BEEDS: Large-Scale Biomedical Event Extraction using Distant Supervision and Question Answering

Xing David Wang[1], **Ulf Leser**[1], **Leon Weber**[1,2]
[1]Computer Science Department, Humboldt-Universität zu Berlin
[2]Max Delbrück Center for Molecular Medicine
{wangxida, leser, weberple}@informatik.hu-berlin.de

Abstract

Automatic extraction of event structures from text is a promising way to extract important facts from the evergrowing amount of biomedical literature. We propose BEEDS, a new approach on how to mine event structures from PubMed based on a question-answering paradigm. Using a three-step pipeline comprising a document retriever, a document reader, and an entity normalizer, BEEDS is able to fully automatically extract event triples involving a query protein or gene and to store this information directly in a knowledge base. BEEDS applies a transformer-based architecture for event extraction and uses distant supervision to augment the scarce training data in event mining. In a knowledge base population setting, it outperforms a strong baseline in finding post-translational modification events consisting of enzyme-substrate-site triples while achieving competitive results in extracting binary relations consisting of protein-protein and protein-site interactions.

1 Introduction

Cellular processes such as DNA damage repair or cell division are realized by the orchestration of simple biochemical events into larger structures called pathways. Pathways play a crucial role in Biology research, for example in network analysis (Barabasi and Oltvai, 2004) or enrichment analysis (Reimand et al., 2019). For these applications, accurate and exhaustive lists of biochemical reactions are crucial. Examples for databases collecting such biochemical events are KEGG (Kanehisa et al., 2002), the Protein Interaction Database (PID, Schaefer et al., 2009) and Reactome (Fabregat et al., 2018). Although pathway knowledge bases strive to include as much information as possible their foremost goal is the correctness of provided data and they mostly rely on manual collection and review of data. Thus, they are notoriously incomplete despite extensive curation efforts (Weber et al., 2020).

In this work, we present BEEDS (**B**iomedical **E**vent **E**xtraction using **D**istant **S**upervision), a novel approach to biomedical event extraction from a large corpus, i.e., PubMed. BEEDS takes questions like *What phosphorylates JAK2?* or *What regulates expression of JAK2?* to find typed interactions between molecular entities and follow up questions like *Which sites does EPO phosphorylate in JAK2?* to expand upon previously found answers, as a basis to recover complex event structures. To answer such questions, BEEDS uses a pipeline of three steps: retrieval, machine reading and entity normalization. In the first step, our model retrieves documents relevant to the query from all PubMed abstracts and PubMed Central full texts. In the second step, we feed the retrieved documents to a transformer-based model to identify and extract answer spans in each document. In the third step, we apply an entity normalizer to map the identified entities to canonical database identifiers before returning them as answers.

As training data for event mining is notoriously scarce, BEEDS applies distant supervision for obtaining a more comprehensive model. Specifically, it extracts biochemical events from curated pathway knowledge bases and transforms these into text annotations, by sourcing text spans containing the pair of proteins from a knowledge base event. This creates a distant supervision training set, as we do not know whether a found text span actually describes the respective event. To the best of our knowledge, this is the first approach for distantly supervised biomedical event extraction. We augment this distantly supervised training set with gold standard text annotations for biomedical event structures from (Kim et al., 2011) and (Ohta et al., 2013). For evaluation, we again make use of pathway knowledge base data by checking how many of their reactions are found by our model.

Proceedings of the BioNLP 2022 workshop, Dublin, Ireland, pages 298–309
May 26, 2022. ©2022 Association for Computational Linguistics

Compared to EVEX (Van Landeghem et al., 2013) as baseline, our experiments indicate that BEEDS is well able to mine biomedical event structures from the literature achieving a rise in recall of about 13 percentage points (pp) when mining for enzyme-substrate-site triples of post-translational modifications (PTMs). In mining of binary relations like protein-protein and protein-site interactions, BEEDS gains about two pp in recall when compared to EVEX.

The rest of this paper is structured as follows: In Section 2, we give a brief overview over related work in event mining. In Section 3, we describe the event extraction task and our used data sets, explain each part of our model pipeline in detail and provide the evaluation setup together with our baseline. In Section 4 and Section 5, we present and discuss our results. In Section 6, we make final remarks and conclude this work.

The code for reproducing this paper is freely available under https://github.com/WangXII/BEEDS.

2 Related Work

The two approaches which are closest to BEEDS are EVEX (Van Landeghem et al., 2013) and PEDL (Weber et al., 2020). Both aim to solve the task of populating pathway knowledge bases with automatically extracted event structures from the literature. EVEX differs from BEEDS as it does not use a retriever component so its document reader has to be applied to every document in PubMed which is expensive in terms of computing resources. Additionally, it is only able to learn from manually labeled, directly supervised data and cannot incorporate noisy, distantly supervised text annotations for training. PEDL's main difference to BEEDS is that it is a relation extraction system and can only extract binary relations but not more complex event structures with three or more participants.

Regarding the formulation of biomedical event extraction as question answering with a document reader, BEEDS builds upon our previous approach introduced in Wang et al. (2020). DeepEventMine (Trieu et al., 2020), another approach for biomedical event extraction, solves the task by employing a multi-layered model structure each responsible for a different step in event construction like entity detection and event merging. However, both these methods only make use of directly supervised training data. Furthermore, they both only cover the machine reading component of biomedical event extraction and have not been applied to large-scale biomedical event extraction.

