Context-aware Medication Event Extraction from Unstructured Text

Noushin Salek Faramarzi[*] **Meet Patel**[*] **Harika Bandarupally**[*] **Ritwik Banerjee**[*][§]
[*] Department of Computer Science
[§] Institute for AI-Driven Discovery and Innovation
Stony Brook University, New York, USA
{nsalekfarama,meppatel,hbandarupall,rbanerjee}@cs.stonybrook.edu

Abstract

Accurately capturing medication history is crucial in delivering high-quality medical care. The extraction of medication events from unstructured clinical notes, however, is challenging because the information is presented in complex narratives. We address this challenge by leveraging the newly released Contextualized Medication Event Dataset (CMED) as part of our participation in the 2022 National NLP Clinical Challenges (n2c2) shared task. Our study evaluates the performance of various pretrained language models in this task. Further, we find that data augmentation coupled with domain-specific training provides notable improvements. With experiments, we also underscore the importance of careful data preprocessing in medical event detection.

1 Introduction

Ensuring the accuracy of a patient's treatment history is essential for delivering high-quality medical care. This allows healthcare professionals to assess the effectiveness of existing treatments, detect possible medication-related problems, and suggest appropriate future treatment options (FitzGerald, 2009). Various forms of treatment changes, however, are often absent from structured electronic data sources, being recorded only in clinical narratives (Turchin et al., 2009). An accurate extraction of medication event information from unstructured data in patients' medical records is thus crucial for a complete understanding of their treatments.

When extracting medication changes from clinical text, it is necessary to take into account various forms of contextual information, due to the narrative and longitudinal nature of clinical documentation. Clinical text often documents events over a patient's medical history, and providers may also record the reasoning behind their medical decisions. These factors result in complex events that cannot be properly captured by extracting medication changes alone, without considering the surrounding clinical context. This is especially true when developing a medication change extraction system to support real-world applications, such as medication timeline generation (Plaisant et al., 2003; Belden et al., 2019) or medication reconciliation (Poon et al., 2006; Cadwallader et al., 2013). Indeed, as Wang et al. (2018) have argued, the use of sophisticated natural language processing (NLP) information extraction (IE) become a necessity when the automatic extraction of relevant medical information is required by large-scale or real-time applications further downstream, such as clinical research and decision support.

This study investigates how to extract information about changes to patient medications from clinical text using the Contextualized Medication Event Dataset (CMED) developed by Mahajan et al. (2021) and subsequently released to the community as a shared task in 2022 National NLP Clinical Challenges (n2c2)[1]. This consists of three tasks: (i) medication extraction, to extract all medication mentions in clinical notes, (ii) event classification, to identify whether a medication change is discussed in an event,and (iii) context classification, to classify the contextual information of a medication change event along five orthogonal dimensions, with each dimension further classified into multiple *attributes* of the event.

For the first medical named entity recognition task, we note that Lee et al. (2020) demonstrate significant improvements with the use of BioBERT, a domain-specific model initialized with BERT and then pretrained on PubMed abstracts and PubMed Central full text articles. Thus, we proceed to use BioBERT as well, providing comparisons against popular general purpose language models like BERT (Devlin et al., 2019). Additionally, we also utilize Bio+Clinial BERT (Alsentzer et al., 2019), another domain-specific model initialized

[1] n2c2.dbmi.hms.harvard.edu/2022-track-1

Proceedings of the 5th Clinical Natural Language Processing Workshop, pages 86–95
July 14, 2023 ©2023 Association for Computational Linguistics

with BioBERT and further pretrained on notes from the MIMIC-III dataset (Johnson et al., 2016).

We evaluate the performance of several pretrained language models for the second and third tasks. Specifically, we examine three popular general purpose models – BERT, RoBERTa (Liu et al., 2019b), and XLNet (Yang et al., 2019) – and one domain-specific pretrained model, Bio+Clinical BERT.

While many downstream natural language understanding tasks are readily successful when a large pretrained language model is tuned for the task in hand, we observe that the accuracy of clinical event detection crucially depends on careful data preprocessing. In particular, identifying the proper linguistic context from surrounding text is of utmost importance. To that end, we develop and employ a sentence detection method tailored to this task, leading to a better performance by all models. We also find that augmenting the data with the DDI (drug-drug interaction) Corpus (Herrero-Zazo et al., 2013) leads to overall improvements in medication change detection and its context classification.

2 Related Work

The first task in contextualized medication event extraction is to extract the mention of medications – clearly, a medical named entity recognition (NER) task. Medical NER, in general, includes identifying other types of entities such as diseases, symptoms, proteins, or patient information (see Pagad and Pradeep (2022) for an overview). To identify medication names in particular, many approaches have been proposed. Early methods relied explicitly on domain ontology or dictionaries (Sanchez-Cisneros et al., 2013), rules (Segura-Bedmar et al., 2008), and subsequently, contextual rules and automatically learned rules (Hamon and Grabar, 2010; Coden et al., 2012). A comprehensive survey of this literature has been conducted by Liu et al. (2015). More recent approaches are hybrid, combining LSTM and its variants with conditional random fields (CRF) or other graphical models (Alfattni et al., 2021; Jouffroy et al., 2021). Even more recent, however, are techniques that utilize Transformer models (*e.g.*, BERT). There is some work to further indicate that combining BERT with BiLSTM-CRF improves medical NER (Yu et al., 2019), while others demonstrate the improvements in using domain-specific pretraining with BERT initialization (Lee et al., 2020).

