Team Yseop at #SMM4H 2024: Multilingual Pharmacovigilance Named Entity Recognition and Relation Extraction

Anubhav Gupta
Yseop
agupta@yseop.com

Abstract

This paper describes three RoBERTa based systems. The first one recognizes adverse drug events (ADE) in English tweets and links them with MedDRA concepts. It scored `F1-norm` of **40** for the Task 1. The next one extracts pharmacovigilance related named entities in French and scored a `F1` of **0.4132** for the Task 2a. The third system extracts pharmacovigilance related named entities and their relations in Japanese. It obtained a `F1` of **0.5827** for the Task 2a and **0.0301** for the Task 2b. The French and Japanese systems are the best performing system for the Task 2[1].

1 Introduction

As of 2008, the European Commission proposed certain measures[2] to protect the public from the harm caused by Adverse Drug Reaction (ADR). Nonetheless, Koyama et al. (2023) observed a global increase in Adverse Drug Event (ADE) related deaths. Thus, pharmacovigilance, i.e. activities involving detection, comprehension, and prevention of adverse effects due to medication is an important subject. With the arrival of internet, the general public started using it to seek and share health related information[3] (Gerber and Eiser, 2001). With the help of natural language processing systems, this publicly available data can be analysed to extract information related to side effects. As a result, it can play a key role in strengthening pharmacovigilance reporting systems.

Note: Even though, the "ICH E2A Clinical safety data management: definitions and standards for expedited reporting"[4] distinguishes between ADR and ADE, in this paper we will use the terms interchangeably to refer to the unintended consequences of taking a prescribed medication.

SMM4H-2024: The 9th Social Media Mining for Health Research and Applications Workshop and Shared Tasks — Large Language Models (LLMs) and Generalizability for Social Media NLP (Xu et al., 2024) proposed 7 shared tasks. We participated in two of them:

- Task 1: Extraction and normalization of adverse drug events (ADEs) in English tweets
- Task 2: Cross-Lingual Few-Shot Relation Extraction for Pharmacovigilance in French, German, and Japanese

2 Task 1 - English

In this task, we have to identify ADEs in a short text, called "tweets", written in English. If ADEs are present, then they have to be mapped to Preferred Terms Id (`ptId`) from the Medical Dictionary for Regulatory Activities (MedDRA)[5].

Due to an error on our part, we did this task with the SMM4H -2023 Task 5 dataset, shared on the task's Google Groups on 06/02/2024.

2.1 Dataset

The training set had 17385 tweets, out of which only 1239 mentioned any ADE. The tweets and the annotations were provided in two separate files. An example of a tweet:

```
SMM4H2022uCZV2SRsCe4vzjFm
@USER__________ have to go to a doc
now to see why i'm still gaining.
stupid paxil made me gain like 50
pounds ?? and now i have to lose it
```

[1] All the models will shared on https://huggingface.co/yseop and the code will be available on https://github.com/yseop/YseopLab

[2] https://ec.europa.eu/commission/presscorner/detail/cs/MEMO_08_782

[3] https://web.archive.org/web/20150924101434/https://www.pushdoctor.co.uk/Resources/PushDoctor-Health-report.pdf

[4] https://database.ich.org/sites/default/files/E2A_Guideline.pdf

[5] https://www.meddra.org/

Proceedings of The 9th Social Media Mining for Health Applications (#SMM4H) Workshop and Shared Tasks, pages 136–141
August 15, 2024

The annotation file had the spans (start and end position in the tweet) and lowest level term[6] id (`llt`) of each ADE mention:

```
SMM4H2022uCZV2SRsCe4vzjFm
ADE 61 68 gaining 10047896
```

A MedDRA dictionary (**llt.asc**) containing the mapping between ADE, `llt`, and preferred term id (`ptid`) was also provided.

For the tweets in quotes we found that the spans were off by 1, we corrected that for the training.

