SSDAU: Structured Semantic Data Augmentation for Joint Entity and Relation Extraction

Anonymous submission

Abstract

Joint Entity and Relation Extraction (JERE) is highly susceptible to weak generalization due to low-quality training data. Data augmentation is a common strategy to enhance model generalization across different domains. However, existing data augmentation methods often overlook text relevance and may disrupt semantic structures and dependencies, making it difficult to generate effective augmented data for improving model generalization. In this paper, we propose **Structured Semantic Data Augmentation (SSDAU)**, a novel method designed to preserve the semantic structure of text during augmentation. SSDAU segments text based on entity labels and employs an encoder to capture semantic features of entities through context awareness. It then performs entity semantic restructuring to generate augmented data. To distinguish semantically similar entities, SSDAU fuses contextualized embeddings with traditional similarity scores. To mitigate potential topic ambiguity and information loss, we apply the BERTTopic model to filter out irrelevant topics, ensuring topic consistency. We evaluate SSDAU on datasets with different annotation types and compare its performance on five representative JERE models against seven popular data augmentation baselines. Experiments demonstrate that SSDAU generates semantically consistent data with superior robustness against ambiguity (8.26% F1 decrease vs. 31.91% for baselines), significantly outperforming all existing methods across all metrics.

1 Introduction

Joint Entity and Relation Extraction (JERE) is widely used for representation learning on text data due to its strong performance in applications such as information retrieval (Lin et al., 2020), question answering (Abdelaziz et al., 2021), and text summarization (Zhong et al., 2020). The generalization performance of JERE models heavily depends on the quality and scale of the training data. A common strategy to enhance generalization is data augmentation. Techniques such as MixUp (Cheng et al., 2020) and back-translation (Xie et al., 2020) enable efficient expansion of the training set by generating new data with subtle perturbations derived from the original samples.

However, a key challenge in applying existing techniques to enhance the generalization of JERE models is that introducing noise or perturbations into the original data may weaken entity relevance (Kambhatla et al., 2022). Training on incorrectly generated data can ultimately degrade JERE models' performance. Additionally, entities are often involved in multiple triples with complex semantic relations and dependencies. Existing data augmentation methods can disrupt the structures and dependencies, leading to issues such as overlapping relations and cascading (Liu et al., 2020).

To address this issue, we propose Structured Semantic Data Augmentation (SSDAU) to preserve the semantic structure of text during data augmentation. Instead of directly perturbing text, SSDAU aligns triplet text to maintain semantic integrity. First, we use a feature-based encoder to segment the text, ensuring that each segment retains the semantics of its neighboring regions. Next, we match segments with similar semantic labels using a decoder. Our approach integrates contextualized embeddings with pretrained pooler weights to differentiate semantically similar but distinct entities and employs topic-aware consistency filtering to prevent error propagation. Finally, we substitute text with high similarity to reorganize the original text, generating augmented data while preserving semantic coherence. To mitigate error propagation, we employ a topic-aware consistency filtering mechanism that scores candidate triples using the BERTTopic model and eliminates those inconsistent with gold-standard semantics.

To assess the effectiveness of SSDAU, we compared its performance on four widely used datasets

with seven baseline methods. The experimental results validate our main finding: SSDAU consistently outperforms other methods in both common and low-quality data scenarios. Our ablation studies further reinforce this conclusion. Even when faced with semantic ambiguity, SSDAU maintains stable performance with an average F1 score decrease of only 8.26% across all datasets, while other baselines suffer substantial degradation. This robust performance extends to overall effectiveness, with SSDAU achieving average precision of 92.03% and F1 score of 91.96% across all datasets, substantially outperforming all baseline methods including recent approaches like ChatIE.

