MedMine

MedMine: Examining Pre-trained Language Models on Medication Mining

Haifa Alrdahi [1*], **Lifeng Han** [1*], **Hendrik Šuvalov** [2*], and **Goran Nenadic**[1]
[1] The University of Manchester, UK
[2] The University of Tartu, Estonia
∗ co-first authors
firstname.lastname @ {manchester.ac.uk | ut.ee}

Abstract

Automatic medication mining from clinical and biomedical text has become a popular topic due to its real impact on healthcare applications and the recent development of powerful language models (LMs). However, fully-automatic extraction models still face obstacles to be overcome such that they can be deployed directly into clinical practice for better impacts. Such obstacles include their imbalanced performances on different entity types and clinical events. In this work, we examine current state-of-the-art pre-trained language models (PLMs) on such tasks, via fine-tuning including the monolingual model Med7 and multilingual large language model (LLM) XLM-RoBERTa. We compare their advantages and drawbacks using historical medication mining shared task data sets from n2c2-2018 challenges. We report the findings we get from these fine-tuning experiments such that they can facilitate future research on addressing them, for instance, how to combine their outputs, merge such models, or improve their overall accuracy by ensemble learning and data augmentation. MedMine is part of the M3 Initiative https://github.com/HECTA-UoM/M3

1 Introduction

Medication mining plays a vital role in clinical natural language processing (ClinicalNLP) applications and digital healthcare settings. For instance, the extraction of medications used in patients' historical electronic health records and the corresponding targeted treatments can be very important features for the cohort selection of certain diseases and treatments. Medications and corresponding adverse drug effects extraction can be studied for future optimised and personalised treatments. Medication extraction itself can also be beneficial to epidemiological studies and term management (Ananiadou and Nenadic, 2006), e.g. to rheumatologists, and the extracted medical terminologies and concepts can be useful for knowledge transformation (Han et al., 2023; Han et al., 2022).

Medication extraction has been an application field of NLP models for decades across statistical and neural NLP methods (Spasić et al., 2010; Sarker et al., 2018). With the breakthrough of advanced learning structures based on Transformer and pre-trained language models (PLMs) BERT (Devlin et al., 2019), researchers have reported new results on clinical terminology mining. In this project, we aim at re-examining current state-of-the-art medication mining and large language models (LLMs) on medication extraction tasks by carrying out fine-tunings. We investigate the strengths and weaknesses of these models, explore the possibility to unify or merge their outputs, and have the goal in mind to develop an augmented medication mining toolkit and platform for open research, which we name as **MedMine**.

To the best of our knowledge, MedMine is the first attempt (or one of the first) to integrate existing state-of-the-art information extraction and LLMs in the healthcare and clinical domain, especially on medication extraction tasks.

2 Related Work

Some relevant work to ours includes the following categories: 1) *Medication mining in social medial* data to monitor medication abuses. For instance, Sarker et al. (2016) investigated the possibility of drug overuse detection using Twitter user posts from the USA on medications including Adderall, oxycodone, and quetiapine, in comparison to controlled medication metformin. They annotated 6,400 tweets manually and trained the classifier using LibSVM and Weka toolkit, which achieved

around 46% of the F1 score on prediction.

2) *Medication mining to identify risks* and improve treatments. For instance, Härkänen et al. (2019) used around 70K medication incident reports from England and Wales and conducted relevant events mining around the risk areas. The outcomes of identified areas included allergic reactions, intravenous administration of antibacterial drugs, and Fentanyl patches, etc. that need more attention. Bereznicki et al. (2008) carried out case studies by looking into asthma patients' medication records using data mining software applications from Australia. Insightful observations were achieved from statistical analysis and intervention letters were sent to patients to visit their GP.

3) *Prediction of next prescribed medications*. Wright et al. (2015) took diabetes patients' medication records and carried out the next prescribed medication prediction task using sequential pattern mining and cSPADE toolkit (Buchta et al., 2023). Out of 161,497 patients' data, their experiment achieved around 90% evaluation scores on the drug class level and 64% at the generic drug level. This work also indicated the usefulness of temporal information in relationship modelling of medications.

