# Enhancing NER by Harnessing Multiple Datasets with Conditional Variational Autoencoders

**Taku Oi** and **Makoto Miwa**
Toyota Technological Institute
**Correspondence:** makoto-miwa@toyota-ti.ac.jp

## Abstract

We propose a novel method to integrate a Conditional Variational Autoencoder (CVAE) into a span-based Named Entity Recognition (NER) model. This approach models shared and unshared information among labels in multiple datasets, thereby easing training on these datasets. Experimental results using multiple biomedical datasets demonstrate the effectiveness of the proposed method, showing improved performance on the BioRED dataset. Our source code for this implementation is publicly available at GitHub[1].

## 1 Introduction

Named Entity Recognition (NER) is a fundamental Natural Language Processing (NLP) task, serving as a crucial first step in information extraction. Although Large Language Models (LLMs) with zero- or few-shot learning have been widely investigated, supervised learning or full fine-tuning remains essential for achieving high performance (Zhong and Chen, 2021; Luo et al., 2023; Wang et al., 2023; Munnangi et al., 2024; Zhou et al., 2024). The performance of such NER models depends on the quantity of labeled data, which is often costly to create manually.

One approach to increase labeled training data is to combine existing labeled datasets (Luo et al., 2022; Islamaj et al., 2021a). However, even datasets targeting the same types of named entities have different type definitions and annotation criteria, making it challenging to merge them into a unified training dataset. For instance, in the biomedical NER datasets NLMChem (Islamaj et al., 2021a) and BioRED (Luo et al., 2022), which both include a *Chemical* entity type, the term "hematoxylin" is labeled as *Chemical* in NLMChem, but not in BioRED. This discrepancy arises because BioRED's annotation guidelines explicitly exclude staining reagents from the *Chemical* label.

A common strategy for utilizing multiple datasets is multi-task learning (MTL) (Wang et al., 2018; Zuo and Zhang, 2020; Rodriguez et al., 2022). MTL shares a base model among tasks while maintaining a separate classification layer for each dataset, which allows for learning without explicitly addressing dataset-specific differences. However, MTL cannot account for relationships among labels in different datasets. Luo et al. (2023) tackled this issue by manually editing additional datasets to align with the target dataset, which improved performance, but the manual editing process is costly.

Span-based approaches are widely used for NER (Sohrab and Miwa, 2018; Ouchi et al., 2020; Nguyen et al., 2023) due to their simplicity and effectiveness. Span representations are critical for these methods, as they directly influence the ability to accurately identify and classify named entities. Nguyen et al. (2023) explored integrating Variational Autoencoders (VAE) to enhance span representations. Inspired by this, we employ a span-based approach with Conditional VAE (CVAE) to model spans using label-specific conditions that are both shared and unshared across multiple datasets.

This study proposes a method that models the relationships between labels in different datasets as label-specific conditions using CVAE, thereby incorporating these conditions to alleviate differences across datasets. The contributions of this paper are as follows:

- We propose a span-based NER method to integrate multiple existing datasets, despite their differing type definitions and annotation criteria.
- We propose integrating CVAE to incorporate and model the relationships between labels from multiple datasets.

[1] https://github.com/tti-kde/CVAE-NER

*Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1107–1117
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