2.2 Model

We augmented data with English texts from the SM-ADE sub-task (Wakamiya et al., 2023) to train a binary classifier that can distinguish between the tweets having ADE from those without ADE. The resulting classifier could get a F-1 score 0.73 on the validation set. This was less than the F1 achieved by Gupta and Rayar (2023)'s multilingual Bert model with similar dataset. Therefore, we did not use the classifier.

We fine-tuned RoBERTa[7] (Liu et al., 2019) for token classification[8] using Huggingface (Wolf et al., 2020). We preprocessed the training data by aligning the annotations with tokens. The first token of an ADE entity was labelled `B-MISC` and remaining tokens were labelled `I-MISC`. The non-ADE tokens in the input text were labelled `O`. While training, the model was evaluated on the validation set using seqeval's[9] `f1_score`.

We kept aside 20% of the training data for validation using scikit-learn's (Pedregosa et al., 2011) stratified `train_test_split` and fine-tuned the model on a) 80% of the train set, and b) on augmented training data consisting of the 80% of the provided training corpus and the non-ADE English data from the SM-ADE sub-task. There was not much of the difference in the performance (see Figure 1), so we did not submit the model trained on the augmented data.

The ADEs detected by the first model (see Table 1 for the parameters used) were searched in the MedDRA dictionary with the help of Sentence-Transformers'[10] (Reimers and Gurevych, 2019) `multi-qa-mpnet-base-dot-v1` model. The results (sub1.zip) were submitted to the leader board.

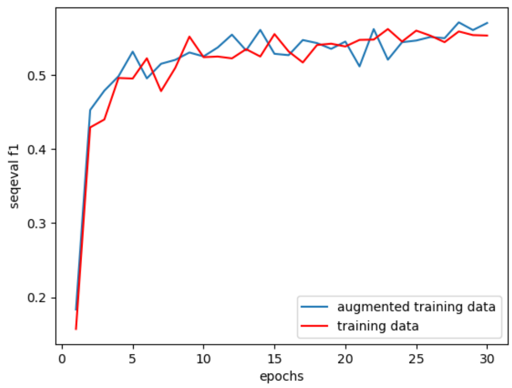

Figure 1: Comparison of the raw training data with the augmented one.

2.3 Test Results

The model obtained a `F1-norm` score of **40.0** compared to **43.9** by the baseline provided by DeepADEMiner (Magge et al., 2021). `-Norm` metrics are calculated by comparing the normalized ADE, i.e. `llt`. Table 2 compares the performance of the model with the baseline and the median of all the submissions.

3 Task 2 - French and Japanese

The goals of this task were Named Entity Recognition (NER) and Relation Extraction (RE). The corpus consisted of, mainly, German and Japanese text taken from online forums and Twitter/X. It also had four French documents translated from German texts.

3.1 Dataset

The French dataset had 4 documents and the Japanese had 392. The documents were annotated in the brat standoff format[11] (Stenetorp et al., 2011). There are 3 entities and 2 relations:

- Entities:
  - DISORDER, a symptom not necessarily related to a drug
  - DRUG
  - FUNCTION, bodily functions
- Relations:

[6] https://www.meddra.org/how-to-use/basics/hierarchy

[7] https://huggingface.co/FacebookAI/roberta-base

[8] https://huggingface.co/docs/transformers/en/tasks/token_classification

[9] https://github.com/chakki-works/seqeval

[10] https://www.sbert.net/index.html

[11] https://brat.nlplab.org/standoff.html

Training Parameters	Task 1	Task 2a	Task 2b
tokenizer max length	512		
learning rate	1e-05		
weight decay	0.001	0.0	0.0
epochs	30	50	15
batch size	16		
machine	ml.g5.xlarge		

Table 1: Non default hyperparameters used for fine-tuning.

System	F1-Norm	P-Norm	F1-NER	F1-Norm-Unseen
`sub1.zip`	40	39.6	47.2	29.5
Median	29.3	33.9	37.6	14.1
Baseline	43.9	39.3	48.1	32.3

Table 2: Performance on the Task 1 test set.

