CVAE による複数データセットからの固有表現抽出

大井 拓　三輪 誠　佐々木 裕
豊田工業大学
{sd23404,makoto-miwa,yutaka.sasaki}@toyota-ti.ac.jp

概要

固有表現抽出などで主流となっている教師あり深層学習で高い性能を達成するには大量のラベル付きデータが必要になる．ラベル付きデータを増やすために既存のラベル付きデータセットを複数組み合わせて使用する方法が考えられるが，データセットごとにラベル付けの違いがあり性能が上がるとは限らないという問題がある．そこで本研究ではデータセットごとのラベル付けの違いを CVAE (Conditional Variational Autoencoder) を用いた損失関数を加えることで弱い制約として固有表現抽出モデルに与える手法を提案する．実験では生物医学文書に対する固有表現抽出において提案手法の有効性を示した．

1　はじめに

固有表現抽出は自然言語処理分野における基礎的な問題であり，文書からの情報抽出の第一段階として重要である．特に深層学習の進歩により，教師あり深層学習を用いた手法が主流となっている [1, 2]．教師あり深層学習を用いた固有表現抽出モデルの性能は一般にラベル付きデータの量に依存するが，ラベル付きデータの作成には高いコストがかかる．

学習に用いるラベル付きデータを増やすために，既存のラベル付きデータセットを組み合わせて使用する方法が考えられる．しかし，同じ固有表現を対象としたデータセットであっても，データセットごとに特徴があるため，既存のラベル付きデータセットを複数組み合わせた学習は単純ではない．具体的にはそれぞれのデータセットごとにラベルの種類の違い，基準の違いによって，ラベル付けされる用語に違いがあるため，同様のデータとして扱うことが難しいという問題がある．例えば，生物医学分野の文章に対する固有表現抽出のデータセットである BC5CDR (BioCreative V CDR task corpus) [3] と NCBI (NCBI Disease corpus) [4] では，ともに “pain” が含まれており，BC5CDR では病名としてラベル付けされているが，NCBI ではラベル付けされていない．

複数のデータセットを組み合わせて使用する既存手法としてマルチタスク学習を行う手法がある [5]．マルチタスク学習では，ベースとなるモデルを共有しつつ，データセットごとに異なる分類層を用意することで，上記のようなデータセットごとの違いを気にせずに学習できる．一方でデータセットごとの違いを気にせず別のものとして扱うため似たラベルや同じラベルを十分に活かせない，データセットの組み合わせに左右され性能が上がるとは限らないという問題がある．

それに対して Luo ら [6] は対象とするデータセットの各ラベルに対して，追加データセットをそれぞれ一つのラベルのみが含まれるように処理して加えることで，マルチタスク学習に対する性能向上を達成した．しかし，Luo らの手法ではモデルが単純である代わりに追加するデータセットの選定と正解ラベルに対する人手による編集が必要になる．

本研究では学習データの増加と対象にできるラベルの数の増加のために，ラベル付けの基準の違いを考慮した既存のデータセットの有効利用を目的とする．そのためにマルチタスク学習中に異なるデータセットで付与されたラベル間の関係を弱い制約として加える手法を提案する．具体的には，与える条件に応じた生成が可能な CVAE (Conditional Variational Autoencoder) の構造に注目し，その損失関数を固有表現抽出モデルに組み込む．

本稿の貢献を以下に示す．

- CVAE の損失関数によって，表現ベクトルに異なるデータセットのラベル間の関係を表す弱い制約をかける固有表現抽出モデルの提案．
- ラベル付けの基準の違いを考慮した複数データセットからの学習の有効性の確認．
- 複数データセットに対して人手でラベルを編集を加えず特定のデータセットに合わせた学習の実現．

