MIMICause: Representation and automatic extraction of causal relation types from clinical notes

Anonymous ACL submission

Abstract

Understanding causal narratives communicated in clinical notes can help make strides towards personalized healthcare. Extracted causal information from clinical notes can be combined with structured EHR data such as patients' demographics, diagnoses, and medications. This will enhance healthcare providers' ability to identify aspects of a patient's story communicated in the clinical notes and help make more informed decisions.

In this work, we propose annotation guidelines, develop an annotated corpus and provide baseline scores to identify types and direction of causal relations between a pair of biomedical concepts in clinical notes; communicated implicitly or explicitly, identified either in a single sentence or across multiple sentences.

We annotate a total of 2714 de-identified examples sampled from the 2018 n2c2 shared task dataset and train four different language model based architectures. Annotation based on our guidelines achieved a high inter-annotator agreement i.e. Fleiss' kappa (κ) score of 0.72, and our model for identification of causal relations achieved a macro F1 score of 0.56 on the test data. The high inter-annotator agreement for clinical text shows the quality of our annotation guidelines while the provided baseline F1 score sets the direction for future research towards understanding narratives in clinical texts.

1 Introduction

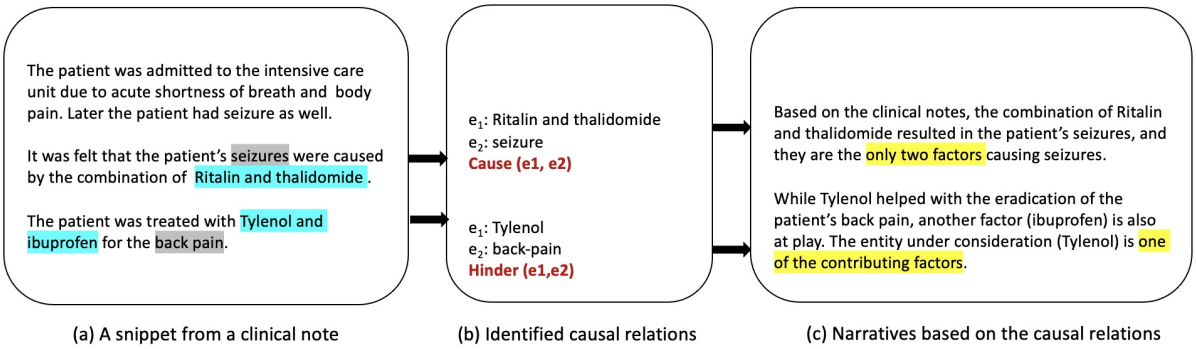
Electronic Health Records (EHRs) have significant amounts of unstructured clinical notes containing a rich description of patients' states as observed by healthcare professionals over time. Our ability to effectively parse and understand clinical narratives depends upon the quality of extracted biomedical concepts and semantic relations.

The contemporary advancements in natural language processing (NLP) have led to an increased interest in tasks such as extraction of biomedical concepts, patients' data de-identification, medical question answering and relation extraction. While these tasks have improved our ability for clinical narrative understanding, identification of semantic causal relations between biomedical entities will further enhance it.

Identification of novel and interesting causal observations from clinical notes can be instrumental to a better understanding of patients' health. It can also help us identify potential causes of diseases and determine their prevention and treatment. Despite the usefulness of identification and extraction of causal relation types, our capability to do so is limited and remains a challenge for specialized domains like healthcare.

The NLP community has been actively working on causality understanding from text and has proposed various methodologies to represent (Talmy, 1988; Wolff, 2007; Swartz, 2014; Hassanzadeh et al., 2019), as well as extract (Mirza and Tonelli, 2014; O'Gorman et al., 2016; Mirza and Tonelli, 2016; Gao et al., 2019; Khetan et al., 2022), causal associations between the events expressed in natural language text. In the healthcare domain, most of the related work can be grouped around the problem of adverse drug effect identification from biomedical scientific articles (Gurulingappa et al., 2012) or clinical notes (Johnson et al., 2016; Liu et al., 2019; Henry et al., 2020; Rawat et al., 2020), and identification of cause, effect and their triggers (Mihaila et al., 2012). There is no work that has yet tried to represent different types of causal associations along with direction (between biomedical concepts) communicated in clinical notes.

