

Text-to-Text Extraction and Verbalization of Biomedical Event Graphs

Giacomo Frisoni*, Gianluca Moro* and Lorenzo Balzani*

Department of Computer Science and Engineering (DISI)
University of Bologna, Via dell'Università 50, I-47522 Cesena, Italy
{giacomo.frisoni, gianluca.moro}@unibo.it
balzanilo@icloud.com

Abstract

Biomedical events represent complex, graphical, and semantically rich interactions expressed in the scientific literature. Almost all contributions in the event realm orbit around semantic parsing, usually employing discriminative architectures and cumbersome multi-step pipelines limited to a small number of target interaction types. We present the first lightweight framework to solve both event extraction and event verbalization with a unified text-to-text approach, allowing us to fuse all the resources so far designed for different tasks. To this end, we present a new event graph linearization technique and release highly comprehensive event-text paired datasets, covering more than 150 event types from multiple biology subareas (English language). By streamlining parsing and generation to translations, we propose baseline transformer model results according to multiple biomedical text mining benchmarks and natural language generation metrics. Our extractive models achieve greater state-of-the-art performance than single-task competitors and show promising capabilities for the controlled generation of coherent natural language utterances from structured data.¹

1 Introduction

In recent years, events have become an influential formalism for modeling complex relations mentioned within the text as semantic graphs (Frisoni et al., 2021, 2022). In bioinformatics, an event generally refers to an interaction between one or more biomedical entities (e.g., proteins, genes, diseases, drugs), each contributing with a specific role (e.g., Theme, Cause, Site). For instance, biomedical events include molecular reactions, organism-level outcomes, and adverse drug reactions. Their expressive power and flexibility have supported

*Equal contribution.

¹The data and the code to reproduce our baseline results are available at <https://github.com/disi-unibo-nlp/bio-ee-egv>

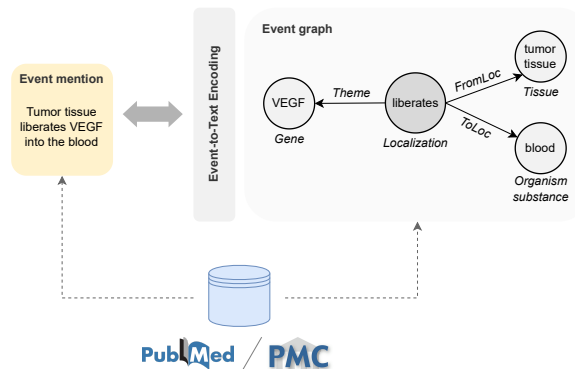


Figure 1: Illustration of an event graph and its textual mention from our datasets. All text-event pairs refer to human-crafted annotations above the biomedical literature (abstracts or full papers).

many practical applications like literature-based knowledge discovery (Wang et al., 2021b), biological network construction (Björne et al., 2010), diagnosis prediction (Zhang et al., 2020c), document summarization (Zhang et al., 2020b), and question answering (Berant et al., 2014).

Text-to-event (or event extraction, EE) and event-to-text (or event graph verbalization, EGV) systems effectively bridge natural language and symbolic representations. They provide a step towards decoupling concept units (*what to say*) from language competencies (*how to say it*) (Mel'čuk, 1973). Strongly linked to natural language understanding, EE is a fundamental task to automatically identify, monitor, and aggregate the relational knowledge disseminated within life science papers, speeding up medical progress and promoting discoveries. Yet, although much attention has been paid to EE, no research efforts have been directed to its inverse task, namely EGV. Even if under-explored, EGV targets the generation of informative text constrained on semantic graphs, holding a lot of potential in applications like conversational agents and summarization systems (Frisoni et al., 2022; Moro and Ragazzi, 2022; Moro et al., 2022).

Most state-of-the-art (SOTA) approaches handle structured prediction by employing task-specific architectures and discriminative models. Ordinarily, they need to be adapted to the target events and their schema, or they are not flexible enough to work with different domains or analysis granularities (sentence- vs document-level). The extraction process is typically divided into subtasks, executed in a pipeline or joint manner, where the output of several classifiers needs to be integrated. Additionally, each task is often associated with its own output space, limiting knowledge sharing and multi-task learning (MTL). Classes (i.e., event, argument role, and entity types) are specified implicitly through numerical indices, and models contain no prior information about their meaning. Furthermore, modern deep learning solutions require a non-trivial amount of examples to train, but event annotations are expensive to produce—a discrepancy that results in multiple, stand-alone, and closed-domain datasets with few records and potentially overlapping labels (Miwa et al., 2013).

On a parallel track, transfer learning has been the pinnacle of the latest breakthroughs in natural language processing (NLP). Large pre-trained language models (PLMs) are powerful backbones that can be fine-tuned for different tasks to achieve impressive performance in wide-ranging applications (Kalyan et al., 2021). PLMs capture contextual information and latent linguistic/relational knowledge (Petroni et al., 2019; Roberts et al., 2020), incorporating syntax and semantics. In that sense, textual representations and conditional generative modeling can be seen as natural ways of encoding different events in a shared predictive space.

In this paper, we design a framework to solve both EE and EGV as text-to-text problems, thus leveraging SOTA PLMs and disposing of the need for complex and hardly adaptable architectures. Concretely, we propose a way to decompose events into text sequences, neatly preserving structure and labels. Above it, we present the Biomedical Text-to-Event (BIO2TE) and Event-to-Text (BIOE2T) datasets, two corpora of textualized biomedical event graphs paired with their mention. Precisely, we aggregate and preprocess gold annotations coming from 10 popular EE benchmarks, intending to systematize the community work matured with public evaluation programs and solving the low coverage issue. Among the exciting multimodal opportunities enabled by these datasets, we show that

out-of-the-box transformer models can effectively learn text \rightarrow event and event \rightarrow text translations (Figure 1). We achieve this symmetry by using the same architecture for parsing and generation, as well as for all event instances, originally belonging to separate EE tasks with independent output spaces. To the best of our knowledge, this is the first study to handle such a variety of event schema without distinct models or additional task-specific modules. Our key contributions are the following:

1. We devise a novel event linearization with a consistent textual output format based on formal grammar (§3).
2. We introduce BIO2TE and BIOE2T, two large-scale biomedical event-text aligned datasets designed to frame the extraction and verbalization of general biomedical events as text-to-text tasks (§4).
3. We experiment EE, EGV, and MTL (§5 and §6). We demonstrate that autoregressive seq2seq models can achieve SOTA performance—previously attained only by discriminative solutions—while being much more flexible and scalable.

2 Related Work

Graph-Text Paired Data. Many graph-text paired datasets have sprung up. Nevertheless, annotating text or semantic graphs is expensive, especially for specific fields like biology. Most of the resources are domain-general and focus on knowledge graphs (KGs). Although there are datasets assembled by crowdsourced human annotators—such as WebNLG (Gardent et al., 2017), one common thread is using NLP tools and automatic alignment heuristics to forge silver pairs massively, e.g., mapping Wikipedia sentences to Wikidata triples (Elsahar et al., 2018; Agarwal et al., 2021) or Wikipedia paragraphs to Freebase subgraphs (Wang et al., 2021a). Predicate linkers and PLMs are already used to inherently construct KGs from the biomedical literature (Geleta et al., 2021), but the relations extracted for each document are generally not openly released. In contrast, we present the first datasets directly pairing scientific sentences to biomedical event graphs, usable as evaluation gold standards thanks to expert user provenance.

Event Extraction. In the NLP field, EE is placed within the more general information extraction (IE)

and structured prediction (SP) areas. Specifically, it aims to interpret and distill free-text chunks into structured, semantic, and fine-grained relations capturing an interplay between many different participants (entities or other events) usually subjected to a state change. EE requires to recognize *triggers* (text spans that clearly testify the occurrence of a real-world event), classify the type of the events for which they act as lawyers, detect involved 0-N *arguments* (entity mentions and corresponding classes, or sub-event triggers), predict their *semantic role*, and establish some optional event-level *modifiers*. For example, a Localization event is indicated in Figure 1 at “liberates”, involving three bio-entities. Notably, end-to-end EE systems usually integrate named entity recognition (NER) and coreference resolution. Compared to the more traditional binary relation extraction, where the goal is deriving subject-relation-object triplets, EE is a more complex task that needs to deal with high-level linguistic phenomena, an avalanche of narrative styles, and syntactic constructions.

Early approaches tackled EE with pipeline architectures to decompose the problem in its sequential subtasks and independently train a classifier for each of them, also relying on gold-tagged entities to eventually ignore NER objectives. Historically, first attempts made use of pattern-based techniques (Cohen et al., 2009) or data-driven methods centered on generalizing classical machine learning algorithms to SP, including, among others, support vector machines (Miwa et al., 2012). More recently, joint and MTL architectures have gained popularity among researchers, training a single model on all EE sub- and linked-tasks simultaneously, benefiting from information sharing and mutually improving local predictions. Deep learning is the main architect of this transition, with many EE efforts rooted in transformers (Ramponi et al., 2020), convolutional (Björne and Salakoski, 2018), recurrent (Li et al., 2019), and graph (Zhao et al., 2021) neural networks (CNNs, RNNs, GNNs). Current SOTA EE solutions train end-to-end neural models on top of the features learned by domain-specific PLMs, such as SciBERT (Beltagy et al., 2019). In this line of work, DeepEventMine (Trieu et al., 2020) presently holds leading performance on most biomedical EE (BEE) benchmarks, with custom discriminative classification layers above SciBERT-encoded intra-sentence spans. Most BEE systems work within the sentence scope, not being able to

scale to entire documents and facts with scattered arguments. Our framework is designed for joint EE, also including the NER subtask², and is not limited to sentence-level extraction in principle.

