# MedBERT: A Pre-trained Language Model for Biomedical Named Entity Recognition

Charangan Vasantharajan* and Kyaw Zin Tun† and Ho Thi-Nga† and Sparsh Jain‡ and Tong Rong§ and Chng Eng Siong†
* University of Moratuwa, Colombo, Sri Lanka, E-mail: charangan.18@cse.mrt.ac.lk
† Nanyang Technological University, Nanyang, Singapore, E-mail: ztkyaw, ngaht, aseschng@ntu.edu.sg
‡ Birla Institute of Technology and Science, Pilani, India, E-mail: jainn.sparsh@gmail.com
§ Singapore Institute of Technology, Singapore E-mail: tong.rong@singaporetech.edu.sg

***Abstract*—This paper introduces MedBERT, a new pre-trained transformer-based model for biomedical named entity recognition. MedBERT is trained with 57.46M tokens collected from biomedical-related data sources, i.e. datasets acquired from N2C2, BioNLP, CRAFT challenges, and biomedical-related articles crawled from Wikipedia. We validate the effectiveness of MedBERT by comparing it with four publicly available pre-trained models on ten biomedical datasets from BioNLP and CRAFT shared tasks.**

**Our experimental results show that models fine-tuned on MedBERT achieve state-of-the-art performance in nine datasets that predict Protein, Gene, Chemical, Cellular/Component, Gene Ontology, and Taxonomy entities. Specifically, the model achieved an average of 84.04% F1-micro score on ten test sets from BioNLP and CRAFT challenges with an improvement of 3.7% and 7.83% as compared to models that were fine-tuned on BioBERT and Bio_ClinicalBERT, respectively.**

## I. Introduction

With the advancement in information technology and medical facilities, the generation of electronic medical records (EMR) has increased significantly [1]. The capability of recognizing various biomedical entities from a huge amount of biomedical documents automatically thus becoming more critical. This task is usually referred to as the task of Named Entity Recognition (NER) in the biomedical domain. NER is traditionally referred to as a Natural Language Processing (NLP) task that extracts required entities, e.g. location, person name, time, from text-based documents. In the biomedical domain, NER is adapted to recognize the biomedical entities of interest such as chemicals, diseases, proteins, cellular components, and molecular functions.

In recent years, transformer-based pre-trained models [2] have been widely adopted in biomedical NER systems. There is a wide range of pre-trained models: from models that are trained on general domain text such as BERT and DistilBERT to models that are trained specifically on the biomedical domain such as Bio_ClinicalBERT and BioBERT that are trained on PubMed and MIMIC III [3] corpus. Both Bio_ClinicalBERT and BioBERT have achieved good results on their matched test sets, i.e. NCBI Disease [4], 2010 i2b2/VA [5], BC5CDR [6], and JNLPBA [7]. However, the models did not perform well on other datasets like BioNLP [8, 9] and CRAFT [10] which have different biomedical disciplines, i.e. Cell Ontology, Gene Ontology, and Sequence Ontology. This may be mainly due to the differences in linguistic characteristics between multiple biomedical disciplines. This motivates us to develop a pre-trained biomedical transformer-based model to cover biomedical disciplines on a larger scale.

This paper introduces a new transformer model for the biomedical domain, namely MedBERT, which is pre-trained by initiating from Bio_ClinicalBERT and adapted on large medical corpora collected from multiple sources, specifically National NLP Clinical Challenges (N2C2)[1], BioNLP Corpus [11], and CRAFT Corpus [12]. We evaluated the effectiveness of using generic versus clinical-based transformer models for identifying biomedical entities from clinical/biomedical texts. Specifically, we used BERT [2], and DistilBERT [13] for generic domain and MedBERT, BioBERT [14], and Bio_ClinicalBERT [15] for clinical domain. Experimental results show that NER models that are fine-tuned from MedBERT outperform models that are fine-tuned on the existing generic and clinical transformers.

In summary, our contributions are as follows:

- We developed and released MedBERT, a state-of-the-art pre-trained transformer-based model for biomedical named entity recognition. The model outperforms BERT, DistilBERT, Bio_ClinicalBERT and BioBERT on 9 out of 10 shared tasks across BioNLP [9, 8] and CRAFT [16] challenges. The model is made publicly available[2].
- We report our experimental results on the effectiveness of using generic and clinical transformer-based models on biomedical named entity recognition tasks. The experimental results are documented and could be easily replicated using our released NERP[3] package. This can stand as a referenced baseline for future development on the task.

