Journal of Research in Science and Engineering (JRSE)
ISSN: 1656-1996 Volume-7, Issue-12, December 2025
DOI: 10.53469/jrse.2025.07(12).04

A Data Synthesis Framework for Few-Shot Biomedical Relation Extraction via Semantic Deconstruction and Re-synthesis

Hai Zhu

School of Business, Nanfang College Guangzhou, Guangzhou, Guangdong, China
zhuh@nfu.edu.cn

Abstract: ***In the field of Biomedical Natural Language Processing (BioNLP), particularly in Drug-Drug Interaction (DDI) relation extraction tasks, the development of high-performance deep learning models has long been constrained by the severe scarcity of high-quality annotated data. Although Large Language Models (LLMs) have demonstrated exceptional capabilities in text generation and understanding, directly applying them to data synthesis in specialized domains often faces severe challenges, including frequent "factual hallucinations," monotonous logical structures, and privacy leakage. Existing data augmentation methods, such as simple synonym replacement or two-stage prompting approaches based on the "generate-filter" paradigm, alleviate the data scarcity issue to some extent but fail to guarantee the logical self-consistency and structural diversity of synthetic data from the source. To overcome this bottleneck, this paper proposes a novel "Semantic Deconstruction and Re-synthesis" (SDR) data synthesis framework. Unlike traditional black-box generation modes, SDR adopts a white-box methodology characterized by "Deconstruction-Planning-Reconstruction." First, a small number of seed samples are deconstructed into fine-grained semantic fragments to construct an extensible semantic fragment repository. Subsequently, an LLM acts as a "Knowledge Architect" to logically recombine these fragments. Crucially, we introduce an In-process Verification mechanism to double-check the abstract factual skeleton before text generation, thereby blocking the propagation of logical fallacies. Extensive experiments on the DDI Corpus dataset demonstrate that a local model (BioBERT) fine-tuned with SDR-synthesized data significantly outperforms existing methods in 1-shot, 5-shot, and 10-shot settings. Notably, in the 10-shot scenario, SDR achieves an F1 score of 60.43%, demonstrating the framework's superior effectiveness and generalization capabilities in low-resource scenarios.***

Keywords: Data Synthesis, Large Language Models, Relation Extraction, Semantic Deconstruction, In-process Verification, Drug-Drug Interaction.

1. Introduction

With the exponential growth of biomedical literature, automatically extracting structured knowledge, such as Drug-Drug Interactions (DDI), from massive unstructured texts has become a critical task in bioinformatics, computational pharmacology, and precision medicine. Relation Extraction (RE), a core component of information extraction, aims to identify semantic relationships between entity pairs in text. However, biomedical texts often contain complex syntactic structures and highly specialized terminology, necessitating large-scale, expert-annotated high-quality datasets to train high-precision RE models. The acquisition of such data is prohibitively expensive and involves legal and ethical issues regarding patient privacy, leading to a severe phenomenon of "data hunger".

To address the issue of data scarcity, Few-Shot Learning (FSL) has emerged. Its core objective is to enable models to adapt quickly to new tasks using only a minimal number of samples (e.g., 1 or 5 samples per class). In recent years, Pre-trained Language Models (PLMs) such as BioBERT [1] and ClinicalBERT [2] have achieved immense success in fully supervised settings. However, in few-shot scenarios, these models often face severe overfitting and drastic performance degradation due to the lack of sufficient samples to adjust their massive parameters.

With the advent of Large Language Models (LLMs) like GPT-4 and Gemini, utilizing LLMs for data augmentation or direct reasoning has become a new research hotspot. However, Tang et al. [3] pointed out that directly using LLMs for Zero-shot clinical text mining yields suboptimal results, with F1 scores far lower than supervised models. More critically, LLMs suffer from inherent "Hallucination" problems, where generated text reads fluently but may contain erroneous biomedical facts (e.g., fabricating non-existent drug side effect mechanisms) [4]. Furthermore, uploading sensitive medical data to closed-source LLM APIs raises non-negligible privacy concerns.

