自動生成した学習データを用いたマルチタスク学習による タンパク質と化学物質間の関係抽出

新城 大希 [1,3]　西川 仁 [1]　徳永 健伸 [1]　牧野 拓哉 [2,3]　岩倉 友哉 [2,3]

[1] 東京工業大学 情報理工学院　　[2] 株式会社富士通研究所

[3] 理研 AIP-富士通連携センター

[1]{shinjo.t.ab@m, hitoshi@c, take@c}.titech.ac.jp

[2]{makino.takuya, iwakura.tomoya}@fujitsu.com

1 はじめに

化学物質の知識は，新材料や新薬の開発，材料を用いた製品開発に用いられており，化学物質の知識を活用するために，論文や特許で数分ごとに報告される化学物質の物性値や化学物質間の相互関係といった情報を構造化する作業が行われている.[1] しかしながら，構造化の作業は専門的な知識を必要とし，現状，人手に大きく頼る作業であるため，日々増え続ける化学物質への対応が困難となっている．

このような問題を解決するための技術の一つとして，化学物質間の関係を自動で抽出する研究が行われている．関係抽出の実現のためには，正解が付与されたコーパスを用いて学習を行う手法が広く研究されている．たとえば，関係抽出用のコーパスである CHEMPROT コーパス [5] においては，約 2,400 件の化学論文のアブストラクトに対して，タンパク質・化学物質のペアとその関係がラベル付けされている（図 1）．さらには、正解が付与されていないラベルなしテキストを用いた精度改善方法も提案されている．BioBERT[6] はバイオ分野の文献を用いて事前学習したのちに CHEMPROT の学習データで fine-tuning することで，CHEMPROT の最高精度である F 値 0.765 を達成している．

BioBERT は高い抽出精度が得られることが報告されているが，次の 2 点から改善の余地がある．一つ目は，ラベル付きデータの量が限定的であるという点である．化学文献からの関係抽出タスクはラベル付与に高度な専門知識を必要とする．そのため，人手作業コストの観点から大規模なラベル付きコーパスを構築することが難しい．二つ目は，目的の関係抽出タスクで

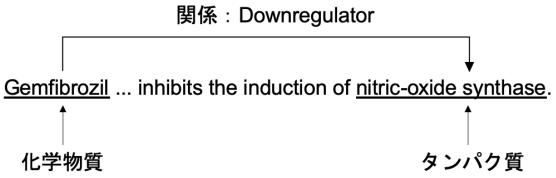

図 1: CHEMPROT のアノテーション例

定義されるラベルは構文的な情報が手掛かりになると期待できる一方で，BioBERT はそのような情報を利用していないという点である．BERT は単語列からなる文を入力として文に対する特徴量を抽出するため、構文的な情報を利用しない．そのため、構文的な情報を与える方法は自明ではない．

本研究では，さらなる精度改善を目標に，ラベルなしテキストから自動作成した補助タスク向けの教師データと，目的タスクの学習データを用いたマルチタスク学習による関係抽出手法を提案する．具体的には，Open IE と呼ばれる，文内の構文的な構造から「2 つのエンティティとその関係」の 3 つ組を抽出する技術を利用して補助タスクの学習データを作成し，関係抽出を行う「主タスク」と Open IE で抽出されたペアかどうかを分類する「補助タスク」を同時に学習し，主タスクの精度向上を図る．本手法を BioBERT 上に実装し，CHEMPROT で評価した．その結果，F 値で 0.773 という最高精度が得られることを確認した．

2 関連研究

CHEMPROT を用いて関係抽出を行なっている研究を紹介する．Lee ら [6] は，BERT[3] の構造に基づき，

[1] https://www.jaici.or.jp/annai/img/20150709_CAS_PressRelase.pdf

生物医学ドメインコーパスである PubMed[2]，PMC[3] を用いて事前学習を行った BioBERT モデルを提案した．BioBERT を CHEMPROT でファインチューニングした結果，CHEMPROT を用いた関係抽出において現時点での最高精度を達成している．

