複数の学習器による知識の蒸留を利用した読影所見用語認識の精度向上

田川 裕輝　　中野 騰久　　尾崎 良太　　西埜 徹
谷口 友紀　　大熊 智子　　中村 佳児

富士フイルム株式会社
{yuki.tagawa, norihisa.nakano, ryota.ozaki, toru.nishino,
tomoki.taniguchi, tomoko.ohkuma, keigo.nakamura}@fujifilm.com

1 はじめに

読影レポートとは放射線科の医師が CT や MRI などの医療画像を元に異常所見や疾患についての読影診断を自然言語によって記述したものである．このレポートの中には医師の知識や経験が表現されており，これを抽出すれば様々な用途に利活用することが期待できる．言語で表現された医師の知識を再利用可能な形にするにはレポートを構造化して疾患や臓器などの観点から記述内容を分析できるようにする必要がある．レポートから所見用語を抽出するための固有表現認識（Named Entity Recognition; NER）はレポート構造化の根幹となる技術である．

NER は情報抽出技術の一つであり，多くの研究が取り組まれている [1]．特に，近年では深層学習の発展に伴い，大量の教師データでモデルを訓練する手法が提案されている [2, 3]．一方で，読影レポートへのラベル付けは専門知識が必要であり，大量の教師データを用意するにはコストがかかる．そのため，我々は大量の教師データを用意せずに所見用語認識の精度向上を目指す．

限られた量の教師データで NER の精度向上を目指した研究に Lai ら [4] や Liang ら [5] の研究がある．これらの研究は限られた教師データで学習したモデルで教師無しデータにラベル付けすることで，弱教師データを生成し，学習データを水増ししている．

Lai らは教師データだけで訓練した BERT [6] (以下 Teacher と呼ぶ) と，Teacher が生成した弱教師データと教師データを混ぜて訓練した BERT (以下 Student と呼ぶ) の精度を比較し，Student が Teacher の精度を上回った場合，Teacher のパラメータを Student のパラメータに置き換えて，再び弱教師データを生成する手法を提案している．またこの処理を複数回繰り返すことで，高品質な弱教師データが生成できたことを報告している．このように，Teacher が学習した知識を Student の学習に利用することを知識の蒸留 [7] という．特に，Teacher と Student に同じモデル構造を利用したものは Self-Distillation [8] や Born-Again Networks [9] と呼ばれる．

図 1 読影レポートとアノテーションデータのサンプル．

生成された弱教師データの質が NER の精度に影響することを考えると，単一の Teacher から弱教師データを生成するのではなく，Teacher に複数のモデルを用意し，それぞれの出力を統合することで，より高品質な弱教師データの生成が期待できる．特に，NER において高い精度が報告されている BERT はドメイン適用するために対象ドメインのコーパスで継続学習するもの [10] や文字単位の語彙を扱うもの [1)] など様々なモデルが提案されている．このような特徴の異なるモデルを Teacher として組み合わせることで，更なる精度向上が期待できる [11, 12]．

本研究では複数のモデルによる知識の蒸留を利用した NER の学習手法を提案し，所見用語認識に適用する．実験では先行研究である Lai らの研究と比べ，提案手法が Micro-F1 評価において高い精度となることを確認する．

2 提案手法

図 2 に提案手法の概要図を示す．提案手法の学習フローを以下に説明する．

1. あるトークン系列 x に対して正解ラベル系列 y がアノテーションされた教師データ $S = \{(x_1, y_1), ..., (x_L, y_L)\}$ を利用して，モデルリ

