BERT を用いた多言語同時学習による疾患分類

朱晨成　Niraj Pahari　嶋田和孝
九州工業大学大学院
zhu.chencheng822@mail.kyutech.jp　nirajpahari@gmail.com
shimada@ai.kyutech.ac.jp

概要

電子カルテが紙ベースの医療記録に代わって使用されるようになり，医療分野における自然言語処理技術の応用が注目されている．しかし，英語以外の言語における関連する分野のデータセットは依然として少ない．マルチタスク学習は，データ不足の問題を緩和できるアプローチであることが示されている．本研究では，BERT に基づく多言語の同時学習モデル（MTL）を構築し，NTCIR-13 MedWeb マルチラベルの疾患分類タスクで精度評価を行う．シングルタスク学習（STL）と比較した結果，MTL モデルは場合によって STL モデルを上回ることが判明した．また，アブレーションテストによって類似性が高い言語のバイリンガル MTL モデルにおける精度の向上を確認した．

1　はじめに

医療分野でのデジタル化の進展に伴い，電子形式の医療記録（EHR）が推進されている．それにより大量のテキストデータが発生し，医療における自然言語処理の研究にも注目が集まっている [1]．しかし，そのドメインの特異性から，専門家によるアノテーションが必要で時間と人件費のコストが大きく，患者のプライバシーに配慮して非公開にされるデータセットも多い [2]．したがって医療ドメインではアノテーションされた学習用コーパスの不足が続いている．特に英語以外の言語で書かれたデータではこの傾向が強く，関連研究の進展が遅い．このような言語資源のアンバランスは，多言語環境に適した技術によって克服することができる．さらに，同じタスクであれば言語を超えた共通の特徴が存在する可能性もある．異なる言語の医学文章を解析することで，語の曖昧性（コロナと COVID-19 など）が解消され，各国の希少疾病の研究などの発展に貢献すると考えられる．

マルチタスク学習は，機械学習の様々な分野で利用されている [3]．自然言語処理の研究においては，少量のラベル付きデータセットを処理し，関連するタスク間の表現を学習するためによく用いられている [4, 5]．医療分野では，固有表現抽出や関係抽出などのタスクにマルチタスク学習を適用する研究が盛んに行われている [6, 7, 8]．多言語を用いた研究においては，マルチタスク学習は主に対訳問題の解決に用いられている [9, 10]．

本研究では，英語・日本語・中国語の 3 つの言語の医療分野におけるマルチラベル分類タスクに対して，BERT に基づくマルチタスク学習のフレームワークを流用し，同時に学習するモデルを提案する（以降 MTL と呼ぶ）．その性能をシングルタスク学習 (STL) との比較する．また，多言語 MTL モデルにおける言語類似度の影響を調べるために，アブレーションテストも行う．

2　関連研究

マルチタスク学習は転移学習の一種である．関連する複数のタスクから有用な情報を共有することで，1 つのタスクのみを学習させるよりもより良い結果を得られ，モデルの汎化性能も高められる [11]．Ruder [3] は，一般的に用いられているマルチタスク学習法をハードパラメータシェアとソフトパラメータシェアの 2 種類に分類した．ハードパラメータシェアでは，複数のタスク間でモデルの重みを共有するが，各タスクの固有出力層を残す．これに対して，ソフトパラメータシェアではタスクごとにそれぞれ独立したモデルと重みを持つ．学習の時，タスク独自の重みを更新する際に，お互いの損失シグナルを重み付けながら導入できる [12]．

医療分野では，マルチタスク学習が様々なタスクで幅広く活用されている．Joshi ら [13] は，BiLSTM に基づくマルチタスク学習モデルを構築し，ツイート文を対象に 3 つの医療関連の分類タスクを行い，