Similar approaches combining a retriever reader model to pose questions directly to a corpus include DrQA (Chen et al., 2017), REALM (Guu et al., 2020) and Lewis et al. (2020). DrQA answers questions posed to a Wikipedia corpus and uses two models, the BM25 algorithm for retrieval (Robertson and Walker, 1994) and a deep learning model consisting of an LSTM (Long short-term memory) for reading. The BM25 algorithm is still a widely used document retrieval algorithm, e.g., in the internal retrieval tool of PubMed, Best Match (Fiorini et al., 2018), where it is complemented by a machine learning model reranking its top 500 retrieved documents. REALM and Lewis et al. follow a similar idea like introduced in DrQA but use dense retrieval methods, i.e., a retriever employing a deep learning model, and unite the retriever and reader components in a joint deep learning model which can be optimized end-to-end. Compared to BEEDS, these systems lack a normalizing component and have neither been applied to event extraction nor in the biomedical domain.

3 Material and methods

3.1 Event types and data sets

BEEDS can extract three types of biomedical events: Post-translational modifications (PTMs), gene expressions and regulation events in general. Regulation events include the former two event types plus other forms of state changes. For PTMs, such as phosphorylation, we extract relation triples of theme, cause and amino acid site. For gene expression and regulation, we extract relation pairs of theme and cause. Themes are always given by a single protein or gene, causes or controllers may also include other types of molecules. For the remainder of the document, we use the terms protein and gene interchangeably. BEEDS neither recognizes event modifiers like negation or speculation, i.e., it may extract negated or speculated events without discerning the negations or speculations themselves, nor the polarities of events, like positive or negative regulation.

BEEDS uses a data set for training that consists of two portions: The first portion is a distantly supervised, knowledge base data set containing presumable descriptions of events from the union of the following seven pathway databases: KEGG, PID, Reactome, HumanCyc (Romero et al., 2005), INOH (Yamamoto et al., 2011), PANTHER (Mi et al., 2017) and NetPath (Kandasamy et al., 2010). The second portion is a directly supervised data set containing gold annotations from the GENIA (Kim et al., 2011) and Pathway Curation (Ohta et al., 2013) challenges; in the following, we call the former the KB data set and the latter the BioNLP data set.

3.2 Question answering for event extraction

For each of the three event types that BEEDS can extract, we define templates to construct the natural language questions from a given query entity. For regulations and gene expression, we define only one template, i.e., to find the controller for a given protein of interest. The template for regulations is:

What regulates [theme entity] ?

where [theme entity] is filled with the protein of interest. For PTMs, we define several question templates, each to extract a different participant in the event structure: One template to find the controller/enzyme of a given event, one to find modified amino acid sites on the given protein, and a third to find modified amino acid sites for a theme-cause pair found in a previous question from the first template. See Table 1 for an example and an overview of the question templates. We call all questions that build upon the answer of a previous question "multi-turn questions" and all other ones "single-turn questions".

We transform all event structures from our two data sets into question-answer pairs. The size of the transformed data sets can be found in Table 2. Note that for the KB data, each canonical protein entity (with a unique database identifier) occurs at maximum once for a combination of event and question type. For the BioNLP data, each occurrence of a protein entity in a different document counts as a separate question. We split the data sets into train, development and test sets across individual theme entities/proteins, e.g., all events with AKT1 as theme go into one split and all events with GSK3B as theme go into another one. To further reduce the danger of information leakage, we also grouped together all proteins belonging to the same function (as defined by Pfam Mistry et al., 2021) and assign them to the same data split, i.e., all AKT proteins (like AKT1, AKT2 or AKT3) are assigned to the same split.

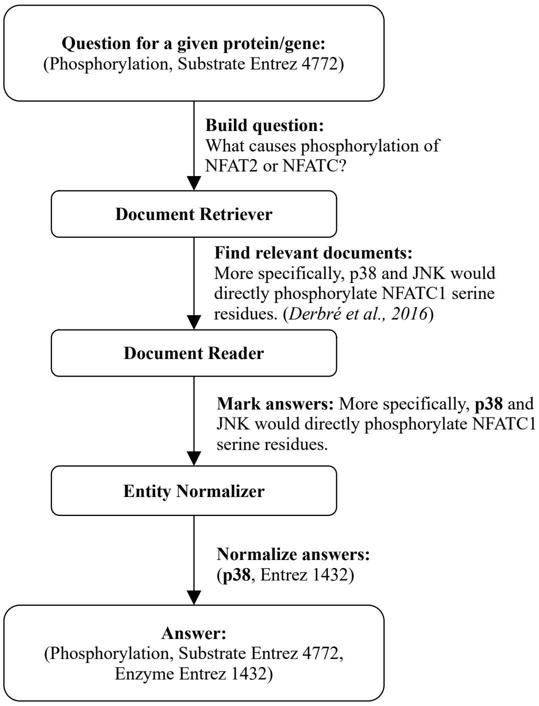
Question for a given protein/gene:
(Phosphorylation, Substrate Entrez 4772)
Build question:
What causes phosphorylation of NFAT2 or NFATC?
Document Retriever
Find relevant documents:
More specifically, p38 and JNK would directly phosphorylate NFATC1 serine residues. (Derbré et al., 2016)
Document Reader
Mark answers: More specifically, p38 and JNK would directly phosphorylate NFATC1 serine residues.
Entity Normalizer
Normalize answers:
(p38, Entrez 1432)
Answer:
(Phosphorylation, Substrate Entrez 4772, Enzyme Entrez 1432)

Figure 1: Model overview

3.3 BEEDS overview

BEEDS implements a pipeline consisting of three main components: the document retriever, the document reader and the entity normalizer. An overview is shown in Figure 1. We now describe each component in detail.

3.4 Document retriever

During document retrieval, we want to select the documents probably relevant to our query, which we define of those containing the query protein and a trigger term for the query event. If retrieval fails, then our subsequent machine reading model has no chance of finding correct answers and events in the provided documents. A reliable document retriever is the BM25 model (Robertson and Walker, 1994) which ranks documents based on their cosine similarity between query and document in TF-IDF representation. We use Apache Lucene[1] to index all documents and to perform the BM25-based retrieval.