Data-to-Text. Data-to-text is the task of generating natural language text conditioned on source content provided in the form of structured data. Different GNNs have been proposed to better encode the input structure in the case of graphs, like Graph Transformers (Koncel-Kedziorski et al., 2019) and DualEnc for KG-triples ordering and verbalization (Zhao et al., 2020). On the other hand, recent works (Kale and Rastogi, 2020; Wang et al., 2021c; Agarwal et al., 2021) have favored seq2seq pre-trained models—with T5 (Raffel et al., 2020) as prominent example—which showcased better grammatical correctness and domain-shift robustness. To the extent of our knowledge, no prior research has attempted to verbalize event graphs. In this paper, we start from these heated evidences to fill the gap.

Seq2seq for Structured Prediction and Graph Verbalization. It has become increasingly popular to cast structured prediction problems as translations between natural languages, linearizing data when necessary and leveraging the transfer learning capacity of a transformer-based PLM. Text-to-text reframing has been applied to many contexts, from general NLP tasks (Raffel et al., 2020) and semantic role labeling (Biloshmi et al., 2021) to relation extraction (Huguet Cabot and Navigli, 2021). Closer to us, TANL (Paolini et al., 2021) and TEXT2EVENT (Lu et al., 2021b) are the only works carrying out this strategy on EE. However, the authors solely consider ACE2005 and ERE, two simplistic newswire datasets with a small type coverage and flat target structures. More importantly, TANL encodes event annotations in the form of augmented text, dividing EE into different subtasks with the lack of support for nested events, modifiers, or event overlapping, which are instead common in biology, thus being not directly applicable to our datasets. Instead, we solve BEE by generating the output graph at once, supporting complex structures and schema. Outside of our work, symmetric parsing and generation have been chiefly explored

²We do not predict relationships between gold entities, as is frequently assumed in other works dependent on extra input annotations or external NER tools that interrupt the backpropagation process. On the contrary, our models are directly trained to recognize target entities and classify their type as a fundamental subtask for the ultimate goal of end-to-end event extraction via text translation.

with AMRs (Konstas et al., 2017; Bevilacqua et al., 2021). The aforementioned publications highlight the relevance of seq2seq models. Not only do they exhibit strong performance, but they also lean on decoding mechanisms rather than predefined type sets, being easily extendable to new or unseen inputs. We also underline that, by conditioning future decoding on previous generations, they implicitly deal with dependencies among graph records (i.e., non-atomic extractions).

3 Event Linearization

Seq2seq models require that both the input and target be presented as a linear sequence of tokens. In this section, we describe our format design concept to reformulate event graphs as strings.

Events have an n-ary and potentially nested structure, with optional modifiers (e.g., “polarity”, “certainty level”) reshaping the described interaction. Like many other relational data, events can be conveniently and naturally modeled as rooted directed acyclic graphs (Frisoni et al., 2021). With this formalization, triggers and entities are nodes, while argument roles define edges. Each trigger, entity, and trigger-trigger/trigger-entity association is assigned to its type according to a predefined ontology³.

We revisit the formulation by (Paolini et al., 2021) and put forward a formal event language designed to be easy and deterministically reversible to event graphs. While being more complex to learn, our linearization comes with the advantage of enclosing entire events in single expressions, reducing the overhead of generating different output sequences for trigger and argument annotations. Each node is surrounded by the special tokens [and], which represent semantic structure indicators. Inside, a sequence of |-separated tags reports the text span, the type (described in natural words), and a list of X=Y relations, where X is the argument role and Y is the target trigger. Note that the same entity can be coupled to different events (triggers) with distinct roles (i.e., double tagging). The root trigger is a source vertex and has not incoming edges. Trigger nodes also specify event-level modifiers as additional X=Y assignments, in the form `Property=Value` (e.g., “Po-

³In this paper, we refer to closed-domain EE settings. Please note that closed-domain EE exclusively searches for target events (e.g., positive/negative regulation, binding, carcinogenesis) with a defined schema. On the contrary, open-domain EE does not assume specific target types and aims to detect general events unsupervised, thereby being more limited.

larity=Negative”). So, the information on event components (i.e., nodes and their embedded interconnections) are all within [...] patterns, which can be nested in case of sub-events. The final string minimizes the number of tokens to be submitted or generated so as to make encoding/decoding more efficient. Event constituents are sorted by their order of appearance in the .a2 for consistency.

We define a context-free grammar (§A.1) and test it with JFLAP (Rodger and Finley, 2006). Figure 2 depicts a practical example of textualized nested bio-event.

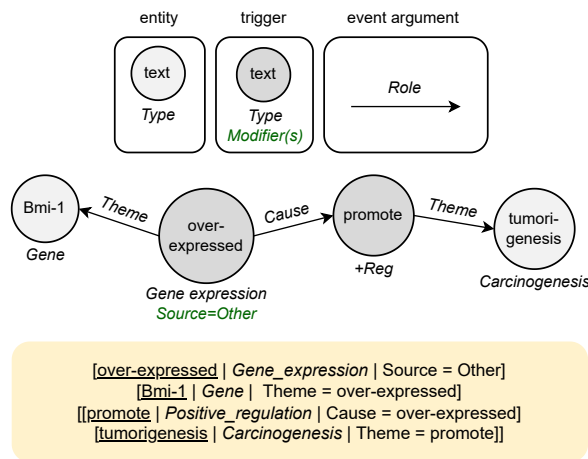


Figure 2: Example of textually linearized event graph.

4 Datasets

Based on §3, we build new corpora suitable for text- or graph-conditioned sequence modeling. Here, we present the construction process of BIOT2E and BIOE2T, together with their main properties.

4.1 Construction Process

4.1.1 Data Collection

Obtaining a large gold dataset of jointly annotated pairs of sentences and event graphs may require years of labor (Kim et al., 2008). We overcome this issue by combining the training sets of 10 influential real-world datasets originally designed for BEE, primarily derived from the ongoing BioNLP-ST series (Kim et al., 2019). Table 1 reports the characteristics of the seed datasets used for BIOE2T and BIOT2E construction⁴. These sources comprise seminal bioinformatics projects like GENIA (Kim et al., 2013a) and well-known tasks meeting

⁴We intentionally focused on fusing existing benchmarks to build a highly-comprehensive biomedical evaluation gold standard (not possible with silver pairs forged with heuristics).

biologists’ needs, including topics such as cancer genetics (Pyysalo et al., 2013) and infectious diseases (Pyysalo et al., 2011). Each focuses on a particular domain, differs in the annotation schema, and consists of human-curated event annotations on top of PubMed abstracts and full papers (English language). The reader should be aware that biomedical benchmarks generally support only two boolean modifiers—negation and speculation. On the flip side, modifiers are essential for a correct event interpretation, even with instances having the same triggers and arguments. Given the potential and uniqueness of modifiers in event data, we include GENIA-MK (Miwa et al., 2012) to manage more sophisticated forms of meta-knowledge during translation. Details on embraced modifiers are available in §A.2.

	Corpus	Domain(s)	#Documents	Annotation Schema
BioNLP-ST*11	Genia Event Corpus (GE08) (Kim et al., 2008)	Humans blood cells transcription factors	1,000 abstracts	35 entity types, 35 event types
	Genia Event 2011 (GE11) (Kim et al., 2012)	See GE08	1,210 abstracts, 14 full papers	2 entity types, 9 event types, 2 modifiers
	Epigenetics and Post-translational Modifications (EPI11) (Ohta et al., 2011)	Epigenetic change and common protein post-translational modifications	1,200 abstracts	2 entity types, 14 event types, 2 modifiers
	Infectious Diseases (ID11) (Pyysalo et al., 2011)	Two-component regulatory systems	30 full papers	5 entity types, 10 event types, 2 modifiers
	Multi-Level Event Extraction (MLEE) (Pyysalo et al., 2012)	Blood vessel development from the subcellular to the whole organism	262 abstracts	16 entity types, 19 event types
	GENIA-MK (Miwa et al., 2012)	See GE08	1,000 abstracts	35 entity types, 35 event types, 5 modifiers (+2 inferable)
BioNLP-ST*13	Genia Event 2013 (GE13) (Kim et al., 2013a)	See GE08	34 full papers	2 entity types, 13 event types, 2 modifiers
	Cancer Genetics (CG13) (Pyysalo et al., 2013)	Cancer biology	600 abstracts	18 entity types, 40 event types, 2 modifiers
	Pathway Curation (PC13) (Ohta et al., 2013)	Reactions, pathways, and curation	525 abstracts	4 entity types, 23 event types, 2 modifiers
	Gene Regulation Ontology (GRO13) (Kim et al., 2013b)	Human gene regulation and transcription	300 abstracts	174 entity types, 126 event types

Table 1: Summary of the biomedical event extraction corpora used for constructing BIOT2E and BIOE2T. All data is in public domain and licensed for research purposes.

4.1.2 Data Preprocessing, Filtering and Sampling

Annotations follow standoff *.a**, *.ann*, or *.xml* formats, where labels are connected to the text spans of the document through (*start*, *end*) character offset pairs. We automatically produce the linearized version of each event graph by parsing and normalizing these files, otherwise having structure

and labeling variants depending on the original dataset. For example, *.ann* files identify modifiers with “A” instead of “M”; GENIA-MK specifies the value of each property and not the active type only (e.g., “Speculation=True” versus “Speculation”); GRO13 supports the recognition of triggers or entities with scattered text spans (e.g., “RFX . . . 3” → “RFX3”). Our encoding formalism constitutes a straightforward approach to control such nuances and unify all EE sources. To force a network to learn the connection between linguistic phenomena and event modifiers, we consistently report the latter in an expanded version, standardizing the names in case of inconsistencies (e.g., “Negation” → “Polarity=Negative”). We eliminate duplicate events, instances with annotation errors (e.g., references to undefined entities) or with nesting cycles. If multiple overlapping linearizations from different datasets correspond to the same event mention, we keep the longest and most complete one.

