Evaluating Data Augmentation for Medication Identification in Clinical Notes

Jordan Koontz
Ixa
UPV/EHU
Donostia, Basque Country, Spain
jkoontz001@ikasle.ehu.eus

Maite Oronoz and **Alicia Pérez**
HiTZ - Ixa
UPV/EHU
Donostia, Basque Country, Spain
maite.oronoz@ehu.eus
alicia.perez@ehu.eus

Abstract

We evaluate the effectiveness of using data augmentation to improve the generalizability of a Named Entity Recognition model for the task of medication identification in clinical notes. We compare disparate data augmentation methods, namely mention-replacement and a generative model, for creating synthetic training examples. Through experiments on the n2c2 2022 Track 1 Contextualized Medication Event Extraction data set, we show that data augmentation with supplemental examples created with GPT-3 can boost the performance of a transformer-based model for small training sets.

1 Introduction

Natural Language Processing (NLP) is an active area of research in healthcare, especially due to the proliferation of Electronic Health Records (EHR). EHRs contain extensive information about individual patients, such as diagnoses with their corresponding International Classification of Disease (ICD) codes, treatment records and test results. While some medication information can be extracted from the structured data in the EHRs, a substantial amount of the medication information resides in text-based narrative clinical notes (Sohn et al., 2014). The information contained in clinical notes can be useful for pharmacovigilance, comparative effectiveness studies, and adverse event detection (Uzuner et al., 2010). The objective of the n2c2 2022 Track 1 Contextualized Medication Event Extraction was to capture multi-dimensional context of medication changes documented in clinical notes. The track was comprised of three subtasks:

- Task 1: [NER] Medication Extraction
- Task 2: [Event] Event Classification
- Task 3: [Context] Context Classification

A prerequisite for understanding medication changes in clinical documents is to successfully identify all mentions of medication in the documents. However, in the clinical domain, a common challenge for training machine learning models is a lack of annotated training data. Annotating clinical notes can be an expensive and lengthy process that requires medical domain experts. In this paper, we set out to evaluate disparate data augmentation techniques to create supplemental training examples with the hope of reducing a dependence on manual annotations while also boosting the performance of a medication identification model.

First, we detail our model architecture comprised of a transformer-based language model and a Conditional Random Fields (CRF) (Lafferty et al., 2001) component for identifying mentions of medication in clinical documents that obtained competitive results on the n2c2 2022 [NER] Medication Extraction subtask. Next, we detail our data augmentation methodology for creating synthetic training examples. Finally, we evaluate the effectiveness of using data augmentation for the task of medication extraction in clinical documents. Moreover, we evaluate the effectiveness of using data augmentation for low-resource medication extraction, i.e. a scenario in which the size of a training set is small.

2 Background

Early systems for medication identification relied chiefly on rule-based techniques. Evans et al. (1996) combine Natural Language Processing (NLP) pre-processing techniques and regular expressions to extract drug-dosage information from clinical narratives. The authors achieve an approximate 80% rate of exact and partial matches on target phrases.

Proceedings of Recent Advances in Natural Language Processing, pages 578–585
Varna, Sep 4–6, 2023
https://doi.org/10.26615/978-954-452-092-2_063

Later, machine learning demonstrated effectiveness for the task of medication identification. Patrick and Li (2010) used a CRF model to identify medications for the 2009 i2b2 medication extraction task. The model used six feature sets, many of them requiring external knowledge (e.g. gazetteers) and hand-crafted features (e.g. morphological patterns).

Currently, neural network architectures, namely transformers, demonstrate state-of-the-art results for medication identification. Alsentzer et al. (2019) fine-tune their domain-specific Bio+Clinical BERT model on the i2b2 2010 concept extraction task (Uzuner et al., 2011), achieving an F1 score of 0.872 for exact matching, outperforming non-domain-specific variants such as BERT (Devlin et al., 2019).

