近傍知識グラフからの埋め込みを統合利用する文書からの遠距離教師あり関係抽出

松原 拓磨　三輪 誠　佐々木 裕
豊田工業大学
{sd19082,makoto-miwa,yutaka.sasaki}@toyota-ti.ac.jp

概要

データベースから入力文書に出現する用語の近傍を含む近傍知識グラフを動的に構築し，その近傍知識グラフから獲得した用語の表現を入力文書に統合して利用する，新たな文書からの遠距離教師あり関係抽出モデルを提案する．提案手法により，遠距離教師あり関係抽出において作成したラベル付きテキストの言語情報のみを用いており，データベースに登録されている他の用語や関係に関する豊富な情報を活用できていない，という問題の解消を目指す．実験では薬物・疾患・遺伝子間の文書単位関係抽出データセットである ChemDisGene データセットにおいて，提案手法における用語の表現の統合による顕著な性能向上を確認した．

1　はじめに

様々な分野において，エンティティ間の新たな関係を報告する文書は日々爆発的に増え続けており，人手での分野データベースへの登録・整備が追いついていない．そのため，文書からの関係の自動抽出が求められており，その高い性能から機械学習を用いた手法が主流となっている．しかし，機械学習では大量のラベル付きデータを必要とするため，ラベル付けにコストがかかる．

このコストの削減のために，Mintz ら [1] は既存のデータベースを用いてラベルなしコーパスに機械的にラベル付けをした遠距離教師データを利用する遠距離教師あり関係抽出を提案した．遠距離教師データは，人手でラベル付けされたコーパスとは異なり，データベースに登録されている用語間の関係をもとにラベル付けをするため，**データベースに直接結び付けられた**ラベル付きコーパスである．しかし，既存の遠距離教師あり関係抽出ではコーパスのみを用いており，データベースに登録されている用語の性質などの他の豊富な情報を利用できていない．

そこで，本研究では，データベースに登録されている情報の関係抽出への活用を目指す．このために，データベース内の入力文書の全用語近傍の情報を**近傍知識グラフ**として動的に構築し，遠距離教師データに統合利用する新規の文書からの遠距離教師あり関係抽出モデルを提案する．近傍知識グラフを動的に構築することで，遠距離教師あり学習における対象用語ペア間の教師ラベルをデータベースの情報から選択的に排除できるとともに，用語ペアごとに参照するデータベース内の情報の量を抑えることができる．本研究の貢献は次の通りである．

- 入力文書の全用語近傍のデータベースの情報を表現する近傍知識グラフの動的構築に基づく，遠距離教師あり関係抽出のための用語の埋め込み表現の獲得手法を提案
- 近傍知識グラフから獲得した用語の表現を入力文書に統合し，利用する遠距離教師あり関係抽出モデルを提案
- 文書単位遠距離教師あり関係抽出データセットである ChemDisGene[2] で，用語の表現を追加することでの性能向上を確認

2　関連研究

2.1　遠距離教師あり関係抽出

遠距離教師あり関係抽出は Mintz ら [1] により提案され，人手でラベル付けされた教師データを必要としない関係抽出モデルの学習方法である．遠距離教師データはデータベースと大量のラベルなしコーパスを用いて作成される．ラベルなしコーパスで共起した 2 つの用語に対して，データベースで 2 つの用語間に関係が登録されているときに，2 つの用語間に関係をラベル付けする．

