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

矢島 雄樹　　三輪 誠　　佐々木 裕
豊田工業大学
{sd15093, makoto-miwa, yutaka.sasaki}@toyota-ti.ac.jp

1 はじめに

医療従事者は日々進歩している医療現場において常に自らの専門分野の最新情報を取り入れ，既存の臨床結果に基づいた医療を実践している．この『根拠に基づく医療』を展開していく上で，薬物とタンパク質間の相互作用に関する情報は重要であり，これらの情報を大量の医学文書から収集する技術も重要な役割を担っている．このために，テキストからの関係抽出が注目されており，その中でも近年様々なタスクにおいて高精度を達成している深層学習を利用する技術が注目されている．

深層学習を利用した相互作用抽出では，大量の訓練データ上で学習されたモデルを用いて，文書に現れる薬物とタンパク質の関係を予測する．ニューラル関係抽出モデルでは，大量のラベル付きデータから関係のモデルを学習をすることで予測精度を向上できるが，その人手でのラベル付きデータの作成には莫大なコストがかかるという問題を抱えている．

このような人手でのラベル付きデータ作成の問題を回避し，少ないコストで大量の訓練データを用意できる遠距離教師有り学習が Mintz ら [1] によって提案されている．大量のラベルなしデータに対して，データベース情報を元に機械的にラベル付けする遠距離教師有り学習により，低コストで大量の教師データの作成が可能となる．しかし，遠距離教師有り学習はデータ作成時に誤ったラベルをつけてしまう問題が残っており，これが学習の妨げとなっている．

本研究では，人手でラベル付けされた教師データと機械的にラベル付けされた遠距離教師データの両方を訓練データとしてモデルの学習に利用することで，教師データを増やし，相互作用抽出精度の向上を目指す．また，この際，データベース中の関係から参照されている論文のみから遠距離教師データを作成することで，精度の高い遠距離教師データの作成を目指す．

2 関連研究

2.1 ニューラル関係抽出モデル

関係抽出は自然言語処理における重要な問題の一つであり，文中に現れる対象エンティティ間の関係を機械的に予測することを目的としている．対象とするエンティティのペアを (h, t) と表し，このペア間に r の関係がある時，(h, r, t) と表現する．

この関係抽出において，深層学習を利用したニューラル関係抽出モデルが近年高精度を達成している．ニューラル関係抽出では，はじめに関係が表現されている文 x の特徴をエンコードし，ここで得られた特徴からペア (h, t) 間の関係 r を予測する．Zeng ら [2] は，文 x のエンコードに畳み込みニューラルネットワーク (CNN) を利用したモデルを提案した．このモデルは文の特徴としてエンティティからの距離を単語の素性として加えることで，文中に現れる各単語自体が持つ特徴だけでなく，位置関係の特徴も考慮している．ニューラル関係抽出では，その学習に大量のラベル付きデータが必要であり，そのデータを人手で用意するのに多大なコストがかかるという問題がある．

2.2 遠距離教師有り学習

遠距離教師有り学習は Mintz ら [1] によって提案された学習方法である．遠距離教師有り学習では，大量な教師データに向けて人手でラベル付けすることを避けるため，大量のラベルなし文書を用意し，この文書に機械的にラベル付けを行う．関係を予測したいペアが共起する文をデータベースから全て取り出し，ラベル付けすることで，コストをかけずに大量の教師データを用意できる．しかし，この手法では，関係を表していない文に対してもラベル付けをしてしまうことがあ

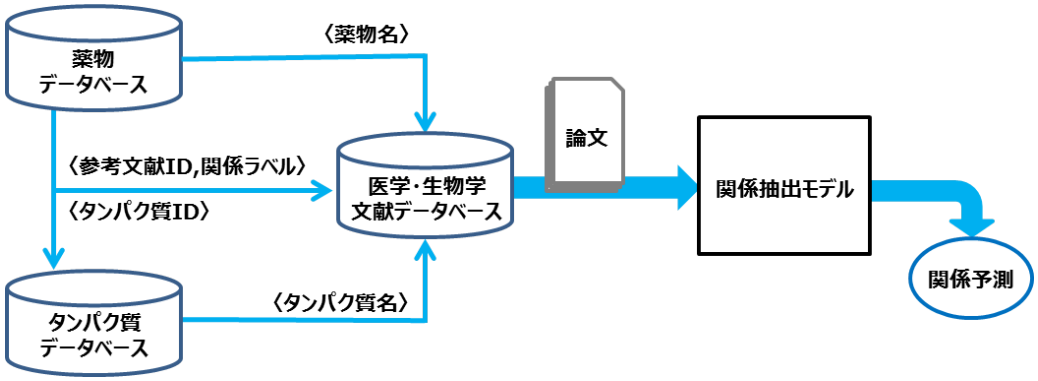

図 1: 提案手法の流れ

るため，誤ったデータを含む教師データを前提とした関係抽出モデルが必要であり，様々な学習手法 [3, 4, 5] が提案されている．

3 提案手法

本研究では，人手でラベル付けされた教師データとデータベースから用意した誤りを含む遠距離教師データの 2 種類のデータを同時に利用して，薬物相互作用抽出のためのモデル学習を行う手法を提案する．薬物相互作用抽出を行うまでの流れは図 1 に示した．本手法は遠距離教師データの作成 (3.1 節) と畳み込みニューラルネットによる関係抽出 (3.2 節) からなる．