大規模なコーパスからエンティティ間の関係を抽出する技術として，Open Information Extraction（Open IE）[1] がある．Open IE では，主語・目的語となるエンティティのペアと，エンティティ間の関係を合わせた 3 つ組を抽出する．例えば，“Shakespeare is author of Hamlet.” という文に対して，Open IE を用いると (Shakespeare, is, author), (Shakespeare, is author of, Hamlet) といった 3 つ組が抽出できる．2 つのエンティティとその関係を抽出するという点において，関係抽出タスクとの関連性があり，後述するマルチタスク学習に利用できると考えられる．

マルチタスク学習とは，主タスクに関連する補助タスクを同時に学習することにより，タスク間で共通する情報を学習させる手法である．Ruder ら [9] によると，複数タスクによる「潜在的なデータ拡張」，タスク間に共通する特徴に着目する「Attention focusing」，タスク間の難易度の差を利用する「盗み聞き」といった要素を，マルチタスク学習の精度向上の根拠としている．本研究では補助タスクにラベル付きデータを使用せず，Open IE によって自動でラベル付きデータを作成する．

3 提案手法

本研究では，後述する主タスクと補助タスクを同時に学習するマルチタスク学習を提案する．提案手法の概念図を図 2 に示す．

なお，CHEMPROT には文をまたいだ関係も存在するが，全体の 1 %未満であるため，本研究では同一文内に出現するタンパク質と化学物質のペアのみを対象とする．したがって，入力には事前にアノテーションされたタンパク質と化学物質のペアを含む 1 文を用いる．各入力文に対して主タスクのラベルと補助タスクのラベルが 1 つずつ付与されている．この入力文を共通のエンコーダーに入力し，エンコーダーの出力を各タスクの分類器の入力に用いる．

訓練時には主タスクと補助タスクの学習を同時に行い，各タスクの損失の重み付き和を最終的な損失関数

[2]http://www.ncbi.nlm.nih.gov/pubmed/
[3]http://www.ncbi.nlm.nih.gov/pmc/

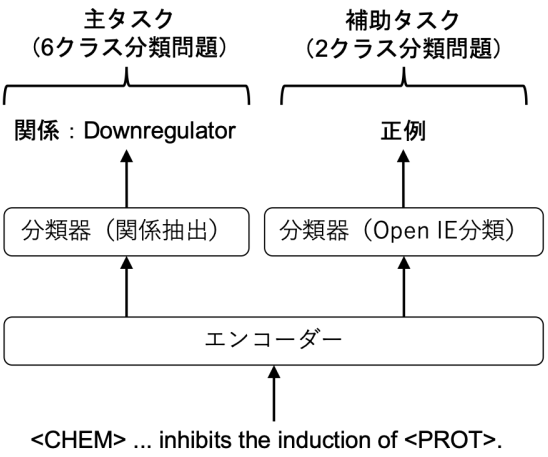

図 2: 提案手法の概念図

とする．主タスクの損失関数 L_m，補助タスクの損失関数 L_a，補助タスクの重みを表すパラメータ w を用いて，最終的な損失関数は式 (1) で表せる．

$$L = (1 - w)L_m + wL_a \tag{1}$$

$w = 0$ のとき，主タスクのみの学習が行われるので，これをベースラインとして用い，マルチタスク学習を行なった場合と比較する．ラベルを予測する際には補助タスクは行わず，主タスクの出力のみを用いる．以下に主タスク，補助タスクの詳細について述べる．

3.1 主タスク

CHEMPROT では，タンパク質と化学物質のペアについて，関係を持つペアに対しては 5 種類の関係のうちどの関係にあてはまるかがアノテーションされている．そのため主タスクでは，タンパク質と化学物質を含む文を入力とし，2 つのエンティティの関係を (正例 5 クラス＋負例 1 クラス) の 6 クラスに分類するタスクを取り扱う．ここで，正例の 5 クラスは「エンティティ間に関係があり，CHEMPROT で定義される 5 種類の関係のうちどれに当てはまるか」を示し，負例は「CHEMPROT で定義される 5 種類の関係のどれにも当てはまらない」ということを示す．主タスクの損失関数には Cross Entropy Loss を用いる．