For BIOE2T (verbalization), we map each textualized graph with its mention. At this juncture, it is essential to clarify that, with the term “event mention”, we refer to the complete sentences that describe all the components of a certain event and therefore contain all the offsets related to its triggers and arguments. Note that an event mention (generation target) can be longer than one sentence. Linearizations of events sharing the same text span (i.e., double tagging) are decomposed in multiple records. Poorly represented event types (less than three occurrences) are discarded. Similarly, single-node (trigger only) events are ignored since predicting entire sentences from such a little context would be unreasonable. Using stratified random sampling, we split data in training, validation, and test sets with a 90-5-5 proportion. We stratify on multiple variables: (i) the source dataset; (ii) the event type (the main one in case of nesting, i.e., the graph root); (iii) the event mention length.

BIOT2E (parsing) is specular, except to include a balanced number of negative examples, manage double tagging by concatenating linearized events, and not be filtered, thus enabling 1:N extractions. We map a PubMed sentence to a target linearization, if present, or to an empty string otherwise. Hence, we perform BEE at a sentence level, but we do not exclude the investigation of document granularities in future works thanks to efficient transformers (Tay et al., 2020)—e.g., LongT5 (Guo et al., 2022). By accommodating these steps, we treat originally

distinct tasks as different datasets of the more general BEE task.

4.2 Data Properties

We devote the last part of this section to quantitatively analyze the composition of BIOE2T and BIOT2E. Basic statistics are shown in Table 2. Note that the total number of unique event, entity, role, and modifier types are ~ 170 , ~ 150 , 19, and 6, respectively, considerably larger than those in previous standalone corpora (Frisoni et al., 2021). Figure 3 shows the distribution of event graph sizes and mention lengths, skewed with a long tail.

		Train		Valid		Test		All	
# Pairs		61,319		3,407		3,407		68,133	
		36,635		2,035		2,036		40,706	
# Event types		166	168	95	90	96	96	166	170
# Entity types		148	141	81	66	85	72	150	142
# Argument role types		19	19	15	16	15	14	19	19
# Modifier types		6	6	6	6	6	6	6	6
# Nodes per event	min	2	1	2	1	2	1	2	1
	mean	4.29	3.65	4.31	3.66	4.30	3.63	4.30	3.65
	max	35	25	35	20	31	19	35	25
# Modifiers per event	min	0	0	0	0	0	0	0	0
	mean	2.43	2.26	2.44	2.35	2.40	2.25	2.42	2.26
	max	5	5	5	5	5	5	5	5
# Sentences per event mention	min	1		1		1		1	
	mean	1.19		1.19		1.19		1.19	
	max	3		3		3		3	
# Tokens per event mention	min	6	2	8	2	11	2	6	2
	mean	58.55	38.57	58.61	37.67	58.66	39.18	58.56	38.56
	max	301	301	212	161	301	174	301	301
# Events per sentence	min	0		0		0		0	
	mean	1.40		1.39		1.43		1.40	
	max	28		15		24		28	

Table 2: Basic statistics about our BIOE2T and BIOT2E (blue text) datasets.

5 Experimental Setup

In this section, we provide the formal definition of text-to-event parsing and event-to-text generation. Then, describe the setup of the experiments we conducted to evaluate our framework in both tasks.

5.1 Tasks

We see event graph extraction and verbalization as bidirectional transduction tasks via conditional generation, similarly to machine translation.

Training an **event parser** means finding a set of parameters θ_P for a model f that predicts an event graph \hat{e} given a text span s :

$$\hat{e} = \operatorname{argmax}_e f(e|s; \theta_P). \quad (1)$$

Training an **event mention generator** require finding a set of parameters θ_G for a model f that predicts a text span \hat{s} given an event graph e :

$$\hat{s} = \operatorname{argmax}_s f(s|e; \theta_G). \quad (2)$$

In both cases, we use the same family of predictors f (i.e., architectural symmetry without dataset- or task-dependent modifications) by means of seq2seq models. We focus on two data settings: (i) multiple datasets for the same task (multi-dataset based on BIO2E and BIOE2T with independent parameters θ_P and θ_G), and (ii) all datasets across different tasks (multi-task with shared parameters).

5.2 Models

Given the above-reframed definition of EE and EGV, we employ encoder-decoder architectures to autoregressively predict the target sequence y conditioned on the input sequence x :

$$p(y|x) = \prod_{i=1}^{|y|} p(y_i|y_{<i}, x), \quad (3)$$

where $y_{<i} = y_1 \dots y_{i-1}$ and $p(y_i|y_{<i}, x)$ is the probability over the target vocabulary \mathcal{V} normalized by $\operatorname{softmax}(\cdot)$. Because most of the tokens in linearized event representations are also natural language words, we investigate two PLMs with different capacities, aiming to reuse their general text and world knowledge: T5-Base (Raffel et al., 2020) and BART-Base (Lewis et al., 2020). Details about models, training, and hardware configurations are listed in §A.3. According to our literature review, T5 and BART are the two leading generative models adopted in this field. Basically, they are both transformer-based models (with a subword vocabulary) pre-trained on massive corpora through a denoising self-supervised task, i.e., reconstruction of artificially corrupted spans. T5 comes pre-trained also on a multi-task mixture of text-to-text supervised tasks, but none of these include language generation from structured data. As remarked by other researchers for AMR (Bevilacqua et al., 2021), we hypothesize that denoising pre-training is beneficial for EE and EGV. Linearized events can be seen as reordered and partially corrupted sentences that a model must reconstruct, and vice versa. Given a training dataset $\mathcal{D} = \{(x_1, y_1), \dots, (x_{|D|}, y_{|D|})\}$, the learning objective is the negative log-likelihood (teacher forcing):

$$\mathcal{L} = - \sum_{(x,y) \in \mathcal{D}} \log p(y|x, \theta). \quad (4)$$

5.3 Evaluation

Parsing. While training is based on a likelihood objective, we assess EE models using standard pre-

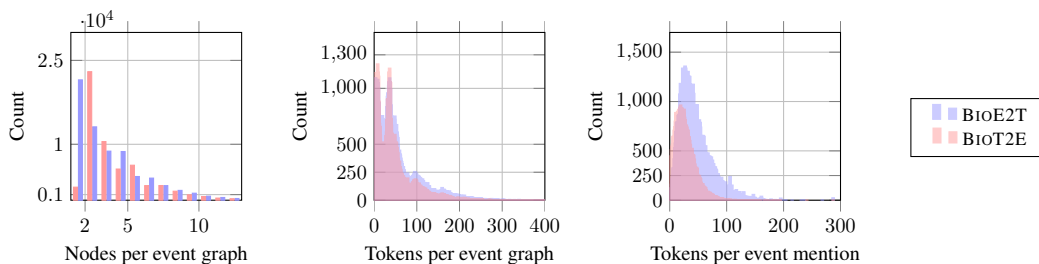


Figure 3: Distribution of instance origins, event graph sizes, and event mention lengths across our BIOE2T and BIOT2E datasets. Tokens refer to the T5 vocabulary.

cision, recall, and F1 scores according to the “approximate recursive matching” criterion (Kim et al., 2011) with string correspondence equality. Since our models stand on free generation, the derived event annotations are not accompanied by offset indices communicating their position in the original documents, preventing the “approximate span matching” relaxation. To avoid introducing error sources affecting results interpretation, we do not apply fragile heuristics such as likelihood-based class predictions (Paolini et al., 2021) or offset reconstruction (Lu et al., 2021b). In fact, (i) BIOT2E has a high type heterogeneity overhead; (ii) many sequences mention multiple events with the same trigger and scattering arguments, making difficult to assume that the matching argument-utterance is the one closest to the trigger.

Generation. To quantitatively compare predictions against ground truth literature sentences on the test set, we use a broad spectrum of natural language generation (NLG) evaluation metrics. We deepen them in §A.4 and refer the reader to (Celikyilmaz et al., 2020) for further details on their properties. In line with previous graph-to-text works, we include BLEURT (Sellam et al., 2020), a recent regression-based measure showing an higher correlation with human judgments than other simple yet widespread n-gram-overlap-based metrics.

6 Results

6.1 Event Extraction

Multi-dataset and Text-to-Text EE. Table 3 summarizes the BEE F1-scores of our end-to-end models trained on BIOT2E when evaluated on the validation set of the individual tasks⁵. We report complete precision and recall results in §A.5. Baseline systems have been assessed on the official

⁵Task organizers’ servers for test set evaluation are currently non-available. Accessed on May 9th, 2022.

BEE datasets—following a standard $\langle .txt, .a1, .a2 \rangle$ structure—provided by each BioNLP shared task. They adopt discriminative architectures, meaning they train a distinct model for each task by relying on benchmark-specific event/entity/role target classes covered in *.a1* and *.a2* files. Despite conducting training on all tasks at once (unified thanks to a text-to-text format), we separately evaluate our models on each validation set, allowing for a fair comparison with the baseline. Thanks to knowledge sharing among several biomedical subareas and seq2seq, we significantly push the state-of-the-art on all the benchmarks, with T5 empirically producing better results than BART. Compared to the solutions previously known in the literature (Frisoni et al., 2021), our framework’s main advantage is a higher recall and generalization capacity. We lay out a detailed error analysis in §A.6.

