データベースの説明文を利用した薬物相互作用抽出

浅田 真生　　三輪 誠　　佐々木 裕

豊田工業大学

{sd17402, makoto-miwa, yutaka.sasaki}@toyota-ti.ac.jp

1 はじめに

薬物相互作用とは，患者に薬物を併用投与する場合，薬物の本来の作用が増強・減弱したり，副作用が増強したりすることである．薬物相互作用は薬学論文等で日々報告されているが，そのデータベースへの登録は人手で行われており，その報告に対してデータベースの整備が追いついていない．このため，文書から薬物相互作用自動抽出を行う研究は重要であり，近年は深層学習を用いた手法 [1, 2] が注目されている．深層学習には薬学専門家により正解付けされたコーパスが必要であるが，コーパス作成はコストが大きいため増やすことが難しい．そこで，現在までに整備されている薬物データベースの情報の薬物相互作用抽出への有効活用が求められている．

多くの薬物相互作用抽出では，薬物エンティティ自身の情報を用いず，エンティティの周辺単語から薬物相互作用を予測する．薬物データベース DrugBank [3] には薬物についての説明文が記載されおり，薬物の説明文を薬物の情報として利用できれば，他の薬物との相互作用を求める助けとなる可能性がある．

本研究では，データベースに記載された薬物の説明文を文献からの薬物相互作用抽出に援用することを目的として，薬物説明文と相互作用を抽出する入力文をそれぞれ畳み込みニューラルネットワーク (Convolutional Neural Network; CNN) で表現し，二つの CNN を同時に学習する手法を提案する．提案したモデルを SemEval-2013 Task 9 データセット [4] で評価したところ，説明文の追加により精度の向上を達成し，説明文の薬物相互作用抽出への有効性がわかった．

2 CNN を用いた薬物相互作用抽出

Liu ら [1] は CNN を用いた薬物相互作用抽出手法を提案し，高い性能を示している．しかし，モデルの学習に使用するコーパスは人手でタグ付けされたものであり，より高い性能のために正解付きコーパスを増やすことはコストが大きい．また，相互作用抽出には薬物エンティティ周辺の単語の情報のみを利用しており，薬物エンティティの情報を利用できていない．

3 提案手法

本研究では，データベース中の薬物の説明文を薬物相互作用抽出に援用するために，薬物相互作用を抽出する入力文およびデータベースに記載された薬物の説明文をそれぞれ CNN で表現し，二つの CNN を同時に学習しながら薬物相互作用抽出を行う手法を提案する．提案した手法の全体図を図 1 に示す．提案手法では，薬物ペアについて述べた文が，どの種類の薬物相互作用を持つかの分類問題を解くことによって薬物相互作用抽出を行う．入力文中の薬物エンティティに対応付けられた外部データベースの薬物エントリから薬物の説明文を獲得して CNN の入力とする．

3.1 薬物説明文の表現

二つの薬物エンティティの説明文をそれぞれ $S^{d1} = \{w_1^{d1}, w_2^{d1}, \cdots, w_{n_1}^{d1}\}$，$S^{d2} = \{w_1^{d2}, w_2^{d2}, \cdots, w_{n_2}^{d2}\}$ とする．S^{d1} を CNN の入力とし，薬物説明文を固定長の実数値ベクトルで表現する．単語 w_i^{d1} の単語ベクトルを $\boldsymbol{w}_i^{d1}$，畳み込みのフィルタサイズの集合を $\{k_1^d, \cdots, k_{L_d}^d\}$，畳み込みのフィルタ数を J_d，畳み込みテンソル，バイアスをそれぞれ $\boldsymbol{W}^{dconv}$，$\boldsymbol{b}^{dconv}$ とし，畳み込みの処理を以下のように行う．

$$\boldsymbol{z}_{i,l}^{d1} = [(\boldsymbol{w}_{i-(k_l-1)/2}^{d1}), \cdots, (\boldsymbol{w}_{i-(k_l+1)/2}^{d1})] \quad (1)$$

$$m_{i,j,l}^{d1} = \mathrm{relu}(\boldsymbol{W}_j^{dconv} \odot \boldsymbol{z}_{i,l}^{d1} + \boldsymbol{b}^{dconv}) \quad (2)$$

ここで $\odot$ は要素積である．得られた表現について max プーリングを行い，S^{d1} の表現を次のように得る．

$$\boldsymbol{h}_l^{d1} = [h_{1,l}^{d1}, \cdots, h_{J_d,l}^{d1}],\ h_{j,l}^{d1} = \max_i m_{i,j,l}^{d1} \quad (3)$$

