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

飯沼 直己　　三輪 誠　　佐々木 裕
豊田工業大学
{sd16005, makoto-miwa, yutaka.sasaki}@toyota-ti.ac.jp

1 はじめに

薬物とタンパク質間の相互作用に関する情報の多くは，論文などの文献において発表される．相互作用の情報を抽出するために薬理学者が逐一論文を読むのはコストがかかる．そのため，テキストからの相互作用抽出の自動化が注目されている．中でも，近年様々なタスクにおいて高い精度を達成している深層学習を用いた相互作用抽出への関心が高まっている．深層学習を利用した相互作用抽出では，ラベル付きデータから関係のモデルを学習することで高い予測精度を達成できるが，データ作成に莫大なコストがかかるという問題を抱えている．

そこで，低コストで大量の教師データの作成を可能にする遠距離教師あり学習がMintsら[5]によって提案されているが，この手法には学習時のノイズとなる誤ったラベルのデータを作成してしまう問題が残っている．そのため，少量の人手の教師データを利用して遠距離教師データのノイズを緩和する手法がBeltagyら[1]によって提案されている．

本研究では，異なる関係クラスを持った少量の人手の教師データと大量の遠距離教師データを対象に，クラスの対応付け（マッピング）を行いながら，マルチタスク学習を行う手法を提案する．教師となる情報を増やした学習により，共通の深層学習モデルの表現をより正確にし，薬物タンパク質間の関係抽出精度の向上を目指す．

2 関連研究

2.1 遠距離教師あり学習

遠距離教師あり学習はMintsら[5]によって提案された，データベースから機械的に文書に関連情報をラベル付けする手法である．これにより，大量の教師データを生成できるが，ラベル付けを誤った教師データを含んでしまう．この誤ったラベルは予測モデルの性能を低下させるノイズとなる．そのため，ノイズの影響を緩和する手法が提案されている．よく用いられる手法として，教師データをデータベース上のペアに対応するインスタンスのバッグで扱うマルチインスタンス学習がある． Zengら[7]はバッグ中の最も良い表現を持つデータにより学習する手法を提案した． Yeら[6]は2つの注意機構を導入し，バッグをまとめたグループにより学習する手法を提案した．他のノイズの影響の緩和手法として，Beltagyら[1]は，少量の人手の教師データによる注意機構を利用する手法を提案した．

2.2 マルチタスク学習

複数のタスクにおいてモデルパラメータを部分的に共有して学習を行うマルチタスク学習[2]という手法がある．これにより，各タスクのモデルを学習する際の情報を増やすことができ，予測精度を向上できる．

3 提案手法

本手法では，関係クラスが別々に定義されている人手の教師データと遠距離教師データの薬物タンパク質間の関係抽出を対象に，マルチタスク学習を行う．特に，人手の教師データと遠距離教師データの異なる関係クラスに対応するため，本手法ではいくつかの遠距離教師データの関係クラスのマッピングとモデルの性能への影響を検証する．本手法により，モデルがより多くの関係表現パターンを学習し，遠距離教師データの教師あり関係抽出への有効利用を目指す．

本手法は大きく遠距離教師データの作成と関係抽出の2つに分けられる．関係抽出についてはさらに，人手の教師データの関係抽出，および遠距離教師データの関係抽出，人手の教師データと遠距離教師データのマルチタスク学習の3つに分けられる．3.1節では，データベースからの遠距離教師データ生成について説明する．3.2節では，人手の教師データの関係抽出について説明する．3.3節では，遠距離教師データの関係抽出について説明する．3.4節では，人手の教師データと遠距離教師データのマルチタスク学習について説明する．

3.1 遠距離教師データの作成

遠距離教師データの作成の概略を図1に示す．遠距離

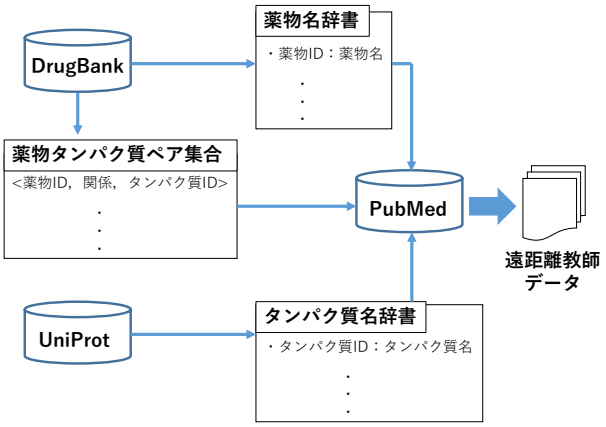

図1　遠距離教師データ作成の概略図

教師データの作成には，薬物データベースDrugBank・タンパク質データベースUniProt・医学文献データベースPubMedの3つのデータベースを利用する．以降，これらのデータベースを用いて遠距離教師データを作成する手順を説明する．