Low-resource. We experiment on the CG13 dataset, using only 1% to 10% of the training data (Figure 4, §A.3). F1-scores in such low-resource regime demonstrate that our framework is data-efficient compared to DeepEventMine, the SOTA discriminative model for BEE.

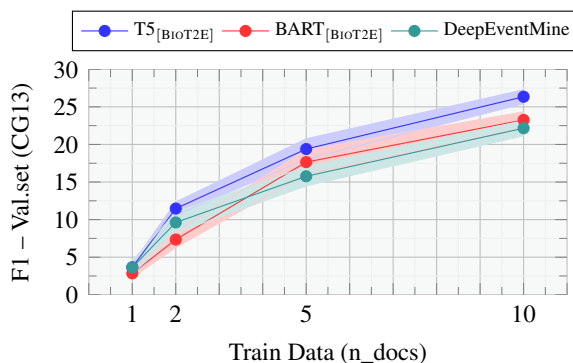


Figure 4: F1 comparison of our proposed models and DeepEventMine with train down-sampling on the CG13 validation set. Mean and standard deviation over 3 runs.

		Datasets									
		GE08	GE11	EPI11	ID11	MLEE	GE13	CG13	PC13	GRO13	GENIA-MK
single-task	Works										
	Shared task winner	43.12†	55.90	56.41	50.10	–	50.74†	55.41†	51.10†	22.00†	–
	Trieu et al. (2020) w/o gold entities	–	56.64	55.81	50.10	51.73	45.95	54.27	50.53	–	–
multi-task	Abdulkadhar et al. (2021)	63.09†	61.74†	–	–	–	58.30†	–	–	–	61.58†
	Ours										
	T5-Base _[BioT2E]	70.74	73.62	84.43	84.13	79.91	81.18	80.10	83.19	81.91	83.24
	BART-Base _[BioT2E]	68.50	69.55	78.79	78.16	73.82	73.84	72.05	73.20	71.79	75.25

Table 3: F1-score (%) performance comparison on the validation set of the most significant biomedical event extraction tasks (eight BioNLP-STs, MLEE, and GENIA-MK). Top: original BioNLP-ST winning results and current SOTA neural systems (with per-task models); Bottom: proposed framework (with multi-task models). † indicates test set results if validation ones are unavailable. The highest scores are bolded. Both our models significantly outperform competitors (student t-test, $p < 0.05$).

6.2 Event Graph Verbalization

NLG metrics. In Table 4, we show the event-to-text results achieved by T5 and BART on the overall BIOE2T test set. Since there are many ways to express the same symbolic concept, we use beam search at inference time to return all the different top beam sequences (i.e., multi-output). To give additional insights on generative performance, we apply metrics to all target-output pairs and not only to the one with the highest log-likelihood. T5-Base performs the best across all the NLG metrics, which—despite capturing different dimensions (grammatical correctness, fluency, informativeness, adequacy, etc.)—prove to be consistent with each other. Interestingly, we observe a relevant score gap between max-likelihood and max-score selection within a beam. This is strong evidence of the decoding strategy impact (often overlooked), also reinforcing the hypothesis that high quality human language does not follow a distribution of high probability next words (Holtzman et al., 2020). Moreover, it should be emphasized how this detachment is much more attenuated with evaluations closer to a semantic level. Sequences generated via beam search tend to be syntactically different (albeit moderate) but semantically similar, underlining the importance of metrics to grasp meaning preservation. From qualitative investigations, both models displays promising abilities in translating modifiers in elements of language (§A.7).

Graph Structure and Output Length Impact.

Figure 5 shows the effect of the event graph size on verbalization, measured with BLEURT, abstractness (Gehrmann et al., 2019), and repetitiveness (Peyrard et al., 2017)⁶. In this experiment, we aver-

⁶Abstractness: percentage of new n-grams in the predictions, compared to the references. Repetitiveness: average

	T5 _[BioE2T]		BART _[BioE2T]	
	MAX_L	MAX_S	MAX_L	MAX_S
BLEU	63.8	69.6 (+5.8)	53.1	59.6 (+6.5)
ROUGE-1	68.8	73.9 (+5.1)	60.0	65.6 (+5.6)
ROUGE-2	61.3	66.7 (+5.4)	49.8	55.7 (+5.9)
ROUGE-L	66.1	71.2 (+5.1)	56.2	61.8 (+5.6)
METEOR	66.6	72.1 (+5.5)	56.3	64.4 (+6.1)
BLEURT	68.9	73.5 (+4.6)	59.8	64.4 (+4.6)
NUBIA	65.2	73.1 (+7.9)	56.2	64.3 (+8.1)
BERTSCORE	94.1	95.0 (+0.9)	92.3	93.4 (+1.1)
BARTSCORE	-2.5	-1.7 (+0.8)	-2.31	-1.31 (+1.0)

Table 4: Event-to-text generation results on the BIOE2T test set. We show the average metric score considering the sequence with maximum likelihood (MAX_L) within a beam and the one obtainable by taking the sequence with maximum ground-truth-match according to the metric of interest (MAX_S). The gap between the two is shown in round brackets.

age the metric score for all the generated sequences and divide the results by the node number. We find that sequence quality increases as the event graph size increases, following a logarithmic function. This behavior is justified by the fact that BIOE2T, differently from other datasets like WebNLG, contains similar text lengths for various graph sizes. When the input is a larger event graph, the model has more contextual information to be leveraged during the generation, approaching the target syntactically and semantically. This thesis is also supported by the decline in abstraction, while repetitiveness is generally low and appears proportional to prediction length.

6.3 Multi-task Setting

Our method naturally allows us to train a single model on multiple datasets covering different NLP tasks, besides EE and EGV. In this setting, we use number of n-grams with at least one repetition in the generated sequences. We scan word-level unigrams.

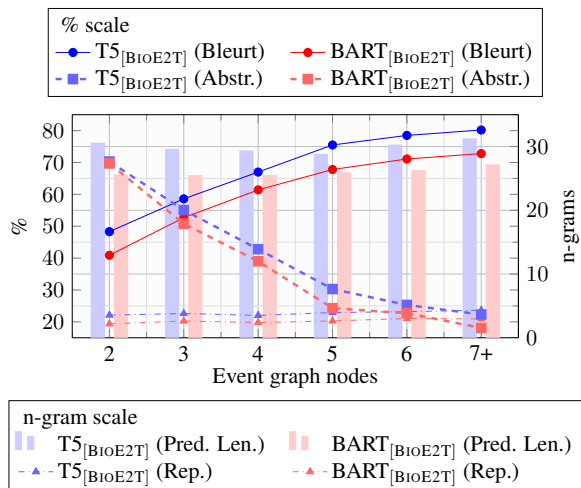


Figure 5: Average BLEURT score, abstractness, repetitiveness, and prediction length compared to the size of the event graph to condition on.

a task-specific prefix (e.g., “extract events:”) to let the model know the requested transformation for each input. In particular, we inspect the performance gap on the PUBMED dataset (Cohan et al., 2018) for single document summarization, revealing a fairly advantage in terms of ROUGE (Table 5). Our intuition is that both EE (with its event-to-text back translation) and summarization tasks aim to distill salient information from massive text, providing complementary features for each other that can be beneficial for general NLP.

	T5-Base (R / P / F1)	T5-Base Event-driven MTL (R / P / F1)
ROUGE-1	27.24 / 58.11 / 32.97	33.46 / 47.94 / 39.41
ROUGE-2	10.84 / 23.82 / 13.15	12.12 / 17.92 / 14.46
ROUGE-L	18.08 / 40.29 / 22.03	22.19 / 32.01 / 26.21

Table 5: Single document summarization performance on PUBMED test set w/o and w/ event-driven MTL.

7 Conclusion

This paper presented the first sequence-to-sequence framework for both biomedical event extraction and verbalization. Concretely, we proposed BIOT2E and BIOE2T, two highly comprehensive datasets with parallel text-event gold annotations, constructed through a novel linearization technique. By training autoregressive language models on them, we achieved an average F1-score of 0.81 on ten benchmarks, making considerable improvements over previously published work. In stark contrast with discriminative solutions, we employed the same architecture to perform previously distinct tasks, exploiting pre-trained knowledge and

label semantics. Experimental results also proved the usefulness of (i) knowledge sharing between different biomedical spheres in event-based tasks, (ii) events in improving model understanding for NLP tasks in general, like document summarization. We hope that our contributions will lead to further progress in natural language understanding and generation as transfer learning becomes even more vital for graph-to-text and text-to-graph translations.

Future directions At the edge of our knowledge, this is the first work that proposes single deep neural models capable of effectively extracting (and back-translating to text) such a variety of biomedical events and their components. This high ontological coverage opens the door to numerous applications and research blueprints. Future work should tackle: (i) document-level granularities; (ii) prompting-based purely generative models (Ma et al., 2022); (iii) text \leftrightarrow graph boosting approaches echoing autoencoders and Cycle-GT (Guo et al., 2020); (iv) few-shot learning; (v) event aggregation towards automatic corpus-level knowledge graph learning (Frisoni et al., 2020a; Frisoni and Moro, 2020; Frisoni et al., 2020c,b); (vi) conversion of events to logic and constrained decoding algorithms (Lu et al., 2021a); (vii) infusion of events in pre-trained language models for tasks like biomedical multi-document summarization (Moro et al., 2022) and information retrieval (Moro and Valgimigli, 2021).