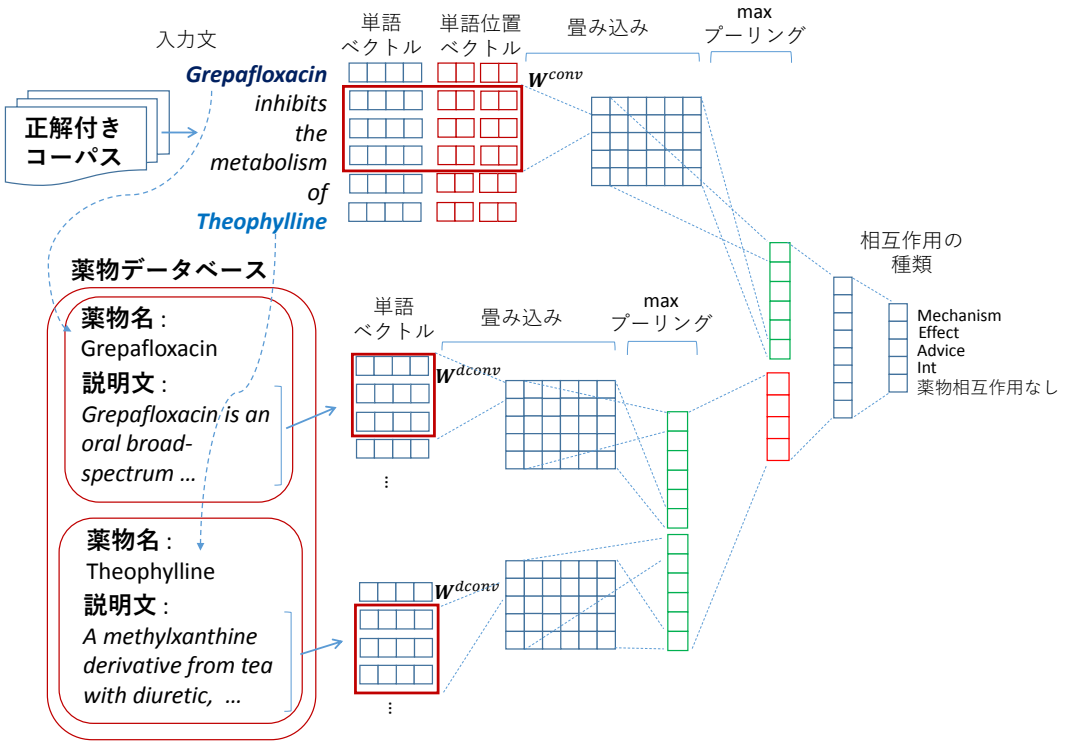

図 1: 提案手法の全体図

$$\boldsymbol{h}^{d1} = [\boldsymbol{h}_1^{d1}; \ldots; \boldsymbol{h}_L^{d1}]. \tag{4}$$

ここで，$[;]$ はベクトルの連結を表す．同様にして，S^{d2} の表現 $\boldsymbol{h}^{d2}$ を得る．二つの説明文の表現を連結し，

$$\boldsymbol{h}^{desc} = \mathrm{relu}(\boldsymbol{W}^{desc}[\boldsymbol{h}^{d1}; \boldsymbol{h}^{d2}] + \boldsymbol{b}^{desc}) \tag{5}$$

と変形することで説明文の表現 $\boldsymbol{h}^{desc}$ を得る．ここで，$\boldsymbol{W}^{desc}$，$\boldsymbol{b}^{desc}$ は重みとバイアスである．

3.2 入力文の表現

二つの薬物エンティティ e_1, e_2 の間の薬物相互作用について述べた入力文 $S = (w_1, w_2, \cdots, w_n)$ が与えられたときに入力文を CNN により実数値ベクトルで表現する．単語 w_i の単語ベクトルを $\boldsymbol{w}_i$，エンティティ e_1, e_2 に対する単語位置ベクトルをそれぞれ $\boldsymbol{w}_{i,1}^p, \boldsymbol{w}_{i,2}^p$ とする．単語ベクトルと単語位置ベクトルを連結して以下のように表現する．

$$\boldsymbol{w}_i = [(\boldsymbol{w}_i^w)^{\mathrm{T}}; (\boldsymbol{w}_{i,1}^p)^{\mathrm{T}}; (\boldsymbol{w}_{i,2}^p)^{\mathrm{T}}]^{\mathrm{T}} \tag{6}$$

