Distant Supervision for Large-Scale Biomedical Event Extraction and Knowledge Base Population

Exposé for master thesis

Xing Wang

September 3, 2020

1 Introduction and Objective

Event extraction together with Named Entity Recognition (NER) and Named Entity Normalization (NEN) is an important task in biomedical Natural Language Processing (NLP). NER deals with identifying physical entities, i.e., proteins, genes and other chemicals, in text, NEN links found entity mentions to entities stored in a database and Event Extraction tries to infer the relationships between the found entities. Different relationship and event types are distinguished by their respective event triggers which are often denoted by verbs, e.g., *phosphorylates* or *catalyzes*. Entities are often denoted by proper nouns like *glucose* or *ATP*. Common event extraction systems learn from annotated documents, usually PubMed abstracts and PMC open access full texts, where the event structures, their participants and their positions in a text are explicitly highlighted. Examples of event extraction systems are TEES [Björne, 2014, Björne and Salakoski, 2018] and EventMine [Miwa and Ananiadou, 2013, Trieu et al., 2020]. The amount of available biomedical data in both text documents and curated data and knowledge bases grows at a fast rate [Gonzalez et al., 2015]. Thus, manual annotation of all documents is not feasible and gold standard data sets for event extraction are rather small. Hence, we want to follow a distant supervision approach in this work and learn from knowledge bases and the plain documents instead of relying on manually labeled texts. After learning to

identify event structures from text we want to populate knowledge bases with the newly found event structures.

2 Material and Methods

2.1 Corpora

In this work, we focus on domain data from biological pathways, i.e., metabolic and signaling pathways, and molecular and genetic interactions on a cell level. We use the Pathway Commons database [Cerami et al., 2010] which includes about 5.000 pathways and more than 2.3 million interactions in 2019 [Rodchenkov et al., 2020]. The data is stored in the BioPAX [Demir et al., 2010] format, a standard to ensure compatibility between different biological databases. We split the database, the known physical entities and their interactions into a training set (for training and development) and a test set (for prediction/evaluation). For implementation, we will use pybiopax[1] to access and transform pathway data in the BioPAX format.

Our text corpus consists of all PubMed abstracts and PMC full texts as of May 2020 (April 2013 when comparing to the EVEX baseline [Van Landeghem et al., 2013]).

2.2 Baseline

EVEX[2] [Van Landeghem et al., 2013] is a large-scale biomedical event extraction system trained on directly supervised data. Van Landeghem et al. [2013] apply a pipeline of NER, NEN and event extraction to all PubMed abstract and PMC full texts from April 2013. They use TEES [Björne, 2014] as an event extraction system which has been trained on various BioNLP shared tasks like GENIA [Kim et al., 2011] and Epigenetics and Post-translational Modifications (EPI) [Pyysalo et al., 2012].

[1] https://github.com/indralab/pybiopax
[2] http://www.evexdb.org

2.3 Question Answering for Event Extraction

Our basic model follows the outline of Chen et al. [2017] who apply large scale question answering on the English Wikipedia[3]. The event extraction task is split into document retrieval and machine reading as question answering. In the first subtask of document retrieval, relevant documents to the question at hand are determined by a retrieval system. In a second step, relevant text evidence from the chosen documents is determined as answer to a given question. Using the information obtained from multiple questions and their answers, we form biomedical event structures. Use of question answering for multi-task learning has been popularized in the recent years [McCann et al., 2018] and been applied to entity-relation extraction by Li et al. [2019]. In our previous work [Wang et al., 2020], we have already successfully applied question answering to biomedical event extraction in presence of directly supervised data.

We apply multi-turn question answering as described by Wang et al. [2020] to form the biomedical event structures. Starting from a physical entity as an event theme, the central subject of an event, we query for corresponding event triggers and corresponding entities as event arguments. Assume the ordered tuple *(conversion, PEP, pyruvate)* from a knowledge base is given and we want to find occurrences of the tuple in text and learn how to detect it. In our question answering model, we learn this with two related questions. In our first question, we ask for event triggers: *What are events involving PEP?*. After having found the event triggers and its event type, we ask for event arguments in our second question: *What are products for the conversion of PEP?*. After having extracted the corresponding product we have found the full *conversion* event structure with theme, product and its corresponding text evidence.

2.4 Model overview

Preprocessing and document retrieval are conducted in the same way for both training and evaluation. For event extraction and sequence labeling, training and evaluation differ: First, a model using a distantly supervised dataset is learned during training and then applied to predict new events during evaluation.

