NICTmed at the NCTIR-16 Real-MedNLP Task

Masao Ideuchi
National Institute of Information and Communications Technology
Japan
FUJITSU LIMITED
Japan
Nara Institute of Science and Technology
Japan
masao.ideuchi@nict.go.jp

Masatoshi Tsuchiya
National Institute of Information and Communications Technology
Japan
Toyohashi University of Technology
Japan
tsuchiya@imc.tut.ac.jp

Yiran Wang
National Institute of Information and Communications Technology
Japan
yiran.wang@nict.go.jp

Masao Utiyama
National Institute of Information and Communications Technology
Japan
mutiyama@nict.go.jp

ABSTRACT

This paper describes NICTmed team's challenge to Subtask1-CR-EN, Subtask1-CR-JA, Subtask3-CR-EN, Subtask3-CR-JA, Subtask3-RR-JA and Subtask3-RR-EN in NTCIR-16 Real-MedNLP. Real-MedNLP is information extraction task in real medical domain text, and approximately 100 annotated real clinical reports in both English and Japanese are given to its participants. Subtask1-CR-EN/JA and Subtask3-CR-EN/JA are the tasks based on case reports (CR), Subtask1 is few-resource Named Entity Recognition (NER) and Subtask3 is information extraction for adverse drug event (ADE). Subtask3-RR-EN/JA are the tasks of case identification (CI) based on radiology reports (RR). We used multilingual BERT (mBERT) and XLM-RoBERTa (XLM-R) to compare how effective the multilingual pre-trained models work in specific domain downstream tasks both English and Japanese. Our experiment used no external data to adjust conditions of English and Japanese experiments. Our experiment showed that the multilingual pre-training models achieved level of accuracy in Japanese as in English, and got rank 3 in Entity F1 of all target entities for Subtask1-CR-JA, top rank in Report-level precision and F1 for Subtask3-CR-JA.

KEYWORDS

Medical Natural Language Processing, Named Entity Recognition, Multilingual Pre-trained Language Model, Adverse Drug Event, Case Identification

TEAM NAME

NICTmed

SUBTASKS

Subtask1-CR-EN
Subtask1-CR-JA
Subtask3-CR-EN (ADE)
Subtask3-CR-JA (ADE)
Subtask3-RR-EN (CI)
Subtask3-RR-JA (CI)

1 INTRODUCTION

Recently, the number of annotated corpora published becomes larger, but Japanese corpora are not enough compared with English, and the number of natural language processing researches use English corpora and pre-training models. The corpora without annotation in Japanese are available such as Japanese Wikipedia[1] and Aozora Bunko[2], and Japanese pre-training models are also available. However, for specific domains and specific tasks, most annotated corpora are available only in English, and difficult to prepare data in Japanese.

When trying to solve a task in a specific domain in Japanese, the most common first step is to find reference systems which solve the similar tasks in the specific domain using pre-trained models with English corpora. In such cases, reproducing the reference systems with a multilingual pre-training model and then applying to a target Japanese corpus is a good practice because the changes from reference system are reduced and problem identification becomes easily when the accuracy is not enough. Especially, a small change in machine learning system may bring a large change[7], therefore, changing and comparing a part of experimental conditions from a reference system help to identify how we can improve accuracy for practical use.

In Real-MedNLP[8], annotated real clinical reports were provided in both Japanese and English. With the multilingual pre-trained models become familiar, the corpora available in multiple languages for the similar tasks are increasing, but annotated corpora in multiple languages for specific domains and tasks are still rare. The bilingual reports in Real-MedNLP are good examples for examininn how well multilingual pre-trained models work in a particular domain. In this paper, we used multilingual pre-trained models that are trained on general corpora and fine-tuned to tasks in a specific domain then we confirmed the accuracy achieved and the differences in accuracy between English and Japanese. As a result, we revealed that the multilingual pre-trained models can work in a specific domain both English and Japanese.