まず，DrugBankの情報を基に薬物タンパク質ペア集合と薬物名辞書を作成する．薬物タンパク質ペア集合とは，関係を持つ薬物とタンパク質のID，またそれらの関係を持つペアの集合である．また，薬物名辞書とは薬物IDと薬物名，その同義語が対応づいた辞書である．次に，UniProtの情報を基にタンパク質IDとタンパク質，その同義語が対応づいたタンパク質名辞書を作成する．最後に，薬物タンパク質ペア集合，薬物名辞書，タンパク質名辞書を利用して辞書マッチによりPubMedから遠距離教師データを作成する．

3.2　人手の教師データの関係抽出

人手のデータの関係抽出には，PCNNにより特徴量抽出を行い，全結合層により予測を行うモデルを用いる．本節では，以降，人手のデータの関係抽出に用いたモデルの説明を行う．

3.2.1　単語埋め込み層

単語埋め込み層によって，入力文の表現$\boldsymbol{x}$を獲得する．$\boldsymbol{x}$は各単語のベクトル表現$\boldsymbol{w}$のリスト

$$\boldsymbol{x} = [\boldsymbol{w}_0^T, \boldsymbol{w}_1^T, ..., \boldsymbol{w}_{N-1}^T] \tag{1}$$

で表される．入力文はパディング処理により長さNとする．各単語のベクトル表現$\boldsymbol{w}$は単語そのものを表現するベクトルとエンティティからの相対距離を表現するベクトルを結合して得られる．単語のベクトル表現はSkip-gramにより獲得したd_w次元のベクトルを用いる．エンティティからの相対距離を表現するベクトルは，d_p次元のベクトルである．従って，各単語のベクトル表現$\boldsymbol{w}$は$d = d_w + 2d_p$次元のベクトルである．この表現$\boldsymbol{x}$をPCNNの入力とする．

3.2.2　畳み込み層

単語埋め込み層から渡された入力xを元にPCNN[7]により，各文の特徴量ベクトル$\boldsymbol{h}$を生成する．

3.2.3　全結合層

全結合層では特徴量ベクトル$\boldsymbol{h}$と全結合層の重み行列$\boldsymbol{W}^{fc}$の内積を計算し，得られたベクトルに対しソフトマックス関数を施すことで各関係が表現されている確率$\boldsymbol{p}$を予測する．全結合層で行われる処理を以下に示す．

$$\hat{\boldsymbol{h}} = \boldsymbol{W}^{fc} \times \boldsymbol{h} + \boldsymbol{b}^{fc} \tag{2}$$

$$\boldsymbol{p}(\boldsymbol{y} \mid \boldsymbol{x}; \boldsymbol{\theta}) = softmax(\hat{\boldsymbol{h}}) \tag{3}$$

ここで，$\boldsymbol{b}^{fc}$は全結合層のバイアス項，$\boldsymbol{\theta}$はパラーメータである．入力文の表現$\boldsymbol{x}$に対して確率$\boldsymbol{p}$が最も高い関係を予測結果とする．

3.3　遠距離教師データの関係抽出

遠距離教師データはノイズを多く含む特徴がある．そのため，関係を予測したいエンティティペアに対してそれらを含む文で構成されるバッグを作成し，バッグ単位で関係を予測するマルチインスタンス学習を行う．ここで，特徴抽出器として3.2節で説明した関係抽出器の単語埋め込み層から畳み込み層までの構造を用いる．以降では，バッグの特徴行列$\boldsymbol{B}$を求める過程を説明する．まず，バッグ内の各文を特徴抽出器に入力し，出力として各文の特徴ベクトル$\boldsymbol{h}_i$を得る．次に，Yeら[6]が提案した注意機構を適用してバッグの特徴行列$\boldsymbol{B} = [\boldsymbol{b}_1, \boldsymbol{b}_2, .., \boldsymbol{b}_r]$を求める．特徴行列$\boldsymbol{B}$の成分の一つである$b_k$の導出過程を以下に示す．

