Towards Semantic Consistency Data Augmentation for Bio-Relation Extraction via Biomedical Notion Infusion

Anonymous ACL submission

Abstract

Biomedical Relation Extraction (Bio-RE) aims to recognize and classify the potential relations between various molecules and biomolecules. The main obstacle in Bio-RE is the scarcity of annotations especially in low-resource relation labels, thus the models cannot fully understand the connection between chemicals and diseases or drug-drug interactions. Existing works usually adopted data augmentation approaches to generate pseudo-annotated instances to alleviate the scarcity of annotations. However, the generated sentences largely ignore the semantic consistency of the biomedical domain and the logical coherence between biomolecules and diseases, causing a fatal phenomenon that the generated sentences introduce counterfactual information when learning the interactions between the drugs or diseases. To this end, this paper proposes a bio-notion-dedicated data augmentation approach that is able to measure intersections between biomedical relation notions and tokens of each instance to generate augmented data with semantic consistency. Experimental results demonstrate that our proposed method could bring 5.61% F1 improvement over SoTA baseline methods on three benchmark Bio-RE datasets in terms of BLURB.

1 Introduction

Biomedical Relation Extraction (Bio-RE) is a fundamental task in Biomedical Natural Language Processing (Bio-NLP) aiming to recognize and classify the potential relations between various molecules and biomolecules. This processing requires understanding the association between the suffering target and disease, and drug-drug interaction based on the given biomedical description to promote the downstream tasks in the biomedical field, ranging from auto-generating medical records to communication with patients (Bravo et al., 2015).

In practice, annotating biomedical data is extremely labor-intensive which requires domain knowledge from experts or doctors, and especially severely impedes the advancement of Bio-RE in low-resource relation labels. (Bravo et al., 2015; Krallinger et al., 2017) More recently, the Large Language Models (LLMs) were utilized for NLP tasks (Ding et al., 2020; Liu et al., 2021; Zhou et al., 2021), but poor performance was achieved in Bio-RE. This is because biomedical data is very sensitive, but insufficient data makes it difficult to pre-train in the biomedical domain. As a result, it faces heavy hallucination issues which is a disaster in the biomedical field. (Ghosh et al., 2023; Hu et al., 2023). In this case, how to bring more data is the key point for biomedical natural language processing tasks (Xu et al., 2022).

A general technique adopted to alleviate the data scarcity problem in the Bio-RE task is Data Augmentation (DA) (Ghosh et al., 2023; Hu et al., 2023). Existing DA methods can be divided into two categories: Amendment Rule-based methods and Generation-based methods. The former one uti-

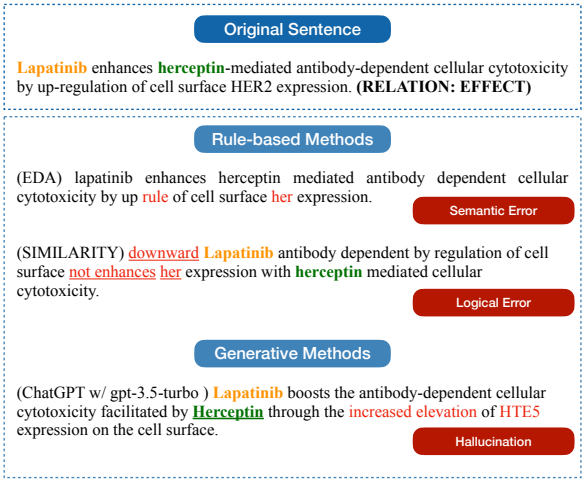

Figure 1: Examples for different data augmentation methods. Existing amend-based methods simply consider the linguistic features to substitute the token or syntax, resulting in severe semantic and logical errors, while generative models face heavy hallucinations. Both approaches introduce model counterfactual information.

lizes similarity or pre-defined rules amending the given input data to obtain similar instances (Zeng et al., 2016; Wei and Zou, 2019) while the latter one adopts generative models to generate pseudo data based on fixed prompts or instruction (Cai et al., 2020; Bayer et al., 2023). Whereas, illustrated by Figure 1, simply amending "`regulation`" to "`rule`" or paraphrasing "`...enhance...by up-regulation...`" to "`downward...not enhance...`" would cause severe semantic and logical errors which cut off the inner connection between diseases and drugs. Besides, generative methods meet heavy hallucination issues, as "`HTE5`" does not even exist. Both existing DA methods could poison Bio-RE models because counterfactual data results in the model's misunderstanding among diseases and drugs, and further damages the confidence of biomedical models.

Targeting this issue, this paper proposes a semantic consistency data augmentation approach via BIomedical NOtion infusion, named BINO. Our approach measures interactions between biomedical relation notions and tokens of each instance to help the model generate sentences with semantic consistency, instead of merely computing linguistic features of *within* or *between* instances. As illustrated in Figure 2, BINO consists of two modules, the Encoder is fine-tuned on modification of reformed input sentences. Different from other text reconstruction methods which generally adopt attention map on linguistic statistics to obtain the attribution words of the key entity, we add specific biomedical notions corresponding to every relation to the Attribution Selector to compute the mutual information between each token and specific biomedical notion to help model learn the semantic consistency among diseases and drugs. With the help of biomedical notions, the selected attributions contribute at both entity-level and target relation-level. Then, to preserve the selective attributions, we mask all other words except the selective common attributions as keywords and feed them into the model which is expected to recover the masked tokens while engraving the selective keywords. These keywords all refer to the indicator between the key entity and the target biomedical relation. As for the decoder, the generated sentences are expected to be formed as professionally as the original biomedical instances in structural and syntax. Hence, we sampled similar sentences from the training set corresponding to the target instance. Then an Extractor is adopted to extract the common exclusive structure among all original instances providing the decoder ability to maintain the logical coherence between biomolecules and diseases to hold the expertise as transcribing from experts.