Our document corpus consists of all currently available PubMed[2] abstracts plus the full texts from the open access portion of PubMed Central[3]. For

[1] https://lucene.apache.org

[2] https://pubmed.ncbi.nlm.nih.gov/about/

[3] https://pubmed.ncbi.nlm.nih.gov/

Event type	Question type	Example
PTM	Cause Site Cause + Site	*What regulates phosphorylation of CLIP1?* *Where is CLIP1 phosphorylated?* *Where does mTOR phosphorylate CLIP1?*
Expression	Cause	*What induces gene expression of MIP-1-beta?*
Regulation	Cause	*What regulates RUNX2?*

Table 1: Overview of extracted events and corresponding question types.

	BioNLP		KB	
	Qu.	Ans.	Qu.	Ans.
Phosphorylation Cause	272	440	674	2,452
Phosphorylation Site	214	404	546	1,792
Acetylation Cause	19	32	72	215
Acetylation Site	8	16	66	159
Ubiquitination Cause	16	19	134	271
Ubiquitination Site	3	4	54	100
Expression Cause	671	813	721	2,868
Regulation Cause	1,878	3,244	1,584	7,171
Single Turn	**3,081**	**4,972**	**3,851**	**15,028**
Phospho. Cause + Site	61	67	1,783	4,247
Acety. Cause + Site	4	4	148	264
Ubiquit. Cause + Site	0	0	87	158
Multi Turn	**65**	**71**	**2,018**	**4,669**
All	**3,146**	**5,043**	**5,869**	**19,697**

Table 2: Number of questions (Qu.) and answers (Ans.) for the BioNLP and the KB data sets after transformation to question-answer pairs.

indexing and retrieval, each PubMed abstract and each paragraph of a PubMed Central full text are considered as one document, resulting in a set of ~140 million documents. An important hyperparameter of BEEDS is the maximal number r of top-ranked documents that are considered as potential answer sources for a given query.

To enhance retrieval performance, we slightly adjust our retrieval queries to obtain better ranking results. In a first step, we remove all tokens from the full question except the tokens for the protein and event type. We then expand the protein with a list of all its known synonyms, e.g., for AKT1 we add PKB-alpha, RAC, protein kinase b alpha etc. This list is extracted from NCBI Gene[4] and helps to cope with the severe synonym problem in protein naming. For the event types, we conduct a similar expansion by including further event triggers as defined in the BioNLP data set. In the end, we receive a list of subjects/objects and predicates as the retrieval query where at least one synonym for each entity has to be matched.

3.5 Document reader

For document reading, we employ BERT (Devlin et al., 2018), a popular transformer-based deep learning model. More specifically we use a pretrained checkpoint of the model called SciBERT (Beltagy et al., 2019). Question answering with BERT is modeled as a sequence labeling task where the input consists of the tokenized question, followed by a special separating token and a tokenized document from the retrieval. In the output sequence, corresponding answers in the tokenized document are marked using the IOB2 tagging notation where B and I stand for the start and middle of an answer token, O for a non-answer token and X for a continuation of a token from a previous word, respectively. Token splits are made automatically by the tokenizer and the X tag signalizes to defer labeling of a subtoken to its respective starting token. This tagging is realized by a fully connected output layer on top of BERT with the output dimension $d \times n$, where d denotes the number of possible sequence labels (4 in our case) and n denotes the maximum sequence length of the input. For each sequence position $i \in \{1, ..., n\}$, we obtain a d-dimensional vector denoting the log probabilities for each possible label. An example of input and output from the BERT document reader is shown in Figure 2. Detailed hyperparameter settings for BEEDS can be found in appendix A.

Generating distantly supervised training instances

As a distinct feature, BEEDS is able to also learn from noisy training annotations extracted from pathway knowledge bases. These samples are created as follows. Given a question-answer pair in the training set, we tag all answer synonyms that are near the question entities (protein, event type

about/

[4] https://ftp.ncbi.nih.gov/gene/DATA/gene_info.gz

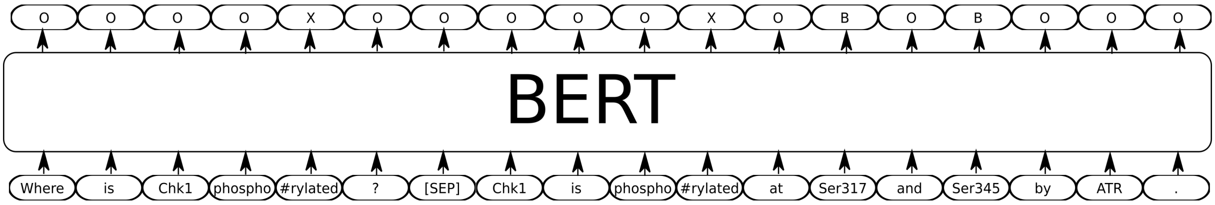

Figure 2: Question answering as sequence tagging. Depiction of the input tokens fed into the BERT model and of the output tags produced.

and possibly amino acid site) as a valid answer similarly to the strategies carried out in Quirk and Poon (2017) and Peng et al. (2017). We define "near" by restricting the number of sentences between a question entity and our answer candidate to three. For amino acid sites, the set of valid synonyms is defined as the full name of the amino acid and its abbreviations in form of one and of three letter codes. For instance, synonyms for the amino acid site Y183 include tyrosine183, Tyr183 and Y183 and the further combinations with either a whitespace, a hyphen or brackets, e.g., Y 183, Y-183 and Y(183).

