小説のセリフを利用した登場人物に紐づく人間関係語の抽出

安田 大朗[1] 安藤 一秋[1]
[1]香川大学創造工学部
{s21t335, ando.kazuaki}@kagawa-u.ac.jp

概要

日常対話を目的とした非タスク指向型対話システムは,ユーザと長期的な信頼関係を構築するために,深く対話を継続する必要がある．システムの実現には，人間同士のコミュニケーションのように共生や共感といった視点が重要であり，ユーザの個々の情報を把握・活用する必要がある．本研究では，ユーザの情報として人間関係に着目する．また，小説の台詞から登場人物の人間関係を抽出できれば，日常対話におけるユーザの人間関係も抽出できるという仮説のもと，日常対話の代替として小説の台詞を活用する．本稿では，台詞に出現する登場人物とその人物に紐づく人間関係語を抽出するモデルを検討する．実験の結果，0.6618 の F 値で人間関係語を抽出できることを確認した．

1 はじめに

近年，高齢者や若者の孤独感を減らす手段としてスマートスピーカーのような雑談対話を目的とした非タスク指向型対話システム（雑談対話システム）への関心が高まっている[1, 2]．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．

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．また，口調の急激な変化などを抑制する観点から，プロンプトに対人関係の情報を与えて発話を制御する研究[5]も実施されている．

本研究では，非タスク指向型対話内に出現する人間関係をユーザの情報として抽出･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．

会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，人間関係を含むデータが必要となる．人間関係を含むデータとして，既存の雑談対話コーパスの利用が考えらえるが，既存のコーパスからは個人情報が削除されており，個人の人間関係を含む会話が利用できない．また，長期的な会話も存在していない．そこで本稿では，小説の台詞に注目する．小説の台詞から登場人物の人間関係を抽出できれば，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．また，人間関係を抽出する手法の初期検討として，台詞に登場する人物とその人物に紐づく人間関係語を抽出するモデルを構築し，性能を評価する．

2 関連研究

小説から人間関係を抽出する研究[6, 7, 8]は，人物相関図の構築を目的として実施されているものが多い．西村らの研究[7]では，物語文から人間関係を抽出するために，人手で作成した文の係り受けを考慮したパターンと半自動で獲得したパターンを用いて,人名とその関係を表す単語を抽出する手法を提案している.実験の結果,人手によるパターンでは 0.290,半自動では 0.340 の F 値が得られたと述べている．

内野らの研究[8]では，人物の関係を表す単語を[MASK]に置換したテンプレートを文末に挿入することで，小説に登場する人物同士の関係性を予測する手法を提案している．実験の結果，人物同士の関係性を最大 0.8 の正解率で予測したと述べている．

3 既存コーパスと小説台詞の比較

日本語日常対話コーパス[9]に，人間関係を含む発話がどの程度存在しているのか調査する．1 発話内に人名および人間関係を表す単語（関係語）が各 1 つ以上含まれているものを関係文候補として抽出し,その数と割合を調査する．また,「小説家になろう」[10]の現実世界を舞台とする小説 18 編から同程度の

台詞を抽出して比較する．発話からの人名抽出には GiNZA を利用する．関係語には，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抽出した 147 語を用いる．

調査結果を表 1 に示す．既存コーパスよりも小説のほうが関係文候補の数が多く，割合も高いことから，多様な人間関係が含まれている可能性がある．また，小説は今後も継続的に公開されるため，量的問題も解消できる．よって，人間関係を抽出するデータとして小説の台詞に注目する．

表 1: 関係文候補の調査

	既存コーパス	小説(18 編)
発話・台詞数	41,737	40,515
関係文候補（件）	248	1,298
関係文候補（%）	0.59	3.20

4 データセットの構築

台詞に登場する人物とそれに紐づく関係語を抽出するモデルの学習・評価で利用するデータセットの構築法について述べる．

4.1 データセットの概要

本研究では，括弧（「」,『』）で囲まれた文を台詞とする．「小説家になろう」の現実世界を舞台とする小説 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 件抽出して利用する．なお，関係語は，実際の対話に出現する人間関係を学習するため，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抽出した 147 単語のうち，日本語日常対話コーパスに頻出する上位 47 語を利用する．

西村らが提案した人間関係の種類[7]を参考に，人間関係を含む台詞を 5 カテゴリに分類する．表 2 にカテゴリの定義と例を示す．なお，本稿における人間関係は，人名とそれに紐づく関係語の一対一の関係を指す．例えば，「山田は友達だよ」には，[山田-友達]という人間関係が存在するとみなす．