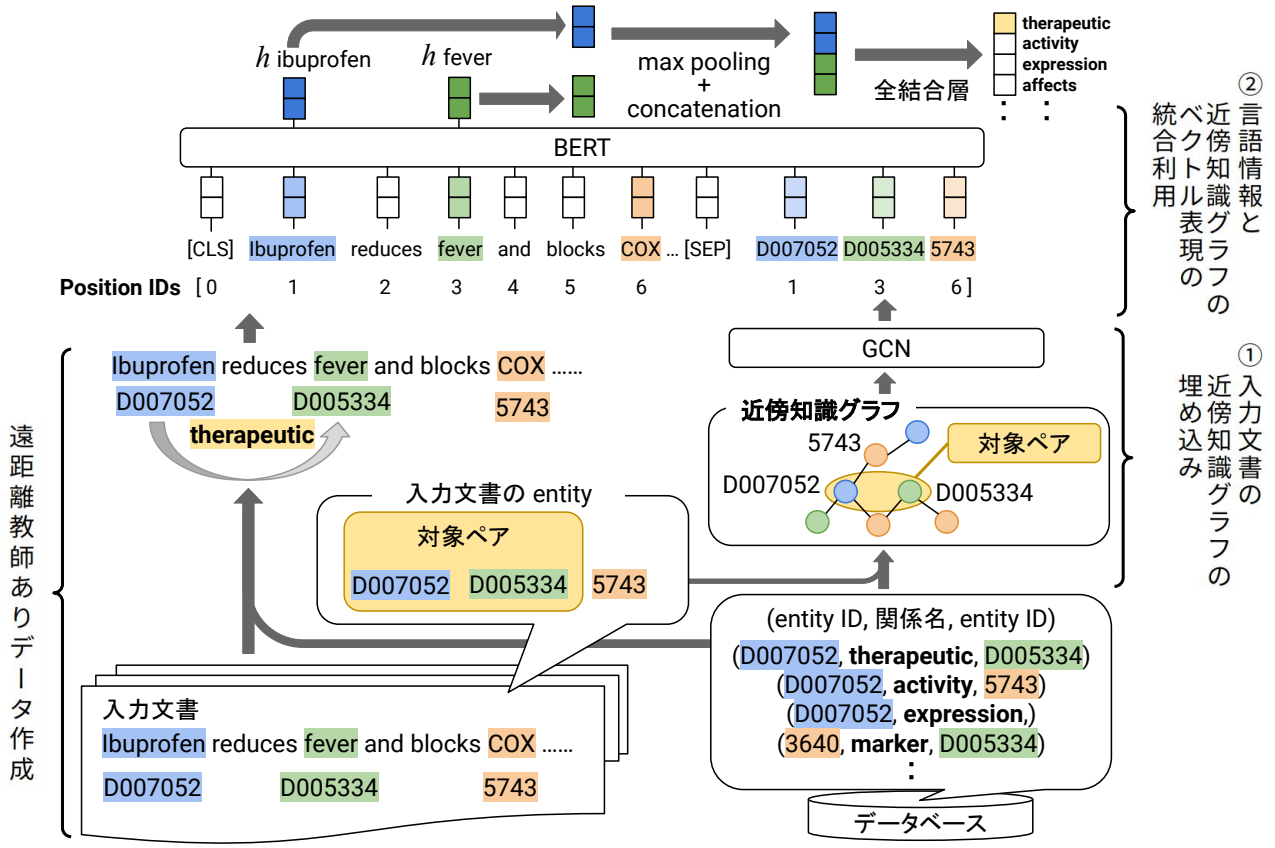

図 1　提案手法の概要．具体例として ChemDisGene [2] のデータを利用．

Zhang ら [2] は遠距離教師あり関係抽出データセットである ChemDisGene を提案し，テキストベースの関係抽出モデルの評価を行った．ChemDisGene は生物医学文献データベース Medline [3] に登録されている文献の題目・要旨に CTD (Comparative Toxicogenomics Database) [4] に登録されている薬物・疾患・遺伝子間の相互作用を割り当てた文書単位での遠距離教師あり関係抽出データセットである．評価のために人手でラベル付けされたテストデータを提供している．CTD は薬物・疾患・遺伝子・突然変異・代謝の相互作用に関する複数のデータベースを統合したデータベースである．PubTator [5] を用いて文献の題目・要旨に対して，用語のスパンと概念クラスのラベル付けがされている．また，関係抽出モデルは事前学習済みモデル BERT (Bidirectional Encoder Representations from Transformers) [6] を用いて，BERT，最大値プーリング，全結合層の構成としている．

2.2　知識グラフの表現学習

知識グラフの各ノードにベクトル表現を与え，ノード間の関係性を学習する手法として，Kipf らが提案したグラフ畳み込みネットワーク (Graph Convolutional Network; GCN) [7] が挙げられる．GCN はグラフ構造を表現するグラフニューラルネットワーク (Graph Neural Network; GNN) の 1 種であり，対象ノードの表現に隣接ノードの表現に重みをかけたものを畳み込むことでノード表現の更新を行う．