In this work, we fill the gap by defining types of semantic causal relations between biomedical entities, building detailed annotation guidelines and annotating a large dataset. Figure 1 shows a snippet of clinical note extracted from the n2c2 dataset (Henry et al., 2020), different sets of annotated



(a) A snippet from a clinical note (b) Identified causal relations (c) Narratives based on the causal relations

Figure 1: (a) A snippet from a clinical note with highlighted biomedical entities identified in the n2c2 dataset. (b) Causal relations identified between the specified biomedical entities (e1 and e2). In the first case, two entities are specified together as e1 for causal relation identification, while the second case specifies only one entity as e1. (c) Narratives based on the causal relations identified between the specified biomedical entities

biomedical entities along with the causal relationship between them, and the corresponding narrative based on the proposed guidelines outlined in Section 3.1.

Even with the inherent complexities of clinical text data (e.g., domain knowledge, short hand by doctors, etc.), following our proposed guidelines, we achieved a high inter annotator agreement of Fleiss' kappa (κ) score of 0.72[1].

2 Related Works

In linguistics, the focus on representing causality has been on understanding interactions between events. Talmy (1988) proposed force-dynamics to decompose the causal interaction between events as "letting", "helping", "hindering" etc. Wolff (2007) built upon force-dynamics by incorporating the theory of causal verbs and proposed the Dynamic-model of causation. Wolff categorised causation in three categories, *"Cause"*, *"Enable"* and *"Prevent"*, and provided a set of causal verbs to express these categories.

Dunietz et al. (2015; 2017) proposed BECauSE Corpus to represent linguistic expressions of causation stated explicitly. BECauSE 1.0 (Dunietz et al., 2015) consists of a cause span, an effect span, and a causal connective span. Their work treats the causal connectives e.g. *because of, so* etc. as the "centerpiece" of causal language, impacting the selection of instances to be annotated. In addition to the types of causation (Consequence, Motivation, and Purpose) and degrees of causation (Facilitate and Inhibit) introduced in BECauSE 1.0, the subsequent work BECauSE 2.0 (Dunietz et al., 2017) extended the annotation scheme to include overlapping relations other than causal. In contrast, our work focuses on both explicit (indicated by connectives) and implicit (lack of connectives) identification of types of causal associations between biomedical concepts as communicated in clinical notes.

More recently, Mostafazadeh et al. (2016b) built upon the work of Wolff and proposed annotation framework *CaTeRS* to represent causal relations between events for commonsense perspective. CaTeRS categorises semantic relations between events to capture causal and temporal relationships for narrative understanding on crowd-sourced ROC-Stories dataset (Mostafazadeh et al., 2016a) but has only 488 causal links. In comparison, our MIMI-Cause dataset is built on actual clinical narratives, i.e., *MIMIC-III Clinical text data* (Johnson et al., 2016) and has 1923 causal observations.

Another interesting decomposition of causation is proposed by Swartz (2014) as a necessary and sufficient condition, but such detailed information is seldom communicated in clinical notes. There have been several other recent attempts of modeling and extracting causality from unstructured text. Bethard et al. (2008) created a causality dataset using the Wall Street Journal corpus and captured the directionality of causal interaction with simple temporal relations (e.g., Before, After, No-Rel) but did not focus on the types of causality between the events. The work of Gorman et al. on Richer Event Description (RED) (Ikuta et al., 2014) describes causality types as cause and precondition and uses negative polarity to capture the context of hinder and prevent. This is in line with the annotation

[1]Upon acceptance of this paper, we will be releasing annotated dataset and code.

guidelines proposed in our current work, but we also defined explicit **Hinder** and **Prevent** causality types along with directionality.

Mirza et al. (2014) proposed the use of explicit linguistic markers, i.e., CLINKs (due to, because of, etc.) to extended TimeML TLINKs (Pustejovsky et al., 2003) based temporal annotations to capture causality between identified events. The resulting dataset had temporal as well as casual relations but still lacks the causality types between events. Hassanzadeh et al. (2019) proposed the use of binary questions to extract causal knowledge from unstructured text data but did not focus on types and directionality of causal relations. More recently, Khetan et al. (2022) used language models combining event descriptions with events' contexts to predict causal relationships. Their network architecture wasn't trained to predict the type or directionality of causal relations. Furthermore, they removed the directionality provided in SemEval-2007 (Girju et al., 2007), and SemEval-2010 (Hendrickx et al., 2009) datasets to evaluate their model on a larger causal relation dataset. Our causality extraction network is built upon their methodology, i.e., *Causal-BERT* but also focuses on directionality as well as types of causality communicated in clinical notes.

