Joint Extraction of Clinical Entities and Relations Using Multi-head Selection Method

Xintao Fang
Graduate School of Information Science and Engineering
Ritsumeikan University
Kusatsu, Japan
gr0475vx@ed.ritsumei.ac.jp

Yuting Song, Akira Maeda
College of Information Science and Engineering
Ritsumeikan University
Kusatsu, Japan
songyt@fc.ritsumei.ac.jp, amaeda@is.ritsumei.ac.jp

***Abstract*—The extraction of entities and relations from unstructured clinical records has been attracting increasing attention. In addition to the existing traditional methods, deep learning methods have also been proposed for entity and relation extraction. However, previous work on clinical entity and relation extraction did not consider the multiple relations between clinical entities, which often exist in clinical texts. To deal with multiple relations, we propose using a multi-head selection method for clinical entity and relation extraction. As pre-trained language models have been shown to be effective for clinical entity and relation extraction, we integrate a pre-trained language model with a multi-head model to jointly extract clinical entities and relations. The experimental results show that the proposed model is effective for entity and relation extraction on both the i2b2/VA 2010 and n2c2 2018 challenge datasets and outperforms the top-ranking systems in the n2c2 2018 challenge. We also evaluate the impact of four existing pre-trained language models on clinical entity and relation extraction performance. The domain-specific pre-trained language model improves the performance of clinical entity and relation extraction. Between BERT and CharacterBERT, which uses a Character-CNN module instead of BERT's wordpiece system to represent entire words, we find that BERT outperforms CharacterBERT on joint extraction of clinical entities and relations.**

***Index Terms*—Clinical record, entity recognition, relation extraction, multi-head selection, pre-trained language model**

I. Introduction

Electronic health records (EHRs) contain large amounts of free-text data which includes detailed records about clinical events. However, this information exists in unstructured text. There has been increasing interest in applying natural language processing (NLP) technologies to extract structured information from EHRs. Extracting available information including entities and their relations is helpful for supporting clinical research and applications such as constructing clinical knowledge graphs.

To extract entities and their relations, traditional pipeline methods [1], [21] first identify entity mentions and then classify relations for the entity pairs. This separated framework makes the task more flexible and easy to deal with. However, it ignores the interaction and correlation between the two sub-tasks and leads to erroneous delivery because each sub-task has an independent model. Due to the drawbacks of the pipeline method, later studies focused more on building joint models to simultaneously extract entities and relations [3], [16], [20], [26], [30]. The joint methods mitigate the aforementioned issues and have achieved state-of-the-art performance on entity and relation extraction.

Entity and relation extraction have also been thoroughly studied in the clinical domain for a long time [12], [13], [25], [27], [28]. For example, the i2b2/VA 2010 challenge presented the tasks of extracting medical concepts (medical problems, tests, and treatments) from patient reports and identifying the relations between the medical concepts. Most proposed methods extract clinical entities and their relations with a pipeline model, while few methods utilize the joint model. Moreover, previous work on clinical entity and relation extraction has not considered the multiple relations between clinical entities, which often exist in clinical texts. As shown in the example in Figure 1, the test "electrocardiogram" revealed two medical problems, "ST elevations" and "J elevations."

We propose extracting clinical entities and their relations by utilizing the multi-head selection-based joint model [3], which can deal with the multi-relations. Given the success of pre-trained language models (e.g., BERT [7], CharacterBERT [8]), we using a pre-trained language model with the multi-head model to jointly extract clinical entities and relations. Specifically, o ur m odel u ses t he p re-trained l anguage model as the encoder layer. Then, we use a conditional random fields (CRF) layer for clinical entity recognition and a sigmoid layer for the relation extraction.

To evaluate the effectiveness of our method, we conducted experiments on two clinical datasets provided by the n2c2 2018 and i2b2/VA 2010 challenges. The results show that our method outperforms the top-ranking systems in the n2c2 2018 challenge. We found that our method correctly extracted 71.8% of the multiple relations in the n2c2 dataset. We also evaluated the impact of four existing pre-trained language models. The results demonstrated that the use of a domain-specific pre-trained language model improves the accuracy of joint extraction of clinical entities and relations. Between BERT and CharacterBERT, which uses a Character-CNN module instead of BERT's wordpiece system to represent entire words, we found that BERT outperforms CharacterBERT on the joint extraction of clinical entities and relations.