Identifying medication change events and classifying their attributes, however, is a significantly less explored problem. This is due largely to the scarcity of annotated resources, but to a lesser extent, also to the complexity of the language used in clinical narratives to describe such events. Initially, research heavily relied on annotated datasets like the 2009 i2b2 and the 2013 DDI datasets (Uzuner et al., 2011; Herrero-Zazo et al., 2013). Some early work focused on very specific events of clinical relevance, such as Liu et al. (2019a), who inspect medication discontinuation, or Sohn et al. (2010), who focus on whether medication was started, stopped, increased, or decreased. In another approach, Pakhomov et al. (2002) introduced temporal information into their labels. In spite of the success on individual datasets, these approaches employ rule-based decisions and classical supervised learning algorithms like support vector machines (SVMs) or maximum entropy modeling, which are unlikely to generalize across multiple datasets with linguistic variation without extensive supervision for each dataset.

For a detailed understanding of treatments, such as extracting the dosage, frequency, or mode of drug administration, or in determining its relation to other phenomena like adverse drug effects, generalizable success in this task carries immense significance. It is thus worth noting that recent methods leveraging neural architectures and models have shown promise in medical event extraction and classification tasks (Narayanan et al., 2022). Lerner et al. (2020) use a neural top-down transition based parser and achieve results comparable to BiLSTM models for medical entity and event detection. Perhaps the closest to our study is the approach of Lybarger et al. (2021), who tune BERT on COVID-19 data to identify various events of clinical significance, such as symptoms, severity, and assertion. This body of work is distinct from ours, however, since it does not delve into classification of event attributes involving complex temporal or conditional expressions.

Several studies (Uzuner et al., 2011; Chapman et al., 2001; Szarvas et al., 2008; Morante, 2010; Albright et al., 2013) have examined the detection of negated medical concepts in clinical text. However, none of them specifically focus on identifying medication change events. Moreover, they have not looked at the combined identification of negation and the actor responsible for that negation. Early

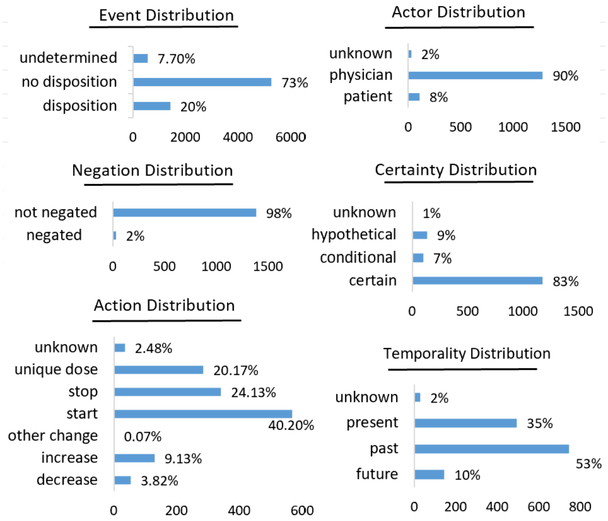

Figure 1: Distribution of labels in CMED (training set).

work on negation detection in clinical texts was based on negation lexicons, and rule-based algorithms using them. Most notable among them is NegEx (Chapman et al., 2001). Although these have been superceded by others who combine lexicons with dependency structures or other linguistic features (Mehrabi et al., 2015), we use an implementation of NegEx built into the popular spaCy[2] library, called negspaCy. Our results (Sec. 6) show that in spite of its simplicity, this approach suffices.

With the *Contextual Medication Event Dataset*, CMED, Mahajan et al. (2021) provide annotated data capturing five orthogonal dimensions of contextual information related to medication change events. Further, they demonstrate the viability of SVMs and Transformer-based models in detecting and classifying these events. Very recently, Ramachandran et al. (2023) have explored an avenue similar to ours, with domain-specific language models based on BERT. In these, it has been noted that sentences that mention multiple drugs are particularly difficult to detect and classify. By contrast, our work investigates data augmentation and task-specific preprocessing in conjunction with the user of domain-specific language models. In particular, we develop and use a custom sentence extraction module in our pipeline, which improves the accuracy of these models on the given tasks.

3 The CMED Dataset

The CMED dataset comprises annotated clinical notes, where each medication mention is assigned one or more event labels from the three categories:

(1) `Disposition`, indicating the mention of a medication change, *e.g.*, "prescribed albuterol for shortness of breath",

(2) `NoDisposition`, indicating that the mediction is mentioned with no indication of change, *e.g.*, "patient continues to take aspirin", and

(3) `Undetermined`, indicating a lack of clarity or evidence regarding medication change, *e.g.*, "Plan: Lasix".

