AdaBioBERT: Adaptive Token Sequence Learning for Biomedical Named Entity Recognition

Sumit Kumar and **Tanmay Basu**
Indian Institute of Science Education and Research Bhopal, Madhya Pradesh, India
{sumit23 | tanmay} @iiserb.ac.in

Abstract

Accurate identification and labeling of biomedical entities, such as diseases, genes, chemical and species, within scientific texts are crucial for understanding complex relationships. We propose Adaptive BERT or AdaBioBERT, a robust named entity recognition (NER) model that builds upon BioBERT (Biomedical Bidirectional Encoded Representation from Transformers) based on an adaptive loss function to learn different types of biomedical token sequence. This adaptive loss function combines the standard Cross Entropy (CE) loss and Conditional Random Field (CRF) loss to optimize both token level accuracy and sequence-level coherence. AdaBioBERT captures rich semantic nuances by leveraging pre-trained contextual embeddings from BioBERT. On the other hand, the CRF loss of AdaBioBERT ensures proper identification of complex multi-token biomedical entities in a sequence and the CE loss can capture the simple unigram entities in a sequence. The empirical analysis on multiple standard biomedical coprora demonstrates that AdaBioBERT performs better than the state of the arts for most of the datasets in terms of macro and micro averaged F1 score.

1 Introduction

The field of Biomedical Named Entity Recognition (NER) has evolved significantly, transitioning from rule-based systems to advanced deep learning methodologies. Early approaches relied heavily on handcrafted rules, dictionaries, and regular expressions to identify biomedical entities such as genes, diseases, and proteins. For instance, He (He et al., 2009) utilized domain-specific lexicons like UMLS to recognize entities. While these rule-based methods provided moderate accuracy, they struggled with the diversity and ambiguity of biomedical terminology, particularly for multi-token entities or novel terms. Their reliance on manual rule creation and limited adaptability hindered scalability (Settles, 2004; Leaman et al., 2015). The advent of machine learning techniques, such as Conditional Random Fields (CRF) (Sutton and McCallum, 2011) and Support Vector Machines (SVM) (Joachims, 1998), marked a shift toward data-driven models. CRF-based systems, like those developed by Settles (Settles, 2004) and Tsai (Tsai et al., 2006), leveraged labeled datasets to train classifiers that captured contextual and sequential information. These models demonstrated greater flexibility and adaptability compared to rule-based approaches. However, they still require extensive manual feature engineering, which limited their effectiveness in handling the complexity of biomedical data. For example, Leaman (Leaman et al., 2015) successfully applied CRF models to extract chemical and disease entities from PubMed abstracts but noted challenges in recognizing infrequent or context-dependent terms.

The introduction of Long Short-Term Memory (LSTM) networks and Convolutional Neural Networks (CNNs) revolutionized the NER tasks. Lample introduced a BiLSTM-CRF framework (Lample et al., 2016), which set new benchmarks for sequence labeling tasks, including NER. (Chiu and Nichols, 2016) extended this approach to biomedical texts, demonstrating the effectiveness of deep learning in capturing sequential dependencies and complex relationships. The emergence of transformer-based models, such as BERT (Devlin et al., 2019) and its biomedical counterpart, BioBERT (Lee et al., 2020), further advanced the capabilities of NER systems. These models employ self-attention mechanisms to capture the context of each word within a sentence, making them particularly effective for complex biomedical texts. BioBERT, which is pre-trained on biomedical corpora, has been effective in recognizing domain-specific entities (Lee et al., 2020). Unlike general-domain models, BioBERT effectively captures intricate relationships between biomedical terms, im-

Proceedings of the 24th Workshop on Biomedical Language Processing (BioNLP 2025), pages 56–62

