複数の事前学習モデルを併用した化学分野の関係抽出

肥合 智史 [1,3]　嶋田 和孝 [1,3]　渡邊 大貴 [2,3]　三浦　明波 [2,3]　岩倉 友哉 [2,3]
[1] 九州工業大学　[2] 株式会社富士通研究所　[3] 理研 AIP-富士通連携センター

{s_hiai, shimada}@pluto.ai.kyutech.ac.jp
{watanabe-taiki, miura.akiba, iwakura.tomoya}@fujitsu.com

1　はじめに

近年，大量の化学分野の論文が出版，公開されている．そして，その化学論文には，新たな化学物質に関する情報や，化学物質間の相互作用等の情報が含まれている．そうした情報を抽出することは有益であるが，日々大量の論文が公開されているため，人手で抽出するのはコストが大きい．よって，大規模なデータに適用できる情報抽出手法へのニーズが高まっている．

化学分野の関係抽出タスクにおいて，BERT モデルによる手法が大きな成果を上げている [2, 8]．Lee ら [8] は，化学分野のコーパスによって事前学習をした BERT モデルを用いた化学分野における関係抽出手法を提案した．そして，化学分野の関係抽出タスクにおいて最高精度を達成した．しかし，BERT モデルサイズは大きくなる傾向にあり，大規模データに対する推論には大きな時間がかかる．この時間を短縮するには，高機能な計算資源の導入や，計算資源を複数用意し並列処理することなどが考えられるが，計算資源導入や運用面でのコストが問題となる．よって，高精度かつより小規模なモデルが必要である．

BERT より軽量なモデルを利用した手法として，ELMo [9] を利用したものがある．ELMo は多層の双方向 LSTM による単語レベルの言語モデルであり，このモデルにより文脈を考慮した単語分散表現を得ることができる．Jin ら [7] は，化学分野の関係抽出タスクにおいて，化学論文コーパスによって事前学習した ELMo による単語分散表現を入力とした関係抽出モデルを提案した．

本研究では，Contextual String Embeddings (CSE) [1] を利用した化学分野の関係抽出手法を提案する．CSE は，ELMo による分散表現と同様に，軽量な言語モデルを用いて得られる文脈を考慮した単語の分散表現である．CSE では，文字レベルの言語モデルが利用され，文字レベルの情報を含んだ単語分散表現が得られる．化学物質名は共通した特徴的な部分文字列を含むことが多く，そのような文字列は化合物の種類や性質を推定する上で重要な手がかりとなることが考えられる．よって，文字レベルの情報は化学分野のタスクにおいて有用であると考えられる．そこで，CSE の文字レベル言語モデルの事前学習に化学分野の大規模コーパスを用いて，関係分類タスクに適用する．

本研究の貢献は次の二点である．

- ELMo 分散表現と CSE の 2 種類の事前学習モデルによる分散表現の利用により化学分野の関係抽出の精度が向上することを確認した．
- GAD 関係分類データセットにおいて，CSE の利用により，最高精度を達成しつつ BERT に基づく関係抽出より高速に処理可能な関係抽出モデルを構築した．

2　関連研究

CSE [1] の文字レベルの言語モデルによって，文字レベルの情報を含んだ文脈を考慮した分散表現を獲得できる．言語モデルは単層の双方向 LSTM によって構築され，BERT より軽量なモデルである．Sharma ら [10] は CSE の言語モデルを化学論文の大規模コーパスによって事前学習し，化学ドメインの固有表現抽出タスクに利用することで，その有効性を確認した．また，Watanabe ら [11] は，化学論文コーパスによって事前学習した CSE の言語モデルを用いたマルチタスク学習モデルによって，固有表現抽出タスクにおいて最高精度を達成した．本手法でも CSE の言語モデルを大量の化学論文コーパスによって事前学習して利用する．Sharma らや Watanabe らは化学分野の固有表現抽出タスクにおける CSE の有効性を確認したが，関係抽出タスクにおけるその効果を検証していない．本研究では，化学ドメインの関係抽出タスクにおいて CSE の有効性を検証する．

3　手法

図 1 に関係抽出手法の全体像を示す．本手法では，事前に CSE の言語モデルを化学論文の大規模コーパスによって訓練する．その言語モデルによって，図 1 下部のように各単語に対応する分散表現を得る．次に，

