Majority or Minority: 固有表現抽出における データの不均衡性に着目した損失関数の提案

根本颯汰 [1]　北田俊輔 [2]　彌冨仁 [1]
[1] 法政大学 理工学部　[2] 法政大学大学院 理工学研究科
{sota.nemoto.5s, shunsuke.kitada.8y}@stu.hosei.ac.jp　iyatomi@hosei.ac.jp

概要

多くの自然言語処理タスクはデータの不均衡の問題に直面しており，実用的な応用がなされている固有表現抽出もその一つである．固有表現抽出は抽出対象の固有表現以外のトークンすべてが ⑥ クラスとなるため，⑥ クラスが大多数を占める不均衡なデータとなっている．本論文では，固有表現抽出における不均衡性に着目した新たな損失関数 majority or minority loss (MoM loss) を提案する．提案手法の核となるアイディアは多数派のクラスである ⑥ クラスのトークンのみを計算対象した loss を従来のモデルの損失関数に追加するものである．実験を通じて MoM loss がマルチクラス，2 クラス分類問わず，言語非依存で性能向上に寄与することを確認した．

1　はじめに

多くの 自然言語処理 (natural language processing; NLP) タスクはデータの不均衡性の問題に直面しており， 固有表現認識 (named entity recognition; NER) もその一つである．具体的に，非固有表現である ⑥ クラスのサンプル数は CoNLL2003 データセット [1] においては，それ以外の固有表現クラスのサンプル数の合計の 4.6 倍，OntoNotes 5.0 データセット [2] においては 7.6 倍である．そのため結果として一般的な機械学習モデルを適用した場合，多数を占める ⑥ クラスへの適合が優先され，他の少数の固有表現クラスの予測性能が低下する恐れがある．したがってこの不均衡な問題に対策することは NER の予測性能の向上において普遍的で重要な課題である．

NER は細かく flat-NER [3] と nested-NER [4] に分けられる．flat-NER はテキスト内の各トークンに 1 つのラベルを割り当てるように訓練する系列ラベリングタスクとして定式化され， 各トークンの固有表現ラベルを予測するマルチ分類問題に帰着される．一方で nested-NER は固有表現の範囲が入れ子になっているタスクで，各トークンに対する質問応答タスクとしてみなして学習する機械読解 (machine reading comprehension; MRC) タスクとして定式化されている．例えば「[大谷翔平] はメジャーリーグで二刀流として活躍している.」に人名を抽出する場合は「本文中に含まれる人名は誰だ」という質問文に対して「大谷翔平」が抽出される．そのため， MRC タスクは質問文に対応したテキスト中の各トークンが抽出対象かどうかの 2 クラス分類問題に帰着される．系列ラベリングタスクで ⑥ クラスが多数派であったように，MRC タスクでも非固有表現のトークンが非常に多いため，こちらも不均衡な問題である．

不均衡性に対する手段として深層学習技術の親和性や導入の手軽さから，機械学習モデルの損失関数を設計する cost sensitive learning が現在の主流の一つである [5, 6, 7]. その中でも focal loss (FL) [5] や dice loss (DL) [7] は既存の有効的な手法として広く知られている．FL はコンピュータビジョン分野で効果が確認された損失関数で，物体検出タスクの背景と検出対象の前景との不均衡性に着目して設計された．NLP タスクにも応用され一定の成果を上げている [8]．また DL は NLP 分野のいくつかのタスクで効果が確認され，評価指標である F1 スコアに近いダイス係数を使用することで不均衡性に対処する損失関数である [7]．しかし 2 クラス分類に基づいて設計された FL や DL は 2 クラス分類問題として扱う MRC タスクには適用可能だがマルチクラス分類である系列ラベリングタスクに適用することは困難である．従って NER を系列ラベリングタスクとして解く際にも適用可能な不均衡性に対処できる損失関数が必要である．