Distant supervision and multi-instance learning

In the classic distant supervision setting as described by Mintz et al. (2009), all automatically generated annotations are assumed to be correct and thus valid learning examples. However, in many settings, including the one described here, examples contain noisy, false positive training examples which may lead to conflicting signals for the learner and degraded model performance (Surdeanu et al., 2012). In the multi-instance learning formulation, Surdeanu et al. (2012) alleviate this problem by relaxing the assumptions on the generated annotations. Instead of assuming every generated annotation to be right, their idea was to require only at least one of the generated annotations to be correct. We follow this idea and thus assume that only at least one of the text snippets per query-answer pair in the KB data set is correct, which means that our model does not need to fit every training example but nevertheless may do so. We call the collection of examples for a given question-answer pair a bag and use the hyperparameter b as maximal bag size ($b = 100$ in BEEDS). If retrieval size r is greater than the maximal bag size b, retrieved documents are split across multiple bags so that no bag exceeds size b.

A sequence annotation during training is deemed correct if the labels for each output token are tagged correctly. In the BioNLP data set, this is simply given by the gold standard tags. For the KB data set, this is given by our generated, distantly supervised annotations. The output tag at position k in the sequence of length n is determined by the tag with the highest output emission score $e_{y_{ik}}(x_i)$. The overall log probability of an output sequence $\mathbf{y}$ given the input sequence $\mathbf{x}$ is determined by the sum of log probabilities of its individual output labels:

$$\log P(\mathbf{y}|\mathbf{x}) = \log \prod_{i=1}^{n} P(y_i|x_i) = \sum_{i=1}^{n} \log P(y_i|x_i)$$

$$= \sum_{k=1}^{n} \max_{k=1,\ldots,d} e_{y_{ik}}(x_i) - \sum_{k=1}^{d} e_{y_{ik}}(x_i)$$

For our learning objective, we separate the whole bag of training examples into a positive and a negative bag. The positive bag contains all the output sequences which have marked at least one answer, i.e., one token at least has a B or I label. The negative bag on the other hand contains all noisy annotations where no token is marked as a potential answer. Applying the multi-instance learning formulation for each bag separately ensures that our model learns when to label an answer with a B or I token instead of just labelling every token with an O token. We apply the multi-instance formulation by calculating the maximum of all sequence log probabilities for both the positive and negative bag. For training stability and optimization purposes, we use the smooth approximation of the maximum function, logsumexp, in our computations instead of the maximum preventing a sparse gradient flow (c.f. Weber et al., 2020). As our objective loss functions are to be minimized instead of maximized, we multiply the resulting probability by -1. We sum up our positive loss ℓ_{pos} and negative loss score ℓ_{neg} to obtain the final objective function $\ell_{distant}$:

$$\ell_{pos} = -\log \sum_{\mathbf{y}_j \in pos} \exp P(\mathbf{y}_j|\mathbf{x}_j)$$

$$\ell_{distant} = \ell_{pos} + \ell_{neg},\ \ell_{neg} \text{ analogous}$$

For directly supervised examples, loss calculation is more straightforward. We use the same formulas but always set the bag size b to 1 which corresponds to a standard sequence labeling loss

$$\ell_{direct} = -\log P(\mathbf{y}|\mathbf{x}).$$

We do not use negative examples and bags for directly supervised examples. We introduce the additional hyperparameter w which is multiplied with each direct loss ℓ_{direct} allowing us to control the relative importance of direct examples in comparison to distantly supervised examples. During each training step, we either choose one directly supervised example or one bag with distantly supervised examples resulting in the final loss

$$\ell = \begin{cases} \ell_{distant} & \text{, if distantly supervised sample,} \\ w \cdot \ell_{direct} & \text{, else.} \end{cases}$$

3.6 Entity normalizer

For entity normalization, we use the existing normalizer PubTator Central[5] from Wei et al. (2019). It provides mention-level and document-level normalizations for proteins in every PubMed and PubMed Central article by mapping mentions to NCBI Entrez Gene identifiers. Because proteins in our knowledge bases are identified using UniProt identifiers, we map the UniProt identifiers to their corresponding Entrez Gene identifiers using UniProt ID mappings[6]. In addition, most of our knowledge bases focus on interactions in human. To handle homologous genes from other species, we use HomoloGene[7] to map genes to their human orthologs (NCBI taxonomy ID 9606) whenever possible.

Entities which we cannot normalize to a gene/protein mention using PubTator Central are normalized to CHEBI[8] identifiers using a simple dictionary lookup. For amino acid site strings, our normalization performs the reverse way as the synonym expansion for sites (see Section 3.5), i.e., we try to transform every possible extracted amino acid sites from text to their canonical symbols. For instance, serine 123 would be normalized to S123.

3.7 Baseline and evaluation

We use EVEX (Van Landeghem et al., 2013) as a strong and still popular baseline for event mining. To allow adequate comparison to our results, we only consider documents published before 2013 for our document retrieval. We have downloaded all EVEX annotations[9] (one annotation file for each PubMed/PubMed Central article) and transformed the extracted events structures into the same question-answering format as used by our model. Mapping of the BioNLP/EVEX events to our event types is straightforward and can be found in the appendix Table 8.

Our evaluation setup consists of two experiments: knowledge base evaluation and sample evaluation. Knowledge base evaluation is a fully automated evaluation where we measure how many of the event structures in the test set of the KB data set are found by each method. As evaluation metrics, we use knowledge base recall and the number of predicted question-answer pairs; note that for those not in the DB data set we cannot decide automatically whether they are correct or not and thus cannot compute a precision. In such a setting, the number of predicted question-answer pairs is helpful to put the achieved recall value into perspective.

Sample evaluation involves manual review of some randomly chosen events extracted by BEEDS and some events extracted by the baseline and allows to estimate precision. A further advantage of this evaluation, though laborious, is that it also considers new predictions, i.e., those events not already present in a knowledge base.

