GPT for Extraction of Biomedical Fields from Clinical Study Texts

Ryan Andrew, Mari Itoh, Masataka Kuroda, Hiroya Takamura, Yayoi Natsume-Kitatani
takamura.hiroya@aist.go.jp
{natsume, m-kuroda, mari, ite001392ryan}@nibiohn.go.jp

Abstract

Generative transformer models are powerful tools used in a wide variety of natural language processing applications. One area in which the application of generative pre-trained transformers (GPT) hold great potential in the field of biomedical research is automated data curation. In this study, we utilized GPTs to systematically extract structured data from scientific articles within PubMed and ClinicalTrials.gov. In our research we used OpenAI's GPT-3 and GPT-4 language models to efficiently populate text fields that have been pre-selected by domain experts in the corresponding biomedical fields and present the outcomes in JSON format. Initial findings suggest that generative transformers such as GPT-3 and GPT-4 hold promise as potent tools for automating data curation in the biomedical domain.

1 Introduction

Scientific papers and databases contain a wealth of knowledge obtained through research to date, and there are high expectations for their utilization for further research activities. However, while information in these resources is described in natural language, data structuring is often required for its utilization. Furthermore, in the case of texts with highly specialized content, such as academic papers, it is difficult for a person who is not an expert in the field to properly extract the necessary knowledge from the text, and the hurdle for manual processing is high. Therefore, there is a high demand for easy and accurate structured data creation from text (defined as data curation in this paper). With this background, the authors have developed a web-based application, BiomedCurator [1], which combines technologies in the field of natural language processing (named entity recognition, entity linking, relation extraction, and text classification). Although BiomedCurator is able to extract information with an accuracy equivalent to SOTA, it is not very versatile because the annotation data used for training covered only the specific field of expertise.

On the other hand, recent rapid progress in LLMs has made it possible to realize data curation by using LLMs, and the fact that the use of LLMs for data curation eliminates the need to create annotation data is a significant advantage. Therefore, in this study, we studied and optimized prompts to properly implement data curation using OpenAI’s GPT language models, and evaluated their performance.

2 Methods

2.1 Prompt Engineering

The focal point of this research is the extraction of fields of interest using a generative transformer model for the purposes of biomedical data curation. The main techniques used in extracting these fields are prompt engineering, providing sample responses and inputs, and specifying the appropriate output format. Each of these was essential in both extracting the appropriate data as well as returning it in a configuration that could easily be transferred to a database.

The prompt engineering phase consisted primarily of providing basic instructions to the model along with a list of fields to be extracted and their corresponding descriptions. Once this was completed, adjustments were made to both the instructions and field descriptions to improve the model’s performance.

Some of the less self-explanatory fields contained sample outputs in addition to or in place of field descriptions in order to provide the model with more information to extract the desired data. Additionally, each of the model self-evaluation scores (further explained in Section 2.2) contained sample inputs along

with the expected evaluation to demonstrate how inputs should be evaluated.

The final piece of prompt engineering was specifying the response format. In order to minimize the amount of preprocessing and formatting that would be required to handle the text responses received from the GPT models, the decision was made to have the outputs formatted in JSON. Additionally, in order to avoid unnecessarily verbose responses – particularly for the fields that can be summed up in a couple of words – the model was instructed to make its response as short as possible. Lastly, in order to mitigate unhelpful formatting patterns that would make dealing with the data more difficult, the model was instructed not to create any nested objects in its JSON response.

```
Identify the following items from the article text:
{feature_descriptions}

The article is delimited with triple backticks.
Format your response as a JSON object with \
{feature_keys} \
as the keys.

If the information isn't present, use "[NA]" \
as the value.

Make your response as short as possible, and do no create
any nested objects in the JSON response.

Article text:
```

