Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 6922–6931
Marseille, 20-25 June 2022

SuMe: A Dataset Towards Summarizing Biomedical Mechanisms

Mohaddeseh Bastan[♠], **Nishant Shankar**[♣], **Mihai Surdeanu**[♡], **Niranjan Balasubramanian**[♠]
[♠] Stony Brook University [♣]Delft University of Technology [♡] University of Arizona
{mbastan, niranjan}@cs.stonybrook.edu
n.shankar@tudelft.nl msurdeanu@email.arizona.edu

Abstract

Can language models read biomedical texts and explain the biomedical mechanisms discussed? In this work we introduce a biomedical mechanism summarization task. Biomedical studies often investigate the mechanisms behind how one entity (e.g., a protein or a chemical) affects another in a biological context. The abstracts of these publications often include a focused set of sentences that present relevant supporting statements regarding such relationships, associated experimental evidence, and a concluding sentence that summarizes the mechanism underlying the relationship. We leverage this structure and create a summarization task, where the input is a collection of sentences and the main entities in an abstract, and the output includes the relationship and a sentence that summarizes the mechanism. Using a small amount of manually labeled mechanism sentences, we train a mechanism sentence classifier to filter a large biomedical abstract collection and create a summarization dataset with 22k instances. We also introduce conclusion sentence generation as a pretraining task with 611k instances. We benchmark the performance of large bio-domain language models. We find that while the pretraining task help improves performance, the best model produces acceptable mechanism outputs in only 32% of the instances, which shows the task presents significant challenges in biomedical language understanding and summarization. [1]

Keywords: Explanation Generation, Text Generation, Summarization, Biomedical NLP, Relation Extraction

1. Introduction

Understanding biochemical mechanisms such as protein signaling pathways is one of the central pursuits of biomedical research (Strötgen and Gertz, 2012) (Arighi et al., 2011; Krallinger et al., 2017; Demner-Fushman et al., 2020). Biomedical research has advanced tremendously in the past few decades, to the point where we now suffer from "an embarrassment of riches". Publications are generated at such a rapid pace (PubMed [2] has indexed more than 1 million publications per year in the past 8 years!) that we need information access applications which can help extract and organize biomedical relations and summarize the biomedical mechanisms underlying them. Developing models that can read biomedical texts and reason about these mechanisms is an important step towards this.

In this paper, we introduce a mechanism summarization task which couples text that discusses elements of biomedical mechanisms with their summaries. The task requires models to read text that presents information about the connection between two target entities and generate a summary sentence that explains the underlying mechanism and the relation between the entities. We see this task from two perspectives. First, summarizing biomedical mechanisms can be seen as part of the broader efforts in extracting (Czarnecki et al., 2012), organizing (Kemper et al., 2010; Kemper et al., 2010; Miwa et al., 2013; Subramani et al., 2015; Poon et al., 2014), and summarizing (Azadani et al., 2018) biomedical literature that are aimed at providing information access tools for domain experts. Second, from an NLP perspective this task can be seen as an explainable relation extraction in a biomedical context, where the explanation is the mechanism that provides information about why the relation holds or how it comes about.

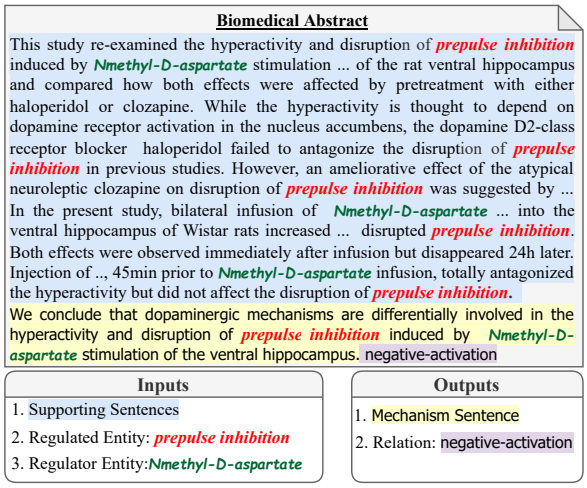

Figure 1: Biomedical Mechanism Summarization Task: Example of an entry from the SuMe dataset. Some supporting text was removed to save space. The input is the supporting sentences with the main two entities. The output is the relation type and a sentence concluding the mechanism underlying the relationship.

A key challenge in addressing such a task lies in creating a large scale dataset necessary for training large neural models. However, building such a dataset manually is a

[1] Code and data is available at SuMe webpage
[2] https://pubmed.ncbi.nlm.nih.gov

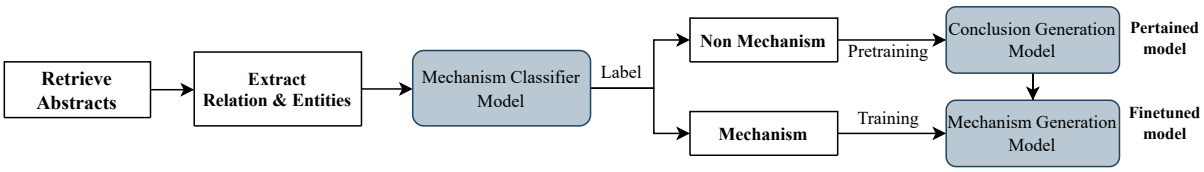

Figure 2: Overview of the semi-automatic bootstrapping process for SuMe creation. We use a mechanism classifier trained with small amount of labeled data to produce weakly-labeled training data for mechanism summarization.

laborious process and requires deep biomedical expertise. To address this, we turn to the structure that exists in biomedical abstracts, make use of related datasets, and devise a semi-automatic bootstrapping process that builds on a relatively small amount of labeling effort from domain experts.

