Medical Relation Extraction with Bio-Clinical BERT and Longformer

W266 Summer 2021

Simon Li

University of California, Berkeley

Abstract

In unstructured clinical documents, the relationships between medical events and temporal expressions (TIMEX) need to be classified to build a timeline of a patient's medical record. This work proposes a BERT-based architecture that uses [CLS] + entity masking to improve the classification of entity pairs in clinical documents. In addition, the Longformer was used to capture document level context of up to 1500 tokens. The architecture in combination with Bio-Clinical BERT and Longformer, produced comparable macro F1-scores of 0.62-0.65 to the baseline model. However, the Longformer with [CLS] + entity masking showed more robustness to the imbalance of data when entities are far apart in a document.

Keywords: Relation Extraction, Deep Learning, Bio-Clinical BERT, Longformer

Introduction

The medical history of a patient is important in determining their diagnosis but are often buried in unstructured clinical documents. As more hospitals begin to capture notes in electronic health records (EHRs), this provides an opportunity to use machine learning to extract medical history out of unstructured text and provide physicians with data to make better decisions.

One challenge inherent in analyzing EHR, is the use of relative expressions that may occur across a document. Consider an example in Figure 1.

"He had [R] **an magnetic resonance imaging** [R] performed on October 18 , 1996 . This showed the unchanged left temporal parenchymal intraparenchymal hematoma and [L] **the old embolic right frontal infarct** [L] , but there were no new lesions and no evidence of his susceptibility studies to suggest old bleeds which might have been consistent with amylase angiopathy . "

Figure 1: Example of entity pairs in clinical text. The tokens [L] and [R] help in locating the entities and are added post text.

Here the two entities are shown in Figure 2, and occur across two sentences.

entity 1: " the old embolic right frontal infarct"
entity 2: " an magnetic resonance imaging"

Figure 2: Entities from the example in Figure 1

To build a timeline, the relationship between the entity pairs are classified to form a directed graph. Typical relationships are AFTER, OVERLAP, and BEFORE. In this specific example, entity 1 happened before entity 2.

[The patient had] "the old embolic right frontal infarct" **BEFORE** " an magnetic resonance imaging" [was performed on him/her]

Figure 3. Example of BEFORE class between two entity pairs

A second challenge in EHR is that data is often limited due to patient privacy. Small datasets thus make it challenging for machine learning models to learn and generalize.

Much of the recent work in relation extraction of EHRs have been rule-based or within-sentence, which leads to limitations [5]. Rule-based approaches require experts to write, which are often complex and domain specific to cover edge cases. Within-sentence approaches are limited as they do not cover entities across sentences and do not capture context from the document level.

To address the limitations with the aforementioned approaches, this body of work is novel in two ways:

- Proposition of a BERT-based architecture that applies [CLS] + entity masking to classify relationships between entities without rules and across sentences
- Application of the Longformer for relation extraction of documents longer than 512 tokens

Related Work

Several machine learning approaches to temporal relation extraction in clinical documents have contributed to this space.

Lee H. *et al* applied SVMs to direct temporal relationships between event/TIMEX pairs only, as opposed to event/event and TIMEX/TIMEX pairs. The SVM system produced an F1-score of 0.67 on the 2012 i2b2 temporal relation dataset [8]. The team showed shortcomings in long sentences and proposed that classification of each subset of temporal relations could be considered a specific task.

Chen T. *et al.* applied 1D-CNN to BERT embeddings and was able to achieve an F1-score of 0.71 on the 2012 i2b2 temporal relation dataset [1]. The team proposed their method as a general approach that did not require developing parse-trees of the text.

Han X. *et al* applied entity masking, entity types and bi-linear layers to BERT embeddings to classify entities and achieved an F1-score of 0.57 on the DocRED dataset [2]. The team suggested that entity masks allow the model to more accurately obtain each entity. Additionally, their work suggested combining the entity embeddings with the entity mask token produced highest performance.