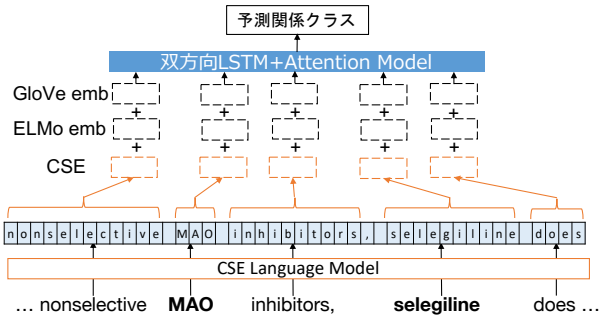

図 1: 提案手法の概要

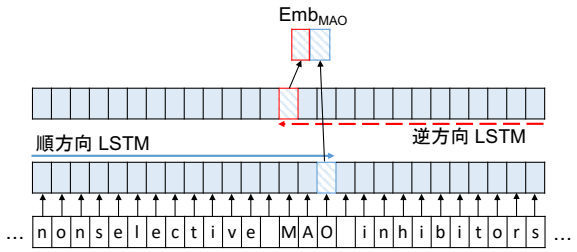

図 2: CSE の言語モデルの概要

図 1 上部のように，CSE と既存の分散表現を連結した入力を用いた関係分類モデルを構築する．CSE における言語モデルとその学習について 3.1 節で説明する．関係分類モデルについて 3.2 節で説明する．

3.1 Contextual String Embeddings

図 2 に CSE の文字レベル言語モデルの概要を示す．まず，双方向 LSTM を利用した文字レベルの言語モデルを化学論文の大規模コーパスによって事前学習する．この言語モデルによって入力文の各単語に対する分散表現が得られる．

図 2 のように，単語の最後の文字を読み込んだ順方向 LSTM の隠れ層のベクトルと，最初の文字を読み込んだ逆方向 LSTM の隠れ層のベクトルを連結する．このベクトルが単語の分散表現である．

我々はこの言語モデルの学習のために，PubMed[1]，PMC[2]，ChemRxiv[3] から取得した科学論文のアブストラクトや本文を用いた．PubMed には約 190 万件，PMC には約 270 万件，ChemRxiv には約 300 万件の論文に関する情報が含まれている．

3.2 関係分類モデル

関係分類モデルには双方向 LSTM+Attention モデル [12] を用いる．そのモデルへの入力として，公開されている既存の二種類の分散表現に加えて，前節で説明した CSE を利用する．既存の二種類の分散表現は，GloVe 分散表現[4]と，ELMo 分散表現[5]である．GloVe 分散表現は化学分野に限らないコーパスである Wikipedia と Gigaword コーパスで学習されたものである．一方，ELMo 分散表現は化学分野のコーパスである PubMed コーパスで学習されたものであるが，単語単位で学習をしており，文字単位の情報は考慮されていない．各単語について，既存の二種類の分散表現と前節の CSE を連結して入力する．出力はどの関係であるかのラベル (予測関係クラス) である．各単語ベクトルを $X = x_1, x_2, ..., x_n$ と表すと，次のようにクラスラベルを予測する．

表 1: 各データセットにおけるサンプル数

データセット	正例数	負例数
GAD	2,801	2,529
ChemProt	9,986	31,298

$$\overrightarrow{h_i} = \overrightarrow{LSTM}(x_i, \overrightarrow{h}_{i-1}) \text{ (順方向)} \quad (1)$$

$$\overleftarrow{h_i} = \overleftarrow{LSTM}(x_i, \overleftarrow{h}_{i+1}) \text{ (逆方向)} \quad (2)$$

$$h_i = [\overrightarrow{h_i}; \overleftarrow{h_i}] \quad (3)$$

h_i は双方向 LSTM モデルの隠れ層のベクトルである．$[\cdot;\cdot]$ は二つのベクトルの連結を表す．この各時点での h_i を基に，次の式で重みを計算する．

$$m_i = \omega^T tanh(h_i) \quad (4)$$

$$a_i = \frac{\exp(m_i)}{\sum_{j=1}^{n} \exp(m_j)} \quad (5)$$

ω は学習可能なパラメータのベクトルである．そして次のように最終的な隠れ層の状態 h^* を計算する．

$$r = \sum_{i=1}^{n} a_i h_i \quad (6)$$

$$h^* = tanh(r) \quad (7)$$

そして次のように，予測ラベル $\hat{y}$ に対応するインデックスを出力する．

$$P(y|X) = softmax(W_s h^* + b_s) \quad (8)$$

$$\hat{y} = \arg\max_y P(y|X) \quad (9)$$

訓練においては，損失関数にはクロスエントロピーロスを用い，最適化手法には SGD [4] を用いる．学習率は 0.1 に設定し，バッチサイズは 32，隠れ層の次元数は 256 に設定する．

[1] http://www.nlm.nih.gov/databases/download/pubmed_medline.html
[2] https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/
[3] https://chemrxiv.org/
[4] https://nlp.stanford.edu/projects/glove/
[5] https://allennlp.org/elmo