また，知識グラフを対象に対象ノード近傍で抽出したサブグラフを利用する手法が提案されている．例えば，Zhang らの提案したサブグラフに GNN を適用する SEAL (learning from Subgraphs, Embeddings and Attributes for Link prediction) [8] はリンク予測タスクにおいて高い性能を報告している．しかし，遠距離教師あり学習において用語の近傍のサブグラフを利用する手法は存在しない．

3　提案手法

本節では，本研究で新たに提案する，データベース内の入力文書近傍の情報を，入力文書に統合し利用する遠距離教師あり関係抽出手法について説明する．3.1 節で入力文書の近傍知識グラフ抽出（図 1-①）について説明し，3.2 節で入力文書と近傍知識グラフのベクトル表現の統合利用（図 1-②）につい

て説明する．

3.1 入力文書の近傍知識グラフの埋め込み

データベースに登録されている入力文書の全用語の近傍となるトリプルを抽出する．具体的にはそれぞれの用語について，一定ホップ数以内で繋がるトリプルの集合を抽出する．次に分類対象となる用語ペアについて，遠距離教師データの正解ラベルの情報となるペア間のトリプルを削除した上で，近傍知識グラフとし，GCN で埋め込みを行う．このようにすることで，データベース内の文書全体の用語の周辺情報と関係情報を利用できるとともに，トリプルの集合を文書ごとに一度事前計算することで関係ごとの近傍知識グラフの計算を簡略化できる．

3.2 言語情報と近傍知識グラフのベクトル表現の統合利用

GCN で埋め込んだエンティティに対応するノードのベクトル表現をテキストと同時に BERT に入力することで，遠距離教師データの言語表現と近傍知識グラフのベクトル表現を統合利用した関係抽出を行う．具体的には，BERT の入力テキストの後ろに “[SEP]” トークンを挟み，それぞれのテキスト内の用語に対応するノードのベクトル表現を用語の位置 ID と一致させながら追加する [9]．以降の関係抽出は既存研究 [2] と同様に行う．具体的には，まず，対象用語ペアの用語のトークンに対応する BERT の最終層のベクトルを取り出し，用語ごとに最大値プーリング層を通してそれぞれの用語の表現を得る．その後，用語の表現を結合し，全結合層を通して，関係ラベルに分類する．本手法では文献 [9] とは異なり，用語ペアに対応するエンティティペアの表現を分類に用いず，文献 [10] と同様，用語ペアの表現を分類に用いる．

学習においては，最適化手法に Adam [11] を用い，交差エントロピーを目的関数として，GCN と BERT を同時に学習する．

4 実験設定

遠距離教師あり関係抽出データセットとして，ChemDisGene [2] の 78,463 件の医学文献をもとに作成された遠距離教師データ (CTD-derived) のうち，76,942 件・1,521 件をそれぞれ訓練・開発データ，523 件の人手でラベル付けされたデータ (Curated) をテストデータとして学習・評価を行った．評価指標はマイクロ平均 F 値を用いた．予測する関係の種類は 14 種類であり，内訳は薬物と疾患の関係 (Chem-Dis) が 2 種類，薬物と遺伝子の関係 (Chem-Gene) が 10 種類，遺伝子と疾患の関係 (Gene-Dis) が 2 種類である．ChemDisGene データセットの統計を付録 A の表 4 に示す．

また，近傍知識グラフのエンティティには CTD の薬物・疾患・遺伝子を用いた．近傍知識グラフには，入力文書に出現する用語に対応するエンティティからそれぞれ 2 ホップ先のノードを含んだものを用いた．100 個以上のエッジを持つノードについてはランダムに 100 個のエッジに制限した．CTD のデータの統計を付録 A の表 5 に示す．

ベースラインモデルは提案モデルから近傍知識グラフの情報を除いた既存研究 [2] と同じモデルである．文献 [2] に従い，BERT には，データセットと近いドメインで事前学習された PubMedBERT [12] を用いた．PubMedBERT の埋め込み次元は 768 次元，テキストの最大長は 512 である．GCN の入力と出力のベクトルはともに 768 次元とし，GCN は 2 層とした．実験環境の詳細は付録 B に示す．