The majority of the state-of-the-art models still have scores below 0.80, which is indicative of the challenge of temporal relation extraction in clinical documents. This work attempts to contribute to this field with the application of [CLS] + entity masking and Longformer.

Data

The 2012 i2b2 dataset is used in this body of work. The dataset includes 310 discharge summaries from the Partners Healthcare and Beth Israel Deaconess Medical Center [3]. Of the 310 discharge summaries, 190 were reserved for training and 120 were used for testing. Each discharge summary contains multiple entity pairs. On average, a discharge summary contained 86 event entities and 12 TIMEX entities.

Eight annotators labeled the relationship between entities to provide golden labels. The three labels are: `(0) AFTER, (1) OVERLAP, and (2) BEFORE.` The classes are moderately unbalanced as shown in Figure 4, where only 10% of the data has a class of `AFTER`.

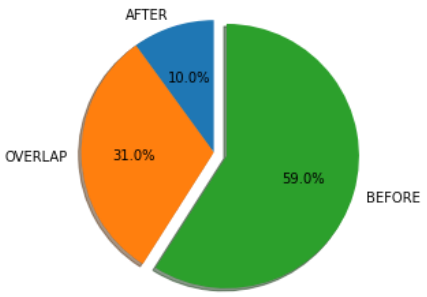
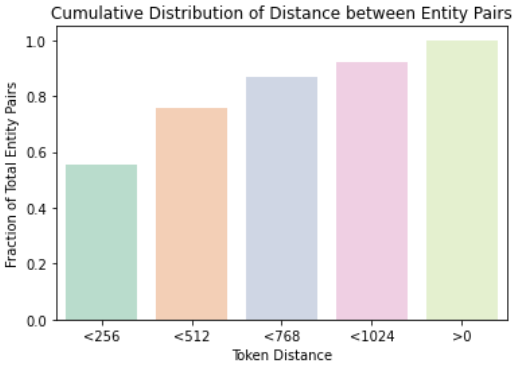


Figure 4: Distribution of relations

The distribution for the number of tokens between entity pairs (distance) are shown in Figure 5. We can see that roughly 25% of entity pairs are more than 512 tokens apart. Note that if we truncated the document to 512 tokens starting from the first token in the document, roughly 35% of at least one entity would be excluded.



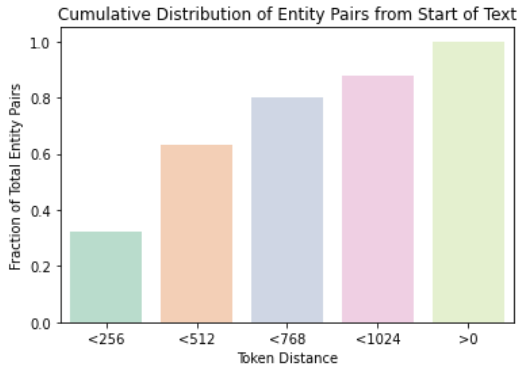


Figure 5: Distribution of number of tokens between entity pairs and from the start of the document.

Additional imbalance is also present in the distance between entity pairs across classes. In Figure 6, we see that `OVERLAP` has the most examples of token distances that are less than 32 tokens. Whereas for distances longer than 64 tokens, the class `BEFORE` is the most dominant. Both imbalances in class and distance between entity pairs challenge a model's ability to learn as they steer a model towards the dominant class.

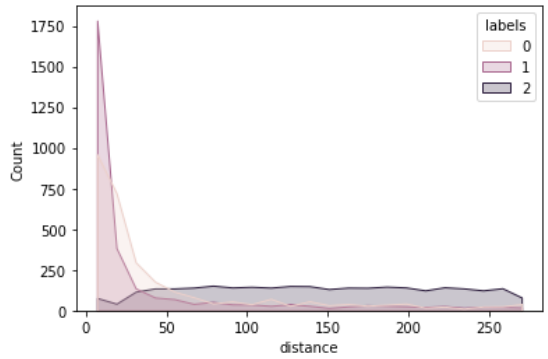


Figure 6: Imbalance among distance between entity pairs.
(0) AFTER (1) OVERLAP (2) BEFORE

Methods

Data Preprocessing

A subset of the data was used in order to manage class imbalance and compare the effect of the proposed architecture and longer context.