STL より良い結果を得た．Hartmann ら [14] は，英語，スペイン語，フランス語のデータセットを処理するために mBERT をベースとして，ハードパラメータシェアによって多言語のマルチタスク学習モデルを構築した．タスクには，製品レビューの否定範囲解析，生物医学テキストの否定範囲解析，症状の有無判断が含まれる．また，このモデルの多言語への汎化能力を評価し，臨床テキストにおけるゼロショットの否定範囲解析が可能であることを示した．

3 データセット

本研究では，多言語テキスト分類タスクにおける MTL フレームワークの性能を調べるため，NTCIR-13 医学自然言語処理 Web 文書（MedWeb）[15] を用いる．MedWeb は，人手で作成された疑似ツイート文からなるデータセットである．作成されたツイートの原文は日本語で，著者たちがそれを英語と中国語に翻訳し，マルチリンガルのデータセットを作成した．ツイート一文に対して，インフルエンザ，下痢，花粉症，咳，頭痛，発熱，鼻水，風邪の 8 つのラベルが付与される．表 1 に各言語の疑似ツイートとそのラベルの例を示す．各疾患・症状には陽性（positive）と陰性（negative）の状態が割り当てられ，それぞれ p と n で示される．1 つの疑似ツイートは複数の症状を表現している可能性があるため，複数のラベルにポジティブなステータスが付与される場合がある．本研究ではこれら 8 つのラベルを同時に分類するマルチラベルタスクに取り組む．

4 手法

本研究では，BERT をベースモデルとして，医療分野におけるマルチラベル分類の STL モデル，及び MTL モデルを構築する．

4.1 STL モデル

STL をベースラインとして，各言語のデータセットに対して，BERT モデルの様々なバリアントを用いて，その言語独自の分類タスクを実行する．使用する BERT バリアントは，5.1 節で説明する．

4.2 MTL モデル

ソフトパラメータシェアで MTL モデルを構築する．本研究では，BERT モデルに基づき，その 12 層の Transformer 隠れ層を分解する．大きく分けるとボトム層，中間層，トップ層である．一般に，Transformer のボトム層は最も線形な語順情報を持ち [16]，中間層は主に依存関係を捉え [17]，タスク間の転移能力が最も強く [18]，トップ層はタスク固有の特徴を学習している [19]．

図 1 に本研究で使用する MTL モデルの概要図を示す．フリーズ層（Freeze）は，重みを更新しない層である．共有層（Share）では，3 つのモデルのパラメータを互いに共有し，3 つのタスクに応じた重みの更新を行う．固有層（Individual）は，各言語のデータセットに固有の特徴を学習し，その重みは対応言語自身の分類タスクに応じて更新される．我々の別の研究において F-S-I の組み合わせが最も効率的であることが示されているため [20]，本研究でもこの設定を採用する．BERT のボトム層をフリーズ層，中間層を共有層，トップ層を固有層とする．最後に，12 層の Transformer の上に線形層を追加して，各言語のマルチラベル分類を行う．

5 実験

5.1 実験設定

BERT バリアント 各言語に対して多言語事前学習モデルと言語独自のモノリンガル BERT を用いる．なお，多言語事前学習モデルは，mBERT [21] と LaBSE [22] を使用する．言語独自のモノリンガル BERT について英語データセットでは *bert-base-uncased*[1] を，日本語データセットでは東北大学で開発された *cl-tohoku/bert-base-japanese-whole-word-masking*[2] を，中国語データセットでは *bert-base-chinese*[3] を使用する．

STL モデル 各言語に対して $Single_{mBert}$，$Single_{LaBSE}$，$Single_{mono}$ の 3 つの STL モデルを用いる．下付き文字の mBERT，LaBSE と mono は各言語のデータセットで利用する BERT バリアントを示している．

MTL モデル 図 1 に示すように $Multi_{mBert}$，$Multi_{LaBSE}$，$Multi_{mono}$ の 3 つの MTL モデルを構築する．また，各 MTL モデルで F-S-I それぞれの層数を変化させ，最適なコンビネーションを探る．初期の組み合わせは 4-4-4（フリーズ層は L1-4，共有層は L5-8，固有層は L9-12）．さらに，1-7-4 と 1-4-7

1) https://huggingface.co/bert-base-uncased
2) https://huggingface.co/cl-tohoku/bert-base-japanese-whole-word-masking
3) https://huggingface.co/bert-base-chinese

表 1　各言語の疑似ツイートとそのラベルの付け方の例.