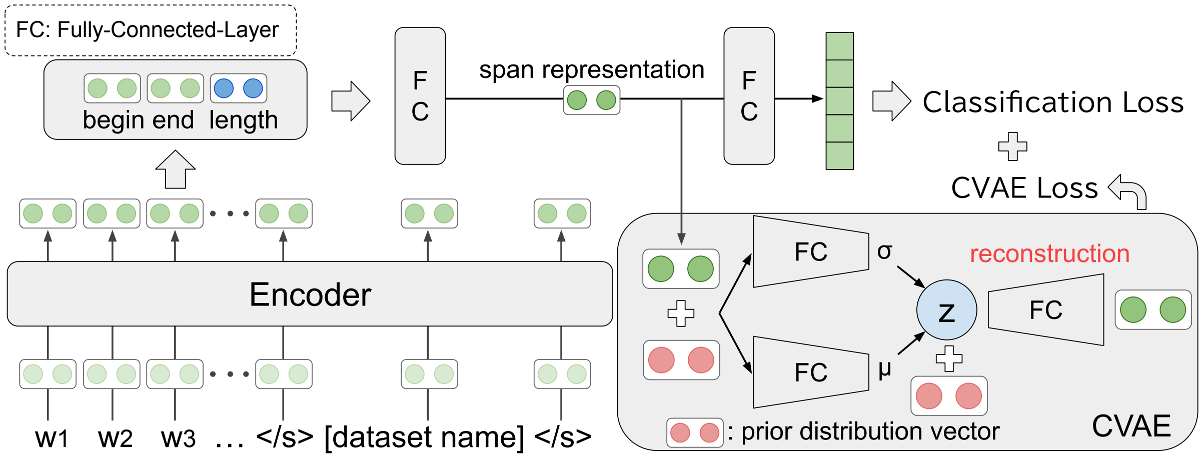

Figure 1: Overview of the proposed method, illustrating the integration of CVAE into a span-based NER model. Within the CVAE component, the prior distribution vector provides information regarding shared and unshared labels to the model.

- We demonstrate the effectiveness of our method through experiments on multiple biomedical datasets, achieving improved performance on the BioRED (Luo et al., 2023) dataset with either of the two encoders used.

## 2 Related Work

### 2.1 Span-based NER model

Span-based models, which directly classify specific text spans within a document, have recently gained attention (Zhong and Chen, 2021; Sohrab and Miwa, 2018) due to their simplicity and ability to model entity spans directly, unlike sequence labeling models that process tokens to represent entities. Zhong and Chen (2021) represent each target span by concatenating the embeddings of the first and last tokens from the encoder with the span length embedding. The span representations are then passed through a fully-connected layer, followed by a Softmax layer for classification. The model demonstrated high performance in NER and achieved state-of-the-art performance in end-to-end entity and relation extraction.

Span representations are crucial in span-based NER models. Ouchi et al. (2020) enhanced interpretability by learning the similarity between span representations and assessing this similarity during prediction. Nguyen et al. (2023) proposed a span-based NER model that incorporates span reconstruction and synonym generation using VAE. Span reconstruction ensures that instance-specific information is retained in the span representations, thereby improving extraction performance.

### 2.2 NER using multiple datasets

MTL (Rodriguez et al., 2022; Zuo and Zhang, 2020) employs a shared encoder and a separate classification layer for each dataset during training. Recent approaches leverage the capabilities of LLMs for MTL (e.g., Zhou et al. (2024)). However, these approaches do not explicitly account for shared label information across diverse datasets.

Luo et al. (2023) trained NER on multiple datasets targeting six types of entities: *Disease*, *Chemical*, *Gene*, *Species*, *Variant*, and *Cell line*, to improve performance on the BioRED dataset. Additional datasets underwent manual editing to ensure each contained only a single label type through label integration, deletion, and adjustment of label spans to align annotation criteria. The method outperformed existing MTL methods but required costly manual annotations. Our approach uniquely addresses the label differences through direct modeling with CVAE, which has not been explored in existing methods.

## 3 Method

This study proposes integrating CVAE into an NER model to incorporate a prior distribution that provides the model with information about shared and unshared labels among the datasets as a label-specific condition. We anticipate that this condition will capture relationships between labels in multiple datasets, enabling the effective use of existing labeled datasets. An overview of the proposed model is illustrated in Figure 1.

We follow Zhong and Chen (2021) to model the representation of a span $x_1, \ldots, x_n$. The representation is constructed by concatenating the representations of the beginning token and end token from the encoder, along with the embeddings of the span length, as shown in Equation (1):

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

where $\mathrm{Linear}(\cdot)$ is a fully-connected layer and $\mathrm{Concat}(\cdot)$ is vector concatenation, $\boldsymbol{x}_i$ is the output representation of a token $x_i$ from the encoder, and $\Phi(n)$ is the embedding of the span length $n$. The dataset name is added as a special token to the input text to ensure the model recognizes inputs from different datasets. The span representation is classified using two fully-connected layers, followed by a Softmax layer. Cross-entropy loss ($L_{CE}$) is employed as the loss function.

In CVAE, the span representation is concatenated with a prior distribution vector, which functions as the conditional parameter during the reconstruction phase. The prior distribution vector is derived from the correct label and encodes the shared and unshared label information across datasets[2]. After concatenating with the prior distribution vector, each span representation is processed through two fully-connected layers to estimate the mean ($\mu$) and variance ($\sigma$) parameters of the latent distribution. A vector $z$, sampled from the estimated distribution, is further concatenated with the prior distribution vector and then processed through two fully-connected layers for reconstruction. The original loss function of CVAE $L_{CVAE}$ (Kingma et al., 2014) is employed, which consists of the reconstruction error and the Kullback-Leibler (KL) divergence, as follows:

$$\begin{aligned} L_{CVAE} = &- \log p(\boldsymbol{h}_{span}|z) \\ &+ KL[q(z|\boldsymbol{h}_{span})||p(z)]. \end{aligned}$$

The prior distribution vector is prepared as trainable parameters. The vector is obtained by concatenating two one-hot vectors: one representing the labels across all datasets, with a "1" at the position corresponding to each label, and the other representing the corresponding (shared) label in the target dataset, including a negative label, with a "1" for each shared label.