8 Ethical Considerations

Largely pre-trained language models that we reference in our study might perpetuate and exacerbate biases and stereotypes hardwired in the training data, risking generating false or misleading information (Zhang et al., 2020a; Nadeem et al., 2021). Healthcare, in particular, requires strong guarantees about the factuality and reliability of predictions, but current state-of-the-art NLP solutions cannot establish such assurance. We acknowledge these issues and caution those who build on our framework to consider the aforementioned implications before deploying systems in the real world. Although automatic extraction of semantic relations from scientific documents is fundamental in the biomedical field, we do not encourage users to employ our models, like previous ones, for critical applications at present performance levels. No sensitive information is contained within our datasets,

which are derived from publicly and openly available PubMed articles. We honor and support the ACL Code of Ethics.

Acknowledgements

We would like to thank all the anonymous reviewers for their constructive feedback and valuable comments. We thank Paolo Italiani for training the DeepEventMine models and assisting us during low-resource experiments.

References

- Sabenabanu Abdulkadhar, Balu Bhasuran, and Jeyakumar Natarajan. 2021. Multiscale laplacian graph kernel combined with lexico-syntactic patterns for biomedical event extraction from literature. *Knowledge and Information Systems*, 63(1):143–173.
- Oshin Agarwal, Heming Ge, Siamak Shakeri, and Rami Al-Rfou. 2021. Knowledge graph based synthetic corpus generation for knowledge-enhanced language model pre-training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3554–3565, Online. Association for Computational Linguistics.
- Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.
- Jonathan Berant, Vivek Srikumar, Pei-Chun Chen, Abby Vander Linden, Brittany Harding, Brad Huang, Peter Clark, and Christopher D. Manning. 2014. Modeling Biological Processes for Reading Comprehension. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1499–1510, Doha, Qatar. Association for Computational Linguistics.
- Michele Bevilacqua, Rexhina Blloshmi, and Roberto Navigli. 2021. One SPRING to rule them both: Symmetric AMR semantic parsing and generation without a complex pipeline. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 12564–12573. AAAI Press.
- Jari Björne, Filip Ginter, Sampo Pyysalo, Jun’ichi Tsujii, and Tapio Salakoski. 2010. Complex event extraction at PubMed scale. *Bioinform.*, 26(12):382–390.
- Jari Björne and Tapio Salakoski. 2018. Biomedical event extraction using convolutional neural networks and dependency parsing. In *Proceedings of the BioNLP 2018 workshop*, pages 98–108, Melbourne, Australia. Association for Computational Linguistics.
- Rexhina Blloshmi, Simone Conia, Rocco Tripodi, and Roberto Navigli. 2021. Generating senses and roles: An end-to-end model for dependency- and span-based semantic role labeling. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pages 3786–3793. ijcai.org.
- Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. 2020. Evaluation of text generation: A survey. *CoRR*, abs/2006.14799.
- Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 615–621, New Orleans, Louisiana. Association for Computational Linguistics.
- K. Bretonnel Cohen, Karin Verspoor, Helen Johnson, Chris Roeder, Philip Ogren, William Baumgartner, Elizabeth White, and Lawrence Hunter. 2009. High-precision biological event extraction with a concept recognizer. In *Proceedings of the BioNLP 2009 Workshop Companion Volume for Shared Task*, pages 50–58, Boulder, Colorado. Association for Computational Linguistics.
- Pierre Colombo, Chloé Clavel, and Pablo Piantanida. 2021. Infoml: A new metric to evaluate summarization & data2text generation. *CoRR*, abs/2112.01589.
- Hady Elsahar, Pavlos Vougiouklis, Arslan Remaci, Christophe Gravier, Jonathon Hare, Frederique Laforest, and Elena Simperl. 2018. T-REx: A large scale alignment of natural language with knowledge base triples. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).
- Giacomo Frisoni., Paolo Italiani., Francesco Boschi., and Gianluca Moro. 2022. Enhancing biomedical scientific reviews summarization with graph-based factual evidence extracted from papers. In *DATA*, pages 168–179. INSTICC, SciTePress.
- Giacomo Frisoni and Gianluca Moro. 2020. Phenomena Explanation from Text: Unsupervised Learning of Interpretable and Statistically Significant Knowledge. In *DATA (Revised Selected Papers)*, volume 1446, pages 293–318. Springer.

- Giacomo Frisoni, Gianluca Moro, and Antonella Carbonaro. 2020a. [Learning Interpretable and Statistically Significant Knowledge from Unlabeled Corpora of Social Text Messages: A Novel Methodology of Descriptive Text Mining](#). In *DATA 2020 - Proc. 9th Int. Conf. Data Science, Technol. and Appl.*, pages 121–134. SciTePress.
- Giacomo Frisoni, Gianluca Moro, and Antonella Carbonaro. 2020b. [Towards Rare Disease Knowledge Graph Learning from Social Posts of Patients](#). In *RiiForum*, pages 577–589. Springer.
- Giacomo Frisoni, Gianluca Moro, and Antonella Carbonaro. 2020c. [Unsupervised Descriptive Text Mining for Knowledge Graph Learning](#). In *IC3K 2020 - Proc. 12th Int. Joint Conf. Knowl. Discovery, Knowl. Eng. and Knowl. Manage.*, volume 1, pages 316–324. SciTePress.
- Giacomo Frisoni, Gianluca Moro, and Antonella Carbonaro. 2021. [A survey on event extraction for natural language understanding: Riding the biomedical literature wave](#). *IEEE Access*, 9:160721–160757.
- Giacomo Frisoni, Gianluca Moro, Giulio Carlassare, and Antonella Carbonaro. 2022. [Unsupervised event graph representation and similarity learning on biomedical literature](#). *Sensors*, 22(1):3.
- Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017. [The WebNLG challenge: Generating text from RDF data](#). In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 124–133, Santiago de Compostela, Spain. Association for Computational Linguistics.
- Sebastian Gehrmann, Zachary Ziegler, and Alexander Rush. 2019. [Generating abstractive summaries with finetuned language models](#). In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 516–522, Tokyo, Japan. Association for Computational Linguistics.
- David Geleta, Andriy Nikolov, Gavin Edwards, Anna Gogleva, Richard Jackson, Erik Jansson, Andrej Lamov, Sebastian Nilsson, Marina Pettersson, Vladimir Poroshin, Benedek Rozemberczki, Timothy Scrivener, Michael Ughetto, and Eliseo Papa. 2021. [Biological insights knowledge graph: an integrated knowledge graph to support drug development](#). *bioRxiv*.
- Mandy Guo, Joshua Ainslie, David C. Uthus, Santiago Ontañón, Jianmo Ni, Yun-Hsuan Sung, and Yinfei Yang. 2022. [Longt5: Efficient text-to-text transformer for long sequences](#). In *NAACL-HLT (Findings)*, pages 724–736. Association for Computational Linguistics.
- Qipeng Guo, Zhijing Jin, Xipeng Qiu, Weinan Zhang, David Wipf, and Zheng Zhang. 2020. [CycleGT: Unsupervised graph-to-text and text-to-graph generation via cycle training](#). In *Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+)*, pages 77–88, Dublin, Ireland (Virtual). Association for Computational Linguistics.
- Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. [The curious case of neural text degeneration](#). In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.
- Pere-Lluís Hugué Cabot and Roberto Navigli. 2021. [REBEL: Relation extraction by end-to-end language generation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2370–2381, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Mihir Kale and Abhinav Rastogi. 2020. [Text-to-text pre-training for data-to-text tasks](#). In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 97–102, Dublin, Ireland. Association for Computational Linguistics.
- Katikapalli Subramanyam Kalyan, Ajit Rajasekharan, and Sivanesan Sangeetha. 2021. [AMMUS : A survey of transformer-based pretrained models in natural language processing](#). *CoRR*, abs/2108.05542.
- Jin-Dong Kim, Claire Nédellec, Robert Bossy, and Louise Deléger, editors. 2019. *Proceedings of The 5th Workshop on BioNLP Open Shared Tasks*. Association for Computational Linguistics, Hong Kong, China.
- Jin-Dong Kim, Ngan L. T. Nguyen, Yue Wang, Jun’ichi Tsujii, Toshihisa Takagi, and Akinori Yonezawa. 2012. [The Genia Event and Protein Coreference tasks of the BioNLP Shared Task 2011](#). *BMC Bioinform.*, 13(S-11):S1.
- Jin-Dong Kim, Tomoko Ohta, Kanae Oda, and Jun’ichi Tsujii. 2008. [From Text to Pathway: Corpus Annotation for Knowledge Acquisition from Biomedical Literature](#). In *APBC*, volume 6 of *Advances in Bioinformatics and Computational Biology*, pages 165–176. Imperial College Press.
- Jin-Dong Kim, Tomoko Ohta, Sampo Pyysalo, Yoshinobu Kano, and Jun’ichi Tsujii. 2011. [Extracting bio-molecular events from literature - the bionlp’09 shared task](#). *Comput. Intell.*, 27(4):513–540.
- Jin-Dong Kim, Yue Wang, and Yamamoto Yasunori. 2013a. [The Genia event extraction shared task, 2013 edition - overview](#). In *Proceedings of the BioNLP Shared Task 2013 Workshop*, pages 8–15, Sofia, Bulgaria. Association for Computational Linguistics.
- Jung-jae Kim, Xu Han, Vivian Lee, and Dietrich Rebholz-Schuhmann. 2013b. [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. Association for Computational Linguistics.

