Teacher-Student 学習を利用したラベル誤りを含むデータにおける固有表現認識の性能向上

田川裕輝 [1] 中野騰久 [1] 尾崎良太 [1] 谷口友紀 [1] 大熊智子 [1] 鈴木裕紀 [2] 木戸尚治 [2] 富山憲幸 [2]
[1] 富士フイルム株式会社 [2] 大阪大学大学院医学系研究科
{yuki.tagawa,norihisa.nakano,ryota.ozaki,tomoki.taniguchi,
tomoko.ohkuma}@fujifilm.com
{y-suzuki,kido,tomiyama}@radiol.med.osaka-u.ac.jp

概要

情報抽出タスクの一つに固有表現認識（NER）がある．医療などの特定の分野を対象に NER の学習データを作成する場合，ラベル付け作業には対象分野の専門知識が必要となり，作業者間で解釈の違いや迷いが生じてしまう．その結果，ラベルにノイズを含むデータが作成され，モデルの性能劣化につながる．本研究ではアンサンブルと Teacher-Student 学習（TS 学習）を利用したラベルノイズを含む学習データにおける NER の学習方法を提案する．4 種類のラベルノイズを定義し，疑似的に作成したノイズを含むデータを用いてその含有率ごとの性能評価実験を行った．その結果，提案手法がノイズの影響を緩和し性能向上に寄与することが確認できた．

ラベル系列	区域	O	区域	O	病変	O		変化	O
トークン系列	右肺	の	S1	に	結節	あり。前回と比較して		著変	なし。

(a) 正解データ

ラベル系列	区域	O	区域		病変	O		変化	O
トークン系列	右肺	の	S1に		結節	あり。前回と比較して		著変	なし。

(b) **Span-shift** NE に対するスパンがずれているデータ

ラベル系列	区域	O	区域	O	病変	O		O	O
トークン系列	右肺	の	S1	に	結節	あり。前回と比較して		著変	なし。

(c) **Entity-missing** NE が抜け落ちているデータ

ラベル系列	区域	O	区域	O	病変	O	性状	O	変化	O	
トークン系列	右肺	の	S1	に	結節	あり。前回と	比較	して	著変	なし。	

(d) **Over-labeling** NE でないスパンにラベルが付いたデータ

ラベル系列	区域	O	区域	O	病名	O		変化	O
トークン系列	右肺	の	S1	に	結節	あり。前回と比較して		著変	なし。

(e) **Label-swapping** 異なるラベルが付いているデータ

図 1　NER におけるラベルノイズの例

1　はじめに

固有表現認識（NER）とはテキストから固有表現（NE）に当たるスパン（NE の開始，終了位置）とスパンに対する NE のラベルを推定する技術である．

近年では深層学習の発展に伴い，学習データを大量に用意し，モデルを学習させる手法 [1, 2] が多い．しかし，ラベルにノイズが含まれている場合，深層学習を利用した手法ではノイズに対しても過学習し，性能劣化につながる問題が指摘されている [3]．

図 1 に医療レポートを例として NER におけるラベルノイズの例を示す．図 1b では“S1”が適切なスパンであるが，実際には“S1 に”というスパンに対してラベルが付いている．このようなノイズを含んだデータが作成される問題は避けることが難しい．

学習データにノイズが含まれている場合の対策にモデルのアンサンブルがある．アンサンブルとは複数の異なるモデルの推論結果を統合し，最終出力を得る手法である．単一のモデルを学習させるとノイズに対して過学習してしまうが，異なるデータやシード値で学習された複数のモデルの出力を統合することで，ノイズの影響を緩和できることが知られている [4]．しかし，推論時に複数のモデルで推論するため，最終出力を得るまでに時間がかかるという問題がある．特に，即時性の必要なアプリでは推論時に複数のモデルを利用する手法は扱いづらい．

本研究では Teacher-Student 学習 （TS 学習）[5] を利用したラベルノイズを含む学習データでの NER の学習方法を提案する．提案手法の概要図を図 2 に示す．提案手法は訓練セットを複数のサブセットに分割し，各サブセットで複数の教師モデルを訓練する．次に，各教師モデルの出力分布を統合した分布に近づけるように単一の生徒モデルを訓練する．提案手法によりノイズの影響が緩和された単一のモデルが得られ，複数のモデルで推論するアンサンブルと比べて最終出力を得るまでの時間が抑えられる．

