Transforming Drug-Drug Interaction Extraction from Biomedical Literature

Dimitrios Zaikis
dimitriz@csd.auth.gr
Aristotle University of Thessaloniki
Thessaloniki, Thessaloniki, Greece

Stylianos Kokkas
kokkassk@csd.auth.gr
Aristotle University of Thessaloniki
Thessaloniki, Thessaloniki, Greece

Ioannis Vlahavas
vlahavas@csd.auth.gr
Aristotle University of Thessaloniki
Thessaloniki, Thessaloniki, Greece

ABSTRACT

Language Models (LM) capture the characteristics of the distribution of words sequences in natural language, learning meaningful distributed representations in the process. Recent advancements in Neural Networks and Deep Leaning have led to rapid progress in this area, greatly attributed to the emergence of the attention mechanism. Specifically, the Transformer architecture that implements attention-based encoder-decoder stacks, has advanced research in numerous Natural Language Processing tasks and produced state-of-the-art pre-trained LMs. One important task is Relationship Extraction (RE), which extracts semantic relationships from text and has significant applications in the biomedical domain, especially in literature pertaining to drug safety and Drug-Drug Interactions (DDI). In DDI extraction, the task is divided into two subtasks, Drug Named Entity Recognition and Relation Classification, consequently identifying drug mentions and classifying the potential effect of drug combinations from literature. Various methods for the extraction of DDIs have been proposed that utilize different architectures, however Transformers-based LMs continue to show the most promise. The overwhelming number of available pre-trained LMs, that each provide their own benefits and disadvantages, renders the selection of a clear baseline extremely difficult. In this paper, we investigate the most relevant LMs for biomedical RE and experiment on the DDI Extraction 2013 dataset. We propose a baseline approach for pre-trained Transformer-based LMs with shallow output architectures to effectively utilize the underlying architecture and introduce a foundation that reaches similar to state-of-the-art performance for both subtasks of DDI extraction.

CCS CONCEPTS

• **Computing methodologies** → *Neural networks*; • **Information systems** → *Information retrieval.*

KEYWORDS

information retrieval, neural networks, drug-drug interactions, transformers, language model

SETN 2022, September 7–9, 2022, Corfu, Greece

ACM ISBN 978-1-4503-9597-7/22/09…$15.00
https://doi.org/10.1145/3549737.3549753

ACM Reference Format:
Dimitrios Zaikis, Stylianos Kokkas, and Ioannis Vlahavas. 2022. Transforming Drug-Drug Interaction Extraction from Biomedical Literature . In *12th Hellenic Conference on Artificial Intelligence (SETN 2022), September 7–9, 2022, Corfu, Greece.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3549737.3549753

1 INTRODUCTION

Literature search is a fundamental step for biomedical researchers, clinicians, database curators and bibliometricians during their scientific discovery process, aiming to find the most relevant information. The exponential growth of published literature coupled with the heterogeneous information sources and the diverse needs of the scientific community render this task difficult and time-consuming. Consequently, these challenges have led to the application of Natural Language Processing (NLP) in the biomedical domain in order to improve biomedical information retrieval [14]. Drug-Drug Interaction (DDI) extraction is a significant biomedical sub-domain and a typical relation extraction task where drug mentions and their pair-wise interactions are extracted from text.

Traditional approaches tackle this task in a pipelined approach where each of the two tasks are completed subsequently, first recognizing drug mentions with Named Entity Recognition (NER) techniques, followed by the classification of the drug pairs with Relation Classification (RC) techniques. Joint approaches, consolidate both NER and RC tasks in a single model or methodology. However, despite the various approaches, state-of-the-art methodologies tend to utilize Transformer-based architectures as their underlying mechanisms. The significant advancements in natural language understanding and the development of Transformer-based Language Models (LM) have led to the application of pre-trained language representations in the DDI task [13]. Although, these methodologies show promising results, their application is subject to numerous challenges, such as choosing the most suitable one out of the overwhelming volume of proposed LMs that are trained on different corpora and determining the best output layer architecture.