The contributions of this paper are as follows:

- This paper proposes a novel data augmentation method named BINO which enhances Bio-RE model by using biomedical notion-infused selective attribution. Compared with existing methods, BINO brings semantic consistency and logical coherence to Bio-RE models by combining a biomedical notion-based encoder and a decoder with a logical structure extractor.
- This paper demonstrates the benefits of alleviating counterfactual issues of Bio-RE models. Our proposed BINO outperforms all other baseline models by 5.61% while augmenting data instances which preserve semantic consistency and maintain logical coherence.
- This paper conducts extensive experiments on three commonly used Bio-RE datasets selected by BLURB: ChemProt, DDI, and GAD, and evaluates them in low-resource settings. The experimental results validate the effectiveness of our proposed method.

2 Related Works

2.1 Biomedical Relation Extraction

Though Large Language Models (LLMs) help several Natural Language Processing (NLP) tasks attaining promising milestones, due to the scarcity of annotated data and yielding experts to involve their knowledge in data annotation is costly, the key challenge of Bio-RE task is performing better results with the limited well-annotated data (Lee et al., 2020; Tinn et al., 2023; Omiye et al., 2024). To address this problem, there are three popular techniques adapted to Bio-RE task, which are domain supervised learning (Beltagy et al., 2019; Lee et al., 2020), indirect supervised learning (Roth, 2017; Xu et al., 2022), and data augmentation (Lee et al., 2021; Hu et al., 2023; Ghosh et al., 2023). Domain supervised learning and indirect supervised learning train models on biomedical-related corpus. The former one is aiming to train on the larger data to obtain better performance while the latter is designed to convert the RE task to another formatting

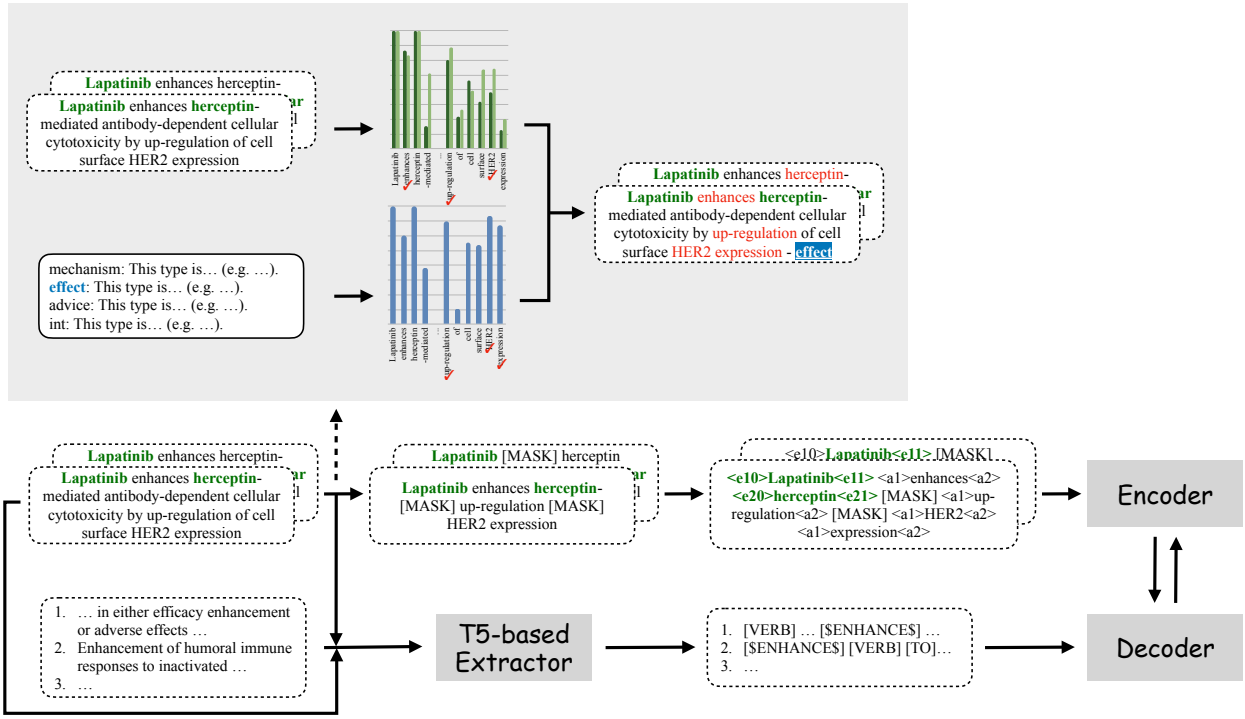

Figure 2: Overview framework of BINO. Biomedical notions and structural information are added to encoder and decoder respectively to help model understand the semantic consistency and logical coherence.

task (e.g., Machine Reading Comprehension, Question Answering, and Natural Language Inference, etc.) to use extra supervision signal promoting the performance of RE models (Xu et al., 2022). The aforementioned approaches are facing two challenges: (a) though given the larger scale corpus, the models cannot understand the interaction among entities and relations since the representation of relations is performed as only indices; (b) the issue of lacking data still obstructs the models, especially for hardly seen instances.