[3] `https://en.wikipedia.org/`

Preprocessing

In text, multiple words may describe the same physical entity in forms of synonyms, e.g., *Na+* and *natrium* refer to the same entity. Hence, we perform NER and NEN on our PubMed corpus with PubTator Central (PTC) [Wei et al., 2019] as a preprocessing step to identify biomedical entities. We replace all entity mentions in texts by their normalized forms, i.e., PTC returns NCBI Gene identifiers to genes and proteins mentioned in a text and we replace these mentions by a common normalized term. In the example of the chemical from Section 2.3, we normalize the synonyms *Phosphoenolpyruvic acid retrieval* and *PEP* to the same name *PEP*. PTC provides their precomputed annotations online, so no additional expenditure of time is needed for NER and NEN. Subsequent document retrieval and event extraction steps are then conducted on the preprocessed and normalized text corpus. For document retrieval, we also apply stemming and stop word removal for faster retrieval performance.

The physical entities as found in the Pathway Commons database do not need to be normalized as they already appear in a common form in the database. We only need to ensure consistent naming between the entities mentioned in the normalized text corpus and in the database, i.e., linking protein IDs from UniProt in Pathway Commons to the corresponding NCBI Gene identifiers and using the same terms for each entity in the subsequent retrieval and question answering steps.

Document Retrieval

EVEX [Van Landeghem et al., 2013] iterates over all given text documents and conducts event extraction on each of them. As the document corpus grows bigger, the computing time required also grows substantially. Instead of applying our event extraction model to the whole corpus, we make use of a retrieval step to filter relevant documents to our given query. For our information retrieval system, we rely on Apache Lucene[4][5] like Yang et al. [2019] and Xie et al. [2020]. Using Lucene with a standard retrieval model, e.g. Okapi BM25, requires no training steps and can be directly applied to given corpora. We use standard Lucene query syntax using the entity keywords from the verbose question described in Section 2.3. For event types and event triggers, we perform automatic synonym expansion to ensure that all events of a given type are found. For entities,

[4]https://lucene.apache.org/
[5]https://lucene.apache.org/pylucene/

synonym expansion is not needed due to the already performed normalization in Section 2.4.

During both training and evaluation, at the start we are only given physical entities as found in the Pathway Commons database. Considering the physical entity *PEP* in the first retrieval step for example, we aim to find all relevant documents containing it. After having found all potential event types associated to *PEP* in the subsequent event extraction step, we repeat a retrieval step once more to query for new documents containing PEP in combination with the just found event types. For instance, having found conversion as an event type involving *PEP* during the event extraction step, a following retrieval query in Lucene syntax for documents involving conversion events with *PEP* would be *PEP AND (conversion OR metabolism OR ...)*. Document retrieval and event extraction alternate during training and evaluation until all event structures revolving around a given physical entity are found.

We will experiment with retrieving documents on different granularity levels, i.e., whole documents, paragraphs or single sentences. Depending on the performance of the document retrieval in preliminary experiments/evaluations we will also test retrieval systems using neural networks as Das et al. [2018], Seo et al. [2019] or Lee et al. [2019].

Event Extraction - Training

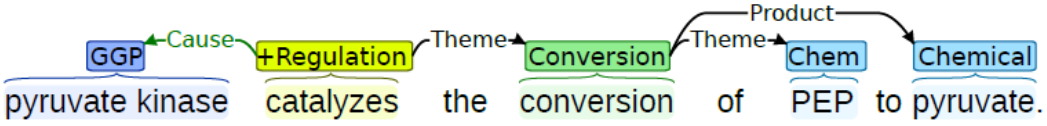

Figure 1: Event representation for a conversion reaction [Ohta et al., 2013]

After having retrieved relevant documents to a question, we use distant supervision [Mintz et al., 2009] to automatically generate labeled training sequences for the question answering step. Given the first question of the example from Section 2.4, and the retrieved document depicted in Figure 1 for the query and knowledge base tuple *(conversion, PEP, pyruvate)* from Section 2.3, we mark all conversion triggers as true answer tokens to be learned, i.e., the word *conversion* is marked as an answer in our example. The sequence labeling model follows IOB2-style where tokens belonging to an answers are labeled with

B and *I* and other tokens are labeled *O*.