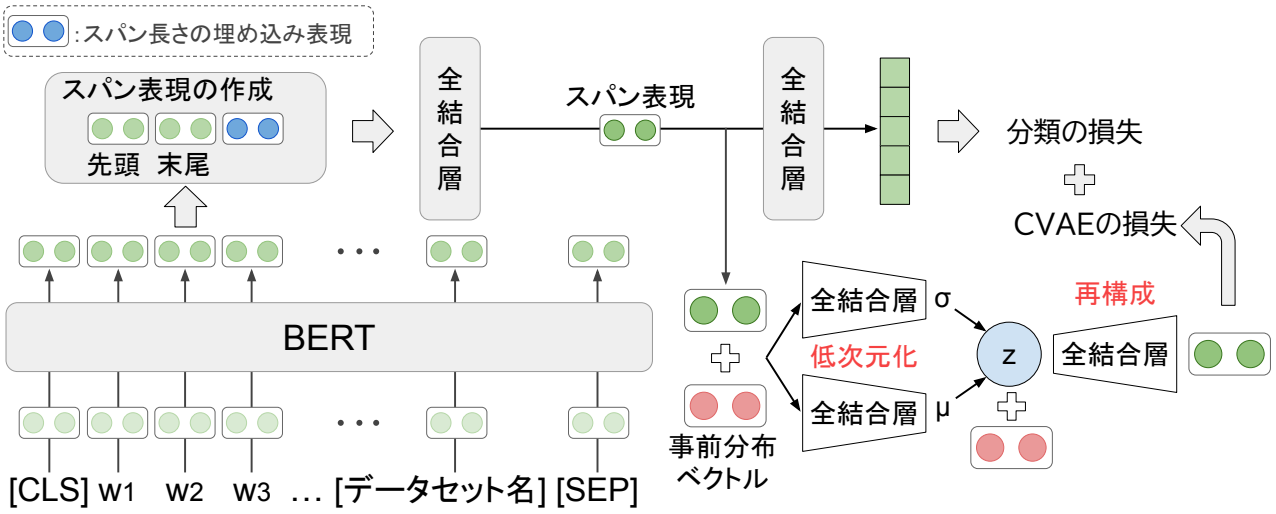

図 1 提案手法の概要図

2 関連研究

2.1 スパンに基づく固有表現抽出

文中の一定範囲の領域であるスパンを対象としたスパンに基づく固有表現抽出モデルが近年注目を浴びている [2, 7]．Zhong ら [2] のモデルでは，文のエンコードの部分で BERT (Bidirectional Encoder Representations from Transformers) [1] における注意機構による文脈を含んだ表現を用いて，対象スパンの先頭・末尾の単語それぞれの埋め込み，長さの埋め込みを連結し，スパンを表現している．Zhong らのモデルはこのスパンの表現を用いて全結合層と Softmax で分類を行う単純なモデルである．このように単純でありながら，end-to-end で固有表現抽出と関係抽出を解くタスクで最高性能を示し，固有表現抽出においても高い性能を示した．

2.2 VAE を用いた固有表現抽出

Nguyen ら [8] はスパンベースの固有表現抽出モデルで，確率分布に圧縮してから再構成する VAE を用いたスパン再構成と同義語生成を組み込んだモデルを提案した．スパンの再構成によってスパン表現にインスタンスの情報が保持され抽出性能が向上することを示した．

2.3 データセットを複数用いた固有表現抽出

Luo ら [6] は Disease, Chemical, Gene, Species, Variant, Cell line の 6 種類の固有表現を対象とし生物医学文書の複数のデータセットを用いた学習を行った．評価には 6 種類すべてを含む BioRED [9] を用いた．追加のデータセットは単一のラベルのみを含むようにラベルの統合や削除，ラベル付けの基準を合わせるためにスパン幅や付いているラベルに対する編集を人手で行った．編集を加えたデータセットと複数のデータセットの負例の違いを考慮するタグ付け方式の提案によって既存のマルチタスク学習による分類性能を超える性能を示した．

3 提案手法

本研究では複数のデータセットにおけるデータの衝突を解消し，既存のラベル付きデータセットを有効に利用することを目的に，各ラベルに対する事前分布を条件として加える CVAE (Conditional VAE) を固有表現抽出モデルに組み込む手法を提案する．提案モデルの全体像を図 1 に示す．スパンベースの固有表現抽出モデルに CVAE を組み込み，与えるラベルの分布によってデータセットの異なるラベルの共有や非共有の関係を表すことを目指す．

まず，固有表現抽出については，Zhong ら [2] のモデルと同様に予測対象となる文を事前学習済みの BERT によってエンコードし，その表現から単語を組み合わせたスパンの表現を作成する．BERT の出力の表現ベクトルをスパンの先頭と末尾で結合し，そこにスパンの長さの埋め込み表現を結合したものをスパンの表現とする．$\boldsymbol{x}_1$ から $\boldsymbol{x}_n$ のスパンの表現を式 1 に示す．

$$\boldsymbol{h}_{span} = \mathrm{Linear}(\mathrm{Concat}(\boldsymbol{x}_1, \boldsymbol{x}_n, \Phi(n))) \quad (1)$$