本論文では，圧倒的多数である ⑥ クラスおよび，少数の各固有表現クラスの識別精度の両立を実現

する majority or minority loss (MoM loss) を提案する．MoM loss は元の機械学習モデルの損失に追加するシンプルかつ汎用的な手法で，多数派のクラスデータのみに対する任意の損失関数として定義される．本論文で扱う NER の系列ラベリングタスクにおいて MoM loss は正解ラベルが $\mathcal{O}$ クラスのトークンに対する損失として定義される．少数派クラスから多数派クラスへの誤識別は，当該少数派クラスの precision の大幅な低下をもたらす．提案する MoM loss は各クラスのデータ数に応じて学習係数や損失関数を調整するような従来の手法と異なり，多数派クラスの誤識別に対して直接ペナルティを課してそれを減らすように学習することで結果として少数派クラスの予測性能を大きく向上させ，全体の識別率の向上に繋げる．また提案する MoM loss は任意の損失関数に適用可能であるため，MRC タスクにおいても適用可能である．

2 MoM Loss

我々は表記方法を紹介した後に一般的な cross entropy (CE) について説明し，次にシンプルでありながら非常に効果的な提案手法 MoM loss について説明する．ここでの説明は我々の提案の主である系列ラベリングタスクとして解くことを想定する．

まず，入力テキスト X は $W = [W_1, W_2, \cdots, W_M]$ の長さ M のトークンに分割される．トークン W_i に対してラベル $T_i \in \mathbb{R}^{|\mathcal{T}|}$ を持つ．これは学習時のラベル $T = [T_1, T_2, \cdots, T_M] \in \mathcal{R}^{M \times |\mathcal{T}|}$ の集合 $\mathcal{T}$ (e.g., B-LOC, I-LOC, B-ORG, ...) として与えられる．各トークンのラベル付けには一般的によく使われる BIO 形式を採用する [9]．この形式を定義する文字はクラスの接頭辞である，先頭の固有表現を表す ($\mathcal{B}$)、それ以外の固有表現の ($\mathcal{I}$)、非固有表現の ($\mathcal{O}$) から成る．モデルは，各トークンについて 予測される確率 $P_i \in \mathbb{R}^{|\mathcal{T}|}$ を出力するように学習される．つまり系列ラベリングタスクは， 教師タグの系列 $T \in \mathbb{R}^{M \times |\mathcal{T}|}$ に近い確率の系列 $P \in \mathbb{R}^{M \times |\mathcal{T}|}$ を推定する．

2.1 Cross Entropy Loss $\mathcal{L}_{\text{CE}}$

CE $\ell_{\text{CE}}(T_{ij}, P_{ij}) = T_{ij} \log(P_{ij})$ はトークン W_i の one-hot に表現のラベル T_i と 予測確率 P_i 間の損失関数 $\mathcal{L}_{\text{CE}}(\cdot, \cdot)$ であり，以下のように定義される:

$$\mathcal{L}_{\text{CE}}(T, P) = -\frac{1}{M} \sum_{i=1}^{M} \sum_{j=1}^{|\mathcal{T}|} \ell(T_{ij}, P_{ij}). \quad (1)$$

T_{ij} と P_{ij} は T_i, P_i ベクトルの j 番目の要素である．

2.2 MoM Loss $\mathcal{L}_{\text{MoM}}$

NER では多数派の $\mathcal{O}$ クラスのサンプルが学習を圧迫し，他の固有表現クラスのサンプルの学習が十分にされず，固有表現クラスの識別精度の低下を招く．従来のデータ数に応じた損失関数の重み付けは一定の成果を上げているが，重みの最適化を正確に行わないと精度が低下する可能性があるため改善の余地が残されていた [5, 7]．

MoM loss の基本的な考え方は，$\mathcal{O}$ クラスの正解ラベル (i.e., $T_{ij} = \mathcal{O}$) に対してのみ損失を計算することである．そのため，非固有表現トークンを $\mathcal{O}$ クラスとみなすことで MRC タスクでも導入が可能となる．我々の MoM loss を以下のように定義する:

$$\mathcal{L}_{\text{MoM}}(T, P) = -\frac{1}{M} \sum_{i=1}^{M} \sum_{j=1;\ T_{ij}=\mathcal{O}}^{|\mathcal{T}|} \ell(T_{ij}, P_{ij}). \quad (2)$$