To alleviate these limitations, we present a baseline approach for each of the two DDI extraction subtasks, proposing Transformer-based LMs with shallow output layer architectures. The task-specific pre-trained Language Models are fine-tuned on the benchmark DDI Extraction 2013 dataset and address the Drug Named Entity Recognition (DNER) and Drug-Drug Interaction Classification (DDIC) respectively. We leverage the underlying semantic and contextual representations in combination with the attention mechanism and shallow output architectures for both tasks, adding minimal complexity to the overall model. In summary, the main contributions of our work are:

- Individual Transformer-based Language Models for both Drug Named Entity Recognition and Relationship Classification tasks.
- Task-specific shallow output architectures for both DNER and RC that leverage the underlying Language Model without added complexity.
- Achievement of comparable to state-of-the-art performance on the DDI Extraction 2013 dataset.

The remainder of this paper is organized as follows. In Section 2, we provide a brief review of related work and explore the Transformer-based approaches used on the DDI extraction dataset. In Section 3, we elaborate on the dataset used and describe our methodology in detail. In Section 4, we present the experimental setup and results. Finally, in Section 5, we present our conclusions and the direction for future research.

2 RELATED WORK

Drug-drug interactions can occur when two or more drugs are co-administered resulting in possible synergetic or antagonistic effects that can prove harmful to the human health. To prevent the negative effects of polypharmacy, several drug interaction sources in the form of databases, individual product information and specialized online resources are available that require careful curation in order to be kept up to date with the latest research. Literature containing new findings in drug interaction related studies are published at an exponential rate, as clinicians are encouraged to publish case reports of new or unusual drug-drug interactions as they are valuable source of information.

The wealth of information hidden in these publications contain abbreviations, biomedical specific language and can have unusual grammatical structures. Consequently, building models that can understand and generate semantically and contextually useful representations is a challenge. Word embeddings established themselves as an essential foundation for learning these complex representations trained on large unlabeled corpora to capturing the implicit semantics.

In the drug-drug interaction domain, the task is treated as a biomedical Relation Extraction (RE) problem where drug mentions are extracted and each drug pair interaction is classified. Initially, Neural Network-based RE approaches leveraged word embeddings and applied Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) architectures in both tasks [2]. CNNs learn features by applying convolutional operations in the word embedding input space while RNNs are temporal-based networks that capture long sequences using an internal memory mechanism. Some studies combined CNNs and RNNs in a hybrid architecture for the classification of DDIs using sentence sequences and shortest distance paths or implement an ensemble method with CNNs, RNNs and Support Vector Machines [19].

Studies treat the drug named entity recognition and the relation classification as either two independent tasks (pipeline) or as a joint task, each subject to its own advantages and disadvantages. The pipelined approach tackles each step individually focusing on each task independently while joint approaches leverage the relevance between them. However, in the DDI domain the term extraction has been used as interchangeably with classification, where researchers focus exclusively on the DDIC task using the gold entity labels from the dataset, completely ignoring the DNER task.

Deep learning's rapid progress has led to the development of new methodologies and models that can be used neural network-based approaches to further improve the performance. Following the trend of general domain NLP, Transformer-based LMs found their way into the biomedical domain as well, achieving state-of-the-art results. In contrast to traditional word embeddings, LMs produce contextualized word representations where each word embedding represents the word based on the entire input sequence, making it depended on the context of the sentence it is contained.

Recent studies show that BERT [4] based architectures outperform other Transformer-based Language Models such as ELMo [15] and GPT-2 [16] in the DDI extraction task. Lee at al. [7] pre-trained BioBERT, which is the BERT architecture trained on large biomedical domain specific corpora, on biomedical NER and relation classification and question answering tasks, achieving a significant improvement in performance in all tasks. Zaikis et al. [24] proposed BERT based architectures for the pipelined DDI extraction using a set of rules and filters on the DNER models output before classifying the interactions between drug pairs. Huang et al. [6] implemented an ensemble of variants in a multi-head selection system that introduces semantically enhanced BERT pre-training and NER corpus and soft label embeddings. Xue et al. [22] implemented a joint NER and RC architecture where the shared task representation encoder is transformed with the use of a dynamic range attention mechanism.

Based on the recent trend in Transformer-based Language Model usage and the dominance of the BERT architecture, determining the best combination of the underlying BERT model and output architecture is very important. In this work, we present the most suitable baseline BERT based architectures with task specific shallow output layers for both Drug Named Entity Recognition and Relation Classification on the DDI Extraction 2013 dataset.

