MIMICause: Defining, identifying and predicting types of causal relationships between biomedical concepts from clinical notes

Vivek Khetan∗†♠  Md Imbesat Hassan Rizvi†♣  Jessica Huber‡♥  Paige Bartusiak‡♦  Bogdan Sacaleanu♠  Andrew Fano♠

∗Accenture Labs, SF  †Indian Institute of Science  ♥Arizona State University  ♦Duke University

{vivek.a.khetan, bogdan.e.sacaleanu, andrew.e.fano}@accenture.com  mdrizvi@iisc.ac.in, jhuber12@asu.edu, paige.bartusiak@cs.duke.edu

Abstract

Understanding of causal narratives communicated in clinical notes can help make strides towards personalized healthcare. In this work, MIMICause, 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 relation achieved a macro F1 score of 0.56 on 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 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.

With the contemporary advancement in natural language processing (NLP), we have witnessed 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 (O’Gorman et al., 2016; Mirza and Tonelli, 2014; Khetan et al., 2022), causal associations between the events expressed in natural language text. A large amount of work in causality representation focuses on decomposition of causal interactions communicated in text data, whereas, the work in causality extraction mostly focuses on causal event extraction or identification of causal relation between already identified events.

There are various proposed representations for causality depending upon the domain and the task. 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; Henry 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.
Table 1: MIMICause sample identification and annotations

| Snippet from Clinical Notes | Text with identified entities | Causal relation identification based on the proposed MIMICause guidelines |
|-----------------------------|-------------------------------|---------------------------------------------------------------------|
| **The patient is a 52 year old woman with past medical history significant for chronic pain on narcotics and benzodiazepines, malabsorption syndrome due to complications of gastric bypass surgery, and severe osteoporosis.** | The patient is a 52 year old woman with past medical history significant for chronic pain on narcotics and benzodiazepines, **malabsorption syndrome** due to complications of gastric bypass surgery, and severe osteoporosis. | complications of gastric bypass surgery enable malabsorption syndrome |
| **The patient is a 52 year old woman with past medical history significant for chronic pain on narcotics and benzodiazepines, malabsorption syndrome due to complications of gastric bypass surgery, and severe osteoporosis.** | The patient is a 52 year old woman with past medical history significant for chronic pain on narcotics and benzodiazepines, **malabsorption syndrome** due to complications of gastric bypass surgery, and severe osteoporosis. | severe osteoporosis enable malabsorption syndrome |
| **The patient is a 52 year old woman with past medical history significant for chronic pain on narcotics and benzodiazepines, malabsorption syndrome due to complications of gastric bypass surgery, and severe osteoporosis.** | The patient is a 52 year old woman with past medical history significant for chronic pain on narcotics and benzodiazepines, **malabsorption syndrome** due to complications of gastric bypass surgery, and severe osteoporosis. | complications of gastric bypass surgery, and severe osteoporosis cause malabsorption syndrome |
| **The patient is a 52 year old woman with past medical history significant for chronic pain on narcotics and benzodiazepines, malabsorption syndrome due to complications of gastric bypass surgery, and severe osteoporosis.** | The patient is a 52 year old woman with past medical history significant for chronic pain on narcotics and benzodiazepines, **malabsorption syndrome** due to complications of gastric bypass surgery, and severe osteoporosis. | complications of gastric bypass surgery, and severe osteoporosis cause malabsorption syndrome |
| **The patient is a 52 year old woman with past medical history significant for chronic pain on narcotics and benzodiazepines, malabsorption syndrome due to complications of gastric bypass surgery, and severe osteoporosis.** | The patient is a 52 year old woman with past medical history significant for chronic pain on narcotics and benzodiazepines, **malabsorption syndrome** due to complications of gastric bypass surgery, and severe osteoporosis. | complications of gastric bypass surgery, and severe osteoporosis cause malabsorption syndrome |

Table 1 shows a snippet of a clinical note extracted from the n2c2 dataset (Henry et al., 2020), different sets of annotated biomedical entities and the causal relation based on the proposed MIMICause 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 for the task of causal relation prediction.

