ニューラル日本語固有表現認識における格フレームの有効性検証

陰山宗一
筑波大学情報学群
s1913548@coins.tsukuba.ac.jp

駒田拓也　乾孝司
筑波大学大学院システム情報工学研究科
{komada@mibel., inui@}cs.tsukuba.ac.jp

1 はじめに

近年，固有表現認識（Named Entity Recognition, NER）課題に対して，LSTM や BERT 等，ニューラルネットワーク（Neural Network，NN）に基づく解析モデルが多く提案されている [1, 2]．これらのモデルでは，NN の特徴であるモデル設計の柔軟さを生かして，解析対象となる文書以外の外部情報を特徴量として容易に取り込むことができる．例えば，Lu ら [3] は解析対象文書の関連画像情報を外部特徴量として取り込むことで NER の性能向上に繋げている．

本研究では，NER 課題で標準的に用いられる Bi-LSTM-CRF モデルに格フレーム情報を外部情報として取り込むことを考え，その有効性を検証する．例文「太郎が筑波大学へ行く」において，述語「行く」のガ格には人間（PERSON）や組織（ORGANIZATION）等をあらわす固有表現が現れやすく，また，ヘ格には場所（LOCATION）をあらわす固有表現が現れやすい．このように述語と格関係にある要素の中には特定の固有表現クラスが現れやすいと考えられ，格フレームを参照することで，この固有表現の現れやすさに関する情報を解析モデルで利用する．

Bi-LSTM-CRF に格フレーム情報を取り込むにあたり，取り込むタイミングとして，LSTM 層よりも前側，後側およびその両方の 3 パターンが考えられる．そこで，3 パターンのうちのどれが組み込むタイミングとして最良であるかも同時に検証する．

2 格フレーム情報

2.1 京大格フレーム

河原ら [4] は約 100 億文の大規模 Web コーパスから格フレームを自動構築する手法を提案しており，構築したものを京大格フレームとして公開している[1)]．本研究では京大格フレームを外部情報として取り込む．

京大格フレームには表 1 のような情報がまとめられている．表 1 は「打つ」の例であり，述語「打つ」と関係をもつ項の情報だけでなく，それらの出現頻度が述語の用法ごとにまとめられている点が特徴である．

2.2 格フレーム内平均ベクトル（MVC）

格フレーム情報を NN モデルの特徴量として用いる場合，その情報をベクトルへ変換する必要がある．本研究では，山城ら [5] が提案した MVC（Mean Vector for Caseframe）特徴量を採用する．

MVC は以下のように計算される．

$$\bar{\boldsymbol{\phi}}_{cf_l^p(c)} = \frac{\sum_{w \in W_{cf_l^p(c)}} count(cf_l^p, c, w) \cdot \boldsymbol{\phi}_w}{\sum_{w \in W_{cf_l^p(c)}} count(cf_l^p, c, w)} \quad (1)$$

ある述語 p に対する京大格フレーム中の格フレーム群を $CF_p = \{cf_1^p, cf_2^p, ..., cf_m^p\}$ としたとき，l 番目の格フレーム cf_l^p がもつ格 c に対応する MVC を $\bar{\boldsymbol{\phi}}_{cf_l^p(c)}$ とする．ここで，格フレーム cf_l^p にまとめられている格 c の要素となる単語の集合を $W_{cf_l^p(c)}$ とする．また，$\boldsymbol{\phi}_w$ を単語 w の分散表現ベクトル，$count(cf_l^p, c, w)$ を単語 w が格フレーム cf_l^p 内の格 c の要素として出現する頻度とする．MVC は cf_l^p 内の格 c として出現する単語の分散表現の重み付き平均で計算される．例えば，表 1 の「打つ/うつ:動 1」のガ格 MVC は以下のように計算される．

$$\bar{\boldsymbol{\phi}}_{cf_{\text{動 1}}^{\text{打つ}}(\text{ガ})} = \frac{253 \cdot \boldsymbol{\phi}_{\text{心臓}} + 134 \cdot \boldsymbol{\phi}_{\text{姿}} + 134 \cdot \boldsymbol{\phi}_{\text{声}} + ...}{253 + 134 + 134 + ...} \quad (2)$$