II. Related work

We first present previous work on entity and relation extraction in the general domain. Because our work focuses on extracting entities and relations from clinical texts, we also present prior work on entity and relation extraction in the medical domain. In addition, we discuss the development of the pre-trained language model BERT and its application in medical domain tasks.

A. Entity and relation extraction in general domain

Two main frameworks are used for entity and relation extraction tasks: the pipeline method [4], [10], [11] and the joint method [2], [19], [20]. The pipeline method treats the task as a pipeline of two sub-tasks (i.e., named entity recognition (NER) and relation classification). This separated framework makes the task easy to deal with, but it ignores the relevance between the two sub-tasks because each sub-task has an independent model. Moreover, the results of entity recognition may affect the performance of relation classification, which leads to erroneous delivery. In contrast, the joint method is expected to simultaneously extract entities and their relations from texts, which helps to mitigate the aforementioned erroneous delivery issues. Early joint models [16], [20] require manually designed features or existing NLP tools (e.g., POS tagger) [14], [24]. With the development of deep learning, CNN-based and RNN-based models have been proposed to jointly extract entities and their relations [16], [19], [30]. Since the emergence of the BERT model, the BERT-based method [23] has achieved state-of-the-art results for entity and relation extraction.

B. Entity and relation extraction in medical domain

Despite the success of the joint method in the general domain, most related studies have preferred the pipeline method for clinical entity and relation extraction. The 2018 n2c2 challenge[1] presented the task of building an end-to-end system that processes raw narrative text to discover the concepts related to medications and adverse drug events and find the potential relations between them. The participants in this challenge proposed deep/machine learning-based methods with hand-crafted features [5], [15]. Although their methods have shown to be effective for extraction, there is still room for improvement as their models did not deal with multiple relations related to one entity.

C. Pre-trained language model

BERT is a multi-layer bidirectional transformer-based language representation model [7] which is pre-trained on a large unlabeled corpus and can be fine-turned on various NLP tasks. BERT has greatly improved the accuracy of most NLP tasks. The BERT model has become the most widely used for building NLP systems, and many variants of BERT model have been proposed [6], [17], [29]. To facilitate research on language representations in a specific domain, some studies [9], [23] initialized BERT with pre-trained BERT provided by Devlin et al. (2018) [7], then continued to pre-train the model using the domain-specific corpus. Peng et al. (2019) [23] found that continuing to pre-train the BERT model on medical corpuses can improve the accurarcy of several medical NLP tasks, such as clinical entity recognition. El Boukkouri et al. (2019) [9] focused on improving language representations in the default "general-domain" to make word embeddings suitable for specialized domains. El Boukkouri et al. (2020) [8] showed that BERT's sub-word tokenization system which relies on a predefined set of wordpiece may not be suitable for the special-domain tasks. They proposed CharacterBERT, which replaces the wordpiece part of $\mathrm{BERT}_{\mathrm{general}}$ model with Character-CNN module. The authors evaluated CharacterBERT on multiple medical domain tasks and attained state-of-the-art performance. Inspired by the success of CharacterBERT and the BERT re-trained on medical domain corpuses, we integrate four released BERT and CharacterBERT models[2] as the encoder of our model to investigate their performance on the the task of entity and relation extraction in clinical domain.

[1]https://portal.dbmi.hms.harvard.edu/projects/n2c2-nlp

[2]https://github.com/helboukkouri/character-bert

III. Proposed method

In this section, we introduce our multi-head selection model integrated with pre-trained language model for clinical entity and relation extraction. We follow the work of Bekoulis et al. (2018) [3] who proposed a joint neural model that formulates the relation extraction as a multi-head selection problem. Because a clinical entity may be involved in multiple relations with other entities (e.g., a treatment entity may cause multiple medical problems), we utilize the multi-head selection model to extract clinical entities and their relations. In addition, we use the pre-trained language model as the encoder and fine-tune it with the multi-head selection model for clinical entity and relation extraction. Our model, shown in Figure 1, consists of a pre-trained language model as an encoding layer to output word representations, a CRF layer for clinical entity recognition, and a sigmoid layer for relation extraction.