2.2 Data Augmentation in Bio-RE

The data augmentation aims to generate pseudo instances corresponding to the given instances in semantics but with diverse syntax. The existing methods can be divided into two categories: (a) Amendment-based methods. This kind of method tries to amend the exact token or order of input sentences according to pre-defined rules to augment the data. Wei and Zou (2019) adopts synonym replacement, random insertion, random swap, and random deletion to augment the original sentences. Beyond the token-level, Lee et al. (2021) interpolates the embeddings and labels of two or more sentences from representation-level. and (b) Generation-based methods. These years, leveraging generative large language models to generate kinds of data has become popular. Anaby-Tavor et al. (2020) and Papanikolaou and Pierleoni (2020) fine-tune multiple generative models for each relation type to generate augmentations. Bayer et al. (2023) proposes a sophisticated generation-based method that generates augmented data by incorporating new linguistic patterns. Beyond that, OpenAI (2023) introduced ChatGPT to make great progress in nearly all kinds of NLP tasks. It has proved to be effective utilizing ChatGPT as a data augmentation technique to enrich the data instances (Van Nooten and Daelemans, 2023).

However, the existing data augmentation methods generally lean to generate instances leaving faithfulness and factuality alone which poison the model in understanding the interaction among biomedical entities and relations, and further, misleading models cannot be adopted to the real scenarios.

3 Methodology

In this section, we give an introduction to our proposed data augmentation method. Figure 2 demonstrates the overview framework of our BINO data augmentation framework. §3.1 formally describe the problem. §3.2 introduce the process of data re-

construction with selected attributions with biomedical notions. §3.3 illustrate the pseudo data generation while how we leverage the biomedical logical syntax to help model augment data like transcribing from biomedical materials.

3.1 Problem Formulation

Biomedical Relation Extraction task aims to predict relation $r \in \mathbb{R}$ by being given a subjective and objective entity pair $[\text{ent}_1, \text{ent}_2]$ in the sentence $s = \{w_i : i = 0, \cdots, n\}$ where n is the length of sentence s, and w_i denotes the i-th token. The data augmentation technique targets the relation r^*, and generates pseudo sentences $s^* = \{w_i : i = 0, \cdots, m\}$ which has the same entity pair $[\text{ent}_1, \text{ent}_2]$ and the corresponding relation r^* but different syntax structure comparing to the given sentence s. Additionally, the pseudo sentences should not break the basic rules of biomedical domain to hold its availability in downstream tasks.

3.2 Data Reconstruction

To train an Encoder to fully understand the meaning of the given sentence, the model is supposed to know the attribution map of the given sentence for the two related entities. This section proposes an Attribution Selector with biomedical notion to hold the semantic consistency of biomedical domain. The Attribution Selector contains two modules, one is designed to obtain the attribution with lexicon mutual information denoted as $\text{attr}_{\text{logits}}$, while another is armed with biomedical notions corresponding to each relation aiming to obtain the semantic-level attribution for the specific biomedical relation denoted as attr_{bio}.

Lexicon-level Attribution. For each sentence s, we separate the subjective entity and objective entity corresponding to the relation r. For each target entity $e = \{w_u, \cdots, w_v\}$, we consider every token except the target entity as candidate keywords $S^* = \{s/e\}$. For every token $w_i \in S^*$ in candidate keywords, we compute the contribution from token w_i to target entity e as the attributed score $\text{attr}(e \leftarrow w_i)$ as follows:

$$\begin{aligned} \text{attr}(e \leftarrow w_i) =& \text{attr}(e) - \text{attr}(e \setminus w_i) \\ \text{attr}(e \setminus w_i) =& \text{attr}(s \setminus w_i) - \\ & \text{attr}(s \setminus \{w_u, \cdots, w_v, w_i\}) \end{aligned} \tag{1}$$

where $\text{attr}(x)$ is a function to obtain the attribution score (e.g., LOO (Lipton, 2018), LIME (Ribeiro et al., 2016), etc.). Then we go through all tokens in candidate keywords for all entities and obtain an attribution map corresponding to the target entity. After obtaining the attribution map, as the relation r^* exists between the two related entities ent_1 and ent_2, the attribution between these two entities ($\text{attr}(e_1 \leftarrow e_2)$ or $\text{attr}(e_2 \leftarrow e_1)$) should be the highest. Hence we fix the attribution map by setting these attributed scores as 1 to be the highest. Then we build an absorber bound to set any attributed score higher than the absorber bound to the highest score. Any lower will be scaled as the same proportion. Through this, we ensure the two related entities have the closest relation in the given sentence. The other tokens have their own attributed score indicating to what extent relating to the target entity. And in this way, we have the fixed attribution map with the lexicon-level information, denoted as $\text{attr}_{\text{logits}} = \{(w_i, \text{attr}(e \leftarrow w_i)) : w_i \in \{s \setminus e\}\}$.