3 Bi-LSTM-CRF（ベースライン）

本研究では，日本語 NER 課題において高い性能が報告されている Misawa ら [6] の，文字ベース Bi-LSTM-CRF をベースラインモデルとする．

Misawa らのモデルの入力となる各トークンは，解析対象となる文書に含まれる j 番目の文字の分散表現を $\boldsymbol{c_j}$，j 番目の文字を含む i 番目の単語の分散表現を $\boldsymbol{w_i}$ としたとき，それらを繋ぎ合わせた

$$\boldsymbol{x}_j = [\boldsymbol{w_i}; \boldsymbol{c_j}] \quad (3)$$

とする．LSTM 層では（4）式のように，ひとつ前のトークンの出力 $\overrightarrow{\boldsymbol{y}}_{j-1}$，$\overrightarrow{\boldsymbol{c}}_{j-1}$ と入力 $\boldsymbol{x}_j$ を受け，出力を

1） http://nlp.ist.i.kyoto-u.ac.jp/index.php

表 1　「打つ」の格フレーム例

格フレーム	ガ格	出現頻度	ヲ格	出現頻度	ニ格	出現頻度
打つ/うつ:動 1	心臓	253	手	19857	料理	1629
	姿	134	終止符	8978	生活	1580
	声	134	舌鼓	8967	歴史	741
	…	…	…	…	…	…

行う．

$$\overrightarrow{\boldsymbol{y}}_j, \overrightarrow{\boldsymbol{c}}_j = LSTM(\boldsymbol{x}_j, \overrightarrow{\boldsymbol{y}}_{j-1}, \overrightarrow{\boldsymbol{c}}_{j-1}) \tag{4}$$

そして，順方向の LSTM の出力 $\overrightarrow{\boldsymbol{y}}_j$ と逆方向の LSTM の出力 $\overleftarrow{\boldsymbol{y}}_j$ を繋ぎ合わせた

$$\boldsymbol{y}_j = [\overrightarrow{\boldsymbol{y}}_j; \overleftarrow{\boldsymbol{y}}_j] \tag{5}$$

が双方向（Bidirectional）LSTM の出力となる．各 $\boldsymbol{y}_j$ は続く CRF 層に渡され，NE ラベルの予測を行う．

4　提案手法

4.1　提案手法 1：前側挿入

1 つ目の提案手法は，LSTM 層の手前で MVC で表現された格フレーム情報を取り込む手法である．具体的には以下の式で示すゲート機構を導入することで，解析対象文書から得られる言語情報と外部情報である格フレーム情報を組み合わせる．なお，このゲート機構は Lu[3] らが提案した Modulation Gate からパラメータを減らして簡略化したものである．

$$\beta = \sigma(W_w \boldsymbol{h}_j + U_w \boldsymbol{v}_j + b_w) \tag{6}$$

$$\boldsymbol{m} = tanh(W_m \boldsymbol{h}_j + U_m \boldsymbol{v}_j + \boldsymbol{b}_m) \tag{7}$$

$$\hat{\boldsymbol{h}}_j = \beta \cdot \boldsymbol{h}_j + (1 - \beta) \cdot \boldsymbol{m} \tag{8}$$

このゲートにおいて，$\boldsymbol{h}_j$ に

$$\boldsymbol{h}_j = \boldsymbol{x}_j = [\boldsymbol{w}_i; \boldsymbol{c}_j] \tag{9}$$

のように文書由来の単語と文字データを割り当てる．また，入力文書を格解析した結果，単語 $\boldsymbol{w}_i$ が述語 p と格 c の関係にある場合は，$\boldsymbol{v}_j$ に

$$\boldsymbol{v}_j = \bar{\boldsymbol{\phi}}_{\boldsymbol{w}_i, cf_l^p(c)} \tag{10}$$

のように格フレーム由来の MVC を割り当てる．格関係にない単語については MVC 情報の代わりにゼロベクトルを割り当てる．このようにして格フレーム情報を取り込んだ $\hat{\boldsymbol{h}}_j$ を（3）式の代わりに LSTM への入力とする．本稿ではこの提案手法を後述の提案手法と区別するために「前側挿入」と呼ぶ．ここで MVC を $\bar{\boldsymbol{\phi}}_{\boldsymbol{w}_i, cf_l^p(c)}$ と表記している理由は，$\bar{\boldsymbol{\phi}}_{cf_l^p(c)}$ が $\boldsymbol{w}_i$ に対応した MVC であることを示すためである．前側挿入モデルのその他の構成はベースラインモデルと同じである．

