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

岡裕二
香川大学大学院工学研究科
s20g460@stu.kagawa-u.ac.jp

安藤一秋
香川大学創造工学部
ando.kazuaki@kagawa-u.ac.jp

1 はじめに

近年，電子書籍や小説投稿サイトの発展により，小説を読む際の場所や時間の制限が緩和されると共に，小説の数も増加し続けている．小説の数が増えることで個人の嗜好にあった小説も増える可能性があるが，膨大な数の小説の中から個人の嗜好にあった作品を発見する労力は増大していると考えられる．書籍を取り扱う EC サイトや小説投稿サイトには，作者やジャンルなど，特定の情報に基づく検索機能が実装されてはいるが，小説の内容に踏み込んだ検索機能は実装されていない．

個人の嗜好は，「ハッピーエンド」や「敵が仲間になる」などの展開に関する嗜好と，「銀髪赤目の少年」や「長身のメイド」などの登場人物に関する嗜好に分けることができる．本研究では，小説内の登場人物情報を体系的に抽出することで，登場人物情報による小説検索，登場人物情報を豊富に取り入れた小説のあらすじ生成，人物関係図の自動生成などを目指している．

筆者らの先行研究 [1] では，商業小説のあらすじテキストにタグ付けしてデータセットを構築し，深層学習モデルと CRF（Conditional Random Fields）を組み合わせた系列ラベリングモデルにより，あらすじテキストから人名，性別，容姿性格，職業などの登場人物情報を自動抽出する手法を提案した．本稿では，あらすじテキストで学習した系列ラベリングモデルを Web 小説の本文に適用し，訓練データおよび提案モデルの小説本文に対する抽出性能について検証する．

2 小説本文に基づくテストデータの構築

小説本文を収集して，テストデータを構築する方法について述べる．

2.1 小説本文の収集

筆者らの先行研究 [1] で構築したデータセットは，商業小説のあらましを利用しているため，本文データを利用することができない．そこで，「小説家になろう」サイトから本文データを収集する．

小説本文の収集には，「小説家になろう」を運営している株式会社ヒナプロジェクトが提供する Web API（なろう小説 API）を用いる．なろう小説 API は，ジャンル別や人気の高い順などの条件により，「小説家になろう」に投稿されている小説の各種データを抽出できるが，本文データを抽出することはできない．そこで，API を使用して小説のジャンル，ID，タイトル，著者名，あらすじを収集し，収集した小説の ID を基に，クローリングによって本文データを収集する．収集対象は，長編小説の本文とする．

2.2 テストデータの構築方法

本稿での評価に用いるテストデータは，次の手順で構築する．なお，タグおよびタグ付け方法については，先行研究 [1] と同様である．

1. 「小説家になろう」から収集した長編小説の本文のうち，三話までのテキストを 1 文ずつ形態素解析する．
2. 各形態素に対して，以下のルールでタグ付けする．タグの形式には IOB ２タグ形式を用いる．
   - 名前に名前タグ（NAME）を付与
     例：西尾，信長，シャルル・マーニュ
   - 性別表現に性別タグ（MF）を付与
     例：男，美男子，美女，乙女，女の子
   - 年齢表現に年齢タグ（AGE）を付与
     例：16 歳，少年，お婆さん，幼い，高校生
   - 容姿や特性表現に状態タグ（STATE）を付与

例：白い髪，元気，高飛車，天才，職人気質

- 職業や立場表現に能力タグ（PRO）を付与
  例：竜飼い，仙女，最高権限者，メンバー，国王
- 組織・種族名に所属タグ（AFF）を付与
  例：鳳凰学園杖術部，日本政府，討伐軍，エルフ
- 以上に当てはまらない人物情報にその他タグ（OTHER）を付与
  例：異星人，神，元凶，気鋭，ペンギン
- 地名や建物名に場所タグ（PLACE）を付与
  例：ムー大陸，日本，パリ，礼拝堂，魔法学校
- 人物関係表現に関係タグ（REL）を付与
  例：兄，親，敵，相棒，結婚
- それ以外のものに O タグを付与

先行研究 [1] と同様，直接的な人物情報ではないが，地名や建物名，人物関係表現についても，今後の応用を考慮してタグ付けする．

3 検証に用いる深層学習モデル

性能評価に用いるモデルについて説明する．

3.1 検証モデル

検証モデルには，先行研究 [1] と同様，Huang らが提案したモデル（BiLSTM-CRF）[2]，Ma らが提案したモデル（BiLSTM-CNN-CRF）[3]，Lample らが提案したモデル（BiLSTM-CRF-L）[4]，Misawa らが提案したモデル（Char-BiLSTM-CRF）[5] の 4 つの深層学習モデルを採用する．

