他文書の予測を知識グラフに蓄積・利用する文書単位関係抽出

松原 拓磨　辻村 有輝　三輪 誠　佐々木 裕
豊田工業大学
{sd23439,sd18602,makoto-miwa,yutaka.sasaki}@toyota-ti.ac.jp

概要

従来の文書単位深層関係抽出では，文書を個別に処理しており，他の文書で予測した関係を利用できていない．そこで，本研究では，再学習せずに予測時に他の文書で予測した関係を知識グラフに関連付けることで，低コストに処理単位を超えて情報を共有する文書単位の関係抽出を提案する．実験では生物医学分野の文書単位関係抽出データセットである BioRED データセットにおいて，他の文書で予測した関係による関係予測への影響を確認した．

1　はじめに

様々な分野で，新たなエンティティ間の関係を報告する文書は増え続け，人手でのデータベースの整備が追いついていない [1, 2, 3]．そのため，データベースの整備の補助を目的として，文書全体の情報を考慮した文書からの関係の自動抽出が求められており，その高い性能から深層学習を用いた手法が主流となっている．しかし，既存の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エンティティの性質などの他の豊富な情報を利用できていない．そのために，これらの豊富な情報の有効性を調査することを目的とした関係抽出が研究がされており，対象のエンティティ間を予測する際に，それらの周辺情報を利用できるため，高い性能を達成している [4, 5]．これらの研究では，知識グラフにおけるエンティティの周辺情報の有用性が示されている一方で，他の文書に記述された関係情報やエンティティの周辺情報は利用されていない．他の文書に記述された情報は，他の文書を処理した際に抽出されているため，その抽出された情報を利用できれば良いと考えられる．しかし，従来の教師あり関係抽出では文書を個別に処理しているため，処理単位を超えた情報の共有が起こらず，予測時に抽出された情報が利用されることはない．

他の文書の情報を利用する手法としては，複数文書を同時に処理する手法 [6] があるが，大量の計算リソースが必要であり，また，他の文書の情報をデータベースの知識と関連付けることもできない．また，半教師あり学習を用いた手法 [7, 8] では，他の文書の情報をモデルが考慮するために再学習が必要である．

そこで，本研究では関係抽出における，知識グラフに関連付けられた他の文書で予測した関係による影響の調査を目的に，**再学習のコストなし**に，**予測時**に他の文書で予測した関係を知識グラフに関連付けることで，低コストに処理単位を超えて情報を共有する**文書単位**の関係抽出を提案する．具体的には，既存のデータベースに登録されている関係情報で表現された知識グラフに，既存の関係抽出モデル [5] を用いて他の文書の予測を追加し，再び対象文書の関係抽出を行う．このとき，他の文書から予測した信頼性の低い情報が知識グラフに追加され，他の文書の予測結果によって動的に変化する知識グラフにモデルの再学習なしに対応できるようにするため，2 つの工夫を行う．1 つは訓練時に他の文書の予測結果を利用するための工夫である，評価時には対象文書以外の文書の予測結果を知識グラフに追加して利用するが，訓練時にはそのような他の文書の予測結果は存在しない．訓練時に予測された関係を利用する状況を再現するために，訓練データ内で交差検証を行うことで擬似的な予測データを作成し，利用する．もう 1 つは，知識グラフに追加する信頼性の低い情報に対する工夫である．知識グラフに追加する予測は正しい情報とは限らないため，関係抽出モデルの予測確率を予測の信頼度として，信頼度の高いものほど知識と同じように扱う．具体的にはエッジの重みで信頼度を表現し，エッジの種類は予測した関係ラベルに対応させる．

本研究の貢献は次の通りである．

- 予測時に，モデルの再学習なしに，他の文書で予測した関係を知識グラフと関連付けることで，他の文書で予測した関係を処理範囲を超えて低コストで利用できる新しい関係抽出手法を提案
- 生物医学分野の文書単位関係抽出データセットである BioRED [9] で，他の文書で予測した関係による関係予測への影響を確認

2 関連研究

2.1 外部知識を利用した関係抽出

他の文書の情報を利用する関係抽出手法として，検索拡張生成 (Retrieval Augmented Generation; RAG) [10] を用いる手法があるが，入力として利用できる知識は限定的であり，検索して取得した文書で表現された関係を抽出し，対象の文書の関係抽出に利用するのには計算コストがかかる．

また，大規模言語モデルで生成した回答をメモリ削減のためにトリプル形式でメモリに保存し，次の回答ではそのトリプルを利用する手法 [11] がある．しかし，この手法では予測を保存し，再び利用するだけであり，予測を知識グラフとして扱って推論していない．