4) *Integrating NLP models for healthcare*. Instead of medication mining, MedCAT is a platform developed by Kraljevic et al. (2021) focusing on diagnoses extraction tasks from clinical text. The integrated models inside MedCAT [1] include statistical ones such as CRF, and neural models such as Transformers, as well as their combinations. MedCAT also has an entity linking function to normalise the extracted entities into existing clinical terminologies databases such as SNOMED CT and UMLS.

Some very relevant shared tasks and workshops (WS) include the n2c2 challenge series [2] which data we will use in this paper for the experimental evaluations, the Bio-medicine and its Applications NLP WS (BioNLP) (Collier et al., 2004; Demner-fushman et al., 2023), Louhi WS on Text and Data Mining of Health Documents (Lavelli et al., 2022; Dalianis et al., 2010). In the UK, we also organised the HealTAC conference series on Healthcare Text Analytics [3].

[1] https://github.com/CogStack/MedCAT

[2] https://n2c2.dbmi.hms.harvard.edu

[3] http://healtex.org 6th Edition this year hosted in the University of Manchester

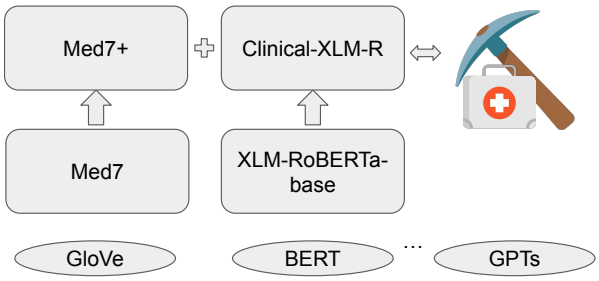

Figure 1: MedMine Illustration: Models already included Glove and RoBERTa embeddings, clinical domain fine-tuning for medication mining.

3 Methodologies and Experiments

In this section, we introduce **MedMine**, the first version of our Medication Mining Integration project.

In MedMine.V0, we try to explore the current large language models (LLMs), especially BERT-based models (Devlin et al., 2019) in comparison to pre-BERT embeddings. The two models we chose are 1) Med7 (Kormilitzin et al., 2021) which is a named entity recognition (NER) model fine-tuned in clinical domain medical records using SpaCy's library of GloVe embedding (Pennington et al., 2014), and 2) XLM-RoBERTa-base which is a multilingual pre-trained RoBERTa model (Conneau et al., 2020) using 100 languages from CommonCrawl corpora. The goal is to achieve fine-tuning on clinical domain data for medication mining tasks as demonstrated in Figure 1.

3.1 Models and Data

The two off-the-shelf models we deployed in our experiments are from HugingFace XLM-RoBERTa-base [4] and Github Med7 [5]. For fine-tuning data, we used the n2c2-2018 track-2 shared task corpus which originally includes 303 letters for training and 202 letters for testing. The corpus was annotated manually by four physician assistant students and three nurses (Henry et al., 2020). The corresponding sentence counts are 46,033 and 30,614 from the original training and testing set (Wu et al., 2022).

This shared task is on Adverse Drug Events and Medication Extraction in Electric Health Records (EHRs). To prepare a new data setting with a validation set, we split the original 505 annotated letters

[4] https://huggingface.co/xlm-roberta-base

[5] https://github.com/kormilitzin/med7

into 70/15/15% for training, development, and testing which corresponds to the following set (353, 76, 76).

3.2 Fine-tuning Parameters

For Med7, the fine-tuning was using 30 iterations (epochs), and for XML-R-Base, the fine-tuning parameters are displayed below.

- batch size = 16
- evaluation strategy = "epoch",
- learning rate=1e-4,
- num train epochs=8,
- weight decay=1e-5
- metric = load metric("seqeval")
- tokenizer = true

3.3 Med7+ Outputs

Med7 has only 7 labels in the outputs including Dosage, Drug, Duration, Form, Frequency, Route, and Strength, but our fine-tuned Med7+ has 9 labels fine-tuned from n2c2-2018 shared task with ADE and Reason labels.

Med7 Deployment using the 7 labels tested on 76 letters is evaluated in Table 1 with precision, recall, f1-score, micro avg, macro avg, and weighted avg. The "support" column is the number of real labels in the reference. The evaluation score shows that Med7 has really low performances on the 'dosage' label achieving 0.11 for Precision, 0.24 for Recall, and 0.15 for F1, even though the number of true labels on dosage (1039) is not much smaller than most of the other labels.