Semantic-level Attribution. The general paradigm of generating pseudo instances is replacing single or multiple keywords regardless of semantic consistency. However, these keywords have special meaning to compose the entire sentence which cannot be replaced or need to be regarded as a whole part. For example, the keyword "*up-regulation...enhance*" means a molecule has positive effectiveness to another molecule and it cannot be placed as "*downward...not enhance*" which breaks the basic biomedical law. To overcome the semantic consistency of biomedical domain gap between the given sentence and the pseudo sentence, we incorporate biomedical notion into Attribution Selector for selecting semantic-level attribution.

Specifically, according to the three main Bio-RE datasets selected by BLURB, each relation type has a specific definition (e.g., relation "*mechanism*" in DDI dataset is defined as "*This type is used to annotate DDIs that are described by their PK mechanism*"[1]). In this way, we consider this definition as a biomedical notion of the corresponding relation in datasets, denoted as r_{bio} for relation r. We then adopt an inference model to evaluate the relativity between the keywords pair and the biomedical notion. For each target entity $e^i = \{w_u^i, \cdots, w_v^i\}$, we consider every token except the target entity

[1]DDI also give an example to clarify each relation. As for "mechanism", the given example is "Grepafloxacin may inhibit the metabolism of theobromine" which indicates the entities "Grepafloxacin" and "theobromine" have the relation of "mechanism".

as candidate keywords $S^* = \{s \setminus \{e^1, e^2\}\}$. For every token $w_i \in S^*$ in candidate keywords, we compute the contribution from (w_i, e^1, e^2) to the relativity between s and r_{bio} as the attributed score $\text{attr}(r_{\text{bio}} \leftarrow w_i)$ as follows:

$$\begin{aligned}\text{attr}(r_{\text{bio}} \leftarrow w_i) =& \text{attr}(e^1, e^2) - \text{attr}(e^1, e^2 \setminus w_i) \\ \text{attr}(e^1, e^2 \setminus w_i) =& \text{attr}(s \setminus w_i) - \\ & \text{attr}(s \setminus \{e^1, e^2, w_i\})\end{aligned} \tag{2}$$

where $\text{attr}(x)$ is the same function as Lexicon-level Attribution. In this case, we set two absorber bounds to limit the attribution score. The highest absorber bound is set to $\text{attr}(s)$ denoting that the whole sentence should be the closest to the given biomedical notion r_{bio} with the value of 1. The lowest absorber bound is set to $\text{attr}(s \setminus \{e^1, e^2\})$ with the value of 0 indicating that the sentence without the subjective and objective entity should have the lowest relativity with the biomedical notion. Then we fix all attribution scores with the same rule as the Lexicon-level Attribution. In this way, we have the fixed attribution map with biomedical notion denoted as $\text{attr}_{\text{bio}} = \{(w_i, attr(r_{\text{bio}} \leftarrow w_i)) : w_i \in \{s \setminus \{e^1, e^2\}\}\}$

Attribution Mask. To obtain the keywords that can represent the lexicon-level attribution while not losing attribution of biomedical domain, we combine the two attribution maps $\text{attr}_{\text{logits}}$ and attr_{bio} by selecting top-n common keywords with the n highest attribution score. Now we have n keywords set K corresponding to a specific sentence s with subjective and objective entities $E = \{e^1, e^2\}$ representing the relation r. Then we mask the other tokens out of the K and the entities E. To make the two entities prominent, as shown in Figure 2, following the Zhong and Chen (2020), we concatenate the two entities behind the original sentence and add label marker ahead and after the entity to explicitly indicate the entity type and give the model hint to recognize the start and the end of the entity, denotes as $e_* = \langle \text{s} : \text{ent}_{\text{type}} \rangle e \langle / \text{s} : \text{ent}_{\text{type}} \rangle$. We denote the new input sentence as $X = \{s_{\text{masked}} | e^1_* | e^2_*\}$.

3.3 Pseudo Data Generation

The proposed model BINO adopts T5 as a pseudo data generator in sequence-to-sequence paradigm. The Encoder is expected to understand the interaction between selective attributions and the subjective and objective entities while also the relation between them. The input $X = \{s_{\text{masked}} | e^1_* | e^2_*\}$ is served as template with masked tokens and the BINO learns to fill the blank to recover the original sentence. Specifically, the encoder $f_{\text{en}}(\cdot)$ takes input tokenized sentence $X = \{x_1, x_2, \ldots, x_n\}$ and obtains the contextualized token embeddings $H = f_{\text{en}}(X) = \{h_1, h_2, \ldots, h_n\}$. The recover decoder $f_{\text{de}}(\cdot)$ aims to recover the masked part of the input sentence with the training object Y:

$$Y = \{y_i : y_i = \max p(y_i | y_{<i}, H)\} \tag{3}$$

As for augmenting data with the same relation type, a natural paradigm is sampling similar sentences from training set and making the decoder part to learn the pattern from similar sentences to generate different instances with the embedding H from trained encoder. However, we argue that the model frequently imitates a similar sentence by using the common show-up words instead of finding the key structure of the similar sentences. In this way, especially in the biomedical field, the sentence would be performed like piling up plenty of terminology but constructed without biomedical logic. In this paper, we propose to leverage biomedical logical syntax with a logical extractor $f_{\text{ex}}(\cdot)$ to help model augment data like transcribing from biomedical materials.

