E-007

系列ラベリングを用いたファンタジー小説のあらすじからの人物情報抽出
Extracting Person Information from Synopsis of Fantasy Novels using Sequence Labeling

岡 裕二† 安藤 一秋†
Yuji Oka kazuaki Ando

1. はじめに

近年，電子小説や小説投稿サイトなどの発展により，小説が容易に読めるようになった．一方，小説数が増加したことで，個人の嗜好に合う作品を探すことが難しくなった．また，いつでもどこでも小説が読める環境が整ったことにより，隙間時間での読書が増加し，前回の読書内容を思い出すための読み返しの機会も増加している．ストーリー展開やキャラクターの特徴に関する個人の嗜好を用いた小説検索やあらすじを自動生成することで，これらの問題を改善する一助になると考える．

本稿では，解決策の実現への第一歩として，系列ラベリングにより，ファンタジー小説のあらすじから，名前，年齢，職業，容姿などといった人物情報と一部の関係表現を抽出する手法について検討する．

2. 人物情報抽出手法の検討

筆者らの先行研究[1]において，深層学習を用いた系列ラベリングモデルによって，ファンタジー小説のあらすじから人物情報を抽出する手法を検討した．本稿では，品詞・品詞細分類の情報と，カタカナに対する前処理を導入することにより，抽出性能への影響を評価する．

以下，あらすじの収集と教師データの作成，深層学習モデル，品詞・品詞細分類の情報，カタカナに対する前処理について述べる．

2.1 教師データ

先行研究[1]と同様，国立情報学研究所（NII）が提供する Webcat Plus[2] で収集したファンタジー小説のあらすじを利用する．2 文以上で構成されるあらすじを対象に，各文を Mecab で単語分割した後，表 1 に示す 10 種類のタグを人手で付与することで，3,524 文からなる教師データを構築した．タグの形式は，IOB2 タグ形式を用いた．

表 1 の地名・建物名に関しては，直接的な人物情報ではないが，小説の舞台（現実世界，パラレルワールド，異世界など）を考慮するために同時に抽出する．また，人物関係表現は，人物関係図のエッジラベルに利用するために抽出を試みる．

2.2 深層学習モデル

本研究では，先行研究[1]で最良性能を記録した，Huang らが提案した BiLSTM-CRF モデル[3]に加え，新たに Char-BiLSTM-CRF モデル[4]を用いて抽出性能を比較する．BiLSTM-CRF モデルは，文中の単語に対する word embeddings を BiLSTM に入力して得られる単語ベクトルを CRF に入力することで，系列ラベリングするモデルである．Char-BiLSTM-CRF モデルは，文字単位でラベリングするモデルであり，文字のベクトルと文字を含む単語のベクトルを結合して BiLSTM に入力する．このモデルは毎日新聞コーパスに対する固有表現抽出において最高性能を達成したモデルである．本稿では，登場人物の情報抽出における Char-BiLSTM-CRF モデルの有効性を検証するために採用した．なお，文字ベクトルには，ランダムに初期化したものを利用する．

表 1. 付与するタグの種類と例

タグ名	タグ付け対象	タグ付け例
NAME	登場人物の名前	西尾，太郎，シャルル・マーニュ
MF	性別表現	男，美男子，乙女
AGE	年齢表現	16歳，お婆さん，幼い
STATE	容姿・性格表現	白い髪，元気，高飛車
PRO	職業・立場表現	騎士，権力者，メンバー
AFF	組織・種族名	日本政府，討伐軍，エルフ
OTHER	その他の人物情報	神，気鋭，ペンギン
PLACE	地名・建物名	ムー大陸，日本，礼拝堂
REL	人物関係表現	兄，相棒，結婚
O	以上に当てはまらないもの	

2.3 品詞・品詞細分類の情報

Aguilar らの研究[5] により，深層学習モデルに品詞情報を追加することで，ソーシャルメディア中のテキストから構築された WNUT2017 データセットに対する抽出性能が向上したと述べられている．本稿においても，深層学習モデルに品詞情報を組み込むことで，抽出性能が変化するかを確認する．

本稿では，品詞と品詞細分類でそれぞれランダムに初期化した品詞ベクトルを用意する．単語ベクトルや文字ベクトルを BiLSTM に入力する際に同時に入力し，モデルの学習とともに品詞ベクトルの値を更新する．次元数は，品詞および品詞細分類でそれぞれ 5 と 10 で実験する．

2.4 カタカナに対する前処理

本稿では，ファンタジー小説のあらすじを対象にすることから，一般の小説に比べて，カタカナ表記の人名や組織名，場所名などが出現する傾向がある．その中でも，カタカナのみで構成される形態素とカタカナ以外の文字も含む形態素が存在する．そこで，Char-BiLSTM-CRF に入力される文字の中で，カタカナを含む形態素について，カタカナのみ（kOnly），またはカタカナ以外の文字も含め形態素全体（kALL）を一纏めの文字列にして入力することで，抽出性能に対する影響を確認する．

3. 評価実験

先行研究[1]で提案した BiLSTM-CRF モデルを Baseline とし，品詞・品詞細分類の情報とカタカナに対する前処理を導入した深層学習モデルとの抽出性能を比較する．抽出性能は，Precision，Recall，F 値を評価尺度に，10 分割交差検証で評価する．人手でタグ付けした結果と機械学習モデルがラベリングした結果を比較し，完全一致した場合の

† 香川大学, Kagawa University

Copyright © 2020 by
The Institute of Electronics, Information and Communication Engineers and
Information Processing Society of Japan All rights reserved.