A. Encoding layer based on pre-trained language model

Unlike Bekoulis et al.'s method [3], which uses bidirectional LSTM as the encoding layer, we utilize the BERT model as the encoding layer. Because the BERT model has a series of variants that can be better adapted to tasks in the medical domain, we use four different pre-trained models [8] as the encoding layer of our model to obtain contextual word representations.

B. CRF layer for clinical entity recognition

A CRF layer is added to the top of the pre-trained language model for clinical entity recognition. The word embeddings obtained from the pre-trained language model are used as the input for the CRF layer. The output of CRF layer is the predicted probability of which category the current word belongs to. In training mode, we use the actual labels for all

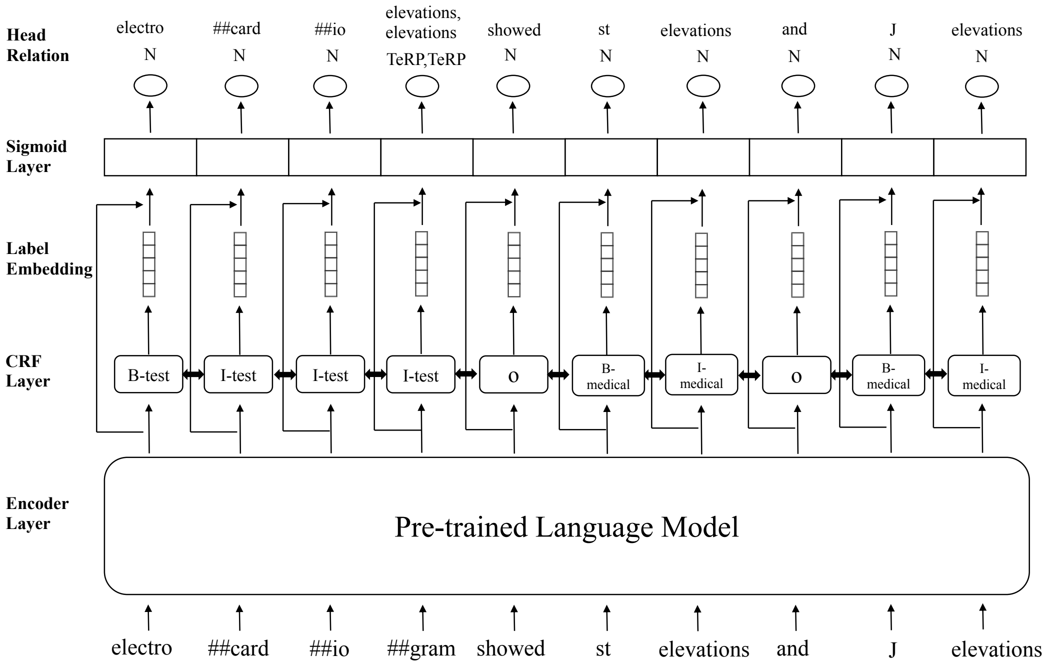

Fig. 1. Proposed multi-head selection model integrated with pre-trained language model for joint extraction of clinical entities and relations.

the words as the input for the next step. In predicting mode, we use the predicted label as the input for the next step.

C. Sigmoid layer for relation classification

In the multi-head selection model, the relation extraction is regarded as a multi-head selection problem. Each word may have multiple relations with other words. The input of the sigmoid layer here is the combination of the word embeddings of the output of pre-trained language model and the vectors of labels that are generated by the CRF layer. For the input token sequence W and a set of relations R, our goal is to identify for each token w_{i} ($w_{\mathrm{i}} \in W$), the most probable head w_{j} ($w_{\mathrm{j}} \in W$), and the most probable corresponding relation label r_{k} ($r_{\mathrm{k}} \in R$). Because the entity may have multiple relations, we use the sigmoid layer to predict relations, and we minimize the cross-entropy loss $Ł_{rel}$ (in Eq. 1) during training phase, where m represents the length of the input sequence.

$$Ł_{rel} = \sum_{i=0}^{m} \sum_{j=0}^{m} -logP(head = w_j, relation = r_{\mathrm{k}} | w_i) \quad (1)$$

For joint extraction of clinical entities and relations, we set the final loss as the sum of the entity recognition loss and five times of the relation classification loss.

