

# Midterm Project Report: Event Argument Extraction

May 30, 2023

Name: Kia Karbasi

UID: 804617014

Confirmation Code: CS263-KWTPEWMM

## 1 Building an Event Argument Ontology

For each of the events, I started by constructing questions that arise from hearing that the event has taken place. Specifically, I tried to come up with when/what/where/how questions that would construct a narrative around the fact that the event has taken place. Moreover, I looked at the in-context tweets that were provided to understand what information could be provided in a tweet about the events. Based on these, I chose appropriate argument roles for each of the events

Since the EAE system in this assignment was specifically indented for text related to the COVID-19 pandemic, on some events (prevent, and control), I chose argument roles that could probably provide information about the event relating to the pandemic. That said, if this EAE system were to be designed for a more general case, I might have chosen different argument roles or different names of argument role for adjacent concepts.

Lastly, "spread" and "infect" events seemed to have some conceptual overlap as the trigger words for tweets labeled as containing them could have been used for both events. Specifically, a tweet about the event "spread" in the in-context set was about a few people getting infected where the event definition is "The process of a disease spreading/prevaling massively at a large scale" which essentially differentiates it from "infect" due to scale, but it is ambiguous what scale. This caused some confusion because depending on how to interpret the scale, one could include or exclude "some origin or source" as an argument role for this event.

## 2 Creating a Gold Annotation

For this section, I identified the words in the tweets that were corresponding to the roles that I defined in section 1.

The ambiguous cases were where multiple word arguments and multiple words for the same argument existed in the tweet. The choice that I had to make was to whether to annotate the multiple words for the same argument as a list or together in the same string, for which I chose the list option.

While annotating the data, I had to go back and update my ontology as I realized that in some tweets, there was some information related to the events that I hadn't considered when designing the ontology.

### **3 Model Prediction on your data**

#### **Zero Shot**

For this section I tried a few zero shot prompts where I didn't use any of the in-context examples and asked chatGPT and only used the event name and event trigger. I also tried providing the roles and asking the model to identify the arguments. I tried using natural language to make sure the model outputs the desired output that was similar to the annotated roles. However, the output of the model was not consistent and sometimes the model's output was not even in the tweet itself.

#### **One Shot**

Then I tried one shot prompts where I included one example from the in-context set. The performance improved, specially in the consistency of the format of the output. I also performed cross validation by trying different tweet from the in-context set as the exemplar. The output of the model became more consistent but in few of the 6 testing scenarios (1 tweet as exemplar and test with the two remaining tweets) the output was not the same as the annotated example, specially when the test tweet had multiple words for the role).

#### **Few Shot**

Finally, I tried few shot (two shot) by using two of the in-context tweets as exemplars. I cross validated the two exemplars and tried switching the order of exemplars to see if I observe consistency in the output. While I learned that the order did not have an impact on the performance, I also learned that choosing the two exemplars with most diversity of the annotation answers (making sure one has an annotated role with a list multiple words and roles with multiple words).

#### **Including information from the Ontology**

For zero shot, one shot, and few shot settings, I tried prompts where I included only the input text and desired output as well as prompts with more information including, providing event information including event definition and event trigger word, event information and role definitions, and entire ontology. Unsurprisingly, the more information and context I provided the model, the performance improved accordingly. However, I realized when I provided the entire ontology in the exemplars, sometimes the model would include words from example sentences in the output. The confusion of the model made sense here because there are many more sentences in the ontology of the two exemplars than the two exemplar tweets. Moreover, including the example sentences greatly increased the overall length of the prompts and in some cases close to the 4096 token limit of the underlying GPT model of chatGPT. I learned that when you prompt chatGPT with a really long prompt (>4096), the model ignores more content from the beginning of the prompt and in my setting, it was detrimental to the performance.

### **Output Format**

In the exemplars, I formatted the answers from annotations as JSON so that the generated answer would be easily readable as a JSON object by the argument extractor.

This made my argument extractor function to be minimal

### **Adding Natural Language Context**

I also tried various ways to instruct the system using natural language in order to provide more context and utilize the model's familiarity with the Event Argument Extraction task and its familiarity of the relation of the task with an ontology through its training, since I hypothesised that it was probably trained over a large amount of text related to EAE as well. In my best prompt, I specifically as the model to "Perform Event Argument Extraction according to the following ontology: ... " Moreover, I learned through trial and error that adding an additional statement after the exemplars and before the actual test tweet improved the performance of the model. This specially helped the model to only provide the desired output and avoid adding additional text to its output. More specifically, I added "You are EAE system..." after the exemplars and before my target tweet and all the other natural language statements that were the same as the exemplars.

### **GPT4 vs GPT3.5**

For the generation I tried both the default free GPT3.5 and the paid GPT4 versions of the chatGPT. The output from the default GPT3.5 was sometimes inconsistent for some of the prompts. Specifically for some of the prompts, the model would sometimes generate different outputs for the same prompt and a few of the outputs were not even in the desired JSON format and looked like the fail-safe responses that chatGPT users are familiar with (I'm sorry, as a large language model...). However, when using the GPT4 version the outputs were almost always consistent for all the prompts no matter how many times I prompted the model. While we don't know much about the implementation details of chatGPT and the underlying GPT4 model, it seems to have superior performance on this task.

