NTCIR 17 Conference: Proceedings of the 17th NTCIR Conference on Evaluation of Information Access Technologies, December 12-15, 2023, Tokyo, Japan
DOI: https://doi.org/10.20736/0002001291

SRCB at the NTCIR-17 MedNLP-SC Task

Hongyu Li
Ricoh Software Research Center (Beijing) Co., Ltd.
China
Hongyu.Li@cn.ricoh.com

Yongwei Zhang
Ricoh Software Research Center (Beijing) Co., Ltd.
China
Yongwei.Zhang@cn.ricoh.com

Yuming Zhang
Ricoh Software Research Center (Beijing) Co., Ltd.
China
Yuming.Zhang1@cn.ricoh.com

Shanshan Jiang
Ricoh Software Research Center (Beijing) Co., Ltd.
China
Shanshan.Jiang@cn.ricoh.com

Bin Dong
Ricoh Software Research Center (Beijing) Co., Ltd.
China
Bin.Dong@cn.ricoh.com

ABSTRACT

Our team SRCB participated in the Social Media Adverse Drug Event Detection (SM-ADE) subtask of NTCIR-17 Medical Natural Language Processing for Social media and Clinical texts (MedNLP-SC). The task focuses on solving the problem of Adverse Drug Event (ADE) detection for social media texts in Japanese, English, French and German, which is a multi-labeling problem aimed at expressing the positive or negative status as an ADE for 22 symptom labels respectively. In this paper, we report our approaches which can be mainly categorized into 3 types according to which task we cast the original task to, including multi-label classification, binary classification and joint entity and relation extraction. Besides, we also conduct optimizations on the approaches that rely on pre-trained transformer language models, with the support of various techniques such as continual pretraining, gradient boosting methods, and transfer learning.

KEYWORDS

Adverse drug event detection, multi-label classification, binary classification, joint entity and relation extraction

TEAM NAME

SRCB

SUBTASKS

SM-ADE-EN
SM-ADE-JA
SM-ADE-FR
SM-ADE-DE

1 INTRODUCTION

NTCIR-17 Medical Natural Language Processing for Social Media and Clinical Texts (MedNLP-SC) is a shared task aimed to promote medical NLP studies focused on analyzing texts from both social media and hospital settings. It consists of two subtasks: Social Media Adverse Drug Event Detection (SM-ADE) [1] and Radiology Report TNM staging (RR-TNM) [2]. Our team mainly focuses on the SM-ADE subtask.

The SM-ADE task defines Adverse Drug Event (ADE) detection as a multi-labeling task, which is to identify a set of symptoms imputed to a drug from short messages written by social media users. In particular, there are 22 defined symptom labels as shown in table 1, some of which are hard to be captured by the language models pre-trained with texts in general domain. The task requires the participants to identify all these symptoms as 1 for positive or 0 for negative, which indicates the status of them as an ADE respectively.

To accomplish this task, we first employ the most straightforward method: a multi-label classification model based on pre-trained language models such as BERT [3], with either a sigmoid layer or a softmax layer at the top. We also consider the original task as a binary classification problem, where the objective is to identify the input sentence to be positive or negative with respect to each of the 22 labels. Besides, we propose to cast the ADE task to entity and relation extraction problem in order to better capture the relation between medications and symptoms. We implement an entity and relation extraction model based on the Universal Information Extraction (UIE) [4] pre-trained model, which reaches the highest performance among the single models. Finally, our submissions of single models or ensemble ones are among the top ranking on the metrics in all 4 tracks of EN, JA, DE, FR.

