遠距離教師データの特徴表現を活用した薬物タンパク質間関係抽出

飯沼 直己　三輪 誠　佐々木 裕
豊田工業大学
{sd20403,makoto-miwa,yutaka.sasaki}@toyota-ti.ac.jp

概要

創薬等の分野で重要な薬物タンパク質間の相互作用に関する情報の深層学習による自動抽出が注目されているが，教師データの作成が高コストであるという問題がある．そこで，低コストで大量の教師データの作成できる遠距離教師あり学習が提案されているが，誤ったラベルのデータが含まれるため，予測性能が低下する問題が残っている．本研究では，教師データを作成するコストを抑えながら，抽出性能の向上を目指し，特徴表現により遠距離教師データを間接的に活用した深層学習モデルの学習手法を提案する．DrugProt コーパスを対象とした実験の結果，遠距離教師データの利用により，F 値のマイクロ平均が 0.8 ポイント向上することを示した．

1　はじめに

「根拠に基づく医療」を展開していく上で，薬物とタンパク質間の相互作用に関する情報は創薬・代謝工学・薬物反応モデリング等の分野で重要である．しかし，相互作用の情報は文献として発表され，文献数は急速に増加しているため [1]，相互作用の情報を決定するために薬理学者が逐一論文を読むことは困難である．そのため，テキストからの相互作用抽出 [2] の自動化が注目されている．この中で，深層学習を利用した手法は高い性能を達成しているものの，教師データ作成に莫大なコストがかかるという問題を抱えている．

低コストで大量の教師データの作成を可能にする遠距離教師あり学習が Mints ら [3] によって提案されている．しかし，この手法には学習時のノイズとなる誤ったラベルのデータを作成してしまう問題が残っている．そのため，Beltagy ら [4] による手法などノイズを緩和する手法が提案されている．

本研究では，自動的に生成した遠距離教師データの，教師あり薬物タンパク質間相互作用抽出への活用方法の提案とその抽出性能の評価を目的とする．作成の容易な遠距離教師データを活用し，教師データ作成のコストを抑えながら，薬物タンパク質間関係抽出の性能向上を目指す．

2　関連研究

2.1　遠距離教師あり学習

遠距離教師あり学習は Mints ら [3] によって提案された，データベースから機械的に文書に関連情報をラベル付けする手法である．これにより，大量の教師データを生成できるが，ラベル付けを誤った教師データを含んでしまう．この誤ったラベルは予測モデルの性能を低下させるノイズとなる．そのため，ノイズの影響を緩和する手法が提案されている．よく用いられる手法として，教師データをデータベース上のペアに対応するインスタンスのバッグで扱うマルチインスタンス学習がある．Zeng ら [5] はバッグ中の正解ラベルに対する予測確率が最も高い表現を持つデータのみを利用して学習する手法を提案した．また Beltagy ら [4] は，少量の人手の教師データによる注意機構を利用する手法を提案した．

2.2　薬物タンパク質間関係抽出

薬物タンパク質間関係抽出は，文献中に記載された薬物とタンパク質間の関係を抽出するタスクである．Gu ら [2] は大規模なニューラルネットワークモデル BERT を大規模な生物医学文献で事前学習したモデル (PubMedBERT) を提案した．

3　提案手法

本手法では，遠距離教師データから得られる特徴表現を活用して人手の教師データの薬物タンパク質間関係抽出を行う．機械的に作成した遠距離教師

データを活用することで，教師データを追加で作成するコストを抑えながら，抽出性能の向上を目指す．

以降，3.1 節でベースとなる関係抽出モデル，3.2 節でデータベースからの遠距離教師データの作成，3.3 節で遠距離教師データの特徴表現の活用手法についてそれぞれ説明する．

3.1 関係抽出モデル

本研究のベースラインとなる人手の教師データのみでモデルを学習する関係抽出モデルについて説明する．まず，入力文を BERT [6] でエンコードし，入力文の文脈を表した特徴量ベクトル $\boldsymbol{h}$ を生成する．BERT は入力文の各トークンに対して表現ベクトルを生成するが，文全体の特徴を含んだ [CLS] トークンの表現ベクトルを特徴表現ベクトルとして用いる．さらに，特徴量ベクトルを基に全結合層とソフトマックス関数により各関係に対する予測確率を表した予測ベクトルを生成する．モデルは予測確率が最大となる関係を予測する．最適化手法は Adam [7] を用い，交差エントロピーを目的関数として最小化するようにモデルを学習する．