ここで ℓ は CE を含む任意の損失関数である．この MoM loss を従来のモデルの損失関数 $\mathcal{L}(\cdot, \cdot)$ に追加する．これによって MoM loss が $\mathcal{O}$ クラスのみの loss を計算するため，固有表現クラスと $\mathcal{O}$ クラスに対する疑似的な 2 段階分類を実現させた．

モデルの訓練時には、従来の CE に加えて MoM loss を加えた以下の損失を最小化する:

$$\mathcal{L}_{\text{total}}(T, P) = \lambda \cdot \mathcal{L}(T, P) + (1 - \lambda) \cdot \mathcal{L}_{\text{MoM}}(T, P), \quad (3)$$

ここでパラメータ λ は $\mathcal{L}(\cdot, \cdot)$ と $\mathcal{L}_{\text{MoM}}(\cdot, \cdot)$ の間のトレードオフを制御するハイパーパラメータである．この $\mathcal{L}_{\text{total}}$ は様々な分野でデータの不均衡に対して有効なマルチタスク学習 [10] の一種として考えることができる [11, 12]．

3 実験

本節では提案する損失関数である MoM Loss の効果を，flat-NER として系列ラベリングタスクと MRC タスクの 2 通りの方法でモデルを学習した．系列ラベリングタスクでマルチクラス分類における提案手法の有効性を示し，MRC タスクで 2 クラス分類に親和性のある先行研究と比較した．

3.1 データセット

我々は日英含む以下の 4 つのデータセットを使用した: CoNLL2003 [1], OntoNotes5.0 [2], 京大ウェブ文書リードコーパス [13], Stockmark-NER-wiki [14].

表 1 各損失関数の Sequence labeling タスクにおける実験結果

	CoNLL03 [1]			Ontonotes 5.0 [2]			KWDLC [13]			Stockmark NER Wiki [14]		
	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1
BERT + CE	90.16	91.86	91.00	87.41	89.07	88.23	70.92	73.96	72.41	77.32	81.04	79.13
BERT + FL	90.33	92.03	91.17	87.62	89.15	88.39	71.88	74.27	73.05	77.79	81.53	79.61
			(+0.17)			(+0.16)			(+0.64)			(+0.48)
BERT + MoM	90.41	92.27	**91.33**	87.39	89.84	**88.60**	72.54	74.13	**73.32**	78.13	81.61	**79.83**
			(+0.33)			(+0.37)			(+0.91)			(+0.70)

これらのデータセットの統計量をまとめたものを付録 A にて議論する．

NER データセットの不均衡性を示すため，$\mathcal{O}$ クラスのサンプル数 $N_{\mathcal{O}}$ と固有表現クラスのサンプル数 $N_{\text{non}\mathcal{O}}$ から不均衡率 ρ を以下のように計算する:

$$\rho_{\mathcal{O}} = N_{\mathcal{O}} / (N_{\mathcal{O}} + N_{\text{non}\mathcal{O}}). \tag{4}$$

すべてのデータセットで $\mathcal{O}$ クラスの占める割合が 8 割を超える非常に不均衡なデータセットであることが確認できる。また，英語のデータセットでは一般的な訓練/開発/評価データの分割を採用し，一方で日本語のデータセットではランダムに 8:1:1 にデータを分割した．さらに詳細なデータセットの内容については付録 A に記載した．

3.2 ベースラインモデル

我々は 系列ラベリングタスクには BERT [3]、MRC タスクには BERT-MRC [15] を使用した．

BERT [3] は，大規模なテキストコーパスから事前学習された Transformer [16] ベースの言語モデルであり，現在 NLP タスクのベースラインモデルとして幅広く利用されている．テキストを入力することで各トークンに対応したマルチクラスの固有表現ラベルを出力する．系列ラベリングタスクを解く際にこのモデルを使用し，日英 4 つのデータセットにて FL [5] を適用したモデルと比較した．

BERT-MRC [15] は BERT を元に NER を MRC タスクとして解くモデルで，当時英語と中国語のベンチマークで当時最先端の記録を更新した．入力テキストから 2 クラスの各固有表現に対応したトークンの場所を出力する．英語のデータセットを用いて FL [5] や DL [7] を適用したモデルと比較した．