2 RELATED WORK

2.1 Multi-label Classification

Multi-label classification is a fundamental task in machine learning, which involves predicting multiple class labels for each instance simultaneously. Over the years, various approaches have been proposed to tackle the challenges posed by multi-label classification. Some traditional approaches convert multi-label classification problem into multiple single-label classification problems, such as binary relevance [5], label power-set [6] and classifier chains [7]. Recent studies have explored deep learning techniques such as neural networks and attention mechanisms for multi-label classification, achieving state-of-the-art performance in various domains. Yang et al. [8] consider multi-label text classification as a sequence generation challenge and employed a sequence generation model with global embedding to address the task of multi-label text classification comprehensively. Qin et al. [9] demonstrated a modification of the RNN sequence model to define probabilities for the set of labels. However, sigmoid and logistic regression loss has been one

Table 1: The 22 selected symptoms describing ADEs which serve as labels for the multi-label classification.

ID	Japanese	English	German	France
01	悪心	nausea	Übelkeit	nausées
02	下痢	diarrhea	Diarrhöe	diarrhée
03	倦怠感	fatigue	Erschöpfung	fatigue
04	嘔吐	vomiting	Erbrechen	vomissements
05	食欲不振	loss of appetite	Anorexie	anorexie
06	頭痛	headache	Kopfschmerzen	maux de tête
07	発熱	fever	Fieber	fièvre
08	間質性肺疾患	interstitial lung disease	Interstitielle Lungenerkrankung	maladie pulmonaire interstitielle
09	肝障害	liver damage	Leberschädigung	problèmes de foie
10	浮動性めまい	dizziness	Drehschwindel	vertiges flottants
11	疼痛	pain	Schmerz	douleur
12	脱毛症	alopecia	Alopezie	alopécie
13	鎮痛剤喘息症候群	analgesic asthma syndrome	Analgetisches Asthma-Syndrom	syndrome d'asthme analgésique
14	腎障害	renal impairment	Nierenerkrankung	insuffisance rénale
15	過敏症	hypersensitivity	Hypersensibilität	hypersensibilité
16	不眠症	insomnia	Insomnie	insomnie
17	便秘	constipation	Constipation	constipation
18	骨髄機能不全	bone marrow dysfunction	Knochenmarkerkrankung	dysfonctionnement de la moelle osseuse
19	腹痛	abdominal pain	Bauchschmerzen	douleur abdominale
20	出血性膀胱炎	hemorrhagic cystitis	Hämorrhagische Zystitis	cystite hémorragique
21	発疹	rash	Ausschlag	exanthème
22	口内炎	stomatitis	Stomatitis	stomatite

of the most common options [10–12]. Especially, it can be combined with the sequence-level text representation for classification of pre-trained language models such as BERT [3].

2.2 Entity and Relation Extraction

Entity and relation extraction is a long-researched Information Extraction (IE) task and traditionally studied as a pipeline composed of two separate tasks of named entity recognition and relation extraction. These pipeline methods suffers from error propagation and lack of interactions between both tasks. In the recent years, there has been a large number of researches studying the joint modeling of extraction of entities and relations [13–16]. Different from these researches, Lu et.al [4] propose a unified text-to-structure generation framework, namely Universal IE (UIE), which can universally model different IE tasks, adaptively generate targeted structures, and collaboratively learn general IE abilities from different knowledge sources. They also release the English and Chinese UIE pre-trained models based on encoder-decoder pre-trained model T5 [17], which shows good performance on IE tasks including entity and relation extraction, even in few-shot settings. In this paper, we use a Japanese version of UIE pre-trained by ourselves.

3 METHODS

3.1 ADE as Multi-label Classification Problem

The most straightforward method we can naturally come up with is to treat the ADE detection task as a multi-label classification task. Our multi-label classification models share a model architecture containing an encoder layer of transformer-based pre-trained language models like BERT and a decoder layer with the target function of sigmoid or softmax function modeling the probabilities of the 22 symptom labels jointly. We employed BCELoss for models featuring sigmoid decoders, while for models with softmax decoders, we explored a range of different loss functions, as outlined below:

a) Cross Entropy Loss: a commonly used loss function for classification problems.
b) Focal Loss [18]: a modification based on the Cross Entropy Loss, which serves to downweight easily classified samples, thereby directing the model's attention towards challenging samples throughout the training process.
c) Label Smoothing Loss [19]: a loss function designed to mitigate overfitting by decreasing the certainty assigned to the correct label.
d) Dice Loss [20]: a F1-score oriented loss, which is consistent with the evaluation metrics.
e) Weighted Loss: we consider that each of the 22 classification tasks should have a different contribution to the process of back propagation, because some labels are easier to classify, while others are not. We define 22 learnable parameters to present the contribution of each label classification to the loss function. The calculation is as follow:

$$loss_{total} = \sum_{k=1}^{22} w_k * loss_k$$

where $loss_{total}$ presents the total loss of the model, w_k presents the learnable parameter, $loss_k$ presents the k_{th} label of 22 labels.

3.2 ADE as Binary Classification Problem

By concentrating on each individual label within the set of 22 symptom labels, we can approach the ADE task as a binary classification problem which involves categorizing the input text as either positive or negative with respect to the specified label. Instead of creating 22 separate models for the classification based on each label, we propose to use a single binary classification model to handle all 22 labels through concatenating the original text with a label-wise prompt as context. In this process, each original text is expanded into 22 new samples, then these samples are fed into the model to determine whether the ADE status of the symptom labels is positive or negative respectively. Our approach involves three primary methods for constructing the label-wise prompt. In addition to simply utilizing the text of the labels themselves, we also utilize the format of Natural Language Inference (NLI) or binary-choice machine reading comprehension. The methods are detailed in Table 2.

We mainly employ a BERT-based architecture for our binary classification models using pre-trained language models shown in Table 3, where we project the [CLS] token representation into a 2-dimensional embedding through a linear transformation. Subsequently, the 2-dimensional embedding are fed into a softmax classifier for binary classification of 1 for positive or 0 for negative.

3.3 ADE as Entity and Relation Extraction Problem

In the NTCIR-17 SM dataset, we observe that for a specified symptom label labeled as positive for the target sentence, there must be at least one corresponding ADE semantic triple of (medication entity, side effect, symptom entity) that could be extracted in the sentence, where the symptom entity is an expression of the specified symptom label. Therefore, as the example in Figure 1 illustrates, we can regard the task of labeling the 22 symptom labels with positive or negative as the task of extracting ADE semantic triples, which is a typical entity and relation extraction problem between medication entities and symptom entities. Due to the fact that there is no annotated entity and relation fine-tuning data, we first automatically construct entity and relation annotation data of medications and symptoms based on the original dataset, and then we fine-tune an Universal Information Extraction (UIE) model to extract ADE semantic triples. The UIE model is a pre-trained model which learnt general IE abilities from different knowledge sources and shows good IE performance in both the common fine-tuning settings and the low-resource settings including the few-shot settings.

3.3.1 Automatic Entity and Relation Data Annotation. Extracting the ADE semantic triples requires annotation data of entity categories including medication and 22 symptom labels as well as the relation type of side effect between the entities. Since the original data does not provide such annotation, we first utilize the few-shot entity extraction ability of the UIE model to automatically annotate medication entities and 22 types of symptom entities, with just a few of human-annotated data. Specifically, we sample 5 positive samples for each symptom label[1] from the training set and annotate all medication entities and symptom entities regardless whether they are one item of a gold ADE semantic triple or not. We fine-tune the UIE model with the 5-shot data, and then use this model to predict the entities for all training samples of the original dataset. However, there is a lot of noise in the automatically annotated data, and the predictions of symptom labels for the same expressions sometimes vary a lot. Therefore, we employ a weighted instance-level ensemble approach to reduce the noise, which uniformly assign the label for each entity mention with the label predicted the most times across different samples. Besides, we also give a weight boost to the symptom labels which is labeled with positive as an ADE in the training set. For the relation annotation, we applied a naive distance-based method that selects the nearest medication entity to each symptom entity of which the corresponding symptom label is labeled as positive in the training set. Finally, we obtained an relatively high-quality automatically annotated entity and relation extraction data which can be used to fine-tune any entity and relation extraction models on the task of extracting ADE semantic triples.

