人工知能学会 インタラクティブ
情報アクセスと可視化マイニング研究会(第29回)
SIG-AM-29-01

ラベルなし対話スクリプトとNext Sentence Predictionを用いたテキストグループチャットにおける返信関係の判別

Identification of Reply-to Relation in Textual Group Chat via Unlabeled Dialogue Scripts and Next Sentence Prediction

SHAN Junjie[1*] 西原 陽子[2] 韓 毅弘[3]
Junjie Shan[1] Yoko Nishihara[2] Yihong Han[3]

[1] 立命館グローバル・イノベーション研究機構
[1] Ritsumeikan Global Innovation Research Organization, Ritsumeikan University
[2] 立命館大学情報理工学部
[2] College of Information Science and Engineering, Ritsumeikan University
[3] 立命館大学大学院情報理工学研究科
[3] Graduate School of Information Science and Engineering, Ritsumeikan University

Abstract: インスタントメッセージ（IM）が他者とのコミュニケーションにおいて重要な役割を果たすようになったことにより，チャットメッセージを分析することで人々のコミュニケーションを支援することを目的とした研究が増え始めている．チャット内では，メッセージ間の関係（返信関係など）が明示されていない．グループチャットにおけるメッセージ間の関係性を明らかにすることが重要となる．本稿では，ラベル付けされていないテキストデータから，グループチャットにおける「返信関係」（reply-to）を，Next Sentence Prediction (NSP) の手法を用いて判別する手法を提案する．まずは，学習データの準備として，ラベル付けされていない対話スクリプトから,「reply-to」関係を持つメッセージと持たないメッセージを自動的にサンプリングする方法を提案する．次に，準備した学習データを用いて，二つのチャットメッセージ間の「reply-to」関係を判別するための NSP モデルを三つのセッティングとパラメータで構築する．構築する NSP モデルは，日本語の事前訓練 BERT モデルに基づいている．最後に，学習したモデルを手作業でラベルを付けた実際のテキストグループチャットのメッセージにより評価した．構築した三つのモデル評価した結果，検証セットで最大 88.82%，テストセットで 69.64%の精度を得られた．

1 はじめに

在宅ワークやリモート学習の普及が進行している現在，オンラインコミュニケーションツールを用いたコミュニケーションが活発に行われるようになっており，実況配信サイト，テレビ会議システム，音声会話ルームなど，様々なものが提供されている．その中，インスタントメッセージ（IM）ソフトウェアは，オンライン時代のコミュニケーションにとって不可欠な存在となっている．人々の情報交換が IM に依存するようになり，IM 上での人々のコミュニケーション活動を支援することを目的とした研究も多く見られるようになった．例えば，コミュニケーションメッセージを通して人々の関係を分析し，良好な関係の維持に役立てる研究 [1] や，IM アプリケーション上の雑談の話題をサポートする研究 [2] などがある．

これらの研究はすべて，詳細的な解析や適用をする前に，各チャットメッセージ間の正しい対応を判別する必要がある．特に「返信関係」の判別が重要となる．チャットソフトウェアのテキストメッセージは，文章や新聞記事などのように，順序付けられて構成されていない場合が多い．チャットメッセージはほとんどの場合において，自由な順序で発言されている．図 1 に示すように，最後のメッセージ「私もです」は最初のメッセージ「土曜日のイベント」に対して返信しており，隣接する 2 番目のメッセージに対する返信ではないことが明らかである．

従来の電子メールなどのオンラインコミュニケーション手段に比べ，テキストのグループチャットには以下のような特徴がある：

*連絡先：立命館グローバル・イノベーション研究機構
〒 525-8577 滋賀県草津市野路東 1 − 1 − 1
E-mail: shan@fc.ritsumei.ac.jp