Addressing these challenges, this paper proposes the SDR (Semantic Deconstruction and Re-synthesis) framework. Our core insight is that high-quality biomedical synthetic data should not be the product of free association by LLMs, but rather a "controlled recombination" based on strict rules and logic. SDR transforms the data generation process from a "black box" to a "white box," with the following specific contributions:

(1) Proposal of the SDR Framework: We introduce a standardized data synthesis pipeline comprising five stages, applying the "Deconstruction-Planning-Reconstruction" paradigm to the high-risk biomedical RE domain for the first time. By constructing a "Semantic Fragment Library," we achieve a balance between "hard content control" and "soft style control" during the data generation process.

(2) Implementation of In-process Verification: Unlike existing "post-generation verification" strategies, SDR innovatively introduces a validation mechanism during the "Structured Fact Synthesis" stage. By utilizing LLMs to check the logical self-consistency of the abstract factual skeleton before text generation, we block the emergence of

factual hallucinations at the Source rather than the Outcome.

(3) Superior Experimental Performance: We conducted extensive experiments on the DDI Corpus dataset. The results indicate that the local BioBERT model trained with SDR-generated synthetic data comprehensively surpasses strong baselines utilizing direct LLM inference in 1-shot, 5-shot, and 10-shot settings. SDR not only resolves privacy and API dependency issues but also provides a generalizable solution for low-resource information extraction in vertical domains.

2. Related Work

This section systematically reviews two categories of work closely related to this study: few-shot relation extraction in the biomedical domain and data synthesis techniques based on Large Language Models

2.1 Biomedical Few-Shot Relation Extraction

Traditional biomedical relation extraction relies primarily on supervised learning. With the development of pre-trained models, BioBERT by Lee et al. and ClinicalBERT by Alsentzer et al. significantly improved downstream task performance through continued pre-training on massive biomedical corpora. However, these models perform poorly in few-shot scenarios.

To cope with the few-shot challenge, Moscato et al. [5] proposed a multi-task learning framework based on MT-DNN, attempting to utilize knowledge from other datasets to assist the current task. Although this method improved recall, it relies heavily on task similarity and is prone to Negative Transfer. Guo et al. [6] introduced a Prompt Learning-based method combined with simple data augmentation (e.g., back-translation, synonym replacement), which enriched data diversity to some extent, but traditional augmentation means cannot generate new samples with significant syntactic structural differences.

With the rise of LLMs, Agrawal et al. [7] explored the potential of LLMs as few-shot clinical information extractors, finding them capable of competing with fully supervised models on specific tasks. Yeh et al. [8] further proposed a template-based Prompt design method, converting relation extraction into a cloze test task. However, recent research by Nagar et al. [9] cast doubt on these findings, pointing out that LLMs are not "Zero-shot Reasoners"; they often exhibit a lack of reasoning ability when dealing with complex biomedical negations and conditional sentences. Additionally, Ma et al. [10] found that LLMs are not good "extractors" in few-shot settings but are better suited as "Rerankers".

2.2 Data Synthesis and Augmentation Based on LLMs

Data augmentation is a classic method for solving data scarcity. Traditional NLP data augmentation techniques include back-translation [11] and contextual augmentation [12]. However, samples generated by these methods are highly dependent on the original samples in terms of semantics and structure, lacking Novelty.

The emergence of LLMs has pushed data augmentation into a new stage of "Data Synthesis." A survey by Wang et al. [13] points out that utilizing LLMs for From-scratch Generation has become a new trend. In the general domain, Zhou et al. [14] proposed the PGA-SciRE framework, comparing "rewriting" and "generation" strategies, finding that while rewriting preserves semantics, it lacks diversity, whereas direct generation offers diversity but contains significant noise.

To address generation quality issues, Gero et al. [15] and Ma et al. [16] explored "Self-Verification" and "Structure-to-Text" generation modes, respectively. Ma et al.'s STAR framework improved extraction performance by generating structure first and then text, but its structure generation process lacks strict logical constraints.