言語	疑似ツイート	Flu	Diarrhea	Hay fever	Cough	Headache	Fever	Runny nose	Cold
en	I have a fever but I don't think it's the kind of cold that will make it to my stomach.	n	n	n	n	n	p	n	p
ja	熱は出てるけどお腹に来る風邪じゃなさそう.								
zh	虽然発烧，但是好像不是肚子着凉的感冒。								

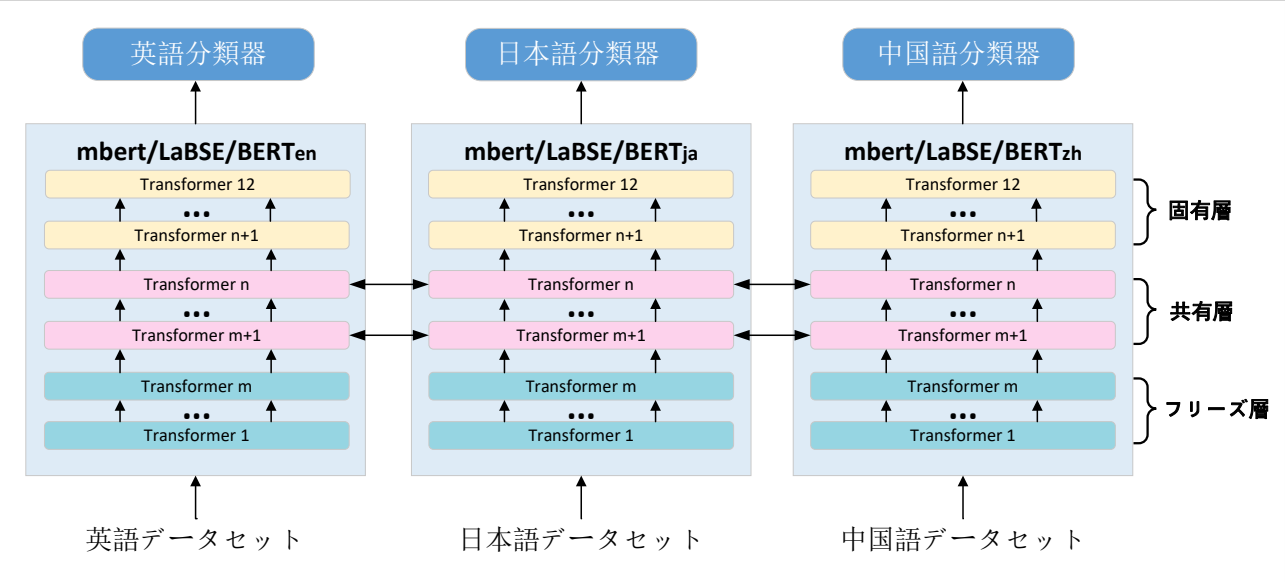

図 1　MTL モデルの概要図.

