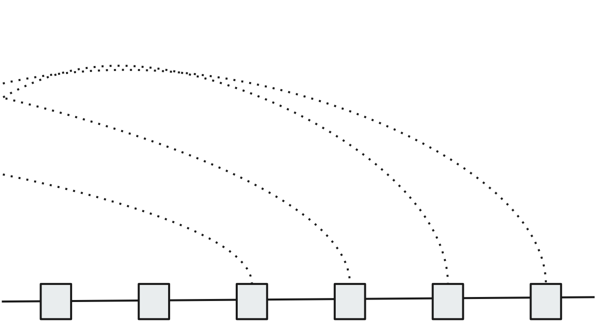
W266 - Medical Relation Extraction with Bio-Clinical BERT and Longformer

August 2, 2021

Simon Li

Medical Relation Extraction

ADMISSION DATE : **10/14/96**
DISCHARGE DATE : **10/27/96** [/T]
date of **birth** ; **September 30 , 1917**
THER PROCEDURES :
arterial catheterization on **10/14/96**
head CT scan on **10/14/96**
HISTORY AND REASON FOR HOSPITALIZATION :
Granrivern Call is a 79-year-old right handed white male with a history of questionable progressive supranuclear palsy **atrial fibrillation** , **deep venous thrombosis** , and **pulmonary embolus** who **presents** with a change in mental status and **a fall** at home . His family stated that he had a **progressive mental decline** over the past **three years** and was initially diagnosed with **Parkinson Zapos;s disease** or features consistent with P**arkinsonism.**



Timeline of patient medical history, by ordering all the entities in the document

Medical Relation Extraction

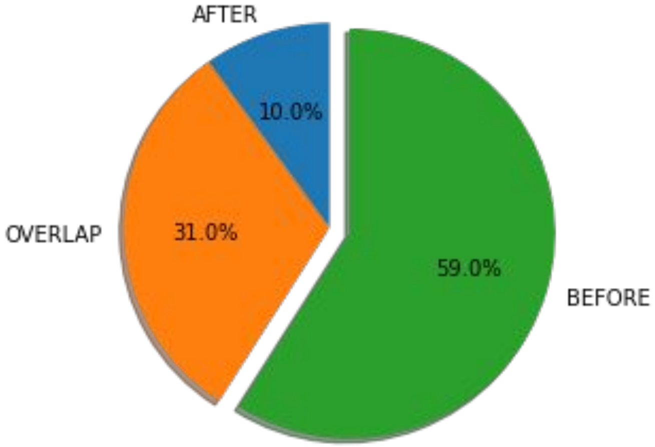
ADMISSION DATE : **10/14/96**
DISCHARGE DATE : **10/27/96** [/T]
date of **birth** ; **September 30 , 1917**
THER PROCEDURES :
arterial catheterization on **10/14/96**
head CT scan on **10/14/96**
HISTORY AND REASON FOR HOSPITALIZATION :
Granrivern Call is a 79-year-old right handed white male with a history of questionable progressive supranuclear palsy **atrial fibrillation** , **deep venous thrombosis** , and **pulmonary embolus** who **presents** with a change in mental status and **a fall** at home . His family stated that he had a **progressive mental decline** over the past **three years** and was initially diagnosed with **Parkinson Zapos;s disease** or features consistent with P**arkinsonism.**

Entity Pair (L, R)	Class
(progressive mental decline, three years)	**OVERLAP**
(arterial catheterization, birth)	**AFTER**
(fall, atrial fibrillation)	**BEFORE**

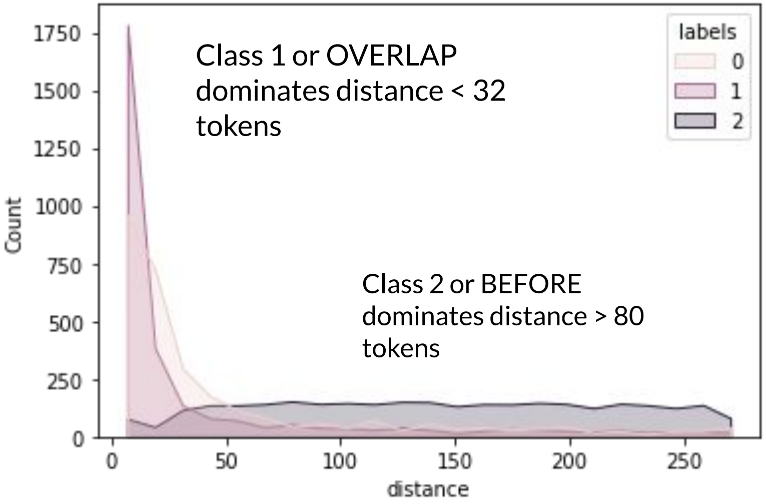
Classification task:

- 3 classes
- Distance between entities vary
- Documents often > 512 tokens

2012 i2b2 Dataset



Moderately unbalanced
(applied resampling to balance before modeling)