2 Related Work

Semantic Match Semantic matching is a subtask of text matching used to retrieve semantically similar texts in search scenarios (Wu et al., 2022). Recent studies show that pre-training semantic classification models can compress large amounts of text and improve the generalization of semantic matching models (Brown et al., 2020). For example, the *Similarities* tool (Zhang Bingyu, 2022) enhances practical applications for text semantic matching, especially in text relation extraction. Based on existing techniques, we improve JERE by incorporating text semantic matching.

Data Augmentation Data augmentation is an effective and efficient method to improve machine learning model performance, especially in data-limited environments (Cashman et al., 2020). Common techniques in NLP include word replacement (Wei and Zou, 2019), word vector replacement (Wang and Yang, 2015), masked language model replacement (Jiao et al., 2020), back translation (Zhang et al., 2020), and adding noise (Min et al., 2020; Yan et al., 2019; Hou et al., 2018). Unlike simple perturbation (Liu et al., 2020) or extra augmentor models (Hou et al., 2021; Hu et al., 2019), we propose sampling-based augmentation to generate data with the same semantic structure while maintaining the logic of the samples.

3 Method

In this section, we first define the problems. Then, we introduce the three main components of SSDAU: 1) data discretization and reconstruction, 2) structured semantic data augmentation, and 3) scoring-based consistency filtering. Figure 1 depicts the overall framework of SSDAU.

3.1 Preliminaries

Given set of sentences $S = \{s_1, s_2, ..., s_N\}$ containing L token and K predefined relations $R = \{r_1, r_2, ..., r_K\}$, we extract entities and relations to construct triples $T = \{(h_i, r_i, t_i)\}_{i=1}^{M}$ in S, where h_i, t_i are the head and tail entities, respectively, N represents the number of sentences, M represents the number of triples. We store the knowledge as a three-dimensional matrix M^{L*K*L}.

Since triplets are the core output format of JERE, we use the triplet as the basic unit of data augmentation and partition the text according to the triplet to obtain three series of text collections. To preserve the contextual semantics of the segmented text, we keep the contextual token l of each segmented text and record the location of each cut point p.

3.2 Data Discretization and Reconstruction

Encoder We use the triplet as the basic unit of data augmentation to eliminate the noise from textual perturbations. We design a text feature-based encoder E (the structure is shown in Figure 2). The input of the encoder is the sentence text S. For each sentence s_i, we locate the specified text block $(q_{h_i}, q_{r_i}, q_{t_i})$ based on the triplet tags $(\rho_{h_i}, \rho_{r_i}, \rho_{t_i})$, and record the context token $(l_{h_i}, l_{r_i}, l_{t_i})$ and its cut position $(p_{h_i}, p_{r_i}, p_{t_i})$. The encoder processes all the input text and gets three output text collections according to the tag types: head entity collection Q_h, tail entity collection Q_t, and relation entity collection Q_r.

Decoder We then design a similarity-based text matching decoder D. The input of decoder D is (Q_h, Q_t, Q_s). The decoder divides the text collections according to the relation types and label types to get $L * K * L$ groups $B = \{B_1, B_2, ..., B_{LKL}\}$, where each group has the same relation type and the same label.

3.3 Structured Semantic Data Augmentation

Discrete Text Matching We designed a text matcher based on the semantic similarity evaluation tool *Similarities* (Zhang Bingyu, 2022) to align the decoder's output. A text block b in an output group $B_i = b_1, b_2, \ldots, b_j$ from the decoder stores the text q, context tokens l, label type ρ, and segmentation position p. We perform matching across all b in different text corpora B_i, incorporating semantic, syntactic, and lexical similarity evaluations, as well as context token similarity assessments. To effectively distinguish between semantically similar but