For each event identified as `Disposition`, the clinical context is provided along five orthogonal dimensions, *viz.*, `action`, `actor`, `negation`, `certainty`, and `temporality`. We describe these next.

(1) `action` refers to the type of change is being made. Its attributes are *start*, *stop*, *increase*, *decrease*, *unique dose*, *other change*, and *unknown*.

(2) `actor` specifies who initiated the action, *physician*, *patient*, or *unknown*.

(3) `negation` indicates whether the action is negated or not.

(4) `temporality` specifies whether the action takes place in the *past*, *present*, or *future*.

(5) `certainty` characterizes the likelihood of the action taking place as *certain*, *hypothetical*, *conditional*, or *unknown*.

The distribution of the event and attribute labels in the training set of this dataset is shown in Fig. 1.

4 Data Preprocessing

We observe that in CMED, most medication mentions are labeled with one event and a set of corresponding attributes. There is, however, a small fraction (< 90 instances in the training set), where the drug mention is labeled with two events and two separate sets of attributes, as noted by Ramachandran et al. (2023) as well. Further, we underscore the frequent presence of sentences containing multiple drug mentions (approx. 78%), with a substantial fraction (over 50%) of such sentences mentioning four or more drugs simultaneously. This aspect of the dataset significantly increases the complexity of detecting and classifying contextual information from clinical sentences. Finally, we note that some events cannot be accurately labeled based solely on the sentence in which they appear, and additional context from neighboring sentences becomes necessary to determine the correct attributes. Table 1 presents examples from CMED showcasing illustrative examples of these phenomena.

[2] spacy.io/

(a) “The patient’s daily dose of **furosemide** was increased from 40mg to 80mg.
and then reduced to 60mg daily.”
LABELS: *increase*, *decrease*

(b) “The healthcare provider started the patient on a new regimen of **metformin** and discontinued the use of **pioglitazone**.”
LABELS: *start*, *stop*

(c) “The healthcare provider instructed the patient to take **acetaminophen**
if their fever rises above 100 degrees.”
LABELS: *conditional*

Table 1: Examples from clinical notes where (a) one drug mention indicates two events with opposite action labels, and (b) two drug mentions, each with their own action labels. Also, (a) and (c) have grammatically valid sentences up until the line break, but the sentence continues. Stopping at the line break will miss the language responsible for the *decrease* and *conditional* labels.

The first step in the medication information extraction task is to prepare the dataset by extracting the sentences containing medication information. However, due to the unstructured and lengthy nature of medical notes, accurately identifying the start and end of a sentence containing a medication mention is challenging. Accordingly, relying solely on tools like, say, spaCy, for their inbuilt sentence parser for this task does not produce satisfactory results. Therefore, we develop a customized approach to accurately identify the sentences that contained medication names, which served as a crucial first step towards performing accurate medication event extraction. Next, we describe the steps of this process.

(i) Abbreviation resolution. Abbreviations such as “Continue” and “Discontinue” are converted to their full forms to facilitate accurate identification of medication mentions in the text. One of the most frequent and important abbreviations is “Discontinue,” which is observed in different forms with various spacing (e.g., “d/c’ed,” “d/c’d,” “d/ c’d,” “d/ c,” “D /c,” etc.). Similarly, “Continue” is abbreviated as “c’d” or “Cont.” Having the full form of these words is important because sentence/token chunkers trained on general purpose language are sensitive to punctuation, and non-standard punctuation as described above may mislead them. For example, if chunking happens in the text “d/ c’d glucophage” as (“d/”, “c’d”, “glucophage”), the model might conceive this text as a continuation rather than discontinuation.

(ii) Coreference resolution. This is an essential step in our text preparation, as it not only improves the clarity of the text but also contributes to more accurate classification of actor attributes. For example, consider a sentence like “The patient was given medication X by their doctor, who also advised them to increase their water intake.” Here, coreference resolution helps to identify that “the patient” and “them” are referring to the same entity, and that “their doctor” and “who” are referring to the same entity. This information is crucial for accurate actor classification, which can inform downstream tasks such as adverse event detection and pharmacovigilance. Therefore, we utilize AllenNLP’s[3] coreference resolution model as part of our text preprocessing pipeline to replace the repeated mentions of entities with their corresponding coreferents.

(iii) Sentence Extraction based on syntactic dependencies. Each sentence is split into its constituent phrases. We then use the spaCy library to parse each phrase into a tree of syntactic dependencies, and identify coordinated conjunction phrases (e.g., “and” or “or” phrases) in the tree. Following that, we construct a list of the longest continuous sequences of words that are dependent on these conjunctions, and remove any conjunctions from the beginning or end of each sequence. This is done by traversing the tree and collecting all conjuncts connected to the root of the tree. Finally, a list of strings representing each identified phrase is combined to form a single string. This string is taken as the sentence that contains the medication mention and its surrounding context. Algorithm 2 is responsible for finding the coordinated conjunction phrases from the parse tree, and Algorithm 1 is responsible for extracting the phrase chunks from a sentence with the aid of dependency parsing.