表 2. 深層学習モデルのハイパーパラメータ

BiLSTM の隠れ層の次元数	128
BiLSTM の層数	1
最大エポックサイズ	50
バッチサイズ	32
学習率	0.001
Dropout rate	0.5
勾配クリッピング	5.0
最適化手法	Adam
Early stopping patience	20

表 3. 事前学習した単語分散表現のハイパーパラメータ

モデル	cbow
次元数	200
Window size	5
ネガティブサンプリング	5
ダウンサンプリング	0.001

みを正解と判断する．表 2 に深層学習モデルで用いたパラメータ，表 3 に事前学習した単語分散表現のパラメータを示す．単語ベクトルには，日本語 Wikipedia の全文で事前学習した分散表現[6] を用いる．単語ベクトルおよび文字ベクトルは，モデルの学習とともに値を更新する．評価には，2.1 で述べた 3,524 文からなるデータセットを用いる．

実験結果を表 4 に示す．表 4 の Model は，上から順に，人物情報の中で最重要と考えられる人名の抽出性能の上位三つのモデル，品詞情報の利用およびカタカナ結合を施していない Char-BiLSTM-CRF（Char），Baseline，カタカナ結合した中での最高性能モデル（Char+kOnly+pos5）である．なお，Model 名の pos@は品詞ベクトルの次元数を表している．品詞と品詞際分類にそれぞれ次元が割り当てられるので，実際に分散表現に結合される次元数は@の二倍となる．数値の太字は各指標での最高値を表している．

表 4 の Char と Baseline の結果より，文字に着目する Char-BiLSTM-CRF モデルは BiLSTM-CRF モデルよりも性能が若干高いことがわかる．また，Baseline に品詞情報を追加したモデルの性能が最良であることがわかる．一方，カタカナ結合した中で最高性能を得たモデル（Char+kOnly+pos5）は，形態素内のカタカナのみを結合したものであるが，Precision が 86.66 と最高値になったものの，F 値は 86.40 と Baseline より低くなった．したがって，NAME の抽出性能の向上には，カタカナ結合があまり寄与しなかったといえる．

4. エラー分析

品詞情報の有無でエラー分析した結果，「言種」という人物情報ではない単語や「チンゲンツァイ国」という実在しない国名に対して，品詞情報を導入したモデルは正確にラベリングできるようになった．逆に，祝いの席を表す「パーティー」を AFF と判断したり，「綾模様」の「綾」を NAME と判断したりするミスが生じた．文脈から判断できるタグをミスしていることから，品詞情報を導入することで，文脈を参照する割合が減少した可能性がある．

表 4. 深層学習モデルの NAME タグ性能

Model	Precision	Recall	F1
Baseline+pos10	85.01	**91.29**	**88.01**
Baseline+pos5	84.71	91.17	87.79
Char+pos5	86.53	89.11	87.78
Char	85.28	90.04	87.58
Baseline	83.17	90.55	86.69
Char+kOnly+pos5	**86.66**	86.19	86.40

BiLSTM-CRF(Baseline)モデルと Char-BiLSTM-CRF モデルを比較すると，Char-BiLSTM-CRF モデルは形態素情報に文字情報も加えて判断するため，「美丈夫」という 3 例しかない表現に対しても適切にラベリングできていた．しかし，「グレてた」の「グレ」を NAME とラベリングしてしまうミスもあり，該当文字が含まれやすいタグに引き寄せられやすくなる可能性がある．品詞情報を追加した Char+pos5 は，品詞情報を基に「グレてた」に対して適切にラベリングできていた．

Char+pos5 と Char+kOnly+pos5 でエラーを比較すると，後者は「エルダント」や「フェアリィ」などの学習データに複数回出現するものには適切にラベリングできている一方，「ガリウス」や「ボーカル」などの一度も教師データに出現しないものはラベリングできない傾向が見られた．形態素のカタカナ部分をまとめることで，固有名詞はカタカナごとの分散表現に惑わされなくなった一方，情報量がカタカナの文字数分から一つに減少することや，未知のカタカナ語のベクトルを学習できないことで汎化性能が落ちた可能性がある．

5. 終わりに

本稿では，深層学習を用いた系列ラベリングモデルに品詞情報やカタカナ結合を導入することで，抽出性能に与える影響を検証した．結果として，BiLSTM-CRF に品詞情報を加えたモデルが最良となったが，NAME の抽出性能としては，1.3 ポイント程度の向上であった．また，カタカナ結合は，NAME の抽出性能の向上には寄与しなかった．今後は，カタカナの有効活用法の検討を継続すると共に，データセットの拡充，人物名と人物情報，人物と人物の紐付け手法，あらすじの自動生成手法の検討を進める．

参考文献

[1] 岡他, “系列ラベリングによる小説のあらすじからの人物情報・関係表現抽出手法の検討”, 2020 年度人工知能学会全国大会（第 34 回）論文集, (2020).
[2] 国立情報学研究所（NII）, Webcat Plus, http://webcatplus.nii.ac.jp/.
[3] Z. Huang, et al., “Bidirectional lstm-crf models for sequence tagging”, arXiv:1508.01991 (2015).
[4] S. Misawa, et al., “Character-based Bidirectional LSTM-CRF with words and characters for Japanese Named Entity Recognition”, in Proc. of the First Workshop on Subword and Character Level Models in NLP (2017).
[5] G. Aguilar, et al., “Modeling noisiness to recognize named entities using multitask neural networks on social media”, in Proc. of the North American Chapter of the Association for Computational Linguistics (2018).
[6] 日本語 Wikipedia エンティティベクトル, http://www.cl.ecei.tohoku.ac.jp/~m-suzuki/jawiki_vector/.

Copyright © 2020 by
The Institute of Electronics, Information and Communication Engineers and
Information Processing Society of Japan All rights reserved.