表 2 に示すカテゴリ 1 は，1 つの人名とそれに紐づく関係語を含む台詞である．カテゴリ 2 は，2 つの人名とそれらの関係語を含む台詞である．カテゴリ 3 は，発話者である主語（代名詞を含む）が出現しないが,発話者と紐づく関係語を含む台詞である．カテゴリ 4 は，複数の関係語が存在する（複数の人間関係が存在する）台詞である．カテゴリ 5 は，上記のいずれかのカテゴリに該当するが，人間関係の解釈には文脈理解が必要であると判断し，4 つのカテゴリと区別したものである．データセットにおいては，以上の 5 カテゴリを正例とする．負例は，人間関係が存在しない台詞とし，カテゴリ 6 に分類する．人手で 2,000 件を分類した結果，カテゴリ 1 の件数が最も多く，次いでカテゴリ 3 が多かった．

台詞には,複数の人名が出現する場合があるため，人名とそれに紐づく関係語を一対一の関係でラベリングができない．そこで，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する．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対一の関係になるようデータを 2 つに拡張する．データの拡張後，カテゴリ間のデータ数の偏りを調整したものを最終的なデータセットとして利用する．データセットにおけるカテゴリの分布を表 3 に示す．最終的に，957 件の正例と 957 件の負例からなるデータセットが構築できた．

4.2 各カテゴリの定性分析

カテゴリ 1 のデータは，「山田は友達だよ」のように，人名と関係語が一対一で存在していることから，抽出しやすいと考えられる．カテゴリ 2 のデータは，「結城と和樹は友達なのね」のように人名が 2 つと関係語が 1 つで構成されるが，カテゴリ 1 と同様，予測する関係語は 1 つであるため，こちらも抽出しやすいデータといえる．カテゴリ 3 のデータは，「今日は妻との結婚記念日」[発話者-妻]のように，表層上は出現していない“発話者”に紐づく関係語を含む．発話者以外の登場人物との関係を示す場合もあるため，カテゴリ 1 と 2 と比較して，抽出は難しいといえる．カテゴリ 4 のデータは，1 つの台詞に 2 つ以上の人間関係を含む．たとえば，「私の兄，和樹に恋人がいた」という台詞の場合，{[私-兄-和樹], [和樹-恋人]}のように異なる人間関係が複数存在するため，抽出は難しいといえる．最後に，カテゴリ 5 のデータは，「結城さんは友達だったけど,これからは恋人」のように，人間関係の抽出に文脈理解が必要であるため，最も難しいといえる．

5 実験

5.1 人名に紐づく人間関係語抽出モデル

構築したデータセットを用いて，台詞に登場する人物とその人物に紐づく人間関係語を抽出するモデ

表 2: 人間関係の 5 カテゴリ

Cat.	定義	例
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	発話者の主語が省略されている人間関係が存在	「今日は妻との結婚記念日」 [発話者-妻]
4	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]
5	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]

表 3: データセットにおけるカテゴリの分布

	1	2	3	4	5	6
拡張前	432	214	317	148	57	624
拡張後	432	426	247	306	111	957
調整後	210	210	216	210	111	957

ルを構築し，抽出性能を評価する．本実験では，ルールに基づく抽出モデル(ルールベースモデル)と，BERT（tohoku-nlp/bert-base-japanese-v3）を用いた抽出モデル（BERT モデル）の 2 つを比較する．

ルールベースモデルには，西原らが提案した文の係り受けを考慮した抽出パターンを利用する．この抽出パターンを表 4 に示す．これらのパターンは，青空文庫の小説 10 編をもとに人手で作成されている．「~」は任意の文字列，「|」は複数の助詞のいずれかにマッチすることを示す．P1，P2，R にマッチした人名や関係語を抽出する．また，言語モデルでの抽出と同じ条件で性能を図るため，人名が抽出できた場合を対象に，その人名に紐づく関係語が抽出できているかを確認する．

BERT モデルへの入力は，“[CLS]台詞[SEP]人名”のように, 台詞と登場する人名を[SEP]で繋げた形式とする. また, 人名に紐づく関係語を抽出するため，BIO タグによる 3 値分類モデルとしてファインチューニングする．台詞は，複数文で構成される場合もあるため，台詞全体を用いた場合と，文単位に分けた場合において，それぞれ性能を評価する．