To manage the class imbalance, stratified random sampling (with replacement) was applied to oversample the number of examples on the minority class `AFTER`. To see the effect of context length on model performance, we only considered examples where the distance between entity 1 and entity 2 was less than 512 tokens. However, the context length varied for each model: 512 tokens for Bio-Clinical BERT and up to 1500 tokens for the Longformer.

As a result of the data processing, roughly 9000 entity pairs (3000 per class) were used for training, however the distribution of distance was left unbalanced. The model's performance was measured on the class and distance unbalanced test dataset.

Entity Tokens

Special tokens called entity tokens, [L] and [R], were added to the documents. The entity tokens allow the model to mask non-entity embeddings and direct entity embeddings towards the classifier. In future work, entity tokens may also be used for masked language modeling where entities can be masked and the model learns to predict the entities [4].

Metrics

The metrics used to compare models are precision, recall, and F1-score for each class given that all classes are equally important in building a timeline . In other research with this dataset, metrics were measured on temporal closure (i.e. whether a timeline was correctly built or not based on correct classification of each entity pair), rather than classes. However, to evaluate the effect of the architecture, only entities within 512 tokens were considered. Entity pairs greater than 512 tokens were excluded, and thus a complete timeline cannot be produced for each document. Therefore the metrics were compared for each class instead.

Models

BERT-based models

Two pretrained models from Hugging Face were used in this work: Bio-Clinical BERT and Longformer. Bio-Clinical BERT was chosen because it was pre-trained on over 880M words from MIMC-III database of critical care clinical notes [6]. The Longformer was chosen to capture long document level context up to 1500 tokens. The Longformer is pre-trained on documents of 4098 tokens in length and is to capture long document context through local attention windows and global attention [7].

Baseline Model Architecture

A baseline was developed using Bio-Clinical BERT and with the standard architecture. Documents and entity pairs were separated by the `[SEP]` token and

fed to BERT. The embeddings from BERT then went through a dropout layer and a softmax classification layer.

```
[CLS] Document [SEP] Entity 1 [SEP]
Entity 2 [SEP]
```

Figure 7: Architecture of Baseline Model

Proposed Architecture

The proposed architecture applies an entity masking to direct embeddings from the entity pair towards the softmax classification layer as shown in Figure 8. The `[CLS]` token is also concatenated with the entity embeddings to provide additional document level context. It was hypothesized that this architecture would allow the model to learn the context of the document through BERT and then focus on the relationship between the entity pairs.

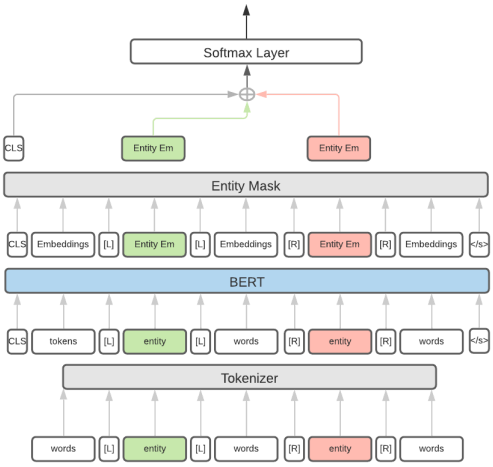


Figure 8: Architecture with Entity Masking and [CLS] token concatenation

Hyperparameters

Most default parameters were used for Bio-Clinical BERT and Longformer. For the Longformer, the window size of 512 tokens was used as experiments with varying window size did not benefit the model (see Figure 13 in Appendix). Global attention was also enabled for the `[CLS]` token and the entity pairs. In both models, a dropout of 0.5 was applied prior to the classification layer.

In the training loop, Adam with a learning rate of 1e-5 was used for the optimizer. Early-stopping was applied as the models tended to overfit beyond 4-5 epochs.

