小説を利用した日本語日常対話コーパス構築のための台詞間の発話応答関係の判定

岩本和真[1]　安藤一秋[1]
[1]香川大学 創造工学部
{s20t301, ando.kazuaki}@kagawa-u.ac.jp

概要

対話コーパスを構築するには，膨大なコストがかかる．この課題を解決する 1 つの方法として，小説内の台詞を利用した対話コーパスの構築が考えられる．しかし，単純に連続する台詞群を 1 会話として抽出するのみでは，1 会話に含まれる発話数が少なくなる課題がある．そこで本稿では，1 会話に含まれる発話数を増やすために，台詞と台詞の間に状況や感情などを説明する文章が存在するが，両台詞に発話応答関係が成立する場合があることに注目し，このような台詞間の発話応答関係の有無を判定する手法について検討する．

1 はじめに

従来の対話コーパスの構築法は，人間同士の対話を録音して書き起こす[1]，SNS の発信・返信を収集する[2]，人手で規範的な対話文を作成する[3]方法などが挙げられる．これらの構築法では，録音のための実験環境の整備や，文章中にある不適切な表現および個人情報の削除，多数の作業者の確保が必要など，コーパス構築に関わるコストが膨大になるという課題がある．また，日本語の対話コーパスは徐々に整備されつつあるものの，英語の対話コーパスと比較すると十分とはいえない．

これらの課題を解決するため，我々は小説テキスト内の台詞に注目する．小説内には，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の対話ととられることができる．それらを利用して機械的にコーパス構築ができると考える．本研究では，小説内にある登場人物の台詞とその周辺情報を活用し，大規模な日本語日常対話コーパスの自動構築を目的とする．

我々の先行研究[3]では，小説内で連続する台詞を 1 会話とみなして構築した疑似小説対話コーパスと日本語日常対話コーパス[4]，JpersonaChat[5]を用いて対話モデルを構築し，実際に対話してもらった被験者に，意味的一貫性と表現的一貫性，満足度についてアンケート評価を実施した．評価の結果，疑似小説対話コーパスで学習したモデルは，既存の対話コーパスで学習したモデルと比較して，受け答えの整合性が取れていないといった意味的一貫性が乏しく，また，口調が統一されていないといった表現的一貫性にも欠けるという結果となった．前者の原因として，既存コーパスの 1 会話に含まれる平均発話数はそれぞれ 7.94，12.35 発話であるのに対し，疑似小説対話コーパスでは 2.94 発話とかなり少ないため，学習時に発話応答関係を正しく学習できていない可能性が考えられる．

小説内には，台詞と台詞の間に状況や感情などを説明する文（以降，状況説明文）が存在しているため，物理的には連続していないが，発話応答の関係が成立しているものが存在している．これらの台詞を 1 つの会話として扱うことにより，1 会話に含まれる発話数を増やすことが可能である．しかし，発話応答が成り立たない場合や鍵括弧で囲まれているが台詞ではない場合もあるため，単に括弧を手掛かりに台詞をすべて繋げるだけでは，コーパスの品質低下を招く．したがって，状況説明文を挟む台詞間の発話応答関係を判定した上で 1 会話にまとめる必要がある．そこで本稿では，対象とする台詞と直前の台詞の間に発話応答関係があるか否かを判定する手法について検討する．

2 関連研究

小説内の台詞を用いた対話コーパスの構築については，いくつかの研究で提案されている．小倉ら[6]は，キャラクタ性を考慮した対話システムの構築を目的に，小説の台詞を利用している．Yulong ら[7]は，小説テキストに対して，パターンマッチングで発話者を特定し，小説内の台詞に話者情報を付与した対話コーパスの構築法を提案している．Yulong らは，提案手法を用いて，4,838 の小説から 19,492 対

話が含まれる対話コーパスを構築している．1 対話は平均 10 発話程度で構成されている．

物語の境界線を判断する研究も実施されている．小林[8]は，物語から場・時・人に関係のある語句を抽出し，抽出した数に対してペナルティを与え，ペナルティが閾値を超えた箇所をシーン境界とする手法を提案している．

図 1 に示すように，小説内には状況説明文を挟んでいるが発話応答の関係が成立する台詞が存在する．本稿では，図 1 の判定対象の台詞と直前の台詞に発話応答関係が存在するか否かを判定する手法について検討する．