Hakala and Pyysalo (2019) combine BERT with a final CRF layer for PharmaCoNER (Gonzalez-Agirre et al., 2019), the first shared task on detecting drug and chemical entities in Spanish medical documents.

Hiba et al. (2023) present an evaluation of fine-tuning pre-trained language models for the task of biomedical entity recognition, namely drug names and symptoms. The authors compare five language models on two biomedical data sets, CADEC and ADE-corpus. Their evaluation results demonstrate that BioBERT (Lee et al., 2020), a language model pretrained on in-domain (biomedical) corpora, outperformed all other models on both data sets and obtrained F1-scores of 0.903 and 0.6873 in the ADE and CADEC corpora, respectively.

For the 2022 n2c2 Medication Extraction subtask, we sought to leverage both an in-domain transformer-based language model, namely Bio+Clinical BERT and a CRF.

3 Material and methods

3.1 Corpus Description

Track 1 of n2c2 2022 used the Contextualized Medication Event data set (CMED) (Mahajan et al., 2022). The corpus is comprised of 500 clinical notes from the i2b2 2014 Heart Disease Risk Factor Challenge data set (Stubbs et al., 2015). The Track 1 data set consists of 9,012 annotated medication mentions over the 500 clinical notes. Moreover, the data set is divided into train (400 notes) and test (100 notes) partitions. In order to train our NER model, we convert the train and test partitions from brat standoff format (Stenetorp et al., 2012) to Inside–outside–beginning (IOB) format (Ramshaw and Marcus, 1995). Table 1 shows a training example from the training partition together with the entities annotated as in IOB format.

Token	Label
METOPROLOL	B-Medication
TARTRATE	O
25	O
MG	O
BID	O

Table 1: Example from the training corpus and its corresponding IOB annotation.

3.2 Model

For our NER model, we used an architecture based on a transformer language model and a CRF. Concretely, we fine-tuned the Bio+Clinical BERT language model. Bio+Clinical BERT was selected due the similarity between its pretraining texts (all note types in MIMIC III v1.4) and the n2c2 corpus. We posited that a language model pretrained on in-domain texts (clinical notes) would be better suited for the task of medication identification than other language models such as BERT. The Bio+Clinical BERT model is followed by a token-level classifier. The tag scores are then fed to a Linear-Chain CRF to maximize the likelihood of selecting the best output label sequence. Table 2 describes the configuration and training of our final model whose parameters were obtained through a grid search.

Encoder model	Bio+Clinical BERT
Dropout	0.25
Maximum sequence length	512
Batch size	8
Epochs	4
Learning rate	0.00001

Table 2: Configuration for our medication identification model.

3.3 Data augmentation

Hoping to produce a model that would generalize well on the challenge's test set, we developed two data augmentation strategies to create synthetic training instances using the following techniques: mention-replacement and a generative model. For the latter, we use few-shot learning with Generative Pre-trained Transformer-3, also referred to as GPT-3 (Brown et al., 2020).

3.3.1 Mention-replacement

Inspired by Dai and Adel (2020), we use a mention-replacement method in which we substitute medication mentions from the original training corpus with medication mentions gleaned from external sources to create novel synthetic instances. To collect additional medication mentions, we had two strategies (depicted in Figure 1):

1. We apply our baseline NER model (trained on the challenge's training set) to a subset of discharge summaries from MIMIC-III (Johnson et al., 2016) to collect medications not present in the original corpus.

2. We collect medication mentions (already annotated) in Spanish from the Chilean Waiting List Corpus (CWLC) (Báez et al., 2020).

The first strategy allows us to create synthetic instances without needing manual annotations from domain experts. We applied our baseline NER model to 596 discharge summaries from MIMIC-III and we obtain 9,149 new medication mentions that do not appear in the original corpus. An example of a synthetic training instance created using this augmentation strategy is shown in Table 3. In an effort to produce a fully automated data augmentation strategy, human intervention was not involved (e.g. entity cleaning and validation) at the cost of permitting errors to be introduced into the training data set.