2 Related Works

Identification of semantic causal relationships from text data is helpful in extraction of causal event chains and can also be leveraged to reason and explain the state of events in a given narrative. Although there are various works defining causation in linguistics, philosophy and psychology, there still isn’t an agreed upon definition for decomposition and representation of causation. While researchers from psychology and philosophy work on causality with the goal of understanding the world; most linguistic researchers focus on representing causality to understand interactions between events.

Talmy (1988) proposed force-dynamics to decompose the causal interaction between events as “letting”, “helping”, “hinder” etc. Wolff (2007) built upon force-dynamics by incorporating the theory of causal verbs and proposed 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 an expression of causation instead of real-world causation or its philosophical meaning. BECauSE 1.0 (Dunietz et al., 2015) consists of a cause span, an effect span, and a causal connective span; but this annotation schema struggles with the distinction between the cause event and the causing agent. For example, in the sentence “I prevented a fire”, based on BECauSE 1.0, instead of an action, the actor “I” is identified as cause. Later, BECauSE 2.0 (Dunietz et al., 2017) proposed the use of means argument to capture cause event and causing agent without struggle. They also proposed three different types of causation (Consequence, Motivation, and Purpose) and also annotated overlapping semantic relations between cause and effect along with causal relations. In con-
trast, our work focuses on identification of types of causal association 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 MIMICause 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 observatory 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 directionality of causal interaction with simple temporal relations (e.g., Before, After, No-Rel) but did not focus on the type 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 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 has temporal as well as causal 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 description with events’ context to predict causal relationship. 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 narratives.

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 bioevent corpora from biomedical scientific articles to capture biomedical causality. Instead of identifying the type (and direction) of causal interaction 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 causal events.

Our work proposes comprehensive guidelines to represent the type 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.

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 presented in Table 2. 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).

However, the provided relationships in the n2c2 dataset are simply defined by identified concepts linked with related medications and hold no semantic meaning. To create the MIMICause dataset, we extracted examples from each entity-pair avail-

---

1We used https://spacy.io/ library with “en_core_web_sm” language model.
Table 2: Examples of Bio-medical concepts/entities in the 2018 n2c2 shared task dataset.

| 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. |

able 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 relationships/associations exist when one or more entities affect another set of entities. The driving concept can be a single entity such as a drug / procedure / therapy or a composite entity such as several drugs / procedures / therapies together.

3.1.1 Direction of causal association

The direction of causal association between entities is captured by the order of entity tags ((e₁, e₂) or (e₂, e₁)) in the defined causal relationships. Either entity can be referred to as e₁ or e₂. 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 (e₁, e₂) while example (2) is annotated with (e₂, e₁).

3.2 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.

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 (e₁, e₂) while example (2) is annotated with (e₂, e₁).

3.2.1 (Un)-certainty of causal association

Establishing real-world causality or the task of causal inference is not in the scope of our current work. Based on our proposed guidelines, a causal association either expressed as speculation or with certainty is annotated similarly.

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**]jof 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.

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.

3.2.2 Types of causal association

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. The patient is a 52 year old woman with past medical history significant for chronic pain on narcotics and benzodiazepines, \(<e_2>\text{malabsorption syndrome}</e_2>\) due to \(<e_1>\text{complications of gastric bypass surgery, and severe osteoporosis}</e_1>\).

In example (5), “malabsorption syndrome” occurred due to two factors viz. “complications of gastric bypass surgery” and “severe osteoporosis”. The entity span covers both of them and they are considered together as a composite entity leading to “malabsorption syndrome”. 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 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 entity under consideration is one of the contributing factors.

6. The patient is a 52 year old woman with past medical history significant for chronic pain on narcotics and benzodiazepines, \(<e_2>\text{malabsorption syndrome}</e_2>\) due to complications of gastric bypass surgery, and \(<e_1>\text{severe osteoporosis}</e_1>\).