- Rik Koncel-Kedziorski, Dhanush Bekal, Yi Luan, Mirella Lapata, and Hannaneh Hajishirzi. 2019. [Text Generation from Knowledge Graphs with Graph Transformers](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2284–2293, Minneapolis, Minnesota. Association for Computational Linguistics.
- Ioannis Konstas, Srinivasan Iyer, Mark Yatskar, Yejin Choi, and Luke Zettlemoyer. 2017. [Neural AMR: Sequence-to-sequence models for parsing and generation](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 146–157, Vancouver, Canada. Association for Computational Linguistics.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. [BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.
- Diya Li, Lifu Huang, Heng Ji, and Jiawei Han. 2019. [Biomedical event extraction based on knowledge-driven tree-LSTM](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1421–1430, Minneapolis, Minnesota. Association for Computational Linguistics.
- Ximing Lu, Peter West, Rowan Zellers, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021a. [Neurologic decoding: \(un\)supervised neural text generation with predicate logic constraints](#). In *NAACL-HLT*, pages 4288–4299. Association for Computational Linguistics.
- Yaojie Lu, Hongyu Lin, Jin Xu, Xianpei Han, Jialong Tang, Annan Li, Le Sun, Meng Liao, and Shaoyi Chen. 2021b. [Text2Event: Controllable sequence-to-structure generation for end-to-end event extraction](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2795–2806, Online. Association for Computational Linguistics.
- Mingyu Derek Ma, Alex Taylor, Wei Wang, and Nanyun Peng. 2022. [DICE: data-efficient clinical event extraction with generative models](#). *CoRR*, abs/2208.07989.
- I. A. Mel'čuk. 1973. [Towards a Linguistic 'Meaning ↔ Text' Model](#), pages 33–57. Springer Netherlands, Dordrecht.
- Makoto Miwa, Sampo Pyysalo, Tomoko Ohta, and Sophia Ananiadou. 2013. [Wide coverage biomedical event extraction using multiple partially overlapping corpora](#). *BMC Bioinform.*, 14:175.
- Makoto Miwa, Paul Thompson, John McNaught, Douglas B. Kell, and Sophia Ananiadou. 2012. [Extracting semantically enriched events from biomedical literature](#). *BMC Bioinform.*, 13:108.
- Gianluca Moro and Luca Ragazzi. 2022. [Semantic self-segmentation for abstractive summarization of long documents in low-resource regimes](#). In *AAAI*, pages 11085–11093. AAAI Press.
- Gianluca Moro, Luca Ragazzi, Lorenzo Valgimigli, and Davide Freddi. 2022. [Discriminative marginalized probabilistic neural method for multi-document summarization of medical literature](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 180–189, Dublin, Ireland. Association for Computational Linguistics.
- Gianluca Moro and Lorenzo Valgimigli. 2021. [Efficient self-supervised metric information retrieval: A bibliography based method applied to COVID literature](#). *Sensors*, 21(19).
- Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. [StereoSet: Measuring stereotypical bias in pretrained language models](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371, Online. Association for Computational Linguistics.
- 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, Sofia, Bulgaria. Association for Computational Linguistics.
- Tomoko Ohta, Sampo Pyysalo, and Jun'ichi Tsujii. 2011. [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. Association for Computational Linguistics.
- Giovanni Paolini, Ben Athiwaratkun, Jason Krone, Jie Ma, Alessandro Achille, Rishita Anubhai, Cícero Nogueira dos Santos, Bing Xiang, and Stefano Soatto. 2021. [Structured prediction as translation between augmented natural languages](#). In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.
- Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and

- Alexander Miller. 2019. [Language Models as Knowledge Bases?](#) In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.
- Maxime Peyrard, Teresa Botschen, and Iryna Gurevych. 2017. [Learning to score system summaries for better content selection evaluation.](#) In *Proceedings of the Workshop on New Frontiers in Summarization*, pages 74–84, Copenhagen, Denmark. Association for Computational Linguistics.
- Sampo Pyysalo, Tomoko Ohta, and Sophia Ananiadou. 2013. [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. Association for Computational Linguistics.
- Sampo Pyysalo, Tomoko Ohta, Makoto Miwa, Hanchchol Cho, Junichi Tsujii, and Sophia Ananiadou. 2012. Event extraction across multiple levels of biological organization. *Bioinform.*, 28(18):575–581.
- Sampo Pyysalo, Tomoko Ohta, Rafal Rak, Dan Sullivan, Chunhong Mao, Chunxia Wang, Bruno Sobral, Jun’ichi Tsujii, and Sophia Ananiadou. 2011. [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. Association for Computational Linguistics.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. [Exploring the limits of transfer learning with a unified text-to-text transformer.](#) *J. Mach. Learn. Res.*, 21:140:1–140:67.
- Alan Ramponi, Rob van der Goot, Rosario Lombardo, and Barbara Plank. 2020. [Biomedical event extraction as sequence labeling.](#) In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5357–5367, Online. Association for Computational Linguistics.
- Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. [How much knowledge can you pack into the parameters of a language model?](#) In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5418–5426, Online. Association for Computational Linguistics.
- Susan H Rodger and Thomas W Finley. 2006. *JFLAP: an interactive formal languages and automata package.* Jones & Bartlett Learning.
- Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. [BLEURT: Learning robust metrics for text generation.](#) In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.
- Pranab Kumar Sen. 1968. Estimates of the regression coefficient based on kendall’s tau. *Journal of the American statistical association*, 63(324):1379–1389.
- Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler. 2020. [Efficient transformers: A survey.](#) *CoRR*, abs/2009.06732.
- Hai-Long Trieu, Thy Thy Tran, Anh-Khoa Duong Nguyen, Anh Nguyen, Makoto Miwa, and Sophia Ananiadou. 2020. DeepEventMine: end-to-end neural nested event extraction from biomedical texts. *Bioinform.*, 36(19):4910–4917.
- Luyu Wang, Yujia Li, Ozlem Aslan, and Oriol Vinyals. 2021a. [WikiGraphs: A Wikipedia text - knowledge graph paired dataset.](#) In *Proceedings of the Fifteenth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-15)*, pages 67–82, Mexico City, Mexico. Association for Computational Linguistics.
- Qingyun Wang, Manling Li, and Xuan Wang et al. 2021b. [COVID-19 literature knowledge graph construction and drug repurposing report generation.](#) In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 66–77. Association for Computational Linguistics.
- Qingyun Wang, Semih Yavuz, Xi Victoria Lin, Heng Ji, and Nazneen Rajani. 2021c. [Stage-wise fine-tuning for graph-to-text generation.](#) In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 16–22, Online. Association for Computational Linguistics.
- Zequi Wu, Michel Galley, Chris Brockett, Yizhe Zhang, Xiang Gao, Chris Quirk, Rik Koncel-Kedziorski, Jianfeng Gao, Hannaneh Hajishirzi, Mari Ostendorf, and Bill Dolan. 2021. [A controllable model of grounded response generation.](#) In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 14085–14093. AAAI Press.
- Haoran Zhang, Amy X. Lu, Mohamed Abdalla, Matthew B. A. McDermott, and Marzyeh Ghassemi. 2020a. [Hurtful words: quantifying biases in clinical contextual word embeddings.](#) In *ACM CHIL ’20: ACM Conference on Health, Inference, and Learning, Toronto, Ontario, Canada, April 2-4, 2020 [delayed]*, pages 110–120. ACM.
- Junsheng Zhang, Kun Li, Changqing Yao, and Yunchuan Sun. 2020b. Event-based summarization method for scientific literature. *Personal and Ubiquitous Computing*, pages 1–10.

Tianran Zhang, Muhao Chen, and Alex A. T. Bui. 2020c. [Diagnostic Prediction with Sequence-of-sets Representation Learning for Clinical Events](#). In *Artificial Intelligence in Medicine - 18th International Conference on Artificial Intelligence in Medicine, AIME 2020, Minneapolis, MN, USA, August 25-28, 2020, Proceedings*, volume 12299 of *Lecture Notes in Computer Science*, pages 348–358. Springer.

Chao Zhao, Marilyn Walker, and Snigdha Chaturvedi. 2020. [Bridging the structural gap between encoding and decoding for data-to-text generation](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2481–2491, Online. Association for Computational Linguistics.

Weizhong Zhao, Jinyong Zhang, Jincui Yang, Tingting He, Huifang Ma, and Zhixin Li. 2021. [A novel joint biomedical event extraction framework via two-level modeling of documents](#). *Inf. Sci.*, 550:27–40.

A Appendix

A.1 Formal Event Grammar

Linearized events follow the formal context-free grammar orderly detailed in Table 6.

A.2 Insights on Event Modifiers

Table 7 recaps the event modifiers covered by our work, their meaning and possible values.

A.3 Training Details and Reproducibility

T5 and BART. We reimplemented T5-Base (~220M parameters, 12-layers, 768-hidden, 12-heads) in Flax (T5X) starting from the Google Research codebase⁷ and built our BART-Base (~139M, 12-layers, 768-hidden, 16-heads) model in PyTorch using the HuggingFace’s Transformers library⁸. For all variants, weights are initialized through the official checkpoints (C4 pre-training for T5). For verbalization, we set the maximum length for event mentions and linearized event graphs to 200 and 400, respectively. For parsing, we extended the linearization maximum length to 650. Instead, we used 1024 and 256 for single-document summarization input/output (truncated). We used BF16 mixed precision and a batch size of 16 (with gradient accumulation every 2 batches) for all models. We employed the Adam optimizer. Following (Raffel et al., 2020), T5 models are fine-tuned with a constant learning rate of 0.001, 1000 warmup steps, and a 0.1 dropout rate. For BART, we used default hyperparameters, but we did not

⁷<https://github.com/google-research/t5x>

⁸https://huggingface.co/transformers/model_doc/bart.html

penalize the model for the generation of repeated ngrams, e.g., multiple opening or closing brackets. We chose the best checkpoints based on the ROUGE score on the validation set; we found that it highly correlates with EE metrics (due to the extractive nature of the task). At prediction time, we used beam search with beam size 4 for generation and greedy decoding for parsing. We trained single-task T5 and BART models for 50 epochs (≈ 40 and ≈ 30 hours per full-training on BIOE2T and BIOT2E, respectively). The estimated⁹ CO2 impact incurred by each model training belongs to the range [6.09, 8.12] kg (carbon footprint). Regarding T5 MTL, we prepended task-specific tags to the input records and performed 10 epochs using a mixture with 100% data proportion sampling for each task.