Firstly, we sample another k sentence $\tilde{s}$ with the same subjective and objective entities e^1 and e^2 with the same relation type r. Then we concatenate the above sentence with the target sentence s as the first one, feeding into the T5-based extractor to extract the key structure $\bar{s}$ of the ten sentence. Then we compute the similarity between the $\bar{s}$ and the given sentences concatenated with the biomedical relation notion denoted as $s|r_{\text{bio}}$ and $\tilde{s}|r_{\text{bio}}$, respectively, till the similarity score is all higher than 80%. We consider the obtained $\bar{s}$ contains the key structure.

Sequentially, we add biomedical notion not only in attribution selector, but also incorporate it into the decoder, to enhance the decoder the ability to generate more biomedically logical pseudo instances, we feed encoder's output H, and target biomedical relation notion r_{bio} and the biomedical key structure $\bar{s}$ to the decoder to generate pseudo instances. In this way, the decoder can used to augment the sentence by maximizing $p(y_i | y_{<i}, H, r_{\text{bio}}, \bar{s})$ to obtain the pseudo sentences s_{aug}:

$$Y = \{y_i : y_i = \max p(y_i | y_{<i}, H, r_{\text{bio}}, \bar{s})\} \tag{4}$$

To increase the diversity and the randomicity of the augmented instances, we shuffle similar sentences every time before feeding into the decoder with the target sentence but still hold the target one as the first. Though similar sentences have the same entities and relations, the meaning and the main idea of the sentence may have discrepancies. To refrain the model from learning noise by adding different main idea sentences in the same batch, we adopt ChatGPT as the judge (Zheng et al., 2023) to evaluate each similar sentence before training the decoder. Only the similarity between target sentence and the similar sentence less than the threshold ω[2] can be fed into the decoder.

4 Experiments

We conduct extensive experiments on three commonly used biomedical relation extraction datasets selected from BLURB: GAD, DDI, and ChemProt and low-resource Bio-RE setting to show the effectiveness of our proposed BINO. Beyond that, we present the analysis of how BINO enhances model to understand the relationship between diseases and drugs and drug-drug interaction.

4.1 Experimental Settings

Datasets. BLURB benchmark contains three sentence-level biomedical relation extraction datasets to evaluate the model's ability to understand the biomedical semantics. **GAD** is a semi-labeled dataset created using Genetic Association Archive and consists of gene-disease associations. The purpose of the dataset is to determine whether there was a gene-disease relationship between the given subjective and objective entities which is a bi-label prediction (Becker et al., 2004). **DDI** is also named as Drug-Drug Interaction indicating the five pre-defined interaction types between different drugs which was specialized in pharmacovigilance built from PubMed abstracts (Herrero-Zazo et al., 2013). **ChemProt** is a disease chemical biology database, which is based on a compilation of multiple chemical–protein annotation resources, as well as disease-associated protein–protein interactions (PPIs) (Taboureau et al., 2010).

Baselines. We make the comparison with the following baseline models categorized by the pre-train corpus as the same as Xu et al. (2022).

- Semantic Scholar: **BioRE-Prompt** (Yeh et al., 2022) and **SciBERT** (Beltagy et al., 2019) are both pre-trained on BERT-based model by providing specific prompt for each relation. The training corpus consists of academic papers selected in Semantic Scholar.

- PubMed articles: **BioBERT** (Lee et al., 2020) was pre-trained on a commercial-collection subset of PMC while **BioLinkBERT** (Yasunaga et al., 2022) adopt link prediction task in pre-training processing which helped model to learn the multi-hop inference.

Beyond, we also compare our BINO with an indirect supervision method **NBR-NLI** (Xu et al., 2022) which converts the relation extraction task to NLI formulation trained on the BioLinkBERT-large backbone. We adopt the version of $\text{NBR}_{\text{NLI+FT}}$ which was fine-tuned on two general domain NLI datasets retaining biomedical domain knowledge and learning relevant NLI knowledge.

4.2 Experimental Results

Comparison with Baseline Models. Table 1 shows the comparison of experimental results over the three Bio-RE datasets. All baseline models achieve performance improvements with the help of data augmentation methods compared with the models only utilize the original training data. Over the other data augmentation methods, our BINO surpassed supervised baseline models 5.61% F1 score performance improvement on average. As for comparison with indirectly supervised models, BINO also enhances the baseline model $\text{NBR}_{\text{NLI+FT}}$ 3.81% F1 score performance improvement.

Considering the ablation study experiments, we evaluate the model by dropping the Lexicon-level Attribution, Semantic-level Attribution, and Structural information Extractor, respectively. Experimental results show that lexicon-level and semantic-level attribution contribute as same importance to perform the augmented data. Without any attribution method to calculate the tokens' contribution would result in around 4% performance dropping. Beyond that, with the help of proposed semantic-level attribution by incorporating biomedical notions, the model could generate pseudo-instances closer to the real data distribution. Besides, the structural information Extractor can also bring over 1.5% performance improvement by providing the

[2]In our experiments, we set ω to 0.8.