本研究の貢献は以下である．i) ラベルノイズを含む学習データにおける NER の性能向上を目指し，アンサンブルと TS 学習を組み合わせた手法を提案する．ii) 4 種類のラベルノイズを疑似的に付与したデータセットにおいて提案手法は高い性能であり，ノイズの影響を緩和できることを確認する．

2 関連研究

NE のラベル付けは難しく，ラベルノイズが混ざったデータが作成されてしまう [6, 7]．NER では大量のデータで学習する手法が多く提案され，高い性能が報告されている [1, 2] が，ノイズが含まれたデータで学習すると，モデルはノイズに対して過学習し，性能が劣化することも知られている [3]．

この問題に対して，Cross-Weigh [6] ではモデルの予測ラベル系列と正解ラベル系列が一致しない場合，その学習サンプルはノイズを含むと推定し，損失値に低い重みを掛けることで，ノイズを含む学習サンプルの学習への影響を小さくしている．この手法はサンプル単位で重み付けするため，同一サンプル内にノイズを含む NE，含まない NE が混ざっている場合に区別して扱えない．一方で，提案手法ではサンプル単位ではなく，複数の教師モデルが予測した各トークンに対するラベル確率分布を利用することで，トークン単位でノイズの影響を抑えながらモデルを学習することができる．

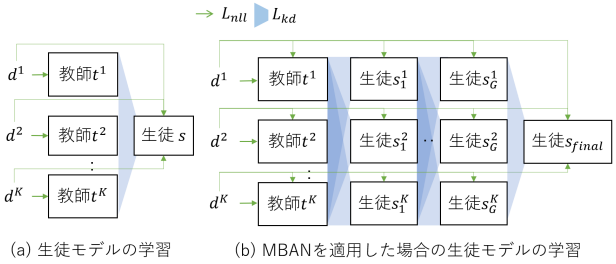
他にもノイズを含む学習データにおける NER の研究 [8, 9, 10] があるが，これらは一部の NE が O ラベルとなるノイズのみを対象としている．実際に NER のラベルノイズを調査した結果，スパンの間違いなども確認された．そこで，本研究では NER のラベルノイズを 4 種類に分類し，各ノイズを疑似的に付与した学習データを用意し，性能を比較する．

3 提案手法

提案手法では複数の教師モデルを訓練した後，それぞれの教師モデルの出力に近づけるように単一の生徒モデルを訓練し，最終的な推論に用いるモデルを獲得する．また，モデルには BERT-CRF を用いる．

3.1 生徒モデルの学習

図 2(a) は生徒モデルの学習方法を示している．まず，学習データ $D = \{(x_i, y_i)\}_{i=0}^{N}$ をランダムに分割し，複数のサブセット $\{d^k\}_{k=1}^{K}, d^k = \{(x_h^k, y_h^k)\}_{h=0}^{H}, K \geq 2$ を作成する．d^k を用いて k 番目の教師モデル t^k を



(a) 生徒モデルの学習　　(b) MBANを適用した場合の生徒モデルの学習

図 2 提案手法の概要図

訓練する．訓練には以下の損失関数 L_{tea^k} を用いる．

$$L_{tea^k} = \sum_{h=0}^{H} L_{nll}(x_h^k, y_h^k), \tag{1}$$

$$L_{nll}(x, y) = -\log p(y|x), \tag{2}$$

ここで，L_{nll} とはトークン系列 x に対するラベル系列 y の負の対数尤度関数である．x_h^k と y_h^k はそれぞれ k 番目のサブセット d_k 中の h 番目のトークン系列とそれに対応する NE ラベル系列である．

次に，訓練した K 個の教師モデル $\{t_k\}_{k=1}^{K}$ の出力分布を統合した分布に近づけるように，以下の損失関数 L_{stu} を用いて生徒モデル s を訓練する．

$$L_{stu} = \sum_{i=0}^{N} \{\alpha \cdot L_{nll}(x_i, y_i) + \beta \cdot L_{kd}(\tilde{p}_i, q_i)\}, \tag{3}$$

$$\tilde{p}_i = \sum_{k=1}^{K} w_i^k \cdot p_i^k, \tag{4}$$

ここで，L_{kd} とは x_i に対する各教師モデル t^k の各トークンに対する出力分布 $p_i^k \in \mathbb{R}^{J \times C}$ を統合した分布 $\tilde{p}_i \in \mathbb{R}^{J \times C}$ と生徒モデル s の出力分布 $q_i \in \mathbb{R}^{J \times C}$ の差異を表す尺度であり，KL-Divergence を利用する．J とはトークン系列長であり，C とは NE ラベル数である．p_i^k と q_i は Forward-Backward アルゴリズムにより算出する．また，w_i^k とは i 番目の学習データ x_i を処理する際の k 番目の教師モデル t^k に対応する重みである．α と β は L_{nll} と L_{kd} のバランスをとるための重みパラメータである．

