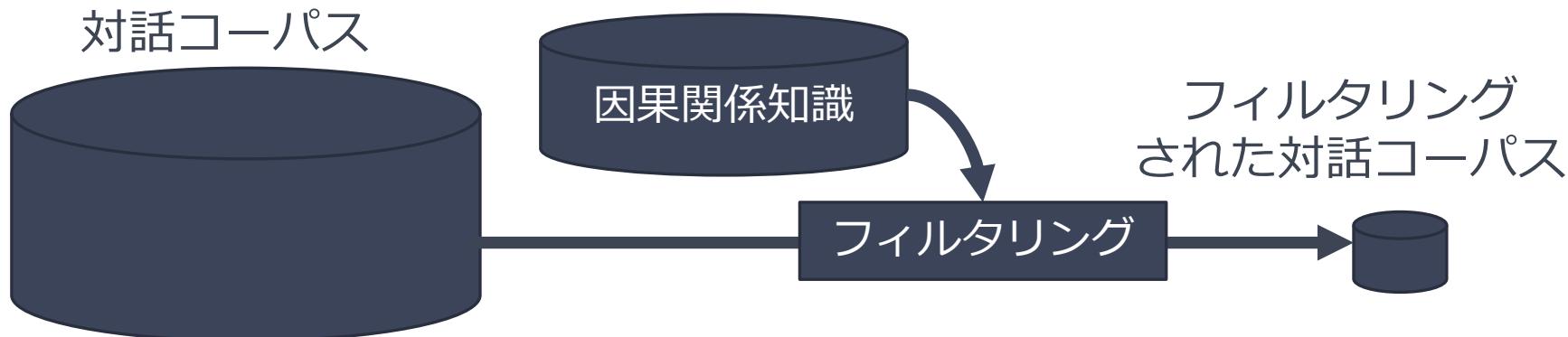
2つの事態間に原因と結果の関係が成立する関係

e.g. ストレスが溜まる（原因） → 発散（結果）

- 雜談対話における発話対に關しても重要
[徳久ら, 2007]
- **間接応答や問い合わせ返し**において先行発話との因果関係が多く成立
- 因果関係による**対話を継続する働き**
(**対話継続性**) 向上の期待

先行研究

因果関係に基づくデータフィルタリング [佐藤ら, 2018]



文脈を考慮した、対話継続性の高い応答を生成

学習データ量が削減されモデルの学習が困難となる可能性



限られたデータにも適用可能な手法の必要性

因果関係を用いた 応答選択

① N-best 応答候補の生成

対話履歴

…
最近ストレス溜
まってるんだよね

① NCM



1. 分かりません
 2. なるほど
 3. それは**発散**した
方が良いですね
- …

応答候補

入力：対話履歴（ユーザ発話）
出力：N-best 応答候補

対話履歴として複数発話を
考慮することも可能

②事態（述語項構造）の抽出

対話履歴

⋮
最近ストレス溜
まってるんだよね

KNP [Kawahara et al., 2006] を用いて
事態（述語項構造）を抽出

1. 分かりません
 2. なるほど
 3. それは**発散**した
方が良いですね
- ⋮

応答候補

⋮
ストレスが
溜まる

対話履歴中の
事態

発散

⋮
応答候補中の
事態

KNP

②

③応答候補のリランキング

因果関係を認定可能な
応答候補を高い順位に
リランキング

複数発話を考慮する
ことも可能

- 1. 分かりません
- 2. なるほど
- 3. それは**発散**した
方が良いですね
- :

応答候補

リランキン
グされた応答候補

因果関係辞書
(ストレスが溜まる
→ 発散
:)

1. それは**発散**した
方が良いですね
2. 分かりません
3. なるほど
- :

リランキン
グ機構

③

ストレスが
溜まる

対話履歴中の
事態

発散
:

応答候補中の
事態

NCM + リランキング機構

対話履歴

⋮
最近ストレス溜
まってるんだよね

① NCM

対話
コーパス

1. 分かりません
2. なるほど
3. それは**発散**した
方が良いですね
⋮

応答候補

リランキン
グされた応答候補

因果関係辞書
(ストレスが溜まる
→ 発散
⋮)

1. それは**発散**した
方が良いですね
2. 分かりません
3. なるほど
⋮

③ リランキン
グ機構

⋮
ストレスが
溜まる

対話履歴中の
事態

発散
⋮

応答候補中の
事態

KNP

②

リランキンギングスコアの計算

NCM から与えられる応答のスコア (対数尤度 < 0)
絶対値が小さいほど高い順位

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

↓
 l
↑
因果関係の重み ($>= 0$)

2つの事態間の相互情報量 ($10 <= lift <= 10,000$)
事態間の因果関係としての結びつきの強さ
高い因果関係スコアを持つ候補を強調

因果関係辞書

ランキングのために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辞書 [柴田ら, 2011] を利用

415,926 件の因果関係知識

各事態は述語項構造を用いて表現

述語：必ず内含 項：必ずしも内含せず

原因 → 結果
述語：溜まる、ガ格：ストレス → 述語：発散 lift: 536.95

$$I(cause; effect) = P(cause, effect) \log \frac{P(cause, effect)}{P(cause)P(effect)}$$

実験

対話履歴の考慮の比較

提案手法は NCM の種類に依らず適用可能

2つのモデルで応答生成時の対話履歴の考慮を比較

- Encoder-Decoder with Attention Model (EncDec) [Luong et al., 2015]

長期の履歴の考慮困難 → 直前1発話のみ考慮

- Hierarchical Recurrent Encoder-Decoder (HRED) [Serban et al., 2016]

長期の履歴の考慮容易 → 直前5発話を考慮

応答の多様性が履歴により制限されランキングに不向きな可能性

ランキング時はどちらの応答も直前1, 5発話を考慮

→ 生成時に考慮できなかつた対話履歴も考慮可能

データセット

Twitter データセットを利用

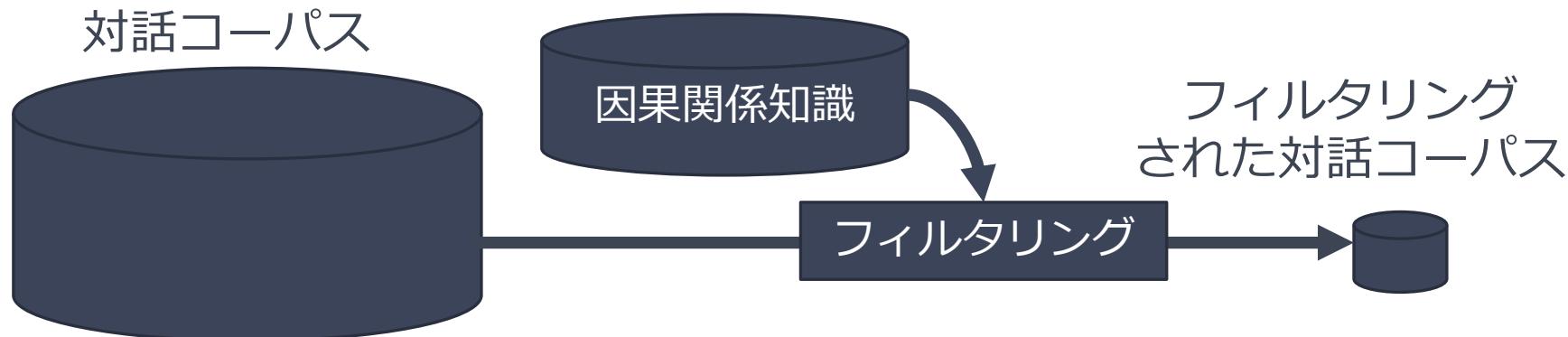
	# dials	$\overline{\text{dial}}$	$\overline{\text{uttr}}$
Train	2,509,836	21.95	22.09
Train (Filtered)	62,700 (2.50%)	6.37	36.92
Valid	63,308	25.55	21.89
Test	58,970	20.29	21.95