Figure 1: Base prompt used for data extraction

```
"drug_therapy": "Drugs or therapies used in the study",
"reference_drug_therapy": "Drugs or therapies referenced in the study",
"treatment_details": "Detailed description of the treatment, including but \
not limited to patient details, drug/therapy, dose/cycles, duration, route, \
schedule, analysis",
"dose": "Dosage of any drugs mentioned",
"route_of_administration": "The method by which the drugs are administered",
"duration": "Duration of the treatment",
"disease_name": "Name of the disease or condition being studied",
"disease_sub_category": "Subtype or stage of the disease",
"Stage": "Stage of the disease, e.g. Stage I, II, III, IV",
"Grade": "Grading of the disease, e.g. Grade I, II, III, IV",
```

Figure 2: Sample data field descriptions provided to the GPT models

2.2 Model Self-Evaluation

Though the primary focus of this research is automated data extraction and curation, a secondary goal was established to assess the capabilities of OpenAI's GPT models in judging the similarity of its own responses compared to human-labeled truth data. If successful, this could lead to a streamlined approach to data curation in the future wherein the need to rely on humans for evaluating AI-extracted data is greatly reduced. Unto this end, a self-evaluation system was implemented which consists of the following labels that were used to score text similarity: "Different," "Match," and "Near Match." Sample inputs and appropriate labels were provided to the model to demonstrate the expected output for each label, and a small subset of the self-evaluations were evaluated by a human to determine the efficacy of the self-evaluation metric.

2.3 Word Mover's Distance

The primary metric used to gauge response similarity was Word Mover's Distance (WMD). "The WMD distance measures the dissimilarity between two text documents as the minimum amount of distance that the embedded words of one document need to 'travel' to reach the embedded words of another document" [2]. It is designed to measure the similarity between two texts based on semantic similarity, utilizing Word2Vec embeddings, rather than strictly overlap-ping vocabulary. The Gensim library was used to generate Word2Vec embeddings based on the text corpus.

With the WMD metric, a lower score indicates higher text similarity and a higher score indicates a higher degree of dissimilarity. The lowest score that can be assigned is zero, which indicates that the responses are identical, but there is no upper limit on the score, and in some cases a score of infinity is calculated which represents zero overlap in the texts.

WMD is particularly useful for evaluating GPT-derived responses compared to human responses, because human-annotated data often uses labels and terminology inferred from the text rather than extracted directly from it, whereas LLMs such as GPT-3 typically do not stray far from the language used in source texts when offering summaries or answering questions posed by the user. In order to effectively evaluate the similarity between these responses something beyond a surface level metric is necessary. Additional metrics used to evaluate response similarity were ROUGE-L, BLEU, and METEOR.

3 Experiments and Results

We conducted experiments on the dataset used in the BiomedCurator project (Sohrab, Mohammad Golam, et al.). We used GPT-3.5 and GPT-4 by OpenAI. In the following, we describe the results of our experiments.

3.1 Self-Evaluation Results

The following trends were observed in the human evaluation of the GPT models' self-evaluation scores.

1. When the ground truth contains some in-formation but the output of GPT is N/A: The result of self-evaluation tends to be "Match" in GPT-3 and "Different" in GPT-4.
2. In the case where the years are different but the numbers themselves are similar, such as 2012 and 2013, it is difficult to be judged as "Different".
3. When there is a difference in the amount of information between the GPT output and the ground truth, GPT-4 tends to output Near Match, while GPT-3 rarely judges Near Match and almost always outputs Match or Different.
4. GPT-4 is closer to human judgment when it comes to self-evaluation.

In addition, many of the fields from which GPT failed to extract knowledge correctly are ambiguously defined, and it is not easy even for humans to judge what to output.

Table 1: Sample responses (GPT-3) compared to human-labeled data. The examples demonstrated here show instances where the model exhibited decent performance on extracting the targeted fields.