We introduce SuMe, a large scale dataset that we construct from abstracts of papers that report on biomedical mechanisms. For a given abstract, we create a task instance that consists of a pair of biochemical entities (regulated and regulator), the relationship between them (positive/negative activation), and supporting sentences that provide information about this relationship, and a sentence that summarizes the mechanism underlying the relation (see Figure 1). Creating such an instance would require a domain expert to read through an abstract and assess if it contains a biomedical mechanism and locate it if so. This process is difficult to scale.

To address this issue, we introduce a semi-automated annotation process to create a large-scale set for development and automatic evaluation purposes and a clean small-scale manually curated subset of instances for manual evaluation. In particular, the necessary entities and relations are extracted using an existing biomedical information extraction system (Valenzuela-Escárcega et al., 2018). To extract mechanism summaries we first collected a small set of mechanism sentences with the help of domain experts. We use this to bootstrap a larger sample by training a mechanism sentence classifier with a biomedical language model (LM) (Kanakarajan et al., 2021) and apply it to a large collection of about 611K abstracts that contained a conclusion sentence about the relationship between a pair of entities. The subset that the classifier identifies as containing mechanism sentences is used to create 22K mechanism summarization instances. Five domain experts manually analyzed a dataset sample of 125 instances to construct a clean partition for manual evaluation purposes. The experts also concluded that the generated dataset has reasonable quality, i.e., 84%. Note that it is common to tolerate some level of noise in the training partitions of automatically constructed NLP datasets. As an example among many, the popular relation extraction dataset by Yao et al. (2010) contains over 20% noise. The overall pipeline is demonstrated in Figure 2. In summary, the contributions of this paper are the following:

- We introduce the SuMe dataset, the first dataset towards summarizing biomedical mechanisms and the underlying relations between entities. The dataset contains 22K mechanism summarization instances collected semi-automatically, an evaluation partition of 125 instances that were corrected by domain experts. We also create a conclusion generation task from the larger set of 611K abstracts which we use as a pretraining task for mechanism generation models.

- We benchmark several state-of-the-art language models for the task of generating the underlying biochemical relations and the corresponding mechanism sentences. We train general domain LMs (GPT2 (Radford et al., 2019), T5 (Raffel et al., 2020a), BART (Lewis et al., 2019)), as well as science domain adapted versions(scientific GPT2 (Papanikolaou and Pierleoni, 2020), and SciFive (Phan et al., 2021)) and benchmark their performance through both automatic evaluation and manual evaluation on curated evaluation samples.

- The evaluation by domain experts suggests that this is a high quality dataset coupled with a challenging task, which deserves further investigation.

- To encourage reproducibility and further research, we release the dataset and the code used during its creation. Both are available at SuMe webpage.

2. Related Work

Deep learning models have been widely used in different NLP applications (Gaonkar et al., 2020; Bastan et al., 2020; Keymanesh et al., 2021; Heidari et al., 2021). Amongst these applications, biomedical NLP is using these models that looks at extracting (Alam et al., 2018; Mulyar et al., 2021; Giorgi and Bader, 2020), organizing (Yuan et al., 2020; Zhao et al., 2020; Lauriola et al., 2021), and summarizing information (Cohan et al., 2018) from scientific literature.

Within this broad context, the mechanism summarization task we introduce broadly relates to previous work in reading and generating information from scientific texts. Most work in this area focus on generating summaries using scientific publication and some times in combination with external information (Yasunaga et al., 2019; DeYoung et al., 2020; Collins et al., 2017)

Some works even seek to generate part of the scientific papers. For example, TLDR (Cachola et al., 2020) introduces a task and a dataset to generate TLDRs (Too

Long; Didn't Read) for papers. They exploit titles and an auxiliary training signal in their model. ScisummNet (Yasunaga et al., 2019) introduces a large manually annotated dataset for generating paper summaries by utilizing their abstracts and citations. TalkSumm (Lev et al., 2019) generates summaries for scientific papers by utilizing videos of talks at scientific conferences. PaperRobot (Wang et al., 2019) generates a paper's abstract, title, and conclusion using a knowledge graph. FacetSum (Cohan et al., 2018) used Emerald journal articles to generate 4 different abstractive summaries, each targeted at specific sections of scientific documents.

In addition to the specifics of the output that we target, our work is different from all these other works because our proposed summarization task is grounded with the underlying biomedical event discussed, rather than focusing on generic summarization, which may lose the connection to the underlying biology that is the core material discussed in these papers. We address mechanism generation, which can be seen as a combination of explainable relation extraction and summarization. There is a huge body of work that addresses explainable methods (e.g., relation extraction (Shahbazi et al., 2020) or explainable QA (Thayaparan et al., 2020)). Many prior works in relation and event extraction treat explanations as the task of selecting or ranking sentences that support a relation (e.g., (Shahbazi et al., 2020; Çano and Bojar, 2020; Yasunaga et al., 2019)). Our work differs from these in that it focuses on *generating mechanisms* underlying a relation from supporting sentences, rather than identifying existing sentences.

3. Mechanism Summarization

Our goal is to develop a task and a dataset that pushes models towards distilling the mechanisms that underlie the relationships between entities from biomedical literature. From a language processing perspective, we can view mechanisms as a form of explanation that justifies the relationship or connection between entities. From a biomedical science perspective, a mechanism provides two types of explanatory information, which we use to characterize mechanism sentences:

Why is the relation true? A sentence can be a mechanism, if it explains *why* the relation exists between the two main entities. For example, one protein (say A) might be up-regulate another (say B), which in turn inhibits yet another protein (say C). This provides the causal reasoning to conclude the relation that protein A inhibits protein C.

How does the relation come about? Another kind of explanatory information is the one that describes the process or manner in which the relation exists between the pair of entities. For example, one protein (say A) may activate another protein (say B) via a specific process.