3 MATERIALS AND METHODS

3.1 Dataset

The DDI Extraction 2013 benchmark corpus, which is a semantically annotated corpus of documents containing sentences describing drug entities and drug-drug interactions from the DrugBank database and MedLine abstracts, was used to evaluate the performance of our proposed method. Two expert annotators manually annotated the corpus with pharmacological substances (drug named entities) and interactions between all possible drug pair combinations. The corpus is made up of 784 DrugBank documents describing drug interactions and 233 MedLine abstracts chosen from the query 'drug-drug interactions' and is split into a single training set and separate test sets for both Drug Named Entity Recognition and Relation Classification tasks.

The corpus contains four classes for the drug named entity types, labeled *drug, group, brand* and *drug_n*, denoting generic drug names, drug group names, branded drug names and active substances that are not approved for human use, respectively. Accordingly, there are four positive classes for the drug-drug interactions, labeled *advice, effect, mechanism* and *int*, describing a recommendation or advice, the impact of the drug-drug interaction, the pharmacokinetic mechanism and an interaction with no additional information, respectively. We arbitrarily assigned the label *no_rel* for the negative class describing the instances where no interaction is found. The corpus statistics make it evident that the dataset is extremely imbalanced in terms of both named entities and interaction types.

Our methodology embraces a minimal preprocessing approach in the form of tokenization only and no negative instance filtering. BERTs WordPiece tokenizer[21] with the cased vocabulary of each BERT-based model was employed for both tasks, representing words that are not in the vocabulary by the frequent subwords each token is made of.

3.2 Transformer architecture

The Transformer architecture [20] follows the encoder-decoder neural network structure without relying on recurrence or convolutions for output generation. Information about the relative positions of the words in the sequence can not be captured and has to be injected with the use of positional encodings of the input embeddings. Therefore, the input vector of the Transformer consists of the summed positional encodings, which are generated using sine and cosine functions of different frequencies, over the input embeddings.

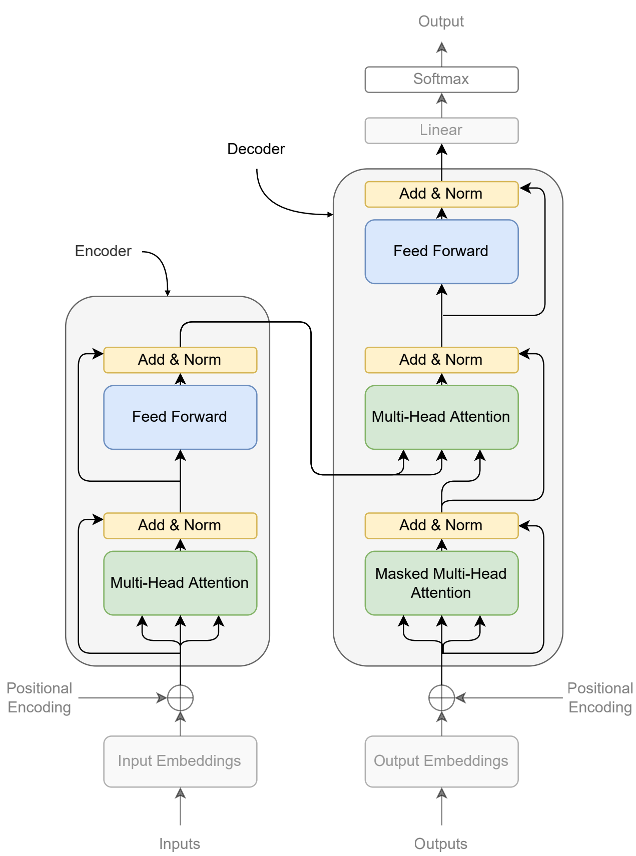

Figure 1: Overview of the Transformer architecture.

The encoder part maps an input sequence to a continuous representation that is passed to the decoder part along with the decoder output of the previous time step to generate the output sequence. The encoder consist of a variable stack of identical layers, where each layer is composed by a multi-head self-attention mechanism and a feed-forward layer with ReLU activation. Each attention layer head receives different linear projections of the queries, keys and values and produces an output simultaneously to generate the final output.

Similarly, the decoder part also consists of a variable stack of identical layers, each composed of three sublayers instead of two. The first sublayer injects the positional encodings in the output of the previous decoder stack and applies multi-head attentions over it. In contrast to the encoder where the attention mechanism attends to all words in the input sequence, the decoder attends only the preceding words. The following two sublayers are similar to the two layers of the encoder part, implementing a multi-head self-attentions and a feed forward network.

Each sublayer for both encoder and decoder parts also implement a residual connections around them and are followed by a normalization layer.