3.2 教師モデルに対する重みの推定

複数の教師モデルの出力分布を統合する際の単純な方法として平均が挙げられる．要するに $w_*^k = \frac{1}{K}$ である．一方で，生徒モデル学習時に学習するサンプルのトークン系列 x_i に対して，その系列と類似したトークン系列を含むサンプルで学習した教師モデルは，その他の教師モデルと比べて，x_i に対して適切なラベル確率分布を推定できると考えられる．そこで，x_i と各教師モデル t^k の学習データであるサブ

セット d^k との類似度を測り，その類似度を重み w_i^k として用いる．トークン系列間の類似を測る尺度には BERTScore[1] [11] を利用する．

3.3 Multiple Born-again Networks の導入

Born-again Networks [12]（BAN）とは教師モデルを基に訓練された生徒モデルが次の世代では教師モデルに入れ替わり，別の生徒モデルを学習する機構である．通常，BAN は単一の教師モデルと生徒モデルを交互に入れ替えながら学習を進めていく．本研究では BAN を複数のモデルで学習を進めるように拡張した Multiple Born-again Networks（MBAN）を提案する．MBAN を適用した提案手法の概要図を図 2(b) に示す．MBAN は複数の教師モデルを基に複数の生徒モデルを入れ替えながら訓練し，最後に単一の生徒モデルを訓練する．事前に定義した世代 G のうち，g 世代目 $(1 \leq g \leq G)$ の k 番目の生徒モデル s_g^k は以下の損失関数 $L_{stu_g^k}$ を用いて訓練する．

$$L_{stu_g^k} = \sum_{h=0}^{H} \{\alpha \cdot L_{nll}(x_h^k, y_h^k) + \beta \cdot L_{kd}(\tilde{p}_{h,g-1}, q_{h,g}^k)\}, \tag{5}$$

ここで，$\tilde{p}_{h,g-1} \in \mathbb{R}^{J \times C}$ とは $g-1$ 世代目の各生徒モデル s_{g-1}^* の各トークンに対する出力分布 $p_{h,g-1}^k \in \mathbb{R}^{J \times C}$ を統合した分布であり，$q_{h,g} \in \mathbb{R}^{J \times C}$ とは生徒モデル s_g^k の出力分布である．次世代 $g+1$ では生徒モデル $\{s_g^k\}_{k=1}^K$ が教師となり，異なる生徒モデル $\{s_{g+1}^k\}_{k=1}^K$ を訓練する．最後に生徒モデル $\{s_G^k\}_{k=1}^K$ を教師として，推論のためのモデルを（3）式を用いて訓練する．

4 実験

提案手法の有効性を検証するため，日本語読影レポートデータセットを用いる[2]．まず，このデータにノイズを疑似的に付与したデータを作成する．次に，そのデータでモデルを学習し，ラベルノイズを含まないテストセットに対する性能を比較する．

4.1 ラベルノイズの種類

本研究では以下の 4 種類のノイズを定義する．また，各ノイズの例を図 1b，1c，1d，1e に示す．

- **Span-shift** NE のスパンが正解とは異なる位置にずれたノイズである．ランダムに NE を選択し，その NE の境界を 1 文字ずらすことにより，ノイズを付与した．ずらす際には NE の右，または左の境界を，増やす，または減らすかの 4 種類の組み合わせの内，ランダムに選択した．[3]
- **Entity-missing** NE のスパンに対して誤って O ラベルが付いたノイズである．ランダムに NE を選択し，その NE のラベルを O ラベルに置換することでノイズを付与した．
- **Over-labeling** NE でないスパンに対して誤って NE ラベルが付いたノイズである．O ラベルが付いている単語のうち，ランダムに選択された単語に，ランダムに選択した O ラベル以外のラベルを付けることでノイズを付与した．
- **Label-swapping** NE のスパンに対して正解とは異なるラベルが付いたノイズである．ランダムに NE を選択し，その NE のラベルを正解のラベルと O ラベル以外のラベルに置換することでノイズを付与した．

