構文情報とラベルなしデータを用いた化学分野の関係抽出

新城大希
東京工業大学 情報理工学院
shinjo.t.ab@m.titech.ac.jp

徳永健伸
東京工業大学 情報理工学院
take@c.titech.ac.jp

牧野拓哉
株式会社富士通研究所
makino.takuya@fujitsu.com

岩倉友哉
株式会社富士通研究所
iwakura.tomoya@fujitsu.com

1 はじめに

新材料や新薬の開発，材料を用いた製品開発に必要不可欠な化学物質の知識は，論文や特許などで日々発表される情報を専門家によって人手で整理し，管理されているのが現状である．[1)] このような人手作業を軽減するために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．

近年の関係抽出タスクでは，正解が付与されたコーパスを構築し，機械学習手法によって自動抽出モデルを作成するアプローチが主流である．たとえば，CHEMPROT コーパス [1] では，2,482 件の化学論文のアブストラクトに対して，タンパク質・化学物質間の関係がアノテーションされている．BioBERT [2] は大規模なバイオ分野の文献を用いて事前学習することで，CHEMPROT において，Wikipedia などの一般的な内容のテキストから事前学習した BERT [3] と比較し，高い精度を達成している．

BioBERT は関係抽出タスクで高い精度を示しているが，次の二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の問題である．化学文献からの関係抽出タスクはラベル付け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，構文的な情報を利用していないという点である．目的の関係抽出タスクでは構文的な情報が手掛かりになると期待できる [4, 5]．しかしながら，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ていない．また，単純に構文解析の結果を利用する場合は実行時間が増加するため，大規模データを効率的に処理することが難しくなる．

この二点をふまえ，本研究では，Open Information Extraction (Open IE) [6] 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」に加え，主タスクが対象とするエンティティのペアが Open IE で抽出されるかどうかを判定する「補助タスク」を導入したマルチタスク学習 [7] を行う．さらに，ラベルなしデータである PubMed に対してラベルを付与し，学習に利用する手法を提案する．提案手法は，ラベルなしデータおよび各タスクの学習データに対して Open IE が構文パタンに基づいて抽出した関係を学習時に利用する．そのため，目的の関係抽出に有効な構文情報を暗黙的に学習することが期待される．また，提案手法は構文情報は学習時のラベル作成のために利用し，抽出時には利用しないため，明示的な構文解析結果の利用で生じる抽出速度の低下を避けることができる．CHEMPROT，GAD，EU-ADR で評価した結果，BioBERT より高い精度が得られ，また，構文解析結果を用いる手法との比較では，高い精度を維持しつつ，最大 29 倍高速に抽出が行えることが示された．

2 関連研究

BERT [3] を分野固有の文書で訓練することによって，その分野のタスクが高性能になることが知られている [2, 8]．また，構文解析結果を考慮することで，一般ドメインの関係抽出において，精度改善が報告されている [4, 5]．Open IE [6] は，構文情報を用いて主語・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三つ組を抽出する．そのため，クラスが定義された教師あり学習とは異なり，種々の関係を抽出できる．