$$\boldsymbol{b}_k = \sum_{j=1}^{m} \alpha_{kj} \boldsymbol{h}_j \tag{4}$$

$$\alpha_{kj} = \frac{\exp(e_{kj})}{\sum_{j'=1}^{m} \exp(e_{kj'})} \tag{5}$$

$$e_{kj} = \boldsymbol{r}_k \boldsymbol{h}_j^\top \tag{6}$$

ここで，$\boldsymbol{r}_k$は全結合層の重み行列$\boldsymbol{W}$のk列目を表す．以上の計算を各関係にわたり行うことでバッグの特徴行列$\boldsymbol{B}$を得る．以降は3.2節と同様に全結合層に$\boldsymbol{B}$を入力し，予測結果を得る．

3.4　マルチタスクモデル

3.2節では人手のデータの関係抽出，3.3節では遠距離教師データの関係抽出について説明した．本手法では，これらをマルチタスク学習する．マルチタスク学習では，単語埋め込み層から畳み込み層までのネットワークのパラメータを共有し，全結合層については人手の教師データ用と遠距離教師データ用のものを用意

したネットワークを学習する．また，人手の教師データについてはインスタンス単位，遠距離教師データについてはバッグ単位でネットワークを学習する．ネットワークを学習する上で最小化する目的関数を以下に示す．

$$L_{batch} = \alpha \frac{1}{n_i} \sum_{k=1}^{n_i} L_k + (1-\alpha) \frac{1}{n_b} \sum_{k=1}^{n_b} L'_k \quad (7)$$

各タスクからクロスエントロピーにより算出した値をハイパーパラメータαによって比重を加えたものを利用する．

3.5 遠距離教師データの関係ラベルマッピング

本手法で使用する人手の教師データの関係クラスは6クラスであるが，遠距離教師データで定義された関係クラスは28クラスである．遠距離教師データの関係クラスのうち，一部は人手の教師データで用いられている関係クラスと名前は一致しているが，基準は異なる．さらに，人手の教師データでは定義されていない関係も存在する．そこで，遠距離教師データの関係予測に対しては以下の4つのタスク設定を検証した．

マッピングなし 遠距離教師データにある28クラスをそのまま利用し，学習．

6クラスマッピング (文字列ベース) 遠距離教師データにある28クラスを文字列一致により選択した6クラスにマッピングし，学習．

2クラスマッピング 遠距離教師データにある28クラスを正例・負例の2クラスにマッピングし，学習．

6クラスマッピング(開発データベース) 遠距離教師データにある28クラスを開発データでの評価により対応付けを行った6クラスにマッピングし，学習．

4 実験

本章では3章で説明した提案手法により行った2つの実験について述べる．ひとつの実験は，遠距離教師データでモデルを学習し，人手の教師データによりテストを行う．もう一つの実験では提案手法の評価を行うとともに，提案手法と人手の教師データのみで学習したベースラインとの性能比較も行う．

4.1節では実験設定について，4.2, 4.3節では2つの実験の結果と考察を述べる．

4.1 実験設定

本節では実験設定について述べる．4.1.1項では実験に使用したデータセット，4.1.2項ではモデルの学習設定について述べる．

4.1.1 データセット

実験では，人手の教師データと遠距離教師データの2種類のデータセットを使用した．以下では，それぞれのデータセットについて説明する．

- 人手の教師データ
  BioCreative VI track 5 CHEMPROTタスク[4]で用意されたデータセットを人手の教師データとして使用する．人手の教師データのデータ数の内訳を表1に示す．データセットでは5つの正例クラスと負例クラスの合計6つ関係クラスのが定義されている．
- 遠距離教師データ
  本手法では3.1節で説明した方法で遠距離教師データを作成した．作成した遠距離教師データを使用し実験を行った．遠距離教師データのデータ数の内訳を表1に示す．データセットでは28の関係クラスが定義されている．

4.1.2 学習設定

提案モデルを学習する際に設定したハイパーパラメータを表2に示す．最適化アルゴリズムにはAdam [3]を用い，人手の教師データにおける開発データのマイクロF値を最大にするようハイパーパラメータをチューニングした．

4.2 遠距離教師データでの学習の評価

まず，遠距離教師データでモデルを学習し，人手の教師データでテストを行った．遠距離教師データは6クラスマッピング（文字列ベース）を使用した．人手の教師データによるテストでは開発データを使用した．関係抽出の結果を表3に示す．マイクロ平均は5つの正例クラスから計算した．この結果から，作成した遠距離教師データは人手の教師データでの関係抽出にとって有用な情報となりうることが分かる．