3.3 BERT-based language models

BERT (Bidirectional Encoder Representations from Transformers) is a Transformer-based Language Model designed to pre-train deep bidirectional representations from large unlabeled corpora by jointly conditioning on both directions, left and right, in all its layers. The BERT architecture aims to generate contextualized language representations and builds on top of the encoder part of the Transformer, implementing a stack of either 12 or 24 encoder layers, with 12 or 16 attention heads respectively.

Similarly to the Transformer architecture, the input representation is constructed by summing the corresponding token, segment and positional embedding. The token embeddings are generated by the WordPiece tokenizer that represents Out-Of-Vocabulary (OOV) words by frequent subwords while the positional embedding follows the same implementation as in the Transformers. Furthermore, BERT is able to take sentence pairs for tasks such as Question-Answering and utilizes the segment embeddings to distinguish between the two sentences and learn unique embeddings for both.

BERT was trained as a generalizable Language Model on a large unlabeled corpus form the English Wikipedia and BookCorpus on two tasks, namely Masked Language Model (MLM) and Next Sentence Prediction (NSP). In MLM the model is trained to predict randomly masked words in a sequence while in NSP the model receives pairs of sentences as input and learns to predict if the second sentence in the pair is the subsequent sentence in the original document. Both approaches allow the model to learn semantically and syntactically relevant word representations which are further improved when fine-tuning on a specific task.

While BERT improved on the performance of state-of-the-art approaches in a wide variety of tasks under the general language understanding, the performance in biomedical domain-specific literature related tasks was often poor. As a result, studies proposed new BERT-based models pre-trained on domain-specific corpora, such as biomedical tests, scientific publications and clinical notes,

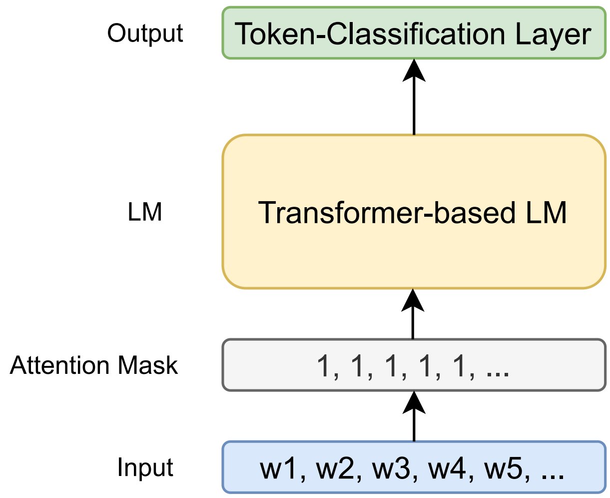

Figure 2: Overview of the Drug Named Entity Recognition architecture.

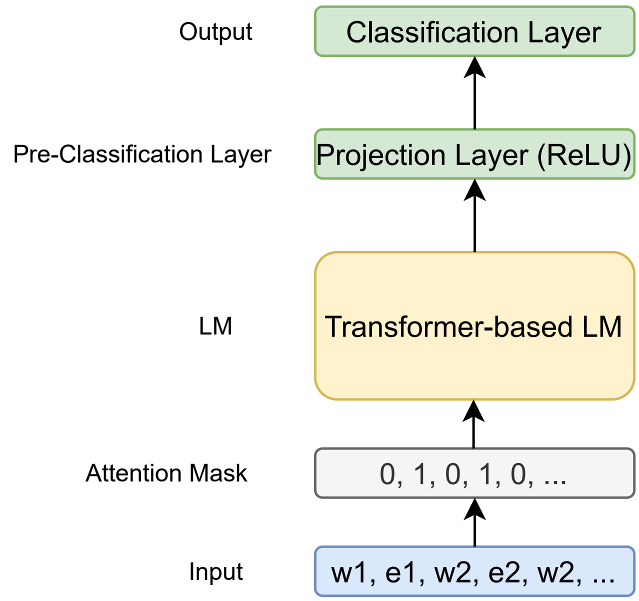

Figure 3: Overview of the Drug-Drug Interaction Classification architecture.

to alleviate these issues. Accordingly, the amount of newly proposed models, pre-trained on a variety of tasks, available to the researchers and the general public increased.

BERT-based models are the foundation of experimentation in this paper due to their versatility and their high performance in downstream tasks as proven by the trend in recent publications. Since BERT's appearance, multiple LMs had been proposed for DDI-specific task that implement a variety of architectures and is fundamental to be compared for a more solid benchmark work.

