



HAL
open science

KEEPHA at n2c2 2022: Track 1

Lisa Raithel, Faith W. Mutinda,, Gabriel H. B. Andrade, Hui-Syuan Yeh,
Tomohiro Nishiyama, Mathieu Lai-King, Shuntaro Yada, Roland Roller, Cyril
Grouin, Agata Savary, et al.

► **To cite this version:**

Lisa Raithel, Faith W. Mutinda,, Gabriel H. B. Andrade, Hui-Syuan Yeh, Tomohiro Nishiyama, et al..
KEEPHA at n2c2 2022: Track 1. 2022 n2c2 Shared Task and Workshop at AMIA Fall Symposium,
Nov 2022, Washington, DC, United States. hal-04297589

HAL Id: hal-04297589

<https://hal.science/hal-04297589v1>

Submitted on 21 Nov 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

KEEPHA at n2c2 2022: Track 1

Lisa Raithel, MSc^{1,2,3*}, Faith W. Mutinda, MSc^{4*}, Gabriel H. B. Andrade, MSc^{4*}, Hui-Syuan Yeh, MSc^{1*}, Tomohiro Nishiyama, BSc⁴, Mathieu Laï-King, MSc¹, Shuntaro Yada, PhD⁴, Roland Roller, PhD², Cyril Grouin, PhD¹, Agata Savary, PhD¹, Aurélie Névéol, PhD¹, Thomas Lavergne, PhD¹, Eiji Aramaki, PhD^{4‡}, Sebastian Möller, PhD^{2,3‡}, Yuji Matsumoto, PhD^{5‡}, Pierre Zweigenbaum, PhD^{1‡}

¹Université Paris-Saclay, CNRS, Laboratoire interdisciplinaire des sciences du numérique, 91405, Orsay, France; ²SLT Lab, DFKI, 10559, Berlin, Germany; ³Technische Universität Berlin, 10623, Berlin, Germany; ⁴Nara Institute of Science and Technology, Nara, Japan; ⁵RIKEN Center for Advanced Intelligence Project, Tokyo, Japan

Abstract

Team KEEPHA¹ submitted results for all three subtasks based on ensembles and inserting position information using different variations of transformer models. We achieve a strict F1 score of 0.95 (lenient F1 is 0.97) in medication extraction, an F1 of 0.94 (micro) and 0.87 (macro) in event classification and a combined F1 of 0.67 in context classification.

Introduction

This abstract describes our participation in the n2c2 2022 shared task 1: contextualized medication event extraction on the CMED corpus [1]. We submitted three runs to subtask 1, six runs to subtask 2, and five runs to subtask 3. All models were trained using the huggingface library [5] and AdamW [4] for optimization. We only report the best models for every subtask.

Subtask 1: Medication Extraction: An ensemble of the best two PubMedBERT [2] models trained with two configurations with five seeds each. We trained the first model on sequence chunks of size 30 and batch size 4, the second one on chunks of size 20 and batch size 8, both with a learning rate of $2e-5$. This approach achieved the best results with a strict F1 score of 0.95 and a lenient F1 score of 0.97. Pre- and post-processing were done with the help of the original brat-related [3] scripts².

*equal contribution; ‡ PIs of project KEEPHA

¹ <https://keepha.lisn.upsaclay.fr/wiki/doku.php?id=partners>

² <https://github.com/spyysalo/standoff2conll> and <https://github.com/nlplab/brat/blob/master/tools/BIOtoStandoff.py>

Subtask 2: Event Classification: For this task, we pre-processed the dataset to generate chunks of text with a 200-character context window around each drug mention. We fed the pre-processed data to a BioClinical BERT [6] model, pretrained on electronic health records. We fine-tuned it for 100 epochs, with AdamW optimized linear decreasing learning rate. It achieved a micro F1 score of 0.93 and a macro F1 score of 0.83 on the gold labeled dataset.

Subtask 3: Context Classification: We pre-processed the dataset as in subtask 2. To obtain rough document-position information, we split each document into three equal-length chunks: head, body, and tail. We appended that information to inputs after a [SEP] token. We produced the results with task-specific ensembles of BioClinical BERT models [6] over five seeds trained with document-position information, one for each dimension of context.

Conclusion

In conclusion, for the first subtask, we found that ensembling models work best, and that remaining false negatives are often mentions that either contain spelling mistakes or that could be both body substances and medication (e.g. “insulin”). For subtask 2, BioClinical BERT model achieved the best performance, especially in classification of Disposition and NoDisposition events, while Undetermined events had a lower score due to low frequency in the training set. For subtask 3, we found that incorporating rough document-position information helps in event classification.

References

1. Mahajan D, Liang JJ, Tsou CH. [Toward understanding clinical context of medication change events in clinical narratives](#). arXiv:201108835 [cs]. 2021
2. Gu Y, Tinn R, Cheng H, Lucas M, Usuyama N, Liu X, Naumann T, Gao J, Poon H. [Domain-specific language model pretraining for biomedical natural language processing](#). ACM Transactions on Computing for Healthcare (HEALTH). 2021 Oct 15;3(1):1-23
3. Stenetorp P, Pyysalo S, Topic G, Ohta T, Ananiadou S, Tsujii J. [brat: a Web-based tool for NLP-assisted text annotation](#). EACL. 2012;102--107
4. Loshchilov I, Hutter F. Decoupled weight decay regularization. International Conference on Learning Representations. 2018;18
5. Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, et al. [Transformers: state-of-the-art natural language processing](#). EMNLP: System Demonstrations. Online: ACL; 2020:38–45.
6. Alsentzer E, Murphy J, Boag W, Weng W-H, Jindi D, Naumann T, et al. [Publicly available Clinical BERT embeddings](#). In: Proceedings of the 2nd Clinical Natural Language Processing Workshop. Minneapolis, Minnesota, USA: ACL; 2019:72–8. [accessed 4 Jun 2020]