Although causality lies at the heart of biomedical knowledge, there are only a handful of works (mostly Adverse Drug Effect (Gurulingappa et al., 2012)) extracting causality from biomedical or clinical text data. One interesting work is BioCause by Mihaila et al. (2012), which annotates existing bio-event corpora from biomedical scientific articles to capture biomedical causality. Instead of identifying the types (and direction) of causal relations in the already provided events of interest, they are annotating two types of text spans, i.e., arguments and triggers. Arguments are text spans that can be represented as events with type Cause, Effect, and Evidence while Trigger spans (can be empty) are connectives between the casual events.

Our work proposes comprehensive guidelines to represent the types and direction of causal associations between biomedical entities, expressed explicitly or implicitly in the same or multiple sentences in clinical notes, and is not covered by any related work.

Concepts/Entities	Examples
Drug	morphine, ibuprofen, antibiotics (or "abx" as its abbreviation), chemotherapy etc.
ADE and Reason*	nausea, seizures, Vitamin K deficiency, cardiac event during induction etc.
Strength	10 mg, 60 mg/0.6 mL, 250/50 (e.g. as in Advair 250/50), 20 mEq, 0.083% etc.
Form	Capsule, syringe, tablet, nebulizer, appl (abbreviation for apply topical) etc.
Dosage	Two (2) units, one (1) mL, max dose, bolus, stress dose, taper etc.
Frequency	Daily, twice a day, Q4H (every 4 Hrs), prn (pro re nata i.e as needed) etc.
Route	Transfusion, oral, gtt (guttae i.e. by drops), inhalation IV (i.e. Intravenous) etc.
Duration	For 10 days, chronic, 2 cycles, over 6 hours, for a week etc.

*The distinction between ADE and Reason concepts is based on whether the drug was given to address the disease (Reason) or led to the disease (ADE).

Table 1: Examples of Bio-medical concepts/entities in the 2018 n2c2 shared task dataset.

3 MIMICause Dataset creation

We used publicly available 2018 n2c2 shared task (Henry et al., 2020) dataset on adverse drug events and medication extraction to build the MIMICause dataset. The n2c2 dataset was used because it is built upon the de-identified discharge summaries from the MIMIC-III clinical care database (Johnson et al., 2016) and has nine different annotations of biomedical entities e.g. Drug, Dose, ADE, Reason, Route etc. The types of biomedical concepts/entities with a few examples as defined in the n2c2 dataset are shown in Table 1.

However, the provided relationships in the n2c2 dataset are simply defined by the identified concepts linked with related medications and hold no semantic meaning. To create the MIMICause dataset, we extracted[2] examples from each entity-pair available in the n2c2 dataset. Our final dataset has 1107 "ADE-Drug" , 1007 "Reason-Drug" and 100 from each of "Strength-Drug", "Form-Drug", "Dosage-Drug", "Frequency-Drug", "Route-Drug" and "Duration-Drug" entity-pair examples.

3.1 Annotation guidelines

Our annotation guidelines are defined to represent nine semantic causal relationships between biomedical concepts/entities in clinical notes. Our guidelines have four types of causal associations, each with two directions, and a non-causal "Other" class. Based on our guidelines, *causal* relationship/association exists when one or more entities affect another set of entities. The driving concept

[2]We used https://spacy.io/ library with *"en_core_web_sm"* language model.

can be a ***single*** entity such as a drug / procedure / therapy or a ***composite entity*** such as several drugs / procedures / therapies considered together.

3.1.1 Direction of causal association

The direction of causal association between entities is captured by the order of entity tags ((e_1, e_2) or (e_2, e_1)) in the defined causal relationships. Either entity can be referred to as e_1 or e_2. The *entity that initiates or drives the causal interaction is placed first* in parenthesis followed by the resulting entity or effect.

1. Odynophagia: Was presumed due to **<e2>mucositis</e2>** from recent **<e1>chemotherapy</e1>**.
2. Odynophagia: Was presumed due to **<e1>mucositis</e1>** from recent **<e2>chemotherapy</e2>**.

Example (1) and (2) are different because the entity references are reversed. Regardless of the entity tags, in the context of the example, "chemotherapy" is the *driving entity* that led to the *emergence* of "mucositis". Therefore, example (1) is annotated with causal direction (e_1, e_2) while example (2) is annotated with (e_2, e_1).

3.1.2 Explicitness / Implicitness of the causal indication

Our guidelines also capture causality expressed both explicitly and implicitly. In example (1), the causality is expressed explicitly using lexical causal connective "due to". Whereas in example (3), the causal association between "erythema" and "Dilantin" can only be understood based on the overall context of all the sentences.

