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　２ 年　 ６ 月 １９ 日現在

機関番号：３３９２４
研究種目：若手研究(B)
研究期間：2017～2019
課題番号：１７Ｋ１２７４１

研究課題名（和文）データベース上での表現学習による薬物関係予測

研究課題名（英文）Drug-drug Interaction Extraction by Representation Learning on Databases

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）・准教授

研究者番号：００５２９６４６

交付決定額（研究期間全体）：（直接経費）　　3,200,000円

研究成果の概要（和文）：日々増え続ける膨大な医学薬学文献から重要な情報を見逃さずに見つけるために，文献に記載された薬物の研究開発・利用に重要な薬物に関する関係情報を自動的に抽出する手法が求められている．本研究ではすでに発見され整理されているデータベースの情報を有効に利用することで，従来，文書のみに着目していた抽出手法に対する精度の向上を図った．データベース中の化学式の情報やタグ付けされていない膨大な文献を利用する深層学習手法を開発し，従来精度70%程度であったシステムに対し，10%ポイントを超える精度の向上を達成した．

研究成果の学術的意義や社会的意義
従来の日本語や英語などの自然言語を対象とした情報抽出の研究では，言語情報のみを利用したものがほとんどであり，データベース情報などは「データベースにあるかどうか」などの特徴として補助的に使われる程度であった．本研究成果は，データベース上に含まれる化学式などの単純には言語と結びつかないような情報を，深層学習を用いて自然言語からの情報抽出に利用可能にし，さらにその情報を使うことで従来手法の精度を向上できたことに意義がある．

研究成果の概要（英文）：To find important information from increasing the medical and pharmaceutical literature, there is a need for a method that automatically extracts relationship information of drugs, which is essential to R&D and use of drugs. In this research, we improved the conventional extraction method, which used to focus only on textual information, by effectively using the database information that is already discovered and organized. We have developed deep learning methods that utilize information on chemical structures in the database and a large amount of unannotated texts, and improved the conventional method, which showed around 70% in extraction performance, by more than 10% points.

研究分野：自然言語処理

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

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

薬物の副作用等の関係は，薬の飲み合わせを始めとして薬の研究開発に重要な情報であるが，多くは医学薬学文献で発表され，文献は日々出版されており，その整理が間に合っておらず，また，その整理に莫大なコストがかかる．このため，関係の自動抽出が自然言語処理・テキストマイニングにおいて注目されている．従来研究では人手で関係を注釈付けした文書を利用した教師あり学習手法が高い精度を上げてきたが，タグ付けのコストが高く，また，一定のタグ付データに対する精度は頭打ちになってきていた．このため，外部のデータベースの情報を使う研究も行われていたが，文字列マッチなど限定的な利用に限られていた．

２．研究の目的

本研究の目的は，医学薬学文献に記述された薬物間の関係（相互作用）について，薬物データベースにすでに登録されている薬物に関する情報を援用することにより，追加のタグ付けコストなしに，自動抽出の精度を向上させることである．

３．研究の方法

研究目的の達成に向けて，データベース中に含まれる薬物間の関係をもとに薬物の表現を試みたが，データベースに関係情報が付与されていない薬物の存在が明らかになったため，薬物を説明するデータベースの他の情報を利用することにした．さらに，データベース外のタグ付けされていないテキストの情報等を利用することで精度の向上を目指した．

(1) テキストからの情報抽出について，畳み込みニューラルネットワークと対象エンティティを考慮した注意機構を開発した．
(2) 薬物データベース内ですべての薬物について共通にタグ付けされている化学式に焦点を当て，化学式を表現するグラフニューラルネットワークについて検討・比較を行い，薬物相互作用抽出において評価を行った．
(3) 薬物データベース内の説明文についても薬物を説明する情報として利用できるように表現の検討を行い，利用した．
(4) 薬物データベース以外に利用可能な注釈付されていない大量の文献から得られる表現の改良のために，自然言語処理の分野で近年急速に進展した巨大な事前学習モデルの利用を検討し，評価を行った．
(5) 文書中に現れる用語とデータベースのエントリの対応付けを解決するためのエンティティリンキングについて検討を行い，評価のために共通タスク n2c2 2019 track 3 に参加した．
(6) 薬物とその他のエンティティの関係についても，薬物とタンパク質の関係に着目して，データベース情報を作った教師データを利用する学習について研究を行った．

４．研究成果

薬物の相互作用抽出については，テキストからの自動抽出における注意機構の考案をまず行い，化学式の利用，説明文の利用，大量のタグ付けされていない文献情報の利用について評価を行った．化学式についてはその表現モデルの評価を行い，また，その情報の薬物関係抽出への利用についても評価を行い，国際会議・国内学会で発表した．説明文の利用についても発表を行った．大量のタグ付けされていない文献情報の利用・説明文の利用については，薬物相互作用抽出の精度を評価した．また，エンティティリンキングについては共通タスクに参加し，薬物・タンパク質の関係抽出については国内会議において発表を行った．

(1) テキストからの薬物相互作用情報の抽出において薬物を利用した注意機構を用いることで，従来手法に比べて，有意な精度向上を実現した（表 1）．

手法	F 値 (%)
Liu et al. 2016	67.01
注意機構の利用	69.12

表 1 テキストからの薬物相互情報抽出における注意機構による精度向上

(2) 化学式の利用については NFP (Neural Finger Print)と GGNN (Gated Graph Neural Networks)を利用した化学式の表現を行った．データベースに含まれる相互作用が記述されているペアとそうでないペアの分類を行ったところ，表 2 に示すように高い精度を実現することができた．この結果は記述されていないペアはランダムに作ったペアであり問題が簡単になっている可能性があるが，化学式の利用可能性を示すことができた．

手法	正解率 (%)
NFP	94.19
GGNN	98.00

表 2 化学式を用いたデータベース上での相互作用分類

また，実際に，(1)の抽出手法を調整したものに，この化学式から得られた表現を追加したとこ

ろ有意な精度の向上を実現でき，化学式の言語処理への利用可能性を世界で初めて示した.

