物語を対象とした登場人物の関係図抽出

内野　太智[1]　　Danushka Bollegala[2]　　Naiwala P. Chandrasiri[1]
[1]工学院大学大学院 [2]リバプール大学
em23006@ns.kogakuin.ac.jp　bollegala@liverpool.ac.uk
chandrasiri@kogakuin.ac.jp

概要

本研究では，小説の物語テキストから登場人物を抽出し，人物関係図を作成するシステムを提案する．小説を選ぶ際に，作成した人物関係図を使用することで，小説を読まなくても内容の全体像の概要を把握でき，好みの内容の小説だけ選択できる．また，物語の進行を忘れた場合にも，それまでの物語の全体像を把握でき，読書の再開を手助けする．本システムで，少しでも読書をする際のストレスになりうる要因を無くすことを目指す．詳細な手法としては，物語テキストから人名を抽出し，登場人物リストを作成した後，GPT-2 を使用して代名詞を最も適切な単語に置き換え，関係性を出力し関係図を作成した．定量評価の結果，代名詞の変換を行った関係図の方が正解率が高いという結果となった．

1 はじめに

近年，読書離れが深刻化している．その中でも特に 20 代を中心に読書をする習慣がない人が多くなっており，問題視されている．その原因の一つとして，読書は多くの時間を取られ，読み進めないとどのような内容かわからないということが考えられる．現代には，多くの娯楽が存在しており，そのほとんどが多くの時間をかけずとも楽しめるものだ．そんな状況だからこそ，内容を忘れてしまったら読み返すしか方法がない読書をする人が減少してしまう原因のひとつとなっていると考察した．このように若者の読書離れは深刻化しているが，近年，スマートフォンやタブレットの普及により，電子出版の市場規模が増加している．そのため，スマートフォンやタブレット等の電子媒体で読書する機会が増えた．電子媒体で読書することは，読書をより身近なものとし，読書離れを解決する方法の一つではないかと考えた．電子媒体で読書をする場合，紙媒体での小説と異なり，複数冊所持しても重たくない．そのため，複数冊を並列して読む人が増えると考えられる．並列して小説を読む場合，物語の進行を忘れてしまうことが増えることが想定される．しかし，現代社会では読書にまとまった時間を確保することが難しく，隙間時間を利用し読書する人が多いため，前のページに戻るために時間を使うことは効果的ではない．そこで，小説の物語テキストから登場人物を抽出し，人物関係図を作成するシステムを作り，読み返しをしなくても本の内容を思い出せるようにしたいと考えた．

2 関連研究

小林らの研究[1]では，物語を既存の辞書などを利用して場所，時間，人物候補を抽出し，これら 3 つのカテゴリ別に数えた語句の異なり数を基準としてシーンを分割する手法を提案している．また，米田らの研究[2]では，物語から局所出現頻度と共起する述語情報を利用して未知の人物名を抽出する手法を提案している．神代らの研究[3]では，発話文と話し手の相対的な位置などを素性とした機械学習により話し手と聞き手を同定し，さらに「わたくしめ」のような人称表現などを素性として人物関係の有無を判定する分類器を学習し，会話文から友好・敵対関係及び上下関係にある人物を抽出する手法を提案している．Srivastava ら[4]は，テキストの文脈の意味を利用するために感情分析を使用し，相互作用に極性を関連付けられることを示した．そして，Chu ら[5]は，BERT を活用したニューラル学習とテキストパッセージの要約を組み合わせた手法が関係抽出に有効だと示した．Shahsavari ら[6]は，読者レビューを使用することにより，物語の枠組みの生成が可能であることを示している．

2.1 人物情報の抽出と体系化について

馬場らの研究[7]では，英米文学の推理小説の物語テキストから形態素解析結果に基づいて人名を抽出

し，体系化では場面における共起頻度を用いて特定二者間の関連度を計算している． その結果，人物相関図を作成できることが明らかにされている．図１は，馬場らが考案した手法の概要を示す．入力は小説テキストを使用し，出力は人物相関図となっている．長方形は処理を表し，円柱は使用する規則や辞書などの資源を示している．また，縣らの研究[8]では，あらかじめ生成した死亡表現リストに基づいた存在状態の判定が人物の情報を抽出に有効であることを示した．

