ニューラル対話モデルにおける品詞に基づく低頻度語処理

原口 洋一 [*1]　村田 真樹 [*2]
[*1] 鳥取大学 工学部 知能情報工学科
[*2] 鳥取大学大学院 工学研究科 情報エレクトロニクス専攻
[*1,*2]{s142044,murata}@ike.tottori-u.ac.jp

1 はじめに

近年，ニューラルネットワークを用いる手法が自然言語処理の多くのタスクで成果を上げている．その中に対話のモデルをニューラルネットワークにより構築したニューラル対話モデルがある [1]．ニューラル対話モデルの学習に用いるデータをコーパスと呼ぶ．

コーパス中での出現頻度が低い要素を低頻度語という．要素 (語彙) が増えるとニューラルネットワークの学習速度が落ちるため，コーパス中で重要でない低頻度語を同一の記号 (ヌルトークン unk_0) へ置き換えることで学習速度の低下を抑える．しかし，ヌルトークン unk_0 となった低頻度語は全て同一の記号となるため，コーパスにおいて本来違う単語が同一単語として扱われる．低頻度語処理の手法として，Copyable Model[2] と低頻度語の高頻度語への置き換え [3] がある．

本研究では，CopyableModel 処理に加え，Copyable-Model で処理できずに残ったヌルトークン unk_0 を少数のグループへ分割することにより，応答の精度を向上させる．分割するためのラベルとして品詞情報を用いる．

また，過去の発話を用いた学習も行う．非タスク指向型の対話は，状況によって許容される応答が変化する．過去の発話を含まないデータの場合「うん」等の，応答が必ずしも必要でない発話に対する応答の評価を行いにくい．そこで，発話の直前の対話 (過去の発話) を学習データとして追加 [4] することで，評価の際に評価者が文脈を知ることができるようにする．

2 先行研究

2.1 Copyable Model

Copyable Model は NMT のための低頻度語処理モデルである [2]．Copyable Model は unk トークンを一種類だけではなく，複数種類のトークンを使用する．これにより未知語となるヌルトークンに，入力と出力で同じ単語であったという情報を残す．

unk_1，unk_2，unk_3 の順でソース文の低頻度語にナンバーを振る．ターゲット文の未知語アノテーションは，ソース文の未知語化した単語と同じ単語があれば，同じトークンに割り当てる．ターゲット文中の低頻度語に対し，原文に同じ単語が無い場合，対応を持たないヌルトークン unk_0 を使用する．入力 (発話) と出力 (応答) の単語に対応関係のない対話モデルではヌルトークン unk_0 が多く出現する．

表 1 は CopyableModel の変換例である．

2.2 低頻度後の高頻度語への置き換え

低頻度語を同義な高頻度語へ置き換える [3]．ターゲット文の，トレーニングデータにおいての低頻度語を高頻度語に言い換えることにより，未知語への変換を削減する．単語の置き換え辞書を作成する必要がある．

3 本研究の手法

3.1 ヌルトークンへの品詞情報の付加

本実験では低頻度語処理の際，Copyable Model の未知語処理を行った後，品詞情報を用いてヌルトークン unk_0 を少数のグループへ分割する．

これにより，低頻度語処理後のコーパスに残すことができる情報が増加する．品詞情報の付加は Mecab で容易に行うことができるため，細分化の方法として採用した．

表 2 は本研究の手法での低頻度語処理例である．

3.2 過去の発話を含めた学習

本実験では連続した対話情報を発話と応答に分割し学習データとしている．通常は発話，応答とも 1 人 (1 発話) の発言で学習するが，入力 (発話) データに対しその前に行われた別人物の発話も付加する．これにより，文脈が見やすくなるため通常のデータより評価が容易になる．また，文脈を学習できる可能性がある．表 3 が通常の対話データで，表 4 が過去の発話を含めた対話データである．