The fine-tuned Med7+ is reported in Table 2 using the same test set where it gives 9 labels. Furthermore, for the existing 7 labels, the Med7+ produced much higher scores than the Med7 baseline model across most of the labels, boosting the individual F1 scores up to around 0.90+ except for the 'duration' label, which is improved from 0.74 to 0.78 (the first section of Table 2). The bottom two sections of Table 2 analyse the evaluation scores when we remove the 'O' label and also the 'Reason' and 'ADE' labels. The bottom score by removing all three labels is to compare our fine-tuned Med7+ directly to the original Med7 baseline on the 7 labels, which says that we have improved the micro, macro, and weighted avg scores all from 0.70s to 0.90s.

Catergory	Pre.	Rec.	F1	Support
form	0.90	0.90	0.90	1696
strength	0.70	0.80	0.75	1639
dosage	0.11	0.24	0.15	1039
drug	0.90	0.77	0.83	3954
route	0.96	0.94	0.94	1341
frequency	0.74	0.79	0.76	1564
duration	0.73	0.75	0.74	139
micro avg	0.71	0.77	0.74	11372
macro avg	0.72	0.74	0.72	11372
weighted avg	0.78	0.77	0.77	11372

Table 1: Outputs from Med7 Deployment on 76 Testing Letters - Evaluation: Type

Catergory	Pre.	Rec.	F1	Support
O	0.0000	0.0000	0.0000	874
reason	0.7276	0.4552	0.5601	927
ade	0.5579	0.2190	0.3145	242
form	0.9229	0.9393	0.9310	1696
strength	0.9749	0.9494	0.9620	1639
dosage	0.9124	0.8816	0.8967	1039
drug	0.9345	0.9135	0.9239	3954
route	0.9580	0.9366	0.9472	1341
frequency	0.8502	0.9399	0.8928	1564
duration	0.8015	0.7554	0.7778	139
accuracy			0.8187	13415
macro avg	0.7640	0.6990	0.7206	13415
weighted avg	0.8454	0.8187	0.8282	13415
Removing label 'O': only using 9 labels				
micro avg	0.9124	0.8758	0.8937	12541
macro avg	0.8489	0.7767	0.8007	12541
weighted avg	0.9044	0.8758	0.8859	12541
Removing labels 'O', 'Reason', and 'ADE': check the improvement on 7 labels				
micro avg	0.9248	0.9240	0.9244	11372
macro avg	0.9078	0.9022	0.9045	11372
weighted avg	0.9262	0.9240	0.9247	11372

Table 2: Outputs from Med7+ fine-tuning using 9labels on 76 Letters - Evaluation: Type

Both Table 1 and Table 2 reported the evaluation using "Lenient Matching" which corresponds to the "Type" category out of the four different Evaluation Strategies introduced from SemEval2013 (Manandhar and Yuret, 2013) [6].

[6] https://paperswithcode.com/dataset/semeval-2013

- Strict: “exact boundary surface string match and entity type”;
- Exact: “exact boundary match over the surface string, regardless of the type”;
- Partial: “partial boundary match over the surface string, regardless of the type”;
- **Type**: “some overlap between the system tagged entity and the gold annotation is required”.

3.4 Clinical-XLM-R Outputs

We name the fine-tuned XLM-RoBERTa-base as Clinical-XLM-R which produces all the 9 labels output as shown in Table 3 using the same evaluation strategy we used for Med7+ “Lenient Matching”.

Acc	Pre	Rec	F1	
96.76%	87.98%	90.14%	89.05%	
Catergory	Pre.	Rec.	F1	Num.
ADE	51.96%	55.09%	53.48%	432
Dosage	85.94%	91.40%	88.59%	803
Drug	93.05%	95.63%	94.32%	8193
Duration	58.40%	67.34%	62.55%	98
Form	91.69%	90.46%	91.07%	1647
Frequency	88.13%	90.09%	89.10%	1838
Reason	59.57%	62.01%	60.77%	1369
Route	90.90%	91.28%	91.09%	1205
Strength	94.83%	96.00%	95.41%	1377

Table 3: Outputs from Clinical-XLM-RoBERTa

In Table 3, the first row is the overall Accuracy, Precision, Recall, and F1 score. The column 'Num.' is the number of entities the model predicted. Some interesting findings from Clinical-XLM-R are listed below:

- 1) ADE, Duration, and Reason are the three most challenging categories to identify, especially since their precision scores are under 60% while others are around 90%.
- 2) Most categories have relatively higher Recall than Precision scores except for ‘Form’ labels, which means the system has more false positive labels than false negative labels. One of the future directions to optimise the model is to assign more attention to restrict false positive labels. However, this also depends on the practical application of medication and related event extraction. For instance, if a High Recall score is preferred to identify all useful information, the model is doing a good job at his stage.
- 3) The tokenizer we used for this model is from the XLM-RoBERTa-base package. We handled documents that were larger than 512 by splitting them up and dividing them into 512-sized chunks. The downside of this is that it might lose some context information from the beginning of the document.

4 Discussion

Looking into both Med7+ and Clinical-XLM-R regarding this specific task on Drugs and related events extraction using n2c2-2018 challenge data, there are some research directions worthy to be explored.

- 1) How to address the label imbalance issue? e.g. Duration and ADE have apparent much lower rates of labels (139, 242) out of 13,415, while other labels have around 1K labels, in addition to ‘Drug’ having around 4K labels. Data augmentation via oversampling or synthetic data generation can be a possible solution toward a more balanced data setting, which hopefully can improve model accuracy on those low-frequency label predictions.
- 2) Med7 has kind of balanced Precision and Recall scores. While Med7+ improved Med7 scores in each category except for the label ‘route’ which stayed almost the same scores (0.96-¿0.96, 0.94-¿0.94, 0.94-0.95), the Precision scores of Med7+ are in general higher than the Recall scores except for ‘form’ and ‘frequency’. This is contradictory to Clinical-XLM-R, which has in general higher Recall scores instead. It is our next step to look into the fine-tuning details of these two models, and hopefully, we can enhance both models from each other’s advantages.
- 3) Clinical-XLM-R boosted the overall Accuracy score to 96.76, in comparison to 81.87 from Med7+. This might indicate that the BERT-based word embeddings learn can learn better performance than GloVe embeddings on this task; however, it is also possible that the multilingual training data XLM-RoBERTa

used is helpful for pre-training, even though Med7 has been pre-trained on their available clinical domain data-set. We need to look into this in detail for model explanation.

5 Conclusion and Future Work

As part of the M3 Initiative (Hussain et al., 2023), we reported the MedMine project progress using clinical domain fine-tuning of two PLMs i.e. Med7 and XLM-R. Med7 itself is already a fine-tuned model using clinical domain data that reports 7 labels, while XLM-RoBERTa is a multilingual pre-trained general domain large language model (LLM). The outcomes of Med7+ and Clinical-XLM-R both demonstrated much higher performances than the Med7 baseline model across most of the 9 original labels using n2c2-2018 shared task data. Comparisons restricted on the 7 labels reported by Med7, Med7+ produced F1 scores (0.9244, 0.9045, 0.9247) vs (0.74, 0.72, 0.77) from Med7 model, for micro avg, macro avg, and weighted avg using our experimental setup.

In future work, we plan to look into example outputs with expert humans regarding different models and investigate the strategies on merging their outputs. We also have an interest in expanding integrated models into GPTs as in Figure 1, and explore different prompt-based learning (PBL) mechanisms for medication extraction using both manual and soft templates, such as mixed template used by Cui et al. (2023) for temporal modelling of medications.

Limitations

In this work, we used n2c2-2018 shared task data, which limits the size of the training and testing set. It is our plan for the next steps to extend the training set and label categories e.g. 'Temporal' by combining more historical n2c2/i2b2 challenge data such as the ones from 2009 and 2012. This will involve merging annotated data sets taking into account their differences in formats.

Ethical Concerns

The data set we used from the n2c2 challenge is anonymised and we have gone through good clinical practice training to access the data, without sharing them with any third parties.

Acknowledgments

We thank Arooj Hussain, Mingyang Li, Yuping Wu, and Warren Del-Pinto for their input on the ongoing M3 project and valuable discussion. This project is partially funded by the University of Manchester Open Research Office via OR Fund Project "An Open-Research Framework on Label Augmentations for Low-Resource Clinical Natural Language Processing using Graph-Based Semi-Supervised Learning" and the UKRI/EPSRC grant EP/V047949/1 "Integrating hospital outpatient letters into the healthcare data space".