The rest of this paper is organized as follows. Related studies are presented in Section II. Subsequently, Sections III and IV describe the proposed pre-trained model and fine-tuning processes respectively. The datasets and the experiment settings

[1] https://portal.dbmi.hms.harvard.edu/projects/n2c2-nlp/
[2] https://huggingface.co/Charangan/MedBERT
[3] https://github.com/Chaarangan/NERP

are presented in Section V followed by the discussion of results. Lastly, we conclude this work in Section VI.

## II. Related Work

Many techniques have been introduced for biomedical information extraction (IE), classification, and named entity recognition. Earliest extraction models [17] [18] [19] relied on the rule-based approach with semantic dictionary to extract drug, gene and protein entities from biomedical records. Knowledge bases such as MetaMap and semantic lexicons were used to identify the drug mentions. These models required domain knowledge and manual work to produce effective rules for the extraction. Hence rule-based models are inefficient to maintain due to the multidisciplinary nature of the medical domain.

As opposed to the deterministic nature of rule-based approaches, Machine Learning (ML) approaches utilize data to generate probabilistic models. One robust classifier of the ML approach is the Support Vector Machine (SVM) where the hyperplane separates the boundary of each distinct class or entity. Studies have explored SVM for biomedical named entity recognition where [20] makes use of subclass splitting, POS tagging, and HMM states features to tackle data sparseness problems. [21] introduced two-phased SVM with a lookup table where one SVM identifies an entity while the other is tasked with classification. Although these early ML techniques have shown to achieve great performance, the datasets used in the experiments are relatively small. When large volumes of documents are available, SVM NER models could not scale proportionally [22].

The rise of deep learning techniques introduced neural network models such as RNN, LSTM, and Transformer that can leverage huge amounts of annotated data to learn a better word representation or create word embeddings [23]. These word embeddings are often utilized as an initial embedding layer, fine-tuned, or can be pre-trained further in-domain medical data by taking advantage of a large amount of un-annotated clinical data [24] [25].

Transformer models such as Bidirectional Encoder Representations from Transformers (BERT) [2] and their variants are widely used word embedding models and have shown to give good results for biomedical entity extraction tasks. BioBERT (BERT for Biomedical Text Mining) [14], pre-trained on large-scaled biomedical data, significantly outperforms previous state-of-the-art models on biomedical NER tasks while Bio_Clinical BERT [15] further improved BioBERT performance by pre-training and fine-tuning with clinical text.

However, there is no study exploring transformers specialized for multiple biomedical entities even though studies from the clinical domain have reported that BioBERT and Bio_ClinicalBERT outperform many benchmarks, e.g. MedNLI, and i2b2. In this study, we systematically examined widely used transformer models for biomedical named entity extraction and developed an open-source clinical transformer-based model specialized for multiple biomedical disciplines.

## III. Proposed MedBERT

### A. MedBERT pre-training Corpus

BioBERT and Bio_ClinicalBERT were trained on biomedical and clinical domain data, and they outperform the traditional BERT on the task of biomedical named entity recognition from medical text and clinical narratives [14, 15]. However, some studies [26, 14, 15] have shown that they did not perform well on data from different biomedical domains data like BioNLP and CRAFT community corpora.

To tackle a wider range of medical context domains, MedBERT is pre-trained on text corpora comprising multiple medical-related data sources, i.e. articles provided by N2C2 challenges (N2C2 Corpus), articles released under the BioNLP project[4] (CRAFT Corpus and BioNLP Corpus), and biomedical related articles crawled from Wikipedia (Wikipedia Corpus). Specifically, N2C2 Corpus is a collection of articles released in N2C2 2018 and N2C2 2022 challenges as supplements to the tasks. CRAFT Corpora consists of 67 full-text open-access biomedical journal articles from PubMed Central that covers a wide range of biomedical domains including biochemistry and molecular biology, genetics, developmental biology, and computational biology. BioNLP Corpus contains the rest of the articles released under the BioNLP project. The articles cover multiple biomedical disciplines such as molecular biology, IE for protein and DNA modifications, biomolecular mechanisms of infectious diseases, habitats of bacteria mentioned, and bacterial molecular interactions and regulations. For Wikipedia Corpus, we acquire the biomedical-related articles in the following steps:

- Firstly, we created a dictionary consisting of ∼5k biomedical named entities. The names are randomly selected from the labeled NER datasets that we selected for the fine-tuning NER presented in Section IV. Each entry in the dictionary is then used as search terms to find the relevant articles in Wikipedia using a python package namely "wikipedia"[5]. As a result, we crawled more than 10k Wikipedia pages, with approximately 794k sentences.
- Secondly, we filtered irrelevant articles that were accidentally crawled from Wikipedia since it contains keywords that are the same as medical terms. For instance, "Arv" is a medical name, but it is also an album name. We used XenC [27] to find the text that is relevant to the in-domain data, i.e. N2C2, BioNLP, and CRAFT, from the out-of-domain dataset, i.e. crawled Wikipedia text. XenC is used to compute the difference between cross-entropy scores of sentences from the crawled Wikipedia text and the in-domain text. We then selected only sentences with a cross-entropy score that was less than 0.55, this is about 70% of total crawled sentences.

Table I presents the number of sentences and words obtained for each corpus after pre-processing. Specifically, N2C2 Corpus contains 1.23M of sentences and 19.25M of words. The

[4]http://bionlp.sourceforge.net/index.shtml
[5]https://pypi.org/project/wikipedia/

TABLE I: In-domain text corpora used for pre-training MedBERT

| Corpus | No. of Sentences | No. of Words |
|---|---|---|
| N2C2 Corpus | 1.23M | 19.25M |
| CRAFT Corpus | 0.26M | 6.90M |
| BioNLP Corpus | 0.45M | 12.77M |
| Wikipedia Corpus | 0.79M | 18.54M |

CRAFT corpora consist of 0.26M sentences and 6.90M words, and BioNLP Corpus contains 0.45M sentences and 12.77M words, respectively. Lastly, we obtained 794k sentences and 18.54M words for the Wikipedia corpus.

### B. Pre-training MedBERT

The overall process of pre-training MedBERT is illustrated in Figure 1.

*1) Word tokenization:* To mitigate the out-of-vocabulary issue that is commonly observed in biomedical text processing, we use WordPiece [28] algorithm to convert each word into a subword sequence. After tokenization, words that are not in the vocabulary are represented by a sequence of subwords, e.g. the word Buspirone is represented by "Bus", "##piron". In a subword, "##" means that the rest of the token should be attached to the previous word, without space for decoding or reversal of the tokenization. To have better coverage of the biomedical terms, the wordpiece vocabulary is derived from the training corpus in Table I.

*2) Model Training:* The network weights of the MedBERT are initialized from the Bio_ClinicalBERT model [15]. In the first stage, we train the MedBERT model on data from N2C2 Corpus, BioNLP Corpus, and CRAFT Corpus for the first 150k steps and checkpoints have been saved every 25k steps. Subsequently, it was trained on Wikipedia Corpus for the next 50k steps, and checkpoints are saved every 25k steps.

When evaluating the development sets, we found that the 200k pre-training steps were optimal. For the rest of this paper, we refer to the model obtained from 200k steps as MedBERT. For the other hyper-parameters, we used the dupe factor at 5 for duplicating input data with different masks and the max predictions per sequence are set at 22. The masked language model probability of 0.15 was used when creating the pre-training input data from the original text corpora. Maximum length, batch size, and learning rate are set at 256, 16, and 0.0001 respectively.

## IV. Fine-tuning BERT-based Models For Biomedical NER

### A. Network Architecture for Fine-tuning

To evaluate the effectiveness of using generic and clinical transformer-based models for the biomedical NER tasks, we employed the same network architecture as depicted in Figure 2. This architecture is inspired by previous works such as [29] and [30]. The architecture has been proven to work best for the task. Specifically, the network consists of four layers as follows:

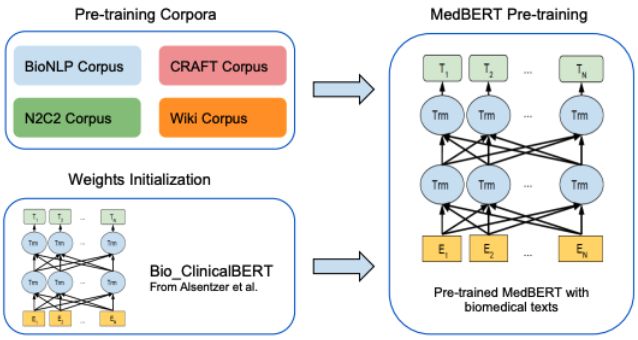



Fig. 1: Overview of the pre-training of MedBERT.