[1)] https://github.com/cl-tohoku/bert-japanese

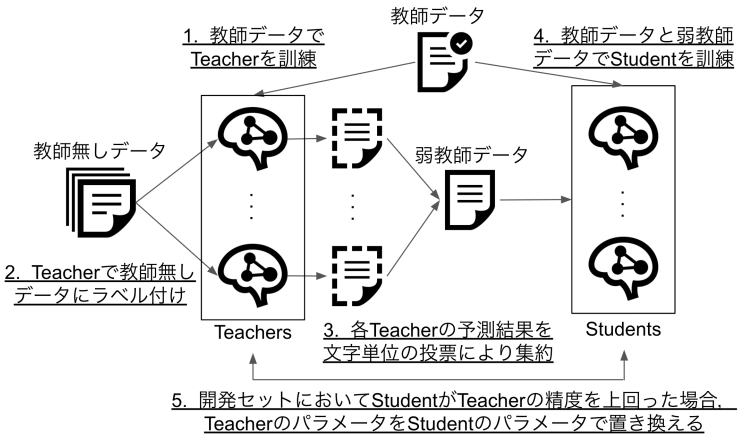

図 2　提案手法の概要図.

スト $F = \{f_{\theta_1}, ..., f_{\theta_i}, ..., f_{\theta_N}\}$ 中の N 個のモデルを訓練し，$Teachers = \{f_{\theta_{t_1}}, ..., f_{\theta_{t_i}}, ..., f_{\theta_{t_N}}\}$ を作成する．ここで f_θ とはパラメータ θ を持つ NER モデルである．

2. 教師無しデータ d のリスト $U = \{d_1, ..., d_j, ..., d_M\}$ に対して，各 Teacher モデルから出力 $O = \{(d_1, \hat{y}_{1,1}), ..., (d_j, \hat{y}_{i,j}), ..., (d_M, \hat{y}_{N,M})\}$ を生成する．ここで，$\hat{y}_{i,j}$ とは Teacher モデル $f_{\theta_{t_i}}$ によって教師無しデータ d_j に対して予測したラベル系列である．
3. 各 Teacher モデルの出力 O を文字単位の多数決で一つの出力結果に統合し，弱教師データ $P = \{(d_1, \tilde{y}_1), ..., (d_j, \tilde{y}_j), ..., (d_M, \tilde{y}_M)\}$ を生成する．
4. 教師データ S と弱教師データ P を混ぜたデータでモデルリスト F 中のモデルを訓練し，$Students = \{f_{\theta_{s_1}}, ..., f_{\theta_{s_i}}, ..., f_{\theta_{s_N}}\}$ を作成する．
5. $f_{\theta_{t_i}}$ と $f_{\theta_{s_i}}$ の開発データに対する精度を比較し，$f_{\theta_{s_i}}$ が $f_{\theta_{t_i}}$ の精度を上回っている場合，$f_{\theta_{t_i}}$ のパラメータ θ_{t_i} を $f_{\theta_{s_i}}$ のパラメータ θ_{s_i} に置き換えることで，$Teachers$ を更新する．
6. 2 から 5 までの処理をどの Teacher モデルも更新されなくなるまで，もしくは事前に定義した回数 I だけ繰り返す．
7. 各 Teacher モデルを利用してテストセットに対して予測する．
8. 各 Teacher モデルの予測結果を 3 と同じく文字単位の多数決により一つの予測結果に集約し，テストセットに対する予測結果を得る．

3 実験

表 1　データセットの統計量.

	訓練	開発	テスト	教師無しデータ
件数	1,000	500	500	10,000
平均文字数	20.85	20.79	22.62	36.32
平均用語数	3.40	3.46	3.59	-

3.1 データセット

本研究では全身を対象とした日本語読影レポートに所見用語ラベルをアノテーションし，データセットを構築した．データセットの統計量を表 1 に示す．表 1 の件数は読影レポート中の 1 行を 1 件としてカウントしている．また，教師無しデータは文字数が多いものから優先して 10,000 件を選択した．

ラベルの種類とその説明を表 2 に示す．図 1 に示すように，病変と病名，変化の表現はレポートではそれらが認められる (Positive; P)，認められない (Negative; N)，疑いや可能性がある (Suspicion; S) といった事実性も同時に記述される．そのため，所見用語ラベルと同時に事実性ラベルもアノテーションした．本研究では所見用語ラベルに事実性ラベルを結合したものをラベル [2] とし，事実性判定も所見用語認識タスクとして同時に解いた．また，BIO 方式 [3] を用いてラベルを付与した．