Huang らの BiLSTM-CRF は，文中の単語に対する word embeddings を Bidirectional LSTM (BiLSTM) に入力し，得られた単語ベクトルを素性の代わりに CRF に入力することで固有表現抽出するモデルである．Ma らの BiLSTM-CNN-CRF と Lample らの BiLSTM-CRF-L は，BiLSTM-CRF に対して，注目単語に含まれる文字情報を利用することで性能向上を図ったモデルである．BiLSTM-CNN-CRF は，注目単語に含まれる文字を CNN に入力して得られた単語ベクトルを word embeddings に結合し，BiLSTM-CRF に入力することで固有表現抽出する．BiLSTM-CRF-L は，BiLSTM-CNN-CRF の CNN の代わりに，注目単語に含まれる文字を Char-BiLSTM に入力して得られた単語ベクトルを word embeddings に結合し，BiLSTM-CRF に入力することで固有表現抽出する．Misawa らの Char-BiLSTM-CRF 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 BiLSTM に入力することで固有表現抽出するモデルである．

表 1 に深層学習モデルで用いたパラメータを示す．Dropout は，BiLSTM への入力の前と後に適用した．表 1 の下部に Ma らの BiLSTM-CNN-CRF と Lample らの BiLSTM-CRF-L で用いたパラメータを示す．単語ベクトルとして用いる分散表現には，日本語 Wikipedia の本文全文で事前学習されたもの [6] を用意した．事前学習に用いたパラメータは表 2 に示すものが使われている．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．

表 1 深層学習モデルのハイパーパラメータ

Common parameters	
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
Parameters of BiLSTM-CNN-CRF	
CNN のフィルタ数	50
CNN の window size	2
Parameters of BiLSTM-CRF-L	
character BiLSTM の隠れ層の次元数	50
character BiLSTM の層数	1

表 2 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

3.2 品詞・品詞細分類情報の活用

Aguilar らの研究 [7] 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．先行研究 [1] と同様，本稿でも，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利用し，有効性を検証する．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．

4 評価実験

4.1 評価方法

小説本文で構築したテストデータに対して，4 つの深層学習モデルと，それぞれのモデルに品詞・品詞細分類の情報を付与したモデルの抽出性能を比較する．抽出性能は，適合率，再現率，F 値を評価尺度に利用する．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みを正解と判断する．

4.2 データセット

訓練データと開発データは，先行研究 [1] で構築した，3,679 文で構成される商業小説のあらすじデータセットを 9:1 に分割して利用した．

テストデータを構築するため，まず，なろう小説 API を用いて，ジャンルがハイファンタジー，またはローファンタジーであり，ランキング上位 2,500 件に入るという条件で収集した長編小説の中からランダムに 8 作品を選出した．そして，各小説の三話までのテキストを MeCab[8] で単語分割し，人手でタグ付けすることにより，3,127 文で構成されるテストデータを構築した．

4.3 実験結果

実験結果を表 3 に示す．表 3 に示すモデルは，品詞・品詞細分類の情報を付与していない 4 つの深層学習モデルと，品詞・品詞細分類の情報を付与したモデルの中で最高性能のモデル（BiLSTM-CRF-pos10）である．モデル名の pos@は，品詞ベクトルの次元数を示す．品詞と品詞細分類それぞれに次元が割り当てられるので，実際に付与される次元数は@の 2 倍となる．太字部分は各ラベルでの最良 F 値である．

表 3 より，PLACE（地名・建物名）と REL（関係表現）を除く全てのラベルで BiLSTM-CRF-pos10 が最良 F 値，または最良に近い F 値であることが確認できる．また，MF（性別表現）と AGE（年齢表現）については，どのモデルでも 9 割近い F 値となった．しかし，他のラベルでは，7 割以下の F 値となっており，特に STATE（容姿・特性表現），AFF（組織・種族名）と PLACE（地名・建物名）は，最良 F 値であっても 5 割以下と非常に低い．