5.2 実験設定

台詞に複数の人名が出現した場合，人名の数だけデータを拡張しているため，訓練データとテストデータで同一の台詞を利用しないように，層化 10 分割交差検証を用いて評価する．それぞれの正例で True Positive，False Positive，False Negative を求め，全体における Precision，Recall，F1 値を計算し，評価指標とする．各検証において，テスト loss が最小になった epoch 時のモデルを採用する．最適化アルゴリズムは AdamW を用いる．

BERT モデルでは，人間関係語をどれだけ検出できたかをカテゴリごとに確認するため，Recall を比較する．また，台詞単位と文単位でそれぞれ学習させ，抽出性能の変化を確認する．

5.3 評価結果

まず，BIO タグの予測性能を表 5 に示す．B タグは，I タグと比較して Recall は高いが，Precision が低い．また，一般的に開始を示す B タグのほうが I タグより学習データに含まれる数が多いにもかかわらず, I タグより Precision が低い. この原因として，I タグは B タグに連続して出現するため予測しやすいことが影響していると考える．

次に，ルールベースモデルと BERT モデルの予測性能を表 6 に示す．ルールベースモデルの Recall は低いが，Precision は最も高い．このことから，ルールベースモデルは，BERT モデルと比較して，高い割合で正しい関係語を抽出できているといえる．一方，台詞単位で学習させた BERT モデル（BERT-台詞）は，recall が最も高く，ルールベースで抽出できなかった関係語を正しく抽出できているといえる．

全てのカテゴリにおいて，台詞単位の方が文単位で学習させるよりも性能が高い結果となった．文単位の場合，成立しない人間関係をうまく学習できていない点や，利用できる情報量が少ないことなどが影響していると考えられる．以上より，学習データの単位は台詞単位がよいことを確認した．

次に，各カテゴリの recall を表 7 に示す．ルールベースは，全てのカテゴリにおいて，BERT モデルより人間関係語の抽出率が低い．台詞単位で学習させたモデルは，抽出候補の関係語が 1 つで誤抽出が少ないカテゴリ 1 とカテゴリ 2 において，特に高い

性能となった．また，カテゴリ 5 についても高い性能となったことから，文脈理解が必要と判断したデータに対しても，BERT モデルは適応できる可能性がある．文単位で学習させたモデルは，台詞単位と比較すると性能が低下している．

表 4: 西原らの抽出パターン

抽出パターン
P1 の P2&R [結城の兄]
P1 の R(の|は)P2 [結城の兄の太郎]
P2(が|は|も)~P1 の~R [結城は和樹の中学の友達]
P1(が|は|も)~R の P2 [結城は時々兄の太郎と]
P1(に|には)P2 という R [結城に太郎という兄が]
P1 が~R、P2 [結城が友人、和樹と]
P1(が|は|も)~P2&R(と|を|に|の) [太郎が公園で妹と]
P2&R(が|は|も)~P1(と|を|に|の) [兄が結城を]

表 5: BIO タグの予測性能

	Precision	Recall	F1
B	0.6318	0.6120	0.6176
I	0.6417	0.5713	0.5940
O	0.9993	0.9994	0.9993

表 6: 全体の性能

	Precision	Recall	F1
Rule	**0.7020**	0.3270	0.4460
BERT-台詞	0.5739	**0.7842**	**0.6618**
BERT-文	0.4970	0.6793	0.5733

表 7: カテゴリ別の Recall

	1	2	3	4	5
Rule	0.3606	0.3761	0.2423	0.3135	0.2817
BERT-台詞	**0.8094**	**0.8871**	**0.8448**	**0.6857**	**0.7757**
BERT-文	0.6788	0.7813	0.7160	0.6111	0.4189