Model	Backbone	Para.	GAD	DDI	ChemProt	Average	Δ
BioREPrompt	-	125M	-	-	67.46	67.46	-
SciBERT-base	-	110M	-	81.32	74.93	78.13	-
BioBERT-base	-	110M	79.83	80.33	76.46	78.87	-
BioLinkBERT-base	-	110M	84.39	82.72	77.57	81.56	-
BioLinkBERT-large	-	330M	84.90	83.35	79.98	82.74	-
w/ EDA (2019)	-	-	84.97	84.03	79.41	82.80	0.06
w/ ParaGraph (2020)	-	-	84.19	83.52	78.99	82.23	-0.51
w/ AdMix (2022)	mBART-base	110M	85.10	84.79	80.26	83.38	0.64
w/ LAMBADA (2020)	GPT-2	117M	85.47	85.19	81.18	83.95	1.21
w/ ChatGPT (gpt-3.5-turbo)	GPT-3	20B	88.56	87.02	83.13	86.27	3.53
w/ BINO	T5-base	220M	**90.21**	**89.54**	**85.29**	**88.35**	**5.61**
w/o $\text{attr}_{\text{logits}}$	T5-base	220M	85.29	85.27	82.03	84.20	1.46
w/o attr_{bio}	T5-base	220M	85.67	86.58	82.54	84.39	1.65
w/o f_{ex}	T5-base	220M	88.12	88.82	83.43	86.79	4.05
NBR-NLI	-	340M	85.86	84.66	80.54	83.69	-
w/ EDA (2019)	-	-	85.99	85.18	81.51	84.23	0.54
w/ ParaGraph (2020)	-	-	86.48	85.39	80.96	84.28	0.59
w/ AdMix (2022)	mBART-base	110 M	87.12	86.97	81.41	85.17	1.48
w/ LAMBADA (2020)	GPT-2	117M	87.32	86.76	81.59	85.22	1.53
w/ ChatGPT (gpt-3.5-turbo)	GPT-3	20B	87.21	87.88	83.57	86.22	2.53
w/ BINO	T5-base	220M	**89.17**	**88.22**	**85.12**	**87.50**	**3.81**
w/o $\text{attr}_{\text{logits}}$	T5-base	220M	87.13	86.45	82.30	85.29	1.60
w/o attr_{bio}	T5-base	220M	87.23	86.71	82.28	85.41	1.72
w/o f_{ex}	T5-base	220M	88.44	88.05	84.31	86.93	3.24

Table 1: Average F1 score over three Bio-RE datasets. The performance of baseline models is obtained from their papers. Backbones and Parameters indicate what the model pre-trained on and how many parameters are involved. The **bold** results indicate the best performance over the baselines.

Model	1%	5%	25%	50%	75%	100%
SUPERVISED MODELS						
BioLinkBERT-large	-	-	-	-	-	83.35
w/ EDA	**83.34**	81.55	76.24	73.83	72.05	71.97
w/ ParaGraph	83.29	82.16	80.44	80.02	79.56	79.38
w/ AdMix	83.21	83.19	81.75	81.58	80.90	80.81
w/ LAMBADA	83.29	82.50	81.43	80.42	81.28	81.25
w/ ChatGPT (gpt-3.5-turbo)	83.09	**83.47**	81.53	80.04	80.25	80.41
w/ BINO	**83.34**	83.25	**83.28**	**82.91**	**82.26**	**82.38**

Table 2: Experimental results on different proportions of augmented data substitution. 1% to 100% indicates substituting specific proportion of original training data with generated instances. The **lower** gap between training on 100% original data and replacement occasion is better which demonstrates the better data quality.

pseudo-instances in the same written style as the original instances. By comparing with advanced generated model ChatGPT with nearly 20B parameters, our proposed BINO surpasses it with only 1% parameter size of gpt-3.5-turbo.

Augmented Data Replacement. Table 2 shows that comparison of experimental results of the BioLinkBERT-large model with different proportions of augmented data replacement. In this experiment, we replace the 1%, 5%, 10%, 25%, 50%, 75%, and even all of original training data (100%) to the augmented data with different DA methods to find out whether the model could truly benefit from the augmented data instead of being affected by original data.

The experimental results demonstrate that our BINO could preserve the original meaning best by comparing with the other popular data augmentation methods. Even after all replacement by

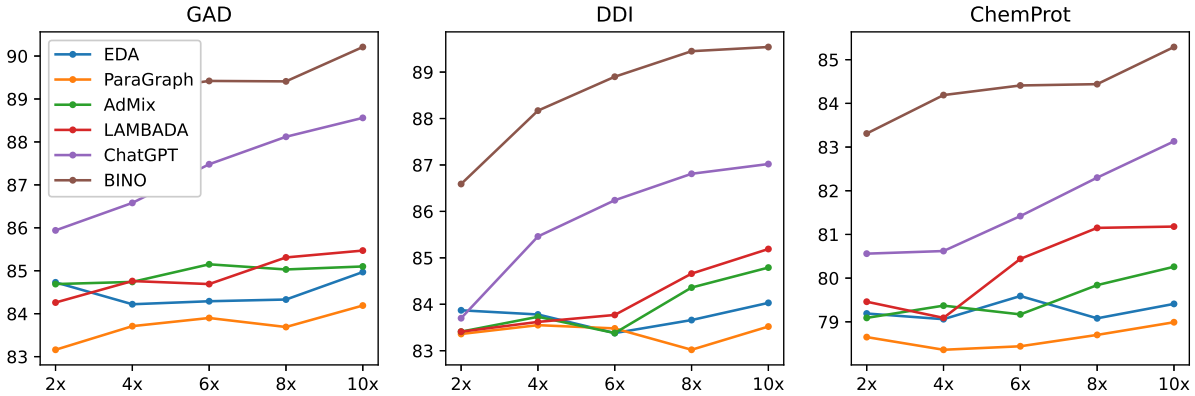

Figure 3: Different augmented proportion on BioLinkBERT-large model. Different proportion indicates the number of generated instances for one original sentence.