過去の発話を含む対話データにおける eos とは，発話の区切り (発話者の切り替え) を示す．

表 1: Copyable Model の変換例

低頻度語	ソース文	ターゲット文
unk_0: 5 月	unk_0 が 待ち遠しい です	こいのぼり を 上げる ん です か ？
unk_0: ぶら下げ,unk_1: 蚊取り線香	unk_1 も unk_0 て	unk_1 は ほしい です ね
unk_1: 神様,unk_2: 試験	unk_1 の unk_2 って 何 よ	unk_1 の unk_2 が ある ん です か ？

表 2: 提案手法の変換例

低頻度語	ソース文	ターゲット文
unk_1:独自，$unk_{動詞}$:重ねる	unk_1 に $unk_{動詞}$ かも しれ ない	unk_1 に 心の哲学 を 研究 なさっ て いる の です ね
$unk_{名詞}$:A(人名)，unk_1: 市ヶ谷	$unk_{名詞}$ ね unk_1 な ん だ よ ね	unk_1 か ー
$unk_{名詞}$:2 月 11 日，unk_1: 建国	$unk_{名詞}$ は なんだろう unk_1 記念 か	unk_1 記念 か

表 3: 対話データ (過去の発話を含まない)

発話データ	応答データ
人口 多い です よ ね 東京 は 魅力 的 だ から	多い です よ ねえ
元気ですか ？	元気です
家事 は 苦手 です か	そうですね

表 4: 対話データ (過去の発話を含む)

発話データ	応答データ
こんにちは ー 海 へ 行き たい ね eos こんにちは ー 行き たい	朝 から 海 に 行き ます か ??
そうですね eos ありがとう ござい ます	どういたしまして
クーラー は 欲しい です ね eos ない の ？	うん

4 実験

4.1 実験環境

コーパスに出現する低頻度語を，複数の手法により処理し，その出力の評価を行う．RNN(LSTM) の文生成モデルである seq2seq[5] を使用する．

本実験では，過去の発話を含まないデータにおける訓練データの発話 (入力側) データ全てにおいて一度しか出てこない形態素を低頻度語と定義する．また，過去の発話は本来の発話 (入力データ) の直前 1 名分の発話とする．低頻度語数に差を出さないために，過去の発話を含むデータの低頻度語は過去の発話を含まないデータの低頻度語と同一にする．

ニューラルネットワークは Open-NMT を使用した [6]．バッチサイズは 64，epoch 数は 50 とした．

表 5 は各手法の低頻度語変換例である．

4.2 使用コーパス

実験で使用するコーパスは，Project Next NLP 対話タスクで収集されたコーパス (雑談対話コーパス)[7] と名大会話コーパス (日本語自然会話書き起こしコーパス)[8] である．コーパスの詳細を表 6 に示す．

また，表 7 はコーパスの用途別データ数である．開発データとは，Open-NMT のプレトレーニングで必要なデータである．

表 6: 使用コーパス

コーパス名称	制作
雑談対話コーパス	対話破綻検出チャレンジ
名大会話コーパス	日本語教育ネットワーク

表 7: コーパス用途別データ数

用途	データ数	用途	データ数
総数	69,452	訓練データ	67,369
開発データ	1,389	テストデータ	694

4.3 実験結果

表 8 は Copyable Model による低頻度語変換数であり，表 9 は Copyable Model の変換時に発生したヌルトークンに，提案手法により品詞情報を付加して行った分類の変換数である．

表 8: Copyable Model による低頻度語変換数

unk 番号	変換数
0(ヌルトークン)	8,701
1	105
2	3

表 9: 提案手法によるヌルトークン変換数

品詞	変換数	品詞	変換数
名詞	6,319	記号	13
動詞	1,508	助詞	10
副詞	390	助動詞	10
形容詞	389	連体詞	8
感動詞	34	接続詞	5
接頭詞	15		

表 5: 各手法の変換例

手法	ソース文 (発話)	ターゲット文 (応答)
原文	電圧 を 換える だけ で 普通 に 今 持っ て いる の が 使える の ?	電圧 が もし 向こう で も オッケー だっ たら
低頻度語を全てヌルトークンへ変換	unk_0 を unk_0 だけ で 普通 に 今 持っ て いる の が 使える の ?	電圧 が もし 向こう で も オッケー だっ たら
Copyable Model	unk_1 を unk_0 だけ で 普通 に 今 持っ て いる の が 使える の ?	unk_1 が もし 向こう で も オッケー だっ たら
提案手法	unk_1 を $unk_{動詞}$ だけ で 普通 に 今 持っ て いる の が 使える の ?	unk_1 が もし 向こう で も オッケー だっ たら