3. patient's wife noticed **<e2>erythema on patient's face</e2>**. On [**3-27**]the visiting nurse [**First Name (Titles) 8706**][**Last Name (Titles)11282**]of a rash on his arms as well. The patient was noted to be febrile and was admitted to the [**Company 191**] Firm. In the EW, patient's **<e1>Dilantin</e1>** was discontinued and he was given Tegretol instead.

3.1.3 (Un)-certainty of causal association

Establishing real-world causality or the task of causal inference is not in the scope of our current work. Our proposed guidelines represent a potential causal association between biomedical entities either expressed as speculation or with certainty in a similar manner.

4. # **<e1>Normocytic Anemia</e1>** - Was 32.8 at OSH; after receiving fluids HCT has fallen further to 30. Baseline is 35 - 40. Not clinically bleeding. Perhaps due to **<e2>chemotherapy</e2>**.

In example (4), causality between biomedical entities is speculated through "Perhaps". While representing speculative causal associations can further enrich narrative understanding; it is not covered in our current work.

3.1.4 Types of causal associations

This section provides detailed guidelines for various types of causal relations (each with two directions) and one non-causal relation ("Other") along with accompanying examples.

- **Cause**(e_1, e_2) or **Cause**(e_2, e_1) – Causal relations between biomedical entities are of these classes if the emergence, application or increase of a **single or composite entity** *exclusively leads* to the **emergence** or **increase** of one or a set of entities.

  5. It was felt that the patient's **<e2>seizures</e2>** were caused by the combination of **<e1>Ritalin and thalidomide</e1>**.

  In example (5), "seizures" occurred due to *two drugs* viz. "Ritalin" and "thalidomide". The entity span covers both of them, and they are considered together as a *composite entity* leading to "seizures". Hence, example (5) is annotated as Cause(e_1, e_2). The annotation would have been different had these entities been considered individually.

  Thus, the "Cause" category is assigned only if the driving entity is responsible in its entirety for the effect. If the specified entity is responsible for the effect in part, then a different causal relation is defined to express this contrast.

- **Enable**(e_1, e_2) or **Enable**(e_2, e_1) – Causal relations between biomedical entities are of these classes if the emergence, application or increase of a **single or composite entity** *leads* to the **emergence** or **increase** of one or

a set of entities in a setting ***where*** a number of factors are at play and the *single or composite entity* under consideration is one of the contributing factors.

6. It was felt that the patient's <e2>seizures</e2> were caused by the combination of <e1>Ritalin</e1> and thalidomide.

Example (6) is the same as example (5) except for the entities in considerations. Both the drugs viz. "Ritalin" and "thalidomide" are contributing to the "seizures". Since the example is considering only "Ritalin", *which is a contributing factor in part*, it is annotated as Enable(e_1, e_2).

With the "Enable" relation type, it can easily be noted that discontinuing only "Ritalin" or "thalidomide" will not lead to the stopping of "seizures". Labelling these samples as "Cause" would have suppressed this detail, and the actions taken based on this would not have been sufficient.

- **Prevent**(e_1, e_2) or **Prevent**(e_2, e_1) – Causal relations between biomedical entities are of these classes if the emergence, application or increase of a **single or composite entity** *exclusively leads* to the **eradication, prevention** or **decrease** of one or a set of entities.

  This class includes the scenario of ***preventing*** a disease or condition from occurring as well as ***curing*** a disease or condition if it has occurred.

  7. You were treated with <e2>tylenol and ibuprofen</e2> for your <e1>back pain</e1>.

  In example (7), "tylenol" and "ibuprofen" are the two different entities used in conjunction to resolve the "back pain". Since the causal relation is to be identified by considering them as a *composite entity*, the example is labelled as Prevent(e_2, e_1). The annotation would have been different had these entities been considered individually.

- **Hinder**(e_1, e_2) or **Hinder**(e_2, e_1) – Causal relations between biomedical entities are of these classes if the emergence, application or increase of a **single or composite entity** *leads* to the **eradication, prevention** or **decrease** of one or a set of entities in a setting ***where*** a number of factors are at play and the *single or composite entity* under consideration is one of the contributing factors.

  Similar to "Prevent", this label also includes the scenario of ***hindering*** a disease or condition from occurring as well as ***curing*** a disease or condition if it has occurred.

  8. You were treated with <e2>tylenol</e2> and ibuprofen for your <e1>back pain</e1>.

  Example (8) is the same as example (7) except for the entities in considerations. Both the entities i.e. "tylenol" and "ibuprofen" are contributing to the resolution of "back pain". Since the example is considering only "tylenol", *individually as a contributing factor in part*, it is annotated as Hinder(e_2, e_1).