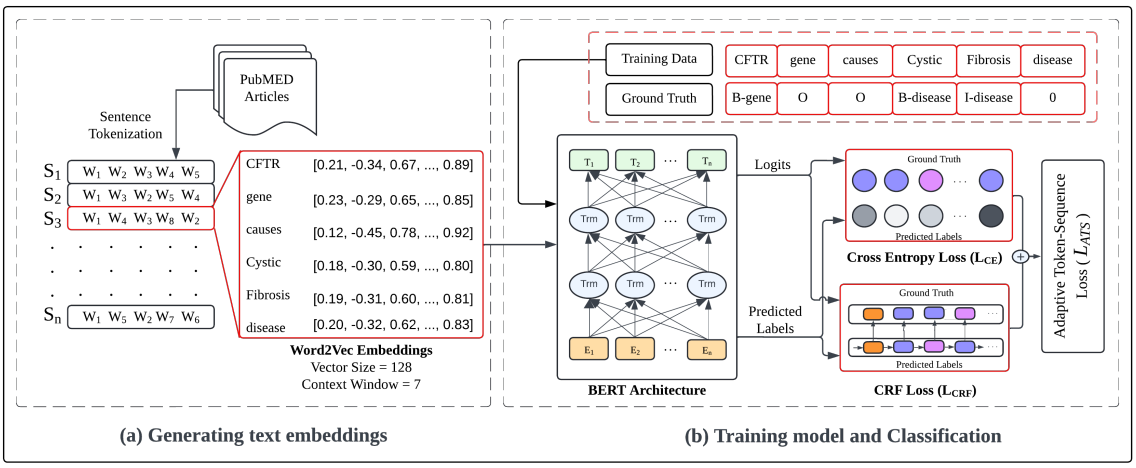

Figure 1: Proposed AdaBioBERT Architecture

proving NER performance in specialized datasets. Despite their effectiveness, transformer-based models often struggle to properly identify the named entities as they need large amount of data for fine-tuning (Chalkidis et al., 2020; Beltagy et al., 2019). Recent advancements have focused on combining the strengths of different loss functions. For example, Ma and Hovy (Ma and Hovy, 2016) introduced a BiLSTM-CRF model that used a fixed combination of CE and CRF loss functions for NER. Similarly, Lample (Lample et al., 2016) employed a fixed-weight combination of CE and CRF loss functions in their BiLSTM-CRF framework, which became a standard approach for NER tasks. However, these methods rely on fixed weighting scheme and cannot distinguish the significance between regular single token biomedical entities like *Nucleolin* and *Agyria*, and rare but important multi-token entities like *lateral sinus thrombosis* and *parietal cortical atrophy* through the loss functions.

Therefore, there is room to improve the quality of the existing methods to properly identify complex multi-token biomedical entities. In this spirit, this paper presents a transformer based Adaptive BioBERT (i.e., AdaBioBERT) NER model, to identify the nuances of complex multi-token biomedical entities by integrating a novel adaptive loss function combining the standard cross entropy and CRF loss functions in the pretrained Bio-BERT model (Lee et al., 2020).

2 Proposed AdaBioBERT Method

AdaBioBERT architecture has two major components: (1) Word2Vec embeddings (Kowsari et al., 2019), which capture semantic relationships between biomedical terms as shown in Fig 1(a) and (2) pretrained BioBERT model to generate rich contextual embeddings using the proposed Adaptive Token-Sequence Loss as shown in Fig 1(b), which dynamically balances token-level and sequence-level predictions.

2.1 Generate Word2Vec Embeddings of PubMED Data

In the first stage, the proposed framework extracts sentences from the freely available PubMED Central(PMC) repository[1], which has mention of any genes or diseases, based on frameworks proposed by (Basu et al., 2021; Guetterman et al., 2018). The objective is to build semantic embeddings of all relevant genes and diseases which are mentioned in the current version of DisGeNET[2] (v24.4) repository. It comprises 26,798 genes and 39,972 diseases and traits (Piñero et al., 2019). Subsequently, we generated word embeddings for these extracted sentences using Word2Vec model (Pennington et al., 2014; Kowsari et al., 2019). Sentences extracted from the PMC repository that build the corpus are tokenized, and then the Word2Vec algorithm generates embeddings for each word, which is represented as a 128-dimensional vector. The context window size of a word is set to 7, meaning the model considers up to seven neighboring words around a target word.