3.2 遠距離教師データの作成

遠距離教師データの作成の概略を図 1 に示す．以前提案した遠距離教師データ作成手法 [8] では薬物データベース DrugBank [9]・タンパク質データベース UniProt [10]・医学文献データベース PubMed [1] を用いたが，本手法ではさらに化学物質データベース Comparative Toxicogenomics Database (CTD) [11] を活用して遠距離教師データを作成する．以降，これらのデータベースを用いて遠距離教師データを作成する手順を説明する．

まず，DrugBank から ID 関係トリプルを作成する．ID 関係トリプルとは薬物の ID・相互作用名・タンパク質の ID のトリプルである．次に，DrugBank, CTD の情報を基に薬物の ID と表層表現を対応付けて薬物名辞書を作成する．同様に UniProt, CTD からタンパク質名辞書を作成する．これらの辞書と ID 関係トリプルから関係トリプルを作成する．さらに，生物医学・科学文献の処理に特化したツールである SciSpacy [12] の文分割・エンティティ抽出により，PubMed 文献を解析し，文献中の専門用語を抽出する (NER). 最後に，関係トリプルと PubMed 文献から抽出された専門用語を用いて辞書マッチにより PubMed 文献から遠距離教師データを作成する．DrugBank 上の相互作用名とタスク上の関係名のマッピングには，付録 A 章の表 2 に示す DrugProt コーパスの関係アノテーションガイドライン [13] の関係の記述に基づいて作成された辞書を用い（例：Inducer は INDIRECT-UPREGULATOR に含まれる），辞書でマッピングできない DrugBank 上の相互作用名のインスタンスはフィルタリングする．

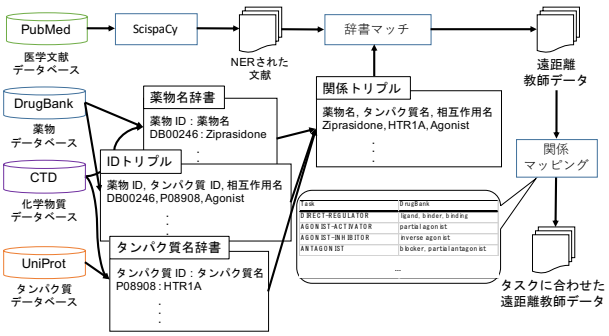

図 1 遠距離教師データの作成

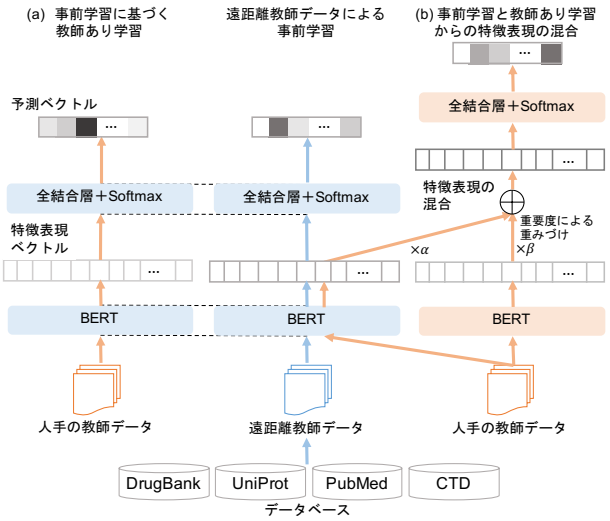

図 2 遠距離教師データの活用

3.3 遠距離教師データを活用した関係抽出

遠距離教師データから得られる特徴表現の活用方法として，遠距離教師データによる事前学習と遠距離教師データと人手の教師データから得られる特徴量の混合の 2 つの手法を提案する．これら 2 つの手法の概要を図 2 に示す．以降，この節では，これら 2 つの手法について説明する．

3.3.1 事前学習としての利用

図 2(a) に示す遠距離教師データを事前学習に活用する手法について説明する．自然言語処理のタスク

において，ドメインの近いデータセットによる事前学習により目的のデータセットでのモデルの性能が向上することが知られている [14]．そこで，遠距離教師データによる事前学習を行う．具体的には，まず遠距離教師データにより 3.1 節で説明した関係抽出モデルを学習し，続いて遠距離教師データにより学習したパラメータを初期値として人手の教師データにより関係抽出モデルを学習する．