4.2 比較手法

モデル[4]の学習方法として以下の手法を比較し，提案手法の有効性を検証する．

1. **Single** 単一のモデルを学習させる手法．
2. **Ensemble** 初期パラメータの異なる複数のモデルを学習し，推論時は各モデルの出力結果を多数決で統合したものを最終出力とする手法．
3. **CW** Wang ら [6] の手法．Encoder には提案手法と同じ事前学習モデルを利用する．
4. **BAN** 単一の教師モデルを学習させ，次世代ではその教師モデルに近づけるように異なる生徒モデルを学習させる手法．
5. **Ours(Ave)** 各サブセットで学習した複数の教師モデルの平均出力分布に近づくように生徒モデルを学習する手法．
6. **Ours(Sim)** 教師モデルの平均出力分布ではなく，3.2 節で説明した教師モデルの加重平均分布に近づくように生徒モデルを学習する手法．
7. **Ours(Ave)+MBAN** 手法 5 に MBAN を加えた手法．
8. **Ours(Sim)+MBAN** 手法 6 に MBAN を加えた手法．

また，Ensemble，CW，提案手法の訓練時のモデル数は 2 とした．

[1] BERTScore の F 値を利用した．利用した事前学習モデルについては付録を参照されたい．

[2] データセットの詳細に関しては付録を参照されたい．

[3] NE のスパンが他の NE と重複することのないように事前にノイズ付与の対象となる NE を選別し，ノイズを付与した．

[4] 本研究ではベースのモデルとして BERT-CRF を利用する．パラメータ，事前学習モデルについては付録を参照されたい．

表 1　Span-shift を付与したデータでの性能.

ノイズの割合	0.2	0.4	0.6	0.8	全体性能
Single	88.90	87.37	80.32	68.86	73.62
Ensemble	90.14	87.03	79.16	68.09	73.36
CW	90.20	87.99	81.36	69.84	75.31
BAN	**90.74**	87.59	**82.45**	70.53	75.88
Ours(Ave)+MBAN	90.37	87.92	81.42	**70.54**	**76.31**
Ours(Ave)	89.72	87.92	77.20	66.91	73.83
Ours(Sim)+MBAN	89.39	**88.75**	81.75	69.64	76.23
Ours(Sim)	89.39	86.31	80.20	66.92	73.00

表 2　Entity-missing を付与したデータでの性能.

ノイズの割合	0.2	0.4	0.6	0.8	全体性能
Single	88.59	80.43	53.89	12.18	65.00
Ensemble	88.88	83.25	54.85	12.69	65.50
CW	89.40	**85.86**	68.37	16.33	69.09
BAN	89.11	85.25	66.90	22.39	69.47
Ours(Ave)+MBAN	89.24	85.28	**71.93**	22.14	**70.69**
Ours(Ave)	88.76	85.69	**71.93**	22.14	69.19
Ours(Sim)+MBAN	**89.78**	85.47	65.45	**29.97**	70.57
Ours(Sim)	88.78	79.16	56.94	8.87	64.62

4.3　実験結果

各データセットで学習したモデルのテストセットに対する Micro-F1 値を表 1，2，3，4 に示す．**太字**は各列で最高性能を意味する．Ours(Ave)+MBAN，Ours(Sim)+MBAN とベースラインである Single，Ensemble を比較すると多くの場合で MBAN を導入した提案手法の性能が高い結果となった．また，全体性能を比較すると全てで MBAN を導入した提案手法の性能が最も高いことがわかる．この結果はノイズを含むデータでの学習において，MBAN を導入した提案手法が効果的であることを示唆している．

一方で，MBAN を導入していない Ours(Ave) や Ours(Sim) はベースラインと比べて性能が向上していない．提案手法は学習データを分割したサブセットで教師モデルを学習するため，1 世代目では教師モデルの学習が進んでおらず，生徒モデルの性能も向上しなかったと考えられる．特に，Ours(Sim) は教師モデルの加重平均分布を基に学習するため，高い信頼度が算出された教師モデルの学習が停滞している場合，生徒モデルは正解のラベル分布とは差異のある分布で学習することとなる．その結果，学習に悪影響を与え，性能が劣化したと考えられる．

表 5 に表 1，2，3，4 の結果から算出したノイズの割合を固定し，各ノイズを均等に含む場合の性能を示す．提案手法の性能はノイズの割合が 0.4 の場合を除き，その他手法と比べて有意な差を確認した．ノイズの割合が 0.4 の場合の性能は CW が最も高い．

表 3　Over-labeling を付与したデータでの性能.