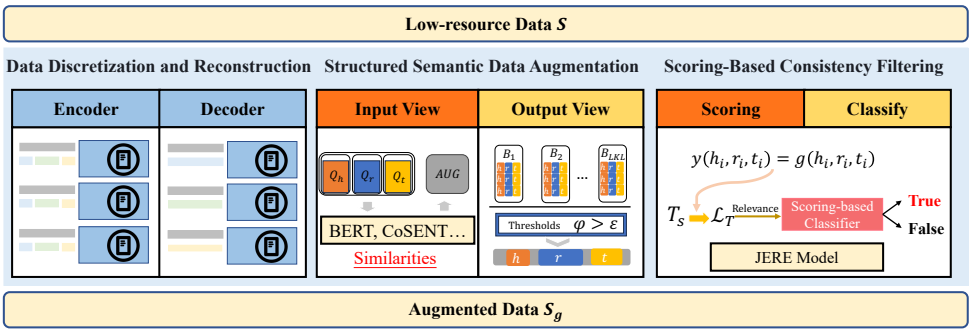


Figure 1: **Overview of SSDAU**. The ***Data Discretization and Reconstruction*** component discretizes the text data S semantically using the *Encoder* and outputs text collections in the form of segmented sets. The *Decoder* then processes these segmented sets to facilitate the ***Structured Semantic Data Augmentation*** component, where the *Input View* is based on similarity matching, while the *Output View* focuses on augmenting the data. Finally, the ***Scoring-based Consistency Filtering*** component uses a structured semantic classifier to filter low-resource data, and the remaining augmented data $\mathcal{L}$ and T are used as augmented data S_g to train a more robust JERE model.

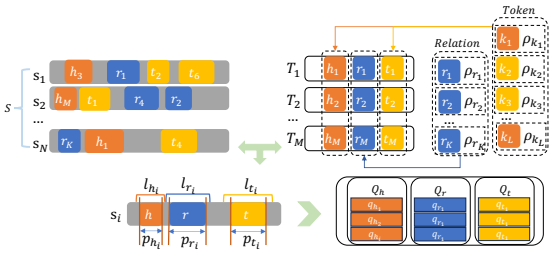


Figure 2: The structure of our feature-based encoder.

distinct entities, we enhance this process by incorporating contextualized [CLS] embeddings from a pretrained BERT encoder and apply pretrained pooler weights to compute entity-level semantic correlation. This correlation score is then fused with the original semantic similarity score to obtain a hybrid similarity measure. The matching results are normalized to a value between 0 and 1 and inserted into a priority queue sorted in descending order of similarity. Finally, for each B_i, we obtain a similarity-based priority queue P_i.

Data Augmentation After completing the similarity matching, we filter out data in the priority queue $P_i = P_1, P_2, ..., P_{KM}$ with a similarity score lower than the threshold ε. For the remaining data, we replace the text content of the corresponding text blocks based on the recorded segmentation position l in each block's information, thereby generating the augmented data.

3.4 Scoring-based Consistency Filtering

To further improve the quality of the augmented data, we employ a BERTTopic model to identify and retain key terms from topic descriptions. We then filter out augmented data associated with irrelevant topics, ensuring the topic coherence of the generated text.

First, we extract all entities and relations from the text. Then, we encode the tokens using BERT (Kenton and Toutanova, 2019), obtaining the corresponding entity tokens $l_1, l_2, \ldots, l_L$. Next, we combine entities and relations in the form of (l_h, r, l_t) and perform triplet extraction using joint entity and relation extraction (Shang et al., 2022). Finally, we apply a function to compute the correlation between the head and tail entities. The scoring function is defined as:

$$h \star t = \phi(W[l_h; l_t]^T + b) \quad (1)$$

where h and t represent the head and tail, respectively. $\star$ denotes circular correlation ($\mathbb{R} \times \mathbb{R}^d \rightarrow \mathbb{R}^d$). $W \in \mathbb{R}^{d_e \times 2d}$ and b are trainable weights and biases, respectively, where d_e denotes the dimension of the entity. $[;]$ is the concatenation operation and $\phi(\cdot)$ represents the ReLU activation function. In our implementation, we use the default pretrained W and b parameters, as recommended by HuggingFace for the BERTTopic model. Our ablation studies confirm that this pretrained initialization consistently outperforms both random and zero initialization.