[1] https://pmc.ncbi.nlm.nih.gov/
[2] https://disgenet.com/DISGENET-Version-24-4-Whats-New

Table 1: Overview of biomedical datasets with training and testing splits

Dataset	Entity Types	Training	Test
BC4CHEMD (Krallinger et al., 2015)	Chemical compounds, drug names	6,000 abstracts	2,000 abstracts
LINNAEUS (Gerner et al., 2010)	Species names	80,000 sentences	10,000 sentences
NCBI-disease (Dogan et al., 2014)	Disease names	793 abstracts	100 abstracts
BC5CDR (Li et al., 2015)	Chemical compounds, diseases	1,000 articles	250 articles
JNLPBA (Kim et al., 2004)	Proteins, DNA, RNA, cell lines and types	2,000 abstracts	204 abstracts
AnatEM (Pyysalo, 2014)	Anatomical entities	1,200 documents	300 documents
BioNLP13GE (Kim et al., 2013)	Gene and gene product	1,500 sentences	500 sentences
Species-800 (Pafilis et al., 2016)	Species mentions	800 abstracts	200 abstracts

2.2 Pretrained BioBERT with Adaptive Token Sequence Loss (L_{ATS})

Let $X = \{x_1, x_2, \ldots, x_T\}$ denote an input sequence of tokens and $Y = \{y_1, y_2, \ldots, y_T\}$ represent the true labels of X, where y_t is a one-hot encoded vector and $y_t = [y_t^1, y_t^2, \cdots y_t^N]$ and $y_t^i \in \{c_1, c_2, \cdots c_N\}$. Here $c_i, \forall i = 1, 2, \ldots, N$ are different classes of biomedical entities. Let us consider $\widehat{Y} = \{\widehat{y_1}, \widehat{y_2}, \ldots, \widehat{y_T}\}$ be the sequence of predicted labels of the input sequence. The predicted probability for the t-th token $x_t \in c_i$ is denoted as $P(x_t \in c_i)$, and $\mathrm{S}(y_t, x_t)$ is the score of the true label sequence y_t given x_t. The L_{ATS} combines Cross-Entropy Loss (L_{CE}) and CRF Loss (L_{CRF}) as follows:

$$L_{ATS} = \alpha \cdot L_{CE} + (1 - \alpha) \cdot L_{CRF}, \quad (1)$$

where α is a learnable weight parameter to make a trade-off between CE loss and CRF loss. Here

$$L_{CE} = -\frac{1}{T} \sum_{t=1}^{T} \sum_{i=1}^{N} y_t^i \log \big(P(x_t \in c_i)\big)$$

is the Cross-Entropy Loss, which captures the sequence with a single biomedical entity and

$$L_{CRF} = -\Big(\mathrm{S}(Y, X) - \log \sum_{\widehat{Y}} \exp(\mathrm{S}(\widehat{Y}), X))\Big)$$

is the CRF Loss, which is used to identify complex multi-label entities in a sequence. L_{ATS} dynamically adjusts the importance of per-token accuracy and sequence coherence through the learnable weight α. The adaptive weight parameter α is updated iteratively after each training epoch using gradient descent, as described in Algorithm 1. When α is close to 1, the model prioritizes individual token predictions, while α close to 0 emphasizes sequence-level coherence for handling multi-token entities and domain-specific terminology. Eventually, the pretrained BioBERT model is fine-tuned using the word embeddings of the genes and diseases generated by the word2vec model in the first stage followed by using L_{ATS}.

Algorithm 1 Adaptive Token-Sequence Loss with Learnable Weight α

Input: Token sequence $X = \{x_1, \ldots, x_T\}$, true labels $Y = \{y_1, \ldots, y_T\}$
Initialize: Model parameters θ, adaptive weight $\alpha \in [0, 1]$, learning rate η
Output: Updated θ, α, and loss L_{ATS}
Compute token-level cross-entropy loss
$L_{CE} \leftarrow \frac{-1}{T} \sum_{t=1}^{T} \sum_{i=1}^{N} y_t^i \log P(x_t \in c_i)$
Compute CRF sequence-level loss
Compute score of true sequence $S(Y, X)$
Compute partition function $Z(X) = \log \sum_{\widehat{Y}} \exp(S(\widehat{Y}, X))$
$L_{CRF} \leftarrow -(S(Y, X) - Z(X))$
Compute adaptive loss
$L_{ATS} \leftarrow \alpha \cdot L_{CE} + (1 - \alpha) \cdot L_{CRF}$
Backpropagation and parameter updates
Compute gradients: $\nabla_\theta L_{ATS}$, $\nabla_\alpha L_{ATS}$
Update parameters:
    $\theta \leftarrow \theta - \eta \cdot \nabla_\theta L_{ATS}$
    $\alpha \leftarrow \alpha - \eta \cdot \nabla_\alpha L_{ATS}$
Return: Final loss L_{ATS}, updated θ, α