Original	
Renaphro	B-Medication
1	O
TAB	O
PO	O
QD	O
Augmented	
pipatz	B-Medication
1	O
TAB	O
PO	O
QD	O

Table 3: Example of data augmentation. The top instance is from the original n2c2 2022 training corpus. The bottom synthetic instance was created by substituting the original medication mention with a new medication identified by our baseline model from MIMIC-III discharge notes.

The second strategy, despite using a corpus already annotated by domain experts (three medical students and one medical doctor), allowed us to evaluate the effectiveness of using code-switched (Spanish and English) training instances. The CWLC is comprised of referrals for several specialty consultations from the waiting list in Chilean public hospitals. We collect 92 medication mentions from 891 sentences.

3.3.2 Few-shot learning with GPT-3 text-davinci-003

GPT-3 has gained attention due to its ability to generate coherent and human-like texts for a given prompt. We sought to evaluate the effectiveness of this 175-billion parameter model (namely text-davinci-003) for generating supplemental training instances. To do so, we provide a few examples of the task at inference time to condition the model as depicted in Table 4. Concretely, the prompt is composed of 3 medications followed by 3 example sentences, and then a final medication to generate a sentence for. The final medication is randomly selected from the 9,149 medication mentions extracted from MIMIC-III clinical notes by our baseline NER model. Using this strategy, we generate 200 sentences and then convert them to IOB format to be used in the model's training.

3.4 Experimental low-resource medication identification

Annotating clinical notes is a lengthy and expensive process that requires medical domain experts. In an experimental setup, we evaluate the effectiveness of data augmentation for low-resource medication identification, i.e. a scenario in which little annotated data is available for training a medication identification model. We simulate a low-resource setting by splitting the n2c2 2022 training set into two partitions. Partition 1 (denoted as Small data set or SM), is comprised of 10% of the sentences from the training set. Partition 2 (denoted as Medium data set or MD) is comprised of 25% of the sentences from the training set. Each partition is then combined with the aforementioned synthetic instances from MIMIC-III, CWLC, and GPT-3.

4 Results

F1-scores, calculated at micro and macro averaged levels, were used in the evaluation using the n2c2

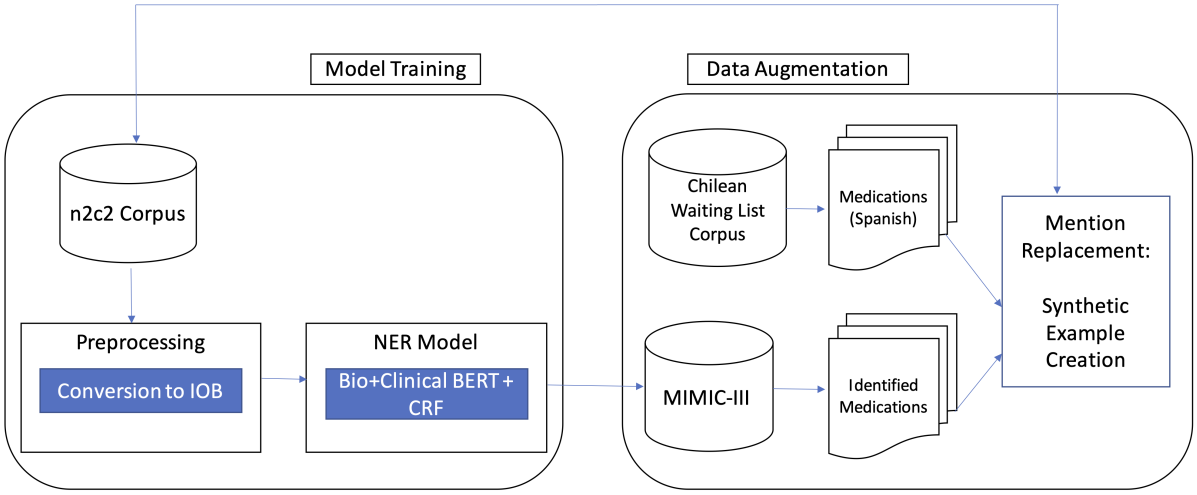

Figure 1: Medication identification system and data augmentation (mention-replacement) architecture diagram.