These provide a way to specify what constitutes a mechanism sentence and help us to locate mechanism sentences in the literature. In particular, we consider abstracts which discuss studies that lead to conclusions about such mechanisms. Typically, these abstracts provide a short set of sentences that describe the goals of the study, the methods used, the experimental observations, the findings, which can be used to substantiate the conclusions that establish the relation of interest, and the mechanism underlying the relation. This suggests a language processing task that tests for ability to understand biomedical mechanisms: given the preceding sentences in the abstract can a model accurately generate the underlying mechanism?

3.1. Task Definition

Given a set of sentences from a scientific abstract (referred to as *supporting sentences*) and a pair of entities (e_i, e_j) that are the focus of the abstract (referred to as *focus entities*), generate the *conclusion sentence* that explains the mechanism behind the pair entities and output a relation that connects these entities (e.g., positive_activation(e_i, e_j)). Figure 1 shows an example of such a tuple of supporting sentences, focus entities, relation, and mechanism sentence. As the example illustrates, mechanism sentences describe some pathway often involving another entity or a process (e.g., *dopaminergic mechanism*), require identifying and combining information from multiple relevant sentences, and non-trivial inferences regarding the relationship between the entities (e.g., recognizing that the different effects on *prepulse inhibition* imply differential involvement).

The task definition suggests what we need to build a dataset. Given an abstract of a scientific literature we need four pieces of information: (1) the two focus entities of the abstract; (2) the relation between entities; (3) sentences from the abstract in support of this relation; and (4) the conclusion sentence where the mechanism underlying the relation is summarized.

4. SuMe Dataset

We aim to create a large scale dataset for the mechanism summarization task defined above. However, identifying instances for this task requires domain expertise and cannot be easily done at scale. Instead, here we employ a bootstrapping process, where we first annotate a small amount of data to build a mechanism sentence classifier that can then helps us collect a large scale dataset for mechanism summarization. The key observation here is that identifying sentences that express a mechanism is a simpler task than the targeted mechanism summarization task, and, thus, should be learnable from smaller amounts of data. We outline the process we use for creating our mechanism summarization dataset, SuMe, and an expert evaluation of its quality next.

4.1. SuMe Construction Process

We construct SuMe using biomedical abstracts from the PubMed open access subset[3]. Starting from 1.1M scientific papers [4] , we use the following sequence of

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

[4] We used all papers available in NIH active directory

bootstrapping steps to construct SuMe. The following steps are also elaborated in Figure 3.

1. Finding Conclusion Sentences: First, we use simple lexical patterns to find abstracts with a clearly specified conclusion sentence. All abstracts which has any form of *conclude* word (*conclusion, concluded, concluding, concludes*, etc.) at the very end of the text are extracted here. We use this matching process to also split the abstracts into the set of supporting sentences (the ones that lead up to the conclusion) and one conclusion sentence (the one that includes the *conclude* word).

2. Extracting Main Entities & Relation Starting with the abstracts which are now in the form of (supporting sentences, conclusion sentence), we then run a biomedical relation extractor, REACH (Valenzuela-Escárcega et al., 2018), which can identify protein-protein and chemical-protein relations between entities. In this work, we focus on the relations where one entity is the controller and another entity is the controlled entity and the relation between them is either *positive/negative activation* or *positive/negative regulation*. If an abstract does not contain any such relation, we keep it for the pretraining step (as described in Section 5.3); otherwise we use it for the main task.

3. Filtering for Mechanism Sentences: We then filter out the instances to only retain those whose conclusion sentences are indeed a mechanism sentence. To this end, we devised a bootstrapping process where we first collect supervised data to train a classifier. To collect likely mechanism sentences we made use of the ChemProt (Peng et al., 2019) relation extraction dataset which contains sentences annotated with positive and negative regulation relations between entities. However, not all of these sentences necessarily explain the mechanism behind these relations. We asked 21 experts (grad students in a biomedical department) to inspect each sentence and rate whether it explains the mechanism behind the ChemProt annotated relation on a four-point Likert scale. For each sentence, an annotator can select between *Clearly a Mechanism, Plausibly a Mechanism, Clearly not a Mechanism,* and *Not Sure.* Each sentence is annotated by three experts and we find the inter-annotator agreement between users to be $\kappa = 73\%$ (Fleiss Kappa (Landis and Koch, 1977)). The final label for a sentence is selected based on the majority voting after combining *Clearly a Mechanism* and *Plausible a Mechanism* labels. Finally, each sentence is labeled as a *Mechanism*, or *Non-Mechanism*. The resulting dataset contained 439 *Mechanism* sentences (264 *Clearly*, 175 *Plausibly*) and 447 *Non-Mechanism* sentences.

Using this small scale mechanism sentence dataset, we train binary classifiers to identify mechanism sentences, where the positive label indicates that the underlying sentence is a mechanism sentence. We fine-tuned multiple transformer-based models: BioBERT (Lee et al., 2020), SciBERT (Beltagy et al., 2019), BiomedNLP (Gu et al., 2020), and BioELECTRA (Kanakarajan et al., 2021) models. Each model is fitted with a non-linear classification layer that takes the output representation for the [CLS] token. The classification layer and top three layers of the transformer are finetuned using the annotated data[5]. We used 80%-20% split for train-test. BioELECTRA performed the best with 74% macro F1 for mechanism sentence classification.

We use this trained mechanism sentence classifier to label all conclusion sentences from the previous step and instances which are predicted to be mechanism sentences are used to create the mechanism generation of the SuMe dataset.

We separate out the abstracts which are predicted to have non-mechanism sentences as additional data. We can define a broader conclusion generation task, which can be be used as a pre-training task for the generative models that eventually use for the mechanism summarization task (as we describe in Section 5.3).