3.4 DDI Extraction Models

Drug-drug interaction extraction from textual data is a typical Relation Extraction task and conceptually consists of two basic steps, finding drug mentions and classifying the interaction of each pair in the context of the sentence. The recognition of drug names, which are task specific entities, is tackled with Named Entity Recognition techniques and the classification of the drug interactions into the task specific categories is tackled with text classification techniques, called DNER and DDI Classification respectively as illustrated in Figures 2 and 3 respectively.

3.4.1 Drug Named Entity Recognition. Drug Named Entity Recognition (DNER) is a sequence-to-sequence chunking task that classifies each token in a sequence of words based on a tagging scheme. The tagging scheme follows the Inside-Outside-Before-End-Single (IOBES) format where B, I and E denote the beginning, the inside parts and the end of a multi-token entity, correspondingly. The O tag indicates that the token is outside of an entity and does not belong to a chunk and the S tag represents a chunk containing a single token. Furthermore, each tag, with the exception of O, has a trailing label for each entity type, 'Drug', 'Group', 'Brand' and 'Drug_n' resulting in a total of 17 classes (i.e. 'I-Drug').

The three main components of the model are the input layer, the Transformer-based LM layers and the output layer. During WordPiece tokenization, OOV words are split based on known subwords, where each subsequent split start with '##' leading to alignment issues with the labels since each original token has an accompanying label. The label re-alignment can be implemented using three strategies, by labeling the splits the same label with the original token, by converting the token into a multi-token entity and using B, I and E or by keeping the label for the first token and labeling the rest of the splits with O. Each approach has its potential advantages and disadvantages but remain outside of the scope of this work.

The LM layers are initialized with their respective pre-trained weights and the output is fed to a Feed Forward layer with a softmax activation for the token classification.

3.4.2 Drug-Drug Interaction classification. Drug-Drug Interaction Classification (DDIC) is a relation classification task that classifies drug entity pairs into predefined categories in the context of the sentence they are contained in. Contrary to typical text classification, information on the drug entities is necessary for the model to correctly learn and accomplish this task. By leveraging the attention mechanism of the Transformer-based architectures, the models attention is directed to the drug entity pair in each sentence with the use of attention masks. The dataset was unrolled, creating multiple instances of each sentence when more than one pair was present, each instance with a unique attention mask vector attending to a single drug pair each time.

Similar to the DNER model, the DDIC models' three main components are the input, Transformer-based LM and output layers. Furthermore, during the tokenization instead of aligning the labels, the attention mask vector is aligned to correctly identify the position of each entity following the same strategy as aforementioned. The concatenated outputs of the LMs pooling layer and the entity specific hidden states are passed to a pre-classification layer with ReLU activation and then to a final classification layer. The LM layers are initialized with their respective weights and the pre-classification layers' output is passed to the final Feed Forward classification layer, generating the class probabilities of the interaction with a softmax activation.

4 RESULTS AND DISCUSSION

4.1 Experimental setting

We trained both DNER and DDIC models for drug mention extraction and DDI classification respectively on the 6976 sentences from 714 abstract documents from both DrugBank and MedLine using the predefined train and test splits. We evaluated each model separately, using the provided DNER and RC test sets and used a subset of the train set to create the validation set for hyperparameter-tuning for each task. Specifically, we evaluated the DNER module with a test set consisting of 665 sentences contained in 112 abstract documents and the Relationship Classification module on 1299 sentences in 181 abstract documents. We experimented with pre-trained weights from BERT as the baseline BERT-based Language Model and carefully selected BERT-based models on both base and large architectures where available.

The base architectures stack 12 encoder layers with a hidden layer size of either 512 or 768, while the large architectures stack 24 encoder layers with a hidden layer size of usually 1024. Padding of the sentence lengths is determined by calculating $padding\_length = mean_{train} + 2 * stdev_{train}$ since the training set sentence lengths follow normal distribution. However, experiments with different maximum lengths were conducted for completeness.

The adaptive optimizer AdamW [11] was used to decouple the weight decay from the optimization step, allowing for separate optimization. The models were trained with a batch size of either 32 or 64, with a learning rate of 0.001 and decay of 0.01 each epoch and the optimal number of epochs was evaluated based on convergence during training. Furthermore, to achieve the task-specific Relation Classification, we utilize the Transformer-based Language Models' attention mechanism to direct the focus to the sentence context in relation to the specific entity pair in each unrolled instance. Finally, Cross Entropy and Binary Cross Entropy loss were used for the DNER and DDIC model respectively.