- *Transformer-based model layer*: The layer allows us to employ different transformer-based models for the evaluation.
- *Bidirectional Long Short-Term Memory layer*: This layer consists of two LSTM networks: one processes the input in a forward direction and another one processes the input in a backward direction and provides contextual information on training samples surrounding each token. It receives the transformer output, processes each sentence token-by-token, and produces an intermediate representation.
- *Fully connected linear layer*: It connects every input neuron to every output neuron to produce the final output. Also, it helps to reshape the output of the transformer model into 3 dimensions, i.e., batch size, number of inputs, and number of outputs. The intermediate representation from the BiLSTM layer is used as an input for the linear layer. Then the reshaped output is passed to the next layer.
- *CRF layer*: CRF is tagging algorithm that learns the task objective and constraints from the model training and makes sure that it returns the most accurate combination of predicted tags.

### B. Selected Pre-trained Transformer Model

*1) Generic Models:* We adopted BERT and DistilBERT as generic transformer-based models for our evaluation. A brief description of the two models is as below:

- **BERT** (bert-base-uncased): [2]: The first and most successful transformer-based model was trained and released by researchers at Google AI Language. It has been pre-trained with masked language modeling and next sentence prediction objectives on a dataset of 800M words from Toronto Book Corpus and 2500M words from English Wikipedia. BERT has been used in a wide range of text processing applications and significantly outperformed traditional systems.
- **DistilBERT** (distilbert-base-uncased) [13]: DistilBERT was released in 2020. It is a smaller, faster, cheaper, and lighter version of BERT. Compared to BERT-base, it has

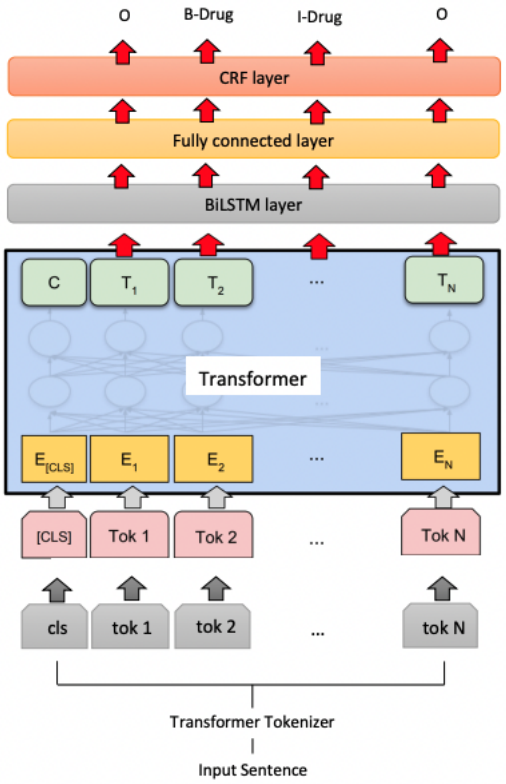



Fig. 2: Network Architecture for Fine-tuning

40% fewer parameters and achieves 0.6% less accuracy than BERT in IMDB sentiment classification tasks.

*2) Clinical Models:* In the past few years, variations of BERT in the clinical domain have been released. Specifically, BioBERT was released by [14] and Bio_ClinicalBERT was released by [15]. In this work, in addition to the proposed MedBERT, we evaluate both BioBERT and Bio_ClinicalBERT.

- **MedBERT**: The proposed pre-trained model for biomedical NER.
- **BioBERT** (biobert-base-cased-v1.2): It is a domain-specific language model pre-trained on a large-scale biomedical domain corpus. It was reported to outperform BERT and other state-of-the-art models in some biomedical tasks.
- **Bio_ClinicalBERT** (Bio_ClinicalBERT): It was initialized from BioBERT & pre-trained on all MIMIC III, a database containing electronic health records from ICU patients at the Beth Israel Hospital in Boston, MA) notes.