Distant supervision is often prone to noisy labeling [Riedel et al., 2010, Surdeanu et al., 2012] as the marked text mentions may not express the given relation. Multi Instance Learning (MIL) aims to solve this problem by relaxing the assumption that all marked text mention express the relation. In MIL, the at-least-once assumption is introduced such that only one of all text mentions must express the relation. A noisy annotation for the knowledge base tuple *(conversion, PEP, pyruvate)* is a text snippet containing *conversion ... PEP*, where the conversion event does not actually refer to *PEP*.

In our sequence labeling setting, we assume that the distantly supervised annotation for at least one retrieved document is correct. We aggregate the label probabilities for each token in a sequence to one single sequence probability using a Conditional Random Field [Hakala and Pyysalo, 2019]. Then, we aggregate the sequence probabilities to one final score using a softmax-like operator [Bansal et al., 2020] for MIL. The loss obtained from this score can be backpropagated through the neural network to learn and update the model weights.

In the question answering model, multiple answers could be valid, e.g., there exists another valid product *C* after a conversion of *P*. If both *pyruvate* and another product *C* are present in the knowledge base, we would mark both mentions as valid answers. For MIL, we aggregate all sequences containing *pyruvate* and *C* in two separate training steps to ensure that both events are learned.

Event Extraction - Evaluation

Like for training, we use the same PubMed abstracts and PMC full texts for evaluation. The Pathway Commons knowledge base with its physical entities (genes and proteins) and their interactions (events) is where training and evaluation data are split up. Whereas both physical entities and the interactions are known during training, only the physical entities are given during evaluation. The interactions have to be predicted and are compared to the stored gold interactions. Sequence labeling during prediction differs from the training steps where we have automatically labeled the retrieved documents given the known interactions from the knowledge base. During evaluation, we apply our learned event extraction model to each found document from the retrieval step and label

the sequences according to it. The extracted answers from all question turns in the event extraction model are then gathered and merged into biomedical event structures, i.e., BioPAX/Pathway Commons interactions and pathways. For our example from Section 2.4, if the physical entity *PEP* and its interaction *(conversion, PEP, pyruvate)* belonged to the evaluation set instead of the training set, then the physical entity *PEP* would be the sole input and we would need to extract the interaction containing the entity *pyruvate* and the event type *conversion* from our pipeline.

We evaluate on two different data subsets from the Pathway Commons knowledge base and the PubMed Corpus. When comparing to the EVEX baseline [Van Landeghem et al., 2013], we limit the PubMed corpus till 2013 (c.f. Section 2.2) and only include the physical entities from Pathway Commons which appear in that PubMed corpus and their corresponding interactions. For the full evaluation mode, we incorporate the whole PubMed corpus and Pathway Commons as of 2020.

For our evaluation format, we use the BioNLP .a* event extraction format [Kim et al., 2011, Ohta et al., 2013]. In contrast to the token-level perspective used in the BioNLP event extraction tasks, we evaluate from a knowledge base perspective which does not need to integrate the positional information of extracted events in a given document.

For the implementation of the event extraction and question answering task, we use the Transformers library by Wolf et al. [2019] and the pretrained neural network models BERT [Devlin et al., 2018] and SciBERT [Beltagy et al., 2019]. In order to compute the attention in long sequences and documents in BERT, we use Longformer [Beltagy et al., 2020].

References

Trapit Bansal, Patrick Verga, Neha Choudhary, and Andrew McCallum. Simultaneously linking entities and extracting relations from biomedical text without mention-level supervision. In *AAAI*, pages 7407–7414, 2020.

Iz Beltagy, Kyle Lo, and Arman Cohan. Scibert: Pretrained language model for scientific text. In *EMNLP*, 2019.

Iz Beltagy, Matthew E Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020.

Jari Björne. *Biomedical Event Extraction with Machine Learning*. PhD thesis, University of Turku, 2014.

Jari Björne and Tapio Salakoski. Biomedical event extraction using convolutional neural networks and dependency parsing. In *Proceedings of the BioNLP 2018 workshop*, pages 98–108, 2018.

Ethan G Cerami, Benjamin E Gross, Emek Demir, Igor Rodchenkov, Özgün Babur, Nadia Anwar, Nikolaus Schultz, Gary D Bader, and Chris Sander. Pathway commons, a web resource for biological pathway data. *Nucleic acids research*, 39(suppl_1):D685–D690, 2010.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. Reading wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879, 2017.