「なあ、聞いてもいいか」　**※直前の台詞群**
「なんですか」　**※直前の台詞（＝直前の台詞群）**
必要なものをセッティングした真昼がこちらに顔を向ける。　**※台詞間の文章群**
「なんで雨の中ブランコ漕いでたんだ。彼氏と揉めたりとかか」　**※判定対象の台詞**
（佐伯さん，『お隣の天使様にいつの間にか駄目人間にされていた件』，小説家になろう，から引用）

図 1　文章を挟むが発話応答関係がある台詞の例[i]

3　データセットの構築

判定手法の学習・評価で利用するデータセットの構築法について述べる．

3.1　データセットの概要

本研究では，「」と『』で囲まれている文章を登場人物の台詞とする．データセットにおける各データは，図 1 に示す「直前の台詞群」，「台詞間の文章群」，「判定対象の台詞」および発話応答関係ラベルで構成する．また，直前の台詞群のうち，最後の台詞を直前の台詞とする．

データセットを構築するための情報源には，「小説家になろう」[9]を利用する．本稿では，日常対話の台詞が多く含まれると考えられる「恋愛」と「推理」ジャンルから 10 小説を収集し，1 小説ごとに任意の 500 件のデータ（計 5,000 件）を人手でラベリングしてデータセットを構築する．

発話応答の関係を示すラベルの定義を表 1 に示す．ラベル 1～3 が正例で, ラベル 4 と 5 は負例を意味する．なお，ラベル 5 の「状況説明文から得られる情報がなければ会話として成り立たない」は，状況説明文に書かれている出来事に対する応答が台詞になっている場合を意味している．例えば，状況説明文“A 氏が黙り込んだ”に対して，「黙りこんでも無駄だよ」のような台詞が続く場合である．また，直前の台詞（群）と判定対象の台詞において，データの範囲内では正例または負例の判断が難しいものはニュートラルとするために，ラベル 6 を設ける．

著者一人でラベリングした結果，ラベルごとのデータ数は表 2 に示すように，正例ラベルが計 2,426 件，負例ラベルは計 2,115 件となった．小説テキストから任意抽出したデータの正負ラベル数がほぼ均等になったことから，物理的には連続していない台詞をまとめることで，1 会話に含まれる発話数を増やせるといえる．

3.2　ラベルの一致度調査

発話応答関係ラベルの妥当性を検証するため，複数人の被験者に一部のデータをラベリングしてもらい，著者が付与したラベルと被験者が付与したラベルの一致度を測定する．対象データは，著者がラベリングしたデータから，ラベルごとに 5 件のデータをランダムに抽出して準備する．被験者 4 名（大学生 2 名，大学院生 2 名）には，各ラベルの定義を説明した上で, 計 25 個のデータに対してラベリングしてもらう．

表 3 には，筆者が付与したラベルと 4 名の被験者が付与したラベル全体の一致度および正例・負例ラベル別における一致度を示す．また，表 4 にはラベルごとの一致割合を示す．表 3 に示す正例・負例の一致度を見ると，被験者の一致度の平均が 89%であることから，正例・負例のラベル分類については妥当性があるといえる．また，表 4 に示すラベルごとの一致度から，ラベル 5 以外の一致度は 80%以上あり，ラベル定義に妥当性があるといえる．なお，ラベル 5 については，一致度が 70%と最も低い結果となった．原因分析の結果，評価データに同一人物の台詞であるが会話としてつながっていないものが存在しており，ラベル 2 とラベル 5 で判断が割れている事例が確認できた．よって，同一人物における台詞に対する判断は人間でも難しいといえる．また，ラベル 5 において「発話応答関係が成立しない」の基準が人によって揺れていると考えられる．

[i] https://ncode.syosetu.com/n8440fe/　から引用

表 1　発話応答関係ラベルの定義

	定義	定義内容
ラベル 1	発話応答関係あり	直前の台詞（群）と判定対象の台詞で発話応答関係が成立し，会話としてつながると判断できる.
ラベル 2	同一人物による会話継続	直前の台詞（群）と判定対象の台詞に発話応答関係はないが，発話者が同一で台詞としてまとめることができる.
ラベル 3	自然な話題遷移	直前の台詞（群）と判定対象の台詞で話題は変化するが，判定対象の台詞が「ところで」などで始まり，自然な話題遷移である.
ラベル 4	台詞以外の用途	言葉の強調などに使われる括弧であり，判定対象の台詞が人物の発話ではない（台詞ではない）.
ラベル 5	発話応答の不成立	直前の台詞群と判定対象の台詞において発話応答の関係が成立しない，また，状況説明文から得られる情報がなければ，発話応答関係が成立しない.
ラベル 6	ニュートラル	直前の台詞群と判定対象の台詞の発話応答が成立するともしないともいえない.