## V. Experiments

### A. Datasets for fine-tuning

Table II shows the distribution of entities for train, development, and test sets of the ten datasets that are selected for fine-tuning transformer models. The datasets are organized by [11] and made publicly accessible on Github[6]. A brief description of the datasets is as follows.

1) *BioNLP13GE*: The dataset was created for GENIA Event Extraction in BioNLP 2013 shared task. It contains 20 full paper articles sourced from the PubMed Central Open Access subset. There are a total of 7721 manually annotated protein names in the dataset [11].
2) *BioNLP13PC*: It consists of 525 PubMed abstracts and annotated close to 16,000 bio-molecular entities.
3) *BioNLP13CG-chem and BioNLP13CG-ggp*: The datasets are built from 600 PubMed abstracts by annotating over 17,000 events. BioNLP13CG-chem contains annotated entities for genes, RNA, and proteins while BioNLP13CG-ggp contains annotated entities for simple, non-repetitive chemical entities.
4) *BioNLP11ID-chem*: The dataset contains primary text content of articles that is annotated with mentions of chemical compounds such as “NaCl”. The article is selected by infectious diseases domain experts from full-text PMC open access documents
5) *CRAFT*: The dataset contains 67 annotated full-text biomedical journal articles from PubMed Central Open Access Subset. The journal articles are drawn from diverse biomedical disciplines and are completely annotated. The annotated entities are Sequence Ontology, Gene Ontology, Cell Type Ontology, Chemical Entities of Biological Interest ontology, NCBI Taxonomy, and Protein Ontology. We use the annotated physical entities from this corpus.
6) *CRAFT-ggp*: The dataset is selected from the CRAFT dataset with only annotated entities and sub-entities for Gene Ontology.
7) *CRAFT-cell*: The dataset is elected from the CRAFT dataset with only annotated entities and sub-entities for Cell Ontology.
8) *CRAFT-chem*: The dataset is selected from the CRAFT dataset with only annotated entities and sub-entities for Chemical Entities of Biological Interest.
9) *CRAFT-species*: The dataset is selected from the CRAFT dataset with only annotated entities and sub-entities for NCBI Taxonomy.

### B. Experiment Settings

For each data set listed in Table II, we use the train, validation sets to fine-tune every pre-trained model and test set for reporting the experimental results. During the fine-tuning process, the same set of hyper-parameters is used. Specifically, a maximum length of 256 was selected and a batch size of 32 was selected across all tasks. We set the learning rate as 0.0001 along with the AdamW optimizer. Finally, we run the experiments on each dataset for 10 epochs.

### C. Discussion of results

Table III shows the experimental results on biomedical entity extraction tasks using the selected transformer models on ten public clinical NER datasets described in Section V-A.

[6]https://github.com/cambridgeltl/MTL-Bioinformatics-2016

TABLE II: Distribution of the biomedical NER datasets used for fine-tuning

| Dataset | Entities | No. of Entities | Distribution | | |
|---|---|---|---|---|---|
| | | | Train | Validation | Test |
| BioNLP13GE [31] | Protein | 12,057 | 3,566 | 4,132 | 4,359 |
| BioNLP13PC [31] | Gene/Protein, Complex, Chemical, Cellular/ Component | 15893 | 7,853 | 2,734 | 5,306 |
| BioNLP13CG-chem [31] | Chemical | 2,270 | 1,097 | 446 | 727 |
| BioNLP13CG-ggp [31] | Gene/Protein | 7,908 | 4,028 | 1,360 | 2,520 |
| BioNLP11ID-chem [32] | Chemical | 973 | 594 | 131 | 248 |
| CRAFT-ggp [10] | Gene/Protein | 16,114 | 8,531 | 2,301 | 5,282 |
| CRAFT-cell [10] | Cell Line | 5,495 | 2,894 | 832 | 1,769 |
| CRAFT-chem [10] | Chemical | 6,802 | 3,377 | 1,044 | 2,381 |
| CRAFT-species [10] | Taxonomy | 7,394 | 3,361 | 1,548 | 2,485 |
| CRAFT [10] | Gene/Protein, Cell Line, Chemical, Taxonomy, Gene Ontology, Sequence Ontology | 57,634 | 29,026 | 9,727 | 18,881 |