[1] https://ja.wikipedia.org/
[2] https://www.aozora.gr.jp/

The tasks for which we submitted results are Subtask1-CR-EN, Subtask1-CR-JA, Subtask3-CR-EN (ADE), Subtask3-CR-JA (ADE), Subtask3-RR-EN (CI), and Subtask3-RR-JA (CI). However, we expected Subtask3-RR-EN (CI) and Subtask3-RR-JA (CI) are supervised classification problem at first, but they are clustering problem in fact and no significant results were obtained in those tasks. Therefore, we report mainly Subtask1-CR-EN, Subtask1-CR-JA, Subtask3-CR-EN (ADE), and Subtask3-CR-JA (ADE) in this paper.

Section 2 describes the definition of the task, and Section 3 describes the data used in our experiment and how they are processed. Section 4 describes the configuration of our system, Section 5 describes the details of the official submissions, and concludes with Section 7.

2 TASK DEFINITION

We formulated Subtask3-CR-EN (ADE) and Subtask3-CR-JA (ADE) as extended labeled tasks of named entity recognition (NER) from Subtask1-CR-EN and Subtask1-CR-JA those are given as NER. Subsection 2.1 describes the definition of an NER task in Subtask1, and Subsection 2.2 describes how we converted given ADE labels to a NER task in Subtask3. For Subtask3-RR, we vectorized documents using the multilingual pre-trained BERT model then applied clustering.

2.1 NER in Subtask1-CR-EN and Subtask1-CR-JA

An NER task is generally treated as a sequence labeling problem with *BIO* formed labels same as CoNLL[9]. The *B* label represent the beginning of an entity, the *I* label represents other tokens inside an entity, and the *O* label represents all other non-entity tokens. If there are multiple types of labels, *B-EntityCategory* and *I-EntityCategory* are used as NER labels for each type of entities. For training, the input text is divided into tokens $X = (x_1, x_2, x_3, ...x_n)$, and NER labels for each token $Y = (y_1, y_2, y_3, ...y_n)$ are given. In predicting, the input text is divided into tokens same as training, then the trained model predicts NER labels for each token.

In Subtask1-CR-EN and Subtask1-CR-JA, we converted text lines in XML files described in Subsection 3.1 to *BIO* labeled data. We considered a method to predict *BIO* for each kind of label and a method to predict all kind of labels in the form of *B-EntityCategory* and *I-EntityCategory* as options. There were some labels that could not be predicted at all since the number of *B* and *I* labels become 1% or less of the total number of tokens. Therefore, we adopted a format in which all kinds of labels are assigned to one file in the form of *B-EntityCategory* and *I-EntityCategory*.

2.2 Subtask3-CR-EN (ADE) and Subtask3-CR-JA (ADE)

Subtask3-CR-EN/JA ADE prediction is the task of predicting the *ADEval* from given "articleID", "tag", and "text". *ADEval* is a categorical number and the value is from 0 to 3 where the higher value corresponds to the higher certainty level. *ADEval* is assigned to the list of NEs, executed medicines and diseases diagnosed as positive, described in each case report. We treated *ADEval* as an additional attribute of each NE and predicted *ADEval* by solving NER task with extended labels.

3 DATA

In Real-MedNLP, two types of reports were distributed, Case Report (CR) and Radiology Report (RR). The use of external data was allowed, but we didn't use external data, except pre-trained models, in order to adjust the experimental conditions in English and Japanese. We focus on CR data and explain the data format in Subsection 3.1 and annotation format in Subsection 3.2, and the relationship with the ADE training data in Subsection 3.3. For RR, data format and annotation format are same as CR. We removed tags for vectorization because the multilingual pre-trained model treats plain text.

3.1 Data Format

The training and test data for CR were provided as XML formated files with multiple reports bundled together for each language and dataset. The XML tags that define the metadata and structure of documents are on separate lines in a XML file, and each report is separated by lines of "`<article>`" and "`</article>`". The other lines of XML files are text lines of reports and the line count of each language and dataset are shown in Table 1.

