E-007

日本語の日常対話コーパスの構築に向けた小説テキストの分析

Analysis of Novel Text for Building a Japanese Daily Dialogue Corpus

岩本 和真† 安藤 一秋†
Kazuma Iwamoto Kazuaki Ando

1. はじめに

大規模言語モデルを用いた対話システムが注目される中，対話モデルを構築，改善するために必要なコーパスの重要性が高まっている．近年，日本語の対話コーパスは少しずつ整備されつつあるが，英語のコーパスと比較して十分とはいえない．対話コーパスの構築法には，人間同士の対話を録音して書き起こす[1]，SNS の発信応答をペアで収集する[2]，人手で対話文を作成する[3]などの方法がある．しかし，これらの構築法には，録音のための環境設定やコーパスからの個人情報の削除など，膨大なコストがかかる．また，対話コーパスにおける対話データの質と量の担保が難しいといった課題もある．

この課題を解決する方法の 1 つに，小説テキストを用いた対話コーパスの構築が考えられる．小説内では，登場人物の台詞が「」や『』などで記述されており，それらが対になっている部分を登場人物同士の対話ととらえることができる[4, 5]．そこで本研究では，小説テキストの登場人物同士の対話とその周辺テキストを活用し，大規模な日常対話コーパスの自動構築を目的とする．

本稿では，事前検討として，小説テキストから対話コーパスが生成できる可能性について分析する．まず，小説テキストに含まれる台詞について定量分析した後，既存の対話コーパスにおける発話の語彙的類似性を比較する．そして，小説テキストから抽出した台詞を単につなげて対話コーパスを構築し，そのコーパスで学習した対話モデルと既存のコーパスで学習したモデルの性能を比較する．

2. 分析

小説テキストから抽出できる台詞について，その規模や内容の観点から対話コーパスとして活用できる可能性を分析する．本稿では，小説家になろう[6]において，日常会話に近い表現が多く使われていると推測した「恋愛」と「推理」ジャンルから無作為抽出した各 20 編と，青空文庫[7]から無作為抽出した長編 20 編を分析対象とする．

本稿では，これらの小説から「」と『』で囲まれている文を台詞とみなして抽出し，単に台詞をつなげた作成したコーパスを疑似対話コーパスとよぶことにする．

2.1 定量分析

小説から収集できる対話の規模を定量的に分析する．本分析では，疑似対話コーパスにおける平均台詞数，台詞の平均連続数，小説あたりの平均トークン数と平均異なり語彙数，台詞あたりの平均トークン長に注目する．なお，台詞のトークン分割には，mecab-ipadic-NEologd で辞書拡張した MeCab を用いる．

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

分析結果を表 1 に示す．まず，表 1 の平均台詞数に注目する．日本語日常対話コーパス[3]の発話数が 41,780 であることから，小説ジャンルによりばらつきはあるものの 5～10 タイトルあれば，同規模の発話数が得られる可能性がある．次に，台詞の平均トークン長は，どの小説ジャンルも最低 16 以上あり，日本語日常対話コーパスの平均トークン長 16.36 と同程度以上ある．よって，分析対象として小説テキストには十分な対話が存在しており，対話コーパスを構築する情報源に活用できる可能性が高い．

次に，台詞の連続性に注目する．平均連続数全体における連続数 0 の占める割合は，それぞれ 47.7%，57.2%，70.2%で最大となった．連続数 0 には単独の台詞や強調に「」が使われる事例なども含まれているため，疑似対話コーパスのままでは対話データとして利用できない．しかし，図 1 に示す例のように，台詞間に状況や感情などを説明する文を挟んでいるが，離れた台詞は対話とみなせる場合がある．したがって，このような台詞を繋げて対話データに利用することで，1 以上の連続数を増やすことができる．

「なにを作っているの？」
鍋で草のようなものをゆでている。
「風邪薬です」
（日向夏『薬屋のひとりごと』から引用）

図 1　状況説明文を挟むが連続する台詞といえる事例

2.2 既存の対話コーパスとの語彙的類似性の比較

既存の 3 つの日常対話コーパスと疑似対話コーパスの語彙的類似性を分析し，疑似対話コーパスの日常対話としての妥当性を確認する．比較対象として，発話者の性格や自己開示的な発話を含むことで発話者のペルソナを反映した JPersonaChat (JPC)[8]，感情的な状況に基づいた共感的な会話で構成される JEmpatheticDialogues (JED)[8]，規範的な日本語表現で構成される日本語日常対話コーパス（JDailyDialogue: JDD）[3]を用いる．コーパスの類似度の計算には，名詞と動詞に基づく Bag-of-Words ベクトルをそれぞれ利用し，コサイン類似度で比較する．なお，コーパスごとに含まれる語彙数が異なるため，頻度上位の 1,000 名詞と，頻度上位の 500 動詞を用いる．

表 2 に比較結果を示す．なお，表中の項目において，コーパス略称に続く N は名詞，V は動詞で意味する．表 2 より，恋愛ジャンルはすべてのコーパスと類似度が高い結果となった．また，青空文庫の類似度は全体的に低いことから，青空文庫より小説家になろうの小説の方が日常対話に適しているといえる．

JPC-N については，どの小説群とも低い値となった．JPC-N と JED-N，JDD-N の類似度はそれぞれ 0.392 と 0.479 であることから，JPC にはない名詞が多く含まれていると考えられる．また，JED-N と JED-V との類似度がすべて高いことから，小説での台詞は共感的な会話が多い可能性がある．