3.3.2 特徴量表現の混合

図 2 (b) に示す遠距離教師データと人手の教師データで各々事前学習したモデルから得られる特徴表現ベクトルを混合する手法について説明する．まず，遠距離教師データと人手の教師データでそれぞれ関係抽出モデルを学習する．関係抽出モデルの特徴抽出器として用いた BERT は入力のトークン系列に対して表現系列 $\boldsymbol{H}$ を出力するため，3.1 節で説明した手法と同様に [CLS] トークンの表現 $\boldsymbol{h}$ を特徴表現ベクトルとする．遠距離教師データ，人手の教師データで学習した関係抽出モデルから得られる特徴表現ベクトルをそれぞれ $\boldsymbol{h_{ds}}, \boldsymbol{h_{sv}}$ とし，以下の式に示すように混合手法として Gate と Concat の 2 種類の手法を提案する．

$$\boldsymbol{h_{Gate}} = \alpha \boldsymbol{h_{ds}} + \beta \boldsymbol{h_{sv}} \quad (1)$$

$$\boldsymbol{h_{Concat}} = [\gamma \boldsymbol{h_{ds}}; \eta \boldsymbol{h_{sv}}] \quad (2)$$

$[\cdot;\cdot]$ はベクトルの結合を示す．$\alpha, \beta, \gamma, \eta$ は各特徴量の重要度を表す重みであり，学習可能なスカラー値のパラメータとする．

Gate は (1) 式に示すように，$\boldsymbol{h_{ds}}, \boldsymbol{h_{sv}}$ を重要度を意味するパラメータで定数倍した後に各々の要素和を計算して混合する．Concat は (2) 式に示すように，$\boldsymbol{h_{ds}}, \boldsymbol{h_{sv}}$ をパラメータで定数倍した後に結合して混合する．これらの手法で混合した特徴表現ベクトル $\boldsymbol{h_{Gate}}, \boldsymbol{h_{Concat}}$ を全結合層とソフトマックス関数により関係を予測する．

4 実験と考察

人手の教師のデータセットとして，5,000 件の医学文献に対して薬物，タンパク質とそれらの関係がタグ付けされた DragProt コーパス [15] を用いて関係抽出を学習し，開発データで評価した．評価指標には F 値を用いた．遠距離教師のデータセットは 3.2 節の手法で作成し，事例数は 400,867 件である．また，実験環境の詳細は付録 B 章に示す．

4.1 関係抽出の性能比較

人手の教師データのみで学習したベースラインと提案手法の抽出性能を比較する実験を行った．ベースラインとして，データセットと近いドメインで事前学習された PubMedBERT，BioRoBERTa-large[16] をベースとした関係抽出モデルを人手の教師データのみで学習したモデルを用いた．BioRoBERTa-large はパラメータサイズが PubMedBERT のおよそ 3 倍の大規模な事前学習モデルであり，DragProt コーパス [15] を用いたコンペティションにおいて外部知識を用いないモデルの中で SOTA を達成し，開発データで 77.46%を記録した [17] モデルである．結果を表 1 に示す．

まず，PubMedBERT をベースラインとして提案手法を適用した場合の性能に着目する．遠距離教師データのみで学習した際の F 値のマイクロ平均は 16.6%と低いものの，全提案手法において，人手の教師データが少ない AGONIST，PRODUCT-OF に対する予測性能が大きく向上した．これは遠距離教師データの特徴表現を活用することで，ノイズの影響を緩和しながら人手の教師データの不足を補えたためだと考えられる．また，特徴量表現の混合により遠距離教師データを利用した Gate, Concat は人手の教師データが多い ANTAGONIST をはじめとした多くの関係に対する予測性能が向上し，マイクロ平均も全体で 0.8 ポイント向上した．このことから，事前学習より特徴量表現の混合の方が遠距離教師データの特徴表現の利用方法として有効であると分かった．

次に，BioRoBERTa をベースラインとして提案手法を適用した場合の性能に注目する．マイクロ平均において，提案手法により 0.5 ポイントの性能向上が得られた．更に，関係ごとの F 値を比較しても，ACTIVATOR，ANTAGONIST，SUBSTRATE を除く全ての関係で性能が向上もしくは維持された．この結果と PubMedBERT の結果から，提案手法はモデルのパラメータサイズに依らず性能向上が得られることが分かった．

4.2 小規模の教師データでの性能比較

本節では，小規模の人手の教師データで学習する際の提案手法の有効性を検証した．本研究では，作成コストの低い遠距離教師データの活用により，教師データを追加で作成するコストを抑えながら薬物

表 1　関係抽出性能比較 (評価指標は F 値)

