

Few-shot プロンプティングによる小説台詞からの人間関係語の抽出 Extracting Human Relationship Terms from Novel Lines using Few-Shot Prompting

安田 大朗[†] 安藤 一秋[‡]
Tomoaki Yasuda Kazuaki ando

1. はじめに

近年、雑談対話システムには、ユーザの短期的な話し相手になるだけでなく、長期的な対話ニーズを満たす雑談相手としての役割が期待されている[1]。雑談対話システムがユーザと長期的な信頼関係を構築するためには、人間らしく自然な対話を実現することが重要である。しかし、対話システムが誤った応答を生成することで、ユーザは困惑し、対話を継続する意欲が削がれる課題がある[2]。このような課題に対して、対話の一貫性を保つペルソナを自動的に更新する研究[3]や、特定の知識に基づき情報を提供することで、回答の質と正確性を向上させる研究[4]が実施されている。

本研究では、非タスク指向の対話中に出現する人間関係をユーザ情報として抽出・活用することで、長期的な信頼関係を築くことができる雑談対話システムの実現を目的とする。システムとの対話の中から人間関係を抽出・活用する方法を検討するためには、人間関係を含む対話データが必要となる。人間関係を含む対話データとして、既存の雑談対話コーパスの利用が考えられるが、既存のコーパスからは個人情報削除されているため、人間関係を含む会話が少ない。また、長期的な会話も含まれていない。たとえば、日本語日常対話コーパス[5]の 1 会話に含まれる平均発話数は 7.94 しかない。

そこで本研究では、小説における登場人物の台詞に注目する。小説の台詞から登場人物や発話者の人間関係が抽出できれば、日常対話におけるユーザの人間関係も抽出できるといえる。著者らの先行研究[6]では、小説の台詞において登場人物と発話者に紐づく関係語(e.g. 友人, 兄)を抽出するモデルを提案した。一方で、複数の人間関係が存在する複雑な文脈において、人名に紐づく関係語を抽出することが難しいという課題があった。そこで、本稿では、大規模言語モデル(LLM)の Few-shot プロンプティングにより、複雑な文脈下において、人名に紐づく関係語を効率的に抽出する手法を検討する。

2. 先行研究で提案した関係語抽出モデル

2.1 データセットの概要

著者らの先行研究[6]では、「小説家になろう」の現実世界を舞台とする小説の台詞を利用し、登場人物と発話者に紐づく関係語を抽出するモデルを提案した。括弧(「」, 『』)で囲まれた文を台詞とし、小説 64 編から、人名と人間関係を表す単語(関係語)を 1 つ以上含む台詞を 2,000 件抽出してデータセットを構築した。関係語は角川類語新辞典と分類語彙表増補改訂版データベースから人手で抽出した 147 単語のうち、日本語日常対話コーパス[5]に頻出する上位 47 語を利用した。なお、アノテーション時には、抽出性能を高めるためにラベルの粒度を統一し、類似する関係

語(e.g. 母, 母親)を「母」に統一してラベルを付与した。

また、人間関係を含む台詞を 6 カテゴリに分類した。表 1 に人間関係を含む台詞のカテゴリを示す。登場人物と発話者の関係語を同時抽出するモデルを学習するため、データセットを登場人物と発話者の 2 種類のデータに分類した。また、提案したモデルは、台詞と人名 1 つを入力し、入力された人名に紐づく関係語を台詞から予測する。そのため、人名が複数存在する台詞は、人名の数だけデータを拡張して正負ラベルを付与した。たとえば、「結城と和樹は友達なのね」では、[結城-友達]、[和樹-友達]のように人名とそれに紐づく関係語を一对一の関係になるようデータを 2 つに拡張した後、カテゴリ間のデータ数の偏りを調整した上でデータセットを構築した。

2.2 データセット内のカテゴリ分布

データセットを構築する際は、登場人物と発話者のデータを結合し、1:1 の割合に調整した。また、それぞれのデータの正例と負例も均一になるように調整した結果、表 2 に示すデータの分布になった。

2.3 関係語抽出モデルの構築

先行研究[6]では、構築したデータセットを用いて BERT¹ を 3 値分類タスクでファインチューニングして関係語抽出モデルを構築した。また、登場人物と発話者でデータの入力方法が異なるため、登場人物のデータについては、[CLS]台詞[SEP]登場人物名として台詞と登場人物名を結合してモデルに入力する。発話者のデータについては、主語が省略されているため、[CLS]台詞[SEP]発話者のように、表層的に出現しない発話者の代わりとして、文字列の「発話者」を入力することで、紐づく関係語を予測する。

3. Few-shot による人間関係語の抽出

本稿では、LLM の Few-shot プロンプティングを用いて、小説の台詞から人名に紐づく関係語を抽出する手法(LLM モデル)を検討する。先行研究[6]で用いた BERT モデルでは、文脈が複雑な台詞において、人名に紐づく関係語の抽出が困難であった。本研究では、これらの台詞に対して、BERT モデルと LLM モデルによる抽出性能を比較検証する。

3.1 実験設定

先行研究[6]では、関係語の表現揺れに対応するため、類似する関係語(e.g. 母, 母親)を「母」に統一してラベルを付与していた。本実験でも、同様の粒度統一を実現するため、LLM モデルの出力に対して、ルールベースの統一処理を適用する。たとえば、抽出された「母親」「お母さん」等の表現を「母」に統一する。LLM に入力するプロンプトを図 1 に示す。対象となる発話と人物名を入力として与え、その人物に紐づく関係語のみを抽出させる。他の人物に紐づく関係語や代名詞、人名を抽出しないように指示を与え、Few-shot で抽出例を明記した。