The above procedure results in a dataset that allows us to define the following mechanism summarization task: Given a set of supporting sentences from an abstract and a pair of entities (e_i, e_j), generate a relation that connects these entities and a sentence that explains the mechanism that was the focus of the study. The statistics of the dataset are shown in Table 1. The dataset consists of three subsets, the training set with about 20k instances which the parameters of the model are trained with, the validation set (Dev) for tuning hyper parameters and choosing the best model, and the test set which is not used until the final evaluation. There is also a small set of 125 instances which is curated by experts and is used as another test set but is not reported in this table.

Dataset	Train	Dev	Test
Abstracts	20765	1000	1000
Avg. #words in conc.	33.7	34.9	33.5
Avg. #words in supp.	187.5	187.9	186.7
Avg. #sent. in supp.	12.15	12.44	12.33
#Unique controller	8094	759	777
#Unique controlled	6684	717	687
#Unique pair entities	19229	988	989
#Unique entities	12685	1357	1364

Table 1: Dataset Statistics: Each dataset contains a number of unique abstracts, a supporting set (supp.), a mechanism sentence (conc.) a pair of entities. The first entity is called the regulator entity (regulator) and the second one is called the regulated entity (regulated)

4.2. SuMe Quality

Our process creates a large scale, albeit, a bootstrapped dataset that can be used to train large language generation models. What is the quality of this dataset? To assess this we asked five biomedical experts to evaluate a random sample of 125 sentences from the dataset. The

[5]All models used are base versions with 768 hidden size and 12 layers. We set the learning rate to be $2e-4$ with a decay of 0.001

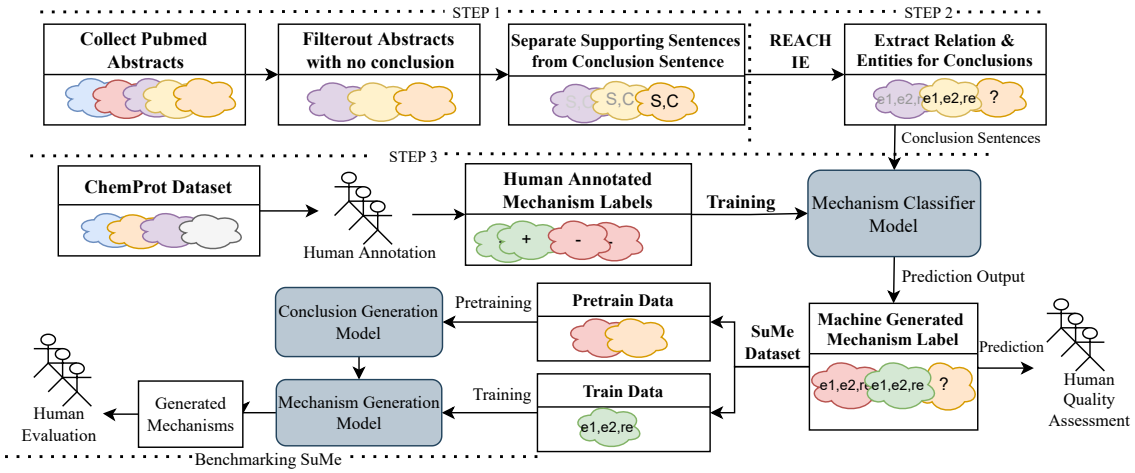

Figure 3: The bootstrapping pipeline for SuMe collection and human evaluation. The main idea behind the pipeline is to collect relatively easy to acquire judgments from domain experts to then bootstrap and generate a weakly-labeled large training corpus. We further assess the quality of the resulting dataset through another round of human evaluation, which also yields a smaller curated evaluation dataset.

experts were given the set of input supporting sentences, the potential mechanism sentence, and the relation between main entities. Our aim is two fold, first to evaluate the quality of the data collection process, second to collect a clean human evaluated dataset which can be used as an extra test set.

The experts were asked to assess errors in the relation label, mechanism, and the need for background knowledge:

1. **Relation Errors:** Is the expected output relation associated with the instance valid?
2. **Mechanism vs. Non-mechanism:** Is the output sentence expected for this sentence an actual mechanism sentence?
3. **Background Knowledge:** Can the information in the output sentence concluded from the information in the input supporting sentences?

The results of the dataset evaluation are shown in Table 2. Only 16% of the data has some error either from relation extraction (question 1) or contains a non-mechanism output sentence (question 2).This evaluation shows that the generated dataset is of reasonable quality, and can serve as a meaningful resource for training models for mechanism summarization. The clean subset that has no relation or mechanism errors is used as an extra test for evaluation. Last, the experts also rated 15% of the instances to require background knowledge (question 3)indicating the fraction of hard instances.

Quality	Correct
Entities & Relation Extraction	90%
Mechanism Sentence Classifier	85%
Instances w/o noise	84%

Table 2: Dataset Quality: We asked three main questions. This table shows what percentage of each category is acceptable. The last question shows what percentage of the sentences are approved in all questions.

5. Evaluation

Our evaluation focuses on the following questions:

1. Benchmarking: What is the performance of generic and domain-adapted large scale language generation models on SuMe?
2. Effect of pretraining: What is the impact of using the additional data via pretraining?
3. Effect of modeling supporting sentences: What is the impact of selecting a subset of supporting sentences?
4. Error analysis: What are the main failure modes of language generation models?

5.1. Experimental Setup

We use SuMe to benchmark language generation models and measure their ability to correctly identify the relation between the focus entities and to summarize the mechanism behind the relation based on the input sentences from the abstract.

Models: We compare pretrained GPT-2 (Radford et al., 2019), T5 (Raffel et al., 2020b), BART (Lewis et al., 2019) models and two domain-adapted models, GPT2-Pubmed (Papanikolaou and Pierleoni, 2020), and SciFive (Phan et al., 2021), which were trained on scientific literature.