4.2　提案手法 2：後側挿入

2 つ目の提案手法は，LSTM 層の直後に格フレーム情報を取り込む手法である．4.1 節で説明したゲートにおいて，$\boldsymbol{h}_j$ に

$$\boldsymbol{h}_j = \boldsymbol{y}_j \tag{11}$$

のように j 番目の LSTM の出力 $\boldsymbol{y}_j$ を割り当てる．また，前側挿入の（10）式と同様に $\boldsymbol{v}_j$ には MVC を割り当てる．このようにして格フレーム情報を取り込んだ $\hat{\boldsymbol{h}}_j$ を CRF 層へ渡す．本稿ではこの提案手法を「後側挿入」と呼ぶ．

4.3　提案手法 3：両側挿入

3 つ目の提案手法は，前側挿入と後側挿入によって 2 つのゲートを同時に適用する手法である．本稿ではこの提案手法を「両側挿入」と呼ぶ．両側挿入モデルの概要図を図 1 に示す．

5　評価実験

5.1　実験設定

実験には，拡張固有表現タグ付きコーパス (ENE コーパス)[7] を用いた．このコーパスのうち，出現頻度の高い 6 つの固有表現クラス（PRODUCT，NUMBER，LOCATION，TIME，ORGANIZATION，PERSON）を解析対象とした．表 2 に実験データの基本情報を示す．

本研究では Misawa ら [6] のモデルをベースラインモデルとして採用しているが，彼らの実験とは以下の点で実験データの設定が異なる．Misawa らは ENE コーパスに収録されているデータの内，新聞記事のみを解析対象にしているが，本研究では全ジャンルのデータを解析対象にしている．また，Misawa らは 4 つの固有表現クラス（PRODUCT，LOCATION，ORGANIZATION，TIME）を解析対象として選択したが，本研究では，格フレーム情報との相性が良いと考

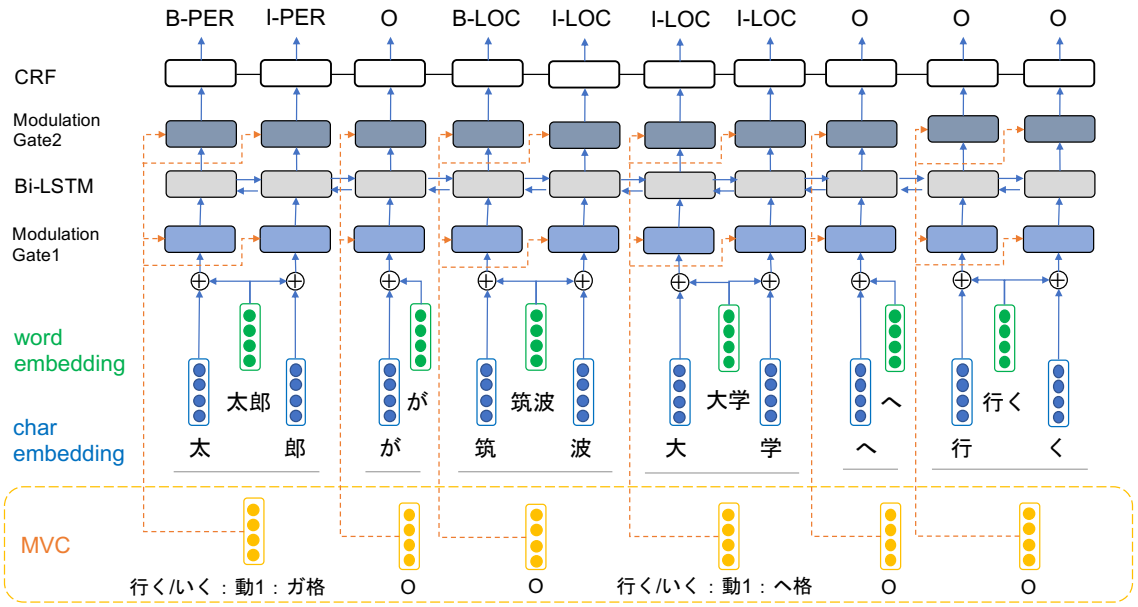

図 1　提案モデル（両側挿入モデル）の概要

表 2　実験データ

	訓練	開発	評価	All
文書数	7,383	901	900	9,184
PRODUCT	76,956	10,130	9,631	96,717
NUMBER	60,469	6,633	7,876	74,978
LOCATION	45,601	5,704	5,207	56,512
TIME	43,001	5,354	5,477	53,832
ORGANIZATION	34,774	4,115	4,028	42,917
PERSON	32,844	4,297	3,989	41,130
ALL	293,645	36,233	36,208	366,086