DeepEventMine. We reimplemented the training script (not released by the authors, accessed on January 16th, 2022), faithfully following the steps listed in the paper (Trieu et al., 2020). For comparison, we modified the original evaluation script to assess predictions without gold entities (not used by our framework).

Low-resource. As outlined in §6.1, we experimented on the CG13 dataset with only a limited portion of the training set available. We selected 1, 2, 5, and 10 PubMed abstracts with an average number of mentioned events. To account for the small dataset size, we fine-tuned on CG13 for a greater number of epochs, proportional to the size of each partition (50x, 25x, 10x, 5x). So, we trained T5 and BART for 2.500, 1.250, 500, and 250 epochs; DeepEventMine for 4.000, 2.000, 800, and 400 epochs. We performed 3 runs (each model being fine-tuned on the same 4 subsets of the training set and then evaluated on the entire validation set).

Hardware Setup. We ran each experiment on a workstation having two Nvidia GeForce RTX 3090 GPUs with 24GB of dedicated memory each, 64GB of RAM, and a Intel® Core™ i9-10900X CPU @ 3.70GHz.

A.4 NLG Evaluation Metrics

Metrics (default parameters and official repository implementation) are summarized in Table 8. As for BARTSCORE and BLEURT, we used the BLEURT-20 and BARTSCORE-CNNM pre-trained models, respectively.

⁹<http://green-algorithms.org/>

Symbol	Description
EV	event
T	trigger
A	argument
TST	text span trigger
TRG	trigger role group
TSE	text span entity
EVT	event type
MG	modifier group
M	modifier name
MV	modifier value
E	entity
ET	entity type
RG	role group
R	role

(a) Symbols in V

Symbol	Type Set
<i>tst</i>	$\wedge[A-Za-z0-9]^+$
<i>tse</i>	$\wedge[A-Za-z0-9]^+$
<i>evt</i>	<i>EventTypes</i>
<i>m</i>	<i>Modifiers</i>
<i>mv</i>	<i>ModifierValues</i>
<i>et</i>	<i>EntityType</i>
<i>r</i>	<i>RoleTypes</i>

(b) Symbols in Σ

Id	Rule	Id ↓	Rule ↓
1	$EV \rightarrow T A$	12	$E \rightarrow [TSE \mid ET \mid RG]$
2	$T \rightarrow [TST \mid EVT \mid MG \mid TRG]$	13	$RG \rightarrow RG \mid RG$
3	$MG \rightarrow MG \mid MG$	14	$RG \rightarrow R = TST$
4	$MG \rightarrow \mid M = MV$	15	$TST \rightarrow tst$
5	$MG \rightarrow \epsilon$	16	$TSE \rightarrow tse$
6	$TRG \rightarrow TRG \mid TRG$	17	$EVT \rightarrow evt$
7	$TRG \rightarrow \mid R = TST$	18	$ET \rightarrow et$
8	$TRG \rightarrow \epsilon$	19	$M \rightarrow m$
9	$A \rightarrow A A$	20	$MV \rightarrow mv$
10	$A \rightarrow E$	21	$R \rightarrow r$
11	$A \rightarrow EV$	22	$EV \rightarrow EVEV$

(c) Rules in R

Table 6: Formal definition of the event grammar $G = \langle V, \Sigma, R, EV \rangle$. V is the finite set of variables (a); Σ is the finite set of terminal symbols and therefore the alphabet of our event language (b); R is the finite set of production rules (c); EV is the start variable. As for (b), each symbol on the left belongs to the type set on the right, which depends on the dataset event schema. The only exception concerns *tst* and *tse* which are alphanumeric strings, reported as regex for notational simplicity. The pipe marker is intended as a character and not as a logic operator.

Modifier	Definition	Possible values
Polarity	The truth value of an event	Positive (<i>default</i>) Negative
Speculation	Whether an event is speculated or not	True, False (<i>default</i>)
Source	Origin of the knowledge expressed by an event	Current paper (<i>default</i>) Other
Manner	The intensity level of an event	High, Low, Neutral (<i>default</i>)
Certainty level	The confidence of an event being expressed	L1 (low confidence), L2 (not complete confidence) L3 (high confidence, <i>default</i>)
Knowledge type	The overarching information expressed by the event	Investigation, Observation, Analysis, Fact, Method, Other (<i>default</i>)

Table 7: Summary of the event modifiers in B10T2E and B10E2T.

Metric	U	S	Strategy	Model(s)
BLEU	✓		N-gram recall	–
ROUGE	✓		N-gram precision	–
METEOR	✓		N-gram overlap w/ synonym match	–
BERTSCORE	✓		Semantic similarity	BERT
BARTSCORE	✓		Conditioned generation for faithfulness, precision, and recall	BART
BLEURT		✓	Human score prediction	BERT RoBERTa
NUBIA		✓	Human score prediction	GPT-2

Table 8: Metrics applied for evaluating event graph verbalization performance. **U**: unsupervised, **S**: supervised, based on the need for human judgments to train. They belong to $[0, 1]$, with the exception of BARTSCORE, whose range is $]-\infty, 0]$. The higher the score, the more valid the hypothesis is.

A.5 Detailed Event Extraction Results

We report detailed event extraction performance for our models in Table 9.

A.6 Error Analysis

A.6.1 Event Extraction

We quantitatively classify errors into three broad categories: format, trigger, and argument errors. Further, we organize the latter two in fine-grained categories: under-prediction (i.e., expected but not predicted), over-prediction (i.e., predicted but not expected), and wrong type. Finally, we distinguish the target type, especially keeping track of multi-event outputs and nested (i.e., complex) events. Table 10 reports the proportions of error types we identified. We notice the most considerable fraction of errors is due to triggers. From a closer look, we found that over-predicted triggers are often linked to generic words used very frequently to indicate specific event types. For instance, similarly to what emerged in previous works (Ramponi et al., 2020), T5_[B10E2T] identifies a positive regulation event anchored at “activated” in the sentence: “Tax [...] maximally activated HTLV-I-LTR-CAT and kappa B-fos-CA” albeit the gold standard does not contain the event in this instance. However, we believe these errors are acceptable from a semantic point of view and sometimes highlight a low-annotation problem within the datasets. As for wrong trigger and argument types, the model tends to generate different but semantically equivalent labels, e.g., “sufficient to restore” instead of “restored”, “Protein_molecule” instead of “Protein”. This issue underlines the need for alternative automatic evaluation metrics operating at the semantic level. Format errors are less frequent, proving that the model can successfully manage bracket [...] rules.

Datasets		GE08	GE11	EPI11	ID11	MLEE	GE13	CG13	PC13	GRO13	GENIA-MK
Works	R	67.71	74.57	92.28	90.28	71.56	73.18	70.83	75.68	74.84	75.79
	P	74.05	72.68	77.81	78.77	90.45	91.15	92.17	92.36	90.45	92.32
	F1	70.74	73.62	84.43	84.13	79.91	81.18	80.10	83.19	81.91	83.24
T5-Base _[BIOE2T]	R	65.23	70.82	87.66	84.22	66.84	66.91	64.79	67.15	66.82	69.62
	P	72.12	68.33	71.56	72.91	82.43	82.36	81.13	80.45	77.56	81.87
	F1	68.50	69.55	78.79	78.16	73.82	73.84	72.05	73.20	71.79	75.25

Table 9: Recall (R), Precision (P), and F1-score (%) performance of T5-Base_[BIOE2T] and BART-Base_[BIOE2T] on the validation set of the most significant biomedical event extraction tasks.

Error Type	Fraction		
	All	Nested	Multi-event
Format	5%	2%	3%
Trigger			
Under-prediction	17%	8%	6%
Over-prediction	28%	16%	5%
Wrong type	10%	3%	4%
Argument			
Under-prediction	13%	7%	4%
Over-prediction	23%	14%	5%
Wrong type	4%	2%	1%

Table 10: Quantitative event extraction error analysis of T5_[BIOE2T]. Average fraction values among the validation sets of all the ten datasets.

A.6.2 Event Graph Verbalization

To further assess the quality of the event-graph-controlled text, we conduct an in-depth human evaluation study for a manual scrutiny of error sources. Following previous works (Colombo et al., 2021), human raters are presented with the source graph, the predicted text, and the ground-truth. They are asked to judge the prediction along six quality criteria with binary rating.

- *Coverage*. Are all the information presented in the event graph included in the text?
- *Compliance*. Does the text contains only the information in the input event graph?
- *Correctness*. Are interactions modeled in the event graph correctly mentioned (correct roles and entity-linkage)?
- *Factuality*. Does the text contains only factual information?
- *Text Structure*. Is the text well-structured, grammatically correct and written in acceptable English?
- *Fluency*. Does the text progress naturally? Is it easy to understand? Is it a coherent whole?

Since the number of events is not balanced with

respect the biomedicine subarea (see Table 1), we randomly sample eight graph-text pairs for each dataset composing the BIOE2T test set (80 in total). The evaluation is performed for T5_[BIOE2T] and BART_[BIOE2T]. For each prediction, we collect scores from 3 expert evaluators and average them.

The average Kendall’s coefficient (Sen, 1968) among all evaluators’ inter-rater agreement is 0.86. Kendall’s coefficient ranges from -1 to 1, indicating low to high association. Considering the subjectivity of the rating task, this number indicated high human agreement for the EGV task.

