Few-shot Learning for Biomedical Event Extraction with large generative language models

Expose for a master thesis

Fabio Barth

December 9, 2022

1 Introduction

Biomedical event extraction (BEE) is a complex but important task in natural language processing of biomedical text (BioNLP) [1]. It can be done by extracting and structuring the information out of a biomedical text and generating a graph out of it. In BEE, a model receives a text (typically with entity annotations) as input and generates as output a graph describing the relations between the entities in form of events.
Recent research has shown that the performance of various NLP tasks is improved by transforming information extraction tasks into translation tasks through a special natural language description (NLD) and training generative language models on it [2]. The used extraction models are inherently label-hungry and generalize poorly across domains [3]. The existing biomedical data sets for event extraction are rather small compared to their complexity [4, 5, 6, 7, 8, 9, 10, 11]. The performance of the language models also relies on the given natural language description [12].
Pre-trained large language models (LLM) achieve already comparable results to state of the art models with just a few (5 to 50) training examples (few-shot learning) [2]. Event Extraction is a good task to use few-shot learning (FSL) on because of the small size of the data sets. Still, there is no research about the performance of LLM with few-shot BEE yet.
For this project, we want to train several models with different NLDs with few-shot settings on event extraction. We want to show the importance of choosing the optimal NLD and compare different LLMs on BEE.

1.1 Goals for this Work

In this project, we will explore whether we can improve FSL for BEE by transforming the BEE task into a text-to-text translation task [13] and applying in-context learning with LLMs to it [2]. For this task, a processing pipeline has to be built that creates example prompts to train the large language model and then evaluates the result of the output.

As a language model, we use GPT-3 [2] or a comparable models like BLOOM [14] or OPT [15]. To make use of the GPT-3 model OpenAI provides an API for prompting [2]. The API can be used to experiment on a pre-trained GPT-3 model but it comes with a service charge.
That is why in the Section 3 a cost overview will be provided for the given experiments.

2 Background and Related Work

2.1 Biomedical Event-Extraction

BEE is a complex task in NLP in which a model has to find events in a text with specific types and their related arguments [3]. A significant distinction between BEE and general domain event extraction is that in BEE, often, events have not only entities as arguments but also other events. In other words, the output graph produced by BEE is often strongly connected. In contrast, in general domain event extraction, it is usually a set of small isolated graphs (one per event) [2].

2.2 Structured Prediction as Translation

When working with a generative LLM like GPT-3 the output is plain text. Therefore, the event extraction task has to be transformed into a translation task for extracting the information. John et al. [2] presented a framework called 'Translation between Augmented Natural Languages' (TANL) to solve structured prediction tasks with a generative language model. This framework allows us to solve tasks like relation extraction, named entity recognition, and event extraction, among others. They used a special language description by annotating the input and output sentence to easily extract the information from the sentence.
The language description follows a strict pattern to extract the information easily in the post-processing. The language description is tested for relation extraction, named entity recognition, and simple event extraction tasks [2].
In preliminary experiments, we found that language models like T5 are capable of performing the event extraction task [12]. Many other experiments have shown, that the underlying NLD does affect the performance of the model [2, 12, 16]. We will experiment with different NLDs to find out which NLD does bring the best performance on BEE.

2.3 Learning with LLMs

The goal of few-shot learning (FSL) is to solve a given task with a small number of examples (5 to 50) [3]. While for humans only a few examples are needed to generally perform a new language task, smaller language models like BERT require task-specific data sets of thousands or tens of thousands of examples for fine-tuning [2].
Recent LLMs achieve SOTA performance in zero-shot (no example given) or few-shot settings because of their pre-training process. Those models are pre-trained on a language-modelling task on a large dataset that can contain around 500 billion tokens [17].
LLMs are pre-trained in multiple languages and some are even trained in a programming language to be applicable in a wider range of contexts [14].

Instead of fine-tuning the model, we use in-context learning to train the model on the BEE task.
In-context learning is a special adaptation of few-shot learning, where the model learns a task with just k input-output examples, without optimizing any parameters. This can be done by parsing the input-output examples and the input example on which the task should be performed all in one prompt into the language model. The model then makes a prediction on the last given input sequence based on the other input-output examples in the prompt. Using in-context learning with LLMs has shown SOTA comparable performances in biomedical information extraction (IE) tasks [18].
This learning process makes it faster to test different NLDs without fine-tuning the whole model. It also is more applicable to the BEE task because of the smaller size of the data sets.

2.4 GPT-3

GPT-3 is an autoregressive language model that achieves strong performance on many NLP data sets in few-shot settings [2]. It is trained on roughly 300 billion tokens from the Common Crawl data set [19] and has 175 billion parameters. It achieves strong performance on many NLP datasets, including translation [2]. That's why using GPT-3 for event extraction on biomedical text could be promising. In preliminary experiments, we found that GPT-3 is capable of translating into the proposed natural language descriptions with just three or four given examples.