Huang et al. [17] achieved a breakthrough in mathematical reasoning with the KPDDS method, which deconstructs mathematical problems into "key points" and then recombines them to generate new problems. The MEPG framework by Zhao and Liu [18] introduced the concept of "Multi-Expert Planning" in image generation. These "Deconstruction - Recombination" ideas provide an important theoretical basis for this paper. However, biomedical texts require not only logical coherence but also strict adherence to pharmacological Factuality. Most current general synthesis methods lack an "In-process Verification" mechanism specifically for domain-specific facts, which is precisely the gap filled by the SDR framework proposed in this paper.

3. Methodology

The proposed SDR framework consists of five cascaded stages: (1) Gold-Standard Seed Set Curation; (2) Semantic Fragment Deconstruction & Expansion; (3) Structured Fact Synthesis & In-process Verification; (4) Fluent Sentence Generation; (5) Dual Quality Assurance & Alignment. Given a minimal seed set $\mathcal{D}_{\text{seed}} = \{(x_i, y_i)\}_{i=1}^{K}$, our goal is to generate a large-scale high-quality dataset $\mathcal{D}_{syn}$, such that a model M trained on $\mathcal{D}_{syn}$ maximizes performance on the test set.

3.1 Stage 1: Gold-Standard Seed Set Curation

High-quality generation begins with high-quality input. Instead of directly using all training data, we design a Prompt to guide the LLM to act as a "BioNLP Expert," selecting the most representative samples from the raw data. We pay special attention to sentence diversity, ensuring the seed set contains various syntactic structures such as active voice, passive voice, conditional clauses, and negations. These seeds provide not only knowledge but also serve as "Style Anchors" for subsequent generation. We use Prompts to guide the LLM to select the most representative K samples from the original training set (where $K \in \{1,3,5,10\}$). Let S_{style} be the linguistic style features of the seed set; our goal is to mine distinct features as much as possible. This stage extracts not only entities and relations but, more importantly, captures the syntactic features of medical texts to serve as style constraints for subsequent generation.

3.2 Stage 2: Semantic Fragment Deconstruction & Expansion

Inspired by KPDDS, we deconstruct sentence x into a set of semantic fragments. We define the deconstruction function $f_{\text{dec}}(x) \rightarrow \{E, T, C\}$, where:

- E (Entity): Includes not only entity names (e.g., *Warfarin*) but also entity types (e.g., *Drug, Protein*) and attributes.
- T(Trigger/Relation): Core verbs or phrases carrying the relationship (e.g., *inhibits, metabolizes, co-administration*) and logical connectors (*although, resulting in*).
- C (Condition): Modifiers limiting the conditions under which the relationship occurs (e.g., *at high doses, in vivo*).

To increase diversity, we introduce an expansion function $f_{\text{exp}}(F_{\text{seed}}) \rightarrow F_{\text{expanded}}$. Leveraging the LLM's knowledge base, we perform synonym replacement on triggers and scenario extension on context constraints, thereby constructing a massive "Biomedical Semantic Fragment Library".

3.3 Stage 3: Structured Fact Synthesis & In-process Verification

This is the core innovation of SDR. Rather than generating text directly, we first synthesize an "Abstract Factual Skeleton." First, we define a synthesis function $f_{\text{syn}}(F_{\text{expanded}}) \rightarrow \mathcal{S}_{\text{fact}}$. The model randomly combines fragments to form new relational logic based on predefined pharmacological rules.

Secondly, traditional generation methods perform checks after generating the complete text, at which point errors are already solidified and difficult to correct. SDR introduces an in-process verification function to intercept errors before the skeleton is "instantiated" into a concrete sentence. The verification function $V_{\text{logic}}(\mathcal{S}_{\text{fact}})$ is defined as follows:

$$V_{\text{logic}}(\mathcal{S}_{\text{fact}}) = \begin{cases} 1, & \text{if Consistency}(\mathcal{S}_{\text{fact}}) \; is \text{ True} \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

For example, if a skeleton contains both "Drug A antagonizes Drug B" and "Drug A potentiates Drug B" without a transitional logic, then $V_{\text{logic}} = 0$, and the skeleton is directly discarded. This step blocks logical hallucinations at the source.

