Evaluating Biomedical Event Extraction on a generative language model using TANL

Expose for student research project

Fabio Barth

May 31, 2022

1 Introduction

Every year more than 1 million biomedical papers are published on the database PubMed [1]. Those scientific papers contain information regarding medical and biomedical research, for instance about protein interactions and the usage of drugs for specific diseases. A main goal in natural language processing of biomedical text is to extract the most useful information and provide it for everyone in a useful format. This, for example, can be done by structuring the information and generating a graph out of it.
Previous works used text-to-graph translation models for this task [2]. But those models had two major drawbacks. The first problem is that they are not scalable to large corpora. The information density of biomedical text can be much higher than in any other text and that caused issues in building the graph. The second major drawback is that the used framework is based on discriminative models. Those models do not produce output text rather than just output labels. This implies that the output-graph is limited by the given labels. One solution to those two problems is to use a generative model that produces output text in a natural language. Those models are scalable for large corpora with high information density and predict output strings for given input text [3].
[4] presented a Framework called Translation between Augmented Natural Languages (TANL) to solve many structured prediction tasks. This framework generates sentences for solving tasks like relation extraction, named entity recognition, event extraction and many more. They used a text-to-text transfer transformer (T5) model as basis for their framework to make it applicable and scalable for many tasks. They also used a natural language description with special annotations as output to keep the results as close to the natural English language as possible. The graph can be build by adapting the language description for the event extraction task and post process the output string from the model. The model has shown good performance on previous event extraction tasks. That's why this framework seems well suited for using it for information extraction on biomedical text [4].

1.1 Goals for this Work

For this project, the TANL framework will be evaluated on event extraction on biomedical text. Therefore, the framework will be embedded in a processing pipeline to parse biomedical text and construct a graph on top of sentences from biomedical publications. Many biomedical data sets are too small to fine-tune a model of that size [3]. Therefore, we propose to use multiple data sets for the same information extraction task. That can be done by pre-processing the corpora for the task into input sequences in a special language description. The result will then be evaluated and compared to the state of the art (SOTA).

The main goal of this project is to figure out if a pre-trained language model is applicable to the task of extracting events in text and whether using multiple data sets at the same time in a multitask fashion is helpful.

2 Background and Related Work

Biomedical event extraction is a complex task in natural language processing with the goal to find events in a text with specific types and their arguments [5]. Events cannot have only entities as arguments but also other events. That's why the output structure of a predicted text is more comparable to a graph than to a simple set of labels as in relation extraction. In the following, I will explain the event extraction task itself and how the TANL framework could be used to get more accurate event predictions.

I will then take a closer look at the TANL framework itself and explain how they used the T5 model to frame a task of translation between Augmented Natural Languages. I will analyse the used language description and compare it to previous language description for the same task.

2.1 Event Extraction

Biomedical text provides important information about protein, gene or drug interactions among others. Those interactions can be interpreted as events. In event extraction the goal is to extract those events from a given text with the entities in each sentence provided as input. An event $e_i \in E$ has a specific type and trigger $t_i \in T$, which is a text span indicating the event (e.g. 'regulates' for a Regulation event) and arbitrarily many arguments $a_i \in A \cup E$. The type of an event describes the interaction between the arguments. Many SOTA frameworks based on discriminative models use an iterative process to first detect event triggers, then predict trigger-argument arcs and finally separate the predictions into distinct events. The TANL framework transforms the event extraction problem into a translation problem. The goal is to extract all events and their arguments in one sentence.

2.2 TANL

Translation between Augmented Natural Languages (TANL) describes the task of translating natural language into an augmented language description and decodes the text into a structured object [6]. The structured object could be, for instance, a graph that represents relations between entities inside the input text.

Therefore, TANL seems to be a useful framework for event extraction. By framing the event extraction task as a translation task, the TANL model could be used to predict relations, events and entities inside of text.
TANL is built with the pre-trained generative language model T5. The results on event extraction and named-entity recognition are comparable to other frameworks [6]. Those results look promising to use the framework for information extraction on biomedical text.
But not only the huge generative language model improves the framework performance. TANL uses an output language description that transfers knowledge about the label semantics. The language description improves the output performance significantly in the few-shot regime [6].

2.3 Natural Language Descriptions

When working with a generative language model instead of discriminative models the impact of the output language description may influence the accuracy and performance of the model. The input format of the corpora and the trained output format may have an impact on the performance of the generative models [6]. The choice of those formats are therefore important to the learning process. The TANL framework is trained on an annotated natural language format.
The language description they used is easy to read for people who are capable of understanding the English language. The authors tried to stay as close as possible to the English language. The model annotates just the task specific parts in the output text inside the sentences. This allows the model to transfer latent knowledge of the task from its pre-training [4]. This leads to an improving performance in the small-data regime.
Another benefit of using a natural language description with annotations like in TANL is that it is easy to understand and to read for humans. Reading information out of the output text is not harder than reading the input text itself. It also can be easily translated in a graph notation and is comparable to other language descriptions.
The language description follows a strict pattern to extract the information easily. The language description can be adapted to multiple tasks. The possible relations and entities are enclosed by the special tokens []. A found entity is described by adding the token from the text with a |-separator and the predicted entity. A relation between two entities X and Y is described as equation $X = Y$.