因果関係に基づくフィルタリング [佐藤ら, 2018] によって学習データは2.50%に減少

上記のデータで学習させた EncDec (Filtered) とも比較

Filtered

因果関係に基づくデータフィルタリング [佐藤ら, 2018]



文脈を考慮した、対話継続性の高い応答を生成

学習データ量が削減されモデルの学習が困難となる可能性



限られたデータにも適用可能な手法の必要性

評価方法

- ビームサーチ結果の多様性 (N -best 応答内の dist)

$$\overline{\text{dist}} = \frac{1}{\text{対話数}} \sum \text{各 } N\text{-best 応答内の dist}$$

dist は応答候補中の N -gram の異なり数の割合

高いほどリランキングの効果が大きい

- リランキングされた応答の割合

リランキングがどの程度適用可能であるかを計測

評価方法 (cont.)

- BLEU

実際の応答との N -gram の一致率

実際の応答は文脈を考慮

システム応答が文脈を考慮しているかをある程度表現

- dist

システム応答が多様であることを意味

- 平均応答長

応答が長いほどその応答は単調でない傾向

ビームサーチ結果の多様性

ビーム幅20の場合を比較 (続く評価でも同一)

	dist-1	dist-2
EncDec	0.38	0.50
HRED	0.35	0.46

EncDec が高い多様性

HRED は多様性が低下

→ 多様性が対話履歴により制限

リランクされた応答の割合

	Reranked (%)
Reference (history 1)	1,566 (2.66)
Reference (history 5)	2,681 (4.55)
EncDec (history 1)	3,231 (5.60)
EncDec (history 5)	3,921 (6.79)
HRED (history 1)	5,401 (9.36)
HRED (history 5)	6,936 (12.02)

実際の応答と対話履歴
との間に因果関係を認定
可能な割合

HRED は生成時にも複数発話を考慮

→ 先行発話との因果関係が成立するものが増加

リランクされた応答の割合

	Reranked (%)
Reference (history 1)	1,566 (2.66)
Reference (history 5)	2,681 (4.55)
EncDec (history 1)	3,231 (5.60)
EncDec (history 5)	3,921 (6.79)
HRED (history 1)	5,401 (9.36)
HRED (history 5)	6,936 (12.02)

実際の応答と対話履歴
との間に因果関係を認定
可能な割合

HRED (history 5) が最も多くの応答をリランク

リランクされた応答の割合はやや低い (12.02%)

BLEU

	BLEU	
EncDec (w/o reranking)	1.46	
EncDec (history 1)	1.67	
EncDec (history 5)	1.69	リランギングによって上昇
HRED (w/o reranking)	0.84	
HRED (history 1)	0.98	
HRED (history 5)	1.05	
Filtered	1.14	フィルタリングによって低下

リランギングにより文脈を考慮した応答を選択

フィルタリングでは文脈を考慮した応答の生成が困難

dist

	dist-1	dist-2
Reference	0.18	0.70
EncDec (w/o reranking)	0.14	0.33
EncDec (history 1)	0.15	0.35
EncDec (history 5)	0.15	0.36
HRED (w/o reranking)	0.06	0.09
HRED (history 1)	0.10	0.15
HRED (history 5)	0.13	0.20
Filtered	0.04	0.16

← 実際の応答から算出

リランキングによって
上昇

フィルタリングによって
低下

リランキングにより多様な応答を選択

フィルタリングでは応答の多様性が減少

平均応答長

	response	
Reference	21.90	
EncDec (w/o reranking)	15.78	
EncDec (history 1)	15.59	
EncDec (history 5)	15.33	リランキング前後で ほぼ同様
HRED (w/o reranking)	16.69	
HRED (history 1)	17.38	
HRED (history 5)	17.69	フィルタリングによって 増加
Filtered	17.98	