  This distinction between "Prevent" and "Hinder" can be useful in scenarios such as identifying conditions that may require the use of multiple drugs for treatment.

- **Other** – We defined the "Other" class to annotate examples with non-causal interaction between biomedical entities. Examples of the "Other" class can either have no relationship between biomedical entities of interest or some other semantic relationship that's not causal. Being non-causal, the "Other" class doesn't have a sense of direction associated with it.

  Based on our guidelines, examples with ambiguous overall context for all the annotators, entities with indirect causal association (an entity leading to a condition which in turn affects another entity) and samples from non-causal entity-pairs in the n2c2 dataset (i.e., Form-Drug, Route-Drug, etc.) are also labelled as "Other".

  9. Patient has tried and failed <e2>Nexium</e2>, reporting it has not helped his <e1>gastritis</e1> for 3 months.
  10. Thus it was believed that the pt's <e1>altered mental status</e1> was secondary to <e2>narcotics</e2> withdrawal.

		Annotation	Count
Causal	e_1 as agent, e_2 as effect	**Cause(e_1, e_2)**	354
		Enable(e_1, e_2)	174
		Prevent(e_1, e_2)	261
		Hinder(e_1, e_2)	154
	e_2 as agent, e_1 as effect	**Cause(e_2, e_1)**	370
		Enable(e_2, e_1)	176
		Prevent(e_2, e_1)	249
		Hinder(e_2, e_1)	185
Other	–	**Other**	791
Total			**2714**

Table 2: *Causal* types and their final counts

11. Atenolol was held given patient was still on <e2>amiodarone</e2> <e1>taper</e1>.

In example (9), "Nexium" was taken to prevent / cure "gastritis" but the expected effect is explicitly stated to be not observed. In example (10), the "altered mental status" is observed due to "narcotics withdrawal", however, the entity span refers only to the "narcotics". Example (11) is from the "Dosage-Drug" entity-pair of the n2c2 dataset and has no causal association between the entities.

Therefore, these examples are annotated as "Other". Similarly, examples with entity-pairs from "Form-Drug", "Strength-Drug", "Frequency-Drug", 'Route-Drug" and "Duration-Drug" are also labelled as "Other".

To summarize, we defined annotation guidelines for nine semantic causal relations (8 Causal + Other) between biomedical entities expressed in clinical notes. Our annotated dataset has examples with both explicit and implicit causality in which entities are in the same sentence or different sentences. The final count of examples for each *causal* type with direction is in Table 2.

3.2 Inter-annotator agreement

It's difficult to comprehend narratives expressed in clinical notes due to the need of domain knowledge, short hand used by the doctors, use of abbreviations (Table 3), context spread over many sentences as well as the explicit and implicit nature of communication.

Three authors of this paper (all with fluency in English language and computer science background) annotated the dataset. Given the nature of our base data (MIMIC-III discharge summaries)

Abbreviation	Expansion	Abbreviation	Expansion
b/o	because of	d/c'd	discontinued
HCV	Hepatitis C Virus	abx	anti-biotics
DM	Diabetes Mellitus	c/b	complicated by
s/p	status post	h/o	history of

Table 3: Clinical abbreviations in the dataset

and the critical importance of our task (causal relations between biomedical entities), the annotators followed the provided guidelines, referred to sources such as websites of Centers for Disease Control and Prevention (*CDC*[3]), National Institute of Health (*NIH*[4]), and *WebMD*[5] to understand domain-specific keywords or abbreviations, and had regular discussions about the annotation tasks.

We performed three rounds of annotation, refining our guidelines after each round by discussing various complex examples and edge cases. We achieved an inter-annotator agreement (IAA) **Fleiss' kappa** (κ) score of 0.72, which indicates substantial agreement and the quality of our annotation guidelines. We did majority voting over the three available annotations to obtain the final gold annotations for our "MIMICause" dataset. In case of disagreements, another author of this paper acted as a master annotator, making the final decision on annotations after discussion with the other three annotators.

A direct comparison of our IAA score with other works is not possible due to differences in the number of annotators, annotation labels, guidelines, reported metrics etc. for different datasets. However, for reference, we discuss IAA scores reported for the task of semantic link annotations, particularly those where κ scores were reported. Of note is the work by Mostafazadeh et al. (2016b) and their annotation framework CaTeRS for temporal and causal relations in ROCStories corpus where the final κ score achieved was 0.51 among four annotators. Similarly, Bethard et al. (2008) reported a κ score of 0.56 and an F-measure (F-1 score) of 0.66 with two annotators labelling for only two relations viz. causal and no-rel. In the clinical domain, Bethard et al. (2017) reported a final IAA agreement (F-1) score of 0.66 on the latest Clinical TempEval dataset (Task 12 of SemEval-2017) labelled by two annotators. However, the relation types in Clinical TempEval are temporal and not

[3] https://www.cdc.gov/
[4] https://www.nih.gov/
[5] https://www.webmd.com/

causal, making the agreement score incomparable.