3 Approach

The first part aims at implementing a text-to-graph transducer with TANL. The framework repository is on github and will not be modified. For the experiments a pre-trained T5-base model is used to keep the framework comparable to related work and the original framework. The goal is to build a pre- and post-processing pipeline for parsing the data and evaluate the performance.
As already explained the model needs as much data as possible for the training to be comparable to SOTA results. Therefore, data parser from the Huggingface framework will be used to download and parse the data sets. We use seven data sets, from several academic authors who provided those data sets for diferent shared tasks between 2011 and 2019 [7, 8, 9, 10, 11, 12, 13, 14]. Huggingface created together with roughly 70 contributors a framework called BigBio to enable easy programatic access to over 120 biomedical data sets [15]. All of the mentioned data sets are included in the framework

and the data can be extracted and parsed in a unified schema. By using that schema all given event extraction corpora can be used to train a single model without building a new pre-processing pipeline for every data set individually [16]. As baslines, we will use the BERT based SOTA framework DeepEventMine [17] and a framework that used a generative model with multi-turn question answering for the event extraction task [18], to which we will compare with the standard metrics of precision, recall and F1 [7].

The general approach for this project will be to build a processing pipeline for the TANL framework and evaluate the results on a large scale data set. The model itself will be based on the pre-trained TANL model which is a T5-base model.

For the processing pipeline four main functions have to be created. An input parser, an output parser, an example creator and an evaluation function have to be written. The input parser will download the data sets and format the corpus for the TANL framework. Therefor, input examples on sentence level will be created with the example creator function. The input text will be modified by adding some extra tags that describe the task and the data set. The entities will be also marked as described in Section 2.3. If the input text is longer than 512 tokens the text will be split into smaller chunks.

The output parser produces output examples for fine tuning the model. Every input sentence will be modified by adding the event annotations. The annotations follow the principles of the given language description by the TANL framework. The output example sentences will then be used to evaluate the predicted output sentence of the TANL model by the given input sentence.

For the evaluation, the evaluation sets of all seven corpora will be used. Every data set will be trained and evaluated separately on the training and evaluation set. The output will be evaluated by extracting the features from the output sentence and rebuild the event graph. The event graph will then be compared to a gold standard. Therefore, the sub-graph isomorphism task for the two graphs has to be solved and evaluated.

The model will be trained on all seven training corpora from the bioNLP shared tasks between 2011 and 2019 separately [7, 8, 9, 10, 11, 12, 13, 14]. To compare the training results with the given baseline we use also all seven evaluation corpora to evaluate the biomedical event extraction with TANL.

References

[1] National Library of Medicine. Scientific literature: Information overload, howpublished = https://www.nature.com/articles/nj7612-457a, note = Accessed: 2021-09-06.

[2] Qipeng Guo, Zhijing Jin, Xipeng Qiu, Weinan Zhang, David Wipf, and Zheng Zhang. Cyclegt: Unsupervised graph-to-text and text-to-graph generation via cycle training, 2020.

[3] 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. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

[4] 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.

[5] Jiangwei Liu, Liangyu Min, and Xiaohong Huang. An overview of event extraction and its applications. *CoRR*, abs/2111.03212, 2021.

[6] Hiroaki Ozaki, Gaku Morio, Yuta Koreeda, Terufumi Morishita, and Toshinori Miyoshi. Hitachi at MRP 2020: Text-to-graph-notation transducer. In *Proceedings of the CoNLL 2020 Shared Task: Cross-Framework Meaning Representation Parsing*, pages 40–52, Online, November 2020. Association for Computational Linguistics.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] Seelam Natasha, Fries Jason, Alan, Kang Sunny, MS, Su Rosaline, Altay Gabriel, Weber Leon, Datta Debajyoti, and Garda Samuele. Bigscience biomedical nlp hackathon. https://github.com/bigscience-workshop/biomedical, 2022.

[16] GENCI Huggingface and IDRIS. Bigscience: A one-year long research workshop on large multilingual models and datasets, 2020.

[17] Hai-Long Trieu, Thy Thy Tran, Khoa NA Duong, Anh Nguyen, Makoto Miwa, and Sophia Ananiadou. Deepeventmine: end-to-end neural nested event extraction from biomedical texts. *Bioinformatics*, 36(19):4910–4917, 2020.

[18] Xing David Wang, Leon Weber, and Ulf Leser. Biomedical event extraction as multi-turn question answering. In *Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis*, pages 88–96, 2020.