In the training data, named entities (NEs) are marked up with XML tags in text lines, and in the test data, no XML tags exist in text lines. The English and Japanese XML files are document-level parallel corpora, which means each report can be paired with the translated report but each sentence cannot be paired with the translated sentence. In many cases, the long sentences in Japanese article are divided in English articles and there are some unnatural NEs because of translation. For example, in some English Time-Date NEs, the words "this" and "study" are included and this is because the words "¡ñ" in Japanese articles are translated to "in this study".

3.2 Annotation Format

There are 11 types of XML tags in CR data, and some types have multiple attributes. There are both essential and optional attributes, however, some of the optional attributes are assigned almost entirely. Therefore, we transfered the XML tags and attributes into 31 types of NER labels according to the actual assignment conditions. The NER labels used in experiments and XML tags and attributes in original data, and the number of labels in the training data are shown in Table 2.

For the sequential labeling problems in Subtask1-CR, the total number of the NER labels are 63, *B* and *I* NER labels for each NE and *O* for other tokens, were used.

3.3 ADE supervised training data

The supervised training data for ADE are given as CSV files with columns of "report ID", "NE class" (medicine or disease), "text" of the NE, and "*ADEval*". Only in Japanese training data, an additional column that contains "standard name" for NEs if available. *ADEval* is a categorical integer values that are on ordinal scale from 0 to 3 that expresses the degree of suspicion of ADEs. $ADEval = 0$ indicates "Unrelated", $ADEval = 1$ indicates "Unlikely", $ADEval = 2$ indicates "Probably", and $ADEval = 3$ indicates "Definitely", respectively. The total count of $ADEval = N$ and count of the reports that contain $ADEval = N$ on given training data are shown in Table 3.

Table 1: Line Count of CR data

Line category	Line count of CR-EN-training	Line count or CR-JA-training	Line count of CR-EN-test	Line count or CR-JA-test
XML tag lines for document attributes	15	15	13	13
article tag lines	298	298	152	152
Text lines of reports	2,160	1,908	1,602	1,524
Total lines count	2,473	2,221	1,767	1,689

Table 2: NER labels for Subtask1 and XML tags and attributes in CR training data

XML tags	Attributes	Count (EN)	Count (JA)	NER labels in Subtask1
d	positive	1698	1698	Disease-Positive
	suspicious	80	80	Disease-Suspicious
	negative	250	249	Disease-Neagtive
	general	302	302	Disease-General
a	-	818	822	Anatomical
f	-	636	637	Feature
c	-	571	571	Change
timex3	date	537	537	Time-Date
	time	53	52	Time-Time
	duration	81	82	Time-Duration
	set	34	34	Time-Set
	age	189	189	Time-Age
	med	430	428	Time-Med
	misc	28	28	Time-Misc
t-test	scheduled	0	0	-
	executed	366	366	Test_Name-Executed
	negated	7	7	Test_Name-Negated
	other	18	18	Test_Name-Other
t-key	-	524	524	Test_Item
t-val	-	428	428	Test_Value
m-key	scheduled	0	0	-
	executed	266	266	Medicine_Name-Executed
	negated	27	27	Medicine_Name-Negated
	other	51	51	Medicine_Name-Other
m-val	-	64	64	Medicine_Value
r	scheduled	28	28	Remedy-Scheduled
	executed	545	544	Remedy-Executed
	negated	30	30	Remedy-Negated
	other	77	77	Remedy-Other
cc	scheduled	1	1	Clinical_Context-Scheduled
	executed	243	242	Clinical_Context-Executed
	negated	3	3	Clinical_Context-Negated
	other	17	17	Clinical_Context-Other

Table 3: ADEval counts in training data

ADEval	Count (EN)	Count (JA)	Report count (EN)	Report count (JA)
0	1,382	1,320	146	146
1	63	61	21	21
2	84	79	21	21
3	175	170	24	24
All	1,704	1,630	147	147

ADEval is heavily biased to 0, and if $ADEval! = 0$, it is biased to 3. In addition, because ADEs were not mentioned in every report, 146 of 147 reports in both English and Japanese include $ADEval = 0$, and 104 reports only include $ADEval = 0$ for all diseases and medicines in the reports.