表 10 は各手法の，過去の発話を含まないデータにおける応答出力結果例である．表 11 は各手法の，過去の発話を含めた学習データにおける応答出力結果例である．過去の発話は本来の発話 (入力データ) の直前 1 名分の発話である．

表 10: 過去の発話を含まないデータにおける各手法の応答出力結果

発話：寝不足 の とき は 頭 の てっぺん を 軽く 押す と いい ん だって	
手法	入出力
Copyable Model	発話：寝不足 の とき は 頭 の unk_0 を 軽く unk_0 と いい ん だって 応答：[A] さん と ね [A] さん が 来 た の ?
提案手法	発話：寝不足 の とき は 頭 の $unk_{名詞}$ を 軽く $unk_{動詞}$ と いい ん だって 応答：そうそう

表 11: 過去の発話を含むデータにおける各手法の応答出力結果

発話：歯医者 の 日 な ん や eos 歯科衛生士 さん は 美人 揃い でし た	
手法	入出力
Copyable Model	発話：歯医者 の 日 な ん や eos unk_0 さん は 美人 unk_0 でし た 応答：そうそう そうそう
提案手法	発話：歯医者 の 日 な ん や eos $unk_{名詞}$ さん は 美人 $unk_{名詞}$ でし た 応答：あ そう な ん だ

以下に評価結果を示す．評価基準は下記の通りである．また，評価例を表 12 に示す．評価は各手法の出力をテストデータから抽出した 200 文で行った．今回の実験では，過去の発話を含むデータ，含まないデータの両方とも，入力より前の文脈 (データ) を見ずに評価し，必要であれば入力より前の文脈を出力に都合の良いように補完する．補完無く評価可能且つ評価○の条件を満たすものを○と評価する．

- 評価：○
  - 現状の発話応答で正しく完結している．
  - 意志表示が必要な発話に対し，自分の意志表示を行う相槌．
- 評価：△
  - 理解可能だが文 (文法) が一部破綻．
  - より前の対話があると仮定すれば (極端な文脈を補完すれば) 正しいと思わる応答．
  - 応答として適切でないが明らかにジョークと取れる応答．
  - 応答として適切だが会話の進展が無い．
  - 会話として適当なオウム返し．
  - 相手に発言を促す相槌．
  - 質問に対する聞き返し．
  - 返答として適切と思われるが一文で完結してない発話．
  - 意図が理解可能な範囲で，期待する内容からずれた返答．
- 評価：×
  - 理解可能だが文が完全に破綻．意味不明．
  - 会話として適当で無いオウム返し，相槌．

表 12: 評価例

評価理由	入出力
○：現状の発話応答で正しく完結している	発話：こんにちは 夏 と いえ ば スイカ だ ね 応答：スイカ は 大好き です ね
△：応答として適切だが会話の進展が無い	発話：奥 が 深い です ね 応答：はい
△：意図が理解可能な範囲で期待する内容からずれた返答	発話：スポーツ は なさい ます か ? 応答：スポーツ は 必要 です
△：会話として適当なオウム返し	発話：マスクメロン は 美味しい です ね 応答：美味しい です ね
×：会話として適当で無いオウム返し，相槌	発話：沖縄 いつ 行く の ? 応答：沖縄

入力に過去の発話を含まないデータで学習したモデルにおける，テスト出力結果の絶対評価を表 13 に示す．

表 13: 評価結果 (入力に過去の発話を含まない)

手法	○	△	×
低頻度語置き換え無し	68	98	34
低頻度語全てをヌルトークン化	81	93	26
Copyable Model	77	86	37
提案手法	75	97	28

入力に過去の発話を含んだデータで学習したモデルにおける，テスト出力結果の絶対評価を表 14 に示す．

表 14: 評価結果 (入力に過去の発話を含む)