次に，小説本文に対する抽出性能と比較するため，先行研究 [1] であらすじデータセットに対して全体の最良性能を得た BiLSTM-CRF の抽出性能を表 4 に示す． 表 4 は，あらすじデータの 8 割を訓練データ，1 割ずつを開発データとテストデータに分割し，十分割交差検証を行った結果の平均を算出している．本稿での最良性能を確認した BiLSTM-CRF-pos10 の抽出性能と比較すると，MF（性別表現）と OTHER（その他の人物情報）以外のラベルにおいて，抽出性能が低下している．AGE（年齢表現）については軽微であるが，それ以外のラベルは大幅な性能の低下となった．

5 考察

あらすじと本文において抽出性能差が生じる要因について考察する．抽出性能が高い MF（性別表現）と AGE（年齢表現）は，他の人物情報と比べて，表現自体に多様性が少ないことから，高い抽出性能が維持されていると考えられる．OTHER（その他の人物情報）に関しては，あらすじに出現する表現が多様であったため，本文に出現する表現を包含でき，結果として，あらすじに対する性能より高くなったと考えられる．STATE（容姿・特性表現）に関しては，容姿に関する記述法が複数あったり，色が異なるだけで別の形態素として認識されることもあるため，抽出性能が低下したと考えられる．NAME（人名），AFF（組織・種族名），PLACE（地名・建物名）に関しては，固有の表現が多く，基本的に同作品や同じ世界を共有しない限り，同じ名前が使われることがないため，抽出性能が低下したと考えられる．また人名に限定すれば「姫」や「友」など他のラベルに付与される可能性の高い形態素も出現することもあるため，性能が低下していると考えられる．

次に，抽出エラーについて分析する．抽出エラーを以下の 5 種類に分類し，各割合を算出する．

- 人物情報に O タグを振る間違い (ne2oMiss)
- 人物情報ではない部分に人物情報タグを振ってしまう間違い (o2neMiss)
- 抽出範囲は正確だが，人物情報タグの種類を間違えている (classMiss)

表 3　実験結果

	BiLSTM-CRF			BiLSTM-CNN-CRF			BiLSTM-CRF-L			Char-BiLSTM-CRF			BiLSTM-CRF-pos10		
	Pre.	Rec.	F1	Pre.	Rec.	F1	Pre.	Rec.	F1	Pre.	Rec.	F1	Pre.	Rec.	F1
NAME	59.12	56.67	57.87	57.69	51.13	54.21	56.06	45.21	50.05	63.75	53.38	58.11	69.20	61.65	**65.21**
MF	94.74	100.0	**97.30**	85.24	98.72	91.49	86.57	99.15	92.43	95.44	98.29	96.84	95.85	98.72	97.26
AGE	85.87	92.94	89.27	86.10	94.71	90.20	88.64	91.76	90.17	89.41	89.41	89.41	87.43	94.12	**90.65**
STATE	24.25	33.06	27.98	22.47	24.49	23.44	27.24	29.80	28.46	20.30	32.65	25.04	29.32	36.73	**32.61**
PRO	44.65	57.14	50.13	40.83	49.70	44.83	29.44	40.48	34.09	41.08	56.85	47.69	46.70	56.85	**51.28**
AFF	41.67	50.00	**45.45**	23.91	55.00	33.33	26.67	40.00	32.00	31.43	55.00	40.00	38.46	50.00	43.48
OTHER	55.40	81.91	66.09	44.92	89.36	59.79	37.50	86.17	52.26	36.77	87.23	51.74	52.83	89.36	**66.40**
PLACE	27.62	71.43	39.84	25.99	70.00	37.91	29.25	70.00	41.26	32.41	75.00	**45.26**	28.74	70.00	40.75
REL	63.69	66.67	65.14	62.93	75.44	68.62	69.27	72.51	70.86	69.57	74.85	**72.11**	66.29	69.01	67.62

表 4　あらすじデータに対する最良モデル（BiLSTM-CRF）の抽出性能

	Pre.	Rec.	F1
NAME	84.24	90.99	87.48
MF	96.06	96.03	95.95
AGE	92.56	92.83	92.62
STATE	58.85	57.72	57.98
PRO	80.39	79.92	80.14
AFF	72.45	71.51	71.86
OTHER	63.62	62.58	63.02
PLACE	73.30	78.78	75.89
REL	82.99	83.93	83.33