IV. Experiments

A. Dataset

We conducted experiments on two clinical datasets provided by the n2c2 2018 [12] and the i2b2/VA 2010 [25] challenges[3].

[3]https://www.i2b2.org/NLP/DataSets/Main.php

- n2c2 2018: This dataset consists of 505 discharge summaries drawn from the Medical Information Mart For Intensive Care III (MIMIC-III) clinical care database. The training data contains 303 discharge summaries and the test data contains 202 discharge summaries. Table I shows the statistics of the n2c2 dataset. This dataset contains nine types of clinical entities and eight types of relations. In the experiments, we randomly split the training data into training and validation sets at a ratio of 4:1.
- i2b2/VA 2010: The original dataset consists of 871 discharge summaries. However, part of the original dataset is not available to the public, so we could only download a subset of the original dataset from the i2b2/VA 2010 challenge website. The available dataset consists of 426 discharge summaries (170 training data and 256 test data). This dataset contains three types of clinical entities (medical problem, test, and treatment) and eight types of relations. In our experiments, we randomly split the 426 discharge summaries into training, validation, and test datasets at a ratio of 3:1:1. Table II shows the statistics of the i2b2 dataset.

TABLE I
NUMBER OF REPORTS, SENTENCES IN N2C2 DATASET

Dataset	Reports	Sentences
Training	303	9105
Test	202	6106

TABLE II
NUMBER OF REPORTS, SENTENCES IN I2B2 DATASET

Dataset	Reports	Sentences
Training	256	3024
Validation	85	935
Test	85	1105

B. Evaluation metrics

We used the exact matching score as our metric for evaluating the experimental results. An entity is considered correct if the boundaries and the entity type are both correct, and the relation is correct when the relation type and the boundaries of the entities are both correct. The experimental results are evaluated on precision, recall, and F1 score.

C. Implementation details

We implemented our model by using PyTorch [22]. The experiments were carried out on a NVIDIA Quadro RTX8000 (GPU). For the n2c2 dataset, we set the maximum sequence length to 300. We used the AdamW optimizer [18] with a learning rate and batch size of 5e-4 and 32, respectively. For the i2b2 dataset, we set the maximum sequence length to 300. We used the AdamW optimizer with a learning rate and batch size of 1e-3 and 32, respectively. We applied a dropout layers of 0.1 to all task-specific layers. We conducted the evaluation for 100 epochs and selected the optimal epoch on the basis of the results on the validation set.

D. Results

To identify the impact of pre-trained language models on clinical entity and relation extraction, we integrated the following four released pre-trained language models into our model respectively. The experimental results are shown in Table III and Table IV.

- $\text{BERT}_{\text{general}}$: A general-domain model obtained by pre-training BERT on a general corpus. It uses the same architecture and wordpiece vocabulary as BERT (base, uncased) [7].
- $\text{CharacterBERT}_{\text{general}}$: A general-domain model obtained by training CharacterBERT on a general corpus. CharacterBERT replaces the wordpiece part of $\text{BERT}_{\text{general}}$ with Character-CNN module.
- $\text{BERT}_{\text{medical}}$: A medical model obtained by re-training $\text{BERT}_{\text{general}}$ on a medical corpus.
- $\text{CharacterBERT}_{\text{medical}}$: A medical model obtained by re-training $\text{CharacterBERT}_{\text{general}}$ on a medical corpus. This is the Character-CNN analog of $\text{BERT}_{\text{medical}}$.