(iv) Sentence separation. Here, the objective is to break sentences with multiple medication names and their corresponding multiple event types. This allows us to accurately identify the events associated with each drug name. We split these sentences into different clausal components. For example, consider the sentence “Started lisinopril 10 mg p.o. daily, substituted for diltiazem.”. Clearly, the verb “started” is associated with the medication “lisinopril 10 mg p.o. daily”, and the verb “substituted” is associated with “diltiazem”. While dependency

[3] allenai.org/allennlp

Algorithm 1 get_conjunction(head)

```
$acc \leftarrow []$, $list\_heads \leftarrow [head]$
while $list\_heads \neq []$ do
    $new\_heads \leftarrow []$
    for $h$ in $list\_heads$ do
        $children \leftarrow$ children of $h$ with dependency tags "conj" or "ccomp"
        if $children \neq []$ then
            append $children$ to $new\_heads$ and $acc$
        end if
    end for
    $list\_heads \leftarrow new\_heads$
end while
return $acc$
```

parsing is capable of distilling these relations, we observe that sentences in the CMED dataset can usually be split into separate clauses where each clause exhibits only one medication change event. In our example, this approach leads to two such simpler expressions, “started lisinopril 10 mg p.o. daily” and “substituted for diltiazem”.

5 Approach

In this section, we explain our technical approach to the tasks of medication mention extraction, medical event identification, and medication event attribute classification. Further, we devote a separate description of the steps we take to detect negation.

5.1 Medication mention extraction

The task of medication mention extraction involves identifying multi-word medication phrases within free-text. As such, it is similar to medical named entity recognition (NER). Following the vast body of work that treats NER as a sequence tagging task, we utilize the beginning-inside-outside (BIO) label prefixes. Typically, medication phrases within CMED are brief, consisting of three or fewer tokens in most cases. Our approach to identifying medication mentions involves the use of BERT-based models, specifically those pretrained on domain-specific data, such as BioBERT and Bio+Clinical BERT. By adding a linear output layer and fine-tuning these models, we improve our ability to predict the specific location of medication references.

To enhance our medication mention extraction model, we experiment with incorporating the DDI (drug-drug interaction) Extraction 2013 corpus (Herrero-Zazo et al., 2013) into our training data. This is a widely recognized corpus comprising sentences from biomedical literature discussing drug-drug interactions, with each sentence annotated to indicate the medications involved in the interaction and the type of their interaction. Employing this corpus allows us to expand the number of medication mentions in our training set, leading to improved performance. The results subsequently obtained, by training only on CMED and then by training on data augmented by the DDI corpus, are shown for comparison in Table 2.

Algorithm 2 get_chunks(sentence)

```
$doc \leftarrow$ parse $sentence$ using spaCy, $chunks \leftarrow []$
for $sent$ in $doc$ do
    $conj\_phrases \leftarrow$ get coordinated conjunction phrases from $sent$'s root using get_conjunction(head)
    for $head$ in $conj\_phrases$ do
        append $head$'s subtree to $chunks$
    end for
end for
sort $chunks$ in ascending order of length
$seen \leftarrow$ empty set, $trimmed\_chunks \leftarrow []$
for $chunk$ in $chunks$ do
    $c2 \leftarrow$ list of unconsumed tokens in $chunk$
    update $seen$ set with indices of tokens in $c2$
    $c3 \leftarrow$ longest continuous sequence of tokens in $c2$
    append longest sequence in $c3$ to $trimmed\_chunks$
end for
$output \leftarrow []$
for $phrase$ in $trimmed\_chunks$ do
    remove any conjunctions at the beginning or end of $phrase$
    join the tokens in $phrase$ to form a string
    remove any leading or trailing commas from the string

    append the string to $output$
end for
sort $output$ in the original order of phrases in $sentence$
return $output$
```

5.2 Identifying negation

Even though prior work on clinical event identification has largely avoided complex negation detection, the task is nevertheless subsumed by research directed at understanding medication changes in clinical notes. In CMED, however, we find negation to be present in a very small proportion of the samples (2%). To correctly handle these instances, we employ negspaCy[4], a Python library that provides pretrained models and tools for detecting negation and other linguistic phenomena in text data. It is specifically designed to identify negated concepts, such as negated medical conditions or treatments, which are commonly encountered in clinical narratives. The library uses a combination of rule-based and statistical methods to identify negation, including the use of dependency parsing, word embeddings, and machine learning algorithms. In our study, we use Med7 (Kormil-

[4] pypi.org/project/negspacy

Table 2: Medication mentions extraction performance on the CMED test set. DDI+CMED is the combined training set of the DDI corpus and the CMED.