Example (6) is same as the example (5) except for the entities in considerations. Both the factors viz. “complications of gastric bypass surgery” and “severe osteoporosis” are contributing to the “malabsorption syndrome”.

Since the example is considering only “severe osteoporosis” which is a contributing factor in part, hence it is annotated as **Enable**\((e_1, e_2)\).

With “Enable” relation type, it can easily be noted that addressing only the “complications of gastric bypass surgery” or “severe osteoporosis” will not lead to the treatment of “malabsorption syndrome”. Labelling these samples as “Cause” would have suppressed this detail and the actions taken.

• **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. with chest and \(<e_1>\text{abdominal pain}</e_1>\) and odynophagia who was found to have circumferential mural thrombus in the supra-renal aorta on cross sectional imaging. At that time the aortic pathology was though to be chronic. Ultimately, her pain resolved with the initiation of a \(<e_2>\text{PPI and GI cocktail}</e_2>\), and was discharged home after a 3 day hospital stay.

In example (7), “PPI” and “GI cocktail” are the two different entities used in conjunction to resolve the “abdominal 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 entity under consideration is one of the contributing factors.

Similar to “Prevent”, this label also includes the scenario of hindering a disease or condi-
... tion from occurring as well as curing a disease or condition if it has occurred.

8. with chest and <e1>abdominal pain</e1> and odynophagia who was found to have circumferential mural thrombus in the supra-renal aorta on cross sectional imaging. At that time the aortic pathology was though to be chronic. Ultimately, her pain resolved with the initiation of a <e2>PPI</e2> and GI cocktail, and was discharged home after a 3 day hospital stay.

Example (8) is same as the example (7) except for the entities in considerations. Both the entities i.e. “PPI” and “GI cocktail” are contributing to the resolution of “abdominal pain”. Since the example is considering only “PPI” individually as a contributing factor in part, it is annotated as Hinder(e2, e1).

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 “Other” class to annotate examples with non-causal interaction between biomedical entities. Examples of “Other” class can either have no relationship between biomedical entities of interest or some other semantic relationship that’s not causal. Being non-causal in nature, 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, marked entities without direct causal association (an entity leading to a condition and that condition affecting other 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.

11. Atenolol was held given patient was still on <e2>amiodarone</e2> and <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 3.

### 3.3 Inter-annotator agreement

It’s difficult to comprehend narrative expressed in clinical notes due to need of domain knowledge, short hand used by the doctors, use of abbreviations (Table 4), context spread over many sentences as well as the explicit and implicit nature of communication.

Given the nature of our base data (MIMIC-III discharge summaries) and critical importance of our task (causal relations between biomedical entities); three authors of this paper (all with fluency...
| 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 |

in English and computer science background) annotated the dataset. They followed the provided guidelines, referred to sources such as websites of Centers for Disease Control and Prevention (CDC\(^2\)), National Institute of Health (NIH\(^3\)) and WebMD\(^4\) 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 (\(k\)) score of 0.717, which is indicative of substantial agreement and the quality of our annotation guidelines.

We did majority voting over the three available annotations to obtain 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 discussing it with the other three annotators.

A direct comparison of our IAA score with other work is not possible due to differences in the number of annotators, annotation labels, guidelines, reported metrics etc. for different datasets. However, for reference, we have discussed IAA scores reported for the task of semantic link annotations, particularly those where \(k\) 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 \(k\) score achieved was 0.51 among four annotators. Similarly, Bethard et al. (2008) reported a \(k\) 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. i.e. To compare IAA scores on clinical datasets, temporal and causal relation classification models using two different language models\(^5\) 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 \(H_0 \in \mathbb{R}^d\) where \(d = 768\) is the dimension of the encoded outputs from BERT-BASE / Clinical-BERT. The formulations of the layers of

