5R-05

犬型ロボットの動作を伴うマルチモーダル応答生成に向けて

寺面 杏優　　梶原 智之　　二宮 崇

愛媛大学大学院理工学研究科

1 はじめに

大規模言語モデル（LLM: Large Language Model）をはじめとする深層学習の技術の進化に伴い、言語情報と非言語情報を統合するマルチモーダル対話システムの開発に大きな期待が寄せられている。本研究では、発話内容（言語情報）とロボット動作（非言語情報）の対応付けに注目し、人間の発話に対して犬型ロボットが動作とともに応答するマルチモーダル雑談対話システムを構築する。

深層学習に基づく対話システムの構築には、名大会話コーパス[1]や日本語日常対話コーパス[2]など、発話文と応答文の対からなる対話コーパスが使用される。これらは、人間同士の対話を収集したものや人間同士の対話を想定して構築されたものであり、犬型ロボットとの対話という状況や犬型ロボットの動作を考慮した日本語の雑談対話コーパスは存在しない。

本研究では、犬型ロボットの動作を伴う発話応答対 8,000 件を収集し、公開[1]する。そして、本コーパスを用いて大規模言語モデルを訓練し、動作とともに応答を出力するマルチモーダル対話システムを構築する。評価実験の結果、我々の対話コーパスを用いた訓練によって、犬型ロボットによる応答生成と動作選択の両方の品質が向上することを確認できた。

2 データセット構築

本研究では、犬型ロボットとの雑談対話システムの訓練のために、図 1 に示すように「人間の発話文・犬型ロボットの応答文・犬型ロボットの動作」の 3 つ組データセットを構築する。

2.1 犬型ロボットの動作

本研究で扱う犬型ロボットは、Unitree Go2[2]である。本ロボットには、あらかじめ 11 種類の動作が登録されており、本研究ではそのうちの 8 種類の動作を対象として、マルチモーダル対話コーパスを構築する。表 1 に動作の一覧を示す。この 8 動作の他にも、Dance2 や Roll around などの動作が登録されているが、他の動作との差別化が難しいことや、動きが大きいため物理的に実験が難しいことが理由で、本研究では表 1 の 8 動作のみを用いる。

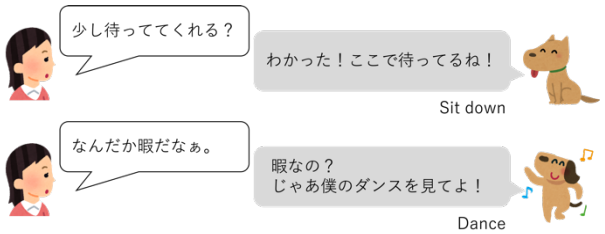

図 1：構築したデータセットの例

2.2 クラウドソーシングによるデータ収集

発話・応答・動作の 3 つ組を人手で作成した。このデータ構築は、まず本稿の第一著者が 800 件（100 組×8 動作）を作成した。その後、収集済みデータとの一貫性と多様性の両方を考慮しつつ、クラウドソーシングを用いて 7,200 件（15 組×8 動作×60 人の作業者）を収集し、合計 8,000 件の 3 つ組データを得た。クラウドソーシングサービスには Lancers[3]を使用し、1 文対あたり 100 円（時給 1,200 円の想定）の報酬を支払った。

データ収集の一貫性を高めるために、犬型ロボットのペルソナや発話スタイルについて、以下のように設定し、事前に作業者に指示をした。

- 人間のように日本語を話します。「ワン」など特殊な語尾を使わないでください。
- ですます調を使わず、カジュアルな表現を使用してください。
- 元気でポジティブな性格です。
- 一人称は「僕」です。

2.3 データ収集における品質管理

データ収集のためにアノテーションツールを構築し、自然言語処理の技術を用いて品質管理を行った。品質管理は、以下のとおり、多様性と一貫性の担保・オウム返しの回避・流暢性の担保の 3 つの観点から実施した。

Toward Multimodal Dialogue with a Dog-Robot Responding in Action
† Ayu Teramen (teramen@ai.cs.ehime-u.ac.jp)
† Tomoyuki Kajiwara (kajiwara@cs.ehime-u.ac.jp)
† Takashi Ninomiya (ninomiya.takashi.mk@ehime-u.ac.jp)
Ehime University

[1] https://github.com/EhimeNLP/direction
[2] https://www.unitree.com/go2
[3] https://www.lancers.jp/

Copyright ©2025 Information Processing Society of Japan.
All Rights Reserved.

