WINNERS AT W-NUT 2020 SHARED TASK-3 LEVERAGING EVENT SPECIFIC AND CHUNK SPAN FEATURES TO EXTRACT COVID EVENTS FROM TWEETS

Ayush Kaushal, Tejas Vaidhya
Indian Institute of Technology Kharagpur
ayushk4@gmail.com, iamtejasvaidhya@gmail.com

Task Introduction

The motivation for the task was to automate the extraction of a variety of COVID-19 related events from Twitter, such as individuals who recently contracted the virus, someone with symptoms who were denied testing and believed remedies against the infection.

The system submission had to identify event-specific questions about factual information or opinion for tweets across five COVID events: Tested Positive, Tested Negative, Can Not Test, Death, Cure.

Approach

- We re-defined the event-specific questions into two types: slot filling and sentence classification.
- We formulated two different multitask architectures for slot filling and sentence classification sub-tasks.
- For both sentence and slot-filling architectures:
  - We incorporated the event-prediction task as auxiliary subtask and fuse its features for all the event-specific subtasks. This additional task aims at identifying the corresponding event to which this tweet belongs.
  - We performed an attention-weighted pooling over the candidate chunk span enabling the model to attend to subtask specific cues.
  - We transferred learn over the domain-specific Bert Embeddings for COVID-Twitter Bert.
- We achieved the state-of-the-art result, **achieving 1st position** in the shared task overall without any ensembles or additional datasets.

Tweet	*Sigh of relief. My wife 's COVID-19 test came back negative today . The Lord has been gracious. One of my favorite pics I took of her. #thankful.[URL]*
Slot Filling	{Who}: My wife's, {Where}: Not Specified, {When}: today, {CloseContact}: Not Specified, {Age}: Not specified, {Duration}: Not Specified
Sentence Classify	{Relation}: Yes, {Gender}: Female
Corresponding Event	{Did someone test negative?} Yes

Table 1: An example tweet from tested negative event.

Architectures

Slot-Filling Architecture

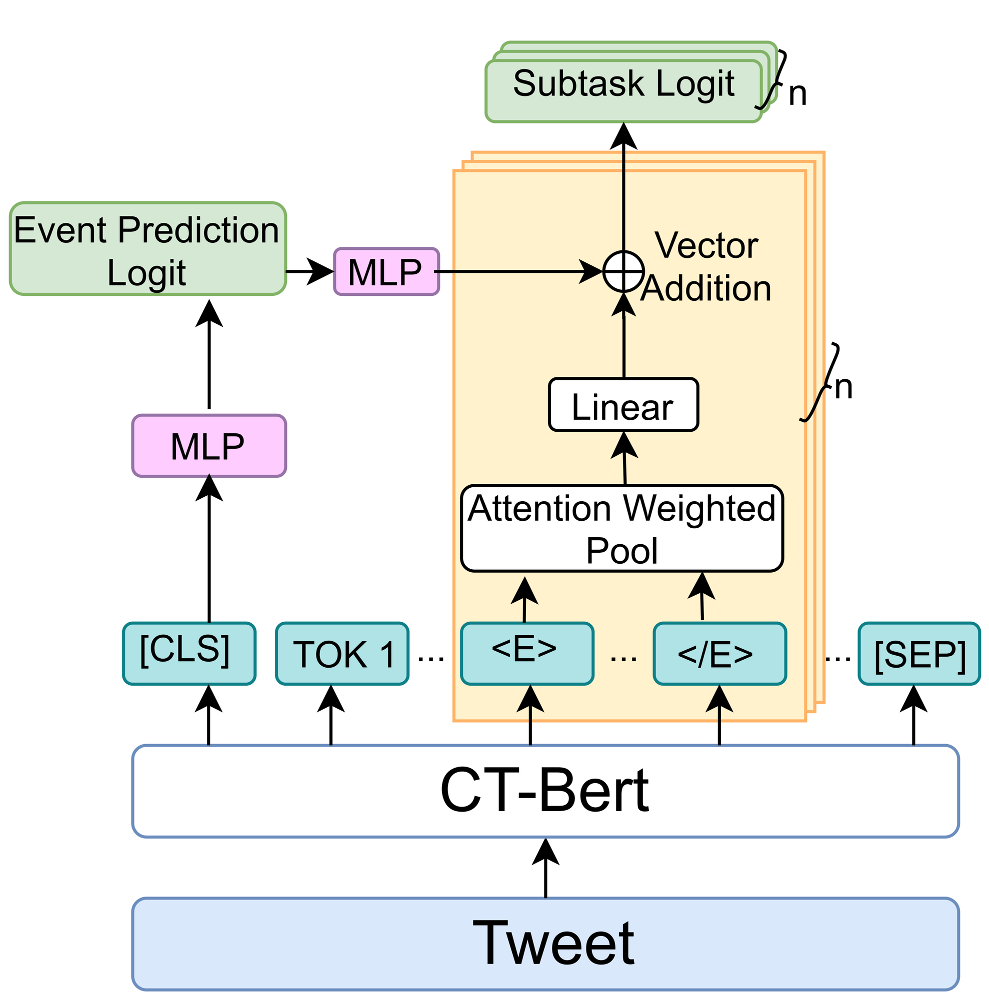

Figure 1: Slot-Filling Model, Here n is the number of slot-filling subtasks.