As a simple example, consider the case where a target corpus tags Chemical$_T$ and Protein$_T$, and a source corpus tags Disease$_S$ and Protein$_S$, with Protein being a shared entity type across datasets. The first component of the prior distribution vector represents the labels across all datasets. For this example, this component would form a 6-dimensional space including negative labels: [O$_T$, Chemical$_T$, Protein$_T$, O$_S$, Disease$_S$, Protein$_S$]. The second component represents the corresponding (shared) label in the target dataset including a negative label. For this example, this component would form a 3-dimensional space: [O, Chemical$_T$, Protein$_T$]. Consequently, the entire prior distribution vector is expressed within a combined 9-dimensional space. For instance, for an entity labeled Protein$_S$ from the source corpus, its prior distribution vector would be [0, 0, 0, 0, 0, 1, 0, 0, 1]. Here, the "1"s are placed at the dimension corresponding to Protein$_S$ in the corpus-specific space and at the dimension corresponding to the shared Protein$_T$ label in the shared target space.

Finally, the weighted sum of the two losses is used as the entire loss $L$ for training:

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

It is important to note that CVAE is not used during inference since it requires correct labels. Instead, the trained classifier alone is used to recognize named entities. Our method thus requires no additional computation during inference, making it suitable for practical applications.

## 4 Experimental Setup

We describe the datasets and models used in the experiments. Detailed settings and hyperparameters are provided in Appendix A.

### 4.1 Datasets

We employed 10 labeled biomedical datasets following Luo et al. (2023). The number of documents and the named entities labeled in these datasets are shown in Table 6 of Appendix B. We used the development set of BioRED for evaluation during training, while the development sets of other datasets were used as training data. Unlike Luo et al. (2023), which edited the spans and target labels of entities in additional datasets to follow the BioRED standards, we utilized the original datasets without any such edits.

### 4.2 Prior distribution vector

As explained in §3, the prior distribution vector provided as a condition to CVAE is constructed

[2] The actual settings for the prior distribution vector will be shown in §4.2.

| Model | Encoder | All | *Disease* | *Chemical* | *Gene* | *Variant* | *Species* | *Cell line* |
|---|---|---|---|---|---|---|---|---|
| Luo et al. (2023)[3] | BERT | 91.26 | 88.07 | 90.98 | 92.40 | 88.51 | 97.50 | 90.53 |
| Single | | 87.86 | 82.33 | 89.42 | 89.63 | 81.38 | 96.53 | 80.90 |
| Multi | | 93.22 | 90.66 | **93.41** | 95.18 | 86.84 | 97.96 | 79.55 |
| Single + CVAE | | 87.28 | 81.93 | 89.76 | 89.64 | 77.80 | 94.74 | 75.00 |
| Multi + CVAE | | 94.19 | 91.88 | 92.98 | **96.47** | 92.80 | 97.48 | 81.82 |
| Single | T5 | 89.89 | 86.52 | 87.99 | 91.19 | 90.56 | 97.73 | 86.67 |
| Multi | | 93.99 | 92.82 | 92.16 | 95.89 | 90.07 | 98.00 | 82.76 |
| Single + CVAE | | 89.87 | 85.81 | 89.45 | 90.72 | 89.18 | 97.62 | **91.84** |
| Multi + CVAE | | **94.36** | **92.84** | 92.89 | 95.77 | **94.29** | **98.23** | 78.57 |

Table 1: F1 scores [%] on the BioRED test set, serving as the primary metric. The ‘Encoder’ column specifies T5-3B for ‘T5’ and PubMedBERT for ‘BERT’. The highest F1 score for each entity type is highlighted in bold.

by concatenating two one-hot vectors. The first vector is a one-hot vector with a dimension of 47, representing all labels listed in Table 6, including a negative label for each dataset. The second vector is a seven-dimensional one-hot vector (six types + negative) representing the shared labels based on the BioRED labels. We use the mapping of Luo et al. (2023), which mapped the labels from the additional datasets to one of the BioRED labels, including the negative label (e.g., treating the Gene and FamilyName labels from NLMGene as equivalent to the Gene label in BioRED). The details of this correspondence are shown in Appendix C.