えられる PERSON クラスを含めたいため，上記で述べた 6 つの固有表現クラスを解析対象とした．

モデルパラメータは Misawa ら [6] にならい設定した．単語分散表現の次元数は 500，文字分散表現の次元数は 50，LSTM のユニット数は 300，バッチサイズは 60 とし，最適化には Adam[8] を用いた．単語，文字分散表現は，BCCWJ[9] と毎日新聞の記事 11 年分（1991-2000 年，2004 年）を用いて GloVe[10] で作成した．MVC 作成に必要となる単語分散表現も同様の手法を用いた．本実験では，MVC の作成に用いた京大格フレームとの対応をとるため，Misawa らとは違い，分かち書きには Juman++[2] を用いた．また，格解析には KNP[3] を用いた．

固有表現のチャンク表現には BIO モデル [11] を用い，完全一致以外は誤りとして，F1 値で解析結果を評価した．

5.2　実験結果

実験結果を表 3 に示す．表の各列における最良値をボールドで示す．結果より，４つのモデルの中では後側挿入が最も高い F1 値となった．F1 値に対する文書単位での並べ替えを行う並べ替え検定を行った結果，後側挿入とベースラインモデルの間には有意水準 0.01 で有意な差があることを確認した．このことから，取り込み方にもよるが，日本語 NER 課題に対して格フレーム情報が有効に作用することが確認できた．一方，前側挿入と両側挿入はベースラインモデルよりも性能が低くなった．MVC 自体は文脈情報から切り離した情報であるため，格フレーム情報を LSTM 層に通過させる場合（前側挿入と両側挿入）と通過させない場合（後側挿入）で，性能差があらわれたと考えられる．

後側挿入での解析結果の例を図 2 および図 3 に示す．これらはベースラインモデルでは正しく予測できなかったが，後側挿入では正しく解析できるようになった例である．例えば，図 2 の事例は，漢数字を含む文字列の例である．ベースラインモデルでは漢数字が含まれている文字列に対して NUMBER クラスを予測してしまう誤りがあるが，後側挿入ではそのような事例の解析が改善されていた．また，図 3 のように，ベースラインモデルでは Other クラスと予測する未抽出誤りがあるが，後側挿入ではそのような事例でも改善がみられた．

2）　http://nlp.ist.i.kyoto-u.ac.jp/?JUMAN%2B%2B
3）　http://nlp.ist.i.kyoto-u.ac.jp/?KNP

表 3　実験結果（各モデルの比較，F1 値）

	PRO	NUM	LOC	TIME	ORG	PER	ALL
ベースライン	69.56	81.74	86.58	88.44	75.28	87.66	80.15
前側挿入	64.78	79.81	85.37	86.80	72.61	86.12	77.48
後側挿入	**71.61**	**82.25**	**86.64**	**88.54**	**76.68**	**88.62**	**81.20**
両側挿入	67.09	80.94	85.55	85.67	75.78	87.24	78.97

表 4　MVC 情報の有無に注目した性能比較（ベースラインと後側挿入）

		PRO	NUM	LOC	TIME	ORG	PER	ALL
MVC 付与無し事例	ベースライン	68.67	81.19	86.32	89.30	76.26	86.85	80.10
	後側挿入	70.50↑	81.91↑	86.22↓	**89.34**↑	**77.76**↑	88.37↑	81.16↑
MVC 付与有り事例	ベースライン	70.58	82.54	87.24	87.55	73.76	88.72	80.24
	後側挿入	**72.85**↑	**82.74**↑	**87.66**↑	87.73↑	75.05↑	**89.01**↑	**81.27**↑
MVC 付与率		0.47	0.39	0.33	0.49	0.40	0.44	0.42