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

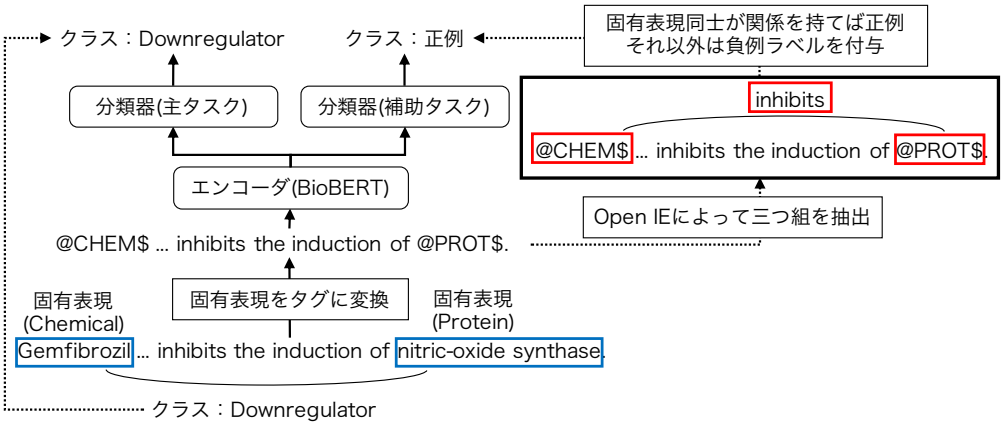

図 1　マルチタスク学習の概念図

3　提案手法

3.1　ベースモデル

本研究では主タスクの関係抽出と後述する補助タスクを同時に学習するマルチタスク学習をベースラインモデルとする [7]．まず関係を判定したい二つのエンティティをタグに変換する．変換後の文に対して Open IE によって補助タスクのラベルを付与し，両タスク共通のエンコーダの入力に用いる．エンコーダとして事前学習済みの BioBERT を用いる．BERT において用いられる [CLS] トークンがエンコーダの出力から得られるので，入力文の分散表現としてそれぞれの分類器の入力に用いる．共通のエンコーダとそれぞれの分類器を学習させ，各タスクの損失の重み付き和を最終的な損失関数とする．以上の手順を図 1 に示す．

主タスクでは関係抽出を行うが，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．補助タスクのラベルを生成するため，入力文に対して Open IE ツール Stanford CoreNLP [9] を用い，（主語，関係，目的語）の三つ組を抽出する．このとき，主語，目的語にそれぞれのエンティティを含む三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．それ以外の場合は負例となる．表 1 に例を示す．したがって，補助タスクではエンティティのペア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．各タスクの損失関数は Cross Entropy Loss を用いる．

3.2　ラベルなしデータの活用

本研究では CHEMPROT などのラベル付きデータに加え，ラベルなしデータである PubMed の活用を提案する．PubMed は生物医学ドメインのデータベースであり，2020 年 9 月時点で約 3,000 万件の文献が収録されている．このうちアブストラクトが取得できる文献をクローリングし，学習に利用する．利用方法は以下の通りである．

1. ラベル付きデータを用いて固有表現抽出器と関係抽出器を学習．抽出器はいずれも BioBERT で使用していたものを利用．
2. PubMed アブストラクトに対して 1 で学習したモデルで固有表現抽出を行い，関係抽出の判定対象のエンティティを含む文を取得．
3. 2 で取得した文に対して関係抽出を行い主タスクのラベルを付与．
4. 2 で取得した文に対して Open IE によって補助タスクのラベルを付与．
5. 主タスクと補助タスクのラベルが付与された PubMed を元のラベル付きデータと同時に学習．

PubMed の利用時は，関係ラベルの正例・負例の比率が元のラベル付きデータと同程度になるよう文を選ぶ．また，関係抽出ラベルを付与した際のスコアが高いものから順に選択し学習に利用する．

4　評価実験

4.1　データセット

評価実験では CHEMPROT に加えて GAD [10]，EU-ADR [11] をデータセットとして用いる．CHEMPROT は関係の種類が 6 クラス定義され

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

ているのに対し，GAD と EU-ADR は 2 クラスが定義されている．いずれのデータセットにおいても BioBERT で使われていた前処理済みのものを実験に用いる．実験ではこれらの CHEMPROT，GAD，EU-ADR に加え，CHEMPROT を関係あり・なしの 2 クラスに変換したデータセットを用いる．以降，本稿では元の 6 クラスの CHEMPROT を CHEMPROT-6，2 クラスに変換した CHEMPROT を CHEMPROT-2 と記す．

4.2 実験設定

CHEMPROT-6，CHEMPROT-2，GAD，EU-ADR の 4 つのデータセットについて，主タスクのみを行う場合とマルチタスク学習を行う場合のそれぞれについて学習時に PubMed を追加して実験を行う．PubMed を利用する際には，元データとの関連性の高い文を利用するために元データに出現するエンティティを含む文のみを用いる手法でも実験を行う．また，構文情報を与える方法として BioBERT の入力に構文木上の二つのエンティティを繋ぐ最短経路のトークンを追加した最短経路手法も比較として実験する．本実験では，各最短経路を BioBERT で encode した結果を結合し，softmax 層にて分類を行う．そのため，この方法では，その都度，構文解析が必要となる．