- CAUSED, the first entity causes the second entity
- TREATMENT_FOR, the first entity remedies the second one

An example of French data with annotation:

```
Salut <user>, pour moi, ça a commencé
à l'âge de <pi>. J'ai suivi une
thérapie et une cure pendant deux ans,
...
Prends soin de toi. <user>

T4   DRUG 123 130    pilules
...
R3   CAUSED Arg1:T10 Arg2:T15
R13  TREATMENT_FOR Arg1:T24 Arg2:T26
```

An example of Japanese data with annotation:

```
881844583344201728:
昼間のレクサプロが副作用ひどくて未
だに気持ち悪い

T1    DRUG 22 27        レクサプロ
T2    DISORDER 28 31    副作用
T3    DISORDER 38 43    気持ち悪い
R1    CAUSED Arg1:T1 Arg2:T2
```

There were some inconsistencies in the data, for example the Japanese term 錠 was not annotated as `DRUG` in most of the documents. Whereas the French equivalent **pilules** (see the example above) and other common nouns such as 薬 and 製品 were.

In the Japanese training, the span of certain entities was updated as shown in the Table 3.

File	Entity	Span
ja_twjp_020-040_0	T1	36 42
ja_twjp_200-220_1	T15	47 55
ja_twjp_240-260_8	T8	140 142
ja_twjp_320-340_4	T1	109 114
ja_twjp_340-360_14	T13	144 151
ja_twjp_440-460_19	T8	73 77
ja_twjp_460-480_18	T6	20 26

Table 3: Annotation files that were corrected.

3.2 Task 2a - French Model

Since there was not enough training data, we decided to use `Mistral-7B-Instruct-v0.1`[12] (Jiang et al., 2023) and DrBERT-CASM2[13] via the `medkit`[14] library. DrBERT (Labrak et al., 2023) was fine-tuned on CASM2 corpus for NER task to produce DrBERT-CASM2. The CASM2 is a private corpus that contains documents from CAS (Grabar et al., 2018).

The Mistral LLM is prompted with the prompt described in Appendix A and parameters `do_sample=False` and `max_new_tokens=256`. If the entities returned by the LLM are in the text they are added to the list of candidates. Then, all the entities extracted by DrBERT-CASM2 are added to the candidates. Lastly, the entities in the candidate list are used to find other substrings in the text.

[12] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.1

[13] https://huggingface.co/camila-ud/DrBERT-CASM2

[14] https://medkit.readthedocs.io/en/stable/index.html

3.3 Task 2a - Japanese Model

Baseline: If we label all occurrences of 副作用 in the training set as `DISORDER`; 薬 and 製品 as `DRUG`, we get a Macro F1 of 0.1658. If we also annotate all consecutive sequence of katakana characters that are not present in the `JMdict_e`[15] (Japanese–Multilingual Dictionary) as `DRUG`, the Macro F1 becomes 0.2842.

We kept aside the 20% of the provided train dataset as validation set using scikit-learn's stratified `train_test_split`. Then on the remaining dataset we fine-tuned `daisaku-s/medtxt_ner_roberta`[16] as token classification task. This model was previously trained on MedTxt-CR dataset (Yada et al., 2022). The seqeval F1 score was better than the baseline and hence it was submitted to the leader-board.

3.4 Task 2b - Japanese Model

The training data contains:

- 390 examples of DRUG causing DISORDER
- 100 examples of DRUG TREATMENT_FOR DISORDER
- 98 examples of DISORDER causing DISORDER
- 20 examples of DRUG causing FUNCTION
- 8 examples of DISORDER causing FUNCTION
- 3 examples of DRUG TREATMENT_FOR FUNCTION

out of 8497 possible relations.