4 Results

4.1 Knowledge base evaluation

We present the results of the knowledge base evaluation in Table 3. Overall BEEDS achieves a ~5pp higher recall than EVEX. The difference is more pronounced in multi-turn questions where BEEDS achieves a recall of 14.30% while EVEX results are close to zero. Note that for knowledge base evaluation of multi-turn questions, we only count

[5] ftp://ftp.ncbi.nlm.nih.gov/pub/lu/PubTatorCentral/

[6] ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/idmapping/idmapping_selected.tab.gz

[7] https://ftp.ncbi.nih.gov/pub/HomoloGene/current/homologene.data

[8] https://www.ebi.ac.uk/chebi/

[9] http://evexdb.org/download/standoff-annotation/

the theme-cause-site triples where the theme-cause pair extracted from the previous single-turn question has been correct, i.e., the theme-cause pair has been curated in one of our knowledge bases. The multi-turn question itself is answered correctly if the whole event triple was extracted correctly. Compared to single-turn questions, recall in multi-turn questions falls off in both approaches, e.g., in BEEDS from 35% to about 14%. In single-turn questions, BEEDS outperforms EVEX in PTMs with a difference of 0pp to 48pp for the different types of PTM, whereas EVEX outperforms BEEDS in expression and general regulations by ~4pp.

Interestingly, BEEDS achieves this higher overall recall with only half of the number of predictions (29,867 versus 56,482). The discrepancy in number of predictions is especially high for the single-turn questions of expressions and regulations. In contrast, BEEDS extracts many more for all other event types. For instance, BEEDS is able to return about 2,000 controller-cause-site triples (given a valid controller-cause pair) whereas EVEX is only able to return 56 of such triples.

4.2 Sample evaluation

We present the results for the sample precision in Table 4. For each model, we randomly sampled 109 predictions and evaluated the correctness of the textual annotations manually (excluding entity normalization). We made sure that the number of each question type and each event type is roughly the same for our model and for the baseline. BEEDS achieves a total sample precision of 49.09% compared to EVEX with 63.30%.

In Table 5, we show events extracted by BEEDS. The first five samples are events not present in any of the knowledge bases showing that the model is able to extract new event structures. The last two are examples of typical errors.

5 Discussion

5.1 Comparison to EVEX

The higher number of predictions in EVEX likely stems from the fact that EVEX for each query analyses all PubMed abstracts whereas BEEDS considers only a limited amount of matches for each question, as controlled by the hyperparameter r (with $r = 1000$ in the experiments). This is especially true for the general regulation type which not only contains PTMs and expression events but also event types like transport or unspecific inhibitions and activations. Nonetheless, the limited amount of documents per question is sufficient for BEEDS to achieve a higher overall recall than EVEX showing that our retriever component is able to extract relevant documents.

For single-turn questions, BEEDS and EVEX achieve similar results. The advantages of BEEDS lie (a) in the important class of PTMs and (b) in multi-turn questions where simple event structures are merged to form larger event structures. Errors propagate in both models, i.e., wrongly extracted theme-cause pairs automatically lead to wrong theme-cause-site pairs, but event merging is more often successful in BEEDS. Multiplying the recalls for the Phosphorylation Cause question and the Phosphorylation Site question for BEEDS results in an expected recall of about 15% for the Phosphorylation Cause and Site question which is almost the exact recall the model achieves with 14.96%. Multiplying the same recalls in EVEX results in an expected recall of about 5% while the actually achieved recall of 0.84% is much lower. However, the higher number of merged events likely leads to a lower sample precision in BEEDS compared to EVEX (~37% versus ~69%).

It may be that recall improvement of BEEDS over EVEX is in part because of the newer GenNormPlus (Wei et al., 2015) normalization algorithm used in BEEDS compared to the older GenNorm (Wei and Kao, 2011) used in EVEX. However, the increase in F1-score performance from GenNorm to GenNormPlus (80.10% to 86.70%, see Wei et al., 2019) does not solely explain the significant discrepancy in recall for the multi-turn questions.

In the sample evaluation, BEEDS achieves much lower results in multi-turn question than in single-turn questions compared to EVEX. We hypothesize that BEEDS is more prone to error propagation than EVEX: Mainly, in extending falsely extracted event pairs to event triples whereas EVEX uses a more conservative approach to event merging. This is in line with our previous results from (Wang et al., 2020) where the machine reading component of EVEX, TEES (Björne and Salakoski, 2011), achieves a slightly worse precision than the machine reading component in BEEDS on the GENIA11 dataset (Kim et al., 2011, 57.65% to 59.33%) and a much better precision on the Pathway Curation dataset (Ohta et al., 2013, 55.78% to 48.74%). The former dataset contains more sim-

	Knowledge Base	BEEDS		EVEX	
	KB Gold	KB Recall	Predictions	KB Recall	Predictions
Phosphorylation Cause	715	36.92	3,175	24.75	3,398
Phosphorylation Site	546	42.12	3,076	19.96	797
Acetylation Cause	25	56.00	39	8.00	7
Acetylation Site	22	22.72	25	22.72	14
Ubiquitination Cause	57	36.84	217	26.31	80
Ubiquitination Site	9	33.33	17	22.22	7
Expression Cause	896	29.01	5,262	32.47	13,580
Regulation Cause	1,901	36.64	16,069	40.87	38,534
Single Turn	**4,171**	**35.81**	**27,880**	**33.03**	**56,426**
Phosphorylation Cause + Site	1,302	14.59	1,946	0.84	55
Acetylation Cause + Site	57	8.77	24	0.00	0
Ubiquitination Cause + Site	18	11.11	17	0.00	1
Multi Turn	**1,377**	**14.30**	**1,987**	**0.79**	**56**
All	**5,548**	**30.47**	**29,867**	**25.03**	**56,482**

Table 3: Results from knowledge base evaluation. Knowledge base (KB) recall values given in percent. In multi-turn questions, we only count the theme-cause-site triples where the extracted theme-cause pair from the previous single-turn question has been correct, i.e., the theme-cause pair has been curated in one of our knowledge bases.