For the n2c2 dataset, we compared the top-ranking systems participators in the n2c2 challenge. The top ranked team, UTHealth//Dalian team (UTH) [27] proposed a BiLSTM-CRF-based joint method to extract entities and relations. Other teams used the pipeline methods. The University of Florida team (UFL) [28] proposed a pipeline method that uses BiLSTM-CRF for entity extraction and SVM for relation classification. The NaCT team [13] also proposed a pipeline method that uses hand-crafted feature-based CRF for entity and relation extraction. The results show that our method outperformed these top-ranking systems. Both BERT and CharacterBERT re-trained on medical corpuses performed more effectively than the top-ranking systems, and the BERT-based models outperformed CharacterBERT-based models on both the n2c2 dataset and the i2b2 dataset.

V. DISCUSSION

A. Multi-relation issues

To investigate the effectiveness of our method for solving the multi-relation problem, we counted the number of sentences containing multiple relations. For the n2c2 dataset, 1709 out of 2387 (71.8%) sentences containing multiple relations were correctly extracted by our method. Some of the sample sentences are shown in Table V. We listed the original text and their golden labels and the prediction by our method. The triplet shows the extracted entities and their relations.

B. General domain vs. Medical domain

As shown in Table III and Table IV, for both the n2c2 and i2b2 datasets, the medical domain pre-trained language models ($\text{BERT}_{\text{medical}}$ and $\text{CharacterBERT}_{\text{medical}}$) were more effective than the general domain pre-trained language models ($\text{BERT}_{\text{general}}$ and $\text{CharacterBERT}_{\text{general}}$). This indicates that the pre-trained language model that is re-trained on the medical domain corpuses can improve the performance of clinical entity and relation extraction.

C. BERT vs. CharacterBERT

We conducted experiments with BERT and CharacterBERT to compare their effectiveness for clinical entity and relation extraction tasks. The results show that the BERT-based model was more effective than CharacterBERT on entity and relation extraction, contrary to the results of El Boukkouri et al. [8].

D. Error analysis

To further examine the performance of our model, we analyzed the results of the most effective $\text{BERT}_{\text{medical}}$-based multi-head selection model. Table VI summarizes the following two errors and provides samples of both.

- The system does not extract correct entities and relation types.
- The system makes additional incorrect predictions.

VI. CONCLUSION AND FUTURE WORK

In this paper, we proposed a multi-head selection model with pre-trained language models for clinical entity and relation extraction. Our method achieved a better result than the top-ranking systems that participated in the n2c2 2018 challenge on the clinical entity and relation extraction task. We also investigated the impact of pre-trained language models on this task. The $\text{BERT}_{\text{medical}}$-based model yielded the highest F1 score in our experiments on the n2c2 and i2b2 datasets. Furthermore, the results showed that pre-trained language model re-trained on medical domain corpuses outperformed

TABLE III
CLINICAL ENTITY RECOGNITION AND RELATION EXTRACTION RESULTS FOR N2C2 DATA

Model	Clinical entity recognition			Relation extraction		
	P	R	F1	P	R	F1
UTH [27]	-	-	89.03	-	-	81.97
UFL [28]	-	-	88.10	83.37	77.73	80.48
NaCT [13]	-	-	88.05	-	-	80.43
$\text{BERT}_{\text{generral}}$	92.26	92.44	92.35	82.37	79.49	80.91
$\text{CharacterBERT}_{\text{general}}$	91.41	91.82	91.61	81.54	80.08	80.81
$\text{BERT}_{\text{medical}}$	93.92	94.26	93.59	84.80	82.16	83.45
$\text{CharacterBERT}_{\text{medical}}$	92.30	93.29	92.79	81.79	84.01	82.89

TABLE IV
CLINICAL ENTITY RECOGNITION AND RELATION EXTRACTION RESULTS FOR I2B2 DATA

Model	Clinical entity recognition			Relation extraction		
	P	R	F1	P	R	F1
$\text{BERT}_{\text{generral}}$	84.45	86.07	85.25	51.93	41.05	45.86
$\text{CharacterBERT}_{\text{general}}$	83.35	83.86	83.60	49.31	42.79	45.82
$\text{BERT}_{\text{medical}}$	91.77	91.74	91.76	57.15	53.08	55.04
$\text{CharacterBERT}_{\text{medical}}$	88.36	87.38	87.87	54.22	53.80	54.01