Author Contributions

HA: Med7 deployment and fine-tuning; HS: XLM-R fine-tuning; LH: supervised the project and drafted the paper; GN: co-supervisor of MedMine and discussion.

References

[Ananiadou and Nenadic2006] Sophia Ananiadou and Goran Nenadic, 2006. *Automatic Terminology Management in Biomedicine*, pages 67–98. Artech House, Inc, United Kingdom.

[Bereznicki et al.2008] Bonnie J Bereznicki, Gregory M Peterson, Shane L Jackson, E Haydn Walters, Kimbra D Fitzmaurice, and Peter R Gee. 2008. Data-mining of medication records to improve asthma management. *Medical journal of Australia*, 189(1):21–25.

[Buchta et al.2023] Christian Buchta, Michael Hahsler, Daniel Diaz, Maintainer Christian Buchta, and Mohammed J Zaki. 2023. Package 'arulessequences'. *Cited on page*.

[Collier et al.2004] Nigel Collier, Patrick Ruch, and Adeline Nazarenko, editors. 2004. *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA/BioNLP)*, Geneva, Switzerland, August 28th and 29th. COLING.

[Conneau et al.2020] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. 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. Association for Computational Linguistics.

[Cui et al.2023] Yang Cui, Lifeng Han, and Goran Nenadic. 2023. MedTem2.0: Prompt-based temporal classification of treatment events from discharge

summaries. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 160–183, Toronto, Canada, July. Association for Computational Linguistics.

[Dalianis et al.2010] Hercules Dalianis, Martin Hassel, and Gunnar Nilsson, editors. 2010. *Proceedings of the NAACL HLT 2010 Second Louhi Workshop on Text and Data Mining of Health Documents*, Los Angeles, California, USA, June. Association for Computational Linguistics.

[Demner-fushman et al.2023] Dina Demner-fushman, Sophia Ananiadou, and Kevin Cohen, editors. 2023. *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, Toronto, Canada, July. Association for Computational Linguistics.

[Devlin et al.2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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, Minneapolis, Minnesota, June. Association for Computational Linguistics.

[Han et al.2022] Lifeng Han, Gleb Erofeev, Irina Sorokina, Serge Gladkoff, and Goran Nenadic. 2022. Examining large pre-trained language models for machine translation: What you don't know about it. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 908–919, Abu Dhabi, United Arab Emirates (Hybrid), December. Association for Computational Linguistics.

[Han et al.2023] Lifeng Han, Gleb , Irina Sorokina, Serge Gladkoff, and Goran Nenadic. 2023. Investigating massive multilingual pre-trained machine translation models for clinical domain via transfer learning. In *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pages 31–40, Toronto, Canada, July. Association for Computational Linguistics.

[Härkänen et al.2019] Marja Härkänen, Jussi Paananen, Trevor Murrells, Anne Marie Rafferty, and Bryony Dean Franklin. 2019. Identifying risks areas related to medication administrations-text mining analysis using free-text descriptions of incident reports. *BMC Health Services Research*, 19:1–9.

[Henry et al.2020] Sam Henry, Kevin Buchan, Michele Filannino, Amber Stubbs, and Ozlem Uzuner. 2020. 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.

[Hussain et al.2023] Arooj Hussain, Haifa Alrdahi, Hendrik Šuvalov, Lifeng Han, Goran Nenadic, Will Dixon, and Meghna Jani. 2023. M3: Extracting medication and related attributes from outpatient letters. In *HealTAC 2023: HEALTHCARE TEXT ANALYTICS CONFERENCE (Posters)*, Manchester, UK, June.

[Kormilitzin et al.2021] Andrey Kormilitzin, Nemanja Vaci, Qiang Liu, and Alejo Nevado-Holgado. 2021. Med7: A transferable clinical natural language processing model for electronic health records. *Artificial Intelligence in Medicine*, 118:102086.

[Kraljevic et al.2021] Zeljko Kraljevic, Thomas Searle, Anthony Shek, Lukasz Roguski, Kawsar Noor, Daniel Bean, Aurelie Mascio, Leilei Zhu, Amos A Folarin, Angus Roberts, Rebecca Bendayan, Mark P Richardson, Robert Stewart, Anoop D Shah, Wai Keong Wong, Zina Ibrahim, James T Teo, and Richard J B Dobson. 2021. Multi-domain clinical natural language processing with MedCAT: The medical concept annotation toolkit. *Artif. Intell. Med.*, 117:102083, July.