3.3.2 Entity and Relation Extraction with UIE. To mitigate error accumulation and align with the UIE preference on entity extraction, we continue utilizing the UIE model to perform the entity and relation extraction task jointly. UIE model uniformly models all IE tasks as text-to-structure tasks. The input and output of UIE during fine-tuning and predicting are illustrated in FIG 2. The UIE model receives the Structural Schema Instructor (SSI) which provides the model with the schema of extracting ADE semantic triples task, along with the target sentence as input. In our case, the schema includes the entity categories of medication (医薬品) and 22 symptom labels (start with "[spot]" token), as well as the relation type of side effect (副作用) (start with "[asso]" token). The output is a sequence constrained by Structured Extraction Language (SEL), which can be decoded into the ADE semantic triples. Finally, for each predicted sample, we label the corresponding symptom labels for the symptom entities with positive.

3.4 Other Methods

3.4.1 Continual Pre-training. Continual pre-training [21] enables the pre-trained language models to better understand content within the domain of the target dataset. In this task, the dataset comprises tweets generated by a T5 model [17], which could be a new semantic domain for the pre-trained language models. Therefore, we employed continual pre-training based on the PLMs we used in different tracks. Specially, in the track of FR and DE, we pre-train the multilingual XLM-RoBERTa-large model with the texts from all 4 tracks.

3.4.2 Gradient boosting as Decoder. Instead of using a Fully Connected layer as the decoder in multi-label classification models, we also try various different gradient boosting models for the decoder, including XGBoost [22], CatBoost [23], and others. These models are more effective on analyzing the features extracted by the pre-trained language models, which leads to slight improvement for some labels.

3.4.3 Transfer Learning. Transfer Learning is a machine learning technique where a model trained on one task is re-purposed and fine-tuned for a related, but different task. We define two different tasks: 1. Classification of the texts into “contains ADE” and “does not contain ADE” 2. Classifying which languages the texts belong

[1] Here, positive samples means the samples where the specified symptom label is labeled as positive.

Table 2: Different prompts used in binary classification

Label-wise Prompt	Model Input
[label]	[CLS] [label] [SEP] [original text]
The text contains drug-induced [label]	[CLS] [original text] [SEP] The text contains drug-induced [label] [SEP]
Is there any adverse drug event related to [label] in this sentence ?	[CLS] [original text] [SEP] Is there any adverse drug event related to [label] in this sentence ? [SEP]

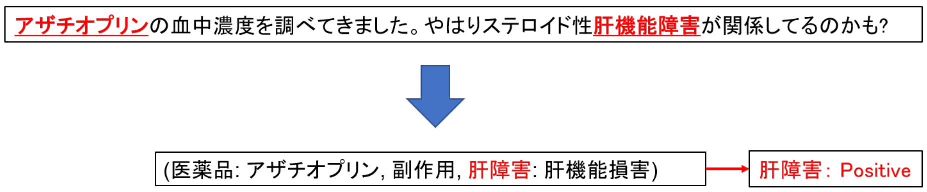

Figure 1: An example of regarding ADE detection as ADE semantic triple extraction

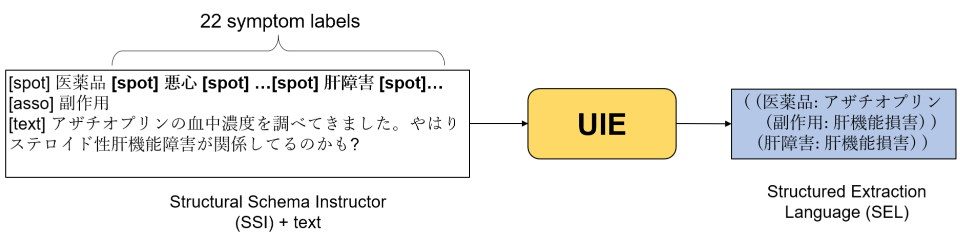

Figure 2: Input and output of Universal Information Extraction (UIE) model

to. The models are first trained on these two tasks and then fine-tuned on the original objective of ADE detection using the SM-ADE dataset. This method shows apparent improvement in our experiments.