ノイズの割合	0.2	0.4	0.6	0.8	全体性能
Single	90.93	90.74	90.00	87.08	78.21
Ensemble	90.95	90.36	89.57	85.74	77.79
CW	**91.38**	**90.95**	90.21	87.40	79.49
BAN	90.99	90.69	**90.59**	86.96	79.69
Ours(Ave)+MBAN	90.82	90.69	90.34	87.59	80.26
Ours(Ave)	90.89	90.57	90.04	85.65	78.60
Ours(Sim)+MBAN	91.31	90.72	89.88	**88.32**	**80.34**
Ours(Sim)	90.84	90.56	89.88	85.24	77.63

表 4　Label-swapping を付与したデータでの性能.

ノイズの割合	0.2	0.4	0.6	0.8	全体性能
Single	89.80	89.22	85.99	66.41	74.48
Ensemble	90.43	89.56	86.45	67.38	74.67
CW	90.16	89.00	86.74	72.13	76.48
BAN	89.70	**89.56**	86.71	71.89	76.77
Ours(Ave)+MBAN	**90.92**	88.38	86.34	**76.02**	**77.82**
Ours(Ave)	90.13	87.79	86.28	70.97	75.60
Ours(Sim)+MBAN	89.84	88.04	**87.00**	72.88	77.31
Ours(Sim)	89.75	88.41	85.48	68.56	74.27

表 5　ノイズの割合を固定し，各ノイズを均等に含む場合の性能. † は提案手法とその他手法の間に有意水準 0.05 で差があることを示している．検定方法は笹野ら [13] に従った.

ノイズの割合	0.2	0.4	0.6	0.8
Single	89.57	87.14	79.48	64.91
Ensemble	90.11	87.66	79.33	64.69
CW	90.29	**88.50**	82.45	67.36
BAN	90.14	88.33	82.53	68.01
Ours(Ave)+MBAN	**90.34**†	88.13	**83.05**†	**69.31**†

2 節で説明したように，CW はサンプル単位でノイズの影響を緩和する手法である．一方で，提案手法は教師モデルの各トークンに対する確率分布を利用して学習するため，トークン単位でノイズの影響を緩和する手法であり，サンプル単位でノイズの影響を緩和する機構は備えていない．そのため，両者を組み合わせることで更なる性能向上が期待できる．

また，表 1，2，3，4 の性能を比較すると，提案手法がその他手法より低い場合もある．今後の課題として，各手法がどのようなノイズに対して有効かなど，手法とノイズ間の関係の分析が挙げられる．

5　おわりに

本研究ではラベルノイズを含む学習データにおけるアンサンブルと TS 学習を組み合わせた NER の学習手法を提案した．また，ラベルノイズを付与したデータにおいて，提案手法が高い性能であることを確認した．今後の課題は各ノイズと各手法間での性能差の詳細な分析である．

参考文献

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[2] Xuezhe Ma and Eduard Hovy. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In **Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 1064–1074, Berlin, Germany, August 2016. Association for Computational Linguistics.

[3] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In **5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings**. OpenReview.net, 2017.

[4] David Opitz and Richard Maclin. Popular ensemble methods: An empirical study. **Journal of Artificial Intelligence Research**, 1999.

[5] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. **arXiv preprint arXiv:1503.02531**, 2015.

[6] Zihan Wang, Jingbo Shang, Liyuan Liu, Lihao Lu, Jiacheng Liu, and Jiawei Han. CrossWeigh: Training named entity tagger from imperfect annotations. In **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)**, pp. 5154–5163, Hong Kong, China, 2019. Association for Computational Linguistics.

[7] Frederick Reiss, Hong Xu, Bryan Cutler, Karthik Muthuraman, and Zachary Eichenberger. Identifying incorrect labels in the CoNLL-2003 corpus. In **Proceedings of the 24th Conference on Computational Natural Language Learning**, pp. 215–226, Online, November 2020. Association for Computational Linguistics.