4 Problem definition and Experiments

We defined our task of causality understanding as the identification of semantic causal relations between biomedical entities as expressed in clinical notes. We have a total of 2714 examples annotated with these 9 different classes (8 causal and 1 non-causal).

4.1 Problem Formalization

We pose the task of causal relation identification as a multi-class classification problem $f : (X, e_1, e_2) \mapsto y$, where X is an input text sequence, e_1 and e_2 are the entities between which the relation is to be identified, and $y \in \mathcal{C}$ is the label from the set of *nine* relations. These samples are taken from the *MIMICause* dataset $\mathcal{D} = \{(X, e_1, e_2, y)_m\}_{m=1}^{m=N}$, where N is the total number of samples in the dataset. The text and entities are mathematically denoted as:

$$X = [x_1, x_2, \ldots, x_{n-1}, x_n] \quad (1)$$

$$e_1 = X[i:j] = [x_i, x_{i+1}, \ldots, x_j] \quad (2)$$

$$e_2 = X[k:l] = [x_k, x_{k+1}, \ldots, x_l] \quad (3)$$

where n is the sequence length, i, j, k and $l \in [1..n]$, $i \leq j$ and $k \leq l$ i.e. entities are subsequences of continuous span within the text X. Additionally, $j < k$ or $l < i$ holds i.e. the entities e_1 and e_2 are non-overlapping and either of these can occur first in the sequence X.

4.2 Models

As a baseline for this dataset, we built our causal relation classification models using two different language models[6] as text encoders (BERT-BASE and Clinical-BERT) and a fully connected feed-forward network (FFN) as the classifier head. The encoder output that captures the bi-directional context of the input text X through the [CLS] token is denoted by $H_0 \in R^d$, where $d = 768$ is the dimension of the encoded outputs from BERT-BASE / Clinical-BERT. The formulations of the layers of the classifier head are given by:

$$K_1 = \mathit{dropout}(ReLU(W_1 H_0 + b_1))) \quad (4)$$

$$K_2 = W_2 K_1 + b_2 \quad (5)$$

$$p = \mathit{softmax}(K_2) \quad (6)$$

[6]We use the implementation of all the encoders from the huggingface (Wolf et al., 2020) repository

where $W_1 \in R^{d' \times d}$, $W_2 \in R^{L \times d'}$, d' was set to 256 and $L = 9$ is the number of labels.

Architectures with additional context introduced between the encoder and classifier head by concatenating averaged representation of the two entities and encoder output were also tried, which led to improved results. The augmented context is denoted by:

$$H_{e_1} = \frac{1}{j-i+1} \sum_{t=i}^{j} H_t \quad (7)$$

$$H_{e_2} = \frac{1}{l-k+1} \sum_{t=k}^{l} H_t \quad (8)$$

$$H' = \mathit{concat}(H_0, H_{e_1}, H_{e_2}) \quad (9)$$

$$H_0 = \mathit{dropout}(ReLU(W_0 H' + b_0)) \quad (10)$$

where i, j, k and l are the start and end indices of the entities, $H_t \in R^d$, $H' \in R^{3d}$, $W_0 \in R^{d \times 3d}$ and the augmented context is assigned back to H_0 for feeding into the classifier head. The architecture details without and with the entity context augmentation are shown in Figure (2) and (3) respectively. An overview of the models is given below:

- ***Encoder (BERT-BASE / Clinical-BERT) with feed-forward network (FFN)*** – The overall architecture as shown in Figure 2 is a simple feed-forward network built on top of a pre-trained encoder. The input sentence is fed as a sequence of tokens to the encoder, with encoder based special tokens such as [CLS] and entity tagging tokens such as $<e1>, </e1>$. The overall sentence context is passed through the fully connected feed-forward network to obtain class probabilities as formulated in equations (4)–(6).

  In addition to the BERT-BASE encoder, we also used the Clinical-BERT encoder to obtain the contextualised representation of our input examples. While BERT is pre-trained on standard corpus such as Wikipedia, Clinical-BERT is pre-trained on clinical notes and provides more relevant representation for our dataset, and hence led to a significant increase in the evaluation metrics.