TABLE V
EXAMPLES OF MULTIPLE RELATIONS CORRECTLY EXTRACTED BY OUR METHOD

Dataset	Text	Gold	Prediction
n2c2	**toprol xl 50mg daily**	(**50mg**, Strength-Drug, **toprol xl**), (**daily**, Frequency-Drug, **toprol xl**)	(**50mg**, Strength-Drug, **toprol xl**), (**daily**, Frequency-Drug, **toprol xl**)
i2b2	**electrocardiogram** at that time showed **st elevations** and **j point elevations in v2 through v4**	(**electrocardiogram**, TeRP, **st elevations**), (**electrocardiogram**, TeRP, **j point elevations in v2 through v4**)	(**electrocardiogram**, TeRP, **st elevations**), (**electrocardiogram**, TeRP, **j point elevations in v2 through v4**)

TABLE VI
EXAMPLES OF INCORRECT EXTRACTIONS BY OUR METHOD

Error	Example	Explanation
Incorrect entities and relation types	**Text**: she is status post 5 doses of iv chemotherapy with intraperitoneal cisplatin. **Gold**:(chemotherapy, Route-Drug, iv) **Prediction**:(chemotherapy, Dosage-Drug, 5 doses) **Text**: he was noted to be in rapid af on several episodes receiving 5mg iv lopressor each time with improvement. **Gold**: (5mg iv lopressor, TrIP, rapid af) **Prediction**: (lopressor, TrAP, rapid af)	Due to the entity prediction error, the relation type was predicted incorrectly
Additional incorrect prediction	**Text**: pt with two episodes of hypoglycemia (fsbs 22 and 27), etiology unclear, pt followed by strickland, who recommend no changes to current insulin regimen **Gold**: (fsbs, TeRP, hypoglycemia), **Prediction**: (current insulin regimen, TeRP, hypoglycemia), (fsbs, TeRP, hypoglycemia) **Text**: her son reports that she then developed a headache and fevers started three days ago which were treated with tylenol. **Gold**:(tylenol, TeRP, a headache), (tylenol, TeRP, fevers) **Prediction**: (tylenol, TeRP, a headache) (tylenol, TeRP, fevers), (a headache, TrWP, fevers)	In addition to the correct entities and relations, the system predicts other incorrect relations.

the pre-trained language model trained on general domain corpuses. We also compared the performance of using CharacterBERT vs. BERT as the encoder layer of our multi-head selection model and found that the BERT model outperformed CharacterBERT in clinical entity and relation extraction.

In the future, we plan to focus on the following research directions to improve the performance on clinical entity and relation extraction:

- We plan to integrate BERT trained on a medical domain corpus from scratch and add medical specialized vocabulary into our model so that we can identify whether the medical specialized vocabulary-based word represen-

tations will be more effective than general vocabulary-based word representations for clinical entity and relation extraction.
- In the training phase of the CRF layer, we used the actual label of entity as the input for the next step, whereas in the predicting phase, we used the predicted label as the input for the next step. As a result, the predicted tokens from training and inference are drawn from different distributions. We plan to modify the model framework to solve the problem of inconsistent data in the current training and predicting phases.

We also plan to conduct experiments on Chinese and Japanese datasets to verify the effectiveness of our method.

REFERENCES

[1] Nguyen Bach and Sameer Badaskar. A review of relation extraction. *Literature review for Language and Statistics II*, 2:1–15, 2007.

[2] Giannis Bekoulis, Johannes Deleu, Thomas Demeester, and Chris Develder. An attentive neural architecture for joint segmentation and parsing and its application to real estate ads. *Expert Systems with Applications*, 102:100–112, 2018.

[3] Giannis Bekoulis, Johannes Deleu, Thomas Demeester, and Chris Develder. Joint entity recognition and relation extraction as a multi-head selection problem. *Expert Systems with Applications*, 114:34–45, 2018.

[4] Ioannis Bekoulis, Johannes Deleu, Thomas Demeester, and Chris Develder. Reconstructing the house from the ad: Structured prediction on real estate classifieds. In *In Proceedings of the 15th Conference on the European Chapter of the Association for Computational Linguistics*, pages 1–6, 2017.