Results and Discussion

In this section, the three models: baseline Bio-Clinical BERT, Bio+Clinical BERT w/ `[CLS]` + entity masking, and Longformer w/ `[CLS]` + entity masking are compared.

Classification Report

The table below shows the metrics on the test set of each class for the three models. The overall performance is similar amongst all three models. The models also have precisions greater than 0.3, which indicates the models do not simply predict the proportion of each class from the training set. However we do see a shift in recall from `BEFORE` to `OVERLAP` in the Longformer, which is explored in the next section.

	Classes	Precision	Recall	F1	macro-F1
Baseline	(0) AFTER	0.350	0.455	0.396	0.649
	(1) OVERLAP	0.718	0.749	0.733	
	(2) BEFORE	0.857	0.784	0.819	
Bio-Clinical BERT w/ [CLS] + entity mask	(0) AFTER	0.345	0.490	0.405	0.650
	(1) OVERLAP	0.673	0.772	0.719	
	(2) BEFORE	0.916	0.752	0.826	
Longformer w/ [CLS] + entity mask	(0) AFTER	0.398	0.426	0.412	0.626
	(1) OVERLAP	0.591	0842	0.695	
	(2) BEFORE	0.936	0.654	0.770	

Figure 9. Performance of each model for each class

Confusion Matrix per Distance

In Figure 10 we explore the effect of distance imbalance with confusion matrices for two distances: (left) 32-64 tokens and (right) 256-512 tokens. Comparing the results from the two distances, show that Longformer is able to learn long distance context despite distance imbalance.

For distances of 32-64 tokens, we can see that Longformer performs slightly better in (1) `OVERLAP` despite (0) `AFTER` containing the majority of the examples. For distances 256-512 tokens, the baseline model is heavily affected by distance imbalance and predicts mostly (2) `BEFORE`. While with the proposed architecture, Bio-Clinical BERT benefits from entity masking and reduces the effect of distance imbalance slightly. The Longformer resists distance imbalance further and leads to better recall in (1) `OVERLAP`.

These findings suggest that the proposed architecture in addition to long document context provided by Longformer may help in learning long distance relationships despite some distance imbalance. The Longformer was able to relate `OVERLAP` for long distances, which is intuitively difficult. However, the

Longformer still has a hard time learning AFTER even though it has more examples than OVERLAP for distances of 256-512 tokens.

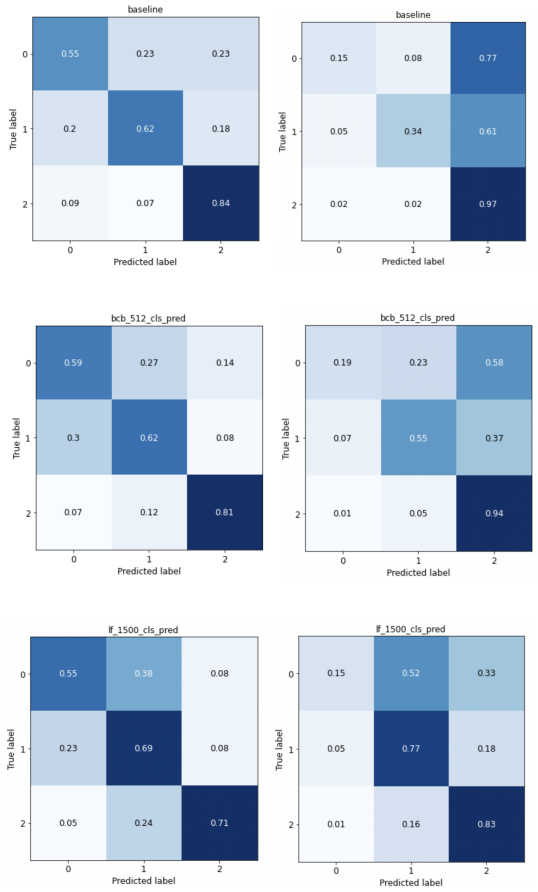


Figure 10. Confusion matrices for each two distances: (left) 32-64 tokens and (right) 256-512 tokens. From top to bottom:baseline, Bio-Clinical BERT w/ [CLS] + entity mask, and Longformer w/ [CLS] + entity mask.