- ***Encoder (BERT-BASE / Clinical-BERT) with entity context augmented feed-forward network (FFN)*** – The overall architecture is shown in Figure 3. While the input with special tokens, encoding and classifier head re-

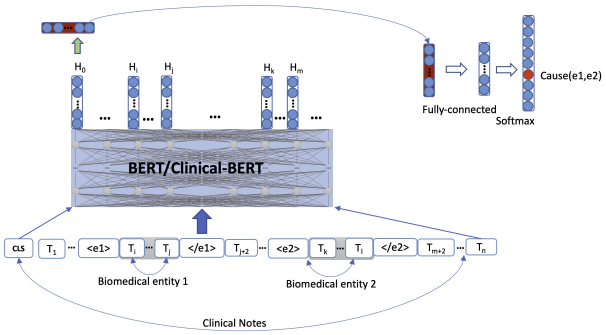


Figure 2: BERT/Clinical-BERT: FFN

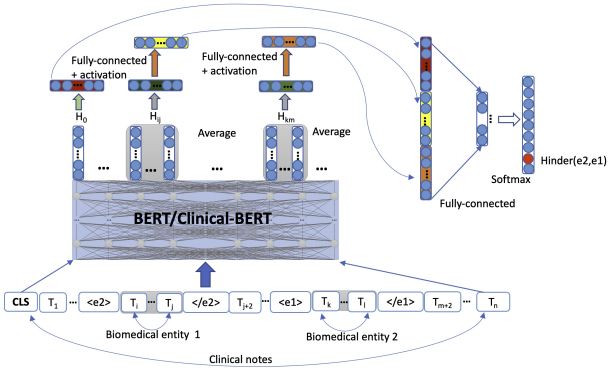


Figure 3: BERT/Clinical-BERT: FFN with entity context

mains the same as discussed earlier, the current architecture also enriches the sentence context with both the entities' context as formulated in equations (7)–(10). The special tokens around the entities ($<e1>$, $</e1>$, $<e2>$, and $</e2>$) are used to identify the tokens related to the individual entities which are then used to obtain the averaged context vector for each entity. These are then concatenated with the overall sentence context and are fed to a fully connected feed-forward network to predict the type of causal interaction expressed in the text.

Similar to our previous discussion, in addition to the BERT-BASE encoder, a pre-trained Clinical-BERT encoder was also used which resulted in the highest evaluation metrics.

	Test	Val	Train
BERT+FFN	0.23	0.25	0.29
Clinical-BERT+FFN	0.27	**0.31**	0.34
BERT+entity context+FFN	0.54	0.27	0.56
Clinical-BERT+entity context+FFN	**0.56**	0.30	**0.70**

Table 4: Macro F1 score on test, val and train dataset

4.3 Results and analysis

We trained all our models on a varied set of hyperparameters and chose the best model from training epochs based on the maximum F1 score on the validation set. For BERT+FFN model, we achieved the best scores with a batch size of 128 and a learning rate of 5e-5. The other three models achieved reported scores with a batch size of 32 and a learning rate of 1e-3. All the models were trained until convergence with the early stopping of 7 epochs with no decrease in validation loss. We used AdamW optimizer with cross-entropy loss for all models.

Table 4 shows performance measures of various models on train/val/test set. Using only the BERT-BASE encoder for the relation identification doesn't yield high scores but concatenating entity context to the BERT's encoded sentence output resulted in significant improvement. Using Clinical-BERT as base encoder resulted in additional improvements, and combining entity contexts with Clinical-BERT as base encoder resulted in the highest F1 score. While Clinical BERT was trained on the MIMIC dataset and might have seen input sequences in the test dataset, it has not seen newly defined causal classes for those sequences.

5 Conclusion

In this work, we proposed annotation guidelines to capture the types and direction of causal associations, annotated a dataset of 2714 examples from de-identified clinical notes and built models to provide a baseline score for our dataset.

Even with the inherent complexities in clinical text data, following the meticulously defined annotation guidelines, we achieved a high inter-annotator agreement, i.e., Fleiss' kappa (κ) score of 0.72. Building various network architectures on top of language models, we achieved a macro F-1 score of 0.56.

An end-to-end NLP pipeline built with models for patients' data de-identification, biomedical entity extraction, and causal relations identification between various biomedical entities will be instrumental in narrative understanding from clinical notes. In the future, we are planning to extend our annotation guidelines to jointly annotate temporal and causal relations to capture the ordering of various causal interactions between biomedical entities over time.

References

Steven Bethard and James H. Martin. 2008. Learning semantic links from a corpus of parallel temporal and causal relations. In *ACL*.