Evaluation Metrics: We conduct both automatic and manual evaluation of the model outputs.

Relation Generation (RG): The models are supposed to first generate the relation type (positive or negative regulation) and then generate the mechanism that underlies this relation. We evaluate the model's output as we would for a corresponding classification task, i.e., the generated relation is deemed correct if it exactly matches the correct relation name. We report F1 numbers for this binary classification task.

Model	RG (F1)	BLEURT	Rouge-1	Rouge-2	Rouge-L
BART	76	42.49	46.54	25.92	35.34
GPT2	74	44.19	46.54	28.32	38.78
T5	72	44.41	48.26	27.63	38.77
GPT2-Pubmed	78	46.33	48.37	29.55	40.19
SciFive	**79**	**47.81**	**52.10**	**32.62**	**43.31**

Table 3: Benchmarking performance of strong language generation models and some domain-adapted models. We present standard automatic evaluations measures for the mechanism sentence generation task along with F1 for the generated relations. The science domain versions of both GPT2 and T5 work better than the original versions.

Mechanism Generation: We evaluate the quality of the generated explanations using two language generation metrics: the widely-used ROUGE (Lin, 2004) scores that rely on lexical overlap, and BLEURT scores (Sellam et al., 2020) which aim to capture semantic similarity between the generated and the gold reference. We use a recent version, the BLEURT-20 model that has been shown to be more effective (Pu et al., 2021) . We compare the generated text as the hypothesis against the actual text as the reference.

Fine-tuning and Training Details: All models are original base models published by HuggingFace that were fine-tuned on the training portion of SuMe for 20 epochs. For each model, we evaluate the average of BLEURT and Rouge-L score on the validation (Dev) set and the one with the highest average is chosen for prediction. The learning rate is set to 6e-5, we use AdamW (Loshchilov and Hutter, 2017) optimizer with $\epsilon = 1e-8$. The input token is limited to 512 tokens, and the generated token is maxed out at 128. We select batch size of 8 with gradient accumulation steps of two.

5.2. Automatic Evaluation Results

Table 3 compares the performance of the five language generation models on both the relation generation (RG) and mechanism generation tasks.

The domain-adapted models, GPT2-Pubmed and SciFive, fare better than fine-tuning the standard pre-trained models for both relation and mechanism generation tasks. SciFive achieves the best performance with more than a 7.5% increase in BLEURT score and more than 9.7% increase in RG F1 over the standard T5 model, highlighting the importance of domain adaptation for the SuMe tasks defined over scientific literature.

The overall numbers (coupled with the human evaluation in Section 5.5) suggest that mechanism generation is a difficult and challenging task.

The models achieve better performance on the relation generation task but there is still a substantial room for improvement here with the best model achieving an F1 of 79. If the model is unable to generate the relation correctly, then the mechanism it generates is not useful. Ideally we want models to correctly generate both the relation and the mechanism that underlies it. We also evaluated the correlation between BLEURT score and relation generation classification score. Our analysis shows that when the model generates an accurate relation, it gets higher BLEURT score while when it generates an incorrect relation, its gets a 10% lower BLEURT score (50.02 vs 45.08)

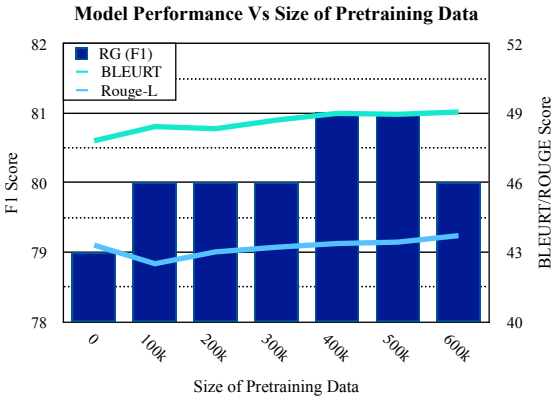

Figure 4: Comparison of relation generation (RG) F1 (left y-axis/blue bars) and the mechanism generation measures (right y-axis/teal+blue curves) against the amount of pretraining. As we increase the size of the pretraining data, the model performance improves.

5.3. Pretraining with Conclusion Generation

Next we analyze the impact of pre-training the models on the related task of generating conclusion (instead of mechanism) sentences, for which we can obtain data at scale without any manual labeling effort. We collected all abstracts from PubMed that ended with a conclusion sentence. We can create training instances on these abstracts in the same format as we did for the mechanism generation instances. The only difference here is the output sentences are conclusion sentences and not necessarily mechanisms. We call this the conclusion generation task. SuMe includes 611K instances of this kind which is an order of magnitude larger than the mechanism summarization instances and can be seen as a form of data augmentation.

We study the effect of this pretraining task by varying the amount of pretraining data. We analyze the impact in terms of the overall effectiveness and the amount of fine-tuning (number of epochs) needed to converge when finetuning.

Pretraining Data Size: We pretrain the SciFive model on the conclusion generation task with increasing amount of data (100K increments), and measure the performance of finetuning the pretrained models on the

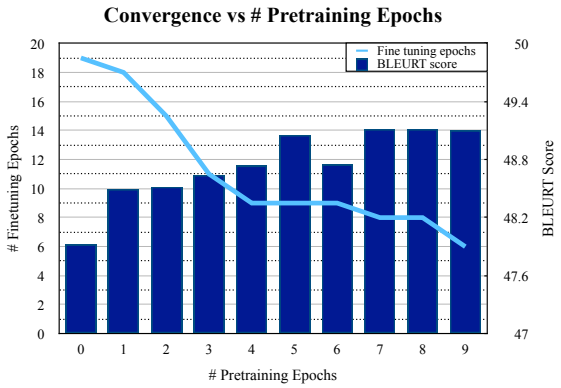

Figure 5: Number of pretraining epochs vs. fine-tuning epochs for each pretrained model until convergence.

mechanism summarization task. Figure 4 shows that performance increases with more data available for pretraining, suggesting that pretraining is beneficial for learning to generate mechanisms.