We enclose the candidate slot inside a start $< E >$ and an end $< /E >$ tokens. The contextualised embeddings for the candidate slot is obtained using attention weighted pooling over Bert embeddings. The logits from the auxiliary task, fused with the pooling slot feature then give the predictions for each slot filling question.

Sentence Classification Architecture

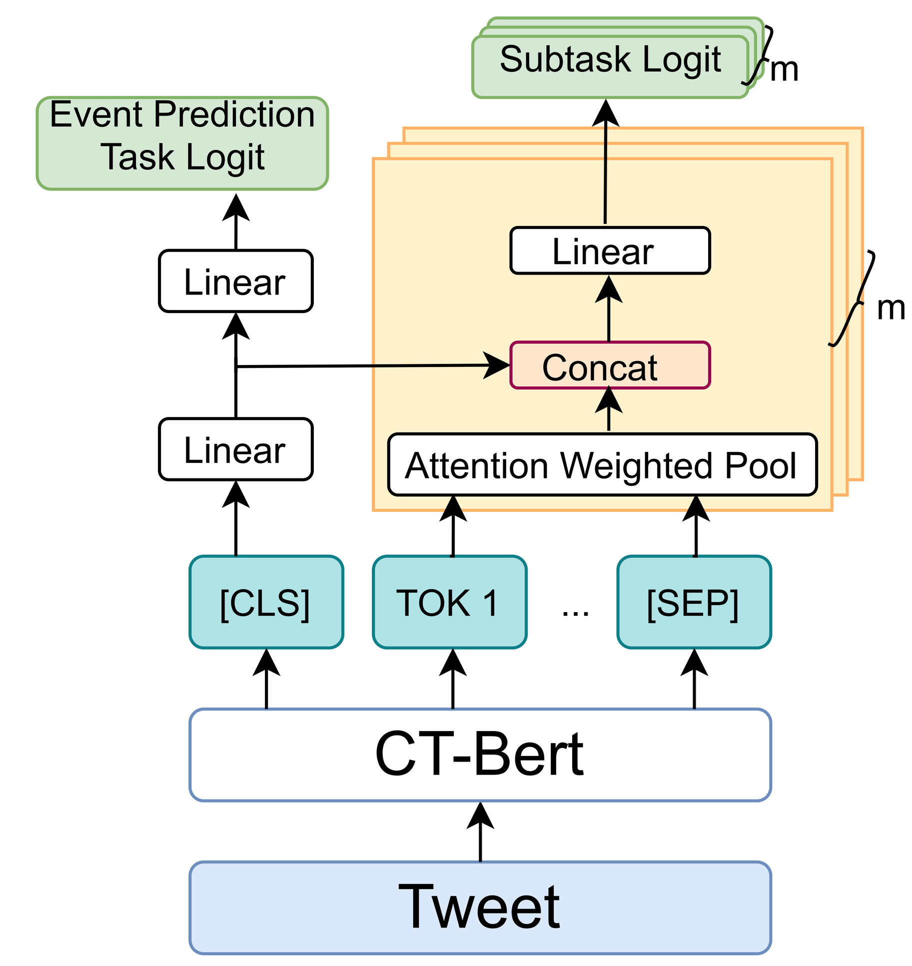

Figure 2: Sentence Classification model, Here m is the number of Sentence Classification subtasks.

The features from the auxiliary task predictions, combined with the Bert sentence vector is used for sentence classification.

Results

- Our system ranks 1st position in the W-NUT 2020 Shared Task-3.
- We also independently rank 1st for 3 events: 'Can Not Test', 'Death', 'Cure'.

Event	F1	Precision	recall
Tested Positive	.66	.80	.58
Tested Negative	.66	.66	.67
Can Not Test	.65	.67	.64
Death	.69	.72	.67
Can Not Test	.63	.75	.53
Overall	**.66**	**.73**	**.60**

Table 2: *Micro averaged scores on the held out test set*

Ablation Studies

- We find that having different architectures for slot-filling and sentence classification results in improved performance.

Model	Micro F1	Macro F1
Bert Separate	**.631**	**.545**
Bert Baseline	.608	.512

Table 3: Results comparing the systems treating the sentence classification and slot-filling subtasks separately vs those treating it similarly.

- Domain-Specific Bert, Attention Weighted pooling and auxiliary task, all results in an improvement over vanilla Bert for both slot filling and sentence classification.

Model	Micro F1	Macro F1
Our (SF)	**.684**	**.558**
Our (SF) w/o pool	.678	.557
Our (SF) w/o CES	.665	.552
CT-Bert	.662	.551
Bert (large)	.610	.529
Bert (baseline)	.612	.528

Model	Micro F1	Macro F1
Our (SC)	**.788**	**.767**
Our (SC) w/o CES	.777	.731
CT-Bert multitask	.760	.717
Bert multitask	.715	.612
LSTM multitask	.614	.543

Table 4: Results of slot-filling models (on left) and sentence classification models on our 70-30 split(on right).