6 エラー分析

カテゴリ 1 とカテゴリ 2 は，抽出する関係語の候補が 1 つのため誤抽出の可能性は少ない．また，「山田の母」のように，人名と関係語が近くに出現する事例の抽出性能が高い．一方，「私はたかしの恋人であって，母親になるつもりはない」（ふん，『2.1 チャンネルスピーカーズ』より引用）という台詞において，人名“私”に紐づく関係語は“恋人”であるが，“母”を誤抽出していた．近くに出現した人間関係を示さない関係語を誤抽出している例が見られた．カテゴリ 3 では，表層的には登場しない発話者の人間関係が存在するため，誤抽出が増える可能性がある．誤抽出の例として「父の状態が危ないようです．今から病院に向かいます．同志ヨシオ」（ベキオ，『麗子の風儀～悪役令嬢と呼ばれましたが、ただの貧乏娘です～』より引用）という台詞において，人名“ヨシオ”に紐づく関係語“同志”以外にも，発話者と人間関係があると思われる関係語“父”を誤抽出していた．カテゴリ 4 は，台詞内に複数の人間関係が存在するため，誤抽出する可能性が高い．「残念ですけど，神主さんの奥さんのお体の方が大事なのでお大事になさって下さい……本当に，はるはるのお父さん，なんでもやってるな」（黒川天理，『そして少女は悪女の体を手に入れる』より引用）という台詞においては，人名“神主”に紐づく関係語“奥さん”を抽出すべきであるが，“父”を誤抽出していた．カテゴリ 5 は，関係語の抽出に文脈理解が必要であると判定した台詞である．「俺と凛香はネトゲで結婚しているんだ．それである日，リアルでも知り合いになって……凛香はリアルでも夫婦だと言ってきたんだよ」（あボーン，『ネトゲの嫁が人気アイドルだった件～クール系の彼女は現実でも嫁のつもりでいる～』より引用）という台詞においては，人名“凛香”に紐づく関係語“知り合い”を抽出すべきであるが，現実の関係ではない“結婚”を誤抽出していた．最後に共通するエラーとして，関係語の一部のみしか抽出できなかった例がある．たとえば，“先輩”として抽出すべき関係語に対して，“先”のみを抽出していた．これは，データ量の少なさが影響しているといえる．

7 おわりに

本稿では，会話の中から人間関係を抽出する手法の初期検討として，台詞に登場する人物とその人物に紐づく人間関係語を抽出するモデルを構築し，性能を評価した．人名に紐づく単純な関係であれば，precision と recall が共に 0.8 以上と高い性能を得た．しかし，カテゴリ 4 のように複数の人間関係を含む台詞については，人名に紐づく関係語の抽出性能が低いため，改善の必要がある．

今後の課題として，複数の人間関係が出現する場合の性能向上や，カテゴリ 3 の表層的に出現しない発話者の人間関係の抽出に取り組む．

参考文献

1. 回想法を模擬した高齢者向け対話システムの構築に関する研究. 中島悠,栂井良太,伊東伸泰,西田昌史,西村雅史. 情報処理学会 第 78 回全国大会 講演論文集, pp.787-788, 2016.
2. 回想法を模した高齢者向け傾聴対話システムの検討. 太田壱成,綱川隆司,遠藤幹也,都築俊宏,西村雅史. 情報処理学会 第 82 回全国大会 講演論文集, pp.465-466, 2020.
3. 対話システムにおけるペルソナの自動生成による更新. 川本稔己,山崎天,佐藤敏紀,船越孝太郎,奥村学. 言語処理学会 第 29 回年次大会 発表論文集, pp.399-404, 2023.
4. 知識グラフの対話システムへの記憶化:学習アプローチの探求. 薛強, 滝口哲也, 有木康雄. 言語処理学会 第 30 回年次大会 発表論文集, pp.1453-1457, 2024.
5. 大規模汎用言語モデルを用いた雑談対話システムの対人関係性に基づく発話制御の検討. 山崎天, 川本稔己, 吉川克正, 佐藤敏紀. 言語処理学会 第 28 回年次大会 発表論文集, pp.1921-1925, 2022.
6. 物語テキストにおけるキャラクタ関係図自動構築. 神代大輔, 高村大也, 奥村学. 言語処理学会 14 回年次大会 発表論文集, pp. 380–383, 2008.
7. 物語テキストを対象とした登場人物の関係抽出. 西原弘真, 白井清昭. 言語処理学会 第 21 回年次大会 発表論文集, pp.628-631, 2015.
8. 物語を対象とした登場人物の関係図抽出. 内野太智, Danushka Bollegala, Naiwala P. 言語処理学会 第 30 回年次大会 発表論文集, pp.450-454, 2024.
9. 日本語日常対話コーパスの構築. 赤間怜奈, 磯部順子, 鈴木潤, 乾健太郎. 言語処理学会 第 29 回年次大会 発表論文集, pp.108-113, 2023.
10. 小説家になろう https://syosetu.com/