5.3　考察

提案手法において，MVC 情報は入力文書中の述語と格関係をもつ一部の単語のみに付与される．例えば，例文「北イタリアのラベンナに住むごく普通の市民」における「ラベンナ」は「住む」とニ格の関係をもつが，「北イタリア」は格関係をもたない．そこで，評価データ中の事例群を MVC が付与できたものと付与できなかったものに二分割し，ベースラインモデルと後側挿入それぞれの F1 値を再評価した．その結果を表 4 に示す．MVC が付与できた割合を表の最下段に示している．また，表中の MVC 付与無し，MVC 付与有りのそれぞれの事例について，ベースラインモデルからみた後側挿入の性能の上昇（↑）と下降（↓）を矢印で示す．

表から，まず，MVC が付与できない事例よりも MVC が付与できる事例の方が，ベースラインモデル，後側挿入共に性能が高いことがわかる．MVC 情報を組み込んでいる後側挿入だけでなくベースラインモデルにおいても性能差が生じていることから，格フレーム情報が得られるような事例群はそうでない事例群に比べて解析が容易な事例が集まっていることが示唆される．

また，表中の矢印の向きに注目すると，多くの固有表現クラスでは後側挿入がベースラインモデルよりも改善されていることがわかるが，MVC 付与率の低い LOCATION クラスでは MVC 付与無し事例群において性能が下がっている．この群の事例を観察すると，LOCATION クラスでは，以下の例文の下線部分のように述語と直接的な格関係をもたないため MVC が付与されない事例が目立っていた．このような事例に関して，MVC 情報の付与方法の改良については今後の課題である．

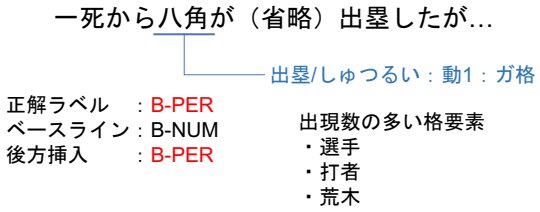

図 2　解析結果の例 1

図 3　解析結果の例 2

- 2 年前の世界選手権（カナダ・ハミルトン）で、（省略）を獲得した日本がお家芸の座を守れるか。
- 米国向けの輸出割合が（省略）まで低下している一方で…

6　おわりに

本研究では日本語 NER 課題おける格フレーム情報の有効性を検証した．評価実験の結果，Bi-LSTM-CRF モデルに格フレーム情報を取り込んだ提案手法（後側挿入）は，格フレーム情報を取り込まない場合に比べて 1 ポイント以上高い F1 値を示した．

今後は，LSTM 系モデルよりも高い性能が報告されている BERT 等の Transformer 系モデルをベースラインモデルとして同様の検証を実施し，格フレーム情報の日本語 NER 課題への有効性を評価する予定である．

謝辞　本研究の一部は JSPS 科研費 JP18K11982 の助成を受けたものです．

参考文献

[1] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 260–270, 2016.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[3] Di Lu, Leonardo Neves, Vitor Carvalho, Ning Zhang, and Heng Ji. Visual attention model for name tagging in multimodal social media. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1990–1999, 2018.

[4] 河原大輔, 黒橋禎夫ほか. 高性能計算環境を用いた web からの大規模格フレーム構築. 情報処理学会研究報告自然言語処理 (NL), Vol. 2006, No. 1 (2006-NL-171), pp. 67–73, 2006.

[5] 山城颯太, 西川仁, 徳永健伸. 大規模格フレームによる解候補削減を用いたニューラルネットゼロ照応解析. 自然言語処理, Vol. 26, No. 2, pp. 509–536, 2019.

[6] 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*, pp. 97–102, 2017.

[7] 橋本泰一, 中村俊一. 拡張固有表現タグ付きコーパスの構築. 言語処理学会　第 16 回年次大会　発表論文集（2010 年 3 月）, pp. 916–919, 2010.

[8] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[9] Kikuo Maekawa, Makoto Yamazaki, Toshinobu Ogiso, Takehiko Maruyama, Hideki Ogura, Wakako Kashino, Hanae Koiso, Masaya Yamaguchi, Makiro Tanaka, and Yasuharu Den. Balanced corpus of contemporary written japanese. *Language resources and evaluation*, Vol. 48, No. 2, pp. 345–371, 2014.

[10] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543, 2014.

[11] 黒橋禎夫. 改訂版 自然言語処理. 一般財団法人 放送大学教育振興会, 2019.