様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

# 科学研究費助成事業　　研究成果報告書

令和　６ 年　 ６ 月 ２０ 日現在

機関番号：３３９２４
研究種目：基盤研究(C)（一般）
研究期間：2020～2023
課題番号：２０Ｋ１１９６２

研究課題名（和文）巨大な異種混合グラフの深層表現学習による薬物関係抽出

研究課題名（英文）Drug relation extraction using deep representation learning on a huge heterogeneous graph

研究代表者
三輪　誠（Miwa, Makoto）

豊田工業大学・工学（系）研究科（研究院）・教授

研究者番号：００５２９６４６

交付決定額（研究期間全体）：（直接経費）　　3,400,000円

研究成果の概要（和文）：薬物データベースをもとに，薬物の名前・説明文・カテゴリ・関係などの薬物固有の情報に加え，薬物と関連するタンパク質などの様々な種類の情報を含む異種混合グラフを作成し、薬物の説明文や化学式の情報も含めて、すべての情報を有効活用できる表現学習を実現した。
また、薬物データベースの情報を文書からの薬物関係抽出に利用する手法について研究を進め、知識グラフから学習した表現を入力する文書と対応付けたものを入力として、文書からの情報抽出を行う単一の統一されたモデルを実現した。当初はF値80%以上の抽出性能を数値目標としていたが、最終的に85.40%というF値を達成し、当初の数値目標を大幅に上回ることができた。

研究成果の学術的意義や社会的意義
外部データベースに含まれる文書や化学式など異なる形式の情報をそれぞれが機能するように統合して、さらに自然言語処理に有効に機能させることができた、という本研究成果は、様々な情報を用いた情報処理の成功例として学術的意義がある。さらに、本研究自体は対象を薬物関係として限定して評価したものの、提案手法自体は薬物関係以外にも利用可能な汎用なものであり、その発展性、応用可能性も高い手法となっている。
また、薬物関係を高精度に文書から抽出する技術は薬物データベースの整備・拡充に貢献できる。薬物データベース上での表現学習により、未知の薬物間の関係の可能性を提示することも可能となり、創薬候補の提示にも利用できる。

研究成果の概要（英文）：Based on a drug database, we created a heterogeneous graph that includes drug-specific information such as drug names, descriptions, categories, and relationships, as well as various types of information, such as proteins related to drugs. We realized representation learning that can effectively utilize all information, including drug descriptions and chemical formula information.
We also studied methods to use information from the drug database for extracting drug relationships from a document. We realized a single model for extracting information from a document, using as input the representations learned from the knowledge graph and mapped to the document to be input. The numerical target was initially set to an F-score of 80% or higher for extraction performance, but in the end, an F-score of 85.40% was achieved, which greatly exceeded the initial target.

研究分野：知能情報学

キーワード：薬物間相互作用　DrugBank　関係抽出　深層学習　知識グラフ　表現学習　BERT　グラフニューラルネットワーク

様　式　Ｃ－１９、Ｆ－１９－１（共通）

１．研究開始当初の背景

副作用などの薬物に関する関係情報は薬物の使用における有害事象リスクの逓減や新規薬物の研究開発において重要である．この情報を整理・共有し，広く利用するために薬物データベースが日々整備され続けている．しかし，情報源である文献データベース MEDLINE は 1 [件/分] のペースで増加し続けており，その整備が間に合っていない．このため，薬物関係情報の自動抽出が自然言語処理・ライフサイエンスの分野で注目されてきた．このような中で，研究代表者は言語外の情報に着目し，薬物の化学式の情報を利用することで性能向上を達成した．しかし，データベース内の一部の表層的な情報の利用に限られており，従来手法ではデータベース中に含まれる様々な薬物情報を十分に活用できているとは言い難い状況であった．

２．研究の目的

本研究の目的は，薬物データベースを中心とした薬物情報を巨大な異種混合グラフとみなした上で，巨大な異種混合グラフを包括的に表現するために，グラフ内の異種の要素・関係を包括的に表現する深層表現学習モデルを確立し，F 値 80%以上の抽出性能を実現することである．

３．研究の方法

研究目的の達成に向けて，薬物データベースをもとに，薬物の名前・説明文・カテゴリ・関係などの薬物固有の情報に加え，薬物と関連するタンパク質などの様々な種類の情報を含む異種混合グラフを作成し、薬物の説明文や化学式の情報も含めて、統合的に表現する表現学習を実現した．さらに，薬物データベースの情報を文書からの薬物関係抽出に利用する手法について研究を進め，知識グラフから学習した表現を入力する文書と対応付けたものを入力として、文書からの情報抽出を行う単一の統一されたモデルを実現した．研究は主に以下の 6 つの項目に分けて実施した．