3 Approach

The proposed project consists of two parts. The first part aims at implementing a processing pipeline for creating input examples for the LLM. Therefore, a data parser from BigBio will be used to download and parse the data sets [20]. We use three data sets, from several academic authors who provided those data sets for different shared tasks between 2011 and 2013 [4, 6, 9]. The processing pipeline builds input and output sequences from the training data sentences and input sequences from development set sentences. When using the TANL natural language description the sequences are annotated as follows:
In the input sequences, the given gold entities are marked with special tokens [ ]. In the output sequence, the event trigger is marked with the same special tokens. The output sequence is a gold prediction for the single example. Every input and output sequence is also labeled with a special head token. The token is either 'Input' or 'Output' depending on the sequence. For every development set sequence, an input prompt will be built. The structure of the input prompt is as follows:

- A head token that specifies the task of the input prompt.
- A number k of training examples with input and its corresponding output sequence from the training data.
- the input sentence from the development set

The prompt is depending on how many previous examples the model should see before predicting the output sequence. Figure 1 shows an example of a single input prompt with $k = 2$ as the number of examples.

Translate:
Example: The dephosphorylated [Axin|Gene_or_gene_product] binds [beta-catenin|Gene_or_gene_product] less efficiently than the phosphorylated form.
Output: The [dephosphorylated|Dephosphorylation|Axin=Theme] Axin [binds|Binding|Axin=Theme|beta-catenin=Theme] beta-catenin less efficiently than the [phosphorylated|Phosphorylation|Axin=Theme] form.

Example: Using a version of [Ace2p|Gene_or_gene_product] tagged with a [c-myc|Gene_or_gene_product] epitope, we show that the protein is excluded from the [nucleus|Cellular_component] of cells during most phases of the mitotic cell cycle.
Output: Using a version of Ace2p tagged with a c-myc epitope, we show that the protein is excluded from the nucleus of cells during most phases of the mitotic [cell cycle|Pathway].

Example: [E2F1|Gene_or_gene_product] uses the [ATM|Gene_or_gene_product] signaling pathway to induce [p53|Gene_or_gene_product] and [Chk2|Gene_or_gene_product] phosphorylation and apoptosis.
Output:

Figure 1: An example prompt with sentences from [9] with the input example and output generation. 'Translate' is the head token that indicates the model that the following input task is a translation task. 'Example:' is the head token of every input sequence to label those sentences as inputs and 'Output:' labels all gold sentences as output examples. The last 'Output:' example ist the one that the model should generate.

After extracting the information from the model's output text, the results will be evaluated with the standard metrics of precision, recall, and F1 [4].
As already mentioned accessing GPT-3 via its API comes with charges. Depending on the model version one thousand input tokens cost between 0,0004 and 0,02 $ Dollars. The average token number of all sentences in the three given data sets is 37 tokens. So if we want to test three different data sets we would have roughly around 37 tokens per sentence and about 12000 sentences [6, 9, 4]. Predicting all development sentences in a k shot setting with $k = 10$ examples (one example is an input and the corresponding output sequence) would cost roughly 100 $.
The experiments will be done with various input example combinations from the test set and various numbers of sequence inputs from the development set.
Our research will attempt to answer the following questions:

- Can we train large language model for BEE with in-context learning in a few-shot setting?
- How do different NLDs affect the performance of the LLM?
- How good is the performance of GPT-3 compared to other LLMs (Galactica [21], Bloom [14], OPT [15], or GPT-J)?

With the first research question, we want to prove that it is possible to train an LLM on BEE with in-context learning in few-shot settings. We, therefore, use the same NLD as in the TANL framework [2] and train an LLM that is comparable with the GPT-3 structure on the BEE task.
In our second experimental set-up, we change the NLD to figure out the importance of the NLD when working with a generative language model. We want to test at least two additional NLDs.
And for our last topic, we want to compare the GPT-3 model with other open-source large language models. We, therefore, want to use at least two state-of-the-art LLMs and train them on BEE with in-context on the best working NLD with few shot settings, as

well as GPT-3.

References

[1] Sophia Ananiadou, Paul Thompson, Raheel Nawaz, John McNaught, and Douglas B. Kell. Event-based text mining for biology and functional genomics. *Briefings in Functional Genomics*, 14(3):213–230, 06 2014.

[2] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.

[3] Rui Feng, Jie Yuan, and Chao Zhang. Probing and fine-tuning reading comprehension models for few-shot event extraction. *CoRR*, abs/2010.11325, 2020.