Number of Epochs: We also compare the impact of the amount of pretraining on the number of epochs needed for convergence in fine-tuning. Figure 5 compares pretrained models with different number of pretraining epochs (x-axis) in terms of their overall effectiveness (BLEURT score bars) and the number of epochs to convergence (Finetuning epochs curve). The figure shows that when we continue pretraining, not only does the resulting model performs better, but it also converges sooner taking fewer number of epochs to reach higher effectiveness. Together these results suggest potential for the auxiliary data available in the SuMe dataset.

5.4. Modeling Supporting Sentences

Will it help to model the subset of sentences within the inputs sentences that provide the best support for generating the mechanism sentence? This kind of an extractive step has been used previously in summarization tasks to reduce the amount of irrelevant information in the input (Narayan et al., 2018; Liu and Lapata, 2019). To understand the utility of this, we built a pseudo-oracle that finds the sentences that have the best overlap (measured via BLEURT score (Sellam et al., 2020)) with the output mechanism sentence. Then, we trained the SciFive model and pretrained version to only use the top few sentences according to BLEURT score such that input size is now half of the original input size. Using this subset instead of the entire subset provides BLEURT score improvements only for the basic SciFive model and the gains reduce when we use the pretrained model. Unlike standard summarization tasks there are fewer completely unrelated sentences in the abstracts and generating the mechanism sentences remains challenging even when we are able to identify the most relevant sentences within this set. This suggests that the task remains hard even when the most important sentences are somehow known to the model.

Supporting Set	BLEURT	Rouge-L
SciFive	47.81	43.31
+Oracle	49	43.07
+Pretraining	49.05	43.72
+Pretraining+Oracle	49.64	43.81

Table 4: The effect of selecting supporting sentences with highest BLEURT score.

5.5. Manual Evaluation

We also conduct a manual evaluation of the outputs from the best model — the SciFive model that was pretrained with the conclusion generation task. We asked three biomedical experts to evaluate output sentences for 100 instances and answer three questions (It took ~ 5 minutes per expert per instance):

1. Does the generated sentence contain a mechanism that explains the relation between the two main entities?
2. Is the information in the generated mechanism supported by the information contained in the supporting set of sentences?
3. Is the generated mechanism factually correct?

The results show that the model learns to generate some mechanism sentence that connects the main entities for most of the instances (79%). It also produces output for which there is support in the input sentences in a majority of the cases (53%). Last, experts found that the output statements to be scientifically correct in many cases (58%). In summary, however, only 32% of the outputs were acceptable for all questions, and were deemed to be good mechanism sentences. This again highlights the significant challenge posed by this task.

5.6. Error Analysis

To understand the frequent failure modes of the model, we manually categorized the errors in 100 outputs that had the **worst BLEURT** scores with the reference mechanism sentences.

1. **Missing Entities (35%)** – The most prevalent issue is the absence of one of the main entities in the generated sentence. Despite this being a necessary feature in all of the mechanism sentences in the training data, the prevalence of this error shows that models find it difficult to track the main entities during generation.
2. **Incorrect Mechanism (24%)** – The model is unable to generate the correct mechanism even though it is able to identify the correct relation and fills in some information that is either unrelated to or unsupported by the input sentences.
3. **Flipped Relation (19%)** – The model predicts the incorrect relation and generates a mechanism that is faithful to this incorrect relation. Improving relation generation is thus an important step for improving mechanism generation.
4. **Non Mechanisms (11%)** – While the model learns to generate mechanism like sentences for the most

Gold Mechanism	Generated Mechanism
On the basis of these observations, we conclude that *IL-18* induces *MCP-1* production through the PI3K/Akt and MEK/ERK1/2 pathways in macrophages.	In conclusion, *IL-18* induces *MCP-1* production in macrophages through the PI3K/Akt and MEK/ERK1/2 pathways.
Taken together, we conclude that *DeltaNp73* negatively regulates *NGF-mediated neuronal differentiation* by transrepressing TrkA.	In conclusion, our results indicate that *DeltaNp73* negatively regulates *NGF-mediated neuronal differentiation* by transcriptionally repressing the expression of TrkA.
In conclusion, the increase in *SNGFR* caused by *ANF* is associated with an increase in glomerular capillary hydraulic pressure and with a blunted maximal tubuloglomerular feedback response.	We conclude that *ANF* inhibits maximal tubuloglomerular feedback by increasing glomerular capillary hydraulic pressure and by decreasing *SNGFR*.
We conclude that, without modulatory factors which play a role in vivo, *NGF* can enhance the synthesis of *tyrosine* hydroxylase n sympathetic ganglia in vitro, provided organ culture conditions which permit optimal survival of adrenergic neurons are selected.	This conclusion was further supported by pulse-labeling of *tyrosine* hydroxylase with [3H]leucine, which showed that *NGF* increased synthesis of *tyrosine* in sympathetic ganglia by approximately 50%.
We conclude that 20 mm alcohol/submaximal CCK as well supramaximal CCK stimulation can trigger pathologic basolateral exocytosis in pancreatic acinar cells via PKC alpha-mediated activation of *Munc18c* , which enables *Syntaxin-4* to become receptive in forming a SNARE complex in the BPM.	We conclude that alcohol can induce a clinically relevant form of pancreatitis by blocking apical exocytosis and redirecting exocytosis to less efficient BPM, mimicking supramaximal CCK stimulation.
We conclude that in the presence of high doses of insulin, *FSH* decreases *aromatase* activity, and an uncoupling of P450 *aromatase* mRNA and *aromatase* activity occurs.	In conclusion, insulin stimulates *aromatase* activity in bovine granulosa cells at low doses but fails to stimulate activity at higher doses of insulin.