Class 1 or OVERLAP dominates distance < 32 tokens

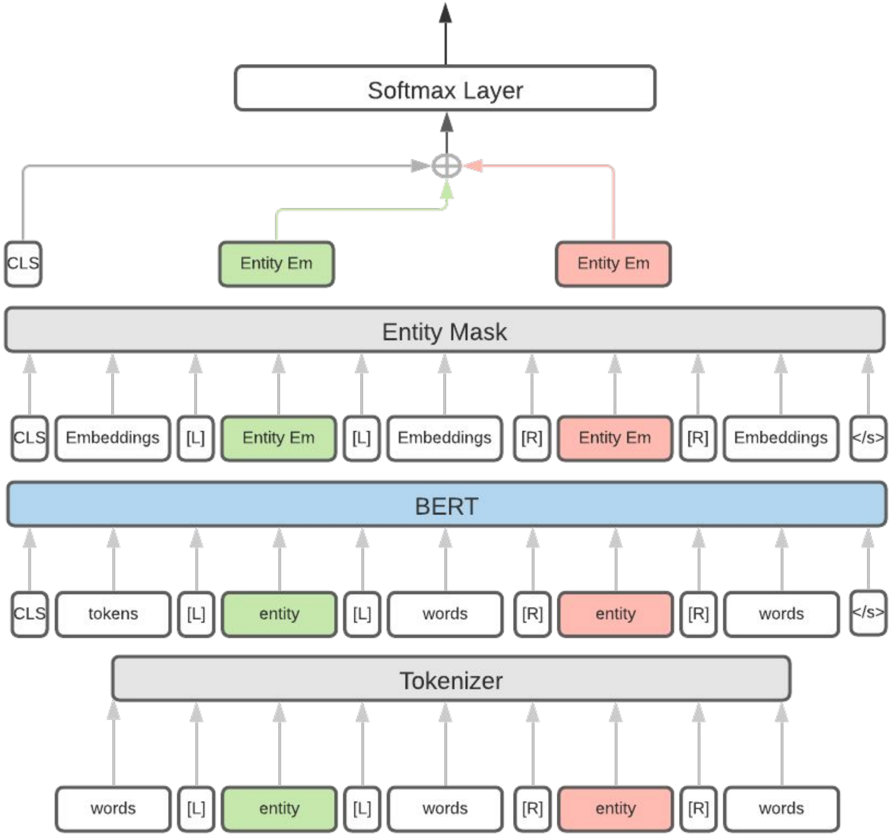
Class 2 or BEFORE dominates distance > 80 tokens

Train dataset with 9000 entity pairs.
Unbalanced amongst distance between entity pairs

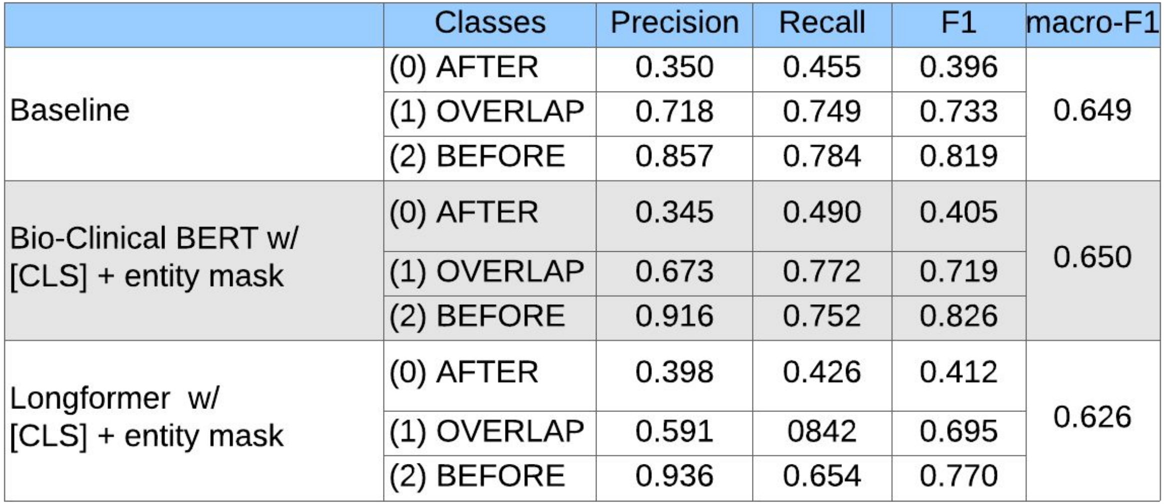
Bio-Clinical BERT and Longformer

Direct embeddings of the entity pairs with [CLS] token towards the classifier.