3.3 比較対象の損失関数

Focal loss (FL) [5] は 2 クラス分類に有効な損失関数であり，物体検出における検出対象の前景と背景の不均衡性に着目した損失関数である．また，マルチクラス分類への拡張も可能であるが，適切なハイパーパラメータの調整が必須となる．

表 2 各損失関数の MRC タスクにおける実験結果

	CoNLL03		
	Prec.	Rec.	F1
BERT-MRC + BCE	92.47	92.19	92.33
BERT-MRC + FL [5]	92.81	92.17	92.49
			(+0.16)
BERT-MRC + DL [7]	92.59	92.47	92.53
			(+0.20)
BERT-MRC + MoM	93.09	92.57	**92.83**
			(+0.50)

Dice loss (DL) [7] は偽陽性と偽陰性を等しく重要視する損失関数で，重み付き F 値に近い損失関数である．しかし、2 クラス分類タスクである MRC タスクでの使用を前提としており，系列ラベリングタスクでは使用できない．ここでは DL の論文内で予測性能の向上に最も貢献した dice coefficient (DSC) を DL として使用した．

今回提案する損失関数をとその他の損失関数である CE や FL, DL を用いた場合の予測性能の比較結果を報告する．

4 実験結果

系列ラベリングタスクにおいて 4 つのデータセットに対する各損失関数を適用した際の予測性能の結果を表 1 に示す．MoM loss は CoNLL2003, OntoNotes 5.0, KWDLC, そして Stockmark-NER-Wiki の 4 つのデータセットに対して 既存の CE によって訓練した BERT より F1 スコアが 0.33%, 0.37%, 0.91%, 0.70% 高い結果となった．提案手法である $\mathcal{O}$ クラスと固有

表 3 Sequence labeling タスクにおける各固有表現ごとの実験結果

	BERT + MoM (**proposed**)			BERT				
CoNLL03	Prec.	Rec.	F1	Prec.	Rec.	F1	test tokens	Diff
LOC	0.9353	0.9324	**0.9338**	0.9302	0.9365	0.9334	1,922	0.0005
MISC	0.7937	0.8464	**0.8192**	0.7887	0.8377	0.8125	918	0.0067
ORG	0.8982	0.9363	**0.9168**	0.9030	0.9287	0.9157	2,496	0.0012
PER	0.9707	0.9816	0.9761	0.9762	0.9787	0.9775	2,769	-0.0013
⓪	0.9975	0.9925	**0.9950**	0.9968	0.9930	0.9949	38,312	0.0001
Accuracy			**0.9835**			0.9832		0.0002
Macro avg.	0.9191	0.9378	**0.9282**	0.9190	0.9349	0.9268	46,417	0.0014
Weighted avg.	0.9840	0.9835	**0.9837**	0.9837	0.9832	0.9834	46,417	0.0002

表現クラスを学習する MoM loss が CE や先行研究の FL よりも優れていることを観測した．

上記の結果を受けて，提案手法と CE の F1 スコアの有意差について，対応ありの t 検定を実施した．各データセットにおける p 値は CoNLL2003 が $6.36e^{-6}$, OntoNotes 5.0 が $3.99e^{-6}$, KWDLC が $1.36e^{-3}$, Stockmark-NER-wiki が $1.06e^{-3}$ となった．有意水準が 0.01 の時，対立仮説が支持されるため提案手法と CE の F1 スコアの有意差が確認された．

表 2 に，MRC タスクにおける CoNLL2003 データセットに対する各モデルの予測性能の結果を示す．MoM loss は既存の binary CE (BCE) によって訓練した BERT-MRC より F1 スコアが 0.50% 高い結果となった．また， MoM loss は FL と DL より 0.34%, 0.30% 高い結果となった．有効性が示されている DL に更に我々のアイディアを追加することで，さらなる予測性能向上を実現した．我々は不均衡性に対処する DL の性質を保ちながら MoM loss が更に予測性能を向上させたことを観測した．

表 3 に，系列ラベリングタスクにおける CoNLL2003 データセットに対する各固有表現ごとの予測性能の結果を示す．MoM loss は地名，その他の固有表現，組織名， ⓪ クラスにおいて 0.04%, 0.67%, 0.11%, 0.01% の予測性能の向上が観測された．一方，人名クラスにおいては 0.14% 低下した．この点について以下のセクションで説明する．