### 4.3 Comparison settings

We employed the encoder part of T5 (Text-To-Text Transfer Transformer)-3B (Raffel et al., 2020) as the encoder of our model. We also employed PubMedBERT (Gu et al., 2021) as the encoder for a fair comparison with Luo et al. (2023). We employ the F1 score as our primary metric. We compared the following settings:

- **Single**
  The span-based NER model trained on a single target dataset.

- **Multi**
  The span-based NER model trained on multiple datasets in a pure MTL setting, where separate classification layers are prepared for each dataset while the base encoder layer is shared.

- **Multi+CVAE**
  The span-based NER model with CVAE that incorporates the prior distribution vector described in §4.2.

- **Multi+CVAE (Fixed)**
  The span-based NER model with CVAE that fixes the prior distribution vector after initialization.

- **Multi+CVAE (Unshared)**
  The span-based NER model with CVAE that initializes the prior distribution vector using only the first one-hot vector (labels in each dataset) in §4.2.

- **Multi+CVAE (Shared label)**
  The span-based NER model with CVAE that initializes the prior distribution vector using only the second one-hot vector (shared label information) in §4.2.

- **Multi+CVAE (Random)**
  The span-based NER model with CVAE that initializes the prior distribution vector randomly.

- **Multi (Shared label)**
  The span-based NER model trained on multiple datasets without CVAE, where shared labels across datasets are assigned to the same output layer.

## 5 Results

The overall performance is summarized in Table 1. Compared with the state-of-the-art (SOTA) model by Luo et al. (2023), our model demonstrates enhanced performance across all label types when using multiple datasets, with the exception of *Cell line*. This *Cell line* result can be attributed to the limited number of instances in the test set, as shown in Table 5 of Appendix B, which makes

[3] The score is taken from the original paper.

| Model | All |
|---|---|
| Multi+CVAE | **94.36** |
| Multi+CVAE (Fixed) | 93.99 |
| Multi+CVAE (Unshared) | 94.12 |
| Multi+CVAE (Shared label) | 94.20 |
| Multi+CVAE (Random) | 94.07 |
| Multi (Shared label) | 92.52 |

Table 2: Ablation study using the T5 encoder.

the evaluation sensitive to variations of individual cases. When trained on a single dataset with CVAE (Single+CVAE), the performance slightly degraded when the BERT encoder was employed. This may be because the representations of different labels are not sufficiently distant due to the interaction between labels caused by the CVAE loss. As for the performance of our models with multiple datasets, both the Multi and Multi+CVAE models achieved better F1 scores regardless of the encoder, compared to the corresponding Single model trained solely on BioRED. Furthermore, the CVAE model improved the F1 score for both encoders, suggesting its potential effectiveness in leveraging CVAE during training. This result indicates that the proposed model could capture some of the shared and unshared information among the datasets. This is further supported by the visualizations in Figure 4 of Appendix D, where instances with shared labels from different datasets have closer embeddings.

The ablation study with modified conditions for the prior distribution vector is shown in Table 2. All settings showed inferior performance compared to the proposed model, demonstrating the effectiveness of using the shared labels and tuning the prior distribution vector during training.

## 6 Conclusion

This study aimed to alleviate the differences in labeling criteria among different datasets to increase the amount of training data. To achieve this, we proposed a method that incorporates a CVAE-based loss function into a span-based NER model, which considers the differences in labeling criteria without the need for manual edits. Our experimental results showed that our CVAE-based approach can leverage multiple datasets while accommodating their labeling discrepancies for either of the two encoders used.

For future work, we will investigate a fully automated method by making the prior distribution vector independent from a specific dataset and eliminating the need for manual definition of the shared vectors, aiming to develop a model that can improve performance on additional datasets as well and be applied to other datasets. Additionally, to further demonstrate broader generalizability, we will evaluate CVAE with other backbone NER models.

## Acknowledgments

The authors gratefully acknowledge the assistance provided by AI tools, specifically GitHub Copilot and Claude 3.5 Sonnet, in supporting our programming and writing processes.

## Limitations

Our model is based on the approach proposed by Zhong and Chen (2021), where the maximum span length is set to 10. As a result, our model cannot extract named entities that exceed this length constraint. Additionally, as our model primarily utilizes an encoder-based architecture, it may not be straightforward to adapt it to LLMs. Although our method does not require manual annotation, it still requires manual effort to define the prior distribution vector. Due to this limitation, we have evaluated our method only on the BioRED dataset, and the application to other datasets has not been investigated.