表 2：ラベルごとのデータ数

1	2	3	4	5	6
1,789	491	146	177	1,938	459

表 3：ラベル全体の一致度

	ラベルの一致度	正例・負例の一致度
被験者 1	0.96	0.96
被験者 2	0.88	0.88
被験者 3	0.60	0.80
被験者 4	0.88	0.92
平均	0.83	0.89

表 4：ラベルごとの一致度

	1	2	3	4	5
被験者 1	1.00	1.00	1.00	1.00	0.80
被験者 2	0.80	1.00	0.80	1.00	0.80
被験者 3	0.60	0.60	0.60	0.80	0.40
被験者 4	0.80	1.00	0.80	1.00	0.80
平均	0.80	0.90	0.80	0.95	0.70

4 評価実験

構築したデータセットを用いて，台詞の発話応答関係を判定する複数のモデルを構築し，判定性能を比較する．本実験では，台詞間の文章の文数を閾値とするルールベースの判定法と，言語モデルを用いた判定法の 2 つを用いる．言語モデルを用いた判定法においては，学習に用いる入力データの範囲が与える影響を確認するため，範囲の異なる 3 つのモデルを構築する．また，ニュートラルであるラベル 6 を正例とした場合と負例とした場合のモデルを構築し，ニュートラルの取り扱い方法も検討する．

4.1 実験設定

ルールベースの判定法では，事前調査に基づき，台詞間の状況説明文が 2 文以下の場合を正例，それ以外を負例と判定する．言語モデルを用いた判定法は，東北大学の BERT[ii]を台詞間に発話応答関係があるか否かの二値分類タスクとして学習して構築する．また，範囲 1：直前の台詞と対象の台詞のみを学習に利用するモデル，範囲 2：直前の台詞群と対象の台詞を用いるモデル，範囲 3：直前の台詞群・台詞間の文章群・対象の台詞全てを用いるモデルを構築する．なお，これら 3 モデルの学習において，ラベル 6 は負例とする．

10 小説を小説単位に分割した 10 分割交差検証を用いて，判定性能を評価する．コーパスの質を担保するためには負例の混入を軽減すべきであると考え，負例の Precision，Recall，F 値を評価指標とする．各検証においてテスト loss が最小になった epoch 時の評価値の平均を各モデルの評価値とする．

4.2 評価結果

ルールベース手法と学習に用いる入力データの範囲を変更した 3 モデルの評価結果を表 5 に示す．表 5 の Detailed_Recall はラベル 6 を除いた負例の再現率である．表 5 から，ルールベース手法の F1 が 64.5%で，Detailed_Recall が 66.4%と最も低い結果となっ

[ii] cl-tohoku/bert-base-japanese-v3 を使用

た．よって，台詞間の発話応答関係は文数のみでは判定が困難であるといえる．言語モデルを用いた判定結果を見ると，範囲 3（直前の台詞群・台詞間の文章群・対象の台詞全て）で学習したモデルがすべての評価指標において最高値を得た．このことから状況説明文が発話応答関係の判定に活用できるといえる．また，範囲 3 におけるラベルごとの Accuracy を確認したところ，ラベル 4 が 86%と最も高く，鍵括弧で囲まれているが台詞でないものをほとんど排除できることを確認した．

次に，範囲 3 で学習したモデルにおいて，ニュートラルであるラベル 6 を正例に加えた場合の判定性能を表 6 に示す．表 6 から，Accuracy（Acc）は正例に加えた方が高くなるが，Recall，F1 は負例のままの方が高い．正例を負例と判定するより，負例を正例と誤判定する方が対話コーパスの質に与える影響が高いといえる．よって，ラベル 6 は負例として学習する方が妥当であるといえる．