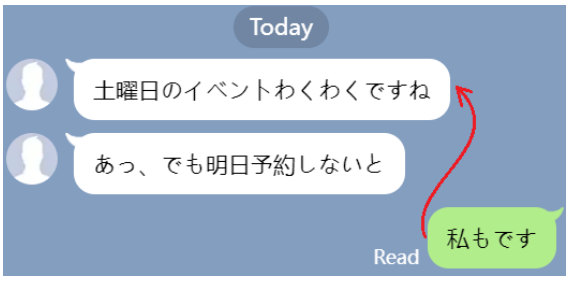

図 1: 自由な順序のチャットメッセージ. 3 番目のメッセージは，隣接の 2 番目メッセージではなく，1 番目のメッセージに対して返信している

- 短文内容で複数回送信
  IM のチャットメッセージは，返信速度を上げるために各メッセージの内容は短く書かれることが多い．結果として，IM チャット，特にグループチャットでは，連続しないあるいは順序性のないメッセージが発生しやすくなる．
- ランダム化と断片化
  グループチャットの中で，人々は各自のトピックや自分のペースで発言し，チャットのメッセージは不規則な順序や断片的なパターンで流れていく．

これらの課題に取り組むために，グループチャットのメッセージと類似したテキスト上の特徴を持ちながら，順序立てて構成された「対話スクリプト」を学習データとして利用する．

本稿では，BERT モデルの NSP（Next Sentence Prediction）手法を活用し，グループチャットのテキストメッセージ間の複数の「返信関係」の判別手法を提案する．提案手法では，主に以下の三つの処理を行う：

1. ラベルなしの対話スクリプトのテキストデータから，「返信関係」を持つメッセージペアと持たないメッセージペアを自動的に取得する方法を提案する．
2. 「返信関係」の識別の効果を検証するために，三つの構造で NSP モデルを構築し，訓練とテストを行う．
3. 実際のグループチャットからテキストメッセージの履歴を三つ収集し，各メッセージ間の「返信関係」を人手でラベル付けて提案手法を評価した．

2 関連研究

2018 年に Google が BERT（bidirectional encoder representation from transformers）モデルを提案して以来 [3]，それに関する研究，またそのサブ機能や構成パーツに関する多くの研究が行われてる．Next Sentence Prediction（NSP）に関する研究もその中の一つである．本来の NSP タスクは，BERT モデルの学習を強化するために提案されたものである．NSP タスクは，2 文のペアを受け入れ，そのペアの 2 番目の文が 1 番目の文に対して論理的または意味的に続くかどうかを予測することを学習する．NSP タスクにより，事前訓練した BERT モデルが文ペアの直接入力を受け取る仕組みが提供されたため，自然言語処理の分野で注目されている．2019 年，Shi らは NSP を暗黙の談話関係分類（implicit discourse relation classification）のタスクに活用し [4]，Liu らは NSP を用いて長文の自動要約を支援する手法を提案した [5]．2021 年に Feng らは，NSP を利用して子供の言語教育における読解問題の難易度を調整することを試みた [6]．Li らは 2022 年にマルチモーダルなエンティティ・リンキングに NSP を利用した [7]．これらの研究は，NSP が文ペアの入力に対する処理能力を持つことを示している．

一方で，テキストチャットの分析に関する研究も幅広く行われている．例えば，チャットメッセージ中の代名詞の共参照解析などを含んでいる「chat disentanglement」はその中の代表的な一つである．このタスクでは，大量の順序が揃っていないチャットメッセージの処理や分析をするために，多数の人手でアノテーションされた特徴を使用する必要がある [8–11]．2017 年，Mehri らは人手でアノテーションされた特徴のあるデータセットを使用して，オンラインのチャットメッセージにおける返信関係の判別を試した [12]．Mehri らは，オンラインのグループチャットから 524 件のメッセージに対して手動で返信関係を表す特徴をアノテーションした．そして，ランダムフォレスト分類器とリカレントニューラルネットワークの 2 つの方法を用いて，返信関係の判別を行った．しかし，このような提案手法は，多くの事前アノテーション作業が必要であるため，チャットメッセージ記録のデータセットが少なくなり，チャットメッセージの内容もより多様なトピックや領域を含みにくくなる．

2019 年に Guo らが最初に人手特徴量を使用しないプレーンテキストのグループチャットメッセージから「返信関係」を判別する手法を発表した [13]．彼らは，グループチャットにおけるメッセージ間の返信関係を入力されたプレーンテキスト情報のみから判別するため，LSTM（Long Short-Term Memory networks）に基づく三つのモデルの提案や検証を行った．その中の一つにエンドツーエンドモデルも含まれる．Guo らが