	PubMed BERT	+ 遠距離	+(a) 事前学習	+(b) Gate	+(b) Concat	BioRoBERTa large	+(b) Concat	学習事例数
INDIRECT-DOWNREGULATOR	76.7	0.0	74.6	77.7	78.7	79.3	79.9	1,330
INDIRECT-UPREGULATOR	73.3	1.9	75.1	73.7	73.6	75.6	76.2	1,379
DIRECT-REGULATOR	65.9	6.1	62.1	66.9	67.7	66.9	69.4	2,250
ACTIVATOR	77.3	5.2	70.6	77.5	76.7	75.7	73.8	1,429
INHIBITOR	84.2	29.4	84.7	84.6	84.3	85.1	86.1	5,392
AGONIST	75.5	6.7	79.7	78.2	77.0	76.1	77.2	659
AGONIST-ACTIVATOR	0.0	0.0	46.2	0.0	0.0	0.0	0.0	29
AGONIST-INHIBITOR	0.0	0.0	80.0	0.0	0.0	0.0	0.0	13
ANTAGONIST	90.6	26.0	89.6	92.2	91.8	91.7	90.2	972
PRODUCT-OF	59.0	10.6	63.7	62.9	62.5	61.2	62.0	921
SUBSTRATE	69.5	13.1	69.1	68.4	69.9	72.7	71.8	2,003
SUBSTRATE_PRODUCT-OF	0.0	0.0	0.0	0.0	0.0	0.0	0.0	25
PART-OF	71.7	0.0	70.6	72.2	71.7	72.8	74.4	886
マイクロ平均	76.2	16.6	75.6	76.8	77.0	77.5	78.0	—

タンパク質間相互作用抽出の性能向上を目指している．4.1 節では人手の教師データを全て使って学習し，提案手法によりベースラインから性能向上が得られることが確認できた．そこで，小規模の人手の教師データで学習する際の有効性を検証するため，人手の教師データの一部のみを使って学習し，ベースラインと提案手法の関係抽出性能を比較した．[100, 200, 500, 1,000] の事例数の各関係の人手の教師データを使ってモデルを学習した際の開発データでの性能を確認した．提案手法には，4.1 節でベースラインから最も性能向上が得られた Concat で特徴表現を混合するモデルを用いた．

結果を図 3 に示す．4.1 節で人手の教師データを全て使って学習した際と同様にベースラインから大幅な性能向上は得られなかったが，全ての事例数で一貫して性能が向上した．このことから，提案手法により遠距離教師データから得られる特徴表現を利用することで，訓練に使用できる人手の教師データの事例数に依らず性能の向上が得られることが分かった．

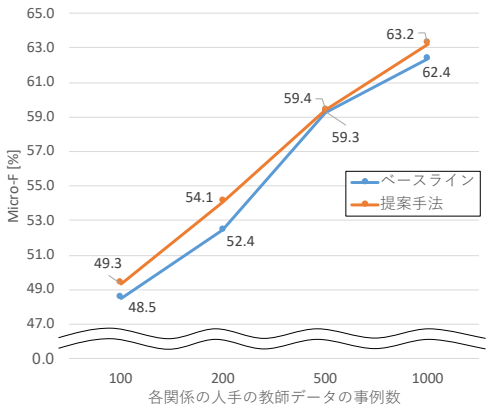

図 3　訓練事例数と Micro-F の関係

5　おわりに

遠距離教師データを低コストで作成し，活用することで，教師データの作成コストを抑えながら薬物タンパク質間相互作用抽出の性能向上を目指した．そこで，2 種類の遠距離教師データの活用手法を提案した．両手法ともに人手の教師データが少ない関係に対する予測性能がベースラインから向上した．さらに，特徴表現混合法では人手の教師データが多い関係に対する予測性能とマイクロ平均も向上し，提案手法の有効性を示すことができた．加えて，モデルのパラメータサイズや訓練に使用できる人手の教師データの事例数に依らず性能向上が得られることを示した．

今後は，さらなる性能の向上に向け，混合特徴表現法での事前学習と特徴表現の混合部分の学習手法を検討する予定である．

謝辞

本研究は JSPS 科研費 JP20K11962 の助成を受けたものです．

参考文献

[1] NCBI Resource Coordinators. Database resources of the national center for biotechnology information. **Nucleic Acids Res.**, 1 2016.

[2] Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. **ACM Transactions on Computing for Healthcare**, Vol. 3, No. 1, p. 1–23, Jan 2022.

[3] Mike Mintz, Steven Bills, Rion Snow, and Daniel Jurafsky. Distant supervision for relation extraction without labeled data. In **Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP**, pp. 1003–1011, Suntec, Singapore, August 2009. Association for Computational Linguistics.

[4] Iz Beltagy, Kyle Lo, and Waleed Ammar. Combining distant and direct supervision for neural relation extraction. 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. 1858–1867, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[5] Daojian Zeng, Kang Liu, Yubo Chen, and Jun Zhao. Distant supervision for relation extraction via piecewise convolutional neural networks. In **Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing**, pp. 1753–1762, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[7] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In **Proceedings of the 3rd International Conference for Learning Representations**, 2015.