TABLE III: Performance of biomedical NERs that fine-tuned and tested on ten selected datasets. Reporting score is F1-Micro. The best scores are in bold, and the second-best scores are underlined.

| Challenge | Test set | BERT | DistilBERT | Bio_ClinicalBERT | BioBERT | MedBERT |
|---|---|---|---|---|---|---|
| BioNLP | BioNLP13GE | 72.92 | 70.41 | 74.54 | <u>78.11</u> | **79.32** |
| | BioNLP13PC | 86.27 | 85.42 | 88.62 | **89.92** | <u>89.32</u> |
| | BioNLP13CG-chem | 71.75 | 69.51 | <u>80.03</u> | 78.69 | **82.06** |
| | BioNLP13CG-ggp | 88.07 | 87.29 | 90.55 | <u>90.59</u> | **91.25** |
| | BioNLP11ID-chem | 34.79 | 33.29 | 34.22 | <u>44.78</u> | **64.17** |
| CRAFT | CRAFT-ggp | 72.86 | 68.03 | 78.59 | <u>80.87</u> | **82.99** |
| | CRAFT-cell | 72.77 | <u>81.65</u> | 53.76 | 80.22 | **84.55** |
| | CRAFT-chem | 77.65 | 77.77 | 82.55 | <u>82.58</u> | **83.92** |
| | CRAFT-species | 92.88 | 95.70 | <u>96.48</u> | 94.47 | **98.25** |
| | CRAFT | 80.31 | 78.52 | 82.76 | <u>83.20</u> | **84.57** |
| **Average** | | 75.03 | 74.76 | 76.21 | <u>80.34</u> | **84.04** |

*1) Generic models vs clinical models:* Experimental results have clearly shown that the clinical models, presented in the last three columns, has higher averaged F1-micro score than the two generic models, i.e. BERT, and DistilBERT on most of the test sets. For example, on the BioNLP13GE test set, the Bio_ClinicalBERT, BioBERT, and MedBERT models outperformed the BERT by 2.22%, 7.12%, and 8.78% respectively in terms of F1-Micro score. The only exception was CRAFT-cell where DistilBERT performs better than Bio_Clinical and BioBERT with 1.43% and 27.89% improvement respectively. For a comparison between generic models, it can be seen that BERT significantly outperformed DistilBERT by 3.22% and 7.10% on BioNLP13CG-chem and CRAFT-ggp test sets. However, they have similar performances on the rest of the test sets.

*2) Performance of MedBERT:* MedBERT achieved the best averaged F1-Micro scores on nine out of ten test sets as compared to the other models. The highest jump is a 19.39% absolute improvement on the BioNLP13CG-chem test set. The highest F1-Micro score was obtained on the CRAFT-species test set at 98.25%. In the exceptional case of the BioNLP13PC test set, MedBERT is only 0.6% behind the best F1-Micro score obtained by BioBERT. But, in terms of individual f1 scores of both models for this dataset, MedBERT only performed slightly poorer than BioBERT in predicting "Complex" entity. This may be due to the lack of pre-training data that are related to the complex.

*3) Comparison between BioBERT and Bio_ClinicalBERT:* Another interesting observation is that BioBERT outperformed Bio_ClinicalBERT on seven out of ten test sets. This is prob-

ably because BioBERT has pre-train and thus has the prior knowledge about the biomedical entities from the PubMed text which most of the test sets are based on. On the other hand, Bio_ClinicalBERT is specialized for clinical notes from MIMIC III which are less likely to be relevant to the test sets.

Overall, the experiment results demonstrated that it is necessary to pre-train transformer-based models using relevant domain text when fine-tuning them for that particular domain, in our case NER for various biomedical disciplines.

## VI. Conclusions

In this study, we proposed MedBERT for biomedical Named Entity Recognition. MedBERT is trained from a diverse biomedical-related corpus in different disciplines. We compared the performance of MedBERT and four other popular pre-trained BERT models by fine-tuned and testing them on ten biomedical datasets. The experimental results show that the MedBERT encodes richer biomedical information, thus it achieves the best average F1-Micro scores on nine out of ten test sets.