台詞間の文章数と正例，負例それぞれの Recall の関係について分析した結果，負例の Recall との相関係数が 0.95 と正の相関があること，また，正例の再現率との相関係数が-0.89 と負の相関があることを確認した．この結果から，範囲 3 は台詞間の文章量を考慮せずに学習しているため，情報量にばらつきがあることが影響していると考えらえる．この問題の解決策として，台詞間の文章において，会話に直接関係のある部分のみを抽出して利用することで，より文脈を考慮した判定ができると考えられる．

表 5：学習に用いるデータの範囲を変更した結果

	Acc	Pre	Recall	F1	Detailed Recall
RULE	0.639	0.651	0.640	0.645	0.664
範囲 1	0.728	0.743	0.702	0.719	0.726
範囲 2	0.750	0.771	0.724	0.744	0.755
範囲 3	**0.791**	**0.806**	**0.770**	**0.787**	**0.821**

表 6：ラベル 6 を正例/負例に加えた場合の結果

	Acc	Precision	Recall	F1
正例	**0.812**	0.798	0.742	0.766
負例	0.791	**0.806**	**0.770**	**0.787**

5 発話応答関係判定の効果検証

提案手法の効果を検証するため，物理的に連続する台詞のみを抽出した場合の 1 会話に 5 発話以上含まれている会話数と，提案手法を適用した後の会話数の変化を確認する．なお，提案手法には，ラベル 6 を負例としたデータおよび範囲 3 で学習したモデルを用いる．また，実験対象には，恋愛ジャンルから新規に収集した 12 小説を含む 20 小説を用いる．

20 小説に対する提案手法の適用前後の会話数の変化を図 2 に示す．提案手法を適用することで，すべての小説において 5 発話以上を含む会話数が増加しており，さらに 14 小説においては会話数が 2 倍以上に増えていることがわかる．1 会話に含まれる平均発話数は，提案手法の適用前は 2.09 発話しかなかったが，適用後は 10.02 発話にまで増加した．

その一方で，複数人による対話が含まれる会話数も増加した．通常の対話コーパスは，1 対 1 の対話で構成されているため，1 対 1 の対話で構成される会話を判定する必要がある．

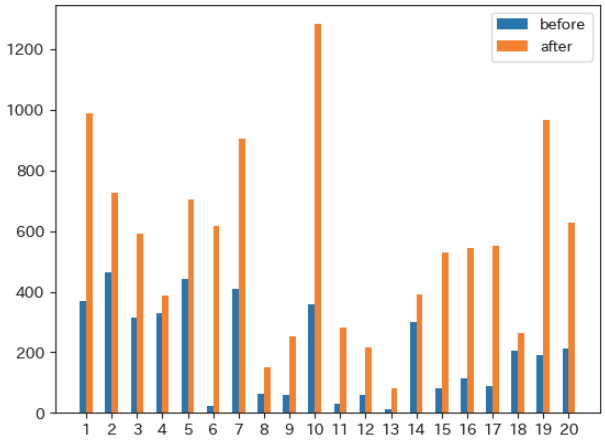

図 2：5 発話以上を含む会話数の変化

6 おわりに

本稿では，小説から対話コーパスを構築する際，単純に連続する台詞群を 1 会話として抽出するのみでは，1 会話に含まれる発話数が少なくなる課題を解決するため，台詞間の発話応答関係の有無を判定する手法について検討した．評価の結果，直前の台詞群・台詞間の文章群・対象の台詞全てを用いて学習した手法が最良となり，F1 値で 78%となることを確認した．また，提案手法を適用することで，70%の小説において，1 会話に 5 発話以上含まれる会話を 2 倍以上に増やすことができた．

今後の課題として，データセットの質の向上と発話応答関係の判定性能の向上を目指す．また，複数人による対話に対応した対話コーパスの構築法についても検討する．

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. 岩本他, “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”, 情報科学技術フォーラム講演論文集, pp.285-286, 2023.
4. 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.
5. H. Sugiyama, et al., “Empirical Analysis of Training Strategies of Transformer-based Japanese Chit-chat Systems”, aiXiv preprint arXiv:2109.05217, 2021.
6. 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, vol.60, No.3, pp967-975, 2019.
7. 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.
8. 小林, “場・時・人に着目した物語のシーン分割手法”, 情報処理学会研究報告, 自然言語処理研究会報告, 2007-NL-179, pp.25-30, 2007.
9. 小説家になろう https://syosetu.com/