Model	DDI		ChemProt	
	TTR	PPL	TTR	PPL
w/ EDA	83.36	96.35	83.45	92.78
w/ ParaGraph	83.92	90.43	85.41	88.75
w/ AdMix	75.23	79.92	78.34	71.93
w/ LAMBADA	73.08	79.86	73.64	69.28
w/ ChatGPT	84.73	76.28	84.62	64.45
w/ BINO	**85.34**	**71.42**	**85.76**	**58.40**
w/o $\text{attr}_{\text{logits}}$	78.41	73.43	80.48	63.96
w/o attr_{bio}	78.96	73.08	80.71	63.50
w/o f_{ex}	83.27	72.71	84.26	61.29

Table 3: Average scores for Type-Token Ratio (TTR), and Perplexity (PPL) over three datasets.

pseudo-instances generated by BINO, the model could achieve 82.38% F1 performance with just 0.97% drop while the EDA method cause nearly 12% performance drop. With the help of incorporating biomedical notions by adding semantic-level attribution, the generated data approach the original data distribution better than only adopting lexicon-level attribution.

Augmented Instances Proportion. To explore the effectiveness of proportion of augmented instances to the model, we conduct experiments on BioLinkBERT-large with different augmented proportions. We vary the multiple of augmented data from double to 10 times the original training data. Figure 3 demonstrates the experimental results. We observe that model would have more performance improvement by increasing the multiplier of augmented data. However, with the continuous increase in the amount of augmented data, the margin decreases sharply and holds when multiplier changes from 8x to 10x.

4.3 Diversity Analysis

We measure the diversity of augmented instances through generally used metrics in Type-Token Ratio (TTR) (Tweedie and Baayen, 1998) and Perplexity (PPL) (Jelinek et al., 1977). The former is used to measure the ratio of the number of different words to the total number of words in the dependency path between two entities for each relation type. Higher TTR generally indicates that the pseudo-instances have a higher diversity at lexicon-level. While the latter PPL is defined as the exponentiated average negative log-likelihood of a sequence. Intuitively, it can be thought of as an evaluation of the model's ability to predict uniformly among the set of specified tokens in a corpus. Lower PPL indicates that the models have a better ability to select different tokens from the corpus. Table 3 shows the average scores for all metrics over three datasets. We observe that our proposed BINO achieves the best performance with 62.53% in PPL and 87.85% in TTR.

5 Conclusions

This paper introduces a bio-notion-dedicated data augmentation method for Bio-RE task named BINO, and Encoder-Decoder framework which uses an Attribution Selector with biomedical notion to hold the semantic consistency of biomedical domain. Besides, a structural extractor is applied to incorporate structural information of biomedical instances into the model to maintain the logical coherence between biomolecules and diseases. Extensive experiments on three selective Bio-RE datasets demonstrate that BINO can effectively preserve the semantic consistency of biomedical domain.

6 Limitations

Biomedical datasets contain massive hyphens and acronyms which impede models' understanding of the whole sentences. Adding a high-quality and dynamic dictionary to the model may be a solution to alleviate this issue. This paper does not discuss the effects of dictionaries on correcting entities' boundaries and data bias. Besides, newly emerging drug entity detection is another challenge. The proposed framework is model-agnostic which can be applied to other models on unseen data groups in theory, however, this paper does not discuss the domain-transferring and adaptation.

7 Ethics Discussion

This work partially adopted generative models, such as gpt-3.5-turbo, to generate pseudo instances for data augmentation. The limitations and hallucination issues have been discussed in this paper.

References

Ateret Anaby-Tavor, Boaz Carmeli, Esther Goldbraich, Amir Kantor, George Kour, Segev Shlomov, Naama Tepper, and Naama Zwerdling. 2020. Do not have enough data? deep learning to the rescue! In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7383–7390.

Markus Bayer, Marc-André Kaufhold, Björn Buchhold, Marcel Keller, Jörg Dallmeyer, and Christian Reuter. 2023. Data augmentation in natural language processing: a novel text generation approach for long and short text classifiers. *International journal of machine learning and cybernetics*, 14(1):135–150.

Kevin G Becker, Kathleen C Barnes, Tiffani J Bright, and S Alex Wang. 2004. The genetic association database. *Nature genetics*, 36(5):431–432.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*.

Àlex Bravo, Janet Piñero, Núria Queralt-Rosinach, Michael Rautschka, and Laura I Furlong. 2015. Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research. *BMC bioinformatics*, 16:1–17.

Hengyi Cai, Hongshen Chen, Yonghao Song, Cheng Zhang, Xiaofang Zhao, and Dawei Yin. 2020. Data manipulation: Towards effective instance learning for neural dialogue generation via learning to augment and reweight. *arXiv preprint arXiv:2004.02594*.