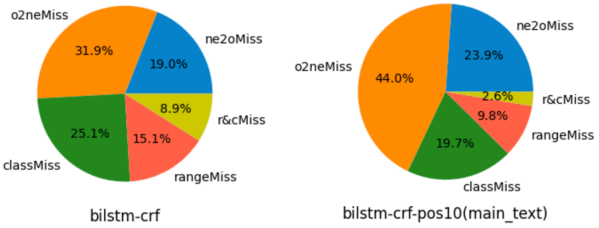

図 1　あらすじと本文での最良性能モデルの抽出ミス割合の比較

- 人物情報タグの種類は正確だが，抽出範囲を間違えている (rangeMiss)
- 人物情報を一部含んでいるが，人物情報タグの種類と抽出範囲を間違えている (r&cMiss)

最良性能モデルによる，あらすじデータに対する抽出ミスの割合と，本文データに対する抽出ミスの割合を図 1 に示す．図 1 より，あらすじに対する抽出ミスと比べて，本文に対する抽出ミスは，人物情報ではない部分に人物情報タグを振ってしまう o2neMiss が多いことが確認できる．これは，「〜学校」などのような固有名詞として出てくる系列にはタグを付与しているが，単なる「学校」などの特定できない一般的な場所名にはタグを付与していないという訓練データに対するタグ付けの問題が影響していると考えられる．また，「主」という文字に対して「あるじ」と読むか「おも」と読むかでタグ付け対象か否かが変わることなども影響していると考えられる．また，ビクリやゴロリ，ガッなどの擬音が NAME（人名）として抽出されるミスが散見された．あらすじでは，短く端的に小説の内容を紹介する必要があるため，副詞的用法が少なく，結果として，本文にしか現れない用法にうまく対応できなかったと考えられる．

6　おわりに

本稿では，あらすじテキストで学習した系列ラベリングモデルを Web 小説の本文に適用し，訓練データおよび提案モデルの小説本文に対する抽出性能を検証した．性能評価の結果，BiLSTM-CRF に品詞・品詞細分類ベクトルを 10 次元ずつ付与したモデル（BiLSTM-CRF-pos10）が最良性能を得ることを確認した．また，あらすじを対象とした人物情報抽出の最良性能と比較した結果，MF（性別表現）と OTHER（その他の人物情報）については，本文を対象とした場合の抽出性能の方が高くなることを確認した．他のラベルにおいては，AGE（年齢表現）の場合は軽微に，その他のラベルについては大幅に抽出性能が低下した．エラー分析の結果，本文を対象にした場合の抽出ミスは，人物情報ではない系列に，人物情報タグを付与するというミスが多いことを確認した．

あらすじデータを訓練データに利用する点については，あらすじには登場しない表現や構文が本文に出現する可能性があるため，すべてに対応できるとはいえない．しかし，一作品あたりの文数が本文よりも圧倒的に少なく，様々な作品を包含することで未知単語を減らすことができる可能性がある．本文データを訓練データに用いたモデルの性能が確認できていない現状，有用性を結論づけることは難しいため，今後検証を継続する必要がある．

参考文献

[1] Yuji Oka and Kazuaki Ando. Extraction of novel character information from synopses of fantasy novels in japanese using sequence labeling. In *Proceedings of the 34th Pacific Asia Conference on Language, Information and Computation*, 2020.

[2] Zhiheng Huang, Wei Xu, and Kai Yu. Bidirectional lstm-crf models for sequence tagging. In *arXiv: 1508.01991*, 2015.

[3] Xuezhe Ma and Eduard Hovy. End-to-end sequence labeling via bi-directional lstm-cnns-crf. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2016.

[4] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Dyer Chris. Neural architectures for named entity recognition. In *Proceedings of the North American Chapter of the Association for Computational Linguistics*, 2016.

[5] Shotaro Misawa, Motoki Taniguchi, Yasuhide Miura, and Tomoko Ohkuma. Character-based bidirectional lstm-crf with words and characters for japanese named entity recognition. In *Proceedings of the First Workshop on Subword and Character Level Models in NLP*, 2017.

[6] 日本語 Wikipedia エンティティベクトル, 2007. http://www.cl.ecei.tohoku.ac.jp/~m-suzuki/jawiki_vector/.

[7] Gustavo Aguilar, A. Pastor López-Monroy, Fabio A. González, and Thamar Solorio. Modeling noisiness to recognize named entities using multitask neural networks on social media. In *Proceedings of the North American Chapter of the Association for Computational Linguistics*, 2018.

[8] Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. Applying conditional random fields to Japanese morphological analysis. In *Proceedings of the Empirical Methods in Natural Language Processing*, 2004.