エンコーダは事前学習済みの BioBERT-Base v1.1[2)] を用いてそれぞれの学習データで fine-tuning を行った．GAD，EU-ADR においては開発データが存在しなかったため，学習データの 10%を開発データとして分割した．各データセットにおいて開発データでハイパーパラメータチューニングを行った．開発データで最も F 値が高くなったハイパーパラメータを用い，学習データに開発データを含めて再度学習したモデルによってテストデータの評価を行った．

CHEMPROT については乱数のシード値を変更した上で 5 回ずつモデルの学習・評価を行い，それらの F 値の平均を評価値として用いる．GAD，EU-ADR については 10 分割交差検証によって F 値を算出する．

2） https://github.com/dmis-lab/biobert

4.3 結果と考察

評価実験の結果を表 2 に示す．BioBERT は Lee et al. [2] の論文中の数値を用いた．

表 2 実験結果（F 値）．+common は PubMed 利用時，教師データに出現したエンティティ間に制限．

CHEMPROT-6					
手法	追加 PubMed データ量				
	0%	100%	200%	500%	1000%
BioBERT	0.7646	-	-	-	-
最短経路	0.7616	-	-	-	-
Single	0.7647	0.7616	0.7650	0.7621	0.7585
+common	-	0.7630	0.7653	0.7619	-
Multi	0.7657	0.7614	0.7638	**0.7666**	0.7636
+common	-	0.7629	0.7664	0.7655	-
CHEMPROT-2					
手法	追加 PubMed データ量				
	0%	100%	200%	500%	1000%
BioBERT	-	-	-	-	-
最短経路	0.7701	-	-	-	-
Single	0.7773	0.7815	0.7855	0.7825	0.7789
+common	-	0.7830	0.7855	0.7843	-
Multi	0.7848	0.7843	0.7848	**0.7882**	0.7850
+common	-	0.7840	0.7878	0.7858	-
GAD					
手法	追加 PubMed データ量				
	0%	100%	200%	500%	1000%
BioBERT	0.7983	-	-	-	-
最短経路	0.8132	-	-	-	-
Single	0.8014	0.8166	0.8182	0.8224	0.8277
+common	-	0.8149	0.8214	0.8313	**0.8418**
Multi	0.8058	0.8049	0.8151	0.8237	0.8281
+common	-	0.8181	0.8193	0.8291	0.8381
EU-ADR					
手法	追加 PubMed データ量				
	0%	100%	200%	500%	1000%
BioBERT	0.7974	-	-	-	-
最短経路	**0.8275**	-	-	-	-
Single	0.8123	0.7825	0.8044	0.8002	0.8159
+common	-	0.8067	0.8116	0.8243	-
Multi	0.7936	0.7867	0.7983	0.8103	0.8187
+common	-	0.8036	0.8256	0.8028	-

ここでマルチタスク学習における主タスクと補助タスクの正例・負例の相関を表 3 に示す．主タスクと補助タスクの相関についてカイ二乗検定を用いて有意水準 0.05 で検定したところ，CHEMPROT-2 では有意差が認められたが，GAD，EU-ADR では有意

表 3　主タスクと補助タスクの相関

CHEMPROT-2		主タスク		
		正例	負例	合計
補助タスク	正例	1,115	245	1,360
	負例	3,015	12,147	15,162
合計		4,130	12,392	16,522

GAD		主タスク		
		正例	負例	合計
補助タスク	正例	545	540	1,085
	負例	1,975	1,736	3,711
合計		2,520	2,276	4,796

EU-ADR		主タスク		
		正例	負例	合計
補助タスク	正例	48	13	61
	負例	187	70	257
合計		235	83	318

表 4　PubMed における主タスクと補助タスクの相関

CHEMPROT-2		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

GAD		主タスク		
		正例	負例	合計
補助タスク	正例	202	400	602
	負例	2,196	1,998	4,194
合計		2,398	2,398	4,796

EU-ADR		主タスク		
		正例	負例	合計
補助タスク	正例	62	16	78
	負例	18	222	240
合計		238	80	318

差は認められなかった．このことから，主タスクと補助タスクの相関が示されている CHEMPROT-2 においてはマルチタスク学習単体で，より大きく F 値が向上したと考えられる．

また，PubMed 追加なしのシングルタスク (Single) とマルチタスク (Multi) の比較から，CHEMPROT-6, CHEMPROT-2，GAD では F 値が向上していることがわかる．