表 1: 複数の「返信先」のある例.
（新着メッセージ D は，過去メッセージの A，B，C の全部に対する返信である）

過去メッセージ:
A: 次の旅行について検討しよう.
B: 美味しい食べ物のある所がいい
C: すてきな景色が見たい
新着メッセージ:
D: じゃ，○○○はどうでしょうか？

提案したモデルは，新着のメッセージの「返信先」となる可能性が最も高い一つの過去メッセージを出力する．或いは，新着メッセージが新しいトピックの始まりとして，すべての過去メッセージに返信していない場合には，新着メッセージが自身の「返信先」として判別し，新着メッセージ自身を出力する．Guo らは実際の中国語のテキストグループチャットのデータを用いて提案モデルを検証し，最大 60%に近い精度を得ることができた．しかし，Guo らのモデルは各新着メッセージに対して,「返信関係」のある過去メッセージは一件しか出力できないため，複数の返信先がある場合に対応が難しい．表 1 が示すように，D からの新着メッセージ 「じゃ，○○○はどうでしょうか？」 の 返信先対象は，過去の 3 つのメッセージ全てになると考えられる．この場合，一つのメッセージに対して複数の「返信先」を持つと考えられる，これはグループチャットにおいてよく見られることである．また，Guo らの研究では，同一のデータセットを用いてモデルの訓練と評価を行なっている．検証用のデータセットを新たに用意し，モデルの汎化性能を評価するべきと考えられる．

そこで，NSP を用いて二文のメッセージペアに「返信関係」を持つかどうかを判別する手法を提案する．また，モデルが特定のデータセットに限定されることを避けるため，ラベルなしのデータから自動的にチャットメッセージペアを収集する手法を提案し，多様なトピックのメッセージデータを取得できるようにする．

3 提案手法

3.1 タスク定義

本稿の課題は，テキストのグループチャットにおいて，二つのチャットメッセージが内容的に「返信関係」を持っているかどうかを判別することを目的とする．具体的には，A と B の二つのチャットメッセージが与えられて，メッセージ B がメッセージ A の内容に対する返信であるかどうかを判別する．内容的にメッセージ B がメッセージ A に返信する確率は，NSP モデルによって式 (1) により計算される：

$$P(reply(B|A)) = Softmax(NSP(A, B)). \quad (1)$$

3.2 対話スクリプトからの「返信関係」メッセージの自動サンプリング

前述したように，一般的なチャットメッセージは，順序と内容のロジックに一貫性がないため，過去の研究で「返信関係」の判別のためのデータセットを準備する時に，各チャットメッセージに対して人手で返信関係を付けることが必要である．本研究では，チャットのテキストメッセージと類似する特徴を持ちながら，かつ順序立てて構成された一つテキストのデータとして映画，ドラマ，舞台劇などの「台本・対話スクリプト」を「返信関係」の学習データとして利用することを提案する．対話スクリプトには以下のような特徴があることを分析・考察した：

- 簡潔で短い.
  文章や新聞記事などに書かれる内容と違って，役者の演出や観客への理解を容易にするため，対話スクリプトの台詞は通常に簡明である必要がある．そのため，使用する言葉は十分に考えられ，台詞のテキストは簡潔になっている．そして，対話スクリプトは考えられて作られた文であるため，普通に考えながら喋る話し言葉とも違って，より短くてわかりやすい文面ではっきりと情報を伝えている．これはチャットのテキストメッセージと類似する特徴である．

- 順序的かつ連続的.
  チャットメッセージと違い，対話スクリプトは，内容に応じてロジカルな順序で構成する必要がある．そのため，一つの対話シーンにおいて，直後の発話が直前の発話内容に対する「返信」である可能性が高くなる．

以上の理由から,「返信関係」を持つメッセージペアを取得するためのソースデータとして，対話スクリプトを採用する．ポジティブ（「返信関係」あり）とネガティブ（「返信関係」なし）のメッセージペアをサンプリングする時に実行する具体的な方法は以下に示す：

1. 対話スクリプトに含むシーン転換のマークを利用して各対話シーンを分割する.
2. 同じ対話シーンで隣接する二つの対話（メッセージ）はすべて「返信関係」を持つと仮定し，それらを「ポジティブ」のペアとして保存する.