3 Experimental Evaluation

3.1 Datasets and Settings

Experimental evaluation was conducted on eight widely used biomedical NER datasets as reported in Table 1. All of these datasets are formatted

Table 2: Macro F1-Scores of AdaBioBERT and State of the Arts

Dataset	SciSpacy	Stanza	Spark NLP	PubMedBERT	BioBERT	AdaBioBERT
BC4CHEMD	71.98	83.25	90.09	91.43	91.72	**95.40**
Linnaeus	79.84	81.73	82.14	85.07	85.47	**87.51**
NCBI Disease	74.82	83.08	84.13	87.83	88.45	**92.68**
BC5CDR	74.47	83.13	83.25	88.67	85.37	**89.83**
JNLPBA	69.35	74.14	76.68	**79.16**	76.18	78.93
AnatEM	74.22	83.35	84.15	90.57	88.14	**94.03**
BioNLP13GE	73.70	82.93	83.24	80.24	84.91	**85.36**
Species800	73.67	81.04	83.14	82.79	81.93	**87.63**

in the IOB (Inside, Outside, Beginning) tagging scheme, ensuring consistency in annotation and format across different biomedical domains. Each dataset is processed by extracting unique labels and tokenized using the AutoTokenizer from Hugging Face's Transformers library, ensuring compatibility with the pre-trained BioBERT model. The Word2Vec embeddings, pre-trained on biomedical literature, are integrated into the model as an additional feature to enhance entity recognition. Our model architecture is based on BioBERT, extended with a CRF layer for structured sequence prediction. A fully connected classifier with dropout is applied to the concatenated BioBERT and Word2Vec embeddings, projecting them onto the label space. The loss function is a weighted combination of CE and CRF loss, where the weight is a trainable parameter optimized during training. The optimizer used is AdamW with weight decay to improve generalization. The model is fine-tuned for 5, 10, 20, 40 epochs with a batch size of 4, 8, 16, 32 using an initial learning rate of 1e-4, 2e-4, 3e-4[3]. A NVIDIA A100 40 GP GPU server is used to implement AdaBioBERT. Evaluation is performed on an evaluation dataset after each epoch, saving the best-performing checkpoint. The trainer relies on mixed precision training and gradient accumulation for efficient computation.

3.2 Results and Discussion

The performance of AdaBioBERT and the state of the arts are reported in Table 2 in terms of macro-averaged F1-score. It can be seen from Table 2 that AdaBioBERT recognizes the biomedical entities better than the state of the arts and it outperforms the other methods for all datasets for macro-averaged F1 scores. Significant improvement of the F1-score of our method can be observed in BC4CHEMD (+3.68 over BioBERT (Lee et al., 2020)), Linnaeus (+ 2.04 over BioBERT), NCBI Disease (+4.23 over BioBERT), BC5CDR(+1.16 over PubMedBERT (Gu et al., 2021)), AnatEM (+3.46 over PubMedBERT), Species800 (+4.49 over SparkNLP) and marginally exceeds BioNLP13GE (+0.45 over BioBERT). Having JLNPBA as an exception where it lags marginally (-0.23 by PubMedBERT) indicating required improvement for recognition of protein, cell line,, and cell type entities in biomedical data. These results suggest that AdaBioBERT excels in biomedical entity recognition tasks where contextual understanding is important. The performance of AdaBioBERT on diverse biomedical entity recognition datasets shows its adaptability and robustness.

Notable improvements in micro F1-score are also reported in Table 3, where AdaBioBERT surpasses the performance in BC4CHEMD (+2.64 over PubMedBERT), Linnaeus (+5.07 over PubMedBERT), NCBI Disease (+7.17 over BioBERT), and AnatEM (+5.41 over PubMedBERT), demonstrating AdaBioBERT's recognition capability in chemical and disease-related entities. Additionally, AdaBioBERT surpasses BioBERT in BioNLP13GE (+2.93), PubMedBERT in BC5CDR (+1.82), SparkNLP in JNLPBA (+2.98), and PubMedBERT on Species800 (+1.54).