Rajarshi Das, Shehzaad Dhuliawala, Manzil Zaheer, and Andrew McCallum. Multi-step retriever-reader interaction for scalable open-domain question answering. In *International Conference on Learning Representations*, 2018.

Emek Demir, Michael P Cary, Suzanne Paley, Ken Fukuda, Christian Lemer, Imre Vastrik, Guanming Wu, Peter D'eustachio, Carl Schaefer, Joanne Luciano, et al. The biopax community standard for pathway data sharing. *Nature biotechnology*, 28(9):935–942, 2010.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Graciela H Gonzalez, Tasnia Tahsin, Britton C Goodale, Anna C Greene, and Casey S Greene. Recent advances and emerging applications in text and data mining for biomedical discovery. *Briefings in bioinformatics*, 17(1):33–42, 2015.

Kai Hakala and Sampo Pyysalo. Biomedical named entity recognition with multilingual bert. In *Proceedings of The 5th Workshop on BioNLP Open Shared Tasks*, pages 56–61, 2019.

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*, pages 7–15. Association for Computational Linguistics, 2011.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. Latent retrieval for weakly supervised open domain question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6086–6096, 2019.

Xiaoya Li, Fan Yin, Zijun Sun, Xiayu Li, Arianna Yuan, Duo Chai, Mingxin Zhou, and Jiwei Li. Entity-relation extraction as multi-turn question answering. *arXiv preprint arXiv:1905.05529*, 2019.

Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. The natural language decathlon: Multitask learning as question answering. *arXiv preprint arXiv:1806.08730*, 2018.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1003–1011, 2009.

Makoto Miwa and Sophia Ananiadou. Nactem eventmine for bionlp 2013 cg and pc tasks. In *Proceedings of the BioNLP Shared Task 2013 Workshop*, pages 94–98, 2013.

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, 2013.

Sampo Pyysalo, Tomoko Ohta, Rafal Rak, Dan Sullivan, Chunhong Mao, Chunxia Wang, Bruno Sobral, Jun'ichi Tsujii, and Sophia Ananiadou. Overview of the id, epi and rel tasks of bionlp shared task 2011. In *BMC bioinformatics*, volume 13, page S2. Springer, 2012.

Sebastian Riedel, Limin Yao, and Andrew McCallum. Modeling relations and their mentions without labeled text. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 148–163. Springer, 2010.

Igor Rodchenkov, Ozgun Babur, Augustin Luna, Bulent Arman Aksoy, Jeffrey V Wong, Dylan Fong, Max Franz, Metin Can Siper, Manfred Cheung, Michael Wrana, et al. Pathway commons 2019 update: integration, analysis and exploration of pathway data. *Nucleic acids research*, 48(D1):D489–D497, 2020.

Minjoon Seo, Jinhyuk Lee, Tom Kwiatkowski, Ankur Parikh, Ali Farhadi, and Hannaneh Hajishirzi. Real-time open-domain question answering with dense-sparse phrase index. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4430–4441, 2019.

Mihai Surdeanu, Julie Tibshirani, Ramesh Nallapati, and Christopher D Manning. Multi-instance multi-label learning for relation extraction. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*, pages 455–465, 2012.

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*, 2020.

Sofie Van Landeghem, Jari Björne, Chih-Hsuan Wei, Kai Hakala, Sampo Pyysalo, Sophia Ananiadou, Hung-Yu Kao, Zhiyong Lu, Tapio Salakoski, Yves Van de Peer, et al. Large-scale event extraction from literature with multi-level gene normalization. *PloS one*, 8 (4):e55814, 2013.

Xing David Wang, Leon Weber, and Ulf Leser. Biomedical event extraction as multi-turn question answering. *(submitted)*, 2020.

Chih-Hsuan Wei, Alexis Allot, Robert Leaman, and Zhiyong Lu. Pubtator central: automated concept annotation for biomedical full text articles. *Nucleic acids research*, 47(W1):W587–W593, 2019.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.

Yuqing Xie, Wei Yang, Luchen Tan, Kun Xiong, Nicholas Jing Yuan, Baoxing Huai, Ming Li, and Jimmy Lin. Distant supervision for multi-stage fine-tuning in retrieval-based question answering. In *Proceedings of The Web Conference 2020*, pages 2934–2940, 2020.

Wei Yang, Yuqing Xie, Aileen Lin, Xingyu Li, Luchen Tan, Kun Xiong, Ming Li, and Jimmy Lin. End-to-end open-domain question answering with bertserini. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 72–77, 2019.