Prompt:
Lipitor → Patient is being treated with Lipitor long acting nitrate → We will continue her on long acting nitrate Advil → She has been taking Advil 200 mg 2 and up to 6 per day ziac →
GPT-3 response:
We have prescribed Ziac for her blood pressure control

Table 4: Example of data augmentation. The top is the prompt composed of three medications, three example sentences, and a final medication to generate a text for. The bottom synthetic instance was generated by GPT-3.

2022 Track1 test data set. The medication extraction subtask employed two kinds of evaluation: strict and lenient matching. For strict matching, the offsets of a span were required to match exactly. For lenient matching, it was sufficient for spans to overlap.

Results for our submission to the n2c2 2022 challenge are presented in Table 5 denoted as Approach I. Our top-performing model on the test data set, 90% in terms of F1 lenient matching, was our baseline (no augmentation). The use of data augmentation with GPT-3 did not form part of our submission to n2c2.

Later, we achieved significant improvements by tuning hyper-parameters and by modifying our postprocessing of the data (e.g. conversion from IOB to Brat standoff format). Improved results post-n2c2 are also included in Table 5 denoted as Approach II.

Once again, our top-performing model, in terms of F1 lenient matching, was our baseline model (without augmentation), with a result of 96% for lenient matching. The model trained with synthetic examples from the CWLC remained the least effective model and it achieved only a modest 1% increase in F1 lenient matching score (90.11%) on the test set with the optimized hyper-parameters.

Moreover, we observed a significant difference between our F1 strict and lenient scores. For all models, we achieved higher F1 lenient scores than strict matching scores. The smallest margin between scores on the test set was for our baseline, with a difference of 4.12%. The differences between the F1 lenient and strict scores were 4.78% and 4.85% for MIMIC-III and CWLC variants respectively.

We also found that the use of data augmentation with GPT-3 did not boost performance on the test set. On the other hand, using examples created by GPT-3 boosted performance in a low-resource setting, demonstrated in Table 6. On the SM partition (10% of the sentences from the n2c2 training set), data augmentation with GPT-3 results in F1-scores of 75.83% and 86.34% for strict and lenient matching respectively. The exclusion of augmentation resulted in F1-scores of 73.96% and 83.90%. The performance boost from data augmentation was less notable on the MD partition (25% of the sen-

tences from the n2c2 training set). Augmentation with GPT-3 resulted in F1-scores of 76.14% and 86.94% while the model trained without augmentation obtained F1-scores of 75.11% and 85.13%. The use of mention-replacement augmentation did not boost performance in the low-resource setting (with the exception of CWLC on the MD partition for F1-strict).

	F1-Strict	F1-Lenient
Approach I		
No augmentation	**87.23**	**90.34**
MIMIC	86.78	89.55
CWLC	86.96	89.11
Approach II:		
No augmentation	**92.22**	**96.34**
MIMIC	90.16	94.94
CWLC	85.37	90.11
GPT-3	84.81	92.37

Table 5: Top: Scores for submissions to the n2c2 2022 Track 1 NER substask measured in terms of F1 strict and lenient matching (test set). The models are: Baseline (no augmentation), MIMIC (data augmentation from MIMIC), and CWLC (data augmentation from the Chilean Waiting List Corpus), and GPT-3 (data augmentation from GPT-3 and MIMIC-III medications). Bottom: Scores for our models improved post-n2c2 2022.