Examples of Correct and Incorrect Prediction

Here we consider a few abbreviated examples in Figure 11 where the Longformer w/ [CLS] + entity masking had difficulty predicting AFTER for distances of 256-512.

```
-----------------------------------
Example: 253_SECTIME13
(true, pred) = (AFTER, OVERLAP)
distance = 302
-----------------------------------
Admission Date : [R] 2017-09-22
[R]....On the newborn nursery the
mother called the nurse for concern of
baby turning dusky while feeding
...The Neonatal Intensive Care Unit ws
notified and the baby was transferred
to [L] the Newborn Intensive Care Unit
[L] for further evaluation.
-----------------------------------
```

```
-----------------------------------
Example: 222_SECTIME7
(true, pred) = (AFTER, BEFORE)
distance = 333
-----------------------------------
ADMISSION: [R] 2012-10-31 [R]

MEDICINE History of Present Illness :
.....For 2 days PTA , he has expressed
a wish to kill himself ( no clear plan
, but fixing things around the house "
so things will be ready when I \'m
gone "). In Monica , pt received [L]
charcoal [L] , 500 mg IV levofloxacin
, 500 mg IV metronidazole for presumed
aspiration pna .
-----------------------------------


-----------------------------------
Example:  442_TL49
(true, pred) = (AFTER, AFTER)
distance = 269
-----------------------------------
Jadiara Harrison is twin #1 [R] born
[R] to a 17 year-old primiparous
mother....She was noted to have apnea
of prematurity , treated with [L]
caffeine [L] until 2017-06-04 .
-----------------------------------
```

Figure 11: Examples for AFTER for distances 256-512

In example 253_SECTIME13, the challenge is understanding when the transition from OVERLAP and AFTER occurs. It is unclear whether the transfer of the baby occurred on the day of admission or subsequent days. This ambiguity between OVERLAP and AFTER could arguably affect human annotation as well.

In example 222_SECTIME14, the header MEDICINE History of Present Illness is the indicator that entity [L] is a past event. Otherwise, it would be difficult to understand whether Monica was a location the patient had previously visited or the hospital the patient is currently in. The model is unable to learn document structure despite the use of global attention on the entities.

In example 442_TL49, the words born and until 2017-06-04 are good indicators that entity [L] happened AFTER entity [R]. Here the model is able to predict correctly given indicating words.

Prediction on Handcrafted Examples

To further explore the Longformer's ability to differentiate classes in the presence of distance imbalance, we study the model predictions on handcrafted examples of 222_SECTIME7. In Figure 12, two variants of the example are present: (top)

original text, (middle) flipping the [R] and [L] entities thus switching AFTER to BEFORE and (c) adding repeated text in between (which adds no additional information) to extend the distance between entities while AFTER to BEFORE are switched.

```
-----------------------------------
original text:
(true, pred) = (BEFORE, BEFORE)
distance = 57
-----------------------------------
"Admission Date : [R] 2012-03-23 [R]
Discharge Date : 2012-03-26 Service :
MEDICINE History of Present Illness :
39 year old male w/ h/o low back pain
on chronic narcotics presents after
being found [L] unresponsive [L] at
home. His daughter awoke him at 7 a.m.
, reports he said he felt cold and
shivery ,vomited several times , then
drove her to school ."
-----------------------------------

-----------------------------------
flip [L] and [R]:
(true, pred) = (AFTER, AFTER)
distance = 57
-----------------------------------
"Admission Date : [L] 2012-03-23 [L]
Discharge Date : 2012-03-26 Service :
MEDICINE History of Present Illness :
39 year old male w/ h/o low back pain
on chronic narcotics presents after
being found [R] unresponsive [R] at
home. His daughter awoke him at 7 a.m.
, reports he said he felt cold and
shivery ,vomited several times , then
drove her to school ."
-----------------------------------

-----------------------------------
flip [L] and [R] + filler text:
(true, pred) = (AFTER, AFTER)
distance = 158
-----------------------------------
"Admission Date : [L] 2012-03-23 [L]
Discharge Date : 2012-03-26 Service :
MEDICINE History of Present Illness :
39 year old male w/ h/o low back pain
on chronic narcotics presents after
h/o low back pain on chronic narcotics
presents after h/o low back pain on
chronic narcotics presents after h/o
low back pain on chronic narcotics
presents after h/o low back pain on
chronic narcotics presents after being
found [R] unresponsive [R] at home.
His daughter awoke him at 7 a.m. ,
reports he said he felt cold and
shivery ,vomited several times , then
drove her to school ."
-----------------------------------
```