表1 各データセットのデータ数

データ	訓練	開発	テスト	合計
人手	6,341	3,544	5,720	15,605
遠距離	43,079	14,115	11,604	68,798

表2 各種ハイパーパラメータ

パラメータ	値
単語ベクトルの次元数	50
単語位置ベクトルの次元数	5
畳み込みのフィルターサイズ	3
畳み込みのフィルター数	230
epoch数	80

4.3 提案手法の評価

提案手法の評価と人手の教師データのみで学習したベースラインの性能比較を行った．提案手法とベースラインの関係抽出の結果を表4, 5に示す．表中の適合率, 再現率, F値は正例クラスのマイクロ平均であり，それぞれの指標で最も高かったものを太字で示した．

まず，提案手法とベースラインの結果を比較する．F値について比較すると，開発データ，テストデータともに提案手法の6クラスマッピング(文字列ベース)がベースラインより高い性能を示した．再現率について比較すると，開発データでは提案手法のマッピングなし，テストデータでは提案手法の6クラスマッピング（開発データベース）がベースラインより高い性能を示した．提案手法では人手の教師データと遠距離教師データをマルチタスク学習し，モデルがベースラインより多くの教師データを学習している．そのため，モデルの再現率が向上したと考えられる．適合率について比較すると，開発データ，テストデータともに提案手法の2クラスマッピングがベースラインより高い性能を示した．しかし，その他提案手法はベースラインより低い性能となった．これは，提案手法では遠距離教師データデータを学習に使用しているため，遠距離距離教師データのノイズの影響により適合率が低下したと考えれられる．

次に，提案手法間の結果を比較する．提案手法では4種類の関係クラスのマッピングを検証した．結果から，関係クラスのマッピングがモデルの性能に影響することが分かった．

表3　遠距離教師データで学習した際のモデルの性能

関係クラス	適合率	再現率	F値（%）
NA	35.69	19.75	25.43
CPR:3	0.00	0.00	0.00
CPR:4	1.35	5.03	7.33
CPR:5	28.41	21.55	24.51
CPR:6	31.41	78.60	44.89
CPR:9	0.00	0.00	0.00
マイクロ平均	27.17	38.24	31.80

表4　開発データでの比較

	適合率	再現率	F値（%）
ベースライン	77.24	72.44	74.76
マッピングなし	74.72	**75.86**	75.28
6 クラスマッピング (文字列ベース)	77.12	74.70	**75.89**
2 クラスマッピング	**78.30**	70.17	74.01
6 クラスマッピング (開発データベース)	74.74	74.74	74.74

表5　テストデータでの比較

モデル	適合率	再現率	F値（%）
ベースライン	68.70	71.55	70.09
マッピングなし	67.09	73.16	69.99
6 クラスマッピング (文字列ベース)	69.44	71.87	**70.63**
2 クラスマッピング	**72.14**	68.81	70.43
6 クラスマッピング (開発データベース)	67.51	**73.68**	70.46

5 おわりに

本研究では，遠距離教師データをマルチタスク学習により援用して人手の教師データの薬物タンパク質相互作用抽出を行う手法を提案した．実験により，作成した遠距離教師データが有用な情報となりうることと遠距離教師データの関係クラスのマッピングがモデルの性能に影響することが分かった．

今後は，関係クラスのマッピングを含めて遠距離教師データのノイズを緩和する手法を検討し，人手の教師データの薬物タンパク質相互作用抽出へのさらなる有効利用を目指す．

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] Beltagy et al. Combining distant and direct supervision for neural relation extraction. In *NAACL*, 2019.

[2] Dong et al. Multi-task learning for multiple language translation. In *ACL*, 2015.

[3] Kingma et al. Adam: A method for stochastic optimization. In *Proceedings of ICLR*, 2015.

[4] Krallinger et al. Overview of the biocreative vi chemical-protein interaction track. In *Proceedings of the BioCreative VI Workshop*, 2017.

[5] Mintz et al. Distant supervision for relation extraction without labeled data. In *Proceedings of ACL*, 2009.

[6] Ye et al. Distant supervision relation extraction with intra-bag and inter-bag attentions. In *NAACL*, 2019.

[7] Zeng et al. Distant supervision for relation extraction via piecewise convolutional neural networks. In *Proceedings of EMNLP*, 2015.