表 2: 実験結果 (太字が最高精度，下線が二番目に高い精度を表す．E-SVM，BioBERT はそれぞれ [3] と [8] における報告値である．SciBERT は [2] において ChemProt の報告値があるが，他の先行研究とは実験設定が異なり，そのまま比較できない値であったため，実験設定を合わせて GAD，ChemProt とも実験し精度を確認した．)

データセット		E-SVM	SciBERT	BioBERT	Baseline1 (GloVe)	Baseline2 (GloVe +ELMo)	Baseline3 (GloVe +CSE)	提案手法 (GloVe +ELMo +CSE)
GAD	P	79.21	**85.54**	76.46	77.31	<u>83.10</u>	78.74	82.64
	R	**89.25**	80.61	<u>87.65</u>	80.01	85.43	87.43	86.36
	F	83.93	82.83	81.61	78.59	<u>84.16</u>	82.83	**84.38**
ChemProt	P	-	**79.73**	<u>77.02</u>	58.13	76.49	67.41	76.05
	R	-	<u>70.85</u>	**75.90**	57.43	65.82	58.79	66.17
	F	-	<u>75.03</u>	**76.46**	57.78	70.76	62.81	70.77

4 実験

本節では，提案手法の評価のための実験とその結果について説明する．4.1 節において，実験に利用する関係分類データセットと比較対象の手法について説明する．4.2 節において，分類結果について説明する．

4.1 実験設定

化学分野の関係分類データセットである，GAD [5]，ChemProt [6] を用いて評価する．各データセットに含まれるサンプル数を表 1 に示す．GAD データセットは遺伝子と疾患間の関係に関するデータセットである．ChemProt データセットにはタンパク質と化合物間の関係に関するデータセットである．ChemProt データセットにおいては，タンパク質と化合物間の関係が 5 種類アノテートされており，この 5 種類の関係のいずれかであるか，もしくはどれでもない負例であるかの合計 6 クラスの分類タスクとなる．表 1 においては 5 種類の関係の合計数を正例数として示している．ChemProt データセットにおいては，評価用データが分けて公開されているが，GAD データセットでは分けられていない．よって，GAD データセットでは 10 分割交差検定によって評価する．また，各データセットにおいて，文章中の関係を判別する対象のエンティティを表現するために，対象のエンティティを特定のタグで囲む方法を採用した[6]．

提案手法は，flair フレームワーク[7]をベースにコードを修正し実装した．

比較対象として，同じ双方向 LSTM+Attention モデルによる分類で，入力として GloVe 分散表現のみを用いた場合 (Baseline1)，GloVe と ELMo 分散表現を用いた場合 (Baseline2) と，GloVe と CSE 分散表現を用いた場合 (Baseline3) での分類も行い，各分散表現の影響を確認する．さらに，大規模モデルである BERT を利用した手法である BioBERT [8]，SciBERT [2] と精度を比較する．BioBERT は BERT モデルを大規模な生物医学分野の論文コーパスによって事前学習したモデルである．SciBERT は計算機科学論文コーパスと生物医学論文コーパスで BERT を事前学習したモデルであり，モデルの入力の単位に科学関係の語彙リストを利用している．また，各データセットの state-of-the-art (SOTA) と比較する．ChemProt データセットにおける SOTA は，BioBERT である．GAD データセットにおける SOTA は，アンサンブル SVM を利用した手法 [3] (E-SVM) である．

4.2 実験結果

表 2 に各データセットにおける分類結果を示す．結果を適合率 (P)，再現率 (R)，F1 値 (F) によって評価した．GAD，ChemProt 両データセットにおいて，CSE を追加した Baseline3 の F1 値は，GloVe のみ利用した Baseline1 の F1 値より高くなった．また，提案手法の F1 値が Baseline1，2，3 より高くなった．よって，既存の GloVe や ELMo といった単語単位の分散表現に加えて，文字単位の情報を含む CSE を利用することは有効であると考えられる．

GAD データセットにおいては，提案手法の F1 値 84.38 が最高精度となり，BioBERT や SciBERT といった大規模モデルによる分類より高精度となった．しかし，ChemProt データセットにおいては，提案手法はどの指標においても SOTA の BioBERT よりも低い精度となった．表 1 に示した通り，ChemProt データセットでは負例数が正例 5 クラス合計数の約 3 倍であり，大きな偏りがある．そして，提案手法の ChemProt の

[6]GAD においては遺伝子，疾患をそれぞれ <gene>，<dise> タグで囲み，Chemprot においてはタンパク質，化合物をそれぞれ <gene>，<chem> タグで囲んだ．

[7]https://github.com/flairNLP/flair

表 3: ChemProt のテストセットの処理時間

手法	時間 (s)
Baseline2 (GloVe+ELMo)	212.51
Baseline3 (GloVe+CSE)	86.53
提案手法 (GloVe+ELMo+CSE)	239.06
BioBERT	680.39