## **4 Evaluation of model**

The evaluation metrics for the EAE task are shown in Table 1 below and actual argument roles (gold and predicted) are shown in Table 2 below.

**Table 1 : Evaluation Metrics**

<b>Tweet</b>	<b>Precision</b>	<b>Recall</b>	<b>F1</b>
Polite request You may not believe the COVID-19 is serious, or even real. That's up to you. But please respect the feelings of people on this platform who've had relatives & loved ones die of it. It's a truly terrifying disease that has killed tens of thousands of people. Thanks.	100.0%	100.0%	100.0%
(user) (user) So you actually believe only 11k Americans have died from Covid. You screenshot your proof from the CDC. I correct you from the same web sight and you tell me Im wrong. Its obvious you cant read stats. Go back to school	100.0%	100.0%	100.0%
Anosmia is the commonest symptom of COVID-19 - can be the only symptom - those with it need COVID testing Our own Prof Hopkins & (user) led much of the research & was discussed (user) COVID-19 Grand Round Gd summary by (user) (url)	100.0%	100.0%	100.0%
(user) (user) (user) (user) How many of those deaths are actually from Covid 19?	100.0%	100.0%	100.0%
Five members of the Canadian Armed Forces working in care homes in Ontario and Quebec have tested positive for #(COVID). Vanessa Lee has the details: (url) (url)	100.0%	100.0%	100.0%
Itll get reported as a #(COVID) death though. He tested positive for the virus #(COVID)	100.0%	50.0%	66.7%
USA: Risk of dying of Covid-19 versus other causes (url)	100.0%	100.0%	100.0%
(user) you let Robert Patillo say there's no treatment for COVID-19 WTF is your problem? You don't even mention this cheap cure that is being used across the world with great success? Terrible job totally unprepared!! Are you saying (user) and her guests are wrong?	100.0%	100.0%	100.0%
We know China lied about the initial spread of #(COVID), putting people all across the world at risk. They now need to come clean about its origins. (url)	33.3%	50.0%	40.0%
Self quarantining & shutting down businesses was done to FLATTEN the curve not to eliminate COVID. We needed time to prepare for increased hospitalization & obtain PPE. Many in the media are acting like we cant open up until COVID is eliminated & that may never happen.,	50.0%	100.0%	66.7%
Where to worry about catching Covid-19, and where not to (url)	100.0%	100.0%	100.0%
I got COVID-19 can you help me?	100.0%	100.0%	100.0%
<b>F1 Score Average:</b>			<b>89.4%</b>

*Table 2 Argument Role Annotations*

Tweet	Model Argument Annotations	Gold Argument Annotations	Correct Argument Prediction
Polite request You may not believe the COVID-19 is serious, or even real. That's up to you. But please respect the feelings of people on this platform who've had relatives & loved ones die of it. It's a truly terrifying disease that has killed tens of thousands of people. Thanks.	{'Cause': 'COVID-19', 'Person': 'tens of thousands of people'}	{'Person': 'tens of thousands of people', 'Cause': 'COVID-19'}	{'Person': 'tens of thousands of people', 'Cause': 'COVID-19'}
(user) (user) So you actually believe only 11k Americans have died from Covid. You screenshot your proof from the CDC. I correct you from the same web sight and you tell me Im wrong. Its obvious you cant read stats. Go back to school	{'Person': '11k Americans', 'Cause': 'Covid'}	{'Person': '11k Americans', 'Cause': 'Covid'}	{'Person': '11k Americans', 'Cause': 'Covid'}
Anosmia is the commonest symptom of COVID-19 - can be the only symptom - those with it need COVID testing Our own Prof Hopkins & (user) led much of the research & was discussed (user) COVID-19 Grand Round Gd summary by (user) (url)	{'Name': 'Anosmia', 'Patient': 'those with it'}	{'Name': 'Anosmia', 'Patient': 'those with it'}	{'Name': 'Anosmia', 'Patient': 'those with it'}
(user) (user) (user) (user) How many of those deaths are actually from Covid 19?	{'Cause': 'Covid 19'}	{'Cause': 'Covid 19'}	{'Cause': 'Covid 19'}
Five members of the Canadian Armed Forces working in care homes in Ontario and Quebec have tested positive for #(COVID). Vanessa Lee has the details: (url) (url)	{'Place': ['Ontario', 'Quebec'], 'Host': 'Five members of the Canadian Armed Forces', 'Disease': 'COVID'}	{'Place': ['Ontario', 'Quebec'], 'Host': 'Five members of the Canadian Armed Forces', 'Disease': 'COVID'}	{'Place': ['Ontario', 'Quebec'], 'Host': 'Five members of the Canadian Armed Forces', 'Disease': 'COVID'}
Itll get reported as a #(COVID) death though. He tested positive for the virus #(COVID)	{'Cause': 'COVID'}	{'Person': 'He', 'Cause': 'COVID'}	{'Cause': 'COVID'}
USA: Risk of dying of Covid-19 versus other causes (url)	{'Place': 'USA', 'Cause': 'Covid-19'}	{'Place': 'USA', 'Cause': 'Covid-19'}	{'Place': 'USA', 'Cause': 'Covid-19'}
(user) you let Robert Patillo say there's no treatment for COVID-19 WTF is your problem? You don't even mention this cheap cure that is being used across the world with great success? Terrible job totally unprepared!! Are you saying (user) and her guests are wrong?	{'Treatment': 'this cheap cure', 'Disease': 'COVID-19', 'Effectiveness': 'great success'}	{'Treatment': 'this cheap cure', 'Effectiveness': 'great success', 'Disease': 'COVID-19'}	{'Treatment': 'this cheap cure', 'Effectiveness': 'great success', 'Disease': 'COVID-19'}
We know China lied about the initial spread of #(COVID), putting people all across the world at risk. They now need to come clean about its origins. (url)	{'Disease': 'COVID', 'Place': 'world', 'Source': 'China'}	{'Disease': 'COVID', 'Place': 'all across the world'}	{'Disease': 'COVID'}
Self quarantining & shutting down businesses was done to FLATTEN the curve not to eliminate COVID. We needed time to prepare for increased hospitalization & obtain PPE. Many in the media are acting like we cant open up until COVID is eliminated & that may never happen.,	{'Action': ['Self quarantining', 'shutting down businesses'], 'Actant': 'Many in the media'}	{'Action': ['Self quarantining', 'shutting down businesses']}	{'Action': ['Self quarantining', 'shutting down businesses']}
Where to worry about catching Covid-19, and where not to (url)	{'Disease': 'Covid-19'}	{'Disease': 'Covid-19'}	{'Disease': 'Covid-19'}