表 1　定量分析の結果

小説群	平均台詞数	台詞の平均連続数						平均トークン数	平均異なり語彙数	台詞の平均トークン長
		0	1	2	3	4	5 以上			
恋愛	8184.0	1928.6	1043.6	555.6	300.1	112.7	99.85	122658.9	7286.0	16.9
推理	3743.3	1177.85	483.6	209.75	95.85	42.1	47.75	71136.9	5448.9	21.0
青空文庫	1021.9	367.6	80.3	24.4	14.75	7.55	28.6	32193.2	3331.9	33.5

表 2　既存コーパスとの類似度評価

小説群	JPC-N	JPC-V	JED-N	JIE-V	JDD-N	JDD-V
恋愛	0.299	0.520	0.416	0.567	0.346	0.497
推理	0.247	0.441	0.350	0.495	0.304	0.451
青空文庫	0.184	0.317	0.241	0.336	0.209	0.304

3. 対話モデルを用いた分析

疑似対話コーパス（小説から抽出した台詞を単につなげたコーパス）で学習した対話モデルと既存のコーパスで学習した対話モデルを用いて，会話の質，一貫性，満足度などを比較し，小説対話コーパスとしての課題点を分析する．

3.1 設定

本稿では，恋愛ジャンルの 2 小説から抽出した台詞をつなげて構築したコーパス（5,699 会話）と JDD（5,256 会話），JPC（4,989 会話）を学習用の対話コーパスとして用いる．対話用の事前学習モデルには，GPT-2（rinna/small）を使用する．生成には直前の 5 発話を入力とし，top-p=50, top-k=0.95 でサンプリングする．

被験者 5 人（学部学生 2 名，大学院生 3 名）にそれぞれのモデルと対話してもらい，アンケート結果（リッカート尺度による 5 段階評価）と対話履歴を評価に用いる．各モデルの性能評価は自動評価と人手評価で実施する．自動評価には，生成された発話の多様性を測る distinct-1 と distinct-2 を用いる．手動評価には，発話単位の関係性（システムの応答が直前の発話とつながりのあるものになっているか），意味・文法性（文法的・意味的なエラーがないか），個性度合（発話が単調ではなく性格・個性を示唆する口調であるか）と，対話全体における意味的一貫性と表現的一貫性，さらに満足度を利用する．手動評価については，それぞれ 5 段間評価の平均を評価値として利用する．

3.2 評価結果

評価結果を表 3 に示す．疑似対話コーパスで構築した対話モデルは，distint-n の値が他のモデルより高いことから，多様性が高いことがわかった．しかし，既存の対話コーパスで構築したモデルと比べて，発話応答の関係性が低く，対話全体における意味的・表現的一貫性がなく，対話全体で矛盾が発生している事例が多いと考えられる．これらの原因としては，連続した台詞の長さと量が不十分であったことが考えられる．特に，意味的・表現的一貫性が低い原因としては，台詞の発話者が特定できず，複数の登場人物の発話が混ざった状態であることや，台詞の連続性を考慮していないことなどが影響していると考えられる．

これらの課題を解決することで，小説テキストから自動で日常会話向けの対話コーパスが構築できるといえる．

表 3　評価に関する結果

	疑似	JDD	JPC
発話応答の関係性	2.44	**3.00**	2.87
意味・文法性	**3.43**	3.28	3.29
個性度合	**3.86**	2.47	3.04
意味的一貫性	1.7	**2.65**	2.35
表現的一貫性	2.85	**4.25**	3.55
満足度	1.8	**2.55**	2.35
distinct-1	**0.44**	**0.44**	0.38
distinct-2	**0.81**	0.77	0.72

4. おわりに

本稿では，小説から抽出した台詞を単につなげた疑似対話コーパスを用いて，小説テキストから対話コーパスを構築できる可能性ついて分析した．定量分析と関連コーパスとの語彙分析，また，対話モデルを用いた分析の結果から，小説テキストから台詞を抽出して，日常会話向けの対話コーパスを構築できる可能性を確認した．その一方で，発話応答の関係性が低く，対話全体における意味的・表現的一貫性も低いことから，疑似コーパスでは，既存のコーパスと比べて質的な問題があることを確認した．

今後は，台詞の発話者特定法や台詞間の連続性の判定法などを検討し，小説テキストから日常会話コーパスを自動構築する手法の実現を目指す．

謝辞

本研究は，香川大学情報メディアセンターの計算資源の支援を受けて実施した．

参考文献

[1] 藤村他，“言語研究の技法　データの収集と分析”，pp.43-72，ひつじ書房，2011.
[2] 別所他，“リアルタイムクラウドソーシングと Twitter 大規模コーパスを利用した対話システム”，情報処理学会研究報告，Vol.2012-NL-206，No.13，pp.1-8，2012.
[3] 赤間他，“日本語日常対話コーパスの構築”，言語処理学会第 29 回年次大会発表論文集，pp.108-112，2023.
[4] 小倉他，“小説対話システム Deep EVE における LSTM を用いたキャラクタ性のある応答文生成”，情報処理学会論文誌，pp967-975
[5] Y. Du 他，“小説からの自由対話コーパスの自動構築”，言語処理学会第 25 回年次大会発表論文集，pp.623-626，2019.
[6] 小説家になろう https://syosetu.com/
[7] 青空文庫 https://www.aozora.gr.jp/
[8] H. Sugiyama, et al., “Empirical Analysis of Training Strategies of Transformer-based Japanese Chit-chat Systems”, aiXiv preprint arXiv:2109.05217, 2021.