[Lavelli et al.2022] Alberto Lavelli, Eben Holderness, Antonio Jimeno Yepes, Anne-Lyse Minard, James Pustejovsky, and Fabio Rinaldi, editors. 2022. *Proceedings of the 13th International Workshop on Health Text Mining and Information Analysis (LOUHI)*, Abu Dhabi, United Arab Emirates (Hybrid), December. Association for Computational Linguistics.

[Manandhar and Yuret2013] Suresh Manandhar and Deniz Yuret, editors. 2013. *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, Atlanta, Georgia, USA, June. Association for Computational Linguistics.

[Pennington et al.2014] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October. Association for Computational Linguistics.

[Sarker et al.2016] Abeed Sarker, Karen O'connor, Rachel Ginn, Matthew Scotch, Karen Smith, Dan Malone, and Graciela Gonzalez. 2016. Social media mining for toxicovigilance: automatic monitoring of prescription medication abuse from twitter. *Drug safety*, 39:231–240.

[Sarker et al.2018] Abeed Sarker, Maksim Belousov, Jasper Friedrichs, Kai Hakala, Svetlana Kiritchenko, Farrokh Mehryary, Sifei Han, Tung Tran, Anthony Rios, Ramakanth Kavuluru, Berry de Bruijn, Filip Ginter, Debanjan Mahata, Saif M Mohammad, Goran Nenadic, and Graciela Gonzalez-Hernandez. 2018. Data and systems for medication-related text classification and concept normalization from Twitter: insights from the Social Media Mining for Health (SMM4H)-2017 shared task. *Journal of the American Medical Informatics Association*, 25(10):1274–1283, 10.

[Spasić et al.2010] Irena Spasić, Farzaneh Sarafraz, John A Keane, and Goran Nenadić. 2010. Medication information extraction with linguistic pattern matching and semantic rules. *Journal of the American Medical Informatics Association*, 17(5):532–535, 09.

[Wright et al.2015] Aileen P Wright, Adam T Wright, Allison B McCoy, and Dean F Sittig. 2015. The use of sequential pattern mining to predict next prescribed medications. *Journal of biomedical informatics*, 53:73–80.

[Wu et al.2022] Yuping Wu, Lifeng Han, Valerio Antonini, and Goran Nenadic. 2022. On cross-domain pre-trained language models for clinical text mining: How do they perform on data-constrained fine-tuning?

Appendix

Because Med7 itself is fine-tuned for n2c2-2018 shared task, we can deploy it directly to test the performances. Here we also list its output scores when we evaluate it using the official n2c2-2018 training data, i.e. the 303 labelled letters. Pay attention that ADE and Reason labels are missing from the model via its direct evaluation.

Catergory	Pre.	Rec.	F1	Num.
Dosage	0.94	0.93	0.93	4221
Drug	0.93	0.89	0.91	16225
Duration	0.82	0.85	0.83	592
Form	0.94	0.92	0.93	6651
Frequency	0.88	0.82	0.85	6281
Route	0.96	0.96	0.96	5476
Strength	0.93	0.94	0.94	6691
Acc.			0.85	49003
Macro avg	0.80	0.79	0.79	49003
Weighted avg	0.87	0.85	0.86	49003

Table 4: Outputs from Med7 Deployment on 303 Letters - Evaluation: Type

The 'Frequency' catergory score has a big drop on Recall with absolute value 0.05 via 'Strict' in comparison to 'Type', otherwise, the recall scores mostly have 0.01-0.02 drops and Precision scores remain the same for other labels.

Catergory	Pre.	Rec.	F1	Num.
Dosage	0.94	0.91	0.92	4221
Drug	0.93	0.88	0.90	16225
Duration	0.81	0.82	0.82	592
Form	0.94	0.91	0.92	6651
Frequency	0.88	0.77	0.82	6281
Route	0.96	0.96	0.96	5476
Strength	0.93	0.94	0.94	6691
Acc.			0.83	49003
Macro avg	0.80	0.77	0.78	49003
Weighted avg	0.87	0.83	0.85	49003

Table 5: Outputs from Med7 Deployment on 303 Letters - Evaluation: Strict