To understand data relationships in detail, we compared diseases and medicines annotated to CR texts and the list of diseases and medicines in ADE supervised training data. We found *ADEval* are given to the diseases and medicines when diseases are diagnosed as positive and medicines are executed in Subsection 3.2, respectively. Moreover, even when the same diseases or medicines were written more equal than two times in a report, the *ADEval* appears only once in the ADE training data. We subcategorized the NER labels of "Disease-Positive" and "Medicine-Executed" in Subtask1 and used new NER labels shown in Table 4 for Subtask3.

Table 4: NER labels detailed in ADE task

NER labels in Subtask1	*ADEval*	NER labels in Subtask3 (ADE)
Disease-Positive	0	Disease-Positive-0
	1	Disease-Positive-1
	2	Disease-Positive-2
	3	Disease-Positive-3
Medicine-Executed	0	Medicine-Executed-0
	1	Medicine-Executed-1
	2	Medicine-Executed-2
	3	Medicine-Executed-3

For the sequential labeling problems in Subtask3-CR, the total number of the NER labels are 69, subcategolized *B* and *I* NER labels for each NE and *O* for other tokens, were used.

4 SYSTEM DETAILS

For the CR task, we reproduced a NER reference system described in Subsection 4.1 at first, and then renewed the implementation on a newer library. After that, we prepared the CR data as described in Subsection 4.2, and fine-tuned with the pre-trained models described in Subsection 4.3. The evaluation results in training data were described in Subsection 4.4. The submission data format was XML, we applied post-processing described in Subsection 4.5 to the NER prediction results. For the RR task, we removed tags from document and vectorized each article by the mBERT model described in Subsection 4.3, then compressed into 3 dimensions by using t-SNE, and applied k-means clustering. The k values in the k-means are determined by silhouette analysis.

4.1 A preliminary experiment

As a preliminary experiment, we reproduced the NER experiments using English open medical data and renewed implementation. We referred the NER implementation and evaluation of BlueBERT[6] that is one of pre-trained BERT[2] model using papers in PubMed[3]. We used BC5CDR[10] data on BLUE benchmark site[4] and confirmed NER for the disease and chemical NEs. Two types of data

[3]https://pubmed.ncbi.nlm.nih.gov/
[4]https://github.com/ncbi-nlp/BLUE_Benchmark

format were used in the preliminary experiment, a Pubtator file in which annotations are defined in a tab-delimited span data after the title and abstract text[5], and a TSV file in which the text is divided into tokens and given NER labels (hereinafter, NER file). The original implementation was worked with Tensorflow v1, so we editted to work with SimpleTransformers. As a result of the preliminary experiment, we couldn't reproduce the accuracy described in the BlueBERT paper by using both pre-trained BlueBERT in GitHub[6] and HuggingFace[7], but the NER on BC5CDR disease and for chemical are achieved about 80 and 90 in F-measure, respectively using multilingual pre-trained BERT model. These scores seemed to be reasonable because the models that are pre-trained with general corpora and the models that are pre-trained with multilingual corpora are known to be slightly inferior than models that are pre-trained with specific domain monolingual corpora. We inherited implementation of this preliminary experiment as a reference system, we switched the data to Real-MedNLP CR text and performed subsequent experiments.

4.2 Data preparation

At first, we converted the CR XML files into Pubtator files by referring to the format and implementation of the preliminary experiment. In CR files, the titles of reports were given as attributes of the `<article>` tags and not annotated. Therefore, we treated all text of a report in CR files as abstracts and made Pubtator files without titles. Then we extracted XML tags and attributes from CR files, with character indices of the start points and end points of the tags as span information.