Precision	Samples	BEEDS	EVEX
Single Turn	80	53.75	61.25
Multi Turn	29	36.66	68.96
All	109	49.09	63.30

Table 4: Precision on sampled text spans

ple events corresponding to single-turn questions and the latter more complex events corresponding to multi-turn questions. Overall, F1-scores of the BEEDS machine reading component in (Wang et al., 2020) and TEES show similar performances in the context of directly supervised tasks: 58.33% for BEEDS compared to 53.30% for EVEX in GENIA11 and 48.29% compared to 51.10% in Pathway Curation, respectively.

Another source of error decreasing the precision for multi-turn questions in BEEDS may be the distantly supervised training examples. Distantly supervised event triples likely contain much more noise than corresponding event pairs as one more entity must be mapped from the database event to potential events in the biomedical literature.

5.2 Importance of the retrieval size

In Table 6, we evaluate the impact of the retrieval size r on the final model performance (columns "BEEDS" versus "BEEDS (100 docs)"). Going from a retrieval size of 100 to 1,000 during evaluation almost doubles the knowledge base recall from 17.77 to 30.29%, implying that a tenfold increase in retrieval size has approximately resulted in a twofold increase in recall. In future work, we plan to perform additional experiments to explore the impact of r.

5.3 Importance of directly supervised data

We evaluate the impact of adding directly supervised data to our training set by evaluating model predictions specifically on the development set of the BioNLP data set. In Table 7, we see a considerable improvement of the ability of the model to extract correct text spans when giving gold annotations during training: On the BioNLP data, the recall increases from 4.84% to 65.23% and the precision improves from 41.46% to 68.45%. In Table 6, we can see similar results when evaluating the KB data set: With access to directly supervised data during training, the knowledge base recall increases from 23.01% to 30.29%.

5.4 Importance of the normalizer

In Table 6, we show results from an experiment where we evaluate how much performance is lost due to insufficient normalization of extracted text spans. We examine this step by constructing a simple dictionary lookup which inverts the mappings from all EntrezGene database identifiers to their respective entity synonyms. Then, we identify answer spans extracted by the machine reading component which have no corresponding normalization in PubTatorCentral. We match these text spans to the corresponding database identifiers from the lookup dictionary. This simple mapping would increase the recall by about a third from 30.29% to

Question Type *Text Evidence*	Substrate(s)	Kinase/Target	Document Source	Correctness
Acetylation Cause	EGID 9126	EGID 850584	(Ben-Shahar et al., 2008)	True
What acetylates SMC3? [...] we show that SMC3 is acetylated in an **ECO1** *-dependent manner [...]*				
Phosphorylation Site	EGID 84335	S183	(Bönig et al., 1996)	True
*Where is PRAS40 or AKT1S1 phosphorylated? PRAS40(***Ser183***) phosphorylation was also inhibited [...]*				
Phosphorylation Cause + Site	EGID 5970, 7124	S276	(Vermeulen et al., 2003)	True
Where does TNF phosphorylate p65? Mutational analysis of p65 revealed **Ser276** *[...] phosphorylation [...] in response to TNF.*				
Expression Cause	EGID 2353	EGID 3586	(Oshiro et al., 2007)	True
What causes expression of FOS or c-FOS? **Interleukin 10** *induced c-FOS expression in human B cells [...]*				
Regulation Cause	EGID 8313	EGID 1869	(Hughes and Brady, 2005)	True
What regulates AXIN2? **E2F1** *up-regulates the expression of the tumor suppressor AXIN2 [...]*				
Phosphorylation Site	EGID 2309	253	(Schwab et al., 2005)	False
*Where is FKHLR1 or FOXO3 phosphorylated? IGF-I induced phosphorylation of FKHR (***Ser 253***), FKHRL1 (Ser 256) [...]*				
Acetylation Cause	EGID 4303	EGID 23411	(Chuang et al., 2011)	False
What acetylates FOXO4? [...] AGE increases FOXO4 acetylation and suppresses expression of the **SIRT1** *protein deacetylase.*				

Table 5: Samples of correctly and wrongly extracted text spans by BEEDS.

KB Dev Set	Questions	Answers	Answers	KB Recall
BEEDS	**452**	**671**	**12,495**	**30.29**
BEEDS (Norm)	479	859	27,435	38.76
BEEDS (Distant)	433	510	7,767	23.01
BEEDS (100 Docs)	414	394	3,807	17.77
KB Gold	681	2,216		

Table 6: Ablation studies on the KB development (dev) set for BEEDS: BEEDS (Norm) estimating the upper bound for the KB recall, BEEDS (Distant) without access to the BioNLP data set and BEEDS (100 Docs) reducing the retrieval size to 100 from 1,000.

BioNLP Dev set	Gold	Preds	Recall	Precision	F1
BEEDS	351	335	65.24	68.35	66.76
BEEDS (Distant)	351	41	4.84	41.46	8.67

Table 7: Performance on the BioNLP dev set with and without access to gold data during training.

38.76% (but would also create many false positives decreasing model precision). This shows room for future optimization of the normalizer.

6 Conclusion

In this work, we have presented BEEDS, a new approach towards large-scale biomedical event extraction. We used question answering to iteratively extend biomedical event structures, first retrieving relevant documents and then applying a machine reader and normalizer to identify answer spans. On a knowledge base population task, BEEDS achieves similar results to an EVEX baseline for events with two participants and a much higher recall than EVEX on PTMs with three participants.

For future work, it remains to be examined how well other current biomedical event extraction approaches like DeepEventMine can be scaled up for large-scale curation efforts and how they compare to our model. We also plan to test other retrieval approaches like dense retrieval methods which might be able to improve the retrieval performance over BM25.