The proposed AdaBioBERT model introduces a novel approach to biomedical NER by integrating Adaptive Token-Sequence Loss with pre-trained contextual embeddings from BioBERT. One of the key technical innovations of AdaBioBERT is

[3]Results are reported for 20 epochs, batch size of 32 and learning rate of 1e-4.

Table 3: Micro F1-Scores of AdaBioBERT and State of the Arts

Dataset	SciSpacy	Stanza	Spark NLP	PubMedBERT	BioBERT	AdaBioBERT
BC4CHEMD	84.55	89.65	93.72	95.17	92.36	**97.81**
Linnaeus	81.74	88.27	86.26	90.22	88.24	**95.29**
NCBI Disease	81.65	87.49	89.13	88.36	89.71	**96.88**
BC5CDR	83.92	88.08	89.73	92.88	90.61	**94.70**
JNLPBA	73.21	76.09	81.29	79.53	77.49	**84.27**
AnatEM	84.14	88.18	89.13	92.04	91.26	**97.45**
BioNLP13GE	77.60	84.34	85.58	89.47	92.66	**95.59**
Species800	74.06	83.35	84.91	86.76	85.31	**88.30**

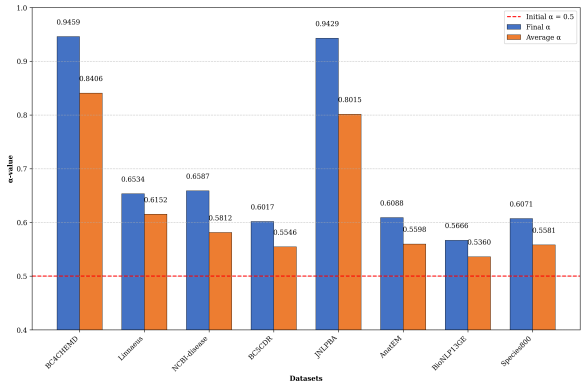

Figure 2: Final and Average α Values for Biomedical NER Datasets

its use of a learnable weight parameter (α) in the L_{ATS} loss function. This parameter enables the model to dynamically adjust the trade-off between token-level and sequence-level objectives during training, which ensures that our model can effectively handle both short, unambiguous entities and longer and complex ones. This flexibility is a significant improvement over the state of the arts that rely on fixed-weight combinations of L_{CE} and L_{CRF}, which may not generalize well across diverse biomedical texts. Additionally, the integration of pre-trained Word2Vec embeddings with BioBERT's contextual embeddings provides a multi-stage transfer learning framework, enhancing the model's ability to capture both semantic and contextual nuances in biomedical texts. The effectiveness of AdaBioBERT for identifying regular single token and complex multi-token entities has been demonstrated in the Table 2 and 3 for almost all datasets. The datasets like Species-800, NCBI Disease, and BC5CDR, where AdaBioBERT outperforms state-of-the-art by significant margins, contain lots of multi-token entities.

The different values of α in Figure 2 show how entity types vary in recognition difficulty. Chemical and gene entities (BC4CHEMD, JNLPBA) have much higher values (>0.94) because they use standard naming patterns that make individual words more important. Disease and anatomy terms (BioNLP13GE, BC5CDR, AnatEM) have lower values (0.56-0.66) because they need more context to understand ambiguous and less consistent names.

4 Conclusion

The potential of the proposed adaptive token-sequence loss with BioBERT embeddings is demonstrated through the extensive empirical analysis. By dynamically adjusting token-level and sequence-level learning through the learnable weight parameter (α), AdaBioBERT improves contextual understanding and multi-token entity recognition. Additionally, the integration of pre-trained Word2Vec embeddings further refines semantic representation in biomedical text. Despite its effectiveness, AdaBioBERT has high computational costs and may struggle with highly specific hierarchical entities. Future work will extend AdaBioBERT to broader biomedical information extraction tasks, including relation extraction, sentence classification, and document classification, to boost knowledge discovery in biomedical research. Codes available at: https://github.com/sumit-kumar-9297/AdaBioBERT-NER.git

5 Acknowledgment

This research is supported by the University Grants Commission (UGC), Government of India, through the Junior Research Fellowship (NTA Ref No: 220520545885).

References

Tanmay Basu, Simon Goldsworthy, and Georgios V Gkoutos. 2021. A sentence classification framework to identify geometric errors in radiation therapy from relevant literature. *Information*, 12(4):139.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. In *Proceedings of EMNLP*.