## Ethical Considerations

This paper utilizes the pre-trained LLM T5 as the base model, which may contain inherent biases due to its pre-training processes. However, we employ only the encoder part of T5 and fine-tune the model for NER, so such potential biases may be mitigated. Furthermore, we use public datasets only for scientific research purposes.

## References

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

Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. 2014. Ncbi disease corpus: A resource for disease name recognition and concept normalization. *Journal of Biomedical Informatics*, 47:1–10.

Martin Gerner, Goran Nenadic, and Casey M Bergman. 2010. Linnaeus: a species name identification system for biomedical literature. *BMC bioinformatics*, 11(1):1–17.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Healthcare*, 3(1).

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Rezarta Islamaj, Robert Leaman, Sun Kim, Dongseop Kwon, Chih-Hsuan Wei, Donald C Comeau, Yifan Peng, David Cissel, Cathleen Coss, Carol Fisher, et al. 2021a. Nlm-chem, a new resource for chemical entity recognition in pubmed full text literature. *Scientific data*, 8(1):91.

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

Durk P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. 2014. Semi-supervised learning with deep generative models. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Jiao Li, Yueping Sun, Robin J. Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J. Mattingly, Thomas C. Wiegers, and Zhiyong Lu. 2016. BioCreative V CDR task corpus: a resource for chemical disease relation extraction. *Database*, 2016:baw068.

Ling Luo, Po-Ting Lai, Chih-Hsuan Wei, Cecilia N Arighi, and Zhiyong Lu. 2022. Biored: a rich biomedical relation extraction dataset. *Briefings in Bioinformatics*, 23(5):bbac282.

Ling Luo, Chih-Hsuan Wei, Po-Ting Lai, Robert Leaman, Qingyu Chen, and Zhiyong Lu. 2023. AIONER: all-in-one scheme-based biomedical named entity recognition using deep learning. *Bioinformatics*, 39(5):btad310.

Monica Munnangi, Sergey Feldman, Byron Wallace, Silvio Amir, Tom Hope, and Aakanksha Naik. 2024. On-the-fly definition augmentation of LLMs for biomedical NER. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3833–3854, Mexico City, Mexico. Association for Computational Linguistics.

Nhung T. H. Nguyen, Makoto Miwa, and Sophia Ananiadou. 2023. 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*, pages 1984–1996, Dubrovnik, Croatia. Association for Computational Linguistics.

Hiroki Ouchi, Jun Suzuki, Sosuke Kobayashi, Sho Yokoi, Tatsuki Kuribayashi, Ryuto Konno, and Kentaro Inui. 2020. Instance-based learning of span representations: A case study through named entity recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6452–6459, Online. Association for Computational Linguistics.

Evangelos Pafilis, Sune P Frankild, Lucia Fanini, Sarah Faulwetter, Christina Pavloudi, Aikaterini Vasileiadou, Christos Arvanitidis, and Lars Juhl Jensen. 2013. The species and organisms resources for fast and accurate identification of taxonomic names in text. *PloS one*, 8(6):e65390.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

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

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

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605.

Xiao Wang, Weikang Zhou, Can Zu, Han Xia, Tianze Chen, Yuansen Zhang, Rui Zheng, Junjie Ye, Qi Zhang, Tao Gui, Jihua Kang, Jingsheng Yang, Siyuan Li, and Chunsai Du. 2023. Instructuie: Multi-task instruction tuning for unified information extraction. *Preprint*, arXiv:2304.08085.

Xuan Wang, Yu Zhang, Xiang Ren, Yuhao Zhang, Marinka Zitnik, Jingbo Shang, Curtis Langlotz, and Jiawei Han. 2018. Cross-type biomedical named entity recognition with deep multi-task learning. *Bioinformatics*, 35(10):1745–1752.

Chih-Hsuan Wei, Alexis Allot, Kevin Riehle, Aleksandar Milosavljevic, and Zhiyong Lu. 2022. tmvar 3.0: an improved variant concept recognition and normalization tool. *Bioinformatics*, 38(18):4449–4451.

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

Zexuan Zhong and Danqi Chen. 2021. A frustratingly easy approach for entity and relation extraction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 50–61, Online. Association for Computational Linguistics.

Wenxuan Zhou, Sheng Zhang, Yu Gu, Muhao Chen, and Hoifung Poon. 2024. UniversalNER: Targeted distillation from large language models for open named entity recognition. In *The Twelfth International Conference on Learning Representations*.