手法	F 値（%）
Liu et al. 2016	67.01
Zheng et al. 2017	71.5
Lim et al. 2018	71.7
テキスト情報のみ	70.16
＋NFP を用いた化学式の利用	72.21
＋GGNN を用いた化学式の利用	72.25

表 3 データベース内の化学式情報を用いた薬物相互作用抽出の改善

(3) 薬物の説明文の利用については，畳み込みニューラルネットワークを用いた説明文の表現を利用することで表 4 のように化学式ほどではないものの精度の向上を達成できた.

手法	F 値（%）
テキスト情報のみ	70.16
＋畳み込みニューラルネットワークによる説明文の利用	71.19

表 4 データベース内の薬物の説明文情報を用いた薬物相互作用抽出の改善

(4) 注釈付されていない大量の文献から得られる表現の改良については BERT モデル・SciBERT モデルを利用することで，大幅な精度向上を実現できることを示した．データベースの情報を用いることでさらに精度の向上が可能であることも示し，文献からの情報とデータベースの情報は相補的な情報であることがわかった.

手法	F 値（%）
CNN	70.16
BERT	79.90
SciBERT	81.19
SciBERT + 説明文	82.04
SciBERT + 化学式	82.32
SciBERT + 説明文 + 化学式	82.45

表 5 大量の文献情報と薬物データベースを用いた薬物相互抽出の改善

(5) 文書中に現れる用語とデータベースのエントリの対応付を解決するエンティティリンキングについては，n2c2 2019 track 3 において，85.2%の正解率を達成し，共通タスクにおいて 2 位と有意な差をつけながら 33 チーム中 1 位を達成した.

チーム	正解率（%）
本チーム	85.26
Kaiser Permanente	81.94
University of Arizona	81.66

表 6 n2c2 2019 track 3 における上位 3 チームの正解率

(6) 薬物とタンパク質の関係抽出においては，既存手法である PCNN にデータベースから作った教師データを追加することで精度の向上が可能であることがわかった.

手法	F 値（%）
PCNN	74.76
PCNN+データベースから作った教師データ	75.89

表 7 薬物タンパク質間相互作用抽出におけるデータベースから作ったデータの利用

５．主な発表論文等

〔雑誌論文〕　計0件

〔学会発表〕　計10件（うち招待講演　0件／うち国際学会　4件）

１．発表者名 浅田真生，三輪誠，佐々木裕
２．発表標題 薬物データベースを統合的に利用する薬物相互作用抽出
３．学会等名 NLP若手の会 第14回シンポジウム
４．発表年 2019年

１．発表者名 飯沼直己，三輪誠，佐々木裕
２．発表標題 遠距離教師データを援用した教師あり薬物タンパク質間相互作用抽出
３．学会等名 言語処理学会第26回年次大会
４．発表年 2020年

１．発表者名 茂里憲之，辻村有輝，三輪誠，佐々木裕
２．発表標題 二段階学習と概念クラスを用いた医療固有表現の正規化
３．学会等名 言語処理学会第26回年次大会
４．発表年 2020年

１．発表者名 Masaki Asada, Makoto Miwa and Yutaka Sasaki
２．発表標題 Using External DB Knowledge in Neural DDI Extraction
３．学会等名 Third International Workshop on Symbolic-Neural Learning（国際学会）
４．発表年 2019年

１．発表者名
Tomoki Tsujimura, Noriyuki Mori, Masaki Asada, Makoto Miwa and Yutaka Sasaki

２．発表標題
TTI-COIN at n2c2 2019 Track 3: Neural Medical Concept Normalization with Two-Step Training

３．学会等名
2019 n2c2/OHNLP Workshop（国際学会）

４．発表年
2019年

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

２．発表標題
Enhancing Drug-Drug Interaction Extraction from Texts by Molecular Structure Information

３．学会等名
the 56th Annual Meeting of the Association for Computational Linguistics (ACL 2018)（国際学会）

４．発表年
2018年

１．発表者名
浅田真生，三輪誠，佐々木裕

２．発表標題
データベースの説明文を利用した薬物相互作用抽出

３．学会等名
言語処理学会第25回年次大会

４．発表年
2019年

１．発表者名
矢島雄樹，三輪誠，佐々木裕

２．発表標題
遠距離教師データを援用した教師有り薬物タンパク質間相互作用抽出

３．学会等名
言語処理学会第25回年次大会

４．発表年
2019年

１．発表者名 Masaki Asada, Makoto Miwa and Yutaka Sasaki
２．発表標題 Extracting Drug-Drug Interactions with Attention CNNs
３．学会等名 the 2017 Workshop on Biomedical Natural Language Processing (BioNLP 2017)（国際学会）
４．発表年 2018年

１．発表者名 浅田真生，三輪誠，佐々木裕
２．発表標題 分子構造を用いた文書からの薬物相互作用抽出
３．学会等名 言語処理学会
４．発表年 2018年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

－

６．研究組織

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