(1) テキストからの情報抽出において，科学分野に特化した大規模事前学習モデル SciBERT を用いた薬物関係抽出を評価・改善した．
(2) 薬物データベースに個別に説明文，化学式の情報を加えたモデルに加え，それらをアンサンブル学習により同時に利用する手法について評価・改善を行った．
(3) 薬物データベースを元に，薬物の名前・カテゴリ・関係などの薬物固有の情報を表現した異種混合グラフについて，表現学習の利用を検討し，グラフ上の隠れたノードを予測するリンク予測タスクを用いて，評価を行った．
(4) (3)で作成した異種混合グラフにさらに説明文や化学式の情報を加えた異種混合グラフについて表現学習の利用を検討し，リンク予測タスクを用いて，評価を行った．
(5) 異種混合グラフから得られた表現を入力する文書と対応付けて入力することで，文書からの薬物関係抽出を行うモデルを提案，実現し，評価を行った．
(6) 文書から関係を抽出する際に必要な文書中で薬物を表す固有表現の抽出，文献情報の表現学習，データベースのエントリと固有表現の対応付け・リンキング，薬物に近いタンパク質や遺伝子など他のエンティティとの関係への利用など，研究を実際の文書に利用する際に必要な手法について調査・研究を行った．固有表現抽出の技術を評価するために共通タスク n2c2 2022 track 1 に参加した．

４．研究成果

(1) 科学分野に特化した大規模事前学習モデル SciBERT を利用した薬物関係抽出について，CNN を利用した手法を評価し，従来の大規模事前学習モデルを用いない手法に比べて大幅に高い性能を達成したモデルについて，さらに CNN を導入することで，性能の向上が可能であることを示した．

| | F 値 (%) |
|---|---|
| Liu et al. 2016 | 67.01 |
| SciBERT+Linear | 81.09 |
| SciBERT+CNN | 81.72 |

表 1 薬物関係抽出の事前学習モデルの影響の評価と CNN を用いた改善

(2) 薬物データベースにおける説明文，化学式の個別の情報に加え，両者をアンサンブル学習によって同時に利用する手法について評価・改善を行った．結果として，(1)で得られた性能を個別の情報で向上できるとともに，両情報を加えることで更なる性能向上を達成した．

| | F 値 (%) |
|---|---|
| SciBERT+CNN | 81.72 |
| SciBERT+CNN+説明文 | 82.91 |
| SciBERT+CNN+化学式 | 82.84 |
| SciBERT+CNN+説明文+化学式 | 84.08 |

表 1 薬物関係抽出の事前学習モデルの影響の評価と CNN を用いた改善

(3) 薬物データベースを元にした異種混合グラフについては，名前・カテゴリなどの表層が利用できる場合について，その表層を入れる手法について検討・検証し，表層の情報を入れることで表 3 のように，異種混合グラフ上で隠された情報を予測するリンク予測タスクにおいて，平均逆順位が向上し，その有効性を示した．

| 表現学習手法 | 名前・カテゴリの表層なし | 表層あり |
|---|---|---|
| TransE | 0.3319 | 0.2883 |
| DistMult | 0.3380 | 0.3011 |
| ComplEx | 0.3242 | 0.3771 |
| SimplE | 0.2973 | 0.3549 |

表 3 異種混合グラフ上でのリンク予測タスクにおける表層の影響の評価．平均逆順位（MRR）での評価．

(4) (3)の異種混合グラフのノードに説明文と化学式の情報を追加する手法を開発し，その追加によるリンク予測タスクにおける影響の評価を行った．結果として，説明文・化学式ともに有効であり，両者を用いたものが最も良い性能を示した．

| 表現学習手法 | ノードのみ | ＋説明文 | ＋化学式 | ＋説明文・化学式 |
|---|---|---|---|---|
| TransE | 0.3114 | 0.2894 | 0.3003 | 0.2877 |
| DistMult | 0.6732 | 0.7702 | 0.7677 | 0.7933 |
| ComplEx | 0.6627 | 0.7874 | 0.7313 | 0.7923 |
| SimplE | 0.6228 | 0.7175 | 0.7156 | 0.7235 |