2.2 近傍知識グラフからの埋め込みを統合利用する関係抽出

著者らは，データベースにおける入力文書に現れるエンティティ近傍の情報を表現した近傍知識グラフと入力文書を統合して利用する関係抽出手法を提案した [5]．概要を図 1 に示す．

まず，近傍知識グラフについては，データベースに登録されている入力文書に現れる全エンティティの近傍にあるトリプルを抽出する．具体的にはそれぞれのエンティティについて，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エンティティペアについて，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，グラフ畳み込みネットワーク (Graph Convolutional Network; GCN) [12] で埋め込みを行う．このようにすることで，文書に現れる全エンティティに対して，データベース内の周辺情報と関係情報を含む表現を獲得する．

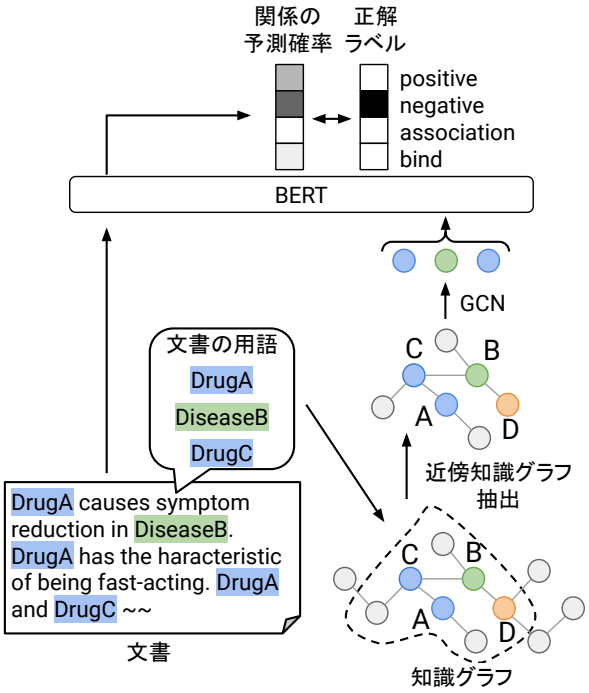

図 1 近傍知識グラフからの埋め込みを統合利用する関係抽出の概要 [5]

次に，近傍知識グラフについて GCN で埋め込んだエンティティの表現を入力文書に統合して，関係抽出を行う．具体的には，入力テキストにそれぞれのテキスト内のエンティティに対応するノードのベクトル表現をエンティティの位置 ID と一致させながら追加し，事前学習済みモデル BERT (Bidirectional Encoder Representations from Transformers) [13] に同時に入力する [14]．そして，BERT の出力のエンティティペアの表現を用いて，関係を分類する．学習においては GCN と BERT を同時に学習する．

3 提案手法

本研究では関係抽出における，知識グラフに関連付けられた他の文書で予測した関係による影響を調査するために，他の文書で予測した関係を知識グラフに関連付けることで処理単位を超えて情報を共有する関係抽出モデルを提案する．提案手法は 2.2 節の知識グラフを用いた関係抽出手法を拡張した手法となっている．

3.1 知識グラフに追加する予測データの作成

訓練時・評価時において，知識グラフに追加する予測データの作成方法について説明する．概要を図 2 に示す．

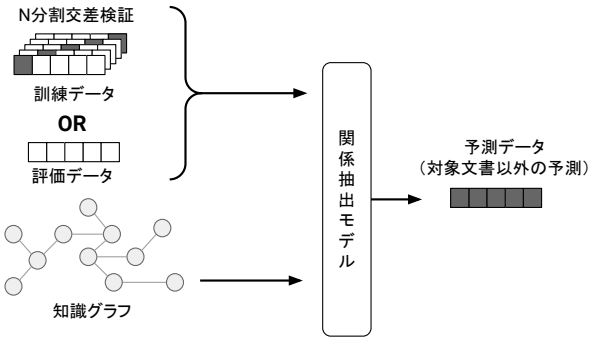

図 2　知識グラフに追加する予測データの作成方法

3.1.1　訓練

本提案手法では，評価時には他の文書の予測結果を知識グラフに追加して利用するが，訓練時にはそのような他の文書の予測結果は存在しない．訓練時に他の文書で予測された関係を利用する状況を再現するために，訓練データ内で N 分割交差検証を行うことで擬似的な予測データを作成し，利用する．具体的には，訓練データを $1 : N - 1$ に分割し，交差検証内の「予測データ」と「訓練データ」として，既存の関係抽出モデル [5] を「訓練データ」で学習し，「予測データ」に対して関係予測を行う．そして，「予測データ」と「訓練データ」を同様に被らないように分割し，再び「予測データ」に対して関係予測を行う．これを N 回繰り返し，訓練データ内の全文書に対する予測データを得る．