表 2: 収集したデータセットの統計データ.

項目	数値
対話スクリプト件数	5
対話シーンの数	49
総発言数	3,043
一発話の平均長さ（文字）	21.24
ポジティブペアの数	1,698
ネガティブペアの数	3,396

3. 異なる対話シーン或いは異なる対話スクリプトにある二つの発話（メッセージ）は,「返信関係」を持たないと仮定した．そこで，異なる台本スクリプトからランダムに二つの発話台詞を選び,「ネガティブ」のペアとして作成する．

ポジティブペアの信頼性の確保とノイズの削減のために，本研究では，話し手が二人しかいない対話スクリプトのみをサンプリングの対象とした．この場合，話し手が変わるときには，概ねに前の話し手の台詞に対する「返信」が出てくる．

本稿では，インターネットから日本語の対話スクリプトを五件収集した，合計 49 の対話シーンがあり，発話の総数は 3,043 である．収集したデータセットの統計情報を表 2 に示す．合計 1,698 組のポジティブの対話ペアが得られた（一部のノイズの入った対話データ（例えば，沈黙を表す「……。」や表情を説明する「(笑)」など）は除去された）．各ポジティブペアに対して二つのネガティブペアを作成し，ネガティブペアの数は 3,396 である．

3.3 NSP モデルの構築

本研究では，三つの NSP モデルを構築し（Orig-NSP, NSP-FA-1L，NSP-1L）,「返信関係」判別の効果を検証する．

3.3.1 Orig-NSP: オリジナルの NSP モデル

はじめに，一般的な BERT の構造からオリジナルの NSP モデルを作成し，検証を行った．図 2 に示すように，事前学習した BERT モデルに [A, B] のメッセージペアを入力し，オリジナルの NSP プーリングを使って [A, B] の「返信関係」を判別する．

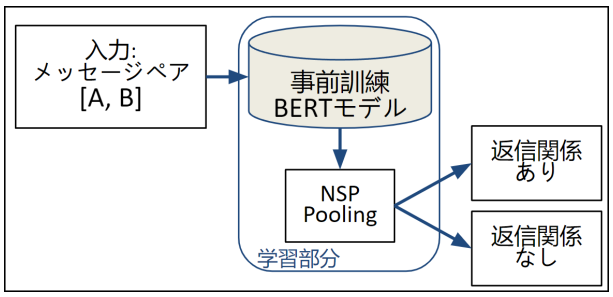

図 2: Orig-NSP: オリジナルの NSP モデル

3.3.2 NSP-FA-1L: NSP モデルでメッセージペアの分散表現を取得

オリジナルの NSP プーリングの後に 128 ユニットを持つ隠れ層を追加し，隠れ層の出力を使って入力メッセージペアの「返信関係」を判別する．このセッティングでは，図 3 に示すように，オリジナルの NSP プーリングと事前学習の BERT モデルを凍結し，追加した隠れ層のみを訓練する．つまり，オリジナルの NSP プーリングからの出力は，入力メッセージペア [A,B] の分散表現（embedding）を取得することのみに使用される．この仕組みは既存研究を参考にし導入する [14].

3.3.3 NSP-1L: オリジナルの NSP モデル＋一つ隠れ層

「NSP-FA-1L」の構造と同様に，オリジナルの NSP 出力の後に 128 ユニットの隠れ層を追加し，オリジナルの NSP 層と事前学習した BERT モデルも凍結しない．図 4 に示すように，この設定では，追加の隠れ層をオリジナルの NSP と BERT モデルとともに，作成したメッセージペアの訓練データを用いて訓練させる．

訓練のセッティングを表 3 に，訓練の結果を表 4 に示す．訓練データ数は 5,094 件（ポジティブ 1,698 件＋

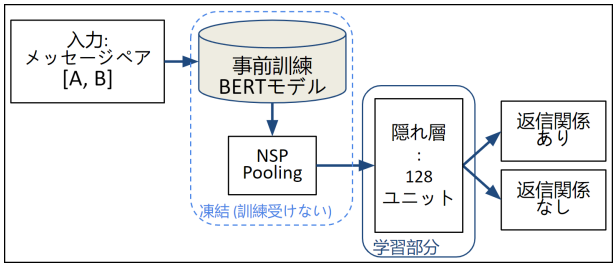

図 3: NSP-FA-1L: NSP モデルでメッセージペアの分散表現を取得.