Dataset	**Model**	**Strict**			**Lenient**		
		P	**R**	**F1**	**P**	**R**	**F1**
CMED	BERT	0.90	0.90	0.90	0.92	0.92	0.92
	BioBERT	0.95	0.95	0.95	0.96	0.95	0.95
	Bio+Clinical BERT	0.93	0.95	0.94	0.95	0.95	0.95
DDI+CMED	BERT	0.90	0.90	0.90	0.92	0.92	0.92
	BioBERT	0.96	0.96	0.96	0.96	0.96	**0.96**
	Bio+Clinical BERT	0.93	0.94	0.93	0.95	0.96	0.94

itzin et al., 2021), a model designed to extract medication-related information from clinical notes, and integrate it with negspaCy. This integration allows us to detect instances where a drug is mentioned in the text but not prescribed.

5.3 Event and attribute classification

As described earlier in Sec 3, a medication mention must be classified as `Disposition`, `NoDisposition`, or `Undetermined`. For those identified as disposition, *i.e.*, indicating a change in medication, the next stage of the pipeline requires identifying the dimensions `action`, `actor`, `negation`, `certainty`, and `temporality` of the event, along with the correct attribute values for each dimension.

For the event and the rest of the attributes, we train a classification model based on transformer-based language models. The event and attribute classification systems assume gold standard medication mentions for model training and comparison. We conduct our experiment using Bidirectional Encoder Representations from Transformer (BERT) models pretrained on general purpose and clinical datasets. Specifically, we use BERT, RoBERTa, Bio+Clinical BERT, and XLNet. The last model, XLNet, is slightly different from the others in that it is an autoregressive Transformer model. We include it with the hope of leveraging the advantages of autoregressive language modeling as well as autoencoding.

Our goal is to classify the medication events using the sentence containing the detected medication mention as context. We use a pretrained Transformer to create a distributed representation, add 0.2 dropout, and use a fully connected layer of size 5 with softmax activation for classification. For fine-tuning with the training and development sets of CMED, we use the Hugging Face transformers package (Wolf et al., 2020). This is a multi-class classification approach, producing predictions at the sentence level for the event as well as its associated dimensional attributes.

This approach does not rely on any explicit knowledge or indication of where the medication is located. During our data preprocessing technique, we ensure that two distinct medications with varying event types are separated into their respective clauses (see Sec 4). This prevents distinct medication mentions from linguistically sharing the same events and event properties. Event classification and the attribute classification are, however, treated as separate tasks. Moreover, each attribution classifier is also trained separately. Thus, if a model is trained to predict the event type of a sentence, it will only be exposed to that specific type of label and will not be able to incorporate information from other label types.

6 Evaluation

To evaluate the accuracy of medication mention extraction systems, we employ two criteria: strict and lenient match. The strict criteria demands an exact match between the predicted and true medication mention spans. The lenient match criteria, on the other hand, considers a predicted medication mention to be correct if at least one token in the predicted mention overlaps with a token in the true mention. While strict criteria may provide a more conservative performance estimate, lenient criteria can identify more correct predictions, but at the expense of higher false positive rates. To evaluate the event and attribute classification systems, we employ precision, recall, and F1 scores, reporting both macro- and micro-averages.

6.1 Medication mention extraction

The performance of BERT, BioBERT, and Bio+Clinical BERT on this task are shown in Table 2. BioBERT achieves the highest F_1 score in strict (0.95) as well as lenient (0.96) evaluation criteria. Bio+Clinical BERT, on the other hand, achieves the highest precision scores (0.95 in both strict and lenient criteria). The slightly lower score of BERT is unsurprising, given its lack of pretraining on domain-specific data. We also note that upon augmenting the training data with the DDI corpus, a slight improvement can be seen in the F_1 score achieved by BioBERT. For Bio+Clinical BERT, however, the results are mixed. The purported advantage of this model is its pretraining on

Task		BERT			RoBERTa			XLNet			Bio+Clinical BERT		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1
Event	**Micro**	0.91	0.91	0.91	0.92	0.92	0.92	0.91	0.91	0.91	0.93	0.93	**0.93**
	Macro	0.85	0.76	0.80	0.84	0.80	0.82	0.85	0.79	0.81	0.90	0.82	0.85
Action	**Micro**	0.78	0.76	0.77	0.82	0.80	0.81	0.79	0.78	0.79	0.83	0.83	**0.83**
	Macro	0.77	0.75	0.76	0.82	0.80	0.81	0.79	0.77	0.78	0.83	0.68	0.72
Temporality	**Micro**	0.75	0.75	0.74	0.69	0.79	**0.81**	0.78	0.70	0.74	0.78	0.70	0.74
	Macro	0.72	0.62	0.68	0.63	0.59	0.61	0.65	0.72	0.68	0.75	0.70	0.70
Certainty	**Micro**	0.85	0.83	0.84	0.86	0.84	0.85	0.87	0.85	**0.86**	0.80	0.64	0.71
	Macro	0.83	0.82	0.81	0.84	0.74	0.83	0.83	0.75	0.85	0.78	0.70	0.70
Actor	**Micro**	0.92	0.91	0.92	0.91	0.89	0.90	0.94	0.92	0.93	0.93	0.93	**0.93**
	Macro	0.71	0.85	0.73	0.84	0.83	0.83	0.76	0.88	0.66	0.84	0.59	0.61

Table 3: Event and attribute classification results (with gold standard medication mentions) on the CMED test set.

biomedical literature as well as clinical notes. We conjecture that the lack of significant improvements is due to the augmentation not by clinical language, but by language from biomedical research literature (MedLine abstracts) and the DrugBank database, which form the DDI corpus.