3.5 Model Ensemble

For each of our model candidates, we conduct 5-fold cross-validation on the training data, resulting in 5 distinct models trained on four folds of the data while validated with the remaining fold. Therefore, we first conduct model ensemble among the 5 models belonging to the same model candidate. Instead of counting on the last checkpoint or the one with the highest overall micro F1 score, we retain the checkpoints for each model where the highest F1 score is achieved on a specific label. During the testing phase, for each test sample, the predictions for each label are determined by the model checkpoint that achieved the highest F1 score for that particular label. This servers as the prediction of that test sample for the respective model. Subsequently, a majority voting strategy is applied, wherein the predictions from the five models are subjected to the prediction with the most number of votes. In this way, we create a prediction file for each model candidate.

For the model ensemble among model candidates, we mainly employ the strategies of majoirty voting and random voting. Majority voting means all model candidates are used in ensemble. And as the final result, it will pick the result which the most number of models agree with for each prediction. While random voting means each time we randomly select a random number of model candidates as one candidate combination and choose the combination that reaches the best evaluation result through multiple experiments. This may work because not all model candidates can contribute to the true value, and sometimes the inconsistency between model candidates drives the result away from the true value.

4 EXPERIMENTS

4.1 Evaluation Metrics

We evaluate the labels on the metrics listed below, according to the instructions provided in the SC-SM overview paper [1].

(1) **Binary Scores**: We evaluate how well the models can detect examples containing ADES, regardless of the symptom labels. Specifically, we calculate the F1 score of classifying a document into the classes “contains ADE” (positive) versus “does not contain ADE” (negative). A document is considered to contain an ADE if a least one symptom class is positive.
(2) **Per ADE Label Scores**: We calculate the F1 score for the class "contains ADE” (positive) across samples.
(3) **Micro/Macro F1 Scores**: We calculate the micro/macro F1 scores across all labels.
(4) **Exact Match Accuracy** We calculate the percentage of exact matches accoss all samples. It will count as an exact

Table 3: Different pretrained language models used in the track of EN, JA, FR, DE

Track	PLM
EN	Pubmed BERT [24]
	Clinical BERT [25]
	BioBERT [26]
	BioLinkBERT[27]
JA	cl-tohoku/BERT-base-japanese
	UIE-large-japanese
FR	xlm-roberta-large [28]
DE	xlm-roberta-large [28]

match when all symptom labels in one sample are perfectly predicted.

We calculate the F1 score for each label of

4.2 Experiment Settings

In order to make better use of the training data for model training and validation, we have implemented a 5-fold cross-validation approach. Due to the severe label imbalance within the training data, it's likely that the minority labels may be entirely absent from one or more of the randomly selected folds, which leads to unreasonable validation results on those labels. To address this problem, we try our best to ensure the minority labels appear the same times in the 5 folds by artificial filtering. For the other labels, we perform random sampling to achieve a comparable distribution of label counts.

Our models utilize various pre-trained language models (PLMs) across different tracks, as indicated in Table 3. For the BERT-based models of multi-label classification or binary classification, we conduct a comparison of several PLMs that have been adapted for the medical domain in the English track. In the Japanese track, we download the Tohouku University's BERT-base-japanese model due to the absence of good Japanese medical PLMs. For France and German tracks, we use XLM-RoBERTa-large model which is a multilingual PLM. Except for UIE-large-japanese, all of the pre-trained language models we used are downloaded from huggingface[2]. We only apply the entity and relation extraction method in the Japanese track, and use a Japanese UIE model (UIE-large-japanese) pre-trained by ourselves, following the pre-training steps of UIE [4] with the processed data from Wikipedia[3] and Wikidata[4] dumps. We currently do not have plans to open-source our Japanese UIE model.