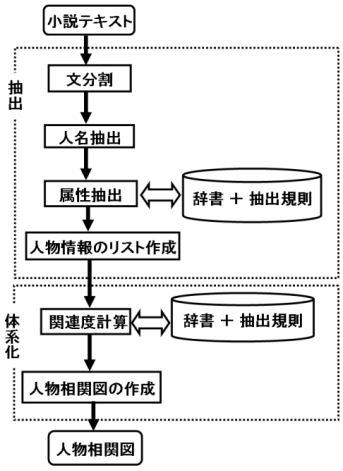


図１　抽出と体系化の概要[7]

Fig. 1　Extraction and Systematization Overview

2.2事前トレーニング済みの言語モデルの改善

Tianyu らの研究[9]では，少数の用例を用いて言語モデルを微調整するための，効果的なファインチューニングする手法が提案された．本研究では，小説テキストに関係性をマスクにしたテンプレートを挿入し，関係性を出力している．

3 システム内容

従来の研究では，登場人物同士の関係性を明らかにする手法は見当たらない.そのため,本研究では,代名詞を登場人物の名前に置き換え，関係性を明らかにする．まず，青空文庫に収録されている書籍から登場人物の名前の抽出を行う．各処理を文単位で行うため，テキストを一文ずつに分割する．次に，形態素解析結果に基づいて人名を抽出し，登場人物リストを作成する．その後，代名詞を登場人物リストに基づいて，変換する．人名同士の関連度は，文単位の名詞リストを作成し，共起単語のペアと出現頻度からなる辞書オブジェクトを参照する.最後に,GPT-2 [10]を使用し，関係性を出力する．

3．1 人名抽出

本研究では，まず馬場らや西原ら[11]の手法を参考に,MeCab[12]を使用し形態素解析を行う. そして,形態素解析の結果, 品詞が「’固有名詞’, ’人名’」と解析された形態素を人名として抽出し，登場人物リストを作成する．その際に，文章中に連続して，品詞が「’固有名詞’，’人名’」と解析された形態素がある場合，その単語は苗字と名前の可能性が高いため，その二つを合わせた単語を名前として扱う．また，人名として登録されていない登場人物の名前を抽出できるように品詞が，「’助詞’」と解析された形態素の中で「’接続助詞’」ではないもののひとつ前の単語も抽出する．

3．2 代名詞の変換

MeCab を使用し形態素解析を行い，品詞が「’代名詞’，’一般’」の単語を”[MASK]”というトークンに変える．その後，その” [MASK]”に登場人物リストの単語を順に入れていく.そして, GPT-2 を使用し，登場人物リストの単語ごとの Perplexity スコアを計算し, Perplexity スコアの最も低い単語を文章に挿入する．

3．2．1 Perplexity スコア

Perplexity は指定されたトークンの並びが発生する確率を変換したものである． 本研究では，Perplexity スコアが低いほど，自然な文とした．Perplexity を計算する式(1)に示す. n はデータセット中の n 番目の単語を表す．$t_{n,k}$は n 番目の単語の正解ラベル，$P_{model}(y_{n,k})$は n 番目の単語に k を予想する確率を表す.

$$ppl = \exp\left(-\frac{1}{N}\Sigma_n\Sigma_k t_{n,k} log p_{model}(y_{n,k})\right) \quad (1)$$

3．3 関係性出力

Tianyu らの研究を参考に，小説テキストを 600 文字ごとに区切り，関係性を”[MASK]”としたテンプレートを文末に挿入する．その後，”[MASK]”に関係性を表す単語を入れる．本研究で，関係性を表す単語として扱う単語は，表 1 に表す．そして，GPT-2 を使用し，各単語の Perplexity スコアを計算し，その中で最も低い単語を文章に挿入する．表 2 に本研

究で使用したテンプレートを示す.

表 1　使用した人物同士の関係性を表す名詞

Table 1 Nouns used to describe relationships between characters

知人	きょうだい	いとこ
恋人	同一人物	親子
夫婦	無関係	

表 2　使用したテンプレート

Table 2 Templates used

[name1 + "と" + name2 + "は" + "[MASK]という関係だ。]
[name1 + "は" + name2 + "と" + "[MASK]という関係だ。]
[name2 + "は" + name1 + "と" + "[MASK]という関係だ。]

3. 4 関係図の作成

本研究では，関係図を作成する際に，ノードを人物名とし，人物関係をエッジとして，人物関係図を表現する．また，関連度の高さが上位 3 位までのエッジは太く表現される．グラフ作成には，networdx[13]を使用する．