また，フィルタサイズの異なる L 個の畳み込みフィルタを用意する．フィルタサイズの集合を式 (7) のように表す．

$$\boldsymbol{k} = \{k_1, k_2, \cdots, k_l, \cdots, k_L\} \tag{7}$$

フィルタサイズが k_l であるとき，単語 w_i とその前後 $\frac{k_l-1}{2}$ 単語のベクトルを連結し，以下のように表現する．

$$\boldsymbol{z}_{i,l} = [(\boldsymbol{w}_{i-(k_l-1)/2})^{\mathrm{T}}, \cdots, (\boldsymbol{w}_{i-(k_l+1)/2})^{\mathrm{T}}]^{\mathrm{T}} \tag{8}$$

畳み込みのテンソルを $\boldsymbol{W}^{conv}$，バイアスを $\boldsymbol{b}^{conv}$，畳み込みのフィルタ数を J とする．$\boldsymbol{z}_i$ に畳み込み処理を行い，以下の結果を得る．

$$m_{i,j,l} = \mathrm{relu}(\boldsymbol{W}_j^{conv} \odot \boldsymbol{z}_{i,l} + \boldsymbol{b}^{conv}) \tag{9}$$

全単語の畳み込みの出力のうち，max プーリングにより以下のように入力文の表現 $\boldsymbol{h}^{sent}$ を得る．

$$\boldsymbol{m}_l = [m_{1,l}, m_{2,l}, \cdots, m_{J,l}], \ m_{j,l} = \max_i m_{i,j,l} \tag{10}$$

$$\boldsymbol{h}^{sent} = [\boldsymbol{m}_1, \cdots, \boldsymbol{m}_L] \tag{11}$$

3.3 データベース中の薬物説明文を利用した薬物相互作用抽出

入力文の表現と薬物の説明文の表現を以下のように連結する．

$$\boldsymbol{h}^{all} = [\boldsymbol{h}^{sent}; \boldsymbol{h}^{desc}] \tag{12}$$

薬物相互作用の予測 $\hat{\boldsymbol{y}}$ を以下のように得る．

$$\boldsymbol{h}^{all1} = \mathrm{relu}(\boldsymbol{W}^{all1}\boldsymbol{h}^{all} + \boldsymbol{b}^{all1}) \quad (13)$$

$$\hat{\boldsymbol{y}} = \mathrm{softmax}(\boldsymbol{W}^{all2}\boldsymbol{h}^{all1} + \boldsymbol{b}^{all2}) \quad (14)$$

ここで，$\boldsymbol{W}^{all1}, \boldsymbol{W}^{all2}$ は重み行列，$\boldsymbol{b}^{all1}, \boldsymbol{b}^{all2}$ はバイアスである．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 実験設定

4.1 正解付きコーパス

正解付きコーパスには SemEval-2013 Task 9 データセット [4] を用いた．このデータセットは，薬物を含んだ文から構成され，どの単語が薬物であるかはあらかじめ特定されている．データセットは以下に示す 4 種類の薬物相互作用が正解付けされており，本実験では薬物のペアの薬物相互作用の有無および，薬物相互作用を持つ場合は 4 種類のうちのどの薬物相互作用を持つかを求める．

Mechanism: 二つの薬物が薬物動態的作用を持つ．
Effect: 二つの薬物が薬力学的作用を持つ．
Advice: 二つの薬物を併用する際の推奨を表す．
Int: 薬物相互作用を持つことのみを表す．

データセットの内訳を表 1 に示す．表 1 より，薬物相互作用を持つペアよりも薬物相互作用を持たないペアが多いことがわかる．入力文は GENIA Tagger[1] により単語分割を行った．また，Liu ら [1] と同様に，ターゲットの二つの薬物はそれぞれ “*DRUG1*”，“*DRUG2*” に，それ以外の薬物は “*DRUGOTHER*” に置き換える前処理を行った．

表 1: 正解付きコーパスの内訳

	訓練データ	評価データ
文数	6,976	1,299
ペア数	27,792	5,716
薬物相互作用あり	4,021	979
Mechanism	1,319	302
Effect	1,687	360
Advice	826	221
Int	189	96
薬物相互作用なし	23,771	4,737

[1] http://www.nactem.ac.uk/GENIA/tagger/

4.2 薬物データベースと説明文

使用した薬物データベース DrugBank [3] には 10,000 件以上の薬物が収録されており，それぞれに薬物の一般的な事実や組成について述べた説明文が付与されている．説明文を GENIA Sentence Splitter[2] により文分割を行い，GENIA Tagger を用いて単語分割を行った．

4.3 薬物エンティティとデータベースエントリの対応付け

コーパス中の薬物エンティティとデータベースエントリとの対応付けは部分文字列一致により行った．文字列には DrugBank の薬物エントリが持つ以下の項目を利用した．