	F1-Strict	F1-Lenient
SM:		
No augmentation	73.96	83.90
MIMIC	69.18	77.44
CWLC	72.36	81.35
GPT-3	**75.83**	**86.34**
MD:		
No augmentation	75.11	85.13
MIMIC	70.97	80.73
CWLC	75.54	85.02
GPT-3	**76.14**	**86.94**

Table 6: Top: Scores measured in terms of F1 strict and lenient matching on the n2c2 test set for the low-resource partition SM. The models are: No augmentation, MIMIC (data augmentation from MIMIC), and CWLC (data augmentation from the Chilean Waiting List Corpus), and GPT-3 (data augmentation from GPT-3 and MIMIC-III medications). Bottom: Scores measured in terms of F1 strict and lenient matching on the n2c2 test set for the low-resource partition MD.

5 Discussion

Fine-tuning the Bio+Clinical BERT language model in conjunction with a CRF, without data augmentation, produces an effective medication identification model, corroborated by our competitive F1 lenient matching score (96%) using Approach II on the n2c2 Track 1 NER substask test set. However, our top-performing model still exhibits some weaknesses, such as its handling of abbreviations. For example, for the target medication *Niacin SR* in the test data set, our model identifies *Niacin* while excluding *SR* (sustained release). Given the input sentence "phoslo 1 tab po tidac" from the test data set, our model identifies *tidac* as a medication mention. Notwithstanding that a Tidac Tablet is a medication used to treat and prevent stomach ulcers, in this context, *tidac* translates to *t.i.d.a.c*, i.e. "three times a day before meals". In addition to abbreviations, we also observed occurrences in which our model struggled to handle multi-word medication mentions. For instance, given the target medication *Multivitamin With Betacarotene*, our model instead identified two unique medications *Multivitamin* and *Betacarotene*.

We also found that the use of data augmentation, when using the full training set, did not improve the performance our model. We achieved F1 lenient matching scores of 94%, 90%, and 92% for our MIMIC, CWLC, and GPT-3 model variants respectively. There are several variables that may have stymied the effectiveness of our data augmentation strategy.

For example, Dai and Adel (2020) demonstrate that a mention-replacement data augmentation method is most effective on the i2b2 2010 concept extraction task when training on a small training corpus comprised of 50 instances. Provided that the CMED data set is comprised of 9,012 annotated medication mentions across 500 clinical notes (400 for training), the baseline training corpus is perhaps ample for training effective medication identification models.

Moreover, our augmentation method may have introduced a significant amount of noise that was ultimately harmful. Applying our baseline model to unannotated discharge summaries resulted in the collection of incorrect and problematic medication mentions. For example, our baseline model recognized *kaopectate / benadryl / lidocaine* as a single medication instead of three unique medications. Our baseline model also identified abstract

concepts in the discharge summaries, such as *narcotic pain medications*, as medication mentions. Terms such as *safetyglide*, *cranberry*, *suction*, and *banana* were incorrectly identified as medications. The quality (e.g. the presence of special characters or medications concatenated with dosage information) of many identified medications in the discharge summaries were also problematic, e.g. *caltrate 600]* - and *simvistatin80mg*.

The effectiveness of our models trained with data augmentation may have also been affected by the randomness of the mention-replacement method. Concretely, the augmentation method makes contextually inappropriate replacements of medication mentions, highlighted in Table 7.

The use of code-switched resources also failed to improve the generalization ability of our baseline model. Notwithstanding that only 92 medication mentions were collected from the CWLC, and hence fewer synthetic examples created than from MIMIC-III, our model trained on the code-switched training corpus resulted in significantly worse results than our baseline model.