The models based on pre-trained transformer language models use a learning rate of 1e-5 and batch size of 16 or 32. We use AdamW as optimizer and employ early stop to avoid over-fitting. The hyper-parameters used for the fine-tuning of our entity and relation extraction models based on UIE-large-japanese is shown in Table 4. We use different hyper-parameters in 5-shot training during automatic entity and relation data annotation and full-data fine-tuning. Rejection noise is a special hyper-parameter introduced by the rejection mechanism of UIE, which trains the model to reject misleading generation of negative entity categories.

[2]https://huggingface.co/
[3]https://www.wikipedia.org/
[4]https://www.wikidata.org/

Table 4: Hyper-parameters for the entity and relation extraction models based on UIE-large-Japanese

Hyper-parameter	Fine-tuning	
	5-shot	Full-data
Learning Rate	1e-4	3e-4
Rejection Noise	0.1	0.1
Global Batch Size	16	32
Schedule	constant	linear
Warmup Rate	0.0	0.06
Epoch	200	50

4.3 Experiment Results

The experiment results on training data of our methods are illustrated in Table 5. We only present the best model results of each method. In addition, since our members are responsible for different methods and different language tracks respectively, there are methods absent from one or more of the tracks. As the most straightforward approach, the multi-label classification method plays a role of a baseline. Gradient boosting method shows a certain level of improvement compared with the baselines. Among the methods, entity and relation extraction method, binary classification method and transfer learning method show superior performance over the others with a considerable margin.

4.4 Submissions

Our submission files comprise the outcomes of model ensemble and individual models, employing various model candidates ranked based on Macro F1 (Macro avg).

Submission-1 (EN): Random voting results of the top-10 model candidates including multi-label classification, binary classification, gradient boosting and transfer learning methods based on Pubmed BERT or BioLinkBERT with or without continual pre-training.

Submission-2 (EN): Majority voting results of the top-10 model candidates including multi-label classification, binary classification, gradient boosting and transfer learning methods based on Pubmed BERT or BioLinkBERT with or without continual pre-training.

Submission-3 (EN): Majority voting results of the all model candidates including all tested methods based on all tested pre-trained transformer language models with or without continual pre-training.

Submission-4 (JA): Single model results of entity and relation extraction method based on UIE-large-japanese.

Submission-5 (JA): Majority voting results of multi-label classification, binary classification and entity and relation extraction methods based on cl-tohoku/BERT-base-japanese or UIE-large-japanese.

Submission-6 (JA): Majority voting results of binary classification, entity and relation extraction methods based on cl-tohoku/BERT-base-japanese or UIE-large-japanese.

Submission-7 (FR)/10 (DE): Random voting results of the top-5 model candidates including multi-label classification, binary

Table 5: Evaluation Results on Training data of Our Methods (Average of 5-fold cross-validation)

	Methods	Metrics				
		Binary Scores (ADE vs. no ADE)	Per Label Scores (Full)	Micro Avg	Macro Avg	Exact Accuracy
EN	Multi-label Classification	0.84	0.90	0.86	0.85	0.862
	Binary Classification	0.92	0.93	0.87	0.86	0.891
	Gradient Boosting as Decoder	0.83	0.91	0.86	0.85	0.868
	Transfer learning	0.91	0.93	0.88	0.87	0.902
JA	Multi-label Classification	0.85	0.90	0.86	0.86	0.866
	Binary Classification	0.92	0.94	0.88	0.87	0.896
	NER&RE	0.93	0.94	0.89	0.88	0.911
FR	Multi-label Classification	0.84	0.89	0.86	0.85	0.864
	Binary Classification	0.91	0.91	0.86	0.86	0.890
	Transfer Learning	0.92	0.92	0.87	0.88	0.901
DE	Multi-label Classification	0.84	0.90	0.85	0.85	0.864
	Binary Classification	0.92	0.92	0.87	0.86	0.890
	Transfer Learning	0.91	0.92	0.86	0.87	0.899

classification and transfer learning methods based on xlm-roberta-large with or without continual pre-training.