分類結果を観察すると，正例の誤分類の約 92%は他の正例クラスではなく，負例クラスへの誤分類であった．よって，提案手法はデータの偏りに弱く，このことが BioBERT と提案手法の再現率の大きな差 (9.73 ポイント) の原因であると考えられる．したがって今後，データの偏りへの対処が必要である．

また，表 3 に Baseline2，3，提案手法と BioBERT の ChemProt のテストセットの処理時間を示す[8]．分類精度では Baseline3 は各データセットで Baseline2 や提案手法より低い F1 値であったが，実行速度は Baseline2 や提案手法よりも 2.5 倍程度高速であった．また，提案手法は BioBERT よりも約 2.85 倍高速であった．GAD データセットの分類精度では BioBERT より提案手法が高い F1 値となっており，提案手法は BioBERT より軽量かつ高精度な分類モデルであることが確認できた．一方 ChemProt データセットでは，提案手法は BioBERT に満たない精度であった．提案手法より F1 値が低かった Baseline3 が提案手法よりさらに高速であることを踏まえると，分類精度と実行速度のトレードオフの関係が確認できた．提案手法では今後，一定の処理速度を保ったまま精度を向上させることが課題である．

5 おわりに

本研究では，複数の言語モデルを組み合わせて軽量で高精度な化学分野の関係抽出手法を提案した．化学論文の大規模コーパスによって事前学習された複数の言語モデルによる分散表現がそれぞれ関係抽出の精度向上に寄与することを確認した．評価実験においては，GAD データセットにおいて最高精度を達成した．また，処理速度の比較において，提案手法が BERT を利用した手法より高速であることを確認した．しかし，ChemProt データセットでは BERT を利用した手法より低精度となったため，精度の向上が課題である．

[8]SciBERT は BioBERT と同じ BERT モデルを用いた手法であり，処理時間は BioBERT と大きく変わらないと考えれるため，ここでは BioBERT のみ評価した．

参考文献

[1] A. Akbik, T. Bergmann, D. Blythe, K. Rasul, S. Schweter, and R. Vollgraf. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *Proceedings of the 2019 NAACL*, pp. 54–59, 2019.

[2] I. Beltagy, K. Lo, and A. Cohan. Scibert: Pretrained language model for scientific text. In *Proceedings of the 2019 EMNLP-IJCNLP*, 2019.

[3] B. Bhasuran and J. Natarajan. Automatic extraction of gene-disease associations from literature using joint ensemble learning. *PLOS ONE*, 13:e0200699, 2018.

[4] L. Bottou. Stochastic gradient learning in neural networks. In *Proceedings of Neuro-Nımes*, p. 91, 1991.

[5] Á. Bravo, J. Piñero, N. Queralt-Rosinach, M. Rautschka, and L. I. Furlong. Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research. *BMC Bioinformatics*, 16:55, 2015.

[6] R. Islamaj Doğan, S. Kim, A. Chatr-aryamontri, C.-H. Wei, D. C. Comeau, R. Antunes, S. Matos, Q. Chen, A. Elangovan, N. C. Panyam, K. Verspoor, H. Liu, Y. Wang, Z. Liu, B. Altınel, Z. M. Hüsünbeyi, A. Özgür, A. Fergadis, C.-K. Wang, H.-J. Dai, T. Tran, R. Kavuluru, L. Luo, A. Steppi, J. Zhang, J. Qu, and Z. Lu. Overview of the biocreative vi precision medicine track: mining protein interactions and mutations for precision medicine. In *the sixth BioCreative challenge evaluation workshop*, Vol. 1, pp. 141–146, 2017.

[7] Q. Jin, B. Dhingra, W. Cohen, and X. Lu. Probing biomedical embeddings from language models. In *Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP*, pp. 82–89, 2019.

[8] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. In *arXiv preprint arXiv:1901.08746*, 2018.

[9] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. Deep contextualized word representations. In *Proceedings of NAACL*, 2018.

[10] S. Sharma and R. Daniel Jr. Bioflair: Pretrained pooled contextualized embeddings for biomedical sequence labeling tasks. In *arXiv preprint arXiv:1908.05760*, 2019.

[11] T. Watanabe, A. Tamura, T. Ninomiya, T. Makino, and T. Iwakura. Multi-task learning for chemical named entity recognition with chemical compound paraphrasing. In *Proceedings of the 2019 EMNLP-IJCNLP*, pp. 6243–6248, 2019.

[12] P. Zhou, W. Shi, J. Tian, Z. Qi, B. Li, H. Hao, and B. Xu. Attention-based bidirectional long short-term memory networks for relation classification. In *Proceedings of the 54th ACL*, pp. 207–212, 2016.