Field	GPT_output	Truth_data
drug_therapy	pemetrexed-carboplatin-bevacizumab, paclitaxel-carboplatin-bevacizumab	Pemetrexed+Carboplatin+Bevacizumab, Paclitaxel+Carboplatin+Bevacizumab
reference_drug_therapy	pemetrexed-bevacizumab, bevacizumab	Paclitaxel+Carboplatin+Bevacizumab, Bevacizumab
disease_name	nonsquamous non-small-cell lung cancer	Lung Cancer, Non-Small Cell
disease_sub_category	advanced	Advanced Non Small Cell Lung Cancer
Stage	Stage IIIB/IV	IIIB, IV
Authors	David R. Spigel, MD, Jyoti D. Patel, MD, Craig H. Reynolds, MD, Edward B. Garon, MD, Robert C. Hermann, MD, Ramaswamy Govindan, MD, Mark R. Olsen, MD, PhD, Katherine B. Winfree, PhD, Jian Chen, PhD, Jingyi Liu, PhD, Susan C. Guba, MD, Mark A. Socinski, MD, and Philip Bonomi, MD	Spigel DR, Patel JD, Reynolds CH, Garon EB, Hermann RC, Govindan R, Olsen MR, Winfree KB, Chen J, Liu J, Guba SC, Socinski MA, Bonomi P.

Table 2: Sample responses (GPT-3) compared to human-labeled data. The examples demonstrated here show instances where the model exhibited decent performance on extracting the targeted fields.

Field	GPT_output	Truth_data
Grade	[NA]	[NA]
author_conclusi	[NA]	Yes
study_type	Phase III	Clinical
patient_number_	[NA]	939
phase	[NA]	III

3.2 WMD and Other Metrics

Of the data fields that were extracted in this experiment, a subset was selected which best demonstrates GPT's strengths and shortcomings on the selected task and dataset. Out of the sample papers analyzed for this paper, many contained values that were left empty or N/A by annotators. The selected fields were chosen to minimize the number of N/A values in the human-annotated fields so as to best illustrate the models' capabilities. The fields that contain a value in the form # / # under the WMD column are fields for which all of the assigned WMD scores were either zero (identical responses) or infinity (completely dissimilar responses) and for which the median and average was deemed not to be useful. For these fields the number before the slash indicates the number of zero-assigned scores, and the number following the slash indicates the number of infinity scores (i.e. zero/infinity).

Table 3: Median Scores for GPT-3

Fields	WMD	ROUGE-L F-measure	BLEU	METEOR
drug_therapy	0.0	0.0000	0.0000	0.1587
treatment_details	1.0431	0.0556	0.0392	0.2089
disease_name	0.6204	0.5455	0.3579	0.5333
study_type	0.0	1.0000	0.8409	0.9990
author_conclusion	1.4219	0.0000	0.0000	0.0092
evidence_statement	1.1174	0.0000	0.0000	0.0490
association	1.2883	0.0000	0.0000	0.1721
Title	0.1361	1.0000	0.9854	0.9990
Authors	0.5632	0.4615	0.3346	0.4753
Year	292/12	1.0000	0.0000	0.0000

Table 4: Score Averages for GPT-3

Fields	WMD	ROUGE-L F-measure	BLEU	METEOR
drug_therapy	0.3976	0.2736	0.2201	0.3403
treatment_details	0.9734	0.1419	0.1459	0.2625
disease_name	0.5677	0.5139	0.3508	0.5449
study_type	0.4238	0.6349	0.5722	0.7288
author_conclusion	1.1911	0.0795	0.0000	0.0946
evidence_statement	1.0964	0.0159	0.0285	0.0875
association	1.1036	0.0981	0.0986	0.3046
Title	0.1633	0.9641	0.8953	0.9627
Authors	0.5453	0.4342	0.3288	0.4771
Year	292/12	0.9638	0.0000	0.0016

Table 5: Median Scores for GPT-4

Fields	WMD	ROUGE-L F-measure	BLEU	METEOR
drug_therapy	0.0	0.2222	0.0841	0.3000
treatment_details	0.9714	0.0769	0.0300	0.1565
disease_name	0.618	0.5455	0.3385	0.4337
study_type	0.0	1.0000	0.8409	0.9990
author_conclusion	0.0	1.0000	0.0000	0.9815
evidence_statement	1.0678	0.1429	0.2710	0.4451
association	0.359	0.6000	0.2780	0.4538
Title	0.134	1.0000	0.9871	0.9990
Authors	0.5632	0.4667	0.3293	0.4748
Year	291/13	1.0000	0.0000	0.0000