\[ X = [x_1, x_2, \ldots, x_{n-1}, x_n] \]
\[ e_1 = X[i : j] = [x_i, x_{i+1}, \ldots, x_j] \]
\[ e_2 = X[k : l] = [x_k, x_{k+1}, \ldots, x_l] \]

where \(n\) is the sequence length, \(i, j, k \text{ and } l \in [1..n]\), \(i \leq j\) and \(k \leq l\) i.e. entities are sub-sequences of continuous span within the text \(X\). Additionally, \(j < k \text{ or } l < i\) 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\(^5\) 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 \(H_0 \in \mathbb{R}^d\) where \(d = 768\) is the dimension of the encoded outputs from BERT-BASE / Clinical-BERT. The formulations of the layers of

\[ n \]

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

4 Problem definition and Experiments

We defined our task of causality understanding as 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 \(D = \{(X, e_1, e_2, y)_m\}_{m=1}^{N}\), where \(N\) is the total number of samples in the dataset. The text and entities are mathematically denoted as:

\[ X = [x_1, x_2, \ldots, x_{n-1}, x_n] \]
\[ e_1 = X[i : j] = [x_i, x_{i+1}, \ldots, x_j] \]
\[ e_2 = X[k : l] = [x_k, x_{k+1}, \ldots, x_l] \]
the classifier head are given by:

\[ K_1 = \text{dropout}(ReLU(W_1 H_0 + b_1)) \quad (4) \]
\[ K_2 = W_2 K_1 + b_2 \quad (5) \]
\[ p = \text{softmax}(K_2) \quad (6) \]

where \( W_1 \in \mathbb{R}^{d' \times d}, \) \( W_2 \in \mathbb{R}^{L \times d'}, \) \( d' \) 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:

\[ H_{e1} = \frac{1}{j - i + 1} \sum_{t=i}^{j} H_t \quad (7) \]
\[ H_{e2} = \frac{1}{l - k + 1} \sum_{t=k}^{l} H_t \quad (8) \]
\[ H' = \text{concat}(H_0, H_{e1}, H_{e2}) \quad (9) \]
\[ H_0 = \text{dropout}(ReLU(W_0 H' + b_0)) \quad (10) \]

where \( i, j, k \) and \( l \) are start and end indices of the entities, \( H_t \in \mathbb{R}^d, H' \in \mathbb{R}^{d'}, W_0 \in \mathbb{R}^{d' \times 3d} \) 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 (1) and (2) 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 1 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 Clinical-BERT encoder to obtain 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. Replacing BERT-BASE with Clinical-BERT showed 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 2. While the input mechanism with special tokens, encoding and classifier head remains the same as discussed earlier, the current architecture also enriches the sentence context with both the entity’s 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 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.
Table 5: Macro F1 score on test, val and train dataset

| Model                                | 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  |

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 best scores with batch-size of 128 and learning rate of 5e-5. The other three models achieved reported scores with a training batch-size of 32 and a learning rate of 1e-3. All the models were trained until convergence with early stopping of 7 epochs with no decrease in validation loss. We used AdamW optimizer with cross entropy loss for all models.

Table 5 shows performance measures of our 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 types and direction of causal association, annotated a dataset of 2714 examples from de-identified clinical notes and built models to provide baseline score for our dataset.

Even with inherent complexities in clinical text data, following the meticulously defined annotation guidelines, we were able to achieve a high inter annotator agreement, i.e., Fleiss’ kappa ($\kappa$) score of 0.72. Building various network architectures on top of language models, we achieved an accuracy of 0.65 and macro F-1 score of 0.56.

In the future, we are planning to extend our annotation guidelines to jointly represent temporal and causal associations in clinical notes. An end-to-end pipeline built with models for patients’ data identification, biomedical entity extraction, and detailed causal and temporal representation between them will help us understand the ordering of various causal associations and enhance our capability for understanding clinical narrative.

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 Savoya, 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.

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. Bhattacharjyia, 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 O’Seaghdha, Sebastian Padó, M. Pennachiotto, 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.

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.

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.

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, Pierrick 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.