Filtered の dist は低い → 多様性は低い

実験結果のまとめ

- EncDec は HRED よりもビームサーチにおいて
高い多様性
- HRED は EncDec よりも生成した応答が
リランギングされる割合が**多い**
- BLEU, dist がリランギングにより**上昇**
- 平均応答長はリランギング前後で**ほぼ同様**
- 大規模な対話データを用意できない場合
フィルタリングでは対話モデルの学習が**困難**

リランキング結果の例 (EncDec, history 5)

Utter	Response (before reranking)	Response (after reranking)	Event causality
リツイートありがとうございます フォローしますね	いえいえありがとうございます ございます	フォローありがとうございます 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

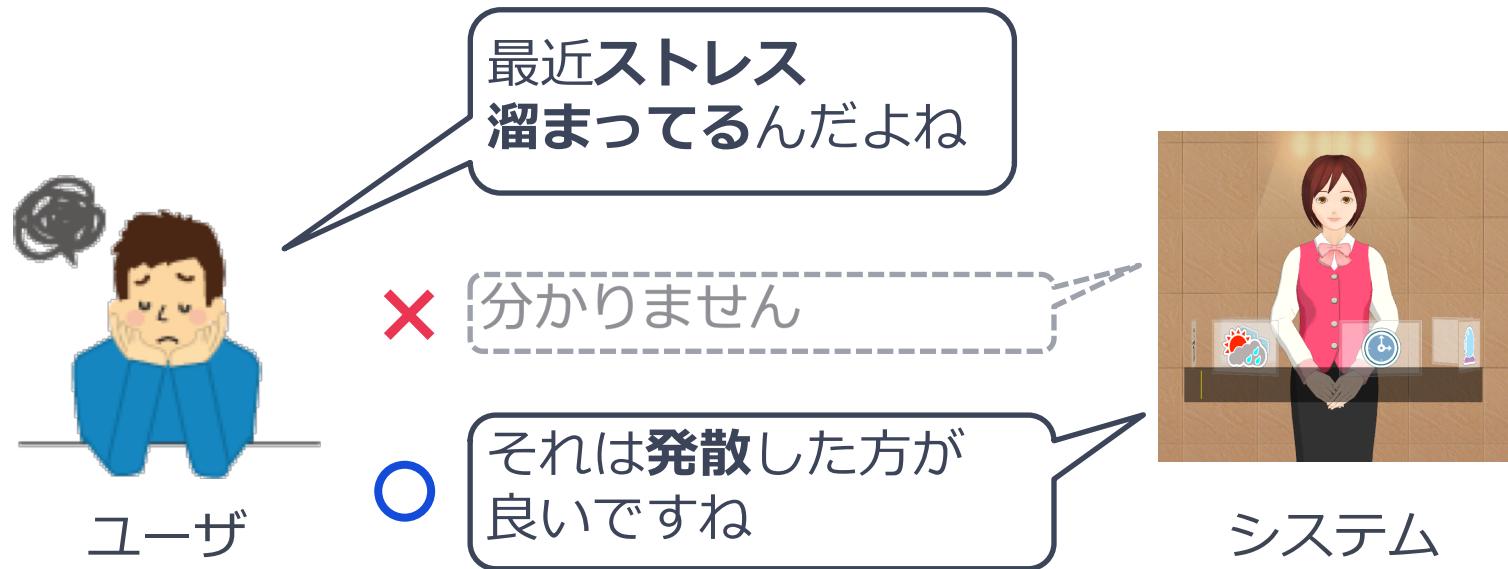
かっこ内の数字はリランキング前の順位

文脈や論理を考慮した、多様な応答を選択

おわりに

まとめ

- 因果関係を用いて文脈や論理を考慮した多様な雑談応答を選択する手法を提案



- 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を確認

今後の課題

- 因果関係のカバレージの向上

リランキング可能な応答の割合はやや低い(12.02%)

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向上

- 大規模な被験者実験

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