Mei Zuo and Yang Zhang. 2020. Dataset-aware multi-task learning approaches for biomedical named entity recognition. *Bioinformatics*, 36(15):4331–4338.

## A Detailed training settings

We listed the number of parameters for the encoder models used in the experiments in Table 3. In experiments using the T5 encoder, we employed Low-Rank Adaptation (LoRA) (Hu et al., 2022) and model parameter quantization during training. In addition, the hyperparameters used in the experiments are summarized in Table 4. In the experiments, we took 16 examples from each dataset. We then performed training with a mini-batch size of 64, effectively including examples from multiple datasets by weight accumulation and combining losses from four datasets before updating the model. The weight of the CVAE loss (i.e., $\alpha$) was selected based on the results from Figure 2, which were obtained from experiments using the T5-base model. To ensure optimal training, the model was evaluated on the development set after each epoch, and training was terminated if no improvement in validation performance was observed for five consecutive epochs. All experiments were conducted on a computing infrastructure equipped with four NVIDIA V100 GPUs. All experimental results reported in this paper are based on a single run with a fixed random seed.

## B Dataset statistics

Table 5 presents the statistics of entity counts in the train, development, and test sets of the BioRED dataset. The statistics of the datasets and the list of named entity labels are summarized in Table 6. Note that all datasets used in this study are in English.

## C Correspondence of labels

The correspondence of the labels of each dataset to the BioRED labels is summarized in Table 7. This correspondence is from Luo et al. (2023). For spans with no correspondence to the BioRED labels, the second one-hot vectors of their prior distribution vectors have no active values.

We visualize the prior distribution vectors after training using t-SNE in Figure 3. The visualization demonstrates that the labels in Table 7 remain clustered around their corresponding BioRED labels, and labels with no correspondence remain far from the BioRED labels. This suggests that the initialization information is mostly preserved throughout the training process.

| Model | #Params |
|---|---:|
| T5-3B | 876,802,048 |
| T5-base | 86,043,264 |
| PubMedBERT | 109,482,240 |

Table 3: Numbers of parameters (#Params) for the encoder models used in the experiments.

| Parameter | Value |
|---|---:|
| Learning rate | 7e-4 |
| $\alpha$ | 1e-4 |
| Mini-batch size | 64 |
| CVAE hidden dim | 150 |
| Max span length | 10 |
| Span length embedding dim | 150 |
| LoRA rank | 32 |

Table 4: Hyperparameter settings

| | Train | Dev | Test |
|---|---:|---:|---:|
| Document | 400 | 100 | 100 |
| All | 13,351 | 3,533 | 3,535 |
| *Disease* | 3,646 | 982 | 917 |
| *Chemical* | 2,853 | 822 | 754 |
| *Gene* | 4,430 | 1,087 | 1,180 |
| *Variant* | 890 | 250 | 241 |
| *Species* | 1,429 | 370 | 393 |
| *Cell line* | 103 | 22 | 50 |

Table 5: Detailed statistics of entity counts across train, development, and test sets of the BioRED dataset (Luo et al., 2022), categorized by the entity types.

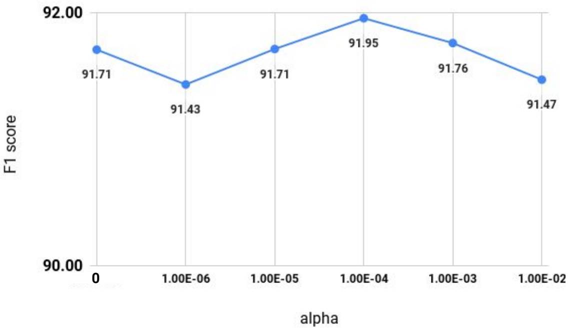

Figure 2: F1 scores [%] on the BioRED development set, plotted against the weight $\alpha$ applied to the CVAE loss, using the T5-base encoder; $\alpha = 0$ indicates the F1 score obtained without the CVAE loss.