6.2 Identifying negation

We evaluate the performance of our negation attribute classification, *i.e.*, label medication change events as *negated* or *not negated*, using Med7 and negspaCy integration. Despite the extremely small support (2% of CMED training set), our method achieves a near-perfect accuracy of 0.98. We also achieve precision, recall, and F_1 (macro average) of 0.82, 0.88, and 0.85, respectively.

6.3 Event and attribute classification

We report the results of event and attribute classification in Table 3, which shows the performance of the four language models BERT, RoBERTa, XLNet, and Bio+Clinical BERT, on the withheld CMED test set. Since this test set contains the gold-standard labels for medication mentions, our evaluations are conducted using the gold standard medication mentions as well.

Similar to results obtained by Ramachandran et al. (2023), all BERT-based language models perform well on these tasks. For event classification, the micro F_1 scores range from 0.91 to 0.93, while for attribution classification, they range from 0.77 to 0.86. In most cases, Bio+Clinical BERT outperforms the other models, achieving the highest F_1 score of 0.93 for event classification and 0.86 for `certainty` classification, as well as the highest precision of 0.94 for `actor` classification. We do report some unexpected success with RoBERTa and XLNet as well, which achieve the highest F_1 in `action` (0.83) and `temporality` (0.81) classification, respectively.

Further, we observe that the macro F_1 scores are generally lower than the micro F_1 scores, indicating that the models struggled with some classes. Specifically, `temporality` and `actor` classification showed lower performance across all models.

6.4 Discussion

When using pretrained language models to extract medication changes from clinical narratives, multiple event annotations for medication mentions can be a significant challenge, leading to prediction errors. For example, the sentence “Lovenox (will clarify timing of surgery and hold accordingly)” has two labels for the event (`undetermined` and `disposition`) for the medication Lovenox, potentially confusing the model. One solution to this issue is to modify the task from a sentence classification task to a multi-label classification task. However, there may be cases where a sentence follows a multi-label scheme, but only one type of annotation is provided. For instance, “DM2: Continue home meds (metformin + insulin), hold when on diet without substantial calories (clears, NPO)” only has the `action` label *start* for the metformin, whereas there is a need for the second attribute label *stop* as well.

During our analysis, we observed instances of incorrect or ambiguous labeling in the annotation, including the `actor` and `temporality` dimensions. For example, in “SL TNG prescribed but not used,” there are two `actor` labels (*patient* and *physician*) for the medication TNG, and in “amox 500 TID x 10d: fluids, steam, acetaminophen,” the medication amox has two `temporality` labels (*past* and *future*). Furthermore, in “We will initiate Zetia to add to the Pravachol,” the event is labeled as `NoDisposition`, despite the word “initiate” clearly suggesting otherwise.

Additionally, we noticed several mistakes in the negation class, such as “Not on beta-blocker ” being labeled as non-negated. The limited number of samples in the negated category, combined with the annotation errors in the test set, has a clear and significant impact on any model. As the model’s training relies heavily on the quality and quantity of the data, a small and incorrectly labeled dataset is particularly harmful. We also noticed several non-English sentences in the training set, such as “Hctz (HYDROCHLOROTHIAZIDE) 12.5 MG (12.5MG CAPSULE take 1) PO QD, Para la presión alta- si se siente muy mareado deje de tomarla y avísele a su médico immediatamente.” While any effort to utilize the advances of natural language processing in clinical applications in multiple languages is laudable, the presence of very few instances of other languages in an otherwise English corpus has a negative impact.

It is noted in the dataset annotation that medication-related information is contained within a single sentence. However, we observe that this is not always the case. There are several instances where information about a single medication event extends beyond a single sentence, requiring the model to analyze multiple sentences in order to identify the relevant context. The dataset includes a number of sentences that are labeled as *undetermined*, many of which are located within an assessment and plan (A/P) section of a medical document. This section can be quite lengthy and contain numerous mentions of medications without specific attributes or events. To correctly classify these undetermined sentences, it is often necessary to look beyond the sentence itself and recursively search for information related to medication events within the A/P section. However, we believe this is a challenging task beyond the ambit of the CMED dataset’s annotation description.

7 Conclusion

Our analysis of CMED and its constituent tasks reveal three main characteristics. First, it is often necessary to consider additional context beyond the specific sentence containing the medication mention to accurately label medication references. This context could include information from previous or subsequent sentences, the patient’s medical history, or other relevant information further away in a document (as often found in the assessment and plan sections) that could impact the interpretation of the medication mention. Second, we observe the prevalence of multiple medication references within a single sentence, which poses a challenge for accurate extraction. Finally, accurate identification of the start and end of a sentence containing a medication mention is also challenging, since standard sentence splitting and tokenization methods often fail in clinical notes, especially if task-specific or domain-specific preprocessing is not done.