Next, we converted the Pubtator files to NER files. For the English text, we treated words as tokens according to the preliminary experiment, and divided some symbols adjacent to words such as periods and commas as other tokens. For the Japanese text, we separated text character by character when using mBERT, and tokenized text using SentencePiece[3] attatched to XLM-R when using XLM-R. In Subtask1, we assigned all the labels described in Table 2 to each token, and in Subtask3, we also used the labels in Table 4. Some part in the CR data, the start point of a word didn't correspond to the start point of an XML tag because spaces between words sometimes included in XML tags. Therefore, we repaired the head of *I* label to *B* label if the NER labels start from *I* label without *B* label then we obtained the NER training data.

4.3 Pretrained models and fine-tuning parameters

We used mBERT and XLM-R[1] as multilingual pre-trained models, mBERT was from official web site[8], and XLM-R was xlm-roberta-base from HuggingFace[9]. The fine-tuning parameters were shown in Table 5.

[5]https://www.ncbi.nlm.nih.gov/research/bionlp/APIs/format/
[6]https://github.com/ncbi-nlp/bluebert
[7]https://huggingface.co/bionlp
[8]https://github.com/google-research/bert/blob/master/multilingual.md
[9]https://huggingface.co/xlm-roberta-base

Table 5: Fine-tuning parameters

Parameters	Values
learning_rate	4e-5
optimizer	AdamW
scheduler	linear_schedule_with_warmup
warmup_steps	0
train_batch_size	16
num_train_epochs	10

4.4 Evaluation in training data

For submission, we used all training data for fine-tuning, but we used divided training data for evaluation in previous experiments. We divided the training data into three part, *train-train*, *train-valid* and *train-test*, and confirmed the accuracy of NER task. Since the number of CR data was so small, *train-test* were the last 10 reports, and *train-valid* were the previous 10 reports, and *train-train* were other reports. The evaluation results based on training data are shown in Table 6.

Table 6: Evaluation result in training data NER

Data	Pretrained models	Precision	Recall	F1
CR-EN Subtask1 NER	mBERT	45.1	50.8	47.8
	XLM-R	45.0	49.2	47.0
CR-JA Subtask1 NER	mBERT	57.4	65.1	61.0
	XLM-R	63.3	68.1	65.6
CR-EN Subtask3 NER	mBERT	47.3	51.6	49.4
	XLM-R	44.8	49.3	46.9
CR-JA Subtask3 NER	mBERT	57.0	62.9	59.8
	XLM-R	60.9	67.2	63.9

4.5 Postprocess for submission

We processed NER results for submission. In Subtask1, submission format was an XML file, and in Subtask3, submission format was a CSV file, we converted the files because prediction results of NER were NER files.

For Subtask1, similar to data preparation, there were some NER results that start from *I* labels, we applied two correction options for to a series of NEs starting from an *I* label. The first option was repairing from *I* to *B* label if the label starts from the *I* label. The second option was exclusion, we converted series of *I* labels without a *B* label to *O* labels. Then we converted NER files to XML files by adding XML tags to the text using span information. Since it was not obvious which correction was better, we submitted the results of two options.

For Subtask3, since the target NEs to be predicted are given in CSV, we determined the predicted value of the ADEval by which NER labels were predicted for the texts of NEs. We merged NER results with weights if multiple NER labels are predicted in texts in NEs. When nothing is predicted by NER, the ADEval is determined by text and weighted score based on the training data.

5 OFFICIAL RESULTS

The experimental settings and Entity-F1 scores of the systems that were submitted to Subtask1-CR are as shown in Table 7.

Table 7: Subtask1-CR submitted systems' results

Language	Pretrained models	ID	NE label correction	F1
English	mBERT	1	*I* to *B*	47.09
		2	*I* to *O*	48.85
	XLM-R	3	*I* to *B*	49.42
		4	*I* to *O*	51.70
Japanese	mBERT	1	*I* to *B*	56.90
		2	*I* to *O*	60.70
	XLM-R	3	*I* to *B*	55.50
		4	*I* to *O*	58.13

The experimental settings and Report-level F1 scores of the systems that were submitted to Subtask3-CR are as shown in Table 8.