Moreover, with a minimal task-specific architectural modification, we hope that MedBERT can be applied to other biomedical NLP tasks through fine-tuning. The pre-trained weights of MedBERT are made publicly available at Google Drive[7] and the source code for the fine-tuning of all transformer models is available at GitHub[8].

## Acknowledgment

This research is supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG- 100E-2018-006). The computational work for this article was partially performed on resources of the National Supercomputing Centre, Singapore (https://www.nscc.sg).

## References

[1] D. Campos, S. Matos, and J. L. Oliveira, "Biomedical named entity recognition: A survey of machine-learning tools," in *Theory and Applications for Advanced Text Mining* (S. Sakurai, ed.), ch. 8, Rijeka: IntechOpen, 2012.

[2] J. Devlin, M.-W. Chang, K. Lee, and K. 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)*, (Minneapolis, Minnesota), pp. 4171–4186, Association for Computational Linguistics, June 2019.

[3] A. Johnson, T. Pollard, L. Shen, L.-w. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Celi, and R. Mark, "Mimic-iii, a freely accessible critical care database," *Scientific Data*, vol. 3, p. 160035, 05 2016.

[4] R. Dogan, R. Leaman, and Z. lu, "Ncbi disease corpus: A resource for disease name recognition and concept normalization," *Journal of biomedical informatics*, vol. 47, 01 2014.

[5] O. Uzuner, B. South, S. Shen, and S. DuVall, "2010 i2b2/va challenge on concepts, assertions, and relations in clinical text," *Journal of the American Medical Informatics Association : JAMIA*, vol. 18, pp. 552–6, 06 2011.

[6] J. Li, Y. Sun, R. Johnson, D. Sciaky, C.-H. Wei, R. Leaman, A. P. Davis, C. Mattingly, T. Wiegers, and Z. lu, "Biocreative v cdr task corpus: a resource for chemical disease relation extraction," *Database*, vol. 2016, p. baw068, 05 2016.

[7] J.-D. Kim, T. Ohta, Y. Tsuruoka, Y. Tateisi, and N. Collier, "Introduction to the bio-entity recognition task at jnlpba," *Proceedings of International Joint Workshop on Natural Language Processing in Biomedicine and its Applications*, 01 2004.

[8] C. Nédellec, R. Bossy, J.-D. Kim, J.-j. Kim, T. Ohta, S. Pyysalo, and P. Zweigenbaum, "Overview of BioNLP shared task 2013," in *Proceedings of the BioNLP Shared Task 2013 Workshop*, (Sofia, Bulgaria), pp. 1–7, Association for Computational Linguistics, Aug. 2013.

[9] J.-D. Kim, S. Pyysalo, T. Ohta, R. Bossy, N. Nguyen, and J. Tsujii, "Overview of bionlp shared task 2011," pp. 1–6, 01 2011.

[10] M. Bada, M. Eckert, D. Evans, K. Garcia, K. Shipley, D. Sitnikov, W. Baumgartner Jr, K. Cohen, K. Verspoor, J. Blake, and L. Hunter, "Concept annotation in the craft corpus," *BMC bioinformatics*, vol. 13, p. 161, 07 2012.

[11] G. Crichton, S. Pyysalo, B. Chiu, and A. Korhonen, "A neural network multi-task learning approach to biomedical named entity recognition," *BMC Bioinformatics*, vol. 18, 08 2017.

[12] K. B. Cohen, K. Verspoor, K. Fort, C. Funk, M. Bada, M. Palmer, and L. E. Hunter, "The colorado richly annotated full text (craft) corpus: Multi-model annotation in the biomedical domain," in *Handbook of Linguistic Annotation*, pp. 1379–1394, Springer, 2017.

[13] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter," *ArXiv*, vol. abs/1910.01108, 2019.

[14] 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," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.

[15] E. Alsentzer, J. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann, and M. McDermott, "Publicly available clinical BERT embeddings," in *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, (Minneapolis, Minnesota, USA), pp. 72–78, Association for Computational Linguistics, June 2019.