Name: 薬物の見出し語
Internatinal-brand: 薬物の製品名
Product: 薬物の製剤名

文字列一致を取る際には，コーパスのエンティティ，薬物エントリの項目どちらも小文字に変換して行った．文字列一致した薬物エントリが複数存在した場合は，一致率が最も高いエントリを使用した．コーパス中の訓練データの 2,242 個の薬物のうち，一致が取れたものは 1,934 個 (86.26%)，評価データの 854 個の薬物のうち，一致が取れたものは 748 個 (87.58%) であった．

4.4 単語ベクトルの事前学習

単語ベクトルの事前学習は Skip-gram [5] によって行った．事前学習に用いたコーパスは PubMed 2014 のデータで，語彙数は 215,840 であった．前処理によって置き換えられた “*DRUG1*” および “*DRUG2*” の単語ベクトルの初期値は “*drug*” と同じ値を用いた．訓練データ中に新しく出現した単語のベクトルは，事前学習した全単語の単語ベクトルの平均値を使用した．訓練データ中の単語のうち，出現頻度が 1 である単語を未知語として扱った．

4.5 学習設定

単語位置ベクトルの初期値は，最小値 −0.01, 最大値 0.01 の一様分布によって決定した．

[2] http://www.nactem.ac.uk/y-matsu/geniass/

最適化アルゴリズムには Adam [6] を用い，学習率は 0.001, ミニバッチサイズ 50 として，ミニバッチ学習を行った．L2 正則化を行い，係数は 0.0001 とした．更新ごとにモデルのパラメータを保存し，予測時は，これまでに保存したパラメータを平均化したモデルを用いた．入力文の CNN と説明文の CNN は単語ベクトルを共有せず，どちらもファインチューニングした．

入力文の表現に用いた CNN と説明文の表現に用いた CNN のハイパーパラメータを表 2 に示す．

5 結果

SemEval 2013 Task 9 コーパスにおける薬物相互作用抽出精度を表 3 に示す．Liu ら [1]，Quan ら [2] は提案手法と同様に CNN を用いて薬物相互作用抽出を行っている．また，Asada ら [7] は薬物エンティティの分子構造情報を用いて，CNN による薬物相互作用抽出を行い，高い精度を示している．表 3 から，データベースに記載された薬物の説明文を利用することにより，薬物相互作用抽出の Precision，Recall，F 値が向上することが分かった．

6 おわりに

本研究では，データベース中の薬物の説明文が薬物相互作用抽出に有用であるか調べるために，相互作用を抽出する入力文，およびデータベース中の薬物の説明文をそれぞれ CNN で表現し，二つの CNN を同時に学習しながら薬物相互作用抽出を行う手法を提案した．モデルを SemEval-2013 Task 9 データセットで学習，評価した結果，薬物の説明文を用いることで薬物相互作用抽出の F 値が向上することが分かった．今後は，薬物データベースの様々な情報を統合的に利用する手法を検討したい．

表 2: 入力文 CNN・説明文 CNN のハイパーパラメータ

パラメータ名	入力文	説明文
単語ベクトルの次元数	200	200
中間層の次元数	500	100
畳み込みのフィルタサイズ	{3,5,7}	{3,5,7}
畳み込みのフィルタ数	100	100

表 3: 薬物相互作用抽出精度

手法	Precision	Recall	F 値 (%)
コーパスのみ	71.97	68.44	70.16
コーパス＋説明文	72.87	71.09	71.97
Liu [1]	75.70	64.66	69.75
Quan [2]	75.99	65.25	70.21
Asada [7]	73.31	71.81	72.55

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] Liu et al. Drug-Drug Interaction Extraction via Convolutional Neural Networks. *Computational and mathematical methods in medicine*, 2016.

[2] Quan et al. Multichannel Convolutional Neural Network for Biological Relation Extraction. *BioMed research international*, 2016.

[3] Wishart et al. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic acids research*, Vol. 46, No. D1, pp. D1074–D1082, 2017.

[4] Segura-Bedmar et al. SemEval-2013 task 9 : Extraction of drug-drug interactions from biomedical texts (DDIExtraction 2013). In *Proceedings of SemEval 2013*, pp. 341–350, 2013.

[5] Mikolov et al. Distributed Representations of Words and Phrases and Their Compositionality. In *Proceedings of NIPS*, pp. 3111–3119, 2013.

[6] Kingma and Ba. Adam: A method for stochastic optimization. In *Proceedings of ICLR*, 2015.

[7] Asada et al. Enhancing Drug-Drug Interaction Extraction from Texts by Molecular Structure Information. In *Proceedings of ACL*, pp. 680–685. ACL, 2018.