E-063

小説テキストを用いた雑談対話コーパスの自動構築手法

An Automatic Construction Method for a Chat-Oriented Dialogue Corpus Using Novel Text

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（LLM）の登場により，対話システムが注目されている．対話システムの構築には，対話コーパスが必要である．近年，日本語の対話コーパスは整備されつつあるが，現状は英語の対話コーパスと比べて少ない．特に雑談対話コーパスの構築には，人同士の会話を録音して書き出す方法[1]やコーパス用に新たに文章を作成する方法[2]，人同士のチャット形式会話を収集する方法[3]などが挙げられる．しかし，人手作業や実験環境の設定などコストが高く，容易に多様性のあるコーパスを構築できない課題がある．この課題を解決するために，本研究では小説の台詞を用いた雑談対話コーパスの自動構築手法を提案する．小説内における登場人物同士の台詞のやり取りを会話とみなすことができ，それらを収集することで大規模な会話コーパスを構築が可能である．また，小説のバリエーションは豊富であり，小説ごとに会話内容も異なるため，既存手法では構築が困難であった多様性のある会話を容易に収集できると考えられる．

本稿では，雑談対話コーパスの構築方針を立て，それに基づいた対話コーパスの構築方法を提案する．また，提案した構築方法によって，小説からどのようなコーパスを構築することが可能であるかを実際に構築したコーパスを用いて定量的に分析し，有用性を考察する．

2. 雑談対話コーパスの構築方針

我々の先行研究[4]において，小説内の記述上連続した台詞を 1 会話として構築した疑似対話コーパスを用いて対話モデルを構築し，モデルの性能を評価した．その結果，疑似コーパスを用いて学習した対話モデルは，キャラクター性が強い特徴がある一方，以下の課題が明らかになった．

1. 1 会話に含まれる発話数が少ないため，会話全体の文脈理解能力に対する学習が不十分である．
2. 小説内に複数の人物が登場することによって，口調の一貫性を保つことが困難である．

1 つ目の課題に対しては，離れた台詞同士を結合して発話数を拡大する．小説には台詞間に状況説明や人物の心情などを説明する地の文が存在する．しかし，地の文を挟む台詞間に発話応答関係が成立すれば，1 つの会話とみなすことができる．そこで，地の文を挟む台詞対に対して発話応答関係を判定し，関係ありと判定された部分を連続した会話として結合することで，1 会話内の発話数を拡大する．

2 つ目の課題に対しては，台詞に発話者情報を付与することで対応する．学習時に発話者を明示または，対話モデルに口調の特徴を学習させることによって，口調の一貫性を担保することができると考えられる．台詞単体では発話者の特定は困難であるが，台詞周辺の地の文や発話者の口調，台詞の内容などから特定が可能といえる．

† 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

3. 雑談対話コーパスの構築手法

構築方針で述べた 2 つの課題に対する解決手法を用いて，小説から雑談対話コーパスを構築する手法を説明する．コーパス構築の手順としては，まず，会話内の発話数を拡大する手法を実行する．その後，発話者を特定する手法を適用することでコーパスを構築する．

3.1 会話内の発話数を拡大する手法

我々の先行研究[4]で提案した発話応答関係を判定する BERT モデルを用いて，前半の台詞と後半の台詞の発話応答関係を判定し，関係があると判定された台詞対を連続している台詞とみなす．判定結果を踏まえて，2 つ以上台詞が連続している台詞群を 1 会話として抽出する．また，判定対象として前後の台詞が離れすぎているものは，発話応答関係が成立しない可能性が高い．そこで，台詞間にある地の文が 7 文以上存在する場合，判定対象外とする．

我々の先行研究[4]でモデルの性能評価を実施した結果，80%程度の F1 値で判定できることを確認した．また，提案手法を用いることで，記述上連続した台詞群と比べて，1 会話に含まれる平均発話数が約 2 倍に増加し，5 発話以上含まれる発話数が 2 倍以上増加したことを確認した．

3.2 台詞の発話者を特定する手法

我々の先行研究[5]で提案した発話者特定モデルを用いて，3.1 で構築した会話内の台詞に対して発話者を特定する．特定する手がかりとして，状況説明文，台詞内の情報，発話者の口調特徴を用いる．まず，ルールベースで発話者を特定し，発話者候補とする．そして，発話者候補を疑似ラベルに用いて人物ごとの口調ベクトルを構築する．その後，口調ベクトルと各台詞の類似度を用いて発話者を特定し，口調ベースの発話者候補とする．最終的にルールベースと口調ベースの発話者候補を統合した結果を発話者とする．

我々の先行研究[5]で性能を評価した結果，70%程度の Precision で話者を特定できることを確認した．

4. 小説コーパスの定量分析

3.で提案した手法を用いて，小説家になろう[6]の恋愛ジャンル 20 小説からコーパスを構築する．そして，構築したコーパス（小説コーパス）について，会話内の発話数と発話者特定の観点から定量的に分析する．

4.1 会話内の発話数に関する分析

1 会話に含まれる発話数の観点で，小説コーパスと既存コーパスを比較，分析する．既存コーパスとしては，規範的な表現で構成されている日本語日常会話コーパス（JDD）

[2]とペルソナ情報が付与されている RealPersonaCorpus（RPC）[3]を用いる．

表 1 に小説コーパスと既存コーパスの平均発話数と最大発話数を示す．なお，小説コーパスの平均発話数は 20 小説の平均値である．表 1 より，平均発話数は既存コーパスと比べて少ないが，最大発話数は最も多い．LLM には，比較的長いテキストを入力することができるため，1 会話に含まれる発話数が多い会話は有用なデータといえる．