[4] Tomoko Ohta, Sampo Pyysalo, Rafal Rak, Andrew Rowley, Hong-Woo Chun, Sung-Jae Jung, Sung-Pil Choi, Sophia Ananiadou, and Jun'ichi Tsujii. Overview of the pathway curation (PC) task of BioNLP shared task 2013. In *Proceedings of the BioNLP Shared Task 2013 Workshop*, pages 67–75, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

[5] Jung-jae Kim, Xu Han, Vivian Lee, and Dietrich Rebholz-Schuhmann. GRO task: Populating the gene regulation ontology with events and relations. In *Proceedings of the BioNLP Shared Task 2013 Workshop*, pages 50–57, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

[6] Jin-Dong Kim, Yue Wang, and Yamamoto Yasunori. The Genia event extraction shared task, 2013 edition - overview. In *Proceedings of the BioNLP Shared Task 2013 Workshop*, pages 8–15, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

[7] Sampo Pyysalo, Tomoko Ohta, and Sophia Ananiadou. Overview of the cancer genetics (CG) task of BioNLP shared task 2013. In *Proceedings of the BioNLP Shared Task 2013 Workshop*, pages 58–66, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

[8] Sampo Pyysalo, Tomoko Ohta, Rafal Rak, Dan Sullivan, Chunhong Mao, Chunxia Wang, Bruno Sobral, Jun'ichi Tsujii, and Sophia Ananiadou. Overview of the infectious diseases (ID) task of BioNLP shared task 2011. In *Proceedings of BioNLP Shared Task 2011 Workshop*, pages 26–35, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.

[9] Jin-Dong Kim, Yue Wang, Toshihisa Takagi, and Akinori Yonezawa. Overview of genia event task in bionlp shared task 2011. In *Proceedings of the BioNLP Shared Task 2011 Workshop*, BioNLP Shared Task '11, page 7–15, USA, 2011. Association for Computational Linguistics.

[10] Tomoko Ohta, Sampo Pyysalo, and Jun'ichi Tsujii. Overview of the epigenetics and post-translational modifications (EPI) task of BioNLP shared task 2011. In *Proceedings of BioNLP Shared Task 2011 Workshop*, pages 16–25, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.

[11] Jin-Dong Kim, Tomoko Ohta, Sampo Pyysalo, Yoshinobu Kano, and Jun'ichi Tsujii. Overview of BioNLP'09 shared task on event extraction. In *Proceedings of the*

BioNLP 2009 Workshop Companion Volume for Shared Task, pages 1–9, Boulder, Colorado, June 2009. Association for Computational Linguistics.

[12] Fabio Barth. Evaluating biomedical event extraction on generative language models [unpublished semester project]., 2022.

[13] Giovanni Paolini, Ben Athiwaratkun, Jason Krone, Jie Ma, Alessandro Achille, Rishita Anubhai, Cicero Nogueira dos Santos, Bing Xiang, and Stefano Soatto. Structured prediction as translation between augmented natural languages. 2021.

[14] Bigscience large open-science open-access multilingual language model. https://bigscience.huggingface.co/blog/bloom. Accessed: 2022-07-30.

[15] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. Opt: Open pre-trained transformer language models, 2022.

[16] Giacomo Frisoni, Gianluca Moro, and Lorenzo Balzani. Text-to-text extraction and verbalization of biomedical event graphs. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2692–2710, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics.

[17] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer, 2019.

[18] Bernal Jiménez Gutiérrez, Nikolas McNeal, Clay Washington, You Chen, Lang Li, Huan Sun, and Yu Su. Thinking about gpt-3 in-context learning for biomedical ie? think again, 2022.

[19] Adam Roberts, Colin Raffel, Katherine Lee, Michael Matena, Noam Shazeer, Peter J. Liu, Sharan Narang, Wei Li, and Yanqi Zhou. Exploring the limits of transfer learning with a unified text-to-text transformer. Technical report, Google, 2019.

[20] Jason Alan Fries, Leon Weber, Natasha Seelam, Gabriel Altay, Debajyoti Datta, Samuele Garda, Myungsun Kang, Ruisi Su, Wojciech Kusa, Samuel Cahyawijaya, Fabio Barth, Simon Ott, Matthias Samwald, Stephen Bach, Stella Biderman, Mario Sänger, Bo Wang, Alison Callahan, Daniel León Periñán, Théo Gigant, Patrick Haller, Jenny Chim, Jose David Posada, John Michael Giorgi, Karthik Rangasai Sivaraman, Marc Pàmies, Marianna Nezhurina, Robert Martin, Michael Cullan, Moritz Freidank, Nathan Dahlberg, Shubhanshu Mishra, Shamik Bose, Nicholas Michio Broad, Yanis Labrak, Shlok S Deshmukh, Sid Kiblawi, Ayush Singh, Minh Chien Vu, Trishala Neeraj, Jonas Golde, Albert Villanova del Moral, and Benjamin Beilharz. Bigbio: A framework for data-centric biomedical natural language processing, 2022.

[21] Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. Galactica: A large language model for science, 2022.