3. 5 評価

本実験では，それぞれの関係性の出力結果をもとに，正解率を算出し，各物語と各関係性での精度を確認した．本研究では，正解として扱う関係性は，20 代前半の男女 3 人が小説を読んで選択したものとする．

人物関係性クラスをL_1からL_nまでとし，クラスL_iと予想されたクラスL_jの件数をC_{ij}をすると，正解率 A は以下の式(2)で表される．

$$A = \frac{\sum_{i=1}^{N} C_{ii}}{\sum_{i=1}^{N}\sum_{j=1}^{N} C_{ij}} \quad (2)$$

4 実験

本研究では，青空文庫に収録されている作品の中で芥川龍之介の「あばばばば」，「秋」，「羅生門」，「藪の中」，「鼻」，太宰治の「律子と貞子」，江戸川乱歩の「日記帳」を使用した．以下の 2 つの実験を行った．

4. 1 実験 1

GPT-2 での代名詞の変換を行わずに人物関係図を作成した．

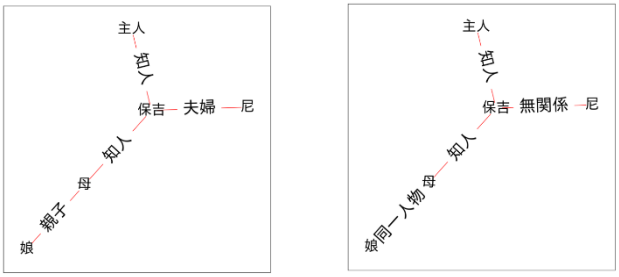
4. 2 実験 2

GPT-2 での代名詞の変換を行い，人物関係図を作成した．

5 実験結果

5. 1 実験 1 結果

GPT-2 を使用し関係性を出力した結果を示す．図 2 に，実験 1 で物語テキストが入力された際に，出力された関係図の例と出力された人物関係図の正解の例を示す．表 3 は，物語ごとの正解率を算出し表にまとめたものである．



(a) 予測結果(出力)　(b) 正解とする結果

図 2　実験 1 で出力された「あばばばば」の人間関係図（a）と正解とする関係図(b)

Fig. 2　Human relationship diagram of "Ababababa" in the experiment 1, (a) Predicted (b) Ground truth

表 3　実験 1 の物語ごとの人間関係の正解率

Table 3　Accuracy of Human relationship extraction for different stories in experiment 1

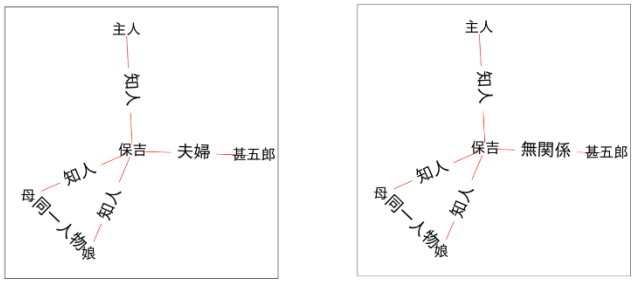
	正解率（%）
あばばばば	50
秋	80
日記帳	50
律子と貞子	50
羅生門	14
藪の中	57
鼻	20

図 2 より，「尼」という物語の登場人物ではない単語が名前として登場人物リストに入っていた．ま

た，図 2（b）より，正しい出力結果である「同一人物」ではなく，「親子」と出力されている．このような出力になってしまう原因のひとつとして，前後の文章で親子に関する会話や描写がされているためだと考えられる．また，表 3 より，「あばばばば」より「秋」の方が正答率が高いことが分かる．理由としては，「秋」で使われる文章が GPT-2 が学習対象としている現代仮名遣いに近いものであるからだと考えられる．

5. 2 実験２結果

GPT-2 を使用し関係性を出力した結果を示す．図 3 に，実験 2 で物語テキストが入力された際に，出力された関係図と出力された人物関係図の正解の例を示す．表 4 は，物語ごとの正解率を算出し表にまとめたものである．



(a) 予測結果(出力)　　(b) 正解とする結果

図 3　実験 2 で出力された「あばばばば」の人間関係図（a）と正解とする関係図(b)

Fig. 3　Human relationship diagram of "Abababababa" in the experiment 2, (a) Predicted (b) Ground truth

表 4　実験 2 の物語ごとの人間関係の正解率

Table 4　Accuracy of Human relationship extraction for different stories in experiment 2