On the other hand, we find that data augmentation using instances generated by GPT-3 can improve F1-scores in a low-resource setting. Concretely, there are two characteristics of GPT-3 that may have contributed to its effectiveness: first, its ability to generate novel human-like sentences, and second, its ability to generate contextually correct sentences (unlike our mention-replacement method). For example, for the input medication *phenylephrine*, GPT-3 generated the sentence "*We can add phenylephrine to help to reduce the congestion*". Phenylephrine is a medication used to relieve nasal discomfort caused by colds, allergies, and hay fever, and therefore GPT-3 is able to create a novel training example with the medication mention used in the proper context.

6 Conclusions

We have described an architecture based on a transformer-based language model (Bio+Clinical BERT) and a CRF for the task of medication identification in clinical notes. Additionally, we have presented a data augmentation strategy for creating synthetic training instances.

Models trained with our proposed data augmentation strategy yielded mixed results on the n2c2 2022 medication identification sub-task. Our model using synthetic examples from MIMIC-III achieved an F1 lenient score of 94% (which places it above the mean score shared by the task organizers), albeit lower than the score obtained by our baseline model. Our model trained with synthetic examples containing medication mentions in Spanish from the CWLC failed to produce competitive results. This model obtained an F1 lenient score of 90% on the test data set, placing it below the mean score shared by the task organizers. On the other hand, our baseline model (without augmentation) achieved competitive results in terms of our F1 lenient matching score (96%) on the n2c2 2022 Track 1 test set. Provided that our chief motivation was to produce an automated data augmentation system (reducing a dependency on costly domain experts), our mention-replacement technique did not contain constraints to ensure the semantic correctness of the substitutes. As a result, errors and biases were likely reinforced during the training of the models with mention-replacement augmentation. Future work should also explore techniques to add restrictions that ensure the semantic correctness of synthetic instances. For example, using publicly available lists of medication names could help to ensure the correctness of the synthetic instances. The use of such lists could also permit an introduction of new medication names in the data for continuous training of models. Even though hyper-parameters were tuned, there are also some architectural changes that may be adjusted in future work. For example, freezing the weights of Bio+Clinical BERT, and hence only training the token classifier and CRF, may be evaluated. Moreover, removing the CRF should also be assessed.

In a low-resource setting, we demonstrate that data augmentation can boost the performance of a medication recognition model. Concretely, we demonstrate that zero-shot learning with GPT-3 is an effective technique for creating novel and contextually correct training examples in the clinical domain for medication identification. This technique could be particularly beneficial in situations where the use of annotators with clinical domain expertise is not feasible. Additionally, a strength of GPT-3 is its ability to generate coherent text in multiple languages, such as Spanish, German, Japanese, and Russian. The generation of synthetic training instances with GPT-3 for medication identification in multiple languages should evaluated in future work. On the other hand, one known weakness of generative models such GPT-3 is their

Original:	Habitrol	patch	and	has	not	smoked	since
	B-Medication:	O	O	O	O	O	O
Augmented:	**dipirona**	patch	and	has	not	smoked	since
	B-Medication:	O	O	O	O	O	O

Table 7: Example of data augmentation with a contextually inappropriate medication mention-replacement. The top instance is from the original n2c2 2022 training corpus. The bottom synthetic instance was created by substituting the original medication mention with a new medication (in Spanish) from the CWLC. Dipirona is painkiller that is commonly given by mouth or by intravenous infusion, but not by patch. Moreover, unlike Habitrol, dipirona is not related to nicotine or smoking.

tendency to hallucinate, i.e. produce factually incorrect text. The ability generate correct medication names from large language models, such as GPT-3, should also be evaluated. The ability of a language model to produce a list of medications related to a given medical problem could reduce dependencies on annotated corpora and external data sources. GPT-3 has other disadvantages, e.g. a pay-per-use system and the collection of user data. Therefore, an evaluation of open-source large language models for the creation of synthetic training instances should be conducted in future work.