The following pre-trained models were selected based on their relevance to the Relation Extraction task for both DNER and DDIC models and are comprised mainly of BERT-based architectures with the exception of Electra and XLNet.

- BioBERT [8] is a domain-specific language representation model pre-trained on large-scale biomedical corpora from PubMed and PubMed Central with almost the same architecture across tasks compared to BERT.
- SciBERT [1] similarly, is a domain-specific language representation model pre-trained on scientific texts from Semantic Scholar and has its own WordPiece vocabulary to match the training corpus.
- BART [9] is a transformer encoder-encoder (sequence-to-sequence) model with a bidirectional (BERT-like) encoder and an autoregressive (GPT-like) decoder. BART is pretrained by corrupting text with an arbitrary noising function, and learning a model to reconstruct the original text.
- RoBERTa [10] iterates on BERT's pre-training procedure, including training the model longer, with bigger batches over more data; removing the next sentence prediction objective; training on longer sequences; and dynamically changing the masking pattern applied to the training data.
- DeBERTa [5] improves the BERT and RoBERTa models using disentangled attention and enhanced mask decoder. It outperforms BERT and RoBERTa on the majority of Natural Language Understanding tasks with significantly less training data.
- DistilBERT [17] is a Transformer-based model, smaller and faster than BERT, which was pre-trained on the same corpus in a self-supervised fashion, using the BERT model as a foundation, that can then be fine-tuned with good performances on a wide range of tasks like its larger counterparts.
- Electra [3] introduced a replaced token detection task for better language representation learning, trained to distinguish input tokens from high-quality negative samples with a Generator and Discriminator architecture.
- XLNet [23] is a unsupervised language representation learning method based on a novel generalized permutation language modeling objective. Additionally, XLNet employs Transformer-XL as the backbone model, exhibiting excellent performance for language tasks involving long context.

We implemented our models with the PyTorch library using the Python programming language and conducted the experiments on a computer with a single RTX 3090 24GB graphics card and a 12-core Intel CPU.

4.2 Experimental results

To evaluate the performance on each task, similar to the related work, we used the micro F1-score, which is the micro-averaged harmonic mean of the Precision and Recall scores. We conducted extensive experiments with the pre-trained models and the hyperparameters for both DNER and DDIC tasks, investigating each model and presenting the best combination of maximum sentence length, batch size and number of epochs for each model on the DDI Extraction 2013 dataset [18], as shown in Table 1. The results show that fine-tuned Transformer-based Language Models are able to achieve very good performance in both Named Entity Recognition and Relation Classification, further validating their generalizability on the various NLP tasks.

We observed that the pre-training domain of each LM in combination with our proposed shallow output architecture, contributed further in achieving better results. However, it is notable that the large architectures show a decrease in performance compared to their base counterparts in the DNER task, attributed to the relatively small dataset and to the need for more training epochs to learn better representations in these architectures resulting, however, in overfitting. In the DDIC task, the use of the large architectures did not contribute to any performance improvements across all models with the exception of XLNet where the XLNet-large increased the F1-score of XLNet-base by 2%. Furthermore, the increase in training time coupled with the added complexity of the additional

Transformer-based layers and of the larger hidden state vectors have the potential of statistical insignificant increase in the overall score. All models utilized a cased vocabulary, where the capitalization of the letters is accounted for since it is of importance in NER tasks in general and in biomedical domain related tasks especially.

In the DNER task, the general domain BART-base and RoBERTa-base models achieved the best performance with a F1-score of 98,2% overall. BART is initially trained by corrupting text with a noising function and learning to reconstruct the original text, improving on its natural language comprehension ability. RoBERTa replicates the BERT architecture and implements small embeddings tweaks and byte-level tokenizer with a different pre-trained scheme. In the DDIC task, the scientific domain SciBERT achieves the best performance with a F1-score of 82,2%, followed by BioBERT-large with a F1-score of 81,6%. Both SciBERT and BioBERT follow the original BERT architecture, trained on scientific data and biomedical data respectively.

The experimental results show that for the DNER task, which is a token level classification task, there are slight negative effects on all models when using domain-specific pre-trained architectures. The best performance improvement is obtained when fine-tuning pre-trained BERT-based models. In DDIC, in contrast to DNER, the performances of the models are enhanced by the weights obtained though pre-training on domain-specific corpora, further indicating the importance of pre-training on task-specific corpora for the Relation Classification task.

