Research Proposal for "Evaluation of Automated Error Analysis with LLMs on three tasks: Schema-Matching, Sentiment Analysis and Price Prediction"

Tim Gutberlet

January 17th, 2024

1 Introduction

Since the release of ChatGPT, Natural Language Processing (NLP) has demonstrated its significance and power. Even stronger models, such as OpenAI's GPT4, have been released in a short span of time. LLMs have demonstrated proficiency in various new tasks, such as entity matching, that previously required significant amounts of fine-tuning [1]. Prompt-Engineering Methods such as Few-Shot Prompting and Automated Error Analysis have been used to better understand answer structures and improve the performance of LLMs. In my thesis I aim to apply these techniques to three new task types: Schema-Matching, Sentiment Analysis and Price Prediction to test if and to what extent error analysis can be used to better understand and potentially improve the performance of solving these tasks with LLMs.

2 Background

Large language models (LLMs) are advanced artificial intelligence systems that have been trained on vast amounts of text data to understand, generate, and respond to human language in a way that is both contextually relevant and coherent. *Prompt engineering* is the practice of strategically formulating input prompts to optimize the performance of LLMs models in generating specific outputs.

Zero-shot prompting involves presenting a task to a model without any prior examples, requiring it to understand and execute the task based solely on the given instructions. *Few-shot prompting*, on the other hand, provides the model with a limited number of examples to guide its understanding and response to a new task.

To enhance the effectiveness of Zero-Shot and Few-Shot prompting, it is essential to comprehend the nature and causes of errors the LLM does [2]. To gain a more profound understanding of arising error patterns, the occurring errors can be clustered into distinct error classes. However, manually understanding and clustering these error classes can be an inefficient and a time and resource-intensive process. In this thesis, I will investigate to what extent and how well LLMs can conduct *error analysis* in an automated way to enhance the effectiveness of Zero-Shot and Few-Shot prompting for different tasks. The thesis will focus on investigating automated Error Analysis in LLMs for three new tasks: Schema Matching, Sentiment Analysis and Price Prediction.

A *schema* is a structured framework, typically hierarchical, that outlines how data are organized within a database. *Schema matching* is the process of aligning one schema with another. Typically, schema matching is executed manually by domain experts. This approach, however, is notably labor-intensive and time-consuming. LLMs could provide a promising method of automation.

Sentiment Analysis involves the computational study of opinions, sentiments, emotions, and attitudes expressed in text. It aims to determine the polarity of a given text — whether the

expressed opinion in a document, sentence, or phrase is positive, negative, or neutral. *Aspect-Based Sentiment Analysis (ABSA)* is an advanced branch of sentiment analysis that focuses on identifying the sentiment towards specific aspects within a text, rather than giving an overall sentiment score to the entire text. LLMs have been shown to perform well on classical Sentiment Analysis, but they make enough errors on ABSA to perform error analysis [3].

Regression, in the context of statistics and machine learning, is an analytical technique used to model and analyse the relationships between variables. It is primarily focused on determining the extent to which a dependent variable changes when one or more independent variables fluctuate. To exemplify this using price prediction, consider the task of forecasting the price of a car. In this scenario, the car price is the dependent variable, while the independent variables could include factors like the make and model of the car, its age, mileage, engine size, fuel type, and other relevant features. As tasks such as price prediction have been historically solved with classical machine learning methods, it poses an interesting question on how well LLMs perform and can analyse their errors in this segment.

3 Goals and Work Plan

Automated error analysis has been demonstrated to be effective in tasks such as entity matching and assessing translation quality. This thesis aims to investigate the usability of automated Error Analysis with LLMs for the three task types. This includes automatically creating error classes on errors made from zero-shot and few-show prompts(random examples vs k-nearest examples), and evaluating the error classes for all tasks. The sentiment analysis and price prediction tasks will also be trained with a simple machine learning classifier, such as logistic regression or Naive Bayes. Based on the results, an LLM will be able to create error classes that are comparable to the error classes of the LLM classifier.

If there is, enough time, I also aim to experiment with using the generated error classes in the advanced prompting method of Prompt Breeding. Prompt breeding is a mechanism that evolves and adapts prompts for a given domain, and has been shown to outperform Chain-Of-Thought and other prompt engineering methods [4].

Datasets. I plan to use one dataset per task. The T2D-SM-NH task from the WDC Schema Matching Benchmark will be used for the *Schema-Matching task*. This task requires matchers to find correspondences between Web tables and tables derived from the DBpedia knowledge graph. [5]. The dataset includes 178 pairs of Web-table/DBpedia-table, covering 28 different DBpedia classes, and 421 correspondences between columns of Web and DBpedia tables.[1] The large-scale Multi-Aspect Multi-Sentiment (MAMS) aspect-term sentiment dataset proposed by Jiang et al. will be used for *Aspect Based Sentiment Analysis*. Each sentence in the dataset contains at least two different aspects with different polarities[6]. The training set comprises 3608 entries, with 1089 positive, 892 negative, and 1627 neutral entries.[2] For the *price prediction task* using LLMs, I will use the 'Vehicles' dataset from Kaggle. This dataset includes information on the features and brands of used cars, with 12 independent variables and 8128 entries.[3]

Models. The experiments will use two different models: OpenAI's GPT3.5-Turbo and GPT4. Both models are accessible through OpenAI's API. Additional tools, such as LangChain, may be useful for development.

Workplan. During the initial 1–2 weeks of my research, I aim to become thoroughly familiar with the existing literature, frameworks, libraries, and APIs mentioned above. Moreover, I will also clean the datasets and will manually check that there are no errors in the test sets. I plan to spend approximately 4–6 weeks performing the experiments, including cleaning and preparing the datasets, executing the queries, and evaluating the results. Structured notes will be taken on all results to facilitate their use in writing the thesis. The results will be written up, and the thesis will be finalized over approximately four weeks. Weekly progress reports will be provided to the supervisor, outlining the results of the experiments, challenges, successes, and next steps.

[1] https://webdatacommons.org/structureddata/smb/
[2] https://github.com/siat-nlp/MAMS-for-ABSA/tree/master/data/MAMS-ACSA/raw
[3] https://www.kaggle.com/datasets/nehalbirla/vehicle-dataset-from-cardekho/data?select=Car+details+v3.csv

References

[1] Ralph Peeters and Christian Bizer. Using ChatGPT for Entity Matching, 2023.

[2] Qingyu Lu, Liang Ding, Liping Xie, Kanjian Zhang, Derek F. Wong, and Dacheng Tao. Toward Human-Like Evaluation for Natural Language Generation with Error Analysis, 2022.

[3] Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Jialin Pan, and Lidong Bing. Sentiment Analysis in the Era of Large Language Models: A Reality Check, 2023.

[4] Chrisantha Fernando, Dylan Banarse, Henryk Michalewski, Simon Osindero, and Tim Rocktäschel. Promptbreeder: Self-Referential Self-Improvement Via Prompt Evolution, 2023.

[5] Christian Bizer. WDC Schema Matching Benchmark (SMB), 2023.

[6] Qingnan Jiang, Lei Chen, Ruifeng Xu, Xiang Ao, and Min Yang. A Challenge Dataset and Effective Models for Aspect-Based Sentiment Analysis. pages 6280–6285, Hong Kong, China, 2019. Association for Computational Linguistics.