3.2 モデル

実験で利用するモデルを以下に説明する．

BERT　日本語 Wikipedia で学習した BERT [1]．語彙は NEologd 辞書 [4] を組み込んだ MeCab によ

[2] 例えば，病変_P，変化_N のようなラベルである．
[3] BIO 方式では固有表現の開始を表す Begin，同一種の固有表現の継続を表す Inside，どの固有表現にも当てはまらない Outside を利用し，固有表現の境界を認識する．
[4] https://github.com/neologd/mecab-ipadic-neologd

表 2　所見用語ラベルとその説明.

ラベル	説明
場所	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“すりガラス影”，“結節”等．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．
時間	具体的な日付や“前回”，“年一度”などの表現が含まれる．
モダリティ	“単純 CT”，“PET”，“造影剤”，などの検査に使用される器材や検査方法．
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．
数量	数詞と助数詞による具体的な寸法や“一部”，“全体”，“最大”などの数量表現．
程度	“明瞭な”，“わずかに”など程度を表す修飾語．状態や形状を含む．

り単語分割した後，Byte Pair Encoding (BPE) により構築している．

Char-BERT　日本語 Wikipedia で学習した文字単位の語彙を持つ BERT [1]．語彙は文字単位の分割処理により構築している．

UTH-BERT [13]　日本語の医療分野のコーパスで学習させた BERT．語彙は万病辞書 [5] を組み込んだ MeCab で単語分割した後，BPE により構築している．

Radio-BERT2K　約 97 万行の日本語読影レポートで学習した BERT．事前学習時は Masked Language Modeling (MLM) [6] で学習した．また，MLM にマスクする単語を単語単位で選択し，選択された単語に対応するサブワードを全てマスクする Whole Word Masking (WWM) を加えた．語彙は MeCab で単語分割した後，BPE により 2,000 の語彙を構築した．前処理として 5 文字以下の行は削除した上で事前学習した．

Radio-BERT4K　**Radio-BERT2K** の語彙数を 4,000 に変更したモデル．その他の設定は **Radio-BERT2K** と同じである．

Radio-AdaptaBERT　**BERT** を日本語読影レポートで継続学習した BERT．継続学習のタスクには WWM を加えた MLM を利用した．語彙は **BERT** と同じものを利用した．

実際に学習させるモデルはこれらのモデルの最終層に CRF [14] を追加したモデルである．学習時はこれらの事前学習モデルで入力系列をエンコードし，得られたベクトル表現から CRF により，正解ラベル系列を出力するようにモデル全体を学習する．

学習フロー 3 と 8 では同じデータに対する複数のモデルの予測結果を一つの結果に集約している．複数の予測結果を一つに結果に集約する手法としてトークン単位の多数決を利用する研究がある [11]．一方で，トークン (またはトークン系列) に対してラベル (またはラベル系列) を予測するモデルの予測結果を集約するには，全モデルでトークン分割の結果が同じでなければ，トークン単位の多数決を適用する事ができない．そこで，各モデルが予測したラベル系列を文字単位のラベル系列に変換し，全モデルで分割単位を揃えた上で文字単位の多数決により予測結果を集約した．

3.3　比較手法

実験で比較する手法を以下に説明する．

Single　教師データのみで学習するベースライン．

Self Distillation (SD)　Lai ら [4] の手法．一つのモデルを Teacher と Student として利用し，弱教師データを生成する．

Collaborative Distillation w/ Multiple Seeds (CDMS)　同じモデル構造で異なる初期シードを設定した複数のモデルを用意し，弱教師データを生成する提案手法．

Collaborative Distillation w/ Different Models (CDDM)　異なる構造を持つ複数のモデルを用意し，弱教師データを生成する提案手法．