3.2 補助タスク

補助タスクでは入力文に Open IE で抽出できる関係が含まれているかどうかを分類する．まず，入力文

に対して Open IE ツール Stanford CoreNLP[8] を用い,（エンティティ1，エンティティ間の関係，エンティティ2）の 3 つ組を抽出する．このとき，エンティティ 1，2 にそれぞれタンパク質と化学物質が含まれている 3 つ組が抽出できた場合，入力文に対する補助タスクのラベルは正例ラベルとなる．それ以外の場合，すなわち Open IE で抽出した 3 つ組の中にタンパク質と化学物質をそれぞれエンティティ1，2 に含むものがない場合は負例となる．表 1 に例を示す．

以上で定義したラベルを用い，補助タスクではタンパク質と化学物質を含む文を入力とし，入力文が正例か負例かの 2 クラスに分類するタスクを取り扱う．補助タスクの損失関数には，主タスクと同様に Cross Entropy Loss を用いる．

4 実験

4.1 データセット

データセットとして CHEMPROT を用いる．CHEMPROT の統計量を表 2 に示す．

前処理は Lim ら [7] と同様のものを用いる．具体的には，以下の手順で前処理を行う．

1. アブストラクト中のタンパク質と化学物質を含む文を抽出する．
2. タンパク質，化学物質をそれぞれ “<PROT>”，“<CHEM>” に置換する．
3. 関係判定対象のタンパク質と化学物質のペアを決め，判定対象外のエンティティを “<OTHER>” に置換する．

図 3 の例では，文中にタンパク質が 2 つ，化学物質が 1 つ存在するため，関係判定対象の文が 2 つになる．

4.2 評価指標

評価には，主タスクの出力結果を用いる．評価指標としては，CHEMPROT で用いられており，一般的な関係抽出タスクの評価指標でもある精度，再現率，F 値を用いる．なお，評価の際は CHEMPROT で用いられている評価スクリプトを使用して値を算出した．

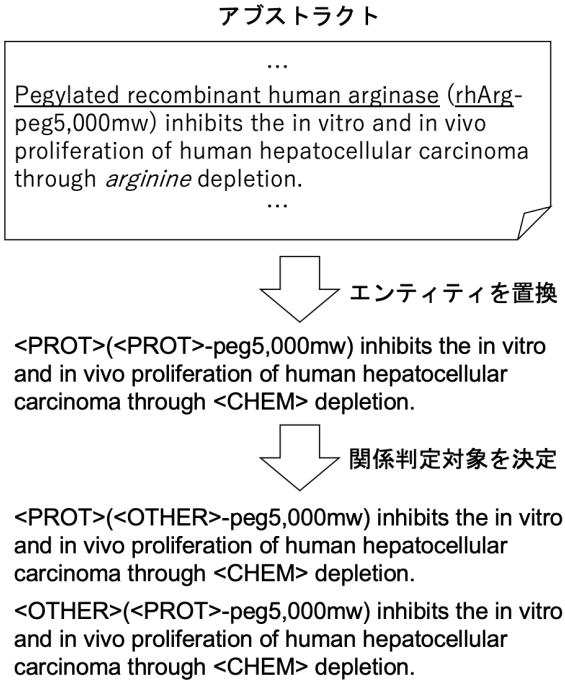

図 3: 前処理の例．下線がタンパク質，斜体が化学物質を表す．

4.3 実験設定

エンコーダーは BioBERT[6] を使用し，事前学習済みのモデルを用いて CHEMPROT によってファインチューニングを行う．式 (1) における補助タスクの重み w は，ベースラインでは主タスクのみでの関係抽出とするため $w = 0$，提案手法では $w = 0.05, 0.10, ..., 1.00$ のうち，開発データで最も F 値の高かった $w = 0.05$ を用い，バッチサイズは 16 とした．モデルの最適化には Adam[4] を用い，Adam の各パラメータに関しては学習率 $\alpha = 0.00005$，$\beta_1 = 0.9$，$\beta_2 = 0.999$，Weight decay $= 0.01$ とした．