I got COVID-19 can you help me?	{'Host': 'T', 'Disease': 'COVID-19'}	{'Host': 'T', 'Disease': 'COVID-19'}	{'Host': 'T', 'Disease': 'COVID-19'}
---------------------------------	--------------------------------------	--------------------------------------	--------------------------------------

The model achieves an overall ~%90 F1 score which is reasonable overall, however looking at the three cases that have brought the score down we can observe the following:

1. In the tweet *"Itll get reported as a #(COVID) death though. He tested positive for the virus #(COVID) "* the model can't seem to make the connection between tested positive for covid and dying from covid. Either because they are in different sentences or the model simply doesn't "understand" the semantics here.
2. In the tweet *"We know China lied about the initial spread of #(COVID), putting people all across the world at risk. They now need to come clean about its origins. (url)"* the model thinks that china is the source of the spread. While that might be true, that information is not mentioned in the tweet itself. That information might be the result of the model's pretraining that might have that information. Moreover, the model fails to label "all across the world" as place of spread. This might be due to the fact that we only used two exemplars here and there were no gold annotated multi-word argument role in them.
3. In the tweet *"Self quarantining & shutting down businesses was done to FLATTEN the curve not to eliminate COVID. We needed time to prepare for increased hospitalization & obtain PPE. Many in the media are acting like we cant open up until COVID is eliminated & that may never happen."* The model seems to think that *"Many in the media"* are the actants behind control actions taken. This might be due to my ontology design and argument role name and description, since "Actant" is a very a very general term. Moreover, the description that I wrote is also very general and refers to individuals taking action.

## 5 Breaking ChatGPT

For this section, I first tried adversarial methods by perturbing the tweets in the eval dataset to change the model output. In some of the examples, I also introduced additional context that would trick the model to include its inherit biases in its output. In another example, I tried introducing additional context for tricking the model to make semantic mistakes due to the similarity of the concepts. For example, in one tweet where the event of interest was death of an individual with trigger word "passed", I introduced another death with a different keyword. However, the model was unable to distinguish the two separate events and attributed the cause of the introduced death event to the original event of interest. I created another example using a similar technique by introducing locations unrelated to the event of interest in order to trick the model to attribute them to the location of the event of interest. In other examples, I perturbed the eval dataset tweets by complicating the semantics and switching around words that also managed to confuse the model and generate incorrect output.

Moreover, I came up with some sentences that demonstrated inherit biases of the model. Specifically, I wrote two sentences in an example that are unrelated, and the event of interest is only in one of them. However, the model attributed the cause of a spread in public schools to massive influx of undocumented immigrants.

Finally, I created examples that would trick the model to hallucinate some information that it has probably learned during its training. For example, I mentioned the word “China” multiple times in an example about the event “spread” though none of these mentions are related to the source of the spread. However, the model identified “China” as the source of the spread which is probably due to lots of articles in its corpus from before 2021 that refer to China as the origin of the pandemic.