Table 6: Score Averages for GPT-4

Fields	WMD	ROUGE-L F-measure	BLEU	METEOR
drug_therapy	0.4355	0.3779	0.3033	0.4322
treatment_details	0.9215	0.1486	0.1501	0.2430
disease_name	0.5803	0.4971	0.3276	0.4764
study_type	0.092	0.9141	0.8077	0.9622
author_conclusion	0.0908	0.7374	0.0000	0.7236
evidence_statement	1.0454	0.1464	0.2667	0.4117
association	0.2759	0.4141	0.3574	0.4885
Title	0.1617	0.9579	0.8875	0.9596
Authors	0.55	0.4433	0.3233	0.4746
Year	291/13	0.9577	0.0000	0.0000

4 Discussion

4.1 Model Performance

Overall, GPT-4 appears to outperform GPT-3 on the task of data extraction and summarization. The fields where this can most clearly be seen are the author conclusion and evidence statement fields. The average WMD scores for the author conclusion are 1.1911 and 0.0908 for GPT-3 and GPT-4 respectively, and the average METEOR scores are 0.0946 and 0.7236. For the evidence statement field, the average WMD scores are 1.1174 and 1.0454, and the average METEOR scores are 0.0490 and 0.4117 for GPT-3 and GPT-4 respectively. GPT-4's improved performance over GPT-3 in these fields demonstrates significant promise for its use in summarizing more complex data, as these fields are among the most complex in terms of text content and annotation difficulty, as most of the content is inferred rather than lifted verbatim from the text.

Additionally, other areas such as the Title and Year proved trivially easy for both models, with each of them scoring highly on the ROUGE-L F-measure metric and mostly zeros on the adjusted WMD metric.

4.2 Difficulties and Obstacles

One of the difficult parts of evaluating each model's performance on the task of data extraction and summarization is that none of the metrics alone paints a complete picture. WMD's most glaring shortcoming is that it is unable to accurately compare numerical values by default. Additional code was needed to assign an appropriate score of zero to identical responses that consisted of only numerical values. The same shortcoming is present in the METEOR and BLEU metrics as well. This poses a potential problem as many of the data fields of interest contain numerical values. Though measures can be taken to compensate for this by overriding incorrectly assigned scores in the case of simple float to float or integer to integer comparisons, it is more difficult when the numbers are embedded in a longer text response.

Another issue with the WMD metric is that it defaults to assigning a score of infinity (indicating no response similarity) between two N/A responses when they should in fact be considered identical responses. The GPT models (primarily GPT-3) also had difficulty with the self-evaluation task when it came to comparing N/A values. With WMD however, it is fairly trivial to override the incorrect assignment of an infinity scoring by performing a simple string comparison and assigning a score of zero if both responses are N/A.

4.3 Conclusions and Future Work

As a method for helping to automate the task of data curation for biomedical applications, LLMs show promise, but there is still much work to do before the can be considered reliable enough to utilize without human supervision. In future research, other techniques such as transfer learning and further prompt tuning may help to improve performance. Additionally, future work remains to be done on other GPT models such as Meta AI's LLaMA that do not rely on third-party servers to use, as much data in the biomedical field is sensitive. With continuous improvements however, LLMs could save researchers valuable time and be an invaluable resource in the task of data curation.

References

1. Sohrab, Mohammad Golam, et al. Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing: System Demonstrations. 2022.
2. Kusner, M., Sun, Y., Kolkin, N. & Weinberger, K.. (2015). From Word Embeddings To Document Distances. *Proceedings of the 32nd International Conference on Machine Learning*. 37:957-966 Available from https://proceedings.mlr.press/v37/kusnerb15.html