表 4 異種混合グラフ上でのリンク予測タスクにおける説明文・化学式の影響の評価．平均逆順位（MRR）での評価．

(5) (4)で得られた情報を薬物関係抽出に利用する手法を開発し，バイオに特化した事前学習モデルである PubMedBERT を用いた評価を行った．PubMedBERT 単体では(2)のモデルの性能に達することはできないのに対し，異種混合グラフ上での表現学習により得られた表現を入力文に対応付けながら利用することで，(1)，(2)の結果を上回る 85.40%という F 値を達成した．

| | F 値 (%) |
|---|---|
| Liu et al. 2016 | 67.01 |
| SciBERT+CNN | 81.72 |
| SciBERT+CNN+説明文+化学式 | 84.08 |
| PubMedBERT | 83.70 |
| PubMedBERT+異種混合グラフの表現 | 85.40 |

表 5 異種混合グラフの表現を用いた薬物関係抽出の評価

(6) 文書中に現れる用語を発見する固有表現抽出については，n2c2 2022 track 1 において，32 チーム中 1 位を達成した．

| チーム | F 値 (%) |
|---|---|
| 本チーム | 97.16 |
| University of Florida | 96.59 |
| IIT Dhanbad など | 95.88 |

表 6 n2c2 2022 track 1 における上位 3 チームの固有表現抽出における F 値

５．主な発表論文等

〔雑誌論文〕　計9件（うち査読付論文　9件／うち国際共著　0件／うちオープンアクセス　5件）

| | |
|---|---|
| １．著者名<br>Asada Masaki、Miwa Makoto、Sasaki Yutaka | ４．巻<br>39 |
| ２．論文標題<br>Integrating heterogeneous knowledge graphs into drug?drug interaction extraction from the literature | ５．発行年<br>2022年 |
| ３．雑誌名<br>Bioinformatics | ６．最初と最後の頁<br>- |
| 掲載論文のDOI（デジタルオブジェクト識別子）<br>10.1093/bioinformatics/btac754 | 査読の有無<br>有 |
| オープンアクセス<br>オープンアクセスとしている（また、その予定である） | 国際共著<br>－ |

| | |
|---|---|
| １．著者名<br>Iinuma Naoki、Miwa Makoto、Sasaki Yutaka | ４．巻<br>- |
| ２．論文標題<br>Improving Supervised Drug-Protein Relation Extraction with Distantly Supervised Models | ５．発行年<br>2022年 |
| ３．雑誌名<br>Proceedings of the 21st Workshop on Biomedical Language Processing | ６．最初と最後の頁<br>161-170 |
| 掲載論文のDOI（デジタルオブジェクト識別子）<br>10.18653/v1/2022.bionlp-1.16 | 査読の有無<br>有 |
| オープンアクセス<br>オープンアクセスではない、又はオープンアクセスが困難 | 国際共著<br>－ |

| | |
|---|---|
| １．著者名<br>Asada Masaki、Gunasekaran Nallappan、Miwa Makoto、Sasaki Yutaka | ４．巻<br>6 |
| ２．論文標題<br>Representing a Heterogeneous Pharmaceutical Knowledge-Graph with Textual Information | ５．発行年<br>2021年 |
| ３．雑誌名<br>Frontiers in Research Metrics and Analytics | ６．最初と最後の頁<br>－ |
| 掲載論文のDOI（デジタルオブジェクト識別子）<br>10.3389/frma.2021.670206 | 査読の有無<br>有 |
| オープンアクセス<br>オープンアクセスとしている（また、その予定である） | 国際共著<br>－ |

| | |
|---|---|
| １．著者名<br>Asada Masaki、Miwa Makoto、Sasaki Yutaka | ４．巻<br>－ |
| ２．論文標題<br>Using Drug Descriptions and Molecular Structures for Drug-Drug Interaction Extraction from Literature | ５．発行年<br>2020年 |
| ３．雑誌名<br>Bioinformatics | ６．最初と最後の頁<br>－ |
| 掲載論文のDOI（デジタルオブジェクト識別子）<br>10.1093/bioinformatics/btaa907 | 査読の有無<br>有 |
| オープンアクセス<br>オープンアクセスとしている（また、その予定である） | 国際共著<br>－ |

| | |
|---|---|
| １．著者名<br>Tsujimura Tomoki、Miwa Makoto、Sasaki Yutaka | ４．巻<br>143 |
| ２．論文標題<br>Large-scale neural biomedical entity linking with layer overwriting | ５．発行年<br>2023年 |
| ３．雑誌名<br>Journal of Biomedical Informatics | ６．最初と最後の頁<br>104433～104433 |
| 掲載論文のDOI（デジタルオブジェクト識別子）<br>10.1016/j.jbi.2023.104433 | 査読の有無<br>有 |
| オープンアクセス<br>オープンアクセスとしている（また、その予定である） | 国際共著<br>－ |