4.4 結果

ベースラインと提案手法に加え，比較のために SPINN [7] と SciBERT [2][4]と BioBERT の元論文 [6] で報告されている v1.1 による結果を表 3 に示す．ベースラインと比較すると，マルチタスク学習によって F 値が向上していることがわかる．

また，主タスクと補助タスクの相関を調べるため，学習データ中の各タスクの正例・負例数を表 4 に示す．

[4]https://paperswithcode.com/paper/scibert-pretrained-contextualized-embeddings

表 1: 補助タスクのラベル付けの例

入力文	Open IE 出力	ラベル
<CHEM> inhibits the induction of <PROT>.	(<CHEM>, inhibits, induction of <PROT>) (<CHEM>, inhibits, induction)	正例
It is derived from <CHEM> and <PROT>.	(It, is derived from, <CHEM> and <PROT>)	負例

表 2: CHEMPROT の統計量

	訓練	開発	テスト
アブストラクト数	1,020	612	800
化学物質数	13,017	8,004	10,810
タンパク質数	12,735	7,563	10,018
関係判定対象文数	16,522	10,362	14,397
関係抽出正例文数	4,133	2,412	3,441
Open IE 正例文数	1,761	1,072	1,521

表 3: 実験結果

	適合率	再現率	F 値
SPINN	0.7480	0.5600	0.6410
SciBERT	-	-	0.7612
BioBERT v1.1	0.7702	0.7590	0.7646
ベースライン	0.7766	0.7510	0.7636
提案手法	0.7732	0.7730	0.7731

なお主タスクでの正例とは，CHEMPROT で定義されている 5 種類の関係に当てはまるものであり，逆にそれらの関係に当てはまらないものは負例となる．

表 4: 主タスクと補助タスクの相関

		主タスク		
		正例	負例	合計
補助タスク	正例	805	956	1,761
	負例	3,328	11,433	14,761
合計		4,133	12,389	16,522

主タスクでの正例発生確率と補助タスクでの正例発生確率でカイ二乗検定を用いて検定した結果，有意差が認められた ($p < 0.01$)．すなわち，主タスクで「関係あり」と判定されるエンティティのペアと，補助タスクにて Open IE によって関係が定義づけられるエンティティのペアの間には相関があり，マルチタスク学習によってこれらに共通の特徴量を学習したと考えられる．

5 おわりに

本研究では，タンパク質と化学物質間の関係抽出課題において，Open IE によって作成したラベル付きデータを用いてマルチタスク学習を行うことを提案した．CHEMPROT コーパスを用いた実験により，提案手法によって関係抽出の F 値を 0.773 に改善することができた．今後の課題としては，CHEMPROT 以外のデータでの提案手法の実験・評価などが挙げられる．

参考文献

[1] Michele Banko, Michael J Cafarella, Stephen Soderland, Matthew Broadhead, and Oren Etzioni. Open information extraction from the web. In *Ijcai*, Vol. 7, pp. 2670–2676, 2007.

[2] Iz Beltagy, Arman Cohan, and Kyle Lo. Scibert: Pretrained contextualized embeddings for scientific text. *arXiv preprint arXiv:1903.10676*, 2019.

[3] 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.

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

[5] Martin Krallinger, Obdulia Rabal, Saber A Akhondi, et al. Overview of the biocreative vi chemical-protein interaction track. In *Proceedings of the sixth BioCreative challenge evaluation workshop*, Vol. 1, pp. 141–146, 2017.

[6] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: pre-trained biomedical language representation model for biomedical text mining. *arXiv preprint arXiv:1901.08746*, 2019.

[7] Sangrak Lim and Jaewoo Kang. Chemical–gene relation extraction using recursive neural network. *Database*, Vol. 2018, , 2018.

[8] Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pp. 55–60, 2014.

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