Table 8: Subtask3-CR submitted systems' results

Language	Pretrained models	ID	Report-level F1
English	mBERT	1	34.48
	XLM-R	2	42.11
Japanese	mBERT	1	48.00
	XLM-R	2	32.00

The experimental settings and normalized mutual info scores of the systems that were submitted to Subtask3-RR are as shown in Table 9.

Table 9: Subtask3-RR submitted systems' results

Language	Pretrained models	ID	Normalized Mutual Info
English	mBERT	1	21.72
Japanese	mBERT	1	17.44

6 ADDITIONAL EXPERIMENTS

Additional experiments were performed on Subtask1-CR using the *train-train*, *train-valid* and *train-test* data used in Subsection 4.4, because the results seem to similar to the official results. First, we checked how much the accuracy changes by detail level of the labels as described in 6.1. Then we use other methods and data on Subtask1-CR and compared with results in Table 6 as the baseline, the results of the experiments are described in Sebsection 6.2 and 6.3.

6.1 Comparison with different level labelings

We examined whether the level of NER labels affect the accuracy of predictions. If the prediction accuracy was clearly deteriorated with detailed NER labels, some additional methods from other viewpoints are necessary, such as increasing the number of cases. We trained 3 models that are trained from different levels of NER labels, then compared prediction results of NEs that have at least 5 cases in the *train-test*. The levels are 1) ADE level that is used for Subtask3-CR (see Subsection 3.3), 2) detail level that is used for Subtask1-CR (see Subsection 3.2), and 3) coarse level that is summarized to "XML tags" in 2. All prediction results are summarized to coarse level for comparison. The results are shown in Table 10.

Table 10: Evaluation results in coarse level NE labels

NEs	count in *train-test*	NE label level		
		ADE level	detail level	coarse level
Disease	81	65.4	65.1	64.3
Anatomical	38	50.0	50.4	53.3
Feature	30	31.1	26.7	27.9
Change	16	28.6	28.6	27.8
Time	56	68.3	65.0	65.4
Test Name	22	79.3	80.7	75.4
Test Item	11	57.1	41.9	34.1
Test Value	7	63.2	46.2	48.0
Medicine Name	5	30.8	30.8	0.0
Remedy	56	56.8	59.4	56.8
Clinical Context	12	27.6	35.7	29.6

From these results, there wer no NEs whose accuracy was significantly deteriorated as the levels of NE labels become more detail. Therefore, subsequent experiments described in the following subsection were conducted using the detail level NE labels. That level is same as Subtask1-CR, which is considered easy to compare with official results.

6.2 Evaluation with other methods

We experimented with NegSampling-NER[10] and NER-BET-CRF[11] implementations and compared with baselines of SimpleTransformer's NER implementation. We changed the Japanese tokenization using pre-trained model's tokenizer. Both methods are based on BERT, the NegSampling-NER is a span-based NER method[4], and the NER-BET-CRF is a method with a CRF layer. The results are shown in Table 11.

The implementation of NegSampling-NER and NER-BET-CRF produced better results than that of SimpleTransformers. In the implementation of NER-BET-CRF, two results of BERT and BET-CRF were produced for comparison, but the accuracy of fine-tuned BERT was greatly improved from the SimpleTransformers implementation. Although detailed analysis of the model was not carried out, the NER-BERT-CRF has the the layer called prediction mask. It seems to be possible to improve the accuracy of SubTask1-CR and SubTask3-CR by changing the method used for the NER task to a more accurate method.

