ChatGPT as a Zero-Shot and Few-Shot Entity Relation Classifier

Benjamin Soli

1 Introduction

The goal of this project is to evaluate ChatGPT's ability to classify entity relations. The data set used comes from the SemEval 2018 Task 7 which involves classifying the entity relations from entities contained within the abstract of a scientific paper. The data set has six possible labels and the ontology for the labels is outlined in (Gábor et al., 2018). The experimental approaches used for this are similar to zero-shot and few-show classification tasks. The goal is evaluate the use of an ontology when prompting a model to perform an entity relation classification task.

The paper is organized follows: Section 2 gives an overview of background works that approach this task and describes the current limitations of recreating those approaches using generative models. This section also describes the alternatives used for these experiments. Section 3 describes the structure of the data set and gives an overview of how it was processed for this task. Section 4 describes the experimental configurations used. Section 5 gives the quantitative results and provides some analysis about their applicability in measuring the performance of text classification with generative model. Section 6 proposes some future work for this area.

2 Background

Recent approaches to zero-shot entity relation classification have attempted to use label descriptions as a form of weak supervision to allow the model to create a better semantic representation of the label space (Chen and Li, 2021). The current benchmark data sets for this task have become the WikiZSL and the FewRel data sets (Chen and Li, 2021), (Wang et al., 2022). However, due to the context window size limitations of gpt-3.5-turbo, it is unfeasible to provide a generative model with a large ontology as a prompt. To explore this model's capability of performing zero-shot and few-shot entity relation classification, the SemEval 2018 Task 7 data (Gábor et al., 2018) has a much more suitably sized ontology which consists of 6 labels. These experiments use the data outlined by sub-task 1.1 for this challenge which aims to predict the label given an entity pair and an abstract in which the entity pair occurs. The approaches in (Chen and Li, 2021), (Wang et al., 2022) involve stratifying the training and test data by label. That is, there are instances in the test data that have a label not found in train. This is unfeasible with the SemEval 2018 Task 7 data because a data instance contains multiple entities and it would be difficult to ensure that that the sets are split according to label. However, this should not be an issue because the model will be provided a very small amount of labeled training data.

3 Data Processing

The data set used comes from the SemEval 2018 Task 7 (Gábor et al., 2018). The data set contains paper titles, abstracts, a list of entities in each abstract, and relation classification with entity pairs. Entities are marked by their character indices in the abstract and relations are marked by their numerical key in a dictionary which maps integers to string labels. Pre-processing was performed to create a tuple representation of the form (LABEL, ARG1, ARG2). The data set was accessed via the datasets library available through huggingface. There is a problem with the test set which contains all empty classifications as this is not the finalized data set post-submission. I was unable to locate a labeled version of the test set through the task's website. Because I do not plan to use the entirety of the training set, I elected to split the training set into a test set where 80 % of the abstracts are within train with the remaining 20 % in test.

4 Experimental Overview

For these experiments, the model used was 'gpt-3.5-turbo'. This is a chat completion model which is unable to be fine-tuned towards any specific task. The goal of these experiments is to explore prompt engineering to elicit output from the model that can be used as classifications. To constrain the models tendency to generate random output, the temperature setting was set to a very low .001.

The first experiment can be framed as a zero-shot classification task. In this approach, the model is prompted with to act as an entity relation classification model and provided with an ontology containing descriptions. Each abstract can contain multiple entity pairs. The data given to the model for each classification contains a prompt outlining the task description along with a summary of the ontology presented in (Gábor et al., 2018), and a data point containing the abstract, and the entity pairs as an ordered list. The model is directed to return an ordered list of labels for each entity pair.

The second experiment is also a zero-shot classification task. The set up is similar to the first one, except the prompt is more detailed and also uses a much more detailed version of the ontology from (Gábor et al., 2018).