5 結果と考察

Curated テストデータを用いて評価した文献からの薬物・疾患・遺伝子間の関係抽出の性能を表 1 に示す．提案手法により近傍知識グラフの情報を用いないベースラインから F 値のマイクロ平均が全体で 1.04%ポイント向上し，一番上がり幅の大きい Chem-Gene: activity-decreases については F 値が 6.81%ポイント向上した．この結果より，近傍知識グラフを用いることで予測性能が向上しており，背景的な情報を考慮した関係抽出ができていると考えられる．

データベースに含まれているトリプル以外の関係への影響を調べるために，CTD に登録されていない事例についての予測を確認したところ，テストデータにおいて，ベースラインでは CTD に登録されていない関係は 1 つも抽出できなかったのに対し，提案手法ではこのような関係を 3 件抽出できていることがわかった．この結果より，CTD の背景的な情報を考慮することで，CTD に登録されていない関係も抽出できるようになったと考えられる．実際に提案手法によって抽出できるようになった表 2 に示す．これらの事例の多くは題目で言及されているものが多かった．

最後に表 3 に，既存研究 [2] との比較を示す．提

表 1 人手でラベル付けされたテストデータ Curated における評価．F 値には 5 回の評価の平均と標準偏差を示した．

	ベースライン			提案手法		
	PubMedBERT			PubMedBERT+近傍知識グラフ		
関係名	P [%]	R [%]	F 値 [%]	P [%]	R [%]	F 値 [%]
Chem-Dis: marker/mechanism	79.65	30.61	44.18 ± 2.29	82.38	31.22	45.16 ± 3.37
Chem-Dis: therapeutic	79.35	23.11	35.60 ± 3.09	79.08	23.65	36.30 ± 2.65
Chem-Gene: activity-decreases	70.93	23.73	35.15 ± 4.19	62.91	31.72	41.96 ± 2.18
Chem-Gene: activity-increases	73.07	30.39	42.74 ± 3.48	77.33	27.34	40.34 ± 3.78
Chem-Gene: binding-affects	73.49	31.79	44.21 ± 4.04	60.27	43.47	48.60 ± 3.70
Chem-Gene: expression-affects	19.05	3.25	5.56 ± 8.33	14.20	3.25	5.04 ± 6.14
Chem-Gene: expression-decreases	77.81	38.30	51.16 ± 1.90	74.67	37.08	49.42 ± 1.59
Chem-Gene: expression-increases	63.21	47.17	53.95 ± 1.32	63.43	46.73	53.81 ± 1.07
Chem-Gene: localization-affects	51.40	43.09	46.75 ± 4.22	59.07	44.72	49.26 ± 5.28
Chem-Gene: metabolic processing-decreases	54.12	57.90	55.89 ± 2.83	57.12	49.12	52.07 ± 1.91
Chem-Gene: metabolic processing-increases	41.52	34.13	37.30 ± 3.39	40.83	45.36	40.04 ± 2.30
Chem-Gene : transport-increases	46.89	36.59	40.61 ± 8.13	44.67	38.21	39.13 ± 3.92
Gene-Dis: marker/mechanism	83.75	20.27	32.27 ± 8.25	83.79	23.88	36.72 ± 6.98
Gene-Dis: therapeutic	53.33	1.63	3.10 ± 3.48	66.67	1.22	2.39 ± 1.94
マイクロ平均	69.51	30.62	42.47 ± 0.16	67.29	32.17	**43.51** ± **0.33**

表 2 CTD に登録されていない関係の正解例

関係	題目・要旨
Gene-Dis: therapeutic Gene: mir-543 Disease: cervical cancer	**miR-543** inhibits **cervical cancer** growth and metastasis by targeting TRPM7. Dysregulation of **miR-543** has been implicated to play crucial roles in various human cancers. · · ·
Chem-Gene: transport-increases Chemical: glucose Gene: Fibroblast growth factor 21	**Fibroblast growth factor 21** secretion enhances **glucose** uptake in mono(2-ethylhexyl)phthalate-treated adipocytes. Previous studies revealed that cellular accumulation of mono(2-ethylhexyl)phthalate (MEHP) disturbed energy metabolism in adipocytes, where **glucose** uptake was significantly increased.· · ·

表 3 Curated テストデータでの既存研究 [2] との比較

	F 値 [%]
PubMedBERT	42.5 ± 0.2
PubMedBERT+近傍知識グラフ	43.5 ± 0.3
PubMedBERT [2]	42.1
PubMedBERT+BRAN [2]	**43.8**