遠距離教師データ作成では，Mints らが提案した遠距離教師有り学習のデータ作成手法をもとに遠距離教師データを作成する．ただし，本研究では，予測したペアが共起する文全てを利用とする既存手法と異なり，データベースにある参照情報をもとに，関係が記述されていることがわかっている文書のみを利用する．この文書中の文上に現れるペアには，データベースで付けられた関係がある可能性が高く，ラベル誤りが抑制された遠距離教師データの作成が期待される．

関係抽出では Zeng ら [2] が提案した畳み込みニューラルネットワークを利用した関係抽出モデルに基づいた関係予測を行うモデルを用いる．そして，人手でラベル付けされた教師データに，作成された遠距離教師データを追加したデータでこのモデルを学習させることで予測精度の向上を図る．

3.1 遠距離教師データの作成

遠距離教師データの作成では，3 つのデータベースとそこに蓄積されたデータのみを用いて，関係抽出モデル学習時に用いるデータを構築する．遠距離教師データの作成過程を図 1 の左側に示す．

データの作成は次のように行う．まず，それぞれの薬物について，薬物の ID と作用するタンパク質の ID とその関係名，その関係が記述された文献の ID の 4 つを 1 つの関係情報として，薬物データベースから関係情報を抽出する．次に，それぞれの関係情報について，関係情報に含まれる文献の ID を使って医学・生物学文献データベース PubMed から論文のタイトル・要旨を取得する．さらに，関係情報に含まれる薬物・タンパク質について薬物データベース・タンパク質データベースから薬物名・タンパク質名を取得する．この薬物名・タンパク質名が同時にタイトル・要旨中の文に出現すれば，その文を取得する．最後に，このような文が取得できたら，この文のエンティティペアがデータベースの関係名を表すペアであるとみなし，薬物，タンパク質，関係名，文を 1 つの遠距離教師データとして取得する．

3.2 関係抽出

人手でラベル付けされた教師データと，3.1 節で作成された遠距離教師データを用いて関係抽出モデルの学習を行う．関係抽出モデルは入力部，畳み込み部，関係分類部で構成されている．

3.2.1 入力部

入力された文 $\boldsymbol{x}$ は，各単語を表現するベクトル $\boldsymbol{w}$ のリスト

$$\boldsymbol{x} = [\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_m] \tag{1}$$

で表現される．この単語ベクトルは 2 つのベクトル，つまり，単語自体を表現するベクトルと単語の位置を表現

するベクトル, から構成される．単語自体を表現するベクトルは word2vec を用いて獲得する d_w 次元のベクトルであり, 位置を表現するベクトルは, 関係を予測したい薬物, もしくはタンパク質からの距離で決まる d_p 次元のベクトルである．これらの表現を並列に連結したベクトルを単語ベクトル $\boldsymbol{w_i}$ ($\boldsymbol{w_i} \in R^d, d = d_w + 2d_p$) とし, このベクトルで表現される文 $\boldsymbol{x}$ を畳み込みニューラルネットワークの入力とする．

3.2.2 畳み込み部

入力層から渡された入力 $\boldsymbol{x}$ を畳み込み層に通し，文 $\boldsymbol{x}$ で表現されている関係 r の特徴を得る．この変換を以下の式で表現する．

$$L = \mathrm{CNN}(\boldsymbol{x}) \tag{2}$$

畳み込み関数 CNN(·) の出力 L の次元数は d_{cnn} である．CNN(·) は周辺 3 単語の単語ベクトルから特徴を取り出す畳み込み層と max プーリング層から成る．

3.2.3 関係分類部

入力文から畳み込み部を通して得られた特徴 L から，文が表現しているペア (h, t) の関係 r を関係分類部を通して予測する．関係分類部は，2 層の全結合層を通して入力文 x がどの関係を表現しているかの確率を以下の式で計算する．

$$H = \mathrm{relu}(W_1 L + \boldsymbol{b}_1) \tag{3}$$

$$p(r|x;\boldsymbol{\phi}) = \mathrm{softmax}(W_2 H + \boldsymbol{b}_2) \tag{4}$$

ここで, $W_1 \in R^{d_h \times d_{cnn}}$, $\boldsymbol{b}_1 \in R^{d_h}$, $W_2 \in R^{n_r \times d_h}$, $\boldsymbol{b}_2 \in R^{n_r}$ である．また，$\boldsymbol{\phi}$ はパラメタ，n_r は予測する関係の数である．学習時はこの確率と正解ラベルとの交差エントロピー損失を最小にするように学習を行う．