図 1 に各小説における発話数ごとの会話数を示す．小説コーパスの平均発話数は既存コーパスより小さいが，図 1 に示すように，各小説から 7〜11 発話以上の会話を一定数収集できるといえる．よって，小説コーパスは，会話の文脈理解に必要な学習データに活用できると考えられる．

表 1: 1 会話に含まれる平均発話数と最大発話数

	平均発話数	最大発話数
小説コーパス	5.55	**101**
JDD	7.94	15
RPC	**30.0**	49

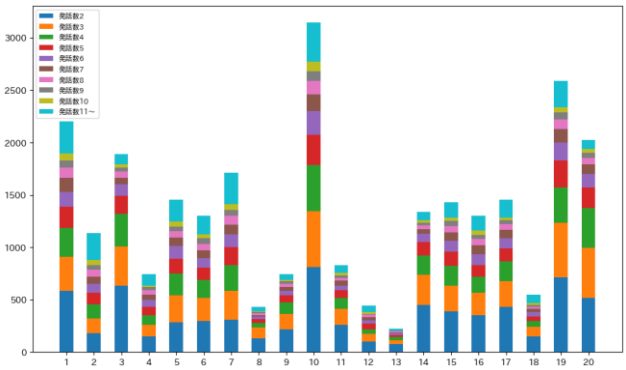

図 1: 各小説の 1 会話に含まれる発話数ごとの会話数

4.2 発話者特定に関する分析

発話者特定により，発話，会話に対してどの程度特定ができたかを分析し，発話者が付与された小説コーパスの特徴と活用方法を考察する．

表 2 に，20 小説全体において人物が特定できた発話数と，会話内の発話がすべて特定できた会話数，1 人物あたりの平均発話数の最大を示す．表 2 から，特定できた発話数が 40%，会話数が 10%であることがわかる．完全に特定できた会話数が少ないことから，少量の小説のみでは口調の一貫性のための対話コーパス構築は困難といえる．しかし，1 人物あたりの平均発話数の最大が 955 発話もあることから，1 小説あたり 1 人物については，Proximal Policy Optimization（PPO）や Direct Preference Optimization（DPO）などの強化学習により，口調の一貫性を持った対話システムが構築できると考えられる．

図 2 に，20 小説全体で発話者が特定できた会話数（一部特定も含む）とその会話内の構成人数の関係を示す．図 2 から，2 人による 2 発話の会話が最も数が多く，発話数が多くなるほど特定することが困難であるといえる．しかし，少量ではあるが 10 発話以上の会話に対しても特定できていることから，発話者情報が付与された発話を多く含む会話が収集可能であるといえる．また，人手では構築が難しい 3 人以上の会話からなるコーパスも小説から構築できるため，複数の小説を用いてコーパスを構築することで，複数人会話に対応する雑談対話システムが実現できるといえる．図 2 の 1 人による会話については，独り言の台詞である場合も考えられるが，発話者特定の誤判定の可能性が高い．発話者特定の Precision が 70%，発話応答関係の F1 値が 80%であるため，構築した小説コーパスにはノイズが存在することに留意する必要がある．今後は，ノイズとなるデータの排除方法が課題となる．

表 2: 発話者特定に関する定量分析の結果

特定発話数(%)	特定会話数(%)	1 人物あたりの平均発話数の最大
60,025 (40.1)	2,772 (10.3)	961.5

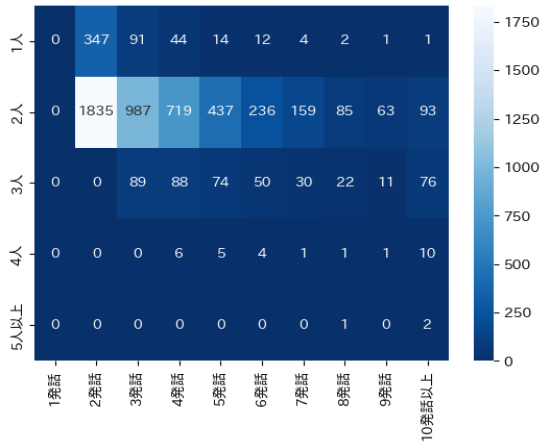

図 2: 特定できた会話の 1 会話の発話数と構成人数

5. おわりに

本稿では，小説の台詞を用いた雑談対話コーパス自動構築手法を提案した．構築したコーパスを分析した結果，各小説から 7〜11 発話以上の会話を一定数収集できることや，複数人の会話からなるコーパスを容易に構築できるなど，既存のコーパスや構築法に対する優位性を確認できた．

今後は，さらなるコーパスの質向上に関する改善手法や構築したコーパスの具体的な活用法について検討する．

参考文献

[1] I. Fujimura, et. al., “Lexical and grammatical features of spoken and written Japanese in contrast: exploring a lexical profiling approach to comparing spoken and written corpora”, Proc of the VIIth GSCP International Conference. Speech and Corpora, pp.393-398 (2013).
[2] 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).
[3] 山下他, “RealPersonaChat: 話者本人のペルソナと性格特性を含んだ雑談対話コーパス”, 言語処理学会第 30 回年次大会発表論文集, pp.2738-2743 (2024).
[4] K. Iwamoto, et al., “A Method for Determining Utterance-Response Relationships Between Japanese Novel Lines for Constructing a Daily Dialogue corpus”, Proc. of 2024 16th IIAI International Congress on Advanced Applied Informatics, pp.391-396 (2024).
[5] 岩本他, “小説における台詞と口調，地の文を活用した台詞の発話者特定手法”, 言語処理学会第 31 回年次大会発表論文集, pp.4166-4171 (2025).
[6] 小説家になろう https://syosetu.com/