Submission-8 (FR)/11 (DE): Majority voting results of the top-5 model candidates including multi-label classification, binary classification and transfer learning methods based on xlm-roberta-large with or without continual pre-training.

Submission-9 (FR)/12 (DE): Majority voting results of the all model candidates including all tested methods based on xlm-roberta-large with or without continual pre-training.

The detailed results for our submissions are listed in Table 6. The results show that the ensemble result of entity and relation extraction, multi-label classification and binary classification, namely submission-5 achieves the best performance among the 16 submissions. And single model result (submission-4) of the entity and relation extraction method based on UIE-large-japanese achieves comparable performance with the ensemble ones.

5 CONCLUSIONS

In this paper, for the task of Adeverse Drug Event Detection shared task of NTCIR-17, we preliminary propose to treat the original task as a multi-label classification problem, binary classification problem or entity and relation extraction problem. In addition, we also utilizing techniques such as continual pre-training of language models, gradient boosting methods and transfer learning to improve the final performance. Besides, we compared the performance of different pre-trained language models, some of which are specialized in medical domain. All of our proposed methods outperform our baselines to varing degrees. In particular, the entity and relation extraction methods based on UIE pre-trained model show its ability to capture the relation between medications and symptoms, leading to the highest scores among the single models that we tested. Finally, we submit our ensemble results and achieve the top ranking across all metrics in all 4 tracks of JA, EN, DE and FR.

REFERENCES

[1] Shoko Wakamiya, Lis Kanashiro Pereira, Lisa Raithel, Hui-Syuan Yeh, Peitao Han, Seiji Shimizu, Tomohiro Nishiyama, Gabriel Herman Bernardim Andrade, Noriki Nishida, Hiroki Teranishi, Narumi Tokunaga, Philippe Thomas, Roland Roller, Pierre Zweigenbaum, Yuji Matsumoto, Akiko Aizawa, Sebastian Möller, Cyril Grouin, Thomas Lavergne, Aurélie Névéol, Patrick Paroubek, Shuntaro Yada, and Eiji Aramaki. Ntcir-17 mednlp-sc social media adverse drug event detection: Subtask overview. In *Proceedings of the 17th NTCIR Conference on Evaluation of Information Access Technologies, NTCIR-17*. National Institute of Informatics (NII), 2023. https://doi.org/10.20736/0002001327.

[2] Yuta Nakamura, Shouhei Hanaoka, Shuntaro Yada, Shoko Wakamiya, and Eiji Aramaki. Ntcir-17 mednlp-sc radiology report subtask overview: Dataset and solutions for automated lung cancer staging. In *Proceedings of the 17th NTCIR Conference on Evaluation of Information Access Technologies, NTCIR-17*. National Institute of Informatics (NII), 2023. https://doi.org/10.20736/0002001328.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

[4] Yaojie Lu, Qing Liu, Dai Dai, Xinyan Xiao, Hongyu Lin, Xianpei Han, Le Sun, and Hua Wu. Unified structure generation for universal information extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5755–5772, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[5] Matthew R Boutella, Jiebo Luob, Xipeng Shena, and Christopher M Browna. Learning multi-label scene classification. *Pattern Recognition*, 37:1757–1771, 2004.

[6] Grigorios Tsoumakas and Ioannis Katakis. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining (IJDWM)*, 3(3):1–13, 2007.

[7] Jesse Read, Bernhard Pfahringer, Geoff Holmes, and Eibe Frank. Classifier chains for multi-label classification. *Machine learning*, 85:333–359, 2011.