Table 5: Examples of the generated outputs by the model. The first three are good outputs where the mechanism is a simple paraphrase of the expected gold mechanism, while the next three illustrate the types of semantic errors we observe. The main entities are makred in *Italics*. The phrase explaining the mechanism in gold data is in blue, in good generation is in green, and in bad generation is in red.

part, it sometimes still fails to produce sentences that contain any mechanism at all.

5. **Multiple pieces of information (11%)** – Some complex mechanisms require combining bits of information from different input sentences. The model generates only a part of such mechanisms.

5.7. Word Analysis

We further analyzed the unigrams of the supporting sentences corresponding to the instances where the model was most confident in its generated mechanism and where it was least confident. The analysis shows that when the words *'binding', 'caused', 'demonstrated', 'dose dependent', 'investigated', 'result',* and *'performed'* are available in the supporting sentences the model can generate explanation sentences with higher quality. This shows that when the supporting sentences convey causal relation and reasoning the model is most confident about generating mechanisms.

Table 5 shows example generated mechanisms. The first three showcase good outputs whereas the next three are examples of incorrect ones. In the good ones, the first is a generated mechanism that is almost identical to the gold mechanism with only a slight syntactic change. The second is a generated mechanism which also conveys the gold mechanism accurately but with a paraphrasing that expands the technical term TRANSPRESSING. In the last three examples with incorrect information, the first shows a bad output which contains a mechanism but not of the relation connecting the main entities. The next is a case where the information is correct but it does not even mention the main entities. The last one is an example one of the entities are missing (*FSH*) and the generated text is about another relation.

6. Conclusions

We introduced SuMe, a dataset for biomedical mechanism summarization. This dataset is coupled with a challenging summarization task, which requires the generation of the relation between main entities as well as a textual summary of the mechanism which explains the reason behind the underlying relation. This dataset is collected using the sentences from actual publication abstracts. We also introduce an easier and scalable pretraining task which improves the baselines by augmenting a larger set of sentences to the main dataset. We evaluated the complexity of the task using multiple state-of-the-art transformer based models. Our evaluation suggests that the proposed task is learnable, but we are far from solving it. The expert analysis also suggests the difficulty and importance of the task.

All in all, we believe that SuMe dataset and associated task are a useful step towards building true information-access applications for the biomedical literature.

Acknowledgments

This work was supported in part by the National Science Foundation under grants IIS-1815358 and IIS-1815948.

7. Bibliographical References

Alam, F., Joty, S., and Imran, M. (2018). Domain adaptation with adversarial training and graph embeddings. *arXiv preprint arXiv:1805.05151*.

Arighi, C. N., Lu, Z., Krallinger, M., Cohen, K. B., Wilbur, W. J., Valencia, A., Hirschman, L., and Wu, C. H. (2011). Overview of the biocreative iii workshop. *BMC bioinformatics*, 12(8):1–9.

Azadani, M. N., Ghadiri, N., and Davoodijam, E. (2018). Graph-based biomedical text summarization: An itemset mining and sentence clustering approach. *Journal of biomedical informatics*, 84:42–58.

Bastan, M., Koupaee, M., Son, Y., Sicoli, R., and Balasubramanian, N. (2020). Author's sentiment prediction. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 604–615, Barcelona, Spain (Online), December. International Committee on Computational Linguistics.

Beltagy, I., Lo, K., and Cohan, A. (2019). Scibert: Pretrained language model for scientific text. In *EMNLP*.

Cachola, I., Lo, K., Cohan, A., and Weld, D. S. (2020). Tldr: Extreme summarization of scientific documents. *arXiv preprint arXiv:2004.15011*.

Çano, E. and Bojar, O. (2020). Two huge title and keyword generation corpora of research articles. *arXiv preprint arXiv:2002.04689*.

Cohan, A., Dernoncourt, F., Kim, D. S., Bui, T., Kim, S., Chang, W., and Goharian, N. (2018). A discourse-aware attention model for abstractive summarization of long documents. *CoRR*, abs/1804.05685.

Collins, E., Augenstein, I., and Riedel, S. (2017). A supervised approach to extractive summarisation of scientific papers. *arXiv preprint arXiv:1706.03946*.

Czarnecki, J., Nobeli, I., Smith, A. M., and Shepherd, A. J. (2012). A text-mining system for extracting metabolic reactions from full-text articles. *BMC bioinformatics*, 13(1):1–14.

Dina Demner-Fushman, et al., editors. (2020). *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*, Online, July. Association for Computational Linguistics.

DeYoung, J., Lehman, E., Nye, B., Marshall, I. J., and Wallace, B. C. (2020). Evidence inference 2.0: More data, better models. *arXiv preprint arXiv:2005.04177*.

Gaonkar, R., Kwon, H., Bastan, M., Balasubramanian, N., and Chambers, N. (2020). Modeling label semantics for predicting emotional reactions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4687–4692, Online, July. Association for Computational Linguistics.

Giorgi, J. M. and Bader, G. D. (2020). Towards reliable named entity recognition in the biomedical domain. *Bioinformatics*, 36(1):280–286.

Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., and Poon, H. (2020). Domain-specific language model pretraining for biomedical natural language processing.

Heidari, M., Zad, S., Hajibabaee, P., Malekzadeh, M., HekmatiAthar, S., Uzuner, O., and Jones, J. H. (2021). Bert model for fake news detection based on social bot activities in the covid-19 pandemic. In *2021 IEEE 12th Annual Ubiquitous Computing, Electronics Mobile Communication Conference (UEMCON)*, pages 0103–0109.