	正解率（%）
あばばばば	80
秋	80
日記帳	50
律子と貞子	60
羅生門	50
藪の中	70
鼻	40

図 3 より，登場人物ではない単語が登場人物リストに入っていることが分かる．表 4 より，実験 1 と比較して実験 2 の正解率がどの物語でも高くなっていることが分かる．

6 考察

名詞の変換を行った後の文章の方が，登場人物同士の関係性をより正しく出力できていることが分かった．しかし，実験 2 での正解率を見ても，80.0%が最大となっている．正解率を下げている原因のひとつに，登場人物の名前ではない単語が人物関係図に出力されていることが考えられる．関係性を表す単語の候補として与えている「無関係」は 1 回も出力されていない．このことから，本研究で使用した人物関係を表す単語の中で，広義的な単語と狭義的な単語があり，それが原因だと考えられる．また，人物関係図を作成する際に使用する登場人物の名前リストの中に，物語テキストで使用された地域や土地の名称や，本の著者など，登場人物ではない名詞も入っている．このことから，登場人物の名前をリストとする際に，人の氏名にも土地の名称にも使われるような名詞は，前後の文脈を考慮する必要があると考えられる．物語ではそれまでに登場していない人物が代名詞を使用し，登場することがある．現在の手法では，そのような場合に間違った登場人物の名前を入れてしまう．また，そのような場合に対応するためには，Perplexity スコアを比較した際に，任意の閾値以上の場合に変換を行わないというシステムにする必要がある．その閾値の設定を正確に行わなくては，間違った代名詞の変換が行われてしまうため，今後適切な閾値を分析する必要もある．

7 むすび

本研究では，まず，物語テキストから人名を抽出して登場人物リストを作成した．その後，代名詞を GPT-2 を使用し，Perplexity スコアの最も低い登場人物リスト内の単語に入れ替え，関係性の出力し関係図を作成した．性能評価の結果，代名詞の変換を行った方が，高い精度で関係図を作成することができた．今後の課題としては，登場人物同士からみた関係性を出力するための手法を考える必要がある．

参考文献

[1] 小林聡，“場・時・人に着目した物語のシーン分割手法”，情報処理学会自然言語処理研究会, pp. 25-30，2007.

[2] 米田崇明，篠崎隆宏，堀内靖雄，黒岩眞吾，“述語情報を利用した小説の登場人物の抽出”，語処理学会第 18 回年次大 会発表論文集，Vol.18，pp.855-858，2012.

[3] 神代大輔，高村大也，奥村学，物語テキストにおけるキャラクタ関係図自動構築，言語処理学会第 14 回年次大会発表 論文集, Vol.14, pp.380-383, 2008.

[4] Shashank Srivastava,Snigdha Chaturvedi，Tom Mitchell，Inferring interpersonal relations in narrative summaries，In 30th AAAI Conference on Artificial Intelligence，2016

[5] Cuong Xuan Chu，Simon Razniewski，Gerhard Weikum，“KnowFi，Knowledge Extraction from Long Fictional Texts”，AKBC，2021

[6] Shadi Shahsavari，Ehsan Ebrahimzadeh，Behnam Shahbazi，Misagh Falahi，Pavan Holur，Roja Bandari，Timothy R，Tangherlini，Vwani P，Roychowdhury，“An Autoated Pipeline for Character and Relationship Extractionfrom Readers Literary Book Reviews on Goodreads.Com”，WebSci ，2020.

[7] 馬場こづえ,藤井敦,“小説テキストを対象とした人物情報の抽出と体系化，”言語処理学会年次発表論文集，13th，pp574-577，2007.

[8] 縣啓示，伊藤雄一，高島和毅，北村喜文，岸野文郎，“物語テキストから進行状況に応じて登場人物の存在状態と関係を推定する手法”，2010.

[9] Tianyu Gao，Adam Fisch，Danqi Chen，“Making Pre-trainbed Language Models Better Few-shot Learners”，Association for Computational Linguistics，2021.

[10] Alec Radford，Jeffrey Wu，Rewon Child，David Luan，Dario Amodei，Ilya Sutskever，”Language Models are Unsupervised Multitask Learners”，2019． (参照:2024-1-9).

[11] 西原弘真，白井清昭，“物語テキストを対象とした登場人物の関係抽出”，言語処理学会第 21 回年次大会，2015.

[12] 工藤拓，“MeCab”http://mecab.sourceforge.net/(参照:2024-1-9).

[13] Network, http://networkx.lanl.gov/contents.html (参照:2024-1-9).