Allow model to find the entities and focus on the relationships between them



Precision, Recall, and F1-Score



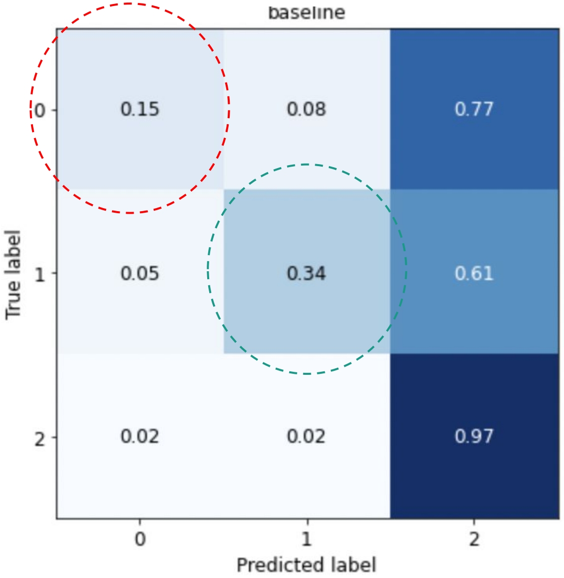
	Classes	Precision	Recall	F1	macro-F1
Baseline	(0) AFTER	0.350	0.455	0.396	0.649
	(1) OVERLAP	0.718	0.749	0.733	
	(2) BEFORE	0.857	0.784	0.819	
Bio-Clinical BERT w/ [CLS] + entity mask	(0) AFTER	0.345	0.490	0.405	0.650
	(1) OVERLAP	0.673	0.772	0.719	
	(2) BEFORE	0.916	0.752	0.826	
Longformer w/ [CLS]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

Resampling did not completely alleviate the problem.

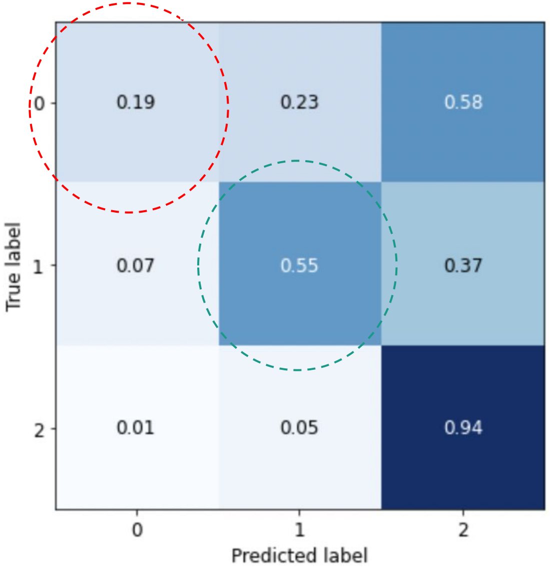
Few examples still make it hard for the learn to learn that class. It appears that Longformer resulted in different performance

Class and Distance Imbalance Highly Affected Model Performance - 256 to 512 tokens between entities

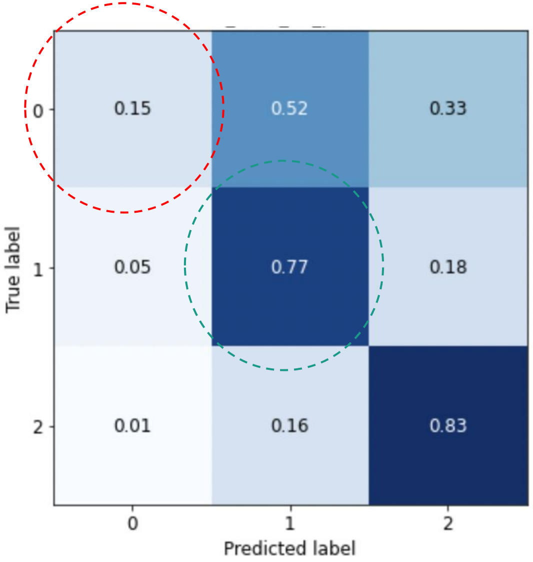
Baseline



Bio-Clinical BERT w/ [CLS] + entity mask



Longformer w/ [CLS] + entity mask



Handcrafted Test of Longformer

"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 ."

original text:
(true, pred) = (BEFORE, BEFORE)
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]:
(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 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 ."

flip [L] and [R] + filler text:
(true, pred) = (AFTER, AFTER)
distance = 158

Summary

Class and distance imbalance made it a challenging task to classify temporal relationships

Macroscopically, F1-score was higher for baseline and Bio-Clinical BERT but is partly inflated due to distance imbalance

On a distance basis, [CLS] + entity mask and Longformer learned OVERLAP class for long distance which is intuitively difficult

Future Work

- Apply more sophisticated resampling techniques to better balance classes and distance
- Study the behavior of global attention scores from Longformer and see if pays attention to keywords and headers
- Apply Longformer for the entirety of the document
- Ensemble approach, different models for different distances or temporal relations

Thank you!

Q/A