Table 11 summarizes the results. We first note a similar trend as in EE, with T5 outperforming BART on most quality axes. We observe that the generators mostly suffer from low compliance issues due to the verbalization of additional information not originally modeled in the input graph. We investigated the reason for this error, finding three main causes: (i) event mentions shared by multiple events, (ii) a low number of nodes, and (iii) superficial and often incomplete dataset annotations—especially on GE08 and ID11. Hence, the error is not attributable to a scarce expressive power of events as semantic representations. Notably, despite the frequent verbalization of further relationships, models generally do not produce fabricated facts, and the output quality is high. Compared to T5, BART is more a victim of hallucinations and tends to paraphrase the text more, mixing patterns seen during training.

A.7 Parsing and Generation Examples

Some input-output examples for the EE and EGV tasks are shown in Table 12 and Table 13, respectively. We emphasize that current end-to-end neural conversation models inherently lack the flexibility to impose semantic control in the response generation process (Wu et al., 2021), justifying the importance of EGV. This control is essential to ensure that users’ semantic intents are satisfied and

Perspectives	Models	
	T5 _[BIOE2T]	BART _[BIOE2T]
Coverage	0.97	0.94
Compliance	0.18	0.22
Correctness	0.96	0.91
Factuality	0.99	0.90
Text Structure	0.97	0.81
Fluency	0.99	0.83

Table 11: Human evaluation scores of the verbalized event graphs on a random sample of 80 instances from the test set. The highest are bolded.

to establish a degree of specificity on generated outputs. Following this line, modifiers offer the concrete opportunity of asking a model not only to verbalize an event but also to do it with a particular writing style. Multiple modifiers can be set at the same time (e.g., “H2A may not be methylated”), allowing great flexibility. EGV is also useful to collect rationales from language models more effectively, revealing what knowledge is stored in their parameters. The qualitative results obtained indicate that the event graphs can indeed steer the language model towards informative content following provided confidence measures or other lexical clue types.

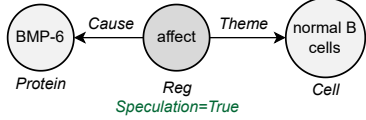
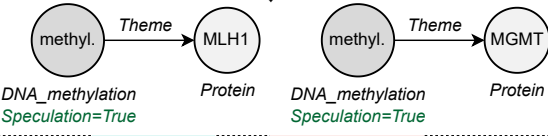

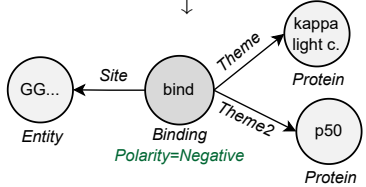
Text	Extracted Event
<p>We wanted to establish whether BMP-6 also could affect the viability of normal B cells.</p>	<p style="text-align: center;"><i>Ground_truth</i></p> <p>[affect Regulation Speculation = True] [BMP-6 Protein Cause = affect] [normal B cells Cell Theme = affect]</p>  <hr/> <p><i>T5</i>_[BioE2T] ✓ <i>BART</i>_[BioE2T] ✓</p>
<p>We analyzed the methylation status of hMLH1 and MGMT using methylation-specific polymerase chain reaction and DNA sequencing analysis.</p>	<p style="text-align: center;"><i>Ground_truth</i></p> <p>[methylation DNA_methylation Speculation = True] [MLH1 Protein Theme = methylation] [methylation DNA_methylation Speculation = True] [MGMT Protein Theme = methylation]</p>  <hr/> <p><i>T5</i>_[BioE2T] ✓ <i>BART</i>_[BioE2T] ✓</p>
<p>We found a lack of PKD1 expression.</p>	<p style="text-align: center;"><i>Ground_truth</i></p> <p>[lack Negative_regulation] [[expression Gene_expression Theme = lack] [PKD1 Protein Theme = expression]]</p>  <hr/> <p><i>T5</i>_[BioE2T] ✓ <i>BART</i>_[BioE2T] ✓</p> <p>[lacking Negative_regulation] [[expression Gene_expression Theme = lack] [PKD1 Protein Theme = expression]]</p>
<p>Conversely, the murine kappa light chain enhancer motif (GGGGACTTTCCG) does not efficiently bind the nuclear p50p50 from tolerant murine P388 macrophages.</p>	<p style="text-align: center;"><i>Ground_truth</i></p> <p>[bind Binding Polarity = Negative] [kappa light chain Protein Theme = bind] [p50 Protein Theme2 = bind] [GGGGACTTTCCG Entity Site = bind]</p>  <hr/> <p><i>T5</i>_[BioE2T]</p> <p>[bind Binding KT = Observation Polarity = Negative Source = Current] [murine kappa light chain enhancer motif DNA_domain_or_region Theme = bind] [p50p50 Protein_complex Theme = bind] [not Negative-Cue Cue = bind]</p> <hr/> <p><i>BART</i>_[BioE2T]</p> <p>[bind Binding Polarity = Negative] [kappa light chain Protein Theme = bind] [p50 Protein_molecule Theme2 = bind]</p>

Table 12: Input-output event extraction examples. Modifiers' influence is highlighted in blue, while deviations (positive or negative) from the ground truth are in red. The check-mark symbol denotes a full-correspondence between the predicted text and the ground truth.

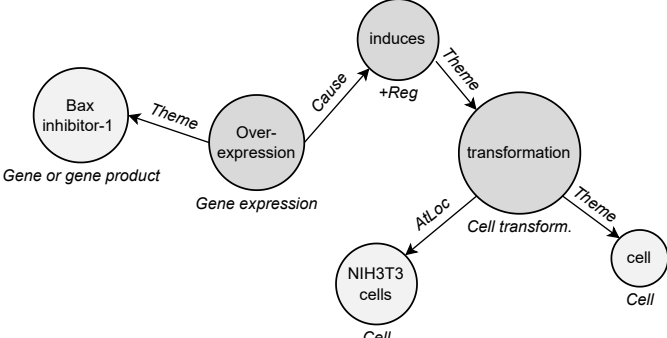
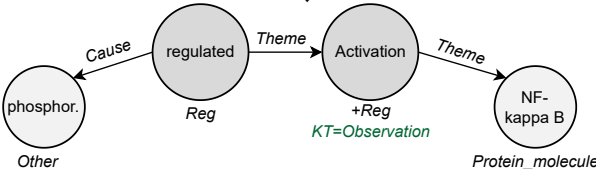
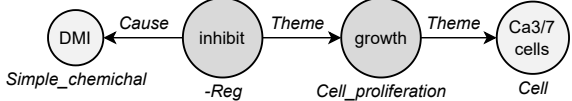
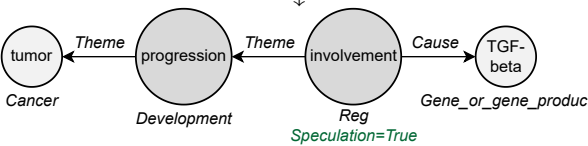
Event	Text
<p>[Overexpression Gene_expression] [Bax inhibitor-1 Gene_or_gene_product Theme = Overexpression] [[induces Positive_regulation Cause = Overexpression Theme = transformation] [transformation Cell_transformation] [NIH3T3 cells Cell AtLoc = transformation] [cell Cell Theme = transformation]]</p> 	<p><i>Ground_truth</i> Overexpression of Bax inhibitor-1 (BI-1) induces cell transformation in NIH3T3 cells.</p> <p><i>T5</i>_[B10E2T] ✓</p> <p><i>BART</i>_[B10E2T] ✓</p>
<p>[regulated Regulation] [[Activation Positive_regulation KT = Observation Theme = regulated] [NF-kappa B Protein_molecule Theme = Activation]] [phosphorylations Other Cause = regulated]</p> 	<p><i>Ground_truth</i> Activation of NF-kappa B in vivo is regulated by multiple phosphorylations.</p> <p><i>T5</i>_[B10E2T] ✓</p> <p><i>BART</i>_[B10E2T] Activation of NF-kappa B is regulated by phosphorylations and rapid degradation of its inhibitor I kappa B alpha.</p>
<p>[inhibit Negative_regulation] [DMI Simple_chemical Cause = inhibit] [[growth Cell_proliferation Theme = inhibit] [Ca3/7 cells Cell Theme = growth]]</p> 	<p><i>Ground_truth</i> DMI is suggested to inhibit the growth of Ca3/7 cells.</p> <p><i>T5</i>_[B10E2T] Therefore, we examined the effect of DMI on Ca3/7 cells, resistant to growth inhibition by the synthetic glucocorticoid flucocinolone acetonide (FA).</p> <p><i>BART</i>_[B10E2T] The growth inhibition of Ca3/7 cells by DMI is neither a consequence nor the sole cause of growth arrest.</p>
<p>[involvement Regulation Speculation = True] [[progression Development Theme = involvement] [tumor Cancer Theme = progression]] [TGF-beta Gene_or_gene_product Cause = involvement]</p> 	<p><i>Ground_truth</i> TGF-beta effects on angiogenesis, stroma formation, and immune function suggest its possible involvement in tumor progression.</p> <p><i>T5</i>_[B10E2T] In the present study, we investigated the possible involvement of transforming growth factor beta (TGF-beta) in tumor progression.</p> <p><i>BART</i>_[B10E2T] An understanding of the molecular basis of TGF-beta-mediated inhibition of angiogenesis and tumor progression will aid in the development of novel therapeutics for the treatment of cancer.</p>

Table 13: Input-output event graph verbalization examples. Modifiers' influence is highlighted in blue. The check-mark symbol denotes a full-correspondence between the predicted text and the ground truth.