3.1.2　評価

評価時には予測結果を知識グラフに追加する．評価を複数回行い，その都度，信頼度が閾値以上であると予測された関係を知識グラフに追加していく．信頼度は，予測の信頼度がそれまでの予測の信頼度を上回った場合に更新する．知識グラフに追加する回数を n 回とする．

3.2　予測データが追加された知識グラフを用いた関係抽出

概要を図 3 に示す．まず，3.1 節で作成した予測データを既存の知識グラフに追加する．このとき，知識グラフに追加する予測は正しい情報とは限らないため，関係抽出モデルの予測確率を予測の信頼度として，信頼度の高いものほど知識と同じように扱う．具体的にはエッジの重みで信頼度を表現し，エッジの種類は予測した関係ラベルに対応させる．また，予測確率が閾値 η 以上の予測を知識グラフに追加する．次に，対象文書の関係予測ペアに対応する知識グラフのエッジを削除して，2.2 節と同様に対象文書の関係抽出を行う．

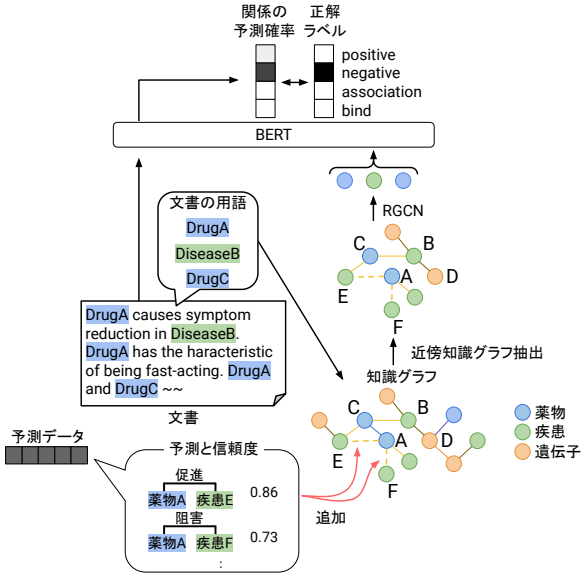

図 3　予測データが追加された知識グラフを用いた関係抽出

2.2 節で説明した既存研究ではエッジの種類を区別しない GCN を用いていたが，知識グラフに追加する関係タイプを区別するために，本研究では GCN の代わりにエッジの種類を区別する RGCN [15] の適用も検討した．エッジの種類を無視する GCN を用いずに，エッジの種類を考慮できる RGCN を用いることで，他の文書で予測した関係情報を区別して考慮することができる．こうして，他の文書で予測した関係を知識グラフに関連付けることで処理単位を超えて情報を共有する関係抽出を実現する．

4　実験設定

文書単位関係抽出データセットである BioRED [9] を対象に学習・評価を行った．BioRED の設定に従い，600 件の医学文献のうち，400 件・100 件・100 件をそれぞれ訓練データ・開発データ・テストデータとした．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 8 種類，対象エンティティは薬物・疾患・遺伝子・変異体の 4 種類である．BioRED データセットの統計を付録 A の表 3 と 4 に示す．

また，知識グラフのエンティティには CTD (Comparative Toxicogenomics Database) [1] 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エンティティに対応するノードからそれぞれ 2 ホップ先のノードを含んだものを用いた．CTD のデータの統計を付録 A の表 5 に示す．ベー

表 1　ベースラインモデルとの比較．($\eta = 0.80, n = 1$)
F 値には 5 回の評価の平均と標準偏差を示した．

	F 値 [%]
テキスト	43.4 ± 0.8
+ 近傍知識グラフ (GCN) [5]	44.5 ± 1.0
+ 近傍知識グラフ (GCN) + 提案手法	44.2 ± 0.9
+ 近傍知識グラフ (RGCN)	43.3 ± 1.1
+ 近傍知識グラフ (RGCN) + 提案手法	42.8 ± 1.2