手法	○	△	×
低頻度語置き換え無し	64	77	59
低頻度語全てをヌルトークン化	73	70	57
Copyable Model	72	69	59
提案手法	78	72	50

また，学習・テスト時に過去の発話を使用しなかったモデルの出力 (応答) を，過去の発話を含むモデルと同様な評価方法により評価した．すなわち，本来の発話データとその直前一発話分の発話を文脈とし，それ以前の文脈は見ずに評価する．絶対評価を表 15 に示す．

表 15: 評価結果 (入力に過去の発話を含まないが，評価では文脈に過去の発話があるとする)

手法	○	△	×
低頻度語置き換え無し	66	74	60
低頻度語全てをヌルトークン化	80	70	50
Copyable Model	70	69	61
提案手法	78	69	53

5 考察

本実験では提案手法での大きな性能向上は見られなかった．CopyableModel は単にヌルトークン unk_0 へ置き換えた場合よりも性能が悪かったため，CopyableModel は対話においてはあまり有効ではないと思われる．そのため，CopyableModel を用いずにヌルトークン unk_0 の品詞置き換えを行うモデルが有効な可能性がある．その手法を用いた実験を今後行いたい．

全手法において，出力結果から「うん」等使用頻度の高い応答が複数あった．例として提案手法の出力数上位 5 文を表 16 に示す．この中で，「うん」はテストデータの総数 694 のおよそ半分を占める．過去の発話を含む学習データにおける出力では，同じ応答の使用頻度は全ての手法において減少した．

「うん」は文脈により意味，役割が変わりやすく，厳密な評価が困難である．表 16 の出力は全て「うん」と同じく文脈により評価がわかれやすい．また，非タスク指向型対話システムにおいて，似たような出力が多くなることは好ましくないと思われる．そのため，出力 (応答) の多様性を評価するためには，評価○，△とした出力の，異なる内容の出力数を調査する必要がある．これは今後の課題とする．

表 16: 提案手法の出力数上位 5 文

過去の発話を含まない		過去の発話を含む	
出力 (応答)	出力数	出力 (応答)	出力数
うん	328	うん	258
ふーん	27	ふーん	31
うーん	16	うーん	22
そうですね	11	何?	13
はい	9	そうですね	11

6 おわりに

ヌルトークンの品詞に基づく細分化と，評価のしやすさの向上のためのデータ調整を行った．

結果として，品詞情報付加によるヌルトークン unk_0 の細分化により，出力に変化は現れたが，精度向上とはならなかった．

今後は，今回の実験における出力文の多様性の調査と，CopyableModel を用いないヌルトークン unk_0 の品詞置き換えモデルの評価を行う．また，対話データが少ないため，データ追加を行いつつ，出力の多様性を増加させる必要がある．

参考文献

[1] 佐藤 翔悦, 吉永 直樹, 豊田 正史, 喜連川 優 : 暗黙の発話状況を考慮したニューラル対話モデル, 言語処理学会第 23 回年次大会発表論文集, 2017.

[2] Ilya Sutskever, Quoc V.Le, Oriol Vinyals, Wojciech Zaremba : Addressing the Rare Word Problem in Neural Machine Translation, ACL 2015, 2015.

[3] 関沢 祐樹, 梶原 智之, 小町 守：目的言語の低頻度語の高頻度語への言い換えによるニューラル機械翻訳の改善, 言語処理学会第 23 回年次大会発表論文集, 2017.

[4] 村田 真樹, 内山 将夫, 白土 保, 井佐原 均 : シリーズ型質問文に対して単純結合法を利用した逓減的加点質問応答システム, システム制御情報学会論文誌, Vol.20, No.8, pp.338-346, 2007.

[5] Ilya Sutskever, Oriol Vinyals, and Quoc V.Le. : Sequence to sequence learning with neural networks., Advances in Neural Information Processing Systems 27(NIPS2014), 2014.

[6] SYSTRAN : OpenNMT, http://opennmt.net/

[7] 対話破綻検出チャレンジ, https://sites.google.com/site/dialoguebreakdowndetection/

[8] 名大会話コーパス, http://mmsrv.ninjal.ac.jp/nucc/