4 実験

4.1 実験設定

提案の有用性を評価するために，人手でラベル付けされた教師データのみの場合と教師データと遠距離教師データを混ぜた 2 種類のデータを訓練データとして，それぞれモデルの学習を行った．それぞれのデータで学習したモデルで関係を予測し，予測結果全てを見たときの適合率，再現率，それらの調和平均をとった F 値を用いて精度の比較をした．ここで，評価データ，テストデータは BCVI のデータを利用した．また，既存研究 [6] に基づいて Positive に分類される関係 5 つはそれぞれ予測し，Negative に分類される関係はまとめて Negative として予測を行った．

4.1.1 タグ付きデータ

訓練，及び評価データには，共通タスクである BioCreative VI Track 5 の CHEMPROT タスク [7] のデータを用いた．CHEMPROT タスクでは，医学・生物学文献データベースである PubMed から集められた論文の要旨を対象に，この文書から薬物とタンパク質の関係を予測する取り組みが行われた．BioCreative VI の ChemProt データ (BCVI) の統計を表 1 に示した．

BCVI は薬物とタンパク質間の関係を 10 種類に分類している．表 2 に分類を示した．

4.1.2 遠距離教師データ

3.1 節に従って, 遠距離教師データ (Distant Supervision from DrugBank; DSDB) を作成した．薬物データベースには薬剤についての詳細情報について収集・統合したデータベースである DrugBank を，タンパク質データベースにはタンパク質の配列や機能に関するデータベースである UniProt を，文献データベースには医学・生物学文献データベースである PubMed をそれぞれ利用した．作成した遠距離教師データ (DSDB) の統計を表 1 に示した．また, DrugBank の関係を表 2 に従い BCVI の分類に置き換えてラベル付けを行った．

4.1.3 学習設定

単語表現ベクトルの次元数 d_w は 100，位置表現ベクトルの次元数 d_p は 5 とする．畳み込み層のフィルタサイズ k は 3 で，フィルタ数 d_{cnn} は 230 とし，畳み込み層，max プーリング層それぞれでゼロパディングを行った．全結合層で隠れ次元数 d_h は 500 とした．

学習はバッチサイズを 32，最適化アルゴリズムは Adam [8]，学習率を 0.001，ドロップアウト率を 0.5，

表 1: データの統計

	BCVI	DSDB
ドキュメント数	1,020	2,673
薬物数	13,017	15,194
タンパク質数	12,753	14,809

epoch 数を 200 回とした．また，単語表現ベクトルはランダムに初期化されたベクトルを利用した．

4.2 結果と考察

BCVI のみ，BCVI+DSDB を訓練データとして学習したモデルで多クラス分類を行った結果を表 3 に示す．訓練データの多い BCVI+DSDB で学習したモデルは BCVI のみで学習したモデルより高い精度を示した．この結果より，人手でラベル付けされた教師データに遠距離教師データを加えた訓練データによって，薬物相互作用抽出の精度を向上できることがわかった．

5 おわりに

本論文では，人手でラベル付けされた教師データに，データベースから作成した遠距離教師データを加えることで，関係抽出の精度を向上させる手法を提案した．データベースの参照情報を用いることで精度の高い教師データを作成できることがわかり，そのデータを既存の教師データに追加することで精度の向上を達成した．

今後は，ハイパーパラメタのチューニングを行い，モデル全体の精度の向上を図る．さらに，遠距離教師データの信頼度 [3] や遠距離教師データからのノイズの軽減 [5] を取り入れ，より精度の高い薬物タンパク質間相互作用抽出を目指す．

表 2: 関係分類

BCVI の分類	属性	DrugBank の分類
CPR:0	Negative	unknown
CPR:1	Negative	part of
CPR:2	Negative	regulator
CPR:3	Positive	activator
CPR:4	Positive	inhibitor
CPR:5	Positive	agonist
CPR:6	Positive	antagonist
CPR:7	Negative	modulator
CPR:8	Negative	cofactor
CPR:9	Positive	substrate
CPR:10	Negative	not

表 3: 予測結果

	適合率	再現率	F 値 (%)
BCVI	46.42	50.41	48.33
BCVI+DSDB	49.63	54.64	52.02

謝辞

本研究は JSPS 科研費 JP17K12741 の助成を受けたものである．

参考文献

[1] Mintz et al. Distant supervision for relation extraction without labeled data. In *Proceedings of ACL-IJCNLP 2009*, pp. 1003–1011, 2009.

[2] Zeng et al. Relation classification via convolutional deep neural network. In *Proceedings of COLING 2014*, pp. 2335–2344, 2014.

[3] Surdeanu et al. Multi-instance multi-label learning for relation extraction. In *Proceedings of EMNLP 2012*, pp. 455–465, 2012.

[4] Zeng et al. Distant supervision for relation extraction via piecewise convolutional neural networks. In *Proceedings of EMNLP2015*, pp. 1753–1762, 2015.

[5] Qin et al. Robust distant supervision relation extraction via deep reinforcement learning. In *ACL 2018*, pp. 2137–2147, 2018.

[6] Peng et al. Extracting chemical–protein relations with ensembles of svm and deep learning models. *Database*, Vol. 2018, No. 1, p. bay073, 2018.

[7] Krallinger 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.

[8] Kingma et al. Adam: A method for stochastic optimization. In *ICLR 2014*, 2014.