PubMed 追加の効果については，Single 行の比較では，すべてのデータセットにおいてデータ追加なしの場合より高い F 値を示している．また，PubMed 利用時に，すべての文を使う場合 (Single, Multi) と教師データに出現するエンティティ 間の関係に制限する場合 (common) の比較では，CHEMPROT では，制限を加えない場合に最高精度が得られ，GAD, EU-ADR では制限が有効であった．データを追加することで F 値は基本的に上がっているものの，一定量追加したあとは F 値が下がる傾向にある．これは学習データを追加する際に関係抽出によってラベルを付与したときのスコアが高いものから順に追加していることに起因すると考えられる．

マルチタスク学習と PubMed の追加を組み合わせた場合については，CHEMPROT-6，CHEMPROT-2 では最も高い F 値を達成しているが，GAD ではシングルタスクでの PubMed 追加手法が，EU-ADR では最短経路手法がそれぞれ最高精度を達成している．PubMed に付与した主タスクと補助タスクのラベルについては，追加データ 100%で使用した追加 PubMed データのラベルの相関を表 4 に示す．主タスクと補助タスクの相関についてカイ二乗検定を用いて有意水準 0.05 で検定したところ，CHEMPROT-2，GAD，EU-ADR のいずれのデータセットにおいても有意差が認められた．このことから，PubMed の学習では両タスクに共通する特徴を学習し，ベースラインよりも高い F 値を達成しているのだと考えられる．

EU-ADR では提案手法がいずれも最短経路手法より F 値が下回っているが，最短経路手法は関係抽出ラベルを付与する際に構文解析の分だけ実行時間がかかることになる．そこで，最短経路手法とその他の手法について，学習データに関係ラベルを付与する際に要する実行時間を計測した[3)]．構文解析を用いないベースラインや提案手法では CHEMPROT, GAD，EU-ADR の順に 74，31，15 秒だったのに対し，最短経路手法では 1,692，451，52 秒だった．これらの結果から，提案手法により，構文情報を学習しつつ，大規模なラベルなしテキストを用いた学習が行えることがわかる．最短経路を利用しない同じ計算量となるモデル間での比較では，EU-ADR においてはマルチタスク学習と PubMed の追加を組み合わせた提案手法が最も F 値が高くなっている．

5　おわりに

本研究では Open IE によるラベル拡張とラベルなし大規模コーパスの利用によって関係抽出タスクの精度の向上を試みた．実験結果から，マルチタスク学習とデータ拡張を同時に行うことで，F 値の向上を確認した．今後の課題として化学ドメインでの文字情報の追加利用や，化学ドメイン以外への本手法の適用などが挙げられる．

3)　計測には Intel Xeon E5-2680 v4 2.4GHz CPU と NVIDIA TESLA P100 for NVlink-Optimized Servers を用いた．

参考文献

[1] 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.

[2] 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.

[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] Yan Xu, Lili Mou, Ge Li, Yunchuan Chen, Hao Peng, and Zhi Jin. Classifying relations via long short term memory networks along shortest dependency paths. In *EMNLP'15*, pp. 1785–1794, 2015.

[5] Makoto Miwa and Mohit Bansal. End-to-End Relation Extraction using LSTMs on Sequences and Tree Structures. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*, 2016.

[6] 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.

[7] 新城大希, 西川仁, 徳永健伸, 牧野拓哉, 岩倉友哉. 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. 言語処理学会第 26 回年次大会, pp. 1555–1558, 2020.

[8] Iz Beltagy, Kyle Lo, and Arman Cohan. SciBERT: A pretrained language model for scientific text. 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. 3615–3620, Hong Kong, China, November 2019. Association for Computational Linguistics.

[9] 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.

[10] Àlex Bravo, Janet Piñero, Núria Queralt-Rosinach, Michael Rautschka, and Laura I Furlong. Extraction of Relations between Genes and Diseases from Text and Large-scale Data Analysis: Implications for Translational Research. *BMC Bioinformatics*, 01 2015.

[11] Erik M. van Mulligen, Annie Fourrier-Réglat, David Gurwitz, Mariam Molokhia, Ainhoa Nieto, Gianluca Trifiró, Jan Kors, and Laura I Furlong. The EU-ADR Corpus: Annotated Drugs, Diseases, Targets, and their Relationships. *Journal of biomedical informatics*, Vol. 45, pp. 879–84, 04 2012.