From the train set, we created a new corpus for relation classification. Similar to Zhong and Chen (2021), we extracted, for each pair of entities, the text between them (entities included). If there is no relation between the pair, it is annotated as 'O', otherwise the label in the train set was used. For the example in Figure 2, we take the text span 抗うつ剤に関しては抵抗がありましたが、安酒を煽るより**100**倍は建設的な精神状態 and annotate it as CAUSED.

We kept aside the 20% of the new corpus as validation set. Then on the remaining dataset we fine-tuned `daisaku-s/medtxt_ner_roberta` as sequence classification task[17].

[15] http://www.edrdg.org/wiki/index.php/JMdict-EDICT_Dictionary_Project

[16] https://huggingface.co/daisaku-s/medtxt_ner_roberta

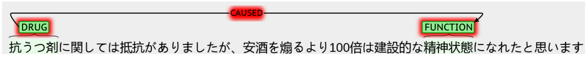

Figure 2: Example of CAUSED relation.

Task	Precision	Recall	F1
Task 2a - Fr	0.6068	0.3133	0.4132
Task 2a - Ja (dev)	0.5873	0.3581	0.4449
Task 2a - Ja	0.5752	0.5903	0.5827
Task 2b - Ja (dev)	0.1852	0.0564	0.0865
Task 2b - Ja	0.0226	0.0449	0.0301

Table 4: Performance on the Task 2 test set.

3.5 Test Results

The results are presented in Table 4. One of the reasons for the bad performance of the Japanese model is tokenization error. For example, in the text `ja_twjp_060-080_19` one of the entities is **DISORDER 47 50** 副作用, however at the given span, the `daisaku-s/medtxt_ner_roberta` tokenizer returns の副作用 as the single token.

4 Conclusion

In Task 1, despite training on the wrong dataset we managed to be in the top 50 percentile. The difficult part was normalization of the ADE using MedDRA dictionary, as a result F1-Norm was lower than F1-NER. For the Task 2, using a model adapted to the clinical domain helped to get the best results. The Task 2b (relation extraction) was challenging, given that the winning team obtained an overall F1 score of 0.0189. For future, We would explore approaches such as GLiNER (Zaratiana et al., 2023) and XMC (D'Oosterlinck et al., 2024) to improve NER in Task 1 and Task 2a.

References

Karel D'Oosterlinck, Omar Khattab, François Remy, Thomas Demeester, Chris Develder, and Christopher Potts. 2024. In-context learning for extreme multi-label classification. *arXiv preprint arXiv:2401.12178*.

Ben S Gerber and Arnold R Eiser. 2001. The patient-physician relationship in the internet age: future prospects and the research agenda. *Journal of medical Internet research*, 3(2):e842.

[17] https://huggingface.co/docs/transformers/en/tasks/sequence_classification

Natalia Grabar, Vincent Claveau, and Clément Dalloux. 2018. CAS: French corpus with clinical cases. In *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis*, pages 122–128, Brussels, Belgium. Association for Computational Linguistics.

Anubhav Gupta and Frédéric Rayar. 2023. Frag at the ntcir-17 mednlp-sc task. In *Proceedings of the 17th NTCIR Conference on Evaluation of Information Access Technologies*, Tokyo, Japan. NII Institutional Repository.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Toshihiro Koyama, Shunya Iinuma, Michio Yamamoto, Takahiro Niimura, Yuka Osaki, Sayoko Nishimura, Ko Harada, Yoshito Zamami, and Hideharu Hagiya. 2023. International trends in adverse drug event-related mortality from 2001 to 2019: An analysis of the world health organization mortality database from 54 countries. *Drug Saf*, 47(3):237–249.