[16] W. Baumgartner, M. Bada, S. Pyysalo, M. R. Ciosici, N. Hailu, H. Pielke-Lombardo, M. Regan, and L. Hunter, "CRAFT shared tasks 2019 overview — integrated struc-

[7] https://huggingface.co/Charangan/MedBERT

[8] https://github.com/Chaarangan/NERP

ture, semantics, and coreference," in *Proceedings of The 5th Workshop on BioNLP Open Shared Tasks*, (Hong Kong, China), pp. 174–184, Association for Computational Linguistics, Nov. 2019.

[17] A. Mykowiecka, M. Marciniak, and A. Kupść, "Rule-based information extraction from patients' clinical data," *Journal of Biomedical Informatics*, vol. 42, no. 5, pp. 923–936, 2009. Biomedical Natural Language Processing.

[18] A. B. Abacha and P. Zweigenbaum, "Automatic extraction of semantic relations between medical entities: Application to the treatment relation," in *Semantic Mining in Biomedicine*, 2010.

[19] D. Hanisch, K. Fundel-Clemens, H.-T. Mevissen, R. Zimmer, and J. Fluck, "Prominer: Rule-based protein and gene entity recognition," *BMC bioinformatics*, vol. 6 Suppl 1, p. S14, 02 2005.

[20] J. Kazama, T. Makino, Y. Ohta, and J. Tsujii, "Tuning support vector machines for biomedical named entity recognition," pp. 1–8, 01 2002.

[21] K.-J. Lee, Y.-S. Hwang, and H.-C. Rim, "Two-phase biomedical NE recognition based on SVMs," in *Proceedings of the ACL 2003 Workshop on Natural Language Processing in Biomedicine*, (Sapporo, Japan), pp. 33–40, Association for Computational Linguistics, July 2003.

[22] S. Karamizadeh, S. M. Abdullah, M. Halimi, J. Shayan, and M. j. Rajabi, "Advantage and drawback of support vector machine functionality," in *2014 International Conference on Computer, Communications, and Control Technology (I4CT)*, pp. 63–65, 2014.

[23] J. Pennington, R. Socher, and C. Manning, "GloVe: Global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, (Doha, Qatar), pp. 1532–1543, Association for Computational Linguistics, Oct. 2014.

[24] M. Habibi, L. Weber, M. Neves, D. L. Wiegandt, and U. Leser, "Deep learning with word embeddings improves biomedical named entity recognition," *Bioinformatics*, vol. 33, pp. i37–i48, 07 2017.

[25] A. Jagannatha and H. Yu, "Structured prediction models for RNN based sequence labeling in clinical text," *CoRR*, vol. abs/1608.00612, 2016.

[26] Y. Peng, S. Yan, and Z. Lu, "Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten benchmarking datasets," in *Proceedings of the 18th BioNLP Workshop and Shared Task*, (Florence, Italy), pp. 58–65, Association for Computational Linguistics, Aug. 2019.

[27] A. Rousseau, "Xenc: An open-source tool for data selection in natural language processing," *The Prague Bulletin of Mathematical Linguistics*, no. 100, pp. 73–82, 2013.

[28] M. Schuster and K. Nakajima, "Japanese and korean voice search," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5149–5152, 2012.

[29] X. Yang, J. Bian, W. R. Hogan, and Y. Wu, "Clinical concept extraction using transformers," *Journal of the American Medical Informatics Association : JAMIA*, vol. 27, p. 1935—1942, December 2020.

[30] K. Hakala and S. Pyysalo, "Biomedical named entity recognition with multilingual BERT," in *Proceedings of The 5th Workshop on BioNLP Open Shared Tasks*, (Hong Kong, China), pp. 56–61, Association for Computational Linguistics, Nov. 2019.

[31] T. Ohta, S. Pyysalo, R. Rak, A. Rowley, H.-W. Chun, S.-J. Jung, S.-P. Choi, S. Ananiadou, and J. Tsujii, "Overview of the pathway curation (PC) task of BioNLP shared task 2013," in *Proceedings of the BioNLP Shared Task 2013 Workshop*, (Sofia, Bulgaria), pp. 67–75, Association for Computational Linguistics, Aug. 2013.

[32] J.-D. Kim, S. Pyysalo, T. Ohta, R. Bossy, N. Nguyen, and J. Tsujii, "Overview of BioNLP shared task 2011," in *Proceedings of BioNLP Shared Task 2011 Workshop*, (Portland, Oregon, USA), pp. 1–6, Association for Computational Linguistics, June 2011.