3.4 Stage 4: Fluent Sentence Generation

For a skeleton $s \in \mathcal{S}_{\text{fact}}$ that passes verification, we utilize the LLM to model the conditional probability $P(y \mid s, S_{\text{style}})$ to "translate" it into a natural language sentence y.

To simulate real data distribution, we require the model to generate multiple syntactic variants (e.g., inverted sentences, relative clauses) for the same s, ensuring the syntactic diversity of $\mathcal{D}_{\text{syn}}$. For example:

Original (Standard Statement): Drug A inhibits the metabolism of Drug B.

Variant A (Passive + Condition): The metabolism of Drug B is significantly inhibited when co-administered with Drug A.

Variant B (Complex Clause): Patients taking Drug A may experience elevated levels of Drug B due to metabolic inhibition.

3.5 Stage 5: Dual Quality Assurance & Alignment

Finally, to ensure the label accuracy of the synthetic data, we implement a strict "Post-hoc Dual Check," defined as $V_{\text{final}}(y, s)$:

$$V_{\text{final}}(y, s) = V_{\text{fidelity}}(y, s) \wedge V_{\text{alignment}}(y, R) \quad (2)$$

Here, the fidelity check V_{fidelity} ensures that the generated text y does not omit key information from skeleton s and does not introduce new entities or erroneous relations. The alignment check $V_{\text{alignment}}$ verifies whether text y can be clearly classified into the target relation R without ambiguity. Only when $V_{\text{final}}(y, s) = 1$ is the sample (y, R) added to the final dataset $\mathcal{D}_{\text{syn}}$.

Through this precise five-stage pipeline, SDR is capable of generating large-scale, high-fidelity, and logically rigorous synthetic datasets.

4. Experiments

4.1 Experimental Setup

We selected the SemEval-2013 DDI (Drug-Drug Interaction) Corpus as our experimental benchmark. This dataset includes two subsets, MedLine and DrugBank, and defines four interaction types: *Mechanism, Effect, Advice,* and *Int*. To simulate real-world low-resource scenarios, we adopted the standard K-shot N-way setting. For each relation type, we randomly sampled 1, 5, and 10 examples from the training set as seeds. Experiments were repeated 5 times, and the average was reported.

We compared SDR with three categories of strong baselines:

Standard Fine-tuning: Clinical-BERT and BioBERT directly fine-tuned on K-shot real data.

LLM Direct Inference: Simple Prompt using the Google Gemini-Flash 2.0 model with basic instructions for zero/few-shot inference.

Complex Prompting Methods: Extract-Verify Prompt, referencing Gero et al., incorporating a self-verification step into the extraction process.

Ours: BioBERT + SDR. We first used Gemini-Flash 2.0 to generate synthetic data via the SDR framework, then used this data to fine-tune the BioBERT model locally.

4.2 Main Results

Table 1 shows the Micro-F1 scores of different methods on the DDI dataset.

Table 1: Micro-F1 Scores (%) on DDI Dataset under Different Shot Settings.

Model / Method	1-shot	5-shot	10-shot
Clinical-BERT	9.55	22.45	31.20
BioBERT	10.62	25.88	35.40
Simple Prompt	39.75	42.60	43.23
Extract-Verify Prompt	43.89	45.12	48.50
BioBERT + SDR (Ours)	**50.85**	**54.75**	**60.43**

As seen in the table, under the 1-shot setting, traditional fine-tuning methods almost completely fail because a single sample is insufficient to support parameter updates in deep neural networks. Even the powerful Extract-Verify method struggles to break through, limited by the capacity ceiling of In-context Learning. In contrast, SDR achieved an F1 score of 50.85%. This proves that SDR's "Semantic Expansion" mechanism effectively infers general patterns from a single instance ("drawing inferences about other cases from one instance"), generating data covering the latent feature space via fragment recombination, thus enabling the local model to learn robust decision boundaries.