表 1：犬型ロボットの動作一覧

動作名	説明
Stretch	伸び
Shake hands	握手
Sit down	お座り
Dance	踊る
Jump forward	前にジャンプ
Cheer	後ろ足で立ち前足でハートを描く
Punch	後ろ足で立ち襲いかかる
Standing up	後ろ足で立ち前足を上下にふる

多様性と一貫性　データの多様性を担保するために、収集済みの発話文や応答文と類似度の高すぎる入力を拒否した。文の類似度は対照学習に基づく文符号化器の SimCSE[4]によって計算し、発話文においては 0.8 以上、応答文においては 0.9 以上の余弦類似度を持つ入力を拒否した。また、収集済みデータとの一貫性を担保するために、入力された応答文に対して類似度の高い順に 10 件の収集済みデータを表示し、矛盾がないことを作業者に確認させた。

オウム返し対策　応答が簡単な事例ばかりが集まることを避けるために、発話文と応答文の類似度が高すぎる入力を拒否した。先述のとおり SimCSE を使用し、0.9 以上の余弦類似度を持つ文対の入力を拒否した。

流暢性　誤字脱字などのノイズを避けるために、流暢性の低すぎる入力を拒否した。文の流暢性は言語モデル GPT-2[5]によって計算し、Perplexity が 300 以上の場合に入力を拒否した。

3 実験設定

モデル　日本語の大規模言語モデルである Swallow[6]および LLM-jp[7]を用いて対話システムを構築した。訓練には LoRA[3]を使用し、ランクを r=4、スケーリング係数を α=8、dropout 率を 0.05、バッチサイズを 8 文対、学習率を 5×10^{-5}、最適化手法を AdamW に設定した。そして、検証用データにおける交差エントロピー損失が 5 エポック連続で改善しない場合に訓練を終了した。

データ　本研究で構築した対話コーパスを用いて大規模言語モデルを訓練した。データは、動作が均等に分かれるように、訓練用に 6,000 件、検証用に 1,000 件、評価用に 1,000 件を使用した。

表 2：応答生成および動作選択の実験結果

	手法	訓練	BLEU	Accuracy
Swallow	先に応答生成	−	1.27	0.299
		✓	4.77	0.435
	先に動作選択	−	1.15	0.272
		✓	4.62	0.446
	同時推定	−	1.59	0.238
		✓	4.90	0.430
LLM-jp	先に応答生成	−	2.38	0.222
		✓	2.59	0.223
	先に動作選択	−	1.78	0.232
		✓	1.87	0.238
	同時推定	−	1.78	0.248
		✓	2.07	0.264

評価　応答生成の性能は BLEU[8]、動作選択の性能は Accuracy を用いて自動評価した。応答生成および動作選択は、以下の 3 つの方法で実施する。

- **先に応答生成**：発話文の入力に対して、まず LLM を用いて応答文を生成する。そして、発話文と応答文の 2 文を改めて LLM に入力し、動作を出力する。
- **先に動作選択**：出力の順番を逆にする。つまり、まず発話文から動作選択を行い、次に発話文と動作の組から応答生成を行う。
- **同時推定**：LLM への入出力を一度のみ行い、発話に対して応答と行動を同時に出力する。

上記の設定で、訓練前後の LLM ベース対話システムの性能を比較し、我々の対話コーパスを用いた訓練の有効性を検証する。

4 実験結果

表 2 に実験結果を示す。LLM の訓練によって一貫して性能が改善していることがわかる。特に Swallow の性能が大きく改善された。これらの実験結果から、動作を伴うマルチモーダル対話システムの開発のために、本研究で構築した対話コーパスが有用であることがわかる。

謝辞　太陽誘電株式会社の支援を受け実施した。

参考文献

[1] Itsuko Fujimura, Shoju Chiba, Mieko Ohso. Lexical and Grammatical Features of Spoken and Written Japanese in Contrast: Exploring a Lexical Profiling Approach to Comparing Spoken and Written Corpora. Proceedings of the VIIth GSCP International Conference: Speech and Corpora, pp.393-398, 2012.

[2] 赤間怜奈, 磯部順子, 鈴木潤, 乾健太郎. 日本語日常対話コーパスの構築. 言語処理学会第 29 回年次大会, pp.108-113, 2023.

[3] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models. ICLR, 2022.

[4] https://huggingface.co/cl-nagoya/sup-simcse-ja-base
[5] https://huggingface.co/rinna/japanese-gpt2-medium
[6] https://huggingface.co/tokyotech-llm/Llama-3.1-Swallow-8B-Instruct-v0.1
[7] https://huggingface.co/llm-jp/llm-jp-3-13b-instruct
[8] https://github.com/mjpost/sacrebleu

Copyright ©2025 Information Processing Society of Japan.
All Rights Reserved.