各モデルのパラメータは **Single** 手法で開発セットに対して最高精度となるパラメータをグリッドサーチで決定した．また，I には 5 を設定した．

3.4　評価方法

手法の評価には Micro-F1 値を用いた．システムが出力したラベル系列と正解ラベル系列を比較し，所見用語の開始点，終点，所見用語ラベルが全て一致した場合に正解とした．

3.5　実験結果

表 3 にテストセットに対する Micro-F1 値を示す．**CDMS** はそれぞれのモデルを異なる 3 つのシードを用意し学習させた結果である．**CDDM** は **BERT**，**Char-BERT**，**Radio-AdaptaBERT** の 3 つのモデルを利

[5] http://sociocom.jp/~data/2018-manbyo/index.html

表 3　テストセットに対する Micro-F1 値．最高値には下線を引いている．

	BERT	Char-BERT	UTH-BERT	Radio-BERT2K	Radio-BERT4K	Radio-AdaptaBERT
Single	88.43	90.06	87.18	87.18	85.84	89.59
SD[4]	89.24	89.80	87.34	87.18	86.17	89.59
	BERT × 3	**Char-BERT** × 3	**UTH-BERT** × 3	**Radio-BERT2K** × 3	**Radio-BERT4K** × 3	**Radio-AdaptaBERT** × 3
CDMS	89.99	89.87	88.67	89.34	88.65	90.31
	BERT + **Char-BERT** + **Radio-AdaptaBERT**					
CDDM	90.44					

用した結果である．値を比較すると，**CDMS** では **Radio-AdaptaBERT** を利用し，学習した場合が最も高い値となった．また，他の **Single** や **SD** と比べても高い値となっていることから，複数のモデルを組み合わせて弱教師データを生成する提案手法は所見用語認識タスクに効果的であったといえる．特に，複数の異なるモデルを組み合わせる **CDDM** が最も高い性能となった．

3.6　分析

各モデルの予測にどの程度のばらつきがあるのかを分析するため，モデル間の予測結果の不一致の割合 [15] を算出した．モデルの予測したラベル系列を文字単位のラベル系列に変換した上で，予測結果の不一致の割合 d_{all} を算出した．さらに，どちらのモデルも所見用語ラベルを予測した事例の内，ラベルが不一致した割合 d_{label} と，予測した所見用語のスパンが不一致した割合 d_{span} [6] を算出した．また，3 つ以上のモデルの組み合わせに対しては全てのモデルの組み合わせに対して不一致の割合を算出し，その平均値を多様性尺度 D とした．

表 4 に **Single** 手法によって訓練されたモデルの予測結果から算出した多様性尺度 D を示す．全ての尺度において，同じモデル構造間の予測結果の値よりも，異なるモデル構造間の予測結果の値の方が高いことがわかる．D_{label} では **BERT** と **Char-BERT** の組み合わせが最も高い値となった．これらのモデルの違いは入力テキストをそれぞれ文字単位で分割するかサブワード単位で分割するかという点である．日本語は単語分割の境界が自明でないため，NER の予測結果は前段の単語分割の結果の影響を受けることが知られており [16]，異なる単語分割器を持つモデルを組み合わせたことが同じモデル構造でシードのみを変えて組み合わせた場合と比べて高い多様性を生んだと考えられる．また，D_{span} においては **Char-BERT** と **Radio-AdaptaBERT** の組み合わせが最も高い値となった．これらのモデルの違いは上述した単語分割方法の違いに加え，事前学習コーパスも異なる．**Radio-AdaptaBERT** は大量の読影レポートで継続学習している．この結果から事前学習時のコーパスの違いも多様性を生むと考えられる．

[6] どちらのモデルも O ラベルを予測した事例は除いた内，(O ラベル，B ラベル)，(O ラベル，I ラベル) の組み合わせで不一致した割合を算出した．スパンの不一致の割合を測るため，所見用語ラベルは考慮せず BIO ラベルのみで不一致を検出した．