To further demonstrate the performance of the pre-trained LMs, we compared the best results for DNER and DDIC with state-of-the-art approaches that tackle both DNER and DDIC tasks, 'Att-BilSTM-CRF + Elmo' [12] and 'TP-DDI-large' [24], as shown in Table 2. First, the results show that the optimal approach to each task manages to outperform the previous models in the DNER task while matching the performance in the DDIC task. However, one important factor to consider is that the reported scores of the state-of-the-art approaches tackle the end-to-end task of DDI Extraction and report the results based on extracted entities in contrast to our approach where tackle each task individually and evaluate both models on the benchmark labels.

5 CONCLUSION

In this paper, we present a baseline approach for each of the two DDI extraction subtasks, Named Entity Recognition and Relation Classification utilizing Transformer-based Language Models with shallow output layer architectures. The pre-trained LMs are fine-tuned and evaluated on the benchmark DDI Extraction 2013 dataset and address the drug named entity recognition and interaction classification respectively. We leveraged the underlying semantic and contextual representations with shallow output architectures for both tasks, utilizing the attention mechanism to inject entity information in the DDI classification task.

We evaluated the performance of the pre-trained Transformer-based models on the DDI Extraction 2013 dataset and compared the best performing models to other state-of-the-art methods. Our experiments show that the pre-trained models with task-specific shallow output layers manage to achieve near state-of-the-art scores by taking advantage of the underlying architecture and the right hyper-parameters, validating the assumption that Transformer-based LMs achieve great performance and need minimal fine-tuning. With this work, we believe a baseline approach can be set to facilitate further improvements on this important domain to aid drug safety dissemination and research. As future work, we plan to extend the study to end-to-end systems, focusing on the mitigation of errors between the two tasks when the interaction classification is carried out on previously predicted drug entities.

REFERENCES

[1] Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: Pretrained Language Model for Scientific Text. In *EMNLP*. arXiv:arXiv:1903.10676

[2] Priyankar Bose, Sriram Srinivasan, William C. Sleeman, Jatinder Palta, Rishabh Kapoor, and Preetam Ghosh. 2021. A Survey on Recent Named Entity Recognition and Relationship Extraction Techniques on Clinical Texts. *Applied Sciences* 11, 18 (2021). https://www.mdpi.com/2076-3417/11/18/8319

[3] Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. In *ICLR*. https://openreview.net/pdf?id=r1xMH1BtvB

[4] Jacob Devlin, Ming Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference* 1 (2019), 4171–4186. Issue Mlm. https://arxiv.org/pdf/1810.04805.pdf

[5] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. DeBERTa: Decoding-enhanced BERT with Disentangled Attention. (6 2020).

[6] Weipéng Huáng, Xingyi Cheng, Taifeng Wang, and Wei Chu. 2019. BERT-Based Multi-Head Selection for Joint Entity-Relation Extraction. In *NLPCC*.

[7] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* (9 2019). https://doi.org/10.1093/bioinformatics/btz682

[8] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. (1 2019). https://doi.org/10.1093/bioinformatics/btz682

[9] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. (10 2019).

[10] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. (7 2019).

[11] Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*. https://openreview.net/forum?id=Bkg6RiCqY7

[12] Ling Luo, Zhihao Yang, Mingyu Cao, Lei Wang, Yin Zhang, and Hongfei Lin. 2020. A neural network-based joint learning approach for biomedical entity and relation extraction from biomedical literature. *Journal of Biomedical Informatics* 103 (2020), 103384. https://doi.org/10.1016/j.jbi.2020.103384

[13] Dinh Phuong Nguyen and Tu Bao Ho. 2020. Drug-Drug Interaction Extraction from Biomedical Texts via Relation BERT. (2020), 1–7. https://doi.org/10.1109/RIVF48685.2020.9140783

[14] Daniel W. Otter, Julian R. Medina, and Jugal K. Kalita. 2021. A Survey of the Usages of Deep Learning for Natural Language Processing. *IEEE Transactions on Neural Networks and Learning Systems* 32, 2 (2021), 604–624. https://doi.org/10.1109/TNNLS.2020.2979670

[15] Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2227–2237. https://doi.org/10.18653/v1/N18-1202

[16] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.

[17] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. (10 2019).