の組み合わせが最も有効的であることが示されているため [20]，計三つの組み合わせで実験を行う．

BERT の実装　最適化アルゴリズムに AdamW，学習率は Transformer 層を 5e-5，線形層を 5e-3 とし，損失関数には Binary Cross Entropy を用いた．Epoch 数は 10，過学習抑制のために EarlyStopping を用いた．各言語について，トレーニングには 1920 文，テストには 640 文を使用する．評価指標は文中の 8 つのラベルがすべて正しく分類されている場合のみを正解とし，精度の計算を行う．

5.2　実験結果

最適な BERT バリアントとレイヤー数の組合せ　表 2 に F-S-I レイヤー数の組み合わせが異なる場合の 3 つの MTL モデルの精度を示す．左端の列の数字は，実験設置のフリーズ層，共有層，固有層の数を示す．太字は各言語での最高精度を意味する．* は各モデルでの各言語の最高精度を意味する．LaBSE を用いた MTL モデルで，3 つの言語に対して平均的に良い結果が得られた．また，1-7-4 層の設定では，すべての BERT バリアントに対して，過半数の言語で最も良い精度であることがわかった．この結果は [20] と同様であり，各言語において固有層が重要であることが示されている．

表 2　異なる F-S-I レイヤー数の組み合わせにおける 3 つの MTL モデルの予測の完全一致精度.

F-S-I	en	ja	zh
Multi$_{mBert}$			
4-4-4	0.827	0.863*	0.845
1-7-4	0.834*	0.850	0.856*
1-4-7	0.830	0.861	0.834
Multi$_{LaBSE}$			
4-4-4	0.827	0.855	0.873*
1-7-4	**0.848***	**0.869***	0.866
1-4-7	0.836	0.866	0.864
Multi$_{mono}$			
4-4-4	0.822	0.847	0.859
1-7-4	0.817	0.850*	**0.877***
1-4-7	0.836*	0.830	0.855

STL との比較　表 2 から 1-7-4 層の設定が MTL モデルに最適な組み合わせであると考えられるため，この設定の各 MTL モデルを用いて STL モデルとの比較を行う．その結果を表 3 に示す．太字は各言語での最高精度を意味する．まず，同じ BERT バリアント（*Single*$_{mBert}$ 対 *Multi*$_{mBert}$，*Single*$_{LaBSE}$

表 3 MTL と STL の予測の完全一致精度比較.

Model	en	ja	zh	Δ	Δ_{mono}
Baseline (STL)					
$Single_{mBert}$	0.794	0.855	0.852		
$Single_{LaBSE}$	0.805	0.861	0.844		
$Single_{mono}$	0.838	0.856	0.873		
MTL(1-7-4)					
$Multi_{mBert}$	0.834	0.850	0.856	+0.014	-0.009
$Multi_{LaBSE}$	**0.848**	**0.869**	0.866	+0.024	+0.005
$Multi_{mono}$	0.817	0.850	**0.877**	-0.008	-0.008

$^{\Delta}$ は同じ BERT バリアントを用いた MTL モデルと STL モデルの分類精度の平均差を示す
$^{\Delta_{mono}}$ は全ての MTL モデルと $Single_{mono}$ の分類精度の平均差を示す

表 4 バイリンガル MTL の予測の完全一致精度.

Model	ja	zh	Δ	Δ_{mono}
Baseline				
$Single_{mono}$	0.856	0.873		
MTL(1-7-4)				
$Multi_{LaBSE}$	0.869	0.866		+0.002
$Multi_{mono}$	0.850	**0.877**		-0.002
MTL(1-7-4) without en				
$Multi_{LaBSE}$*(-en)*	**0.873**	0.866	+0.005	**+0.005**
$Multi_{mono}$*(-en)*	0.869	0.870	**+0.006**	**+0.005**

$Single_{mono}$，$Multi_{LaBSE}$，$Multi_{mono}$ は表 3 の値と同じ
$^{\Delta}$ は同じ BERT バリアントを用いたバイリンガル MTL モデルと元の MTL モデルの分類精度の平均差を示す
$^{\Delta_{mono}}$ は全ての MTL モデルと $Single_{mono}$ の分類精度の平均差を示す

対 $Multi_{LaBSE}$，$Single_{mono}$ 対 $Multi_{mono}$）を前提として，言語間の分類精度の平均差を互いに比較する（Δ）．全体として，STL よりも MTL の方が優れているが，その効果は必ずしも大きくはなかった．次に，全ての MTL モデルを STL モデルで最も精度が高かった $Single_{mono}$ と比較する（Δ_{mono}）．結果，LaBSE を用いた MTL モデルのみ $Single_{mono}$ よりもわずかに精度が向上している．なお，この精度は臨床医学テキストで事前学習した BERT モデルを用いた NTCIR-13 MedWeb タスクの日本語データセットでの予測精度を上回る結果となった [23].