| | |
|---|---|
| １．著者名<br>Tsujimura Tomoki、Yamada Koshi、Ida Ryuki、Miwa Makoto、Sasaki Yutaka | ４．巻<br>144 |
| ２．論文標題<br>Contextualized medication event extraction with striding NER and multi-turn QA | ５．発行年<br>2023年 |
| ３．雑誌名<br>Journal of Biomedical Informatics | ６．最初と最後の頁<br>104416～104416 |
| 掲載論文のDOI（デジタルオブジェクト識別子）<br>10.1016/j.jbi.2023.104416 | 査読の有無<br>有 |
| オープンアクセス<br>オープンアクセスとしている（また、その予定である） | 国際共著<br>－ |

| | |
|---|---|
| １．著者名<br>Matsubara Takuma、Miwa Makoto、Sasaki Yutaka | ４．巻<br>- |
| ２．論文標題<br>Distantly Supervised Document-Level Biomedical Relation Extraction with Neighborhood Knowledge Graphs | ５．発行年<br>2023年 |
| ３．雑誌名<br>Proceedings of the 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks | ６．最初と最後の頁<br>363-368 |
| 掲載論文のDOI（デジタルオブジェクト識別子）<br>10.18653/v1/2023.bionlp-1.33 | 査読の有無<br>有 |
| オープンアクセス<br>オープンアクセスではない、又はオープンアクセスが困難 | 国際共著<br>－ |

| | |
|---|---|
| １．著者名<br>Yamada Koshi、Miwa Makoto、Sasaki Yutaka | ４．巻<br>- |
| ２．論文標題<br>Biomedical Relation Extraction with Entity Type Markers and Relation-specific Question Answering | ５．発行年<br>2023年 |
| ３．雑誌名<br>Proceedings of the 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks | ６．最初と最後の頁<br>377-384 |
| 掲載論文のDOI（デジタルオブジェクト識別子）<br>10.18653/v1/2023.bionlp-1.35 | 査読の有無<br>有 |
| オープンアクセス<br>オープンアクセスではない、又はオープンアクセスが困難 | 国際共著<br>－ |

| | |
|---|---|
| １．著者名<br>Ida Ryuki、Miwa Makoto、Sasaki Yutaka | ４．巻<br>- |
| ２．論文標題<br>Biomedical Document Classification with Literature Graph Representations of Bibliographies and Entities | ５．発行年<br>2023年 |
| ３．雑誌名<br>Proceedings of the 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks | ６．最初と最後の頁<br>385-395 |
| 掲載論文のDOI（デジタルオブジェクト識別子）<br>10.18653/v1/2023.bionlp-1.36 | 査読の有無<br>有 |
| オープンアクセス<br>オープンアクセスではない、又はオープンアクセスが困難 | 国際共著<br>－ |

〔学会発表〕　計20件（うち招待講演　0件／うち国際学会　6件）

| |
|---|
| １．発表者名<br>Koshi Yamada, Ryuki Ida, Tomoki Tsujimura, Kohei Makino, Makoto Miwa and Yutaka Sasaki |
| ２．発表標題<br>Span-based and Question Answering-based Medication Event Extraction |
| ３．学会等名<br>2022 n2c2/OHNLP Shared Task and Workshop（国際学会） |
| ４．発表年<br>2022年 |

| |
|---|
| １．発表者名<br>Iinuma Naoki、Miwa Makoto、Sasaki Yutaka |
| ２．発表標題<br>Improving Supervised Drug-Protein Relation Extraction with Distantly Supervised Models |
| ３．学会等名<br>Proceedings of the 21st Workshop on Biomedical Language Processing（国際学会） |
| ４．発表年<br>2022年 |

| |
|---|
| １．発表者名<br>片桐脩那，井田龍希，三輪誠，佐々木裕 |
| ２．発表標題<br>テキスト情報の表現を利用した文献グラフの表現学習 |
| ３．学会等名<br>言語処理学会第29回年次大会 |
| ４．発表年<br>2023年 |