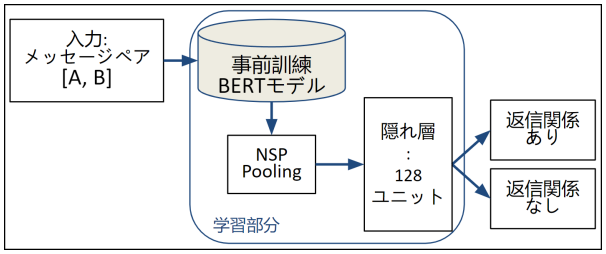

図 4: NSP-1L: オリジナルの NSP モデル＋一つ隠れ層.

表 3: 訓練のセッティング.

項目	数値
訓練メッセージペア数	5,094
形態素分割レベル	文字
最大入力長さ	128
Batch サイズ	64
エポック	10
検証セット割合	0.1
学習率	5e-5
最適化アルゴリズム	Adam

ネガティブ 3,396 件）のメッセージペアである．入力するメッセージペアは全部文字レベルに区切り，最大長さは 128 文字（A と B の合計長さ）である．検証セットのサイズは 10 %である．三つの NSP モデルは，東北大学により公開された事前学習済みの日本語 BERT モデル [1] で初期化した．10 エポックの訓練を行った後，検証セットにおける Orig-NSP，NSP-FA-1L，NSP-1L の精度はそれぞれ 87.06%，78.24%，88.82%であった．

4 評価実験

4.1 評価実験手順

実際的なグループチャットの記録を用いて，学習したモデルの評価実験を行った．評価実験の手順は以下の通りであった：

1. 三つのグループチャットからチャットメッセージの記録を収集し，各メッセージ間の「返信関係」を人手でアノテーションした．三つのチャットは全て 30 分程度の長さであった．

2. チャット記録中の各メッセージとその前の N 個のメッセージを二つずつペアに組んで，訓練済み NSP モデルに入力する．NSP モデルにより各メッセージペアの「返信関係」を予測する．例えば，チャットメッセージの系列が $Seq = [m_{i-N}, ..., m_{i-2}, m_{i-1}, m_i]$ である場合，与えられたメッセージ m_i に対して，各 $[m_{i-n}, m_i](n \in [1, N])$ を一つのメッセージペアとして組み合わせる．

3. 人手でアノテーションされたメッセージペアのラベルと NSP モデルの予測結果から，精度を算出する．各メッセージペア $[m_{i-n}, m_i]$ が「返信関係」を持つかどうかについて，事前にチャットの文脈内容によりアノテーションを行った．(m_i が m_{i-n} への返信である場合は 1 を，ではない場合は 0 にする．)

[1] https://github.com/cl-tohoku/bert-japanese/tree/v2.0

表 4: 訓練結果.

モデル	訓練セットでの損失	訓練セットでの精度	検証セットでの損失	検証セットでの精度
Orig-NSP	0.0389	95.56%	0.6259	87.06%
NSP-FA-1L	0.4974	76.20%	0.4594	78.24%
NSP-1L	0.0807	97.25%	0.3888	**88.82%**

表 5 は，収集された評価実験用のグループチャット記録の統計データである．三つのグループチャットから収集したメッセージの総数は 444 件，各グループのチャットメンバーは 3 名であった．本研究では，手順 2. の N を 5 に設定し，2,052 件のメッセージペアの評価データを取得した（複数「返信先」のペアは 388 件）．N は，あるメッセージに対して「返信」可能なメッセージの範囲を表す．N の大きさは，チャットグループのメンバー数と正の相関があると考えられる．チャットグループのメンバーが増えれば増えるほど，二つ「返信関係」のあるメッセージ間の「距離」が長くなる可能性が高くなる．今回収集したグループチャットのメッセージでは，各チャットグループのメンバーが 3 名であり，すべてのメッセージの「返信先」が過去 5 通以内であることが確認された．故に，N を 5 に設定して訓練済み NSP モデルの評価を行った．