Yanis Labrak, Adrien Bazoge, Richard Dufour, Mickael Rouvier, Emmanuel Morin, Béatrice Daille, and Pierre-Antoine Gourraud. 2023. DrBERT: A Robust Pre-trained Model in French for Biomedical and Clinical domains. In *Proceedings of the 61th Annual Meeting of the Association for Computational Linguistics (ACL'23), Long Paper*, Toronto, Canada. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Arjun Magge, Elena Tutubalina, Zulfat Miftahutdinov, Ilseyar Alimova, Anne Dirkson, Suzan Verberne, Davy Weissenbacher, and Graciela Gonzalez-Hernandez. 2021. DeepADEMiner: a deep learning pharmacovigilance pipeline for extraction and normalization of adverse drug event mentions on Twitter. *Journal of the American Medical Informatics Association*, 28(10):2184–2192.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Pontus Stenetorp, Goran Topić, Sampo Pyysalo, Tomoko Ohta, Jin-Dong Kim, and Jun'ichi Tsujii. 2011. Bionlp shared task 2011: Supporting resources. In *Proceedings of BioNLP Shared Task 2011 Workshop*, pages 112–120, Portland, Oregon, USA. Association for Computational Linguistics.

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. 2023. 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*, Tokyo, Japan. NII Institutional Repository.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Dongfang Xu, Guillermo Lopez Garcia, Lisa Raithel, Rolland Roller, Philippe Thomas, Eiji Aramaki, Shuntaro Yada, Pierre Zweigenbaum, Sai Tharuni Samineni, Karen O'Connor, Yao Ge, Sudeshna Das, Abeed Sarker, Ari Klein, Lucia Schmidt, Vishakha Sharma, Raul Rodriguez-Esteban, Juan Banda, Ivan Flores Amaro, Davy Weissenbacher, and Graciela Gonzalez-Hernandez. 2024. Overview of the 9th social media mining for health applications (#SMM4H) shared tasks at ACL 2024. In *Proceedings of The 9th Social Media Mining for Health Research and Applications Workshop and Shared Tasks*, Bangkok, Thailand. Association for Computational Linguistics.

Shuntaro Yada, Yuta Nakamura, Shoko Wakamiya, and Eiji Aramaki. 2022. Real-mednlp: Overview of real document-based medical natural language processing task subtasks. In *Proceedings of the 16th NTCIR Conference on Evaluation of Information Access Technologies*, Tokyo, Japan. NII Institutional Repository.

Urchade Zaratiana, Nadi Tomeh, Pierre Holat, and Thierry Charnois. 2023. Gliner: Generalist model for named entity recognition using bidirectional transformer. *Preprint*, arXiv:2311.08526.

Zexuan Zhong and Danqi Chen. 2021. A frustratingly easy approach for entity and relation extraction. In *North American Association for Computational Linguistics (NAACL)*.

A Mistral Prompt Template

The LLM input used for NER is a few-shot prompt containing two examples. It uses the text from fr_1025_lifeline_v1_FR_1971_1_1647857960 and fr_1069_lifeline_v1_FR_6168_1_1648459053 as examples. The entities and the relations for each example is used as shown below:

<s>[INST] From the medical report in French below, extract all the mentions of entities DRUG, DISORDER and the relationships CAUSED and TREATMENT_FOR in brat format.
TEXT 1[/INST] DRUG pilules

```
DISORDER    angoisses
FUNCTION    règles
...
DISORDER    humeur
FUNCTION    hormones
CAUSED  Arg1:pilules    Arg2:angoisses
...
TREATMENT_FOR  Arg1:Insidon   Arg2:humeur
```

</s>[INST] From the medical report in French below, extract all the mentions of entities DRUG, DISORDER and the relationships CAUSED and TREATMENT_FOR in brat format.
TEXT 2 [/INST] DRUG mirtazapine

```
DISORDER    problèmes de sommeil
FUNCTION    dors
...
CAUSED  Arg1:antiémétiques  Arg2:somnolence
...
TREATMENT\_FOR Arg1:zolpidem Arg2:troubles du sommeil
```

</s>[INST] From the medical report in French below, extract all the mentions of entities DRUG, DISORDER and the relationships CAUSED and TREATMENT_FOR in brat format.
TEXT FROM THE TEST SET [/INST]