As the shot count increases from 1 to 10, the performance gain of Extract-Verify slows down. This is because prompting methods are constrained by context window length and struggle to fully utilize information from more samples. Conversely, the performance of the SDR method shows sustained growth as the shot count increases. This indicates that SDR can effectively mine the rich semantic patterns contained in more seed samples and amplify them into large-scale training signals.

5. Analysis and Discussion

5.1 Ablation Study

To investigate the contribution of individual components within the SDR framework, we conducted an ablation study under the 10-shot setting. The results are shown in Table 2.

Table 2: Ablation Study of SDR Components (10-shot, DDI Corpus).

Variant	Micro-F1 (%)
Full SDR Framework	**60.43**
w/o Semantic Expansion	55.39
w/o In-process Verification	58.80
w/o Dual QA	59.21

Table 2 highlights that Semantic Expansion is crucial; removing it resulted in a 5.04% performance drop. This suggests that merely reordering original sentences (similar to traditional data augmentation) is insufficient. The "brainstorming" capability of the LLM—introducing external knowledge such as synonyms and related mechanism descriptions—plays a decisive role in covering the long-tail distribution of the test set. Removing In-process Verification also caused a performance decline. This strongly supports our view: Logical errors must be intercepted before generation. Once fluent but logically flawed text (i.e., hallucination) is generated, subsequent QA steps are often deceived by its superficial fluency and fail to eliminate it precisely. In-process verification effectively acts as a "logical firewall".

5.2 Case Study

To visually demonstrate the advantages of SDR, we selected a specific DDI instance for comparative analysis.

- Original Seed: "Co-administration of drugs metabolized by CYP2D6 may require dose adjustment." (Type: Advice)
- Test Sample: "Although Warfarin is typically metabolized by CYP2C9, simultaneous use with Amiodarone significantly inhibits this pathway, leading to potential toxicity." (Type: Mechanism & Advice)

The Extract-Verify method incorrectly predicted Metabolism or Effect. The reason is that the model focused excessively on the keyword metabolized and failed to correctly process the complex transitional and negative logic of "Although... inhibits...", leading to a misjudgment of the sentence's core intent (mechanism of inhibition).

In Stage 2 of SDR, metabolized by was expanded to inhibits the pathway of; in Stage 3, a skeleton containing Constraint: Negation/Concession was constructed. SDR generated the following synthetic sample for training:

- "Despite being a substrate for CYP3A4, the drug's clearance is halted when introduced with strong inhibitors, suggesting a mechanism based on enzymatic competition."

Consequently, BioBERT correctly predicted the relationship as Mechanism. Since the local model encountered a large volume of complex sentence structures generated by SDR containing logical connectors like Although/Despite during training, it learned to capture the deep semantic dependencies of sentences rather than relying solely on keyword matching.

5.3 Error Analysis

Although SDR performs excellently, the improvement in certain categories (such as Int, general interactions) is relatively small. The main reason is that the definition of such relationships is rather vague. SDR's "structured" generation tends to produce sentences with clear definitions and logic (such as Mechanism), and it struggles to simulate the vague, atypical expressions found in the Int class. Future work could explore introducing noise injection mechanisms to simulate non-standard expressions in real data.

6. Conclusion

Addressing the challenges of data scarcity and privacy in biomedical relation extraction tasks, this paper proposes the SDR (Semantic Deconstruction and Re-synthesis) data synthesis framework based on Large Language Models. By formalizing the data generation process into a white-box flow of "Deconstruction-Planning-Reconstruction" and introducing a critical "In-process Verification" mechanism, SDR successfully overcomes the issues of factual hallucination and logical monotony inherent in traditional

LLM generation. Experiments demonstrate that SDR can generate high-quality, diverse synthetic data in few-shot scenarios, significantly improving downstream model performance and exhibiting excellent scalability.