References

Albert-Laszlo Barabasi and Zoltan N Oltvai. 2004. Network biology: understanding the cell's functional organization. *Nature reviews genetics*, 5(2):101–113.

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

Tom Rolef Ben-Shahar, Sebastian Heeger, Chris Lehane, Philip East, Helen Flynn, Mark Skehel, and Frank Uhlmann. 2008. Eco1-dependent cohesin acetylation during establishment of sister chromatid cohesion. *Science*, 321(5888):563–566.

Jari Björne and Tapio Salakoski. 2011. Generalizing biomedical event extraction. In *Proceedings of BioNLP Shared Task 2011 Workshop*, pages 183–191.

H Bönig, D Körholz, B Pafferath, C Mauz-Körholz, and S Burdach. 1996. Interleukin 10 induced c-fos expression in human b cells by activation of divergent protein kinases. *Immunological investigations*, 25(1-2):115–128.

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

Peter Y Chuang, Yan Dai, Ruijie Liu, Helen He, Matthias Kretzler, Belinda Jim, Clemens D Cohen, and John C He. 2011. Alteration of forkhead box o (foxo4) acetylation mediates apoptosis of podocytes in diabetes mellitus. *PLoS One*, 6(8):e23566.

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

Antonio Fabregat, Steven Jupe, Lisa Matthews, Konstantinos Sidiropoulos, Marc Gillespie, Phani Garapati, Robin Haw, Bijay Jassal, Florian Korninger, Bruce May, et al. 2018. The reactome pathway knowledgebase. *Nucleic acids research*, 46(D1):D649–D655.

Nicolas Fiorini, Kathi Canese, Grisha Starchenko, Evgeny Kireev, Won Kim, Vadim Miller, Maxim Osipov, Michael Kholodov, Rafis Ismagilov, Sunil Mohan, et al. 2018. Best match: new relevance search for pubmed. *PLoS biology*, 16(8):e2005343.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. REALM: Retrieval-augmented language model pre-training. *arXiv preprint arXiv:2002.08909*.

Benjamin M Gyori, John A Bachman, Kartik Subramanian, Jeremy L Muhlich, Lucian Galescu, and Peter K Sorger. 2017. From word models to executable models of signaling networks using automated assembly. *Molecular systems biology*, 13(11):954.

Thomas A Hughes and Hugh JM Brady. 2005. E2f1 up-regulates the expression of the tumour suppressor axin2 both by activation of transcription and by mrna stabilisation. *Biochemical and biophysical research communications*, 329(4):1267–1274.

Kumaran Kandasamy, S Sujatha Mohan, Rajesh Raju, Shivakumar Keerthikumar, Ghantasala S Sameer Kumar, Abhilash K Venugopal, Deepthi Telikicherla, J Daniel Navarro, Suresh Mathivanan, Christian Pecquet, et al. 2010. Netpath: a public resource of curated signal transduction pathways. *Genome biology*, 11(1):1–9.

Minoru Kanehisa et al. 2002. The kegg database. In *Novartis Foundation Symposium*, pages 91–100. Wiley Online Library.

Jin-Dong Kim, Yue Wang, Toshihisa Takagi, and Akinori Yonezawa. 2011. 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.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Huaiyu Mi, Xiaosong Huang, Anushya Muruganujan, Haiming Tang, Caitlin Mills, Diane Kang, and Paul D Thomas. 2017. Panther version 11: expanded annotation data from gene ontology and reactome pathways, and data analysis tool enhancements. *Nucleic acids research*, 45(D1):D183–D189.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. 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.

Jaina Mistry, Sara Chuguransky, Lowri Williams, Matloob Qureshi, Gustavo A Salazar, Erik LL Sonnhammer, Silvio CE Tosatto, Lisanna Paladin, Shriya Raj, Lorna J Richardson, et al. 2021. Pfam: The protein families database in 2021. *Nucleic Acids Research*, 49(D1):D412–D419.

Tomoko Ohta, Sampo Pyysalo, Rafal Rak, Andrew Rowley, Hong-Woo Chun, Sung-Jae Jung, Sung-Pil Choi, Sophia Ananiadou, and Jun'ichi Tsujii. 2013. Overview of the pathway curation (pc) task of bionlp shared task 2013. In *Proceedings of the BioNLP Shared Task 2013 Workshop*, pages 67–75.

Noriko Oshiro, Rinako Takahashi, Ken-ichi Yoshino, Keiko Tanimura, Akio Nakashima, Satoshi Eguchi, Takafumi Miyamoto, Kenta Hara, Kenji Takehana, Joseph Avruch, et al. 2007. The proline-rich akt substrate of 40 kda (pras40) is a physiological substrate of mammalian target of rapamycin complex 1. *Journal of Biological Chemistry*, 282(28):20329–20339.

Nanyun Peng, Hoifung Poon, Chris Quirk, Kristina Toutanova, and Wen-tau Yih. 2017. Cross-sentence n-ary relation extraction with graph lstms. *Transactions of the Association for Computational Linguistics*, 5:101–115.

Chris Quirk and Hoifung Poon. 2017. Distant supervision for relation extraction beyond the sentence boundary. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1171–1182.

Jüri Reimand, Ruth Isserlin, Veronique Voisin, Mike Kucera, Christian Tannus-Lopes, Asha Rostamianfar, Lina Wadi, Mona Meyer, Jeff Wong, Changjiang Xu, et al. 2019. Pathway enrichment analysis and visualization of omics data using g: Profiler, gsea, cytoscape and enrichmentmap. *Nature protocols*, 14(2):482–517.

Stephen E Robertson and Steve Walker. 1994. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *SIGIR'94*, pages 232–241. Springer.

Pedro Romero, Jonathan Wagg, Michelle L Green, Dale Kaiser, Markus Krummenacker, and Peter D

Karp. 2005. Computational prediction of human metabolic pathways from the complete human genome. *Genome biology*, 6(1):1–17.