5 議論

CoNLL2003 データセットにおける全 3,453 件のテストケースのうち損失関数を CE から MoM loss に変更して正解した例は 35 件確認された．そのうち固有表現クラスと ⓪ クラス間の正例は 10 件で，固有表現間での正例は 25 件であった．またその 25 件中，サンプル数の少ない固有表現クラスからサンプル数の多い固有表現クラスに予測ラベルが変わったことで正解となった例は 22 件であった．この 35 件中 22 件もサンプル数の少ない固有表現クラスからサンプル数の多い固有表現クラスに予測ラベルが変わった理由として， MoM loss が ⓪ クラスに対する不均衡性を考慮することで，モデルが ⓪ クラス以外の固有表現クラスのみの分布に従ったためだと考えられる．従って，表 3 において人名クラスのみ精度が下がったのは ⓪ クラスを除いた固有表現クラスの中で最もサンプル数の多い人名クラスに過学習したためだと推測する．

6 おわりに

我々は固有表現抽出のデータセットの不均衡性に着目した新しい損失関数 MoM loss を提案した．この提案手法は多数派のクラスのトークンのみを計算する loss を従来のモデルの損失関数に追加する手法で，言語非依存なシンプルかつ系列ラベリングや MRC タスクなどの解き方に左右されない効果的な損失関数である．評価結果から以下の 2 つが確認できた．1) BERT において日英 4 つのデータセットを使用して系列ラベリングタスクとして解いた際に言語に依存しない予測性能の向上を確認した．2) BERT-MRC において英語のデータセットを使用して MRC タスクとして解いた際，先行研究の損失関数である FL や DL を上回る予測性能の向上を確認した．

参考文献

[1] Erik Tjong Kim Sang and Fien De Meulder. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In **Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003**, pp. 142–147, 2003.

[2] Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Hwee Tou Ng, Anders Björkelund, Olga Uryupina, Yuchen Zhang, and Zhi Zhong. Towards robust linguistic analysis using ontonotes. In **Proceedings of the Seventeenth Conference on Computational Natural Language Learning**, pp. 143–152, 2013.

[3] 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, 2019.

[4] Xiaoya Li, Jingrong Feng, Yuxian Meng, Qinghong Han, Fei Wu, and Jiwei Li. A unified mrc framework for named entity recognition. **arXiv preprint arXiv:1910.11476**, 2019.

[5] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In **Proceedings of the IEEE international conference on computer vision**, pp. 2980–2988, 2017.

[6] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In **Proceedings of the IEEE/CVF conference on computer vision and pattern recognition**, pp. 9268–9277, 2019.

[7] Xiaoya Li, Xiaofei Sun, Yuxian Meng, Junjun Liang, Fei Wu, and Jiwei Li. Dice loss for data-imbalanced NLP tasks. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 465–476, Online, July 2020. Association for Computational Linguistics.

[8] Tu Dinh Tran, Minh Nhat Ha, Long Hong Buu Nguyen, and Dien Dinh. Improving multi-grained named entity recognition with bert and focal loss. **ICIC Express Letters, Part B: Applications**, Vol. 12, No. 3, pp. 291–299, 2021.

[9] Lance A Ramshaw and Mitchell P Marcus. Text chunking using transformation-based learning. In **Natural language processing using very large corpora**, pp. 157–176. Springer, 1999.

[10] Rich Caruana. Multitask learning. **Machine learning**, Vol. 28, No. 1, pp. 41–75, 1997.

[11] Yu Zhang, Ying Wei, and Qiang Yang. Learning to multitask. **Advances in Neural Information Processing Systems**, Vol. 31, , 2018.

[12] Alexander Spangher, Jonathan May, Sz-Rung Shiang, and Lingjia Deng. Multitask semi-supervised learning for class-imbalanced discourse classification. In **Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing**, pp. 498–517, 2021.