ここで，Linear(·) は全結合層，Concat(·) はベクトルの結合，$\Phi(n)$ はスパンの長さ n に対する埋め込み

表 1 使用データセット概要（サイズに full が付いているものはフルテキスト，付いていないものは abstract のみ）

データセット	サイズ	対象ラベル
BioRED [9]	600	Disease, Chemical, Gene, Species, Variant, Cell line
BC5CDR [3]	1500	Disease, Chemical
BioID [10]	570 full	CellLine
GNormPlus [11]	694	FamilyName, Gene, DomainMotif
Linnaeus [12]	100 full	Species
NCBIdisease [4]	793	DiseaseClass, SpecificDisease, CompositeMention, Modifier
NLMchem [13]	150 full	Chemical, NonStandardRef, OTHER
NLMgene [14]	550	Gene, FamilyName, Cell, DomainMotif, ChromosomeLocation
SPECIES800 [15]	800	Species
tmVar3 [16]	500	Gene, Species, Disease, DNAMutation, ProteinMutation, OtherMutation, CellLine, AcidChange, SNP, DNAAllele, ProteinAllele

表現を表す．また，データセットが異なる入力をモデルが認識するように入力文にデータセット名を特殊トークンとして加える．作成したスパン表現を 2 層の全結合層と Softmax で分類を行う．このとき複数のデータセットを用いる場合は，予測しているインスタンスのデータセットが対象としているラベルのみを考慮して，分類するように，対応する次元の出力以外に対してマスクをかけた上で分類を行う．損失としては，交差エントロピー損失 (L_{CE}) を用いる．

また，CVAE については，スパン表現に再構成時の条件となる事前分布ベクトルを加えて次元圧縮を行う．事前分布ベクトルは正解ラベルを基にデータセットが異なるラベルの共有と非共有を表すようなベクトルを用いる．各スパン表現に事前分布ベクトルを繋げて二種類の全結合層に入力し平均 (μ) と分散 (σ) を予測する．この分布からサンプリングした z に事前分布ベクトルを繋げたベクトルを全結合層に入力し次元を増やすことで再構成する．CVAE の損失としては，元のスパン表現と再構成したスパン表現の平均二乗誤差と予測した分布をガウス分布に近づける KL ダイバージェンスを用いる．

$$L_{CVAE} = \log p(\boldsymbol{h}_{span}|z) + KL[q(z|\boldsymbol{h}_{span})||p(z)]$$

最後に，この二つの損失の重み付け和を最終的な損失として，学習を行う．

$$L = \alpha L_{CE} + L_{CVAE}$$

推論時は正解ラベルを必要とする CVAE は使用せず，学習した分類器でスパンの分類を行い，固有表現を抽出する．

4 実験設定

4.1 データセット

複数データセットでの学習における提案手法の有効性を検証するために，Luo ら [6] と同様に生物医学文書を対象としてラベル付けされている 10 個のデータセットを用いて評価を行った．これらのデータセットの文書数とラベル付けしている固有表現を表 1 に示す．学習時の設定も Luo らに従い，BioRED の開発データを用いて学習中の評価を行い，それ以外のデータセットの開発データは学習データとして使用した．Luo らは追加するデータセットに対して BioRED の基準に合うようにスパン幅や対象とするラベルに編集を加えて使用しているが，本稿の実験では編集前のデータセットを用いた．

4.2 事前分布ベクトル

提案手法で CVAE に条件として与える事前分布ベクトルは 2 つの one-hot ベクトルを結合して用いた．1 つ目は表 1 に記した全ラベルとそれぞれの負例ラベルを合わせた 47 次元の中でそのスパン表現に対応する正解ラベルの次元に 1 が入った one-hot ベクトルである．2 つ目は BioRED のラベルを基準に，追加データセットの共有するラベルを表すような 6 次元の one-hot ベクトルである．具体的には，Luo らが追加データセットを使用する上でそれぞれのデータセットのラベルを BioRED の 6 種類のラベルに集約して使用しているため（NLMgene の Gene, FamilyName ラベルを BioRED の Gene と同様として扱っている）その対応する次元に 1 が入っている．

表 2 BioRED テストデータによる性能評価（F 値 [%]）

モデル（データセット）	全ラベル	Disease	Chemical	Gene	Variant	Species	Cell line
baseline (BioRED)	89.86	84.79	92.26	91.85	90.28	**98.11**	**97.09**
baseline (ALL)	93.43	91.64	92.39	95.70	95.12	97.14	77.11
Luo ら (ALL) [6][1)]	91.26	88.07	90.98	92.40	88.51	97.50	90.53
提案手法 (BioRED)	90.00	84.94	93.11	92.61	87.75	97.75	83.67
提案手法 (ALL)	94.18	**93.48**	92.29	96.39	95.18	97.69	76.54
+label token（ALL）	**94.51**	92.85	**93.25**	**96.71**	**95.82**	97.47	81.32