[8] Kun Liu, Yao Fu, Chuanqi Tan, Mosha Chen, Ningyu Zhang, Songfang Huang, and Sheng Gao. Noisy-labeled NER with confidence estimation. In **Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 3437–3445, Online, June 2021. Association for Computational Linguistics.

[9] Zhanming Jie, Pengjun Xie, Wei Lu, Ruixue Ding, and Linlin Li. Better modeling of incomplete annotations for named entity recognition. In **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**, pp. 729–734, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[10] Stephen Mayhew, Snigdha Chaturvedi, Chen-Tse Tsai, and Dan Roth. Named entity recognition with partially annotated training data. In **Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)**, pp. 645–655, Hong Kong, China, November 2019. Association for Computational Linguistics.

[11] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In **International Conference on Learning Representations**, 2020.

[12] Tommaso Furlanello, Zachary Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. Born again neural networks. In Jennifer Dy and Andreas Krause, editors, **Proceedings of the 35th International Conference on Machine Learning**, Vol. 80 of **Proceedings of Machine Learning Research**, pp. 1607–1616, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

[13] 笹野遼平, 黒橋禎夫. 大域的情報を用いた日本語固有表現認識. 情報処理学会論文誌, Vol. 49, No. 11, pp. 3765–3776, nov 2008.

表 6　データセットの統計量

訓練サンプル数	1,533
開発サンプル数	327
テストサンプル数	321
サンプルあたりの平均ラベル数	12.38
サンプルあたりの平均文字数	77.06

表 7　固有表現ラベルとその定義

ラベル	定義
解剖区域	主には部位を示す表現.「中間層」,「末端」,「底部」,「一部」などの画像中の位置を表す名詞も含まれる.
病変	画像情報を客観的に観察して得られる病変情報「腫瘤」,「結節」など. 所見にはそれが「認める」,「認めない」など断定的な表現で記述されている.
病名	観察された病変から診断される病名「肺炎」,「肺癌」など.「疑われる」など断定的ではない表現で記述される.
変化	「著変」,「増大」など, 前回診断との比較から得られた観察結果.
形状変化	臓器や部位などの形状の変化に関する表現. 臓器に対する「腫大」,「萎縮」など.
計測値	計測された大きさや値に関する表現。「15mm」や「10 * 8mm 大」など.
計測項目	病変のサイズや造影剤の吸収量などの計測した項目に関する表現.
撮影条件	撮像方法, 撮像タイミングに関する表現.「単純 CT」,「早期相」など。
性状	病変/形状変化の状態や性質などの特徴に関する表現.「[性状] を伴う [病変]」といった文脈や「〜性」,「〜状」という表現で病変と複合して記述されることが多い.

A　データセットの詳細

本研究で利用した日本語読影レポートデータセットの統計量を表 6 に, ラベルの定義を表 7 に示す. 病変, 形状変化, 変化, 性状に関しては実際にそれが生じている（Positive）のか否か（Negative）の事実性が“認められる”や“なし”などの表現でレポートに記載される. この事実性も固有表現と同時に推定するために, NE ラベルに事実性を表す Positive と Negative を結合したもの (例えば, 病変 P, 変化 N といったもの) をラベルとして用意し, モデルを学習させた. また, NER の学習には NE の開始を意味する **B**egin, NE の連続を意味する **I**nside, NE ではないことを意味する **O**ther によって NE の境界を識別する BIO 方式を採用した.

B　モデルのパラメータについて

本研究ではベースモデルとして BERT-CRF を利用した. BERT-CRF の学習率は 5.0×10^{-5}, バッチサイズは 16, α は 0.5, β は 0.5 とした. これらのパラメータはいくつかの候補に対して開発セットに対する性能を基にグリッドサーチにより決定した. また, MBAN を適用した提案手法, Cross-Weigh, BAN の世代数 G は 1 から 5 を試し, 最も開発セットに対して性能が高い世代のモデルを評価のためのモデルとして選択した.

C　事前学習モデルについて

本研究では BERT-base [1] を事前学習モデルとして利用した. この BERT は約 686 万文からなる日本語読影レポートを用いて Masked Language Model で訓練したモデルである. また, 読影レポートを NEologd 辞書 [5] を組み込んだ MeCab で単語分割した後, 文字単位のサブワードに分割することで 3,852 の語彙を構築した.

[5] https://github.com/neologd/mecab-ipadic-neologd