評価結果を表 6 に示す．また，対話スクリプトのデータで訓練を行わないオリジナルの事前学習 BERT モデルの効果も評価した．その結果を「No-Training」として記載する．評価実験の結果により，「NSP-FA-1L」のモデル構築が最も高い精度の 69.64%が得られた．

5 考察

評価実験の結果により，対話スクリプトのデータセットで訓練した三つの NSP モデルは全て，訓練を実行していないオリジナルの事前学習 BERT モデルより精度

表 5: 評価実験のグループチャットの統計データ.

グループ	メッセージ数	メンバー数
グループ 01	237	3
グループ 02	115	3
グループ 03	92	3
合計	444	9

表 6: 評価結果.

モデル	精度
No-Training	49.8%
Orig-NSP	56.63%
NSP-FA-1L	**69.64%**
NSP-1L	62.77%

が高いことが分かった．この結果から，対話スクリプトを利用してグループチャットメッセージの「返信関係」を判別する提案手法が有効であることが示された．

提案した三つのモデルは全て，評価データでの精度が検証セットでの精度より低くなった．これは，主に訓練データが不足しているためと推測される．提案手法の試行として，本研究ではより小規模なデータセットのみを収集してモデル訓練を行った．学習データには，1,698 組のポジティブと 3,396 組のネガティブのメッセージペアしか含まれておらず，BERT のような巨大なモデルを十分に学習させるのは困難なものと考えられる．今後も，より多くの対話スクリプトを収集し，訓練データの量を増やすことを目指している．

モデル NSP-FA-1L は，検証セットでの精度は一番低かったが（78.24%），評価データでは最も高い精度を得た（69.64%）．これは，モデル NSP-FA-1L の効果を反映したものではなく，訓練セットのデータ分布が原因であると考えられる．訓練セットにはネガティブのペアがポジティブのペアの 2 倍であるので，隠れ層 1 つだけの訓練では過剰適合になりやすくと考えられる（NSP-FA-1L モデルは元の BERT モデルと NSP の部分を全部凍結している）．そのため，NSP-FA-1L モデルは他の二つのモデルに比べて，より多くなネガティブの判別結果が出やすくなる．

NSP-1L モデルの精度は，検証セットと評価データの両方で Orig-NSP モデルの精度より上回った．このことから，元の NSP プーリングの後により小さな隠れ層を追加する手法が，モデル訓練の補助に有効であることが示された．BERT モデルにあるオリジナルの 768 個の隠れユニットを持つ NSP プーリング（図 2，3，4 にある「NSP Pooling」の部分）のノイズを軽減し，モデルがより特定のタスクに集中できるようになったと考えられる．

表 7: 正解・不正解の判別結果についての分析.

	一番目メッセージ（A）の平均長さ	二番目メッセージ（B）の平均長さ	入力メッセージペアの平均長さ
正解結果	12.72	**14.28**	27.0
不正解結果	11.05	**9.89**	20.95

検証セットでの精度が最も高い NSP-1L モデルを選択し，それが「返信関係」を正しく判別した/判別しなかったメッセージペアに対して分析した．表 7 には，正解と不正解の結果における入力ペアの一番目メッセージ（A）と二番目メッセージ（B）の平均長さ（単位：文字）をそれぞれ示している．誤って判別されたペアの二番目メッセージ（B）の平均長さは，正しく判別されたペアの二番目メッセージより短くなることが確認された（9.89 文字と 14.28 文字）．これは，今回定義したタスクが，独立した二つのメッセージだけを入力として「返信関係」を判別していて，文脈情報を含んでいないためと考えた．そのため,「はい」,「オッケー」,「もちろん」,「賛成」，あるいは「www」などのような，共通の応答として使える短いメッセージたちは，前のメッセージに「返信関係」を持つと判別されやすくなると考えられる．この問題を根本的に解決するには，文脈情報を追加する必要があると考えられ，今後，この方向でモデルの改善を進めていく予定である．