SDR not only provides a low-cost data production paradigm for BioNLP but its concept of "In-process Verification" also offers important insights for LLM applications in other vertical domains requiring high factuality (such as law and finance). Future work will focus on extending the SDR framework to more complex tasks such as entity recognition and event extraction.

References

[1] J. Lee et al., "BioBERT: a pre-trained biomedical language representation model for biomedical text mining," Bioinformatics, vol. 36, no. 4, pp. 1234–1240, Sep. 2019, doi: 10.1093/bioinformatics/btz682.

[2] E. Alsentzer et al., "Publicly Available Clinical BERT Embeddings," 2019, arXiv. [Online]. Available: https://arxiv.org/abs/1904.03323

[3] R. Tang, X. Han, X. Jiang, and X. Hu, "Does Synthetic Data Generation of LLMs Help Clinical Text Mining?," 2023, arXiv:2303.04360.

[4] B. Jimenez Gutierrez et al., "Thinking about GPT-3 In-Context Learning for Biomedical IE? Think Again," in Findings of the Association for Computational Linguistics: EMNLP 2022, Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.findings- emnlp.329.

[5] V. Moscato, G. Napolano, M. Postiglione, and G. Sperli, "Multi-task learning for few-shot biomedical relation extraction," Artificial Intelligence Review, vol. 56, no. 11, pp. 13743–13763, 2023.

[6] B. Guo, D. Zhao, X. Dong, J. Meng, and H. Lin, "Few-shot biomedical relation extraction using data augmentation and domain information," Neurocomputing, vol. 595, p. 127881, 2024.

[7] M. Agrawal, S. Hegselmann, H. Lang, Y. Kim, and D. Sontag, "Large language models are few-shot clinical information extractors," in Proceedings of EMNLP 2022, 2022.

[8] H. S. Yeh, T. Lavergne, and P. Zweigenbaum, "Decorate the examples: A simple method of prompt design for biomedical relation extraction," in LREC 2022, pp. 3780–3787, 2022.

[9] A. Nagar et al., "LLMs are not Zero-Shot Reasoners for Biomedical Information Extraction," in The Sixth Workshop on Insights from Negative Results in NLP, 2025.

[10] Y. Ma, Y. Cao, Y. Hong, and A. Sun, "Large Language Model Is Not a Good Few-shot Information Extractor, but a Good Reranker for Hard Samples!," in Findings of the Association for Computational Linguistics: EMNLP 2023, Association for Computational Linguistics, 2023.

[11] Q. Xie, Z. Dai, E. Hovy, M.-T. Luong, and Q. V. Le, "Unsupervised Data Augmentation for Consistency Training," 2019, arXiv. [Online]. Available: https://arxiv.org/abs/1904.12848.

[12] S. Kobayashi, "Contextual Augmentation: Data Augmentation by Words with Paradigmatic Relations," 2018, arXiv:1805.06201.

[13] K. Wang, et al., "A Survey on Data Synthesis and Augmentation for Large Language Models," 2024, arXiv:2410.12896.

[14] Y. Zhou, S. Shan, H. Wei, Z. Zhao, and W. Feng, "PGA-SciRE: Harnessing LLM on Data Augmentation for Enhancing Scientific Relation Extraction," 2024, arXiv:2405.20787.

[15] Z. Gero et al., "Self-Verification Improves Few-Shot Clinical Information Extraction," 2023, arXiv: 2306.00024.

[16] M. D. Ma et al., "STAR: Boosting Low-Resource Information Extraction by Structure-to-Text Data Generation with Large Language Models," AAAI, vol. 38, no. 17, pp. 18751–18759, 2024.

[17] Y. Huang et al., "Key-Point-Driven Data Synthesis with Its Enhancement on Mathematical Reasoning," AAAI, vol. 39, no. 23, pp. 24176–24184, 2025.

[18] Y. Zhao and L. Liu, "MEPG: Multi-Expert Planning and Generation for Compositionally-Rich Image Generation," 2025, arXiv:2509.04126.