Kanakarajan, K. r., Kundumani, B., and Sankarasubbu, M. (2021). BioELECTRA:pretrained biomedical text encoder using discriminators. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 143–154, Online, June. Association for Computational Linguistics.

Kemper, B., Matsuzaki, T., Matsuoka, Y., Tsuruoka, Y., Kitano, H., Ananiadou, S., and Tsujii, J. (2010). Pathtext: a text mining integrator for biological pathway visualizations. *Bioinformatics*, 26(12):i374–i381.

Keymanesh, M., Elsner, M., and Parthasarathy, S. (2021). Privacy policy question answering assistant: A query-guided extractive summarization approach. *CoRR*, abs/2109.14638.

Krallinger, M., Pérez-Pérez, M., Pérez-Rodríguez, G., Blanco-Míguez, A., Fdez-Riverola, F., Capella-Gutierrez, S., Lourenço, A., and Valencia, A. (2017). The biocreative v. 5 evaluation workshop: tasks, organization, sessions and topics.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Lauriola, I., Aiolli, F., Lavelli, A., and Rinaldi, F. (2021). Learning adaptive representations for entity recognition in the biomedical domain. *Journal of biomedical semantics*, 12(1):1–13.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., and Kang, J. (2020). Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Lev, G., Shmueli-Scheuer, M., Herzig, J., Jerbi, A., and Konopnicki, D. (2019). Talksumm: A dataset and scalable annotation method for scientific paper summarization based on conference talks. *arXiv preprint arXiv:1906.01351*.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Liu, Y. and Lapata, M. (2019). Text summa-

rization with pretrained encoders. *arXiv preprint arXiv:1908.08345*.

Loshchilov, I. and Hutter, F. (2017). Fixing weight decay regularization in adam. *CoRR*, abs/1711.05101.

Miwa, M., Ohta, T., Rak, R., Rowley, A., Kell, D. B., Pyysalo, S., and Ananiadou, S. (2013). A method for integrating and ranking the evidence for biochemical pathways by mining reactions from text. *Bioinformatics*, 29(13):i44–i52.

Mulyar, A., Uzuner, O., and McInnes, B. (2021). Mt-clinical bert: scaling clinical information extraction with multitask learning. *Journal of the American Medical Informatics Association*, 28(10):2108–2115.

Narayan, S., Cohen, S. B., and Lapata, M. (2018). Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. *arXiv preprint arXiv:1808.08745*.

Papanikolaou, Y. and Pierleoni, A. (2020). Dare: Data augmented relation extraction with gpt-2. *arXiv preprint arXiv:2004.13845*.

Peng, Y., Yan, S., and Lu, Z. (2019). Transfer learning in biomedical natural language processing: an evaluation of bert and elmo on ten benchmarking datasets. *arXiv preprint arXiv:1906.05474*.

Phan, L. N., Anibal, J. T., Tran, H., Chanana, S., Bahadroglu, E., Peltekian, A., and Altan-Bonnet, G. (2021). Scifive: a text-to-text transformer model for biomedical literature. *arXiv preprint arXiv:2106.03598*.

Poon, H., Quirk, C., DeZiel, C., and Heckerman, D. (2014). Literome: Pubmed-scale genomic knowledge base in the cloud. *Bioinformatics*, 30(19):2840–2842.

Pu, A., Chung, H. W., Parikh, A. P., Gehrmann, S., and Sellam, T. (2021). Learning compact metrics for mt. In *EMNLP*.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2020a). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2020b). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67.

Sellam, T., Das, D., and Parikh, A. P. (2020). Bleurt: Learning robust metrics for text generation. *arXiv preprint arXiv:2004.04696*.

Shahbazi, H., Fern, X., Ghaeini, R., and Tadepalli, P. (2020). Relation extraction with explanation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6488–6494.

Strötgen, J. and Gertz, M. (2012). Temporal tagging on different domains: Challenges, strategies, and gold standards. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 3746–3753, Istanbul, Turkey, may. European Language Resource Association (ELRA).

Subramani, S., Kalpana, R., Monickaraj, P. M., and Natarajan, J. (2015). Hpiminer: A text mining system for building and visualizing human protein interaction networks and pathways. *Journal of Biomedical Informatics*, 54:121–131.

Thayaparan, M., Valentino, M., and Freitas, A. (2020). A survey on explainability in machine reading comprehension. *arXiv preprint arXiv:2010.00389*.

Valenzuela-Escárcega, M. A., Babur, Ö., Hahn-Powell, G., Bell, D., Hicks, T., Noriega-Atala, E., Wang, X., Surdeanu, M., Demir, E., and Morrison, C. T. (2018). Large-scale automated machine reading discovers new cancer driving mechanisms. *Database: The Journal of Biological Databases and Curation*.

Wang, Q., Huang, L., Jiang, Z., Knight, K., Ji, H., Bansal, M., and Luan, Y. (2019). Paperrobot: Incremental draft generation of scientific ideas. *arXiv preprint arXiv:1905.07870*.

Yao, L., Riedel, S., and McCallum, A. (2010). Collective cross-document relation extraction without labelled data. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1013–1023.

Yasunaga, M., Kasai, J., Zhang, R., Fabbri, A. R., Li, I., Friedman, D., and Radev, D. R. (2019). Scisummnet: A large annotated corpus and content-impact models for scientific paper summarization with citation networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7386–7393.

Yuan, J., Jin, Z., Guo, H., Jin, H., Zhang, X., Smith, T., and Luo, J. (2020). Constructing biomedical domain-specific knowledge graph with minimum supervision. *Knowledge and Information Systems*, 62(1):317–336.

Zhao, L., Wang, J., Cheng, L., and Wang, C. (2020). Ontosem: an ontology semantic representation methodology for biomedical domain. In *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 523–527. IEEE.