まとめると，提案手法（MTL）は各言語でチューニングされた最適なモデルに対していつも高い精度が得られるとは限らない（例えば $Multi_{LaBSE}$ と $Single_{mono}$ における zh）．共有されたモデル間の Transformer のパラメータがより多くのノイズを生成するため，MTL がその能力を発揮できないことが原因であると考えられる．しかし，複数を組み合わせて同時に学習するメリットも明らかになった．

5.3 アブレーションテスト

Conneau ら [24] の研究では，言語間の類似性が mBERT，XML などの多言語事前学習モデルの cross-lingual transfer 能力に影響し，類似性の高い言語ほどタスクの精度が高くなることが示されている．そこで本研究では，言語間の類似性の影響を調査するため，3 言語内の 1 つを除いてバイリンガル MTL モデルを作成し，アブレーションテストを行った．言語学的に見ると英語と中国語は文法構造が似ており，日本語と中国語は表層形が似ている．また，英語と日本語の類似性はこの組み合わせの中では最も低いと考えられる．英語-中国語，日本語-中国語，英語-日本語の 3 つのペアでテストを行う．その結果を表 4 に示す．without ja と without zh の組み合わせもテストを行ったが，精度の向上が見られなかったため，スペースの関係上割愛する．日本語-中国語ペアはテストを行なった全てのグループの中で最も良い精度が得られた．その平均精度は，3 言語を用いた MTL よりも優れている（Δ）．また，ベースラインで最も性能が良かった $Single_{mono}$ と比較してもわずかに優れていた（Δ_{mono}）．加えて，LaBSE を用いたバイリンガル MTL モデルは，日本語において全ての実験の中で最も高い精度を示した．言語間の相関の度合い，特に表層形の類似度が，多言語 MTL モデルの性能に影響を与えることがわかる．

6 まとめ

本研究では，多言語医療関連マルチラベル分類タスクに対して，Transformer に基づくソフトパラメータシェアアプローチによる MTL モデルを構築した．MTL モデルは全体的に STL モデルより優れているが，言語や組み合わせによっては各言語独自の BERT モデルの精度を上回るには至らない場合もあった．また，アブレーションテストを行い，日本語の分類タスクと中国語の分類タスクのみからなるバイリンガル MTL モデルが最も良い結果を得られた．この結果により，言語間の表層形の類似性が多言語の MTL モデルの性能に影響を与えることが示された．

今回英語，日本語，中国語の 3 言語のみ用いた．今後はスペイン語やフランス語など，英語の表層形に類似度高い言語のデータセットも取り込みたい．

参考文献

[1] Eiji Aramaki, Shoko Wakamiya, Shuntaro Yada, and Yuta Nakamura. Natural language processing: from bedside to everywhere. **Yearbook of Medical Informatics**, 2022.

[2] Shuntaro Yada, Yuta Nakamura, Shoko Wakamiya, and Eiji Aramaki. Real-mednlp: Overview of real document-based medical natural language processing task. In **Proceedings of the 16th NTCIR Conference on Evaluation of Information Access Technologies**, pp. 285–296, 2022.

[3] Sebastian Ruder. An overview of multi-task learning in deep neural networks. **arXiv preprint arXiv:1706.05098**, 2017.

[4] Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. Adversarial multi-task learning for text classification. **arXiv preprint arXiv:1704.05742**, 2017.

[5] Anders Søgaard and Yoav Goldberg. Deep multi-task learning with low level tasks supervised at lower layers. In **Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)**, pp. 231–235, 2016.