１．発表者名
松原拓磨，三輪誠，佐々木裕

２．発表標題
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

３．学会等名
言語処理学会第29回年次大会

４．発表年
2023年

１．発表者名
大井拓，三輪誠，佐々木裕

２．発表標題
ラベル内容のエンコードとラベル間の制約に基づく補助コーパスを用いた固有表現抽出

３．学会等名
言語処理学会第29回年次大会

４．発表年
2023年

１．発表者名
井田龍希，三輪誠，佐々木裕

２．発表標題
文書外の書誌情報と用語情報を組み込んだ文書分類

３．学会等名
言語処理学会第29回年次大会

４．発表年
2023年

１．発表者名
山田晃士，三輪誠，佐々木裕

２．発表標題
複数の質問形式を利用した分類型の質問応答による薬物タンパク質間関係抽出

３．学会等名
言語処理学会第29回年次大会

４．発表年
2023年

１．発表者名
Naoki Iinuma, Masaki Asada, Makoto Miwa, and Yutaka Sasaki

２．発表標題
TTI-COIN at BioCreative VII Track 1 Drug-protein interaction extraction with external database information

３．学会等名
BioCreative VII Workshop（国際学会）

４．発表年
2021年

１．発表者名
山田晃士，三輪誠，佐々木裕

２．発表標題
項の表現に着目した質問応答による関係分類

３．学会等名
言語処理学会第28回年次大会

４．発表年
2022年

１．発表者名
浅田真生，三輪誠，佐々木裕

２．発表標題
薬学知識グラフ上のヘテロな情報を利用した文献からの薬物相互作用抽出

３．学会等名
言語処理学会第28回年次大会

４．発表年
2022年

１．発表者名
飯沼直己，三輪誠，佐々木裕

２．発表標題
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

３．学会等名
言語処理学会第28回年次大会

４．発表年
2022年

１．発表者名
Takuma Matsubara, Makoto Miwa, Yutaka Sasaki

２．発表標題
Distantly Supervised Document-Level Biomedical Relation Extraction with Neighborhood Knowledge Graphs

３．学会等名
The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks（国際学会）

４．発表年
2023年

１．発表者名
Koshi Yamada, Makoto Miwa, Yutaka Sasaki

２．発表標題
Relation Extraction with Entity Type Markers and Relation-specific Question Answering

３．学会等名
The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks（国際学会）

４．発表年
2023年

１．発表者名
Ryuki Ida, Makoto Miwa, Yutaka Sasaki

２．発表標題
Biomedical Document Classification with Literature Graph Representations of Bibliographies and Entities

３．学会等名
The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks（国際学会）

４．発表年
2023年

１．発表者名
鬼頭泰清，牧野晃平，三輪誠，佐々木裕

２．発表標題
固有表現抽出における大規模言語モデルのLoRAファインチューニングの学習設定の調査

３．学会等名
言語処理学会第30回年次大会(NLP2024)

４．発表年
2024年

１．発表者名
松原拓磨，辻村有輝，三輪誠，佐々木裕

２．発表標題
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

３．学会等名
言語処理学会第30回年次大会(NLP2024)

４．発表年
2024年

１．発表者名
大井拓，三輪誠，佐々木裕

２．発表標題
CVAEによる複数データセットからの固有表現抽出

３．学会等名
言語処理学会第30回年次大会(NLP2024)

４．発表年
2024年

１．発表者名
井田龍希，三輪誠，佐々木裕

２．発表標題
文献グラフにおける多項関係の埋め込み

３．学会等名
言語処理学会第30回年次大会(NLP2024)

４．発表年
2024年

１．発表者名
山田晃士，三輪誠，佐々木裕

２．発表標題
複数の形式・表現の質問を利用した多角的な関係抽出

３．学会等名
言語処理学会第30回年次大会(NLP2024)

４．発表年
2024年

１．発表者名
辻村有輝，三輪誠，佐々木裕

２．発表標題
IDレベル関係抽出における不要な文の自動選択

３．学会等名
言語処理学会第30回年次大会(NLP2024)

４．発表年
2024年

〔図書〕　計0件

〔産業財産権〕

〔その他〕
tticoin/PharmaHKG-Text
https://github.com/tticoin/PharmaHKG-Text
DESC_MOL-DDIE
https://github.com/tticoin/DESC_MOL-DDIE
HKG-DDIE
https://github.com/tticoin/HKG-DDIE
n2c2-2022 track1
https://github.com/tti-coin/n2c2-2022_track1
bio-linking
https://github.com/tti-coin/bio-linking

６．研究組織

| | 氏名<br>（ローマ字氏名）<br>（研究者番号） | 所属研究機関・部局・職<br>（機関番号） | 備考 |
|---|---|---|---|

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

| 共同研究相手国 | 相手方研究機関 |
|---|---|