案手法は PubMedBERT+BRAN に比べて平均は若干低い F 値となったがほぼ同等の性能となっており，本提案手法は PubMedBERT+BRAN にも導入が可能であることを考えると，良好な結果を得られていると言える．

6 おわりに

本研究では，データベースに登録されている豊富な情報の有効利用を目的として，入力文書周辺の近傍知識グラフの表現を遠距離教師データに統合利用する遠距離教師あり関係抽出モデルを提案した．提案した手法を ChemDisGene データセットで学習・評価を行った結果，近傍知識グラフの表現を追加することでマイクロ F 値が 1.04%ポイント向上し，最先端モデルとほぼ同等の性能を達成した．また，提案手法ではテキスト情報だけでのモデルでは抽出ができなかった，データベースに登録されていない関係を抽出することができていることを確認した．

今後は，最先端モデルへの本手法の適用を行うとともに，抽出結果をデータベースからのトリプルに追加し，再びモデルを学習するなど，文献内の情報とデータベース内の情報の連携を深め，データベース内の情報のさらなる有効利用を目指す．

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] 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.

[2] Dongxu Zhang, Sunil Mohan, Michaela Torkar, and Andrew McCallum. A distant supervision corpus for extracting biomedical relationships between chemicals, diseases and genes. In **Proceedings of the Thirteenth Language Resources and Evaluation Conference**, pp. 1073–1082, Marseille, France, June 2022. European Language Resources Association.

[3] NCBI Resource Coordinators. Database resources of the national center for biotechnology information. **Nucleic Acids Res.**, 1 2018.

[4] 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, 2021.

[5] Chih-Hsuan Wei, Alexis Allot, Robert Leaman, and Zhiyong Lu. PubTator central: automated concept annotation for biomedical full text articles. **Nucleic Acids Research**, Vol. 47, No. W1, pp. W587–W593, 05 2019.

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

[7] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. **ICLR**, 2017.

[8] Muhan Zhang and Yixin Chen. Link prediction based on graph neural networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, **Advances in Neural Information Processing Systems**, Vol. 31, pp. 5171–5181. Curran Associates, Inc., 2018.

[9] Zexuan Zhong and Danqi Chen. A frustratingly easy approach for entity and relation extraction. In **Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 50–61, Online, June 2021. Association for Computational Linguistics.

[10] Masaki Asada, Makoto Miwa, and Yutaka Sasaki. Integrating heterogeneous knowledge graphs into drug–drug interaction extraction from the literature. **Bioinformatics**, Vol. 39, No. 1, 11 2022. btac754.

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

[12] 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 Trans. Comput. Healthcare**, Vol. 3, No. 1, oct 2021.

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

[14] 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.

A　データセットの統計

データセットの統計を表 4 と 5 に示す．

表 4　文書単位の遠距離教師あり関係抽出データセット ChemDisGene．テストのみ人手でタグづけされている．

データ	論文	薬物	疾患	遺伝子	関係
訓練	76,942	7,187	2,413	5,391	167,005
開発	1,521	759	283	852	3,290
テスト	523	670	318	887	3,833

表 5　本実験で使用したデータベース CTD の統計

ドメイン	unique head	unique tail	トリプル数
Chem-Gene	14,346	53,832	2,274,465
Chem-Disease	10,249	3,285	104,186
Gene-Disease	8,807	5,857	33,449

B　実験環境

本実験を行った環境について説明する．Python [13] のバージョン 3.8.10 と深層学習ライブラリである PyTorch [14] のバージョン 1.9.0 を用いて実装し，表 6 に示すハードウェア環境で実験した．本実験のハイパーパラメータを表 7 に示す．Curated の関係抽出スコア表 1 は遠距離教師データで学習したモデルを 5 個用意し，それぞれのモデルでテストしたスコアの平均と標準偏差を報告した．

表 6　本実験に用いたハードウェア

内容	項目
OS	Ubuntu 20.04.2 LTS
GPU	NVIDIA TITAN V
GPU メモリ	12GB
CPU	Intel(R) Core(TM) i9-7900X
メモリ	128GB

表 7　本実験におけるハイパーパラメータ

ハイパーパラメータ	値
エポック数	10
学習率	1e-5
ドロップアウト率	0.1
weight decay	1e-4