スラインモデルは知識グラフの変化しない既存手法の関係抽出モデル [5] である．訓練時には 10 分割交差検証 (N=10) により，知識グラフに追加する予測データを作成した．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 [16] 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．文献 [9] では関係ペアごとに BERT へテキストを入力しており，予備実験でも BERT へテキストを入力するタイミングを関係ごとではなく，文書ごとにすることで F 値は低下することは分かっているが，計算時間短縮のため，文書ごとに BERT へテキストを入力した．既存研究 [5] で用いた GCN と本提案手法で用いる RGCN 両方での評価を行った．RGCN の関係タイプは（薬物，薬物），（薬物，疾患），（薬物，遺伝子），（疾患，遺伝子），（遺伝子，遺伝子）の 5 種類である．また，入力と出力のベクトルはともに 768 次元とし，RGCN は 2 層とした．実験環境の詳細は付録 C に示す．

5　結果と考察

提案手法とベースラインモデルの比較を表 1 に示す．η と n のチューニング結果は付録 B に示す．F 値のマイクロ平均での評価において，GCN・RGCN ともに+近傍知識グラフと+近傍知識グラフ +提案手法のスコアに差はなく，提案手法とベースラインの間に抽出性能の差は確認できなかった．また，GCN より RGCN の方が F 値が低くなる要因としては，BioRED データセットが十分なサイズではなかったことによる RGCN の過学習が一つの要因として考えられる．そのため，今後の課題としては大規模なデータセットでの実験を行うことで，原因を究明する．

知識グラフに追加する回数 n に対する F 値のマイクロ平均の結果を表 7 に示す．知識グラフに追加する回数が増えるほど，F 値が小さくなっている．これは知識グラフに追加するノイズが増えることにより，間違った推論をするようになったと考えられる．

表 2　提案手法により改善した事例（上）“関係なし””と間違えていたが，“Negative””と予測．
提案手法により改悪した事例（下）“Negative””と予測し，正解していたが，“関係なし””と予測．

題目・要旨
THP (薬物) exhibited an antipsychotic-like profile by potentiating haloperidol-induced catalepsy, reducing **amphetamine** (疾患) -induced hyperactivity and reducing apomorphine-induced ···
Findings suggest a potential for **H(3)-receptor antagonists** (薬物) in improving the refractory cases of **schizophrenia** (薬物). ···

事例数では，平均で 90.2 件改善し，105.8 件改悪していた．この結果から，評価時に他の文書の予測を知識グラフと関連付けることによる予測の変化を確認した．改善・改悪した事例を表 2 に示す．改善した事例について，THP（薬物）と amphetamine（疾患）の間の関係を “関係なし””と間違えていたが，“Negative””と予測し，正解するようになった．これは他の文書の予測を知識グラフにエッジとして加えたことにより，他の文書のエンティティの周辺情報を考慮できたからであると考察する．また，改悪した事例について，antagonists（薬物）と schizophrenia（薬物）の間の関係を “Negative””と予測し，正解していたが，“関係なし””と予測し，間違えるようになった．これは間違った予測が知識グラフに追加された影響であると考察する．

6　おわりに

本研究では，関係抽出における知識グラフに関連付けられた他の文書で予測した関係による影響の調査を目的として，他の文書で予測した関係を知識グラフに関連付けることで処理単位を超えて情報を共有する関係抽出モデルを提案した．提案した手法を BioRED データセットで学習・評価を行った結果，抽出性能に大きな差は確認できなかったが，再学習を行わず，評価時に他の文書の予測を知識グラフと関連付けることによる予測の変化を確認した．

他の文書で予測した関係を知識グラフとより関連づけるために，近傍知識グラフの範囲を超えたグラフ内のエンティティ間の関係を予測するリンク予測を行うなど，文献内の情報と知識グラフの情報の連携を深め，他の文書で予測した関係の有効利用を目指す．

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] 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.

[2] David S Wishart, Yannick D Feunang, An C Guo, Elvis J Lo, Ana Marcu, Jason R Grant, Tanvir Sajed, Daniel Johnson, Carin Li, Zinat Sayeeda, Nazanin Assempour, Ithayavani Iynkkaran, Yifeng Liu, Adam Maciejewski, Nicola Gale, Alex Wilson, Lucy Chin, Ryan Cummings, Diana Le, Allison Pon, Craig Knox, and Michael Wilson. DrugBank 5.0: a major update to the DrugBank database for 2018. **Nucleic Acids Research**, Vol. 46, No. D1, pp. D1074–D1082, 11 2017.