Ilias Chalkidis, Michail Fergadiotis, Roger Stradling, Nikolaos Pappas, and Prodromos Malakasiotis. 2020. Transformer-based models for legal and biomedical document classification: A comparative study. In *Proceedings of ACL*.

Jason PY Chiu and Eric Nichols. 2016. Named entity recognition with bidirectional lstm-cnns. *Transactions of the Association for Computational Linguistics*, 4:357–370.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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*, pages 4171–4186.

Rezarta Islamaj Dogan, 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. The linnaeus corpus: Annotated species names in biomedical literature. *Bioinformatics*, 26(18):i343–i349.

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).

Timothy C Guetterman, Tammy Chang, Melissa DeJonckheere, Tanmay Basu, Elizabeth Scruggs, and VG Vinod Vydiswaran. 2018. Augmenting qualitative text analysis with natural language processing: Methodological study. *J Med Internet Res*, 20(6):e231.

Longhua He, Mehmet Kayaalp, David Megginson, and Thomas Rindflesch. 2009. A rule-based approach for biomedical named entity recognition using umls. *BMC bioinformatics*, 10(1):1–11.

Thorsten Joachims. 1998. Text categorization with support vector machines: Learning with many relevant features. In *Machine Learning: ECML-98*, pages 137–142, Berlin, Heidelberg. Springer Berlin Heidelberg.

Jin-Dong Kim, Tomoko Ohta, Yoshimasa Tsuruoka, Yuka Tateisi, and Nigel Collier. 2004. Introduction to the bio-entity recognition task at jnlpba. *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications*, pages 70–75.

Jin-Dong Kim, Yue Wang, Toshihisa Takagi, and Atsushi Yonezawa. 2013. Overview of the bionlp 2013 shared task on genomic entity normalization. *Proceedings of the BioNLP Shared Task 2013 Workshop*, pages 1–7.

Kamran Kowsari, Kiana Jafari Meimandi, Mojtaba Heidarysafa, Sanjana Mendu, Laura Barnes, and Donald Brown. 2019. Text classification algorithms: A survey. *Information*, 10(4):150.

Martin Krallinger, Florian Leitner, Obdulia Rabal, Miguel Vazquez, Julen Oyarzabal, and Alfonso Valencia. 2015. Overview of the biocreative iv chemical and drug name recognition challenge. *Database*, 2015.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of NAACL-HLT*.

Robert Leaman, Rezarta Islamaj Dogan, and Zhiyong Lu. 2015. Taggerone: joint named entity recognition and normalization with semi-markov models. *Bioinformatics*, 31(8):1361–1370.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Jiao Li, Yijia Sun, Robin J Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, et al. 2015. Overview of the biocreative v chemical disease relation (cdr) task. *Proceedings of the BioCreative V challenge evaluation workshop*, pages 154–166.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional lstm-cnns-crf. *arXiv preprint arXiv:1603.01354*.

Evangelos Pafilis, Sune P Frankild, Lucia Fanini, Sarah Faulwetter, Christina Pavloudi, Aikaterini Vasileiadou, et al. 2016. The s800 corpus: A resource for species name recognition in biomedical texts. *Biodiversity Data Journal*, 4.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Janet Piñero, Juan Manuel Ramírez-Anguita, Josep Saüch-Pitarch, Francesco Ronzano, Emilio Centeno, Ferran Sanz, and Laura I Furlong. 2019. The disgenet knowledge platform for disease genomics: 2019 update. *Nucleic Acids Research*, 48(D1):D845–D855.

Sampo et al. Pyysalo. 2014. Anatem: A dataset for anatomical entity recognition in biomedical texts. *Journal of Biomedical Semantics*, 5(1):1–12.

Burr Settles. 2004. Biomedical named entity recognition using conditional random fields and rich feature sets. *Proceedings of the COLING*, pages 104–107.

Charles Sutton and Andrew McCallum. 2011. An introduction to conditional random fields. *Foundations and Trends in Machine Learning*, 4(4):267–373.

Richard TI Tsai, Shaojun Wu, and Tao-Hsien Chuang. 2006. A machine learning approach to biomedical named entity recognition. In *Proceedings of the 5th International Workshop on Bioinformatics*.