Steven Bethard, Guergana Savova, Martha Palmer, and James Pustejovsky. 2017. SemEval-2017 task 12: Clinical TempEval. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*.

Jesse Dunietz, Lori S. Levin, and J. Carbonell. 2015. Annotating causal language using corpus lexicography of constructions. In *LAW@NAACL-HLT*.

Jesse Dunietz, Lori S. Levin, and J. Carbonell. 2017. The because corpus 2.0: Annotating causality and overlapping relations. In *LAW@ACL*.

Lei Gao, Prafulla Kumar Choubey, and Ruihong Huang. 2019. Modeling document-level causal structures for event causal relation identification. In *NAACL*.

R. Girju, Preslav Nakov, Vivi Nastase, Stan Szpakowicz, Peter D. Turney, and Deniz Yuret. 2007. Semeval-2007 task 04: Classification of semantic relations between nominals. In *SemEval@ACL*.

Harsha Gurulingappa, A. Rajput, A. Roberts, J. Fluck, M. Hofmann-Apitius, and L. Toldo. 2012. Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. *Journal of biomedical informatics*, 45 5:885–92.

O. Hassanzadeh, D. Bhattacharjya, M. Feblowitz, Kavitha Srinivas, M. Perrone, Shirin Sohrabi, and Michael Katz. 2019. Answering binary causal questions through large-scale text mining: An evaluation using cause-effect pairs from human experts. In *IJCAI*.

Iris Hendrickx, S. Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, M. Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. 2009. Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *SemEval@ACL*.

Sam Henry, K. Buchan, Michele Filannino, A. Stubbs, and Özlem Uzuner. 2020. 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. *Journal of the American Medical Informatics Association : JAMIA*.

Rei Ikuta, Will Styler, Mariah Hamang, Timothy J. O'Gorman, and Martha Palmer. 2014. Challenges of adding causation to richer event descriptions. In *EVENTS@ACL*.

Alistair E. W. Johnson, T. Pollard, Lu Shen, Li wei H. Lehman, M. Feng, M. Ghassemi, Benjamin Moody, Peter Szolovits, L. Celi, and R. Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3.

Vivek Khetan, Roshni Ramnani, Mayuresh Anand, Subhashis Sengupta, and Andrew E. Fano. 2022. Causal bert: Language models for causality detection between events expressed in text. In *Intelligent Computing*, pages 965–980, Cham. Springer International Publishing.

Feifan Liu, Abhyuday Jagannatha, and Hong Yu. 2019. Towards drug safety surveillance and pharmacovigilance: current progress in detecting medication and adverse drug events from electronic health records. *Drug safety*, 42(1):95–97.

C. Mihaila, Tomoko Ohta, Sampo Pyysalo, and S. Ananiadou. 2012. Biocause: Annotating and analysing causality in the biomedical domain. *BMC Bioinformatics*, 14:2 – 2.

Paramita Mirza and Sara Tonelli. 2014. An analysis of causality between events and its relation to temporal information. In *COLING*.

Paramita Mirza and Sara Tonelli. 2016. CATENA: CAusal and TEmporal relation extraction from NAtural language texts. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 64–75, Osaka, Japan. The COLING 2016 Organizing Committee.

N. Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, P. Kohli, and James F. Allen. 2016a. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *NAACL*.

N. Mostafazadeh, Alyson Grealish, Nathanael Chambers, James F. Allen, and Lucy Vanderwende. 2016b. Caters: Causal and temporal relation scheme for semantic annotation of event structures. In *EVENTS@HLT-NAACL*.

Timothy J. O'Gorman, Kristin Wright-Bettner, and Martha Palmer. 2016. Richer event description: Integrating event coreference with temporal, causal and bridging annotation.

J. Pustejovsky, J. Castaño, R. Ingria, R. Saurí, R. Gaizauskas, A. Setzer, G. Katz, and Dragomir R. Radev. 2003. Timeml: Robust specification of event and temporal expressions in text. In *New Directions in Question Answering*.

Bhanu Pratap Singh Rawat, Abhyuday Jagannatha, Feifan Liu, and Hong Yu. 2020. Inferring adr causality by predicting the naranjo score from clinical notes. In *AMIA Annual Symposium Proceedings*, volume 2020, page 1041. American Medical Informatics Association.

Norman Swartz. 2014. The concepts of necessary conditions and sufficient conditions.

Leonard Talmy. 1988. Force dynamics in language and cognition. *Cognitive Science*, 12(1):49–100.

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

P. Wolff. 2007. Representing causation. *Journal of experimental psychology. General*, 136 1:82–111.