Carl F Schaefer, Kira Anthony, Shiva Krupa, Jeffrey Buchoff, Matthew Day, Timo Hannay, and Kenneth H Buetow. 2009. Pid: the pathway interaction database. *Nucleic acids research*, 37(suppl_1):D674–D679.

Tracy S Schwab, BB Madison, AR Grauman, and Eva L Feldman. 2005. Insulin-like growth factor-i induces the phosphorylation and nuclear exclusion of forkhead transcription factors in human neuroblastoma cells. *Apoptosis*, 10(4):831–840.

Mihai Surdeanu, Julie Tibshirani, Ramesh Nallapati, and Christopher D Manning. 2012. 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.

Hai-Long Trieu, Thy Thy Tran, Khoa NA Duong, Anh Nguyen, Makoto Miwa, and Sophia Ananiadou. 2020. Deepeventmine: End-to-end neural nested event extraction from biomedical texts. *Bioinformatics*.

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. 2013. Large-scale event extraction from literature with multi-level gene normalization. *PloS one*, 8(4):e55814.

Linda Vermeulen, Gert De Wilde, Petra Van Damme, Wim Vanden Berghe, and Guy Haegeman. 2003. Transcriptional activation of the nf-κb p65 subunit by mitogen-and stress-activated protein kinase-1 (msk1). *The EMBO journal*, 22(6):1313–1324.

Xing David Wang, Leon Weber, and Ulf Leser. 2020. 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.

Leon Weber, Kirsten Thobe, Oscar Arturo Migueles Lozano, Jana Wolf, and Ulf Leser. 2020. Pedl: extracting protein–protein associations using deep language models and distant supervision. *Bioinformatics*, 36(Supplement_1):i490–i498.

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

Chih-Hsuan Wei and Hung-Yu Kao. 2011. Cross-species gene normalization by species inference. *BMC bioinformatics*, 12(8):1–11.

Chih-Hsuan Wei, Hung-Yu Kao, and Zhiyong Lu. 2015. Gnormplus: an integrative approach for tagging genes, gene families, and protein domains. *BioMed research international*, 2015.

Satoko Yamamoto, Noriko Sakai, Hiromi Nakamura, Hiroshi Fukagawa, Ken Fukuda, and Toshihisa Takagi. 2011. Inoh: ontology-based highly structured database of signal transduction pathways. *Database*, 2011.

A Implementation Details

For implementation, parsing of the knowledge base event structures is done by INDRA[10] (Gyori et al., 2017). Mapping the event types in INDRA to our custom types is straightforward, events with a substrate and an enzyme without a corresponding event type in BEEDS are just mapped to the regulation event type.

The retrieval size r for our noisy training sets is 100. During evaluation in the development and test sets, we have found out that a larger retrieval size improves the recall considerably (see Table 6), so $r = 1000$ there. The bag size for multi-instance learning is $b = 100$. The additional weight factor that we multiply directly supervised examples with is $w = 4$. Model training is halted using the early stopping criterion.

Weight parameter of BERT are initialized to the configuration of the pretrained SciBERT (Beltagy et al., 2019) checkpoint. Maximum sequence length for a document is 384, longer documents are truncated so that the question entities remain in the document. Further hyperparameters to the BERT model are a learning rate of 2e-5, the proportion of warmup steps set to 0.1 and a weight decay of 0.01. Dropout probability for every weight in the network is set to 0.1, we use one step for gradient accumulation and a maximum norm of one before we apply gradient clipping. Input parameters to the AdamW optimizer use the default values of β_1 = 0.9, β_2 = 0.999 and ϵ = 1e-8.

B Transformation of BioNLP and EVEX data

In Table 8, we report the mapping from EVEX and BioNLP event types to our event types in BEEDS.

C Binding and Complex events

During our model development, we have also experimented with extracting protein complexes of either two (question type complex pair) or three participants (question type complex triple). The number of gold knowledge base question answer

[10] https://indra.readthedocs.io/en/latest/modules/statements.html

EVEX/BioNLP event types	BEEDS event types
REGULATION of (de-)phosphorylation	Phosphorylation
REGULATION of (de-)acetylation	Acetylation
REGULATION of (de-)ubiquitination	Ubiquitination
REGULATION of gene expression, transcription	Expression
All REGULATIONs including above	Regulation

Table 8: Mapping of EVEX/BioNLP event types to our event types. REGULATION refers to one of the four regulation types in EVEX: Catalysis, Regulation, Positive Regulation and Negative Regulation.

pairs is much larger than for the other event types. This is most likely due to the worse evidence for protein complexes curated in the pathway knowledge bases compared to the evidence of the other question types as many complex relations are determined automatically by transitive nature between separate protein complexes. A sample question for complex pairs would be *"What protein is in complex with AKT-1?"*. A corresponding sample question for complex triples would be *"What protein is in complex with AKT-1 and AKT-2?"*.

Complex pair	Questions	Answers in KB	Answers	Recall
BEEDS	997	1,494	27,880	35.81
EVEX	938	1,378	56,426	33.03
KB Gold	1,074	4,171		

Table 9: Single-turn question Complex pair.

Complex triple	Questions	Answers in KB	Answers	Recall
BEEDS	106	432	1,914	0.05
EVEX	1,334	630	4,818	0.07
KB Gold	20,453	832,875		

Table 10: Multi-turn question Complex triple.

We report the results for the single-turn question of complex pairs in Table 9 and for the multi-turn question of complex triples in 10. As the number of gold knowledge base question answer pairs is much higher in these two question types, the resulting recall values are much lower for both BEEDS and EVEX. EVEX has access to whole PubMed during prediction time, so the number of predictions is much higher than in BEEDS which translates into a larger recall value for both question types.