[8] 飯沼直己, 三輪誠, 佐々木裕. 遠距離教師データを援用した教師あり薬物タンパク質間相互作用抽出. 言語処理学会 第 26 回年次大会, 2020.

[9] Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, Sajed T, Johnson D, Li C, Sayeeda Z, Assempour N, Iynkkaran I, Liu Y, Maciejewski A, Gale N, Wilson A, Chin L, Cummings R, Le D, Pon A, Knox C, and Wilson M. Drugbank 5.0: a major update to the drugbank database for 2018. **Nucleic Acids Res.**, 1 2018.

[10] The UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. **Nucleic Acids Research**, Vol. 49, No. D1, pp. D480–D489, 11 2020.

[11] Allan Peter Davis, Cynthia J Grondin, Robin J Johnson, Daniela Sciaky, Jolene Wiegers, Thomas C Wiegers, and Carolyn J Mattingly. Comparative Toxicogenomics Database (CTD): update 2021. **Nucleic Acids Research**, Vol. 49, No. D1, pp. D1138–D1143, 10 2020.

[12] Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. In **Proceedings of the 18th BioNLP Workshop and Shared Task**, pp. 319–327, Florence, Italy, August 2019. Association for Computational Linguistics.

[13] Obdulia Rabal, Jose Antonio López, Astrid Lagreid, and Martin Krallinger. DrugProt corpus relation annotation guidelines [ChemProt - Biocreative VI], June 2021.

[14] Iz Beltagy, Kyle Lo, and Arman Cohan. Scibert: Pretrained language model for scientific text. In **EMNLP**, 2019.

[15] Martin Krallinger, Obdulia Rabal, Antonio Miranda-Escalada, and Alfonso Valencia. DrugProt corpus: Biocreative VII Track 1 - Text mining drug and chemical-protein interactions, June 2021.

[16] Patrick Lewis, Myle Ott, Jingfei Du, and Veselin Stoyanov. Pretrained language models for biomedical and clinical tasks: Understanding and extending the state-of-the-art. In **Proceedings of the 3rd Clinical Natural Language Processing Workshop**, pp. 146–157, Online, November 2020. Association for Computational Linguistics.

[17] Wonjin Yoon, Sean Yi, Richard Jackson, Hyunjae Kim, Sunkyu Kim, and Jaewoo Kang. Using knowledge base to refine data augmentation for biomedical relation extraction. November 2021.

[18] Guido Van Rossum and Fred L. Drake. **Python 3 Reference Manual**. CreateSpace, Scotts Valley, CA, 2009.

[19] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In **Advances in Neural Information Processing Systems 32**, pp. 8024–8035. Curran Associates, Inc., 2019.

[20] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In **Proceedings of the 25rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**, 2019.

表 2　関係マッピングの辞書

DrugBank	タスク
ligand, binder, binding	DIRECT-REGULATOR
partial agonist	AGONIST-ACTIVATOR
inverse agonist	AGONIST-INHIBITOR
blocker, partial antagonist	ANTAGONIST
inducer, stimulator	INDIRECT-UPREGULATOR
product of	PRODUCT-OF
activator	ACTIVATOR
inhibitor	INHIBITOR
agonist	AGONIST
antagonist	ANTAGONIST
substrate	SUBSTRATE

表 3　実験に用いたハードウェア

内容	項目
OS	Ubuntu 18.04.4 LTS
GPU	NVIDIA Tesla V100 DGXS
GPU メモリ	32GB
CPU	Intel(R) Xeon(R) CPU E5-2698 v4 @ 2.20GHz
メモリ	256GB

表 4　パラメータ探索範囲

ハイパーパラメータ	値
学習率	[1e-6, 1e-4]
ドロップアウト率	[0.0, 0.5]
weight decay	[1e-10, 1e-3]

付録

A　関係名のマッピング

3.2 節で説明した DrugBank 上の相互作用名とタスク上の関係名のマッピングに用いた辞書を表 2 に示す．

B　実験環境

4 章で示した実験を行った環境について説明する．実装には，Python3 [18] のバージョン 3.8.1 と深層学習ライブラリである PyTorch [19] のバージョン 1.9.0 を用い，表 3 に示すハードウェア環境で実験した．学習時のハイパーパラメータは表 4 の探索範囲からオープンソースのハイパーパラメータ自動最適化フレームワーク Optuna[20] を用いて，開発データでの F 値のマイクロ平均が最大になるようにパラメータ探索をして決定した．