Hannah Chen, Yangfeng Ji, and David Evans. 2020. Finding friends and flipping frenemies: Automatic paraphrase dataset augmentation using graph theory. *arXiv preprint arXiv:2011.01856*.

Bosheng Ding, Linlin Liu, Lidong Bing, Canasai Kruengkrai, Thien Hai Nguyen, Shafiq Joty, Luo Si, and Chunyan Miao. 2020. Daga: Data augmentation with a generation approach for low-resource tagging tasks. *arXiv preprint arXiv:2011.01549*.

Sreyan Ghosh, Utkarsh Tyagi, Manan Suri, Sonal Kumar, S Ramaneswaran, and Dinesh Manocha. 2023. Aclm: A selective-denoising based generative data augmentation approach for low-resource complex ner. *arXiv preprint arXiv:2306.00928*.

María Herrero-Zazo, Isabel Segura-Bedmar, Paloma Martínez, and Thierry Declerck. 2013. The ddi corpus: An annotated corpus with pharmacological substances and drug–drug interactions. *Journal of biomedical informatics*, 46(5):914–920.

Xuming Hu, Aiwei Liu, Zeqi Tan, Xin Zhang, Chenwei Zhang, Irwin King, and Philip S Yu. 2023. Gda: Generative data augmentation techniques for relation extraction tasks. *arXiv preprint arXiv:2305.16663*.

Fred Jelinek, Robert L Mercer, Lalit R Bahl, and James K Baker. 1977. Perplexity—a measure of the difficulty of speech recognition tasks. *The Journal of the Acoustical Society of America*, 62(S1):S63–S63.

Chang Jin, Shigui Qiu, Nini Xiao, and Hao Jia. 2022. Admix: A mixed sample data augmentation method for neural machine translation. *arXiv preprint arXiv:2205.04686*.

Martin Krallinger, Obdulia Rabal, Saber A Akhondi, Martın Pérez Pérez, Jesús Santamaría, Gael Pérez Rodríguez, Georgios Tsatsaronis, Ander Intxaurrondo, José Antonio López, Umesh Nandal, et al. 2017. Overview of the biocreative vi chemical-protein interaction track. In *Proceedings of the sixth BioCreative challenge evaluation workshop*, volume 1, pages 141–146.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Kenton Lee, Kelvin Guu, Luheng He, Tim Dozat, and Hyung Won Chung. 2021. Neural data augmentation via example extrapolation. *arXiv preprint arXiv:2102.01335*.

Zachary C Lipton. 2018. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57.

Linlin Liu, Bosheng Ding, Lidong Bing, Shafiq Joty, Luo Si, and Chunyan Miao. 2021. Mulda: A multilingual data augmentation framework for low-resource cross-lingual ner. 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 5834–5846.

Jesutofunmi A Omiye, Haiwen Gui, Shawheen J Rezaei, James Zou, and Roxana Daneshjou. 2024. Large language models in medicine: The potentials and pitfalls: A narrative review. *Annals of Internal Medicine*.

OpenAI. 2023. Gpt-4 technical report.

Yannis Papanikolaou and Andrea Pierleoni. 2020. Dare: Data augmented relation extraction with gpt-2. *arXiv preprint arXiv:2004.13845*.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Dan Roth. 2017. Incidental supervision: Moving beyond supervised learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Olivier Taboureau, Sonny Kim Nielsen, Karine Audouze, Nils Weinhold, Daniel Edsgärd, Francisco S Roque, Irene Kouskoumvekaki, Alina Bora, Ramona Curpan, Thomas Skøt Jensen, et al. 2010. Chemprot: a disease chemical biology database. *Nucleic acids research*, 39(suppl_1):D367–D372.

Robert Tinn, Hao Cheng, Yu Gu, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2023. Fine-tuning large neural language models for biomedical natural language processing. *Patterns*, 4(4).

Fiona J Tweedie and R Harald Baayen. 1998. How variable may a constant be? measures of lexical richness in perspective. *Computers and the Humanities*, 32:323–352.

Jens Van Nooten and Walter Daelemans. 2023. Improving dutch vaccine hesitancy monitoring via multi-label data augmentation with gpt-3.5. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 251–270.

Jason Wei and Kai Zou. 2019. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. 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 6382–6388.

Jiashu Xu, Mingyu Derek Ma, and Muhao Chen. 2022. Can nli provide proper indirect supervision for low-resource biomedical relation extraction? *arXiv preprint arXiv:2212.10784*.

Michihiro Yasunaga, Jure Leskovec, and Percy Liang. 2022. Linkbert: Pretraining language models with document links. *arXiv preprint arXiv:2203.15827*.

Hui-Syuan Yeh, Thomas Lavergne, and Pierre Zweigenbaum. 2022. Decorate the examples: A simple method of prompt design for biomedical relation extraction. *arXiv preprint arXiv:2204.10360*.

Wenyuan Zeng, Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2016. Incorporating relation paths in neural relation extraction. *arXiv preprint arXiv:1609.07479*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*.

Zexuan Zhong and Danqi Chen. 2020. A frustratingly easy approach for entity and relation extraction. *arXiv preprint arXiv:2010.12812*.

Ran Zhou, Xin Li, Ruidan He, Lidong Bing, Erik Cambria, Luo Si, and Chunyan Miao. 2021. Melm: Data augmentation with masked entity language modeling for low-resource ner. *arXiv preprint arXiv:2108.13655*.