この対応を付録 A に示す．

4.3 実験に使用したモデル

今回の実験ではすべてのモデルのエンコーダに生物医学文書によって事前学習された PubMedBERT [17] を使用した．

また，比較を行ったモデルの概要を以下に示す．

- **baseline** Zhong らのモデル．複数データセットでの学習では BERT を共有し最終層をデータセットごとに持つマルチタスク学習を行った．
- **提案手法** CVAE による損失関数を加えたモデル．入力文にそれぞれデータセット名の特殊トークンを加えた．CVAE の事前分布ベクトルについては 4.2 節に記載したものを用いた．
- **+label token** 提案手法から入力文に加える特殊トークンをデータセット名ではなくそのデータセットで対象としているラベル名（その個数分加える）に変更したモデル．

5 結果と考察

提案手法と baseline モデル，Luo らのモデル [6] の比較を表 2 に示す．結果から BioRED の学習データのみで学習した場合と比べて表 1 に示した 10 個のデータセットすべての学習データを用いて学習した場合に baseline，提案手法でともに F 値が向上している．さらに baseline での上がり幅に対して提案手法での上がり幅が大きいことから，提案手法が単純なマルチタスク学習である baseline よりも複数データセットを追加した学習に有効であるとわかる．

また，表 3 は 10 個のデータセットで学習したモデルに対して BioRED 以外のデータセットのテストデータで評価を行った結果である．追加したデータセットの評価では baseline が高い性能を示す傾向にある．これは提案手法における事前分布ベクトルが BioRED を基準としており，その制約によって追加したデータセットでの評価に合わなくなっている可能性がある．

表 3 追加データセットでのテスト評価（F 値 [%]）

データセット	baseline	提案手法	+label token
BC5CDR	90.20	**91.16**	90.91
BioID	**92.20**	91.89	91.02
GNormPlus	78.41	78.15	**78.55**
Linnaeus	**90.52**	84.80	84.15
NCBIdisease	**83.61**	80.13	81.21
NLMchem [2)]	-	-	-
NLMgene	85.22	85.57	**85.94**
SPECIES800	**79.17**	76.02	76.02
tmVar3	89.94	**91.22**	89.38

6 おわりに

本研究では学習データの増加と対象にできるラベルの数の増加のために異なるデータセットのラベル付けの基準の違いを考慮したモデルの実現を目指し，CVAE を用いた損失関数を固有表現抽出モデルに組み込む手法を提案した．実験では追加するデータセットに基準とするデータセットに合わせて人手で編集を行った既存手法より提案手法が高い性能を示した．

一方で提案手法で弱い制約として与えた事前分布ベクトルによって基準としている BioRED のアノテーションによった学習をしており追加したデータセットでの評価では性能が下がる傾向にあった．今後の課題としては事前分布ベクトルを学習可能なパラメータとすることで追加するデータセットに対しても性能向上が可能なモデルを目指す．

謝辞

本研究は JSPS 科研費 JP20K11962 の助成を受けたものです．

1） 論文中の値
2） テストデータ中に BERT の入力長を超える文があったため今回は評価していない

参考文献

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[2] Zhong et al. A frustratingly easy approach for entity and relation extraction. In **NAACL-HLT**, pp. 50–61, Online, June 2021. Association for Computational Linguistics.

[3] Li et al. BioCreative V CDR task corpus: a resource for chemical disease relation extraction. **Database**, Vol. 2016, , 05 2016. baw068.

[4] Doğan et al. NCBI disease corpus: A resource for disease name recognition and concept normalization. **JBI**, 2014.

[5] Nicholas E. Rodriguez, Mai Nguyen, and Bridget T. McInnes. Effects of data and entity ablation on multitask learning models for biomedical entity recognition. **Journal of Biomedical Informatics**, Vol. 130, p. 104062, 2022.

[6] Ling Luo, Chih-Hsuan Wei, Po-Ting Lai, Robert Leaman, Qingyu Chen, and Zhiyong Lu. AIONER: all-in-one scheme-based biomedical named entity recognition using deep learning. **Bioinformatics**, Vol. 39, No. 5, p. btad310, 05 2023.

[7] Mohammad Golam Sohrab and Makoto Miwa. Deep exhaustive model for nested named entity recognition. In **Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing**, pp. 2843–2849, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[8] Nhung T. H. Nguyen, Makoto Miwa, and Sophia Ananiadou. Span-based named entity recognition by generating and compressing information. In **Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics**, pp. 1984–1996, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics.