[10]https://github.com/LeePleased/NegSampling-NER
[11]https://github.com/Louis-udm/NER-BERT-CRF

Table 11: Evaluation results with other methods

Method	Language	F1
baseline1 (SimpleTransformers NER with mBERT)	English	47.8
	Japanese	61.0
baseline2 (SimpleTransformers NER with XLM-R)	English	47.0
	Japanese	65.6
NegSampling-NER	English	53.3
	Japanese	65.7
NER-BERT-CRF (BERT)	English	62.6
	Japanese	70.7
NER-BERT-CRF (BERT-CRF)	English	65.1
	Japanese	72.4

6.3 Evaluation with translated NE data

Replacing some words with another language works well with multilingual model in translation task[5]. Therefore, we replaced the words of NEs with another language and treat that as additional training corpus. We aligned sentences of the documents and treated NEs as translation dictionary if the NE with same tag and attribute was only one in paired sentences. The experiments with the methods described in Subsection 6.2 and the additional training data are shown in Table 12.

Table 12: Evaluation results with NE translated data

Method	Language	F1	diff
SimpleTransformers NER with mBERT	English	51.5	+4.3
	Japanese	60.0	-1.0
SimpleTransformers NER with XLM-R	English	49.6	+2.6
	Japanese	66.1	+0.5
NegSampling-NER	English	54.4	+1.1
	Japanese	65.6	-0.1
NER-BERT-CRF (BERT)	English	64.4	+1.8
	Japanese	68.9	-1.8
NER-BERT-CRF (BERT-CRF)	English	66.6	+1.5
	Japanese	70.1	-2.3

In English with all methods, F1 scores were improved over 1.0. In Japanese, the F1 score of XLM-R was slightly improved, but F1 scores of other methods using mBERT were deteriorated. Since there are various combinations in replacement of other languages, further verification is necessary.

7 CONCLUSIONS

We confirmed that the accuracy of fine-tuned multilingual pretrained models in Japanese can be higher or at least as high as those of English with only a small amount of data. In Japanese, characteristic substrings are used in disease names or anatomical names, therefore the difficulty of the task was less difficult than in English. This results suggest that translated texts may make easier to solve specific domain tasks. By additional experiments, we found the accuracy of base NER tasks can be improved with other methods and translated articles with annotation can be useful for even monolingual NER task.

REFERENCES

[1] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzman, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. *Unsupervised Cross-lingual Representation Learning at Scale.* Association for Computational Linguistics, Online. 8440–8451 pages. https://doi.org/10.18653/v1/2020.acl-main.747

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding *(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)).* Association for Computational Linguistics, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[3] Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (2018), 66–71. https://www.aclweb.org/anthology/D18-2012

[4] Yangming Li, lemao liu, and Shuming Shi. 2021. Empirical Analysis of Unlabeled Entity Problem in Named Entity Recognition. In *International Conference on Learning Representations.* https://openreview.net/forum?id=5jRVa89sZk

[5] Xiao Pan, Mingxuan Wang, Liwei Wu, and Lei Li. 2021. Contrastive Learning for Many-to-many Multilingual Neural Machine Translation *(Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)).* Association for Computational Linguistics, 244–258. https://doi.org/10.18653/v1/2021.acl-long.21

[6] Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. *Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets.* Association for Computational Linguistics, Florence, Italy. 58–65 pages. https://doi.org/10.18653/v1/W19-5006

[7] D. Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, Michael Young, Jean-Francois Crespo, and Dan Dennison. 2015. Hidden technical debt in Machine learning systems. , 2503–2511 pages.

[8] Yada Shuntaro, Nakamura Yuta, Wakamiya Shoko, and Aramaki Eiji. 2022. RealMedNLP: Overview of REAL document-based MEDical Natural Language Processing Task. *Proc. of the 16th NTCIR Conference on Evaluation of Information Access Technologies* (2022).

[9] Erik F. Tjong Kim Sang. 2002. *Introduction to the CoNLL-2002 Shared Task: Language-Independent Named Entity Recognition.* https://aclanthology.org/W02-2024

[10] Wonjin Yoon, Chan Ho So, Jinhyuk Lee, and Jaewoo Kang. 2019. CollaboNet: collaboration of deep neural networks for biomedical named entity recognition. *BMC Bioinformatics* 20, 10 (2019), 249. https://doi.org/10.1186/s12859-019-2813-6