## D Visualization of instance embeddings

Figure 4 visualizes the embeddings of instances in the development set of the BioRED dataset and the test sets of the nine additional datasets using

| Datasets | Size | Labels |
|---|---:|---|
| BioRED (Luo et al., 2022) | 600 | *Disease*, *Chemical*, *Gene*, *Variant*, *Species*, *Cell line* |
| BC5CDR (Li et al., 2016) | 1500 | *Disease*, *Chemical* |
| BioID (Arighi et al., 2017) | 570 full | *Cell line* |
| GNormPlus (Wei et al., 2015) | 694 | *FamilyName*, *Gene*, *DomainMotif* |
| Linnaeus (Gerner et al., 2010) | 100 full | *Species* |
| NCBIdisease (Doğan et al., 2014) | 793 | *DiseaseClass*, *SpecificDisease*, *CompositeMention*, *Modifier* |
| NLMChem (Islamaj et al., 2021a) | 150 full | *Chemical*, *NonStandardRef*, *OTHER* |
| NLMGene (Islamaj et al., 2021b) | 550 | *Gene*, *FamilyName*, *Cell*, *DomainMotif*, *ChromosomeLocation* |
| SPECIES800 (Pafilis et al., 2013) | 800 | *Species* |
| tmVar3 (Wei et al., 2022) | 500 | *Gene*, *Species*, *Disease*, *DNAMutation*, *ProteinMutation*, *OtherMutation*, *Cell line*, *AcidChange*, *SNP*, *DNAAllele*, *ProteinAllele* |

Table 6: Summary of the datasets used. Datasets with "full" in the size column contain full-text data, while the others contain only abstracts.

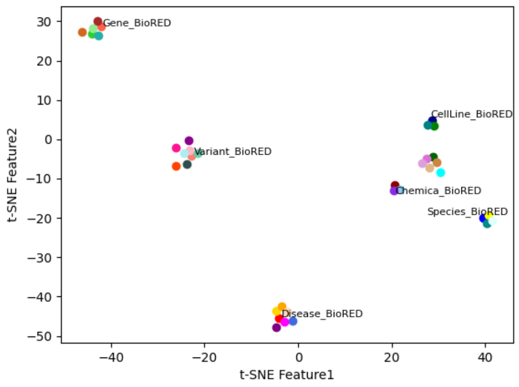

Figure 3: Visualization of prior distribution vectors after training by t-SNE.

t-SNE (van der Maaten and Hinton, 2008). The instances are taken from 300 sentences for each dataset. Compared to the two figures on the left side by the Multi model, the two on the right side using the proposed Multi+CVAE model show a tendency for *Disease* and *Species* instances from different datasets to have closer embeddings. This tendency is particularly pronounced for *Species*, which may contribute to improved performance for *Species* in BioRED.

## E Evaluation on the additional datasets

We evaluated both Multi and Multi+CVAE models on nine additional datasets, with results shown in Appendix E. Our proposed method improved performance over the Multi baseline on only two of the nine datasets. As discussed in the Conclusion, this limited improvement is attributed to using prior distribution vectors specifically designed for BioRED, highlighting an area that requires future investigation.