We especially underscore the importance of data preprocessing when training or fine-tuning models for the medical domain. In this work, for example, we perform abbreviation resolution, coreference resolution, syntactic dependency-based sentence extraction, and a custom sentence extraction with phrase chunking.

Similar to other recent findings, our study demonstrates that pretrained language models are extremely effective in complex clinical information extraction, when fine-tuned on carefully chosen domain data. Overall, our study affirms the utility of Transformer-based models, particularly BioBERT and Bio+Clinical BERT, in medication information extraction from clinical notes. We also exhibit the additional advantage of training such models with augmented domain data.

References

Daniel Albright, Arrick Lanfranchi, Anwen Fredriksen, William F. Styler IV, Colin Warner, Jena D. Hwang, Jinho D. Choi, Dmitriy Dligach, Rodney D. Nielsen, James Martin, Wayne Ward, Martha Palmer, and Guergana K. Savova. 2013. Towards comprehensive syntactic and semantic annotations of the clinical narrative. Journal of the American Medical Informatics Association, 20(5):922–930.

Ghada Alfattni, Maksim Belousov, Niels Peek, Goran Nenadic, et al. 2021. Extracting Drug Names and Associated Attributes From Discharge Summaries: Text Mining Study. JMIR medical informatics, 9(5):e24678.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jin, 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.

Jeffery L. Belden, Pete Wegier, Jennifer Patel, Andrew Hutson, Catherine Plaisant, Joi L. Moore, Nathan J. Lowrance, Suzanne A. Boren, and Richelle J. Koopman. 2019. Designing a medication timeline for patients and physicians. Journal of the American Medical Informatics Association, 26(2):95–105.

Justin Cadwallader, Kenneth Spry, Justin Morea, AL Russ, Jon Duke, and Michael Weiner. 2013. Design of a Medication Reconciliation Application. Applied Clinical Informatics, 4(01):110–125.

Wendy W. Chapman, Will Bridewell, Paul Hanbury, Gregory F. Cooper, and Bruce G. Buchanan. 2001. A Simple Algorithm for Identifying Negated Findings and Diseases in Discharge Summaries. Journal of Biomedical Informatics, 34(5):301–310.

Anni Coden, Daniel Gruhl, Neal Lewis, Michael Tanenblatt, and Joe Terdiman. 2012. SPOT the drug! An unsupervised pattern matching method to extract drug names from very large clinical corpora. In Proceedings of the IEEE 2nd International Conference on Healthcare Informatics, Imaging and Systems Biology, pages 33–39, San Diego, CA, USA.

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.

Richard J. FitzGerald. 2009. Medication errors: the importance of an accurate drug history. Br J Clin Pharmacol, 67(6):671–675.

Thierry Hamon and Natalia Grabar. 2010. Linguistic approach for identification of medication names and related information in clinical narratives. Journal of the American Medical Informatics Association, 17(5):549–554.

María Herrero-Zazo, Isabel Segura-Bedmar, Paloma Martínez, and Thierry Declerck. 2013. The DDI corpus: An annotated corpus with pharmacological substances and drug–drug interactions. Journal of Biomedical Informatics, 46(5):914–920.

Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Liwei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovitz, Leo A. Celi, and Roger G. Mark. 2016. MIMIC-III, a freely accessible critical care database. Scientific Data, 3:160035.

Jordan Jouffroy, Sarah F Feldman, Ivan Lerner, Bastien Rance, Anita Burgun, Antoine Neuraz, et al. 2021. Hybrid deep learning for medication-related information extraction from clinical texts in French: MedExt algorithm development study. JMIR Medical Informatics, 9(3):e17934.

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.

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.

Ivan Lerner, Jordan Jouffroy, Anita Burgun, and Antoine Neuraz. 2020. Learning the grammar of drug prescription: recurrent neural network grammars for medication information extraction in clinical texts. arXiv preprint arXiv:2004.11622.

Feifan Liu, Richeek Pradhan, Emily Druhl, Elaine Freund, Weisong Liu, Brian C Sauer, Fran Cunningham, Adam J Gordon, Celena B Peters, and Hong Yu. 2019a. Learning to detect and understand drug discontinuation events from clinical narratives. Journal of the American Medical Informatics Association, 26(10):943–951.

Shengyu Liu, Buzhou Tang, Qingcai Chen, and Xiaolong Wang. 2015. Drug Name Recognition: Approaches and Resources. Information, 6(4):790–810.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv preprint arXiv:1907.11692.

Kevin Lybarger, Mari Ostendorf, Matthew Thompson, and Meliha Yetisgen. 2021. Extracting COVID-19 diagnoses and symptoms from clinical text: A new annotated corpus and neural event extraction framework. Journal of Biomedical Informatics, 117:103761.