[†] 香川大学大学院創発科学研究科 Graduate School of Science for Creative Emergence, Kagawa University

[‡] 香川大学創造工学部 Faculty of Engineering and Design, Kagawa University

表 1: 人間関係を含む台詞のカテゴリ

	Cat.	定義	例
登場人物	1	1つの人名と関係語が存在	「山田は友達だよ」 [山田-友達]
	2	2つの人名と関係語が存在	「結城と和樹は友達なのね」 [結城-友達-和樹]
	3	複数の人間関係が存在	「私の兄, 和樹に恋人がいた」 [私-兄-和樹] [和樹-恋人]
	4	文脈理解が必要な人間関係	「結城さんは友達だったけど, これからは恋人」 [結城-恋人]
発話者	5	発話者の人間関係のみ存在	「今日は妻との結婚記念日」 [発話者-妻]
	6	発話者と登場人物の人間関係が存在	「妻とショッピング中に, 友人の和樹に会った」 [発話者-妻] [和人-友人]

表 2: 構築したデータセット内のデータ分布

データ種類	登場人物				発話者		負例
Cat	1	2	3	4	5	6	7
調整後	100	100	100	76	188	188	752

You are a competent Japanese assistant.
Given a dialogue and a target person, extract all human relationship terms (e.g., '妹', '友達') that directly refer to the target person. Ignore terms referring to others, as well as personal names and pronouns. Output as a json array (e.g., ['友達']). If none, return [].

Examples

user: 「美咲, 私の友達に会う?」, name 「私」

assistant: ['友達']

user: 「美咲, 私の友達に会う?」, name 「美咲」

assistant: []

user: 「凛の知り合いね. 私は姉と来た」, name 「凛」

assistant: ['知り合い']

user: 「凛の知り合いね. 私は姉と来た」, name 「私」

assistant: ['姉']

Input

user: 「{line}」, name 「{person}」, assistant:

図 1: プロンプトの内容

LLM には, Gemma 2-9b-it² を利用し, 複雑な文脈を含む台詞 (カテゴリ 3) のデータ 500 件に対して, 先行研究で提案した BERT モデルとの性能を比較する. 評価指標には, Precision, Recall, F1 を用いる.

3.2 評価結果

評価結果を表 4 に示す. LLM モデルは, BERT モデルと比較して, 全体的に高い性能を示した. 特に, Precision は LLM モデルが大幅に高く, 複雑な文脈においても正しい関係語を正確に抽出できている. また, Recall も BERT モデルを上回っており, 多様な関係語を抽出できていることが確認できる. これらの結果から, LLM の Few-shot プロンプティングは, 複雑な文脈に含まれる関係語の抽出に有効であることが示唆された.

3.3 抽出傾向とエラー分析

分析の結果, LLM モデルは, BERT モデルより, 多様な関係語を抽出する傾向を確認した. たとえば, 「幼馴染」や「保護者」など BERT モデルの学習データには含まれない関係語も抽出していた. また, 「バイト仲間」のように

¹ <https://huggingface.co/tohoku-nlp/bert-base-japanese-v3>

² <https://huggingface.co/google/gemma-2-9b-it>

表 3: 性能の比較

	Precision	Recall	F1
BERT	0.5432	0.6213	0.5796
LLM (Gemma2-9b-it)	0.7097	0.6667	0.6875

具体的な関係表現についても抽出していた. これらの結果は, LLM モデルの文脈理解の深さを反映しているといえる. 一方で, 「皇の令孫」のように人名と関係語が混同した誤抽出も存在していた. また, 「先生」や「結婚」「味方」など一部の関係語については, 抽出漏れも確認した. これらの関係語は, 人間関係を表すか否かの判定が周囲の文脈に依存するため, 抽出が困難であった可能性がある.

4. おわりに

本稿では, 先行研究で提案した BERT モデルと LLM の Few-shot プロンプティングを用いた関係語抽出モデルの性能を比較した. 実験の結果, LLM モデルは BERT モデルと比較して, 全体的に性能が向上することを確認した. LLM モデルは, 文脈理解の柔軟性が高いことから, BERT モデルで抽出できなかった複数の人間関係が存在する複雑な文脈に対しても, 高い抽出性能を得た. 一方, 本研究で人名との混同や文脈依存性による課題も確認された.

今後の課題として, LLM モデルが抽出する多様な関係語を管理・活用するためには, 関係語の粒度や表現の統一を含むスキーマ設計が必要となる.

参考文献

- [1] J. Xu, et al., "Beyond Goldfish Memory: Long-Term Open-Domain Conversation," in Proc. of the 60th Annual Meeting of the Association for Computational Linguistics, pp.5180-5197, (2022).
- [2] B. Martinovsky, et al., "The error is the clue: Breakdown in human-machine interaction," In Proc. ISCA Workshop on Error Handling in Spoken Dialogue Systems, pp.11-16, (2003).
- [3] X. Xu, et al., "Long Time No See! Open-Domain Conversation with Long-Term Persona Memory," In Findings of the Assoc. for Computational Linguistics: ACL 2022, pp. 2639-2650, (2022).
- [4] Y. Gao, et al., "Retrieval-Augmented Generation for Large Language Models: A Survey," arXiv:2312.10997, (2023).
- [5] 赤間他, "日本語日常対話コーパスの構築", NLP2023 発表論文集, pp.108-112, (2023).
- [6] 安田他, "小説台詞からの登場人物と発話者に紐づく人間関係語の抽出", JSAI2025 論文集, 4 pages, (2025).