[13] Masatsugu Hangyo, Daisuke Kawahara, and Sadao Kurohashi. Building a diverse document leads corpus annotated with semantic relations. In **Proceedings of the 26th Pacific Asia Conference on Language, Information, and Computation**, pp. 535–544, 2012.

[14] Takahiro Omi. stockmarkteam/ner-wikipedia-dataset: Japanese named entity extraction dataset using wikipedia, 2021.

[15] Xiaoya Li, Jingrong Feng, Yuxian Meng, Qinghong Han, Fei Wu, and Jiwei Li. A unified MRC framework for named entity recognition. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 5849–5859, Online, July 2020. Association for Computational Linguistics.

[16] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. **Advances in neural information processing systems**, Vol. 30, , 2017.

[17] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Transformers: State-of-the-art natural language processing. In **Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations**, pp. 38–45, 2020.

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

表 4 各データセットの統計量．全てのデータセットを，文書ごとに分割し件数と Eq. 4 で実施した不均衡性を示す ⓪ クラスの割合である $\rho_{⓪}$ を示す．

	Lang.	# train	# dev	# test	# class	# ⓪ classs	# non ⓪ classs	$\rho_{⓪}$
CoNLL03 [1]	En	14,041	3,250	3,453	9	248,818	53,993	0.8217
OntoNotes 5.0 [2]	En	75,187	9,603	9,479	37	1,441,685	190,310	0.8834
KWDLC [13]	Ja	12,836	1,602	1,613	17	236,290	16,694	0.9340
NER Wiki [14]	Ja	4,274	535	534	17	80,944	17,552	0.8218

A　データセット

English CoNLL2003　[1] は 1996 年 8 月から 1997 年 8 月までのロイター通信のニュース記事からなるデータセットである．また，学習データと評価データは,1996 年 8 月末のファイルから 10 日分のデータで，テストデータは 1996 年 12 月のテキストである．固有表現は人名や地名の他に組織名などがある．さらに，総トークン数が 302,811 に対し非固有表現のトークンが 82.17%である．

English OntoNotes5.0　[2] は 18 個の固有表現からなる，ニュース，電話での会話，ウェブサイト，放送，トークショーなどの様々なソースから構成されている大規模コーパスである．固有表現は人名や地名の他に日時や金額，法律名などがある．さらに，総トークン数が 1631,995 に対し非固有表現のトークンが 88.34%である．

京大ウェブ文書リードコーパス　[13] はニュース記事，百科事典記事，ブログ，商用ページなどのウェブ文書の冒頭 3 文に対してアノテーションしたテキストコーパスである．コーパスの規模は約 5,000 文書で 15,000 文に相当する．固有表現は人名や地名の他に組織名や日付，時間，金額などがある．さらに，総トークン数が 252,984 に対し非固有表現のトークンが 93.40%である．

Stockmark-NER-wiki　[14] は日本語版 Wikipedia から抜き出した文に対して，アノテーションをした，全体で約 4,000 件ほどのデータセットである．固有表現は人名や地名の他に施設名や製品名，法人名などがある．さらに，総トークン数が 98,496 に対し非固有表現のトークンが 82.18%である．

B　実装の詳細

BERT　は Huggingface の Transformers [17] を使用した．事前学習済みモデルにおいて英語は bert-base，日本語は東北大 BERT を使用した．モデルは Adam [18] を用い, 学習率は $2e^{-5}$, バッチサイズは 64 で各データセット 10 エポックの微調整をした. focal loss [5] と MoM loss の各パラメータは機械学習モデルにおけるベイズ最適化のパッケージである optuna により最適化をした. 最適なモデルのチェックポイントは評価データでの F1 スコアに基づく．我々は、異なるランダムシードを用いた 10 回の実行の F1 スコアの平均値を報告する．

BERT-MRC　は論文 [15] 中で公開されている実装を使用した．Adam [18] を用い，学習率は $3e^{-5}$, バッチサイズは 32 で各データセット 10 エポックの微調整をした. その他の focal loss [5] や dice loss [7] などの各パラメータは dice loss [7] の実装に従った. 最適なモデルのチェックポイントは評価データでの F1 スコアに基づく．我々は、異なるランダムシードを用いた 5 回の実行の F1 スコアの平均値を報告する．