Diwakar Mahajan, Jennifer J Liang, and Ching-Huei Tsou. 2021. Toward Understanding Clinical Context of Medication Change Events in Clinical Narratives. In AMIA Annual Symposium Proceedings, volume 2021, page 833. American Medical Informatics Association.

Saeed Mehrabi, Anand Krishnan, Sunghwan Sohn, Alexandra M. Roch, Heidi Schmidt, Joe Kesterson, Chris Beesley, Paul Dexter, C. Max Schmidt, Hongfang Liu, and Mathew Palakal. 2015. DEEPEN: A negation detection system for clinical text incorporating dependency relation into NegEx. Journal of Biomedical Informatics, 54:213–219.

Roser Morante. 2010. Descriptive analysis of negation cues in biomedical texts. In LREC, volume 2010, pages 1429–1436.

Sankaran Narayanan, Kaivalya Mannam, Pradeep Achan, Maneesha V Ramesh, P Venkat Rangan, and Sreeranga P Rajan. 2022. A contextual multitask neural approach to medication and adverse events identification from clinical text. Journal of Biomedical Informatics, 125:103960.

Naveen S. Pagad and N. Pradeep. 2022. Clinical Named Entity Recognition Methods: An Overview. In International Conference on Innovative Computing and Communications, pages 151–165, Singapore. Springer Singapore.

Serguei V Pakhomov, Alexander Ruggieri, and Christopher G Chute. 2002. Maximum entropy modeling for mining patient medication status from free text. In Proceedings of the AMIA Symposium, page 587. American Medical Informatics Association.

Catherine Plaisant, Richard Mushlin, Aaron Snyder, Jia Li, Dan Heller, and Ben Shneiderman. 2003. LifeLines: using visualization to enhance navigation and analysis of patient records. In The craft of information visualization, pages 308–312. Elsevier.

Eric G Poon, Barry Blumenfeld, Claus Hamann, Alexander Turchin, Erin Graydon-Baker, Patricia C McCarthy, John Poikonen, Perry Mar, Jeffrey L Schnipper, Robert K Hallisey, et al. 2006. Design and Implementation of an Application and Associated Services to Support Interdisciplinary Medication Reconciliation Efforts at an Integrated Healthcare Delivery Network. Journal of the American Medical Informatics Association, 13(6):581–592.

Giridhar Kaushik Ramachandran, Kevin Lybarger, Yaya Liu, Diwakar Mahajan, Jennifer J Liang, Ching-Huei Tsou, Meliha Yetisgen, and Özlem Uzuner. 2023. Extracting Medication Changes in Clinical Narratives using Pre-trained Language Models. Journal of Biomedical Informatics, page 104302.

Daniel Sanchez-Cisneros, Paloma Martínez, and Isabel Segura-Bedmar. 2013. Combining Dictionaries and Ontologies for Drug Name Recognition in Biomedical Texts. In Proceedings of the 7th International Workshop on Data and Text Mining in Biomedical Informatics, DTMBIO '13, pages 27—-30, New York, NY, USA. Association for Computing Machinery.

Isabel Segura-Bedmar, Paloma Martínez, and María Segura-Bedmar. 2008. Drug name recognition and classification in biomedical texts: A case study outlining approaches underpinning automated systems. Drug Discovery Today, 13(17):816–823.

Sunghwan Sohn, Sean P Murphy, James J Masanz, Jean-Pierre A Kocher, and Guergana K Savova. 2010. Classification of medication status change in clinical narratives. In AMIA Annual Symposium Proceedings, volume 2010, page 762. American Medical Informatics Association.

György Szarvas, Veronika Vincze, Richárd Farkas, and János Csirik. 2008. The BioScope corpus: annotation for negation, uncertainty and their scope in biomedical texts. In Proceedings of the workshop on current trends in biomedical natural language processing, pages 38–45.

Alexander Turchin, Maria Shubina, Eugene Breydo, Merri L Pendergrass, and Jonathan S Einbinder. 2009. Comparison of information content of structured and narrative text data sources on the example of medication intensification. Journal of the American Medical Informatics Association, 16(3):362–370.

Özlem Uzuner, Brett R. South, Shuying Shen, and Scott L. DuVall. 2011. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. Journal of the American Medical Informatics Association, 18(5):552–556.

Yanshan Wang, Liwei Wang, Majid Rastegar-Mojarad, Sungrim Moon, Feichen Shen, Naveed Afzal, Sijia Liu, Yuqun Zeng, Saeed Mehrabi, Sunghwan Sohn, et al. 2018. Clinical information extraction applications: a literature review. Journal of biomedical informatics, 77:34–49.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 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.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. XLNet: Generalized Autoregressive Pretraining for Language Understanding. Advances in neural information processing systems, 32.

Xin Yu, Wenshen Hu, Sha Lu, Xiaoyan Sun, and Zhenming Yuan. 2019. BioBERT Based Named Entity Recognition in Electronic Medical Record. In 2019 10th International Conference on Information Technology in Medicine and Education, ITME, pages 49–52, Qingdao, Shandong, China. IEEE.