[3] Elisabeth Coudert, Sebastien Gehant, Edouard de Castro, Monica Pozzato, Delphine Baratin, Teresa Neto, Christian J A Sigrist, Nicole Redaschi, Alan Bridge, and The UniProt Consortium. Annotation of biologically relevant ligands in UniProtKB using ChEBI. **Bioinformatics**, Vol. 39, No. 1, 12 2022. btac793.

[4] 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.

[5] Takuma Matsubara, Makoto Miwa, and Yutaka Sasaki. Distantly supervised document-level biomedical relation extraction with neighborhood knowledge graphs. In Dina Demner-fushman, Sophia Ananiadou, and Kevin Cohen, editors, **The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks**, pp. 363–368, Toronto, Canada, July 2023. Association for Computational Linguistics.

[6] Yuan Yao, Jiaju Du, Yankai Lin, Peng Li, Zhiyuan Liu, Jie Zhou, and Maosong Sun. CodRED: A cross-document relation extraction dataset for acquiring knowledge in the wild. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, **Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing**, pp. 4452–4472, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[7] Qingyu Tan, Ruidan He, Lidong Bing, and Hwee Tou Ng. Document-level relation extraction with adaptive focal loss and knowledge distillation. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, **Findings of the Association for Computational Linguistics: ACL 2022**, pp. 1672–1681, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[8] Qi Sun, Kun Huang, Xiaocui Yang, Pengfei Hong, Kun Zhang, and Soujanya Poria. Uncertainty guided label denoising for document-level distant relation extraction. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, **Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 15960–15973, Toronto, Canada, July 2023. Association for Computational Linguistics.

[9] Ling Luo, Po-Ting Lai, Chih-Hsuan Wei, Cecilia N Arighi, and Zhiyong Lu. BioRED: a rich biomedical relation extraction dataset. **Briefings in Bioinformatics**, Vol. 23, No. 5, 07 2022. bbac282.

[10] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, **Advances in Neural Information Processing Systems**, Vol. 33, pp. 9459–9474. Curran Associates, Inc., 2020.

[11] Ali Modarressi, Ayyoob Imani, Mohsen Fayyaz, and Hinrich Schütze. Ret-llm: Towards a general read-write memory for large language models, 2023.

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

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

[14] 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.

[15] Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling. Modeling relational data with graph convolutional networks. In **The Semantic Web: 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3–7, 2018, Proceedings 15**, pp. 593–607. Springer, 2018.

[16] 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.

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

[18] 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

表 3　関係ラベルごとの関係ペア数の統計

	train	dev	test
Association	2,192	560	635
Positive_Correlation	1,089	352	325
Bind	61	19	9
Negative_Correlation	763	216	171
Comparison	28	5	6
Conversion	3	-	1
Cotreatment	31	10	14
Drug_Interaction	11	-	2

表 4　エンティティペアごとの関係ペア数の統計

	train	dev	test
(疾患, 遺伝子)	1,045	293	295
(薬物, 遺伝子)	621	154	148
(薬物, 疾患)	812	236	189
(疾患, 変異体)	538	184	171
(薬物, 薬物)	308	86	94
(変異体, 変異体)	12	3	10
(遺伝子, 遺伝子)	792	195	240
(薬物, 変異体)	49	11	16

表 5　本実験で使用したデータベース CTD の統計

エンティティペア	エッジ数
（薬物，遺伝子）	2,274,465
（薬物，疾患）	104,186
（疾患，遺伝子）	33,449

B　チューニング結果

表 6　予測確率の閾値 η に対する F 値のマイクロ平均 ($n = 1$)

予測確率の閾値 η	F 値 [%]
0.5	42.1 ± 1.7
0.6	42.6 ± 1.3
0.7	42.8 ± 1.5
0.8	42.8 ± 1.2
0.9	42.7 ± 1.0

表 7　知識グラフに追加する回数 n に対する F 値のマイクロ平均 ($\eta = 0.80$)
$n = 0$ は知識グラフの変化しないベースラインモデル

知識グラフに追加する回数 n	F 値 [%]
0	43.3 ± 1.1
1	42.8 ± 1.2
2	42.3 ± 1.2
3	42.0 ± 0.8

C　実験環境

本実験を行った環境について説明する．Python [17] のバージョン 3.10.6 と深層学習ライブラリである PyTorch [18] のバージョン 2.1.0 を用いて実装した．本実験のハイパーパラメータを表 8 に示す．

表 8　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	30
学習率	1e-5
ドロップアウト率	0.25
GCN・RGCN の入力と出力の埋め込み次元	768
GCN・RGCN の層数	2
サブグラフのホップ数	2