[5] Fenia Christopoulou, Thy Thy Tran, Sunil Kumar Sahu, Makoto Miwa, and Sophia Ananiadou. Adverse drug events and medication relation extraction in electronic health records with ensemble deep learning methods. *Journal of the American Medical Informatics Association*, 27(1):39–46, 2020.

[6] Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. ELECTRA: Pre-training text encoders as discriminators rather than generators. *The International Conference on Learning Representations*, 2020.

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

[8] Hicham El Boukkouri, Olivier Ferret, Thomas Lavergne, Hiroshi Noji, Pierre Zweigenbaum, and Jun'ichi Tsujii. CharacterBERT: Reconciling elmo and bert for word-level open-vocabulary representations from characters. In *International Conference on Computational Linguistics*, pages 6903–6915, 2020.

[9] Hicham El Boukkouri, Olivier Ferret, Thomas Lavergne, and Pierre Zweigenbaum. Embedding strategies for specialized domains: Application to clinical entity recognition. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 295–301, 2019.

[10] Katrin Fundel, Robert Küffner, and Ralf Zimmer. RelEx—relation extraction using dependency parse trees. *Bioinformatics*, 23(3):365–371, 2007.

[11] Harsha Gurulingappa, Abdul Mateen-Rajpu, and Luca Toldo. Extraction of potential adverse drug events from medical case reports. *Journal of biomedical semantics*, 3(1):1–10, 2012.

[12] Sam Henry, Kevin Buchan, Michele Filannino, Amber Stubbs, and Ozlem Uzuner. 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. *Journal of the American Medical Informatics Association*, 27(1):3–12, 2020.

[13] Meizhi Ju, Nhung TH Nguyen, Makoto Miwa, and Sophia Ananiadou. An ensemble of neural models for nested adverse drug events and medication extraction with subwords. *Journal of the American Medical Informatics Association*, 27(1):22–30, 2020.

[14] Nanda Kambhatla. Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 178–181, 2004.

[15] John Lafferty, Andrew McCallum, and Fernando CN Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *In Proceedings of 18th International Conference on Machine Learning*, 2001.

[16] Qi Li and Heng Ji. Incremental joint extraction of entity mentions and relations. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 402–412, 2014.

[17] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[18] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2018.

[19] Makoto Miwa and Mohit Bansal. End-to-end relation extraction using LSTMs on sequences and tree structures. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1105–1116, 2016.

[20] Makoto Miwa and Yutaka Sasaki. Modeling joint entity and relation extraction with table representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1858–1869, 2014.

[21] David Nadeau and Satoshi Sekine. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26, 2007.

[22] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. PyTorch: An imperative style, high-performance deep learning library. In *Advances in neural information processing systems*, pages 8026–8037, 2019.

[23] Yifan Peng, Shankai Yan, and Zhiyong 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*, pages 58–65, 2019.

[24] Bryan Rink and Sanda Harabagiu. Utd: Classifying semantic relations by combining lexical and semantic resources. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 256–259, 2010.

[25] Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556, 2011.

[26] Yucheng Wang, Bowen Yu, Yueyang Zhang, Tingwen Liu, Hongsong Zhu, and Limin Sun. TPLinker: Single-stage joint extraction of entities and relations through token pair linking. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1572–1582, 2020.

[27] Jun Xu, Hee-Jin Lee, Zongcheng Ji, Jingqi Wang, Qiang Wei, and Hua Xu. Uth_ccb system for adverse drug reaction extraction from drug labels at tac-adr 2017. In *Proceedings of the Text Analysis Conference*, 2017.

[28] Xi Yang, Jiang Bian, Ruogu Fang, Ragnhildur I Bjarnadottir, William R Hogan, and Yonghui Wu. Identifying relations of medications with adverse drug events using recurrent convolutional neural networks and gradient boosting. *Journal of the American Medical Informatics Association*, 27(1):65–72, 2020.

[29] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32, 2019.

[30] Suncong Zheng, Yuexing Hao, Dongyuan Lu, Hongyun Bao, Jiaming Xu, Hongwei Hao, and Bo Xu. Joint entity and relation extraction based on a hybrid neural network. *Neurocomputing*, 257:59–66, 2017.