[8] Pengcheng Yang, Xu Sun, Wei Li, Shuming Ma, Wei Wu, and Houfeng Wang. Sgm: Sequence generation model for multi-label classification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3915–3926, 2018.

[9] Kechen Qin, Cheng Li, Virgil Pavlu, and Javed Aslam. Adapting rnn sequence prediction model to multi-label set prediction. 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 3181–3190, 2019.

[10] Thibaut Durand, Nazanin Mehrasa, and Greg Mori. Learning a deep convnet for multi-label classification with partial labels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 647–657, 2019.

[11] Jingya Wang, Xiatian Zhu, Shaogang Gong, and Wei Li. Transferable joint attribute-identity deep learning for unsupervised person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2275–2284, 2018.

[12] Shan Lin, Haoliang Li, Chang-Tsun Li, and Alex ChiChung Kot. Multi-task mid-level feature alignment network for unsupervised cross-dataset person re-identification. *CoRR*, abs/1807.01440, 2018.

[13] Jue Wang and Wei Lu. Two are better than one: Joint entity and relation extraction with table-sequence encoders. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1706–1721, 2020.

[14] Suncong Zheng, Feng Wang, Hongyun Bao, Yuexing Hao, Peng Zhou, and Bo Xu. Joint extraction of entities and relations based on a novel tagging scheme. In

Table 6: Submission Results on Test Data of Our Methods

Submission		Metrics				
		Binary Scores (ADE vs. no ADE)	Per Label Scores (Full)	Micro Avg	Macro Avg	Exact Accuracy
EN	Submission-1	0.87	0.90	0.81	0.78	0.869
	Submission-2	0.87	0.90	0.81	0.76	0.855
	Submission-3	0.82	0.87	0.75	0.75	0.790
JA	Submission-4	0.88	0.91	0.82	0.78	0.870
	Submission-5	0.88	0.91	0.82	0.79	0.878
	Submission-6	0.88	0.91	0.81	0.75	0.872
FR	Submission-7	0.87	0.90	0.81	0.76	0.866
	Submission-8	0.86	0.89	0.79	0.72	0.845
	Submission-9	0.82	0.87	0.74	0.73	0.779
DE	Submission-10	0.87	0.91	0.82	0.80	0.864
	Submission-11	0.84	0.89	0.78	0.76	0.820
	Submission-12	0.83	0.88	0.76	0.74	0.798

Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1227–1236, 2017.

[15] Giannis Bekoulis, Johannes Deleu, Thomas Demeester, and Chris Develder. Adversarial training for multi-context joint entity and relation extraction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2830–2836, 2018.

[16] Changzhi Sun, Yeyun Gong, Yuanbin Wu, Ming Gong, Daxin Jiang, Man Lan, Shiliang Sun, and Nan Duan. Joint type inference on entities and relations via graph convolutional networks. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1361–1370, 2019.

[17] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer, 2020.

[18] Tsung Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *arXiv e-prints*, 2017.

[19] Rafael Müller, Simon Kornblith, and Geoffrey Hinton. When does label smoothing help?, 2020.

[20] Carole H. Sudre, Wenqi Li, Tom Vercauteren, Sebastien Ourselin, and M. Jorge Cardoso. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 240–248. Springer International Publishing, 2017.

[21] Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. Don't stop pretraining: Adapt language models to domains and tasks, 2020.

[22] Tianqi Chen and Carlos Guestrin. XGBoost. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, aug 2016.

[23] Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. Catboost: unbiased boosting with categorical features, 2019.

[24] Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23, 2021.

[25] Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *CoRR*, abs/1904.05342, 2019.

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

[27] Michihiro Yasunaga, Jure Leskovec, and Percy Liang. Linkbert: Pretraining language models with document links. In *Association for Computational Linguistics (ACL)*, 2022.

[28] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online, July 2020. Association for Computational Linguistics.