[18] Isabel Segura-Bedmar, Paloma Martínez, and María Herrero-Zazo. 2013. SemEval-2013 Task 9 : Extraction of Drug-Drug Interactions from Biomedical Texts (DDIExtraction 2013). (June 2013), 341–350. https://aclanthology.org/S13-2056

Table 1: Comparison of pre-trained weights with our proposed shallow output architecture for Drug Named Entity Recognition (DNER) and Drug-Drug Interaction Classification (DDIC). Bold emphasis denotes the best performances and the second best are underlined.

Pre-trained model	Task	Max Length	Batch Size	Epochs	F1 Score	Precision	Recall
BERT-base	DNER	60	32	4	96,9	96,4	97,3
	DDIC	70	64	3	81,2	79,1	83,5
BERT-large	DNER	70	32	7	95,6	96,4	94,8
	DDIC	70	32	1	81,2	79,2	83,4
BioBERT-base	DNER	50	32	7	96,7	97,1	96,2
	DDIC	70	32	2	81,2	79,1	83,5
BioBERT-large	DNER	50	32	7	96,7	95,4	97,4
	DDIC	70	64	1	<u>81,6</u>	80,2	83,8
SciBERT	DNER	60	64	3	<u>98,1</u>	98,1	98,1
	DDIC	70	64	1	**82.4**	81,2	84,4
BART-base	DNER	70	64	5	**98,2**	98,2	98,2
	DDIC	70	64	2	81,5	80,4	83,6
BART-large	DNER	50	32	2	95,8	94,5	97,2
	DDIC	70	32	2	81,6	80,5	83,7
RoBERTa-base	DNER	70	64	5	**98,2**	98,2	98,2
	DDIC	70	32	1	81,5	80,4	83,6
RoBERTa-large	DNER	70	32	7	96,9	96,5	97,4
	DDIC	70	32	1	81,5	80,4	83,6
DeBERTa-base	DNER	50	32	4	<u>98,1</u>	98,1	98,1
	DDIC	70	32	2	80,3	78,1	82,5
DeBERTa-large	DNER	70	32	4	96,3	96,4	97,7
	DDIC	70	32	2	80,3	78,1	82,5
DistilBERT	DNER	70	64	4	96,2	98,1	94,5
	DDIC	70	32	2	79,5	78,2	80,8
Electra	DNER	50	32	8	96,3	94,6	98,1
	DDIC	70	32	2	80,1	79,4	80,8
XLNet-base	DNER	70	32	10	94,3	93,9	94,7
	DDIC	70	32	3	77,9	80,1	79,7
XLNet-large	DNER	60	32	6	95,4	96,2	94,5
	DDIC	70	32	2	79,1	77,9	80,4

Table 2: Comparison of the best performing pre-trained models with state-of-the-art. The best scores are denoted with bold.

	DNER			DDIC		
Model	F1	P	R	F1	P	R
Att-BilSTM-CRF + Elmo	89,5	93,2	86,1	75,1	75,0	75,2
TP-DDI-large	97,1	97,4	96,8	**82,4**	86,4	78,8
BART-base	**98,2**	98,2	98,2	81,5	80,4	83,6
SciBERT	98,1	98,1	98,1	**82,4**	81,2	84,4

[19] Víctor Suárez-Paniagua, Renzo M. Rivera Zavala, Isabel Segura-Bedmar, and Paloma Martínez. 2019. A two-stage deep learning approach for extracting entities and relationships from medical texts. *Journal of Biomedical Informatics* 99 (2019), 103285. https://doi.org/10.1016/j.jbi.2019.103285

[20] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. (2017). arXiv:1706.03762 [cs.CL]

[21] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. (2016), 1–23. arXiv:1609.08144 http://arxiv.org/abs/1609.08144

[22] Kui Xue, Yangming Zhou, Zhiyuan Ma, Tong Ruan, Huanhuan Zhang, and Ping He. 2019. Fine-tuning BERT for Joint Entity and Relation Extraction in Chinese Medical Text. *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (2019), 892–897.

[23] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. XLNet: Generalized Autoregressive Pretraining for Language Understanding. (6 2019).

[24] Dimitrios Zaikis and Ioannis Vlahavas. 2021. TP-DDI: Transformer-based pipeline for the extraction of Drug-Drug Interactions. *Artificial Intelligence in Medicine* 119 (2021), 102153. https://doi.org/10.1016/j.artmed.2021.102153