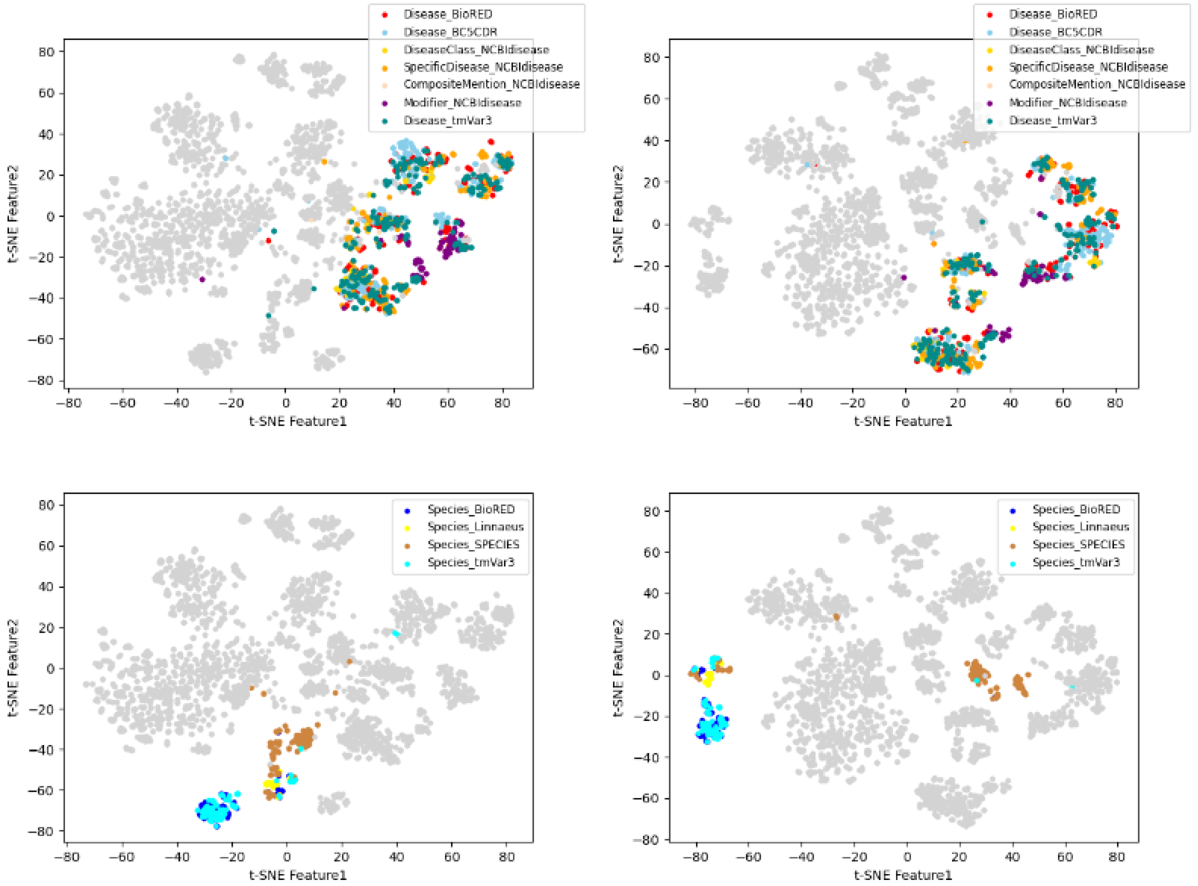

Figure 4: t-SNE Visualization of instance embeddings from the BioRED development set and additional datasets. Upper left: *Disease* Multi+CVAE, upper right: *Disease* Multi, lower left: *Species* Multi+CVAE, lower right: *Species* Multi

| Dataset | Label | Corresponding label in BioRED |
|---|---|---|
| BC5CDR (Li et al., 2016) | *Disease* | *Disease* |
| | *Chemical* | *Chemical* |
| BioID (Arighi et al., 2017) | *Cell line* | *Cell line* |
| GNormPlus (Wei et al., 2015) | *Gene* | *Gene* |
| | *FamilyName* | *Gene* |
| Linnaeus (Gerner et al., 2010) | *Species* | *Species* |
| NCBIdisease (Doğan et al., 2014) | *DiseaseClass* | *Disease* |
| | *SpecificDisease* | *Disease* |
| | *CompositeMention* | *Disease* |
| | *Modifier* | *Disease* |
| NLMchem (Islamaj et al., 2021a) | *Chemical* | *Chemical* |
| NLMGene (Islamaj et al., 2021b) | *Gene* | *Gene* |
| | *FamilyName* | *Gene* |
| SPECIES800 (Pafilis et al., 2013) | *Species* | *Species* |
| tmVar3 (Wei et al., 2022) | *DNAMutation* | *Variant* |
| | *ProteinMutation* | *Variant* |
| | *OtherMutation* | *Variant* |
| | *AcidChange* | *Variant* |
| | *SNP* | *Variant* |
| | *DNAAllele* | *Variant* |
| | *ProteinAllele* | *Variant* |
| | *Disease* | *Disease* |
| | *Chemical* | *Chemical* |
| | *Gene* | *Gene* |
| | *Species* | *Species* |
| | *Cell line* | *Cell line* |

Table 7: Correspondence of labels of the additional datasets to BioRED labels. The labels *DomainMotif* in GNormplus, *NonStandardRef* and *OTHER* in NLMChem, and *Cell*, *DomainMotif*, and *ChromosomeLocation* in NLMGene have no correspondence to the BioRED labels.

| Dataset | Multi | Multi+CVAE |
|---|---|---|
| BC5CDR (Li et al., 2016) | 91.42 | **91.59** |
| BioID (Arighi et al., 2017) | **90.71** | 90.01 |
| GNormPlus (Wei et al., 2015) | **80.97** | 80.68 |
| Linnaeus (Gerner et al., 2010) | **94.69** | 93.84 |
| NCBIdisease (Doğan et al., 2014) | **85.77** | 84.11 |
| NLMChem (Islamaj et al., 2021a) | **82.74** | 82.12 |
| NLMGene (Islamaj et al., 2021b) | **85.12** | 84.76 |
| SPECIES800 (Pafilis et al., 2013) | **81.12** | 77.64 |
| tmVar3 (Wei et al., 2022) | 85.29 | **91.87** |

Table 8: F1 scores [%] on test data of the additional datasets.