6 結論

本稿では，ラベル無し対話スクリプトのデータセットと Next Sentence Prediction タスクを用いて，グループチャットのテキストメッセージにおける「返信関係」を判別する手法を提案した．グループチャットにおけるメッセージは，複数の「返信先」を持つ可能性があると考えられ，それに応じて複数返信先のメッセージを判別できる NSP モデルを活用する手法を提案した．小規模なデータセットで提案手法を検証した．データセットとしては，対話スクリプトから作成した訓練データが 5,094 件のメッセージペアと実際のグループチャット記録からの検証データが 444 件である．三つのセッティングでモデルを構築して検証を行った結果，検証セットで最大 88.82%，検証データで 69.64%の精度を得た．

引き続き，訓練データの収集を続けて，モデルの性能改善を行う予定である．

謝辞

本研究の一部は，立命館グローバル・イノベーション研究機構，第 4 期拠点形成型 R-GIRO 研究プログラム「「心の距離メータ」を用いたフィジカル・サイバー空間における人間関係構築技術の開発」の支援を受けて実施されました．感謝の意を表します．

参考文献

[1] 西原陽子, 砂山渡, 谷内田正彦: 発話テキストからの人間の仲の良さと上下関係の推定, 電子情報通信学会論文誌, Vol. 91, No. 1, pp. 78–88 (2008).

[2] Andalibi., Nazanin., Frank, Bentley., and Katie, Quehl.: Multi-channel topic-based mobile messaging in romantic relationships, *Proceedings of the ACM on Human-Computer Interaction*, 1.CSCW, pp. 1–18 (2017).

[3] Jacob, Devlin., Ming-Wei, Chang., Kenton, Lee., and Kristina, Toutanova.: Bert: Pre-training of deep bidirectional transformers for language understanding, *In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, vol. 1, pp. 4171–4186 (2019).

[4] Shi, Wei., Vera, Demberg.: Next sentence prediction helps implicit discourse relation classification within and across domains, *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 5790–5796 (2019).

[5] Liu, Jingyun., Jackie, CK, Cheung., and Annie, Louis.: What comes next? Extractive summarization by next-sentence prediction, arXiv preprint arXiv:1901.03859, (2019).

[6] J, Feng., J, Mostow.: Towards Difficulty Controllable Selection of Next-Sentence Prediction Questions, *EDM*, (2021).

[7] L, Li., Q, Wang., B, Zhao., et al.: Pre-Training and Fine-Tuning with Next Sentence Prediction for Multimodal Entity Linking, *Electronics*. Vol. 11, pp. 21–34 (2022).

[8] Elsner., Micha., and Eugene, Charniak.: Disentangling chat, *Computational Linguistics*, Vol. 36, No. 3, pp. 389–409 (2010).

[9] Elsner, Micha., Eugene, Charniak.: You talking to me? a corpus and algorithm for conversation disentanglement, *Proceedings of ACL-08: HLT*, pp. 834–842 (2008).

[10] L, Wang., D, W, Oard.: Context-based message expansion for disentanglement of interleaved text conversations, *Proceedings of human language technologies: The 2009 annual conference of the North American chapter of the association for computational linguistics*, pp. 200–208 (2009).

[11] J, Kim., W, Lee., et al.: Optimized combinatorial clustering for stochastic processes, *Cluster Computing*, Vol. 20, No. 2, pp. 1135–1148 (2017).

[12] S, Mehri., G, Carenini.: Chat disentanglement: Identifying semantic reply relationships with random forests and recurrent neural networks, *Proceedings of the Eighth International Joint Conference on Natural Language Processing*, Vol. 1, pp. 615–623 (2017).

[13] G, Guo., C, Wang., J, Chen., et al.: Who is answering whom? Finding “Reply-To” relations in group chats with deep bidirectional LSTM networks, *Cluster Computing*, Vol. 22, No. 1, pp. 2089–2100 (2019).

[14] J, Shan., Y, Nishihara., A, Maeda., and R, Yamanishi.: Question Generation for Reading Comprehension Test Complying with Types of Question, *Journal of Information Science and Engineering*, Vol. 38, No. 3, pp. 571–589 (2022).