Figure 12. Handcrafted examples: (top) original example, (middle) the [L] and [R] tokens are switched, and (bottom): extending distance between entities with no additional information

We see that the Longformer w/ [CLS] + entity masking correctly predicts the flip in AFTER to BEFORE and maintains this prediction even after extending the distance between the tokens into a region where BEFORE is class dominant.

Conclusion

Classifying temporal relations in medical clinical text is a challenging task due to class imbalance, distance imbalance, and the limited number of examples. However, extraction of such information is critical in building a timeline of a patient's history in order to improve diagnosis.

This body of work explored the use of [CLS] + mask entity masking with Bio-ClinicalBERT and Longformer. Both models produced similar results to the baseline when compared macroscopically. However the Longformer produced promising results when the distances between entities were long, even under some distance imbalance.

Future Work

Several approaches should be considered in future work.

- Studying the behavior of global attention scores to understand what information the model is using to classify the entity pairs
- Handling of distance imbalance with resampling techniques such as SMOTE
- Studying Longformer performance for the entire dataset and document, which can be up to 2500 tokens in length.
- Ensemble approach, catering models for different temporal relationships and distances

Acknowledgements

I would like to thank my instructor, Joachim Rahmfeld, and TA, Zachary Alexander, for their brainstorming and feedback on this project.

References

1. Tao, C., Li, H. 2019. A General Approach for Improving Deep Learning-Based Medical Relation Extraction using Pre-trained Model and Fine-tuning

2. Han, X., Wang, Lei. 2020. A Novel Document-Level Relation Extraction Method based on BERT and Entity Information.
3. Sun, W., Rumshisky, A.,& Uzuner, O. 2013. Evaluating Temporal Relations from Clinical Text: 2012 i2b2 Challenge.
4. Soares, L. *et al*. 2019. Matching the Blanks: Distributional Similarity for Relation Learning
5. Sheikhalishah, S. *et al.* 2019. Natural Language Processing of Clinical Notes on Chronic Diseases: Systematic Review
6. Alsentzer, E. *et al*. 2019. Publicly Available Clinical BERT Embeddings.
7. Beltagy, I *et al.* 2020. Longformer: The Long Document Transformer
8. Lee, H. *et al.* 2018. Identifying Direct Temporal Relations between Time and Events from Clinical Notes
9. Li, S. 2021. Github Repository. https://github.com/sli0111/w266-2021-Medical-Relationship-Extraction-with-Bio-Clinical-BERT-and-Longformer

---

Appendix

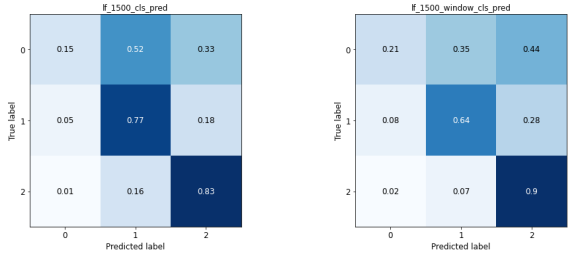


Figure 13. Confusion matrices for distances of 256 to 512. (Left): Longformer w/ window size=512, (Right): Longformer w/ window size that increases from 32 to 512 (bottom to top layers).