[9] Ling Luo, Po-Ting Lai, Chih-Hsuan Wei, Cecilia N Arighi, and Zhiyong Lu. Biored: a rich biomedical relation extraction dataset. **Briefings in Bioinformatics**, Vol. 23, No. 5, p. bbac282, 2022.

[10] Cecilia Arighi, Lynette Hirschman, Thomas Lemberger, et al. Bio-id track overview. In **BioCreative VI Workshop**, pp. 28–31, Bethesda, MD, USA, 2017. BioCreative.

[11] Chih-Hsuan Wei, Hung-Yu Kao, Zhiyong Lu, et al. Gnormplus: an integrative approach for tagging genes, gene families, and protein domains. **BioMed research international**, Vol. 2015, , 2015.

[12] Martin Gerner, Goran Nenadic, and Casey M Bergman. Linnaeus: a species name identification system for biomedical literature. **BMC bioinformatics**, Vol. 11, No. 1, pp. 1–17, 2010.

[13] Rezarta Islamaj, Robert Leaman, Sun Kim, Dongseop Kwon, Chih-Hsuan Wei, Donald C Comeau, Yifan Peng, David Cissel, Cathleen Coss, Carol Fisher, et al. Nlm-chem, a new resource for chemical entity recognition in pubmed full text literature. **Scientific data**, Vol. 8, No. 1, p. 91, 2021.

[14] Rezarta Islamaj, Chih-Hsuan Wei, David Cissel, Nicholas Miliaras, Olga Printseva, Oleg Rodionov, Keiko Sekiya, Janice Ward, and Zhiyong Lu. Nlm-gene, a richly annotated gold standard dataset for gene entities that addresses ambiguity and multi-species gene recognition. **Journal of biomedical informatics**, Vol. 118, p. 103779, 2021.

[15] Evangelos Pafilis, Sune P Frankild, Lucia Fanini, Sarah Faulwetter, Christina Pavloudi, Aikaterini Vasileiadou, Christos Arvanitidis, and Lars Juhl Jensen. The species and organisms resources for fast and accurate identification of taxonomic names in text. **PloS one**, Vol. 8, No. 6, p. e65390, 2013.

[16] Chih-Hsuan Wei, Alexis Allot, Kevin Riehle, Aleksandar Milosavljevic, and Zhiyong Lu. tmvar 3.0: an improved variant concept recognition and normalization tool. **Bioinformatics**, Vol. 38, No. 18, pp. 4449–4451, 2022.

[17] Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. **ACM Trans. Comput. Healthcare**, Vol. 3, No. 1, oct 2021.

[18] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, **Advances in Neural Information Processing Systems**, Vol. 32. Curran Associates, Inc., 2019.

[19] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations**, pp. 38–45, Online, October 2020. Association for Computational Linguistics.

A 追加データセットのラベルと BioRED のラベルの対応

Luo らが複数のデータセットに対して行った編集を基に各データセットのラベルと BioRED のラベルの対応を表 4 に示す.

データセット名	BioRED のラベル	対応ラベル
BC5CDR [3]	Disease	Disease
BioID [10]	CellLine	CellLine
GNormPlus [11]	Gene	Gene, FamilyName
Linnaeus [12]	Species	Species
NCBIdisease [4]	Disease	DiseaseClass, SpecificDisease, CompositeMention, Modifier
NLMchem [13]	Chemical	Chemical
NLMgene [14]	Gene	Gene, FamilyName
SPECIES800 [15]	Species	Species
tmVar3 [16]	Variant	DNAMutation, ProteinMutation, OtherMutation, AcidChange, SNP, DNAAllele, ProteinAllele

表 4 追加データセットのラベルと BioRED のラベルの対応

B ハイパーパラメータ

本研究における実験で用いたハイパーパラメータを表 5 に示す．エンコーダとそれ以外の全結合ニューラルネットワークなどのパラメータで異なる学習率を用いている.

表 5 各モデルのハイパーパラメータ

	提案手法	baseline
学習率（エンコーダ）	2e-6	6e-6
学習率（その他）	8e-5	9e-5
α	100	-
エポック数	100	100
batch size	48	48

C 実験環境

実験のためのプログラミング言語は Python 3.7.11，機械学習ライブラリとして PyTorch [18] のバージョン 1.11.0，事前学習モデルを使用するために Transformers [19] のバージョン 3.0.2 を用いた．また，計算機では CPU に Intel(R) Xeon(R) CPU E5-2698 v4 及び Intel(R) Xeon(R) W-3225，GPU に NVIDIA Tesla V100 DGXS 32GB 及び NVIDIA RTX A6000 を用いた.