The third experiment attempts to "fine-tune" the model using techniques common in supervised NLP. The model is given a prompt similar to the first experiment. It is also provided with one training data instance at a time. After it gives its predictions, the model is given the correct labels. This approach is highly limited by the size of training instances. Using a training set too large will exceed the allotted context window size of 4,096 tokens allowed by the model. To account for this, I only tested data set sizes of 1, 5, 10.

5 Results

This section will give the results for the model including a break down of performance by label. The labels in the ontology are COMPARE, MODEL-FEATURE, PART_WHOLE, RESULT, TOPIC, and USAGE. In some configurations, the model hallucinates new label categories. Those are included in the results table.

It should also be noted that using this approach does not guarantee the model will give a classification for every entity pair in the data instances. In these cases, I elected to ignore those labels in scoring the model. How likely the configuration is to generate incorrectly formatted results should be used as a metric to evaluate the model. At this time, I am unsure what the best approach would be for incorporating this into a meaningful quantitative analysis. For this reason, I only use precision, recall, F1, and accuracy, along with the macro and micro scores. The test set has a total of 289 entity pairs for classification.

Experiment 1 Results:

Label	Prec.	Rec	F1	Support
COMPARE	.29	.36	.32	14
MODEL-F	.00	.00	.00	0
MODEL-FEATURE	.43	.35	.39	79
PART_WHOLE	.42	.50	.46	60
RESULT	.32	.61	.42	18
TOPIC	.05	.12	.07	8
USAGE	.64	.45	.53	97
accuracy			.43	276
macro avg	.31	.34	.31	276
micro avg	.48	.43	.44	276

For this experimental configuration, the model hallucinates a label. It may be possible the model truncated the label name due to reaching the token limit for generating predictions, but this is unlikely because no other model has this issue and each used the same test set and same token size. It should also be noted that this configuration is unable to be scored for 13 entity pairs due to incorrectly formatted output. Visual inspection of those missed entities shows that the model did not provide enough labels for the number of entity pairs in the training instances.

Experiment 2 Results:

Label	Prec.	Rec	F1	Support
COMPARE	.42	.36	.38	14
MODEL-FEATURE	.40	.38	.39	71
PART_WHOLE	.33	.32	.32	60
RESULT	.38	.88	.54	17
TOPIC	.21	.57	.31	7
USAGE	.57	.42	.48	96
accuracy			.42	265
macro avg	.33	.49	.40	265
micro avg	.44	.42	.42	265

For this experimental configuration, it is interesting that the model does not hallucinate any labels. This configuration does exclude 23 entity pairs. The inspection of these errors shows that model over-generated its output in some cases. Rather than just providing labels, it also provided explanations in some cases. The prompt explicitly instructs it not to do so, but the prompt for this configuration is much more detailed so it seems the model is more likely to inconsistently follow instructions.

Experiment 3 Results (training instances = 1):

Label	Prec.	Rec	F1	Support
ASPECT	.00	.00	.00	0
COMPARE	.29	.43	.34	14
MODEL-FEATURE	.36	.38	.37	82
PART_WHOLE	.38	.49	.43	63
RESULT	.50	.83	.62	18
TOPIC	.21	.50	.30	8
USAGE	.58	.26	.36	99
accuracy			.40	284
macro avg	.33	.41	.35	284
micro avg	.44	.40	.39	284

For these results, the model hallucinates an entirely new label. It also only misses five entity pairs in the training set. In this case, the model did not produce a label for each of the entities in the training instance.

Experiment 3 Results (training instances = 5):

Label	Prec.	Rec	F1	Support
COMPARE	.50	.36	.42	14
MODEL-FEATURE	.40	.67	.50	83
PART_WHOLE	.40	.38	.39	65
RESULT	.57	.89	.70	18
TOPIC	.21	.38	.27	8
USAGE	.65	.22	.33	101
accuracy			.44	289
macro avg	.46	.48	.43	289
micro avg	.50	.44	.42	289

Using 5 training instances, the model produces a classification for each entity pair and hallucinates no new labels. There also seem to modest improvements to the overall performance.