Acknowledgements

This work was elaborated within the framework of LOTU (TED2021-130398B-C22) funded by MCIN/AEI/10.13039/501100011033, European Comission (FEDER), and by the European Union "NextGenerationEU"/PRTR. Besides, this work was partially funded by the Spanish Ministry of Science and Innovation (DOTT-HEALTH/PAT-MED PID2019-106942RB-C31 and Antidote PCI2020-120717-2); by the Basque Government (IXA IT-1570-22); and by EXTEPA within Misiones Euskampus 2.0.

References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Pablo Báez, Fabián Villena, Matías Rojas, Manuel Durán, and Jocelyn Dunstan. 2020. The Chilean Waiting List Corpus: a new resource for clinical Named Entity Recognition in Spanish. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 291–300, Online. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners.

Xiang Dai and Heike Adel. 2020. An Analysis of Simple Data Augmentation for Named Entity Recognition. pages 3861–3867.

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. Association for Computational Linguistics.

D A Evans, Nicolas D Brownlow, William R. Hersh, and Emily M. Campbell. 1996. Automating concept identification in the electronic medical record: an experiment in extracting dosage information. *Proceedings : a conference of the American Medical Informatics Association. AMIA Fall Symposium*, pages 388–92.

Aitor Gonzalez-Agirre, Montserrat Marimon, Ander Intxaurrondo, Obdulia Rabal, Marta Villegas, and Martin Krallinger. 2019. PharmaCoNER: Pharmacological Substances, Compounds and proteins Named Entity Recognition track. In *Proceedings of the 5th Workshop on BioNLP Open Shared Tasks*, pages 1–10, Hong Kong, China. Association for Computational Linguistics.

Kai Hakala and Sampo Pyysalo. 2019. Biomedical Named Entity Recognition with Multilingual BERT. In *Proceedings of the 5th Workshop on BioNLP Open Shared Tasks*, pages 56–61, Hong Kong, China. Association for Computational Linguistics.

Chanaa Hiba, El Habib Nfaoui, and Chakir Loqman. 2023. Fine-tuning transformer models for adverse

drug event identification and extraction in biomedical corpora: A comparative study. In *Digital Technologies and Applications*, pages 957–966, Cham. Springer Nature Switzerland.

Alistair Johnson, Tom Pollard, and R Mark III. 2016. MIMIC-III clinical database. *Physio Net*, 10:C2XW26.

John Lafferty, Andrew Mccallum, and Fernando Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. pages 282–289.

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

Diwakar Mahajan, Jennifer Liang, and Ching-Huei Tsou. 2022. Toward Understanding Clinical Context of Medication Change Events in Clinical Narratives. *Proc. American Medical Informatics Association. AMIA Annual Symposium*, 2021:833–842.

Jon Patrick and Min Li. 2010. High accuracy information extraction of medication information from clinical notes: 2009 i2b2 medication extraction challenge. *Journal of the American Medical Informatics Association : JAMIA*, 17:524–7.

Lance Ramshaw and Mitch Marcus. 1995. Text Chunking using Transformation-Based Learning. In *Third Workshop on Very Large Corpora*.

Sunghwan Sohn, Cheryl Clark, Scott Halgrim, Sean Murphy, Christopher Chute, and Hongfang Liu. 2014. Medxn: an open source medication extraction and normalization tool for clinical text. *Journal of the American Medical Informatics Association : JAMIA*, 21.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. BRAT: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations Session at EACL 2012*, Avignon, France. Association for Computational Linguistics.

Amber Stubbs, Christopher Kotfila, Wang Qi, and Ozlem Uzuner. 2015. Identifying risk factors for heart disease over time: Overview of 2014 i2b2/UTHealth shared task Track 2. *Journal of Biomedical Informatics*, 58.

Ozlem Uzuner, Imre Solti, and Eithon Cadag. 2010. Extracting medication information from clinical text. *J Am Med Inform Assoc*, 17:514–518.

Ozlem Uzuner, Brett South, Shuying Shen, and Scott DuVall. 2011. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association : JAMIA*, 18:552–6.