表 4　教師無しデータに対する予測結果から算出した多様性尺度 D_{all}，D_{label}，D_{span}．w/ 3seeds とはモデルを異なる 3 つのシードで学習したことを表す．各尺度毎に最高値には下線を引いている．

モデル群	D_{all}	D_{label}	D_{span}
BERT w/ 3seeds	.054	.081	.014
Char-BERT w/ 3seeds	.050	.074	.014
Radio-AdaptaBERT w/ 3seeds	.044	.064	.014
BERT，**Char**	<u>.060</u>	<u>.086</u>	.017
Char，**Radio-Adapta**	.059	.085	<u>.018</u>
BERT，**Radio-Adapta**	.055	.082	.016
BERT，**Char**，**Radio-Adapta**	.058	.085	.017

このように異なるモデル構造を組み合わせることで予測結果に多様性が生まれることがわかった．多様性は複数のモデルを組み合わせて性能向上を測るアンサンブル学習において，性能を測る尺度として期待できるという分析がある [17]．表 3 の結果からも，多様性の低い異なるシードで学習したモデルを組み合わせる **CDMS** より，多様性の高い異なるモデルを組み合わせる **CDDM** のほうが高い精度となった．これらの結果から多様な予測をするモデルを組み合わせて弱教師データを生成する **CDDM** は所見用語認識の性能向上に寄与したと考えられる．

4　おわりに

本研究では所見用語認識の精度向上を目指し，複数モデルによる知識の蒸留を利用した NER の学習手法を提案した．実験ではベースラインと先行研究 [4] と比較し，提案手法が高い性能となった．

今後は前段の単語分割の結果や事前学習コーパスの違いなどが，後段の所見用語認識の精度に与える影響を分析したい．

参考文献

[1]Vikas Yadav and Steven Bethard. A survey on recent advances in named entity recognition from deep learning models. In *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 2145–2158, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics.

[2]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, San Diego, California, June 2016. Association for Computational Linguistics.

[3]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.

[4]Tuan Lai, Trung Bui, Doo Soon Kim, and Quan Hung Tran. A joint learning approach based on self-distillation for keyphrase extraction from scientific documents. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 649–656, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics.

[5]Chen Liang, Yue Yu, Haoming Jiang, Siawpeng Er, Ruijia Wang, Tuo Zhao, and Chao Zhang. Bond: Bert-assisted open-domain named entity recognition with distant supervision. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2020.

[6]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.

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

[8]Jianping Gou, Baosheng Yu, Stephen John Maybank, and Dacheng Tao. Knowledge distillation: A survey. *arXiv preprint arXiv:2006.05525*, 2020.

[9]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.

[10]Xiaochuang Han and Jacob Eisenstein. Unsupervised domain adaptation of contextualized embeddings for sequence labeling. 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. 4238–4248, Hong Kong, China, November 2019. Association for Computational Linguistics.

[11]Zornitsa Kozareva, Boyan Bonev, and Andres Montoyo. Self-training and co-training applied to spanish named entity recognition. In *Mexican International conference on Artificial Intelligence*, pp. 770–779. Springer, 2005.

[12]Pierre Lison, Jeremy Barnes, Aliaksandr Hubin, and Samia Touileb. Named entity recognition without labelled data: A weak supervision approach. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1518–1533, Online, July 2020. Association for Computational Linguistics.

[13]Yoshimasa Kawazoe, Daisaku Shibata, Emiko Shinohara, Eiji Aramaki, and Kazuhiko Ohe. A clinical specific bert developed with huge size of japanese clinical narrative. *medRxiv*, 2020.

[14]John Lafferty, Andrew McCallum, and Fernando CN Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001.

[15]David B Skalak, et al. The sources of increased accuracy for two proposed boosting algorithms. Citeseer.

[16]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, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[17]Kagan Tumer and Joydeep Ghosh. Theoretical foundations of linear and order statistics combiners for neural pattern classifiers.