[6] Shweta Yadav, Srivatsa Ramesh, Sriparna Saha, and Asif Ekbal. Relation extraction from biomedical and clinical text: Unified multitask learning framework. **IEEE/ACM Transactions on Computational Biology and Bioinformatics**, Vol. 19, No. 2, pp. 1105–1116, 2020.

[7] Andriy Mulyar, Ozlem Uzuner, and Bridget McInnes. Mt-clinical bert: scaling clinical information extraction with multitask learning. **Journal of the American Medical Informatics Association**, Vol. 28, No. 10, pp. 2108–2115, 2021.

[8] Zhaoying Chai, Han Jin, Shenghui Shi, Siyan Zhan, Lin Zhuo, and Yu Yang. Hierarchical shared transfer learning for biomedical named entity recognition. **BMC bioinformatics**, Vol. 23, No. 1, pp. 1–14, 2022.

[9] Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. Multi-task learning for multiple language translation. In **Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)**, pp. 1723–1732, 2015.

[10] Yiren Wang, ChengXiang Zhai, and Hany Hassan Awadalla. Multi-task learning for multilingual neural machine translation. **arXiv preprint arXiv:2010.02523**, 2020.

[11] Yu Zhang and Qiang Yang. A survey on multi-task learning. **IEEE Transactions on Knowledge and Data Engineering**, 2021.

[12] Michael Crawshaw. Multi-task learning with deep neural networks: A survey. **arXiv preprint arXiv:2009.09796**, 2020.

[13] Aditya Joshi, Sarvnaz Karimi, Ross Sparks, Cécile Paris, and C Raina MacIntyre. Does multi-task learning always help?: An evaluation on health informatics. In **Proceedings of the the 17th annual workshop of the Australasian language technology association**, pp. 151–158, 2019.

[14] Mareike Hartmann and Anders Søgaard. Multilingual negation scope resolution for clinical text. In **Proceedings of the 12th International Workshop on Health Text Mining and Information Analysis**, pp. 7–18, 2021.

[15] Shoko Wakamiya, Mizuki Morita, Yoshinobu Kano, Tomoko Ohkuma, and Eiji Aramaki. Overview of the ntcir-13: Medweb task. In **NTCIR**, 2017.

[16] Yongjie Lin, Yi Chern Tan, and Robert Frank. Open sesame: getting inside bert's linguistic knowledge. **arXiv preprint arXiv:1906.01698**, 2019.

[17] Jesse Vig and Yonatan Belinkov. Analyzing the structure of attention in a transformer language model. **arXiv preprint arXiv:1906.04284**, 2019.

[18] Anna Rogers, Olga Kovaleva, and Anna Rumshisky. A primer in bertology: What we know about how bert works. **Transactions of the Association for Computational Linguistics**, Vol. 8, pp. 842–866, 2020.

[19] Yaru Hao, Li Dong, Furu Wei, and Ke Xu. Visualizing and understanding the effectiveness of bert. **arXiv preprint arXiv:1908.05620**, 2019.

[20] Niraj Pahari and Kazutaka Shimada. Multi-task learning using bert with soft parameter sharing between layers. In **2022 Joint 12th International Conference on Soft Computing and Intelligent Systems and 23rd International Symposium on Advanced Intelligent Systems**, pp. 1–6, 2022.

[21] 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.

[22] Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. Language-agnostic bert sentence embedding. **arXiv preprint arXiv:2007.01852**, 2020.

[23] Yoshimasa Kawazoe, Daisaku Shibata, Emiko Shinohara, Eiji Aramaki, and Kazuhiko Ohe. A clinical specific bert developed using a huge japanese clinical text corpus. **Plos one**, Vol. 16, No. 11, p. e0259763, 2021.

[24] Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. Emerging cross-lingual structure in pretrained language models. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 6022–6034, Online, July 2020. Association for Computational Linguistics.