Label	Prec.	Rec	F1	Support
COMPARE	.39	.50	.44	14
MODEL-FEATURE	.41	.67	.51	83
PART_WHOLE	.44	.43	.44	65
RESULT	.58	1.00	.73	18
TOPIC	.12	.12	.12	8
USAGE	.67	.22	.33	101
accuracy			.46	289
macro avg	.44	.49	.43	289
micro avg	.51	.46	.43	289

Using 10 training instances, the model again produces correctly formatted output. There is also a slight improvement in the performance of the model.

While the results are overall poor, it is impressive what the model is able to do with very limited data about the task. The results indicate that the combination of using an ontology for weak supervision combined with supplying data instances in a manner similar to supervised learning yields the highest quality results. This approach to using large generative models for text classification also highlights some issues with this application. Traditional supervised approaches are incapable or hallucinating new labels and tend to improve their performance when trained with more data. The amount of data for training this model was limited to the size of the context window allowed by the model. This limits how much insight can be derived from the slight improvement when using 5 training samples versus 10 training samples. It is interesting to note how including more training data does prompt the model to perform the task exactly as specified in the prompt. While it does not seem to improve the model's understanding of what the labels mean, it does improve the model's overall understanding of the task.

6 Future Directions

From this project, it is clear that the standard way of evaluating classification tasks do not measure the performance of generative models performing these tasks as well as it measures the performance of supervised learning. Some key factors that should be assessed in future research on this task are how to account for instances where the model does not generate correctly formatted output. In the case of these models, the overall metrics were very close, but models that hallucinate output or do not generate the correct number of output labels should be penalized in some meaningful way.

It is also important to note the data instances for this task are relatively large for a classification task, given that they consist of a full paragraph of text along with a list relevant entities. It is likely a task with a similar sized ontology and smaller data samples would greatly benefit from this approach of fine-tuning gpt-3.5-turbo as if it were a supervised model because they would be able to include more data without filling the model's context window.

One advantage of using supervised learning is that is possible to get output in form of a probability distribution. This can give some insight into the model's confidence for its classification. This is not possible for this generative approach. While it is possible to prompt the model to ask why it gave a classification for a specific instance, this approach cannot scale to a larger data set. Supervised learning is also guaranteed to provide the correct number of labels. While it is possible to reprompt the generative model, at this time it was difficult to define a procedure for doing this methodically. For this reason, incorrectly formatted predictions were excluded from scoring. Future work should focus on reprompting techniques to avoid this issue.

Lastly, as the computational capabilities increase and larger context windows become feasible, it would be interesting to revisit this approach with more data. This would hopefully yield more meaningful results and more clearly differentiate the impact of larger training data sets on this model's performance.

7 Conclusion

This project has attempted to use gpt-3.5-turbo as an entity relation classification model using various prompts. I also attempted to combine supervised learning and prompt engineering. The results show that training data improves the models understanding of to perform the task but does not significantly improve the model's performance when it comes to selecting the correct label. The model's tendency to hallucinate and perform the task incorrectly created issues in evaluating the performance of the model and devising metrics with this in mind is left to future work. Lastly, due to the nature of these models, the data cannot be distilled into a numerical representation as with supervised learning. This creates a limit on the amount of training data the can be provided to the model which limits how much insight can be derived from assessing the model's performance according the the size of the input data set.

References

Chih-Yao Chen and Cheng-Te Li. 2021. Zs-bert: Towards zero-shot relation extraction with attribute representation learning.

Kata Gábor, Davide Buscaldi, Anne-Kathrin Schumann, Behrang QasemiZadeh, Haïfa Zargayouna, and Thierry Charnois. 2018. SemEval-2018 task 7: Semantic relation extraction and classification in scientific papers. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 679–688, New Orleans, Louisiana. Association for Computational Linguistics.

Shusen Wang, Bosen Zhang, Yajing Xu, Yanan Wu, and Bo Xiao. 2022. RCL: Relation contrastive learning for zero-shot relation extraction. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2456–2468, Seattle, United States. Association for Computational Linguistics.