7V-07

BERTによる小説の構造化手法の検討

上間　翔 †
琉球大学大学院理工学研究科 †

當間　愛晃 ‡
琉球大学工学部知能情報コース ‡

1　目的と背景

電子化されたテキストには様々な種類がある。新聞や論文、特許といったものから、漫画や小説、雑誌などといったものもある。馬場ら [1] はこの論文といったデータを情報伝達テキスト、小説といったデータを娯楽・芸術テキストと称している。

上記の情報伝達テキストではキーワードによって検索を行うが、小説の場合、自分の検索したい作品が上記のように明確でない場合もある。仮に、明るいキャラクターが登場する話を読みたいとする。そういった場合では、既存のジャンルによる検索のみでは対応することができない。

このように、自分の好みのキャラクターや展開のテキストを検索するためには、話の展開やキャラクターの性格といった情報を検索システムに組み込む必要がある。岡 [2]、馬場 [1] らの研究では、登場人物の名称、性別表現、年齢表現、容姿・性格表現といった属性を抽出している。しかし、これらの研究では性格に関する属性の抽出において、性能が低くなっている。

先行研究を踏まえ本研究では娯楽・芸術テキストを対象とし、性格に関する表現の抽出を目的とし、先行研究で出された性能を上回ることを目指す。

2　関連研究

小説における登場人物を抽出する研究は、岡 [2] が行っている。岡の研究では、小説家になろうに掲載されているファンタジー小説に対し、岡の先行研究で用いられた [3]BiLSTM-CRF-pos10 モデルを用いることによって、登場人物の属性を抽出している。抽出する属性は、以下の通りである。

- NAME：登場人物の名前
- MF：性別表現
- AGE：年齢表現
- STATE：容姿・性格表現
- PRO：職業・立場表現
- AFF：組織・種族名
- OTHER：その他の人物表現
- PLACE：地名・建物名
- REL：人物関係表現
- O：以上に当てはまらないもの

本研究においても上記と同じ属性セットの抽出を目的とし、BERT 及び BertForTokenClassification を用いることで構造化を試みる。

3　提案手法

BERT を用いて、ラベル分類を行う。分類にあたっては BertForTokenClassification を用い、ラベリング形式としては IO 方式とする。

学習を行うにあたって、抽出する属性を抽出することができなかった場合、損失関数の値が増えるように式 (1) によりペネルティの設定を行なった。また式 (1) において、x は BertForTokenClassification の出力、y は正解ラベル、C はクラスの数、N はミニバッチの値、P は O でない属性に対し O と予測した数である。

$$
\begin{aligned}
l(x, y) &= L = \{l_1, ..., l_N\}^T \\
l_n &= -w_{y_n} \ln \frac{\exp(x_{n,y_n})}{\sum_C^{c=1} \exp(x_{n,c})} + P
\end{aligned}
\tag{1}
$$

A consideration of novel structuring by BERT

† Sho Uema, Graduate School of Engineering and Science, University of the Ryukyus
‡ Naruaki Toma, Computer Science and Intelligent Systems Program, School of Engineering, Faculty of Engineering, University of the Ryukyus

4 実験

4.1 実験手順

- 手順1：娯楽・芸術テキストのデータセットの準備。小説家になろう API を用いることにより、小説のテキストデータを取得する。取得するデータの条件は、岡の先行研究 [3] と同様とした。
- 手順2：手順1で得られたデータに対してアノテーションを行う。1文書から40文抽出し、得られたラベルの数を以下に示す。
  - 登場人物の名前：1
  - 性別表現：30
  - 年齢表現：6
  - 容姿・性格表現：29
  - 職業・立場表現：0
  - 組織・種族名：0
  - その他の人物表現：3
  - 地名・建物名：0
  - 人物関係表現：4
- 手順3：ファインチューニングを行う。学習にあたっては、エポック数は10とし、トレーニングデータを6割、バリデーションデータを2割、テストデータを2割とした。

4.2 結果

テストデータにおける3ラベルに対する抽出結果を表1に示す。

表1　キャプションを記述

	precision	recall	F1 score
STATE	62.50	71.42	66.66
MF	57.14	50.00	53.33
NAME	0.00	0.00	0.00
先行研究の STATE	14.24	26.02	18.41

4.3 考察

STATE については、先行研究の precision を上回る数値となった。しかし、この STATE において、テストに用いられたデータ7つのうち、髪について表現しているものが5つであった。表現されている情報が偏っているため、体格や性格といった別の表現を性格に抽出できているかは不明である。

MF についても同様のことが考えられる。MF を表現する際に、男や女という単語が多く使用されるため、抽出が比較的容易であったのではないかと考える。

対照的に、NAME については、precision、recall、F1 score 全て 0% となっている。名前というものは多くの場合、登場人物によって異なるものである。名前を表現するパターンが多かったため、抽出が困難になったと考える。

5 今後の展望

今回は用いたデータが1作品から抽出したものであったので、データ数が少なくなった。それにより、考察で述べたように、表現される情報に偏りが発生し、本来抽出すべき表現を抽出できていない可能性が考えられる。この問題はデータ数を増やすことにより、今後対応したいと考える。

また名前については、正しく予測できたデータ数は0件であった。登場人物の名前というのは固有名詞にあたる。そのほかにも、容姿を表現する際には名詞である体の部位に対し、形容詞を用いることによって表現することも考えれる。このように、属性を抽出するにあたって品詞情報が関わる場合も存在する。そのため、品詞情報を考慮した分類を検討する必要があると考える。

参考文献

[1] 馬場こづえ, 藤井敦. (2007). 小説テキストを対象とした人物情報の抽出と体系化. 言語処理学会第 13 回年次大会発表論文集, 13, pp.574-577.

[2] 岡裕二, 安藤一秋. ”小説本文から抽出した人物情報の構造化手法の検討.” IEICE Conferences Archives. The Institute of Electronics, Information and Communication Engineers, 2021., pp215-216

[3] 岡他, (2021, August). 小説あらすじを用いて学習した系列ラベリングモデルに よる小説本文からの人物情報抽出の性能検証. 言語処理学会第 27 回年次大会., pp8-17