We then incorporate the highly evaluated entity pairs with the relations and use the relational representation function $R \in \mathbb{R}^{d_e \times 4K}$. The vector function is defined as follows:

$$\upsilon_{(l_h, r_k, l_t)}{}_{k=1}^{K} = R^T \phi(drop(W[l_h; l_t]^T + b)) \quad (2)$$

where υ represents the score vector and $drop(\cdot)$ refers to the dropout strategy (Srivastava et al., 2014).

Next, we add the scoring vector υ to the softmax function to predict the corresponding labels. The formulated triples are presented as follows:

$$\zeta_{triple} = -\frac{\sum_{i,j,k} logP(y_{(l_i,r_k,l_j)}, g_{(l_i,r_k,l_j)}|S)}{L \times K \times L} \quad (3)$$

where $g_{(l_i,r_k,l_j)}$ represents the gold tag obtained from annotations. We match all triplets with the golden-label triplets to compute the topic score for each triplet. This topic-aware consistency filtering mechanism effectively mitigates error propagation by scoring candidate triples and eliminating those inconsistent with gold-standard semantics, ensuring robust performance even under semantic ambiguity. Finally, we select the high-scoring triplets as the topic relationships for the text. Augmented data in which these topic relationships have been replaced is filtered out, ensuring that the final augmented data remains both topic-relevant and structurally coherent.

4 Experiment

4.1 Experimental Setup

Baseline We compare SSDAU with seven commonly used data augmentation methods, including *word substitution* (WS) (Wei and Zou, 2019), *back translation* (BT) (Xie et al., 2020), *noise introduction* (NI) (Fanghua Ye, 2022), *same-tag semantic noise* (SSN) (Yan et al., 2019), *generative models* (GM) (Hou et al., 2021), Mixup (Hu et al., 2019), and ChatIE (Wei et al., 2023).

Protocol We select five models for three different types of JERE tasks: Multi-module Multi-Step (PRGC (Zheng et al., 2021), CasRel (Wei et al., 2020)), Multi-module One-Step (TPLinker (Wang et al., 2020), SPN4RE (Sui et al., 2020)), and One-module One-Step (OneRel) (Shang et al., 2022).

We use the following metrics to measure the effectiveness, performance, and adaptability of SSDAU: precision (Prec), F1-score (F1), and Intersection over Union (IoU). Additionally, we conduct significance tests (t-test) to verify the statistical significance of our improvements over other methods.

Implementation We conducted all experiments on a single server equipped with an Intel Xeon Gold 6248 2.50GHz CPU, two Tesla V100 SXM2 32GB GPUs, and Ubuntu 18.04.6 operating system. We reused the pre-trained BERT model (base-cased English) from Huggingface . For parameter initialization in the BERTTopic model, we use the default pretrained W and b parameters as recommended by HuggingFace.

Table 1: The number of augmented samples produced by SSDAU at various thresholds on different datasets.

Dataset	ε	Head	Relation	Tail	Sum.
NYT^*	$0.5 \sim 0.6$	15,062	243	11,300	26,605
	$0.6 \sim 0.7$	9,439	38	4,631	14,108
	$0.7 \sim 0.8$	1,825	19	1,365	3,209
	$0.8 \sim 0.9$	2,927	0	1,137	4,064
	$0.9 \sim 1.0$	960	0	1,546	2,506
$WebNLG^*$	$0.5 \sim 0.6$	7,082	2,742	8,116	17,940
	$0.6 \sim 0.7$	3,933	1,946	5,342	11,221
	$0.7 \sim 0.8$	2,049	2,162	1,557	5,768
	$0.8 \sim 0.9$	814	2,005	1,021	3,840
	$0.9 \sim 1.0$	5,463	890	2,929	9,282
NYT	$0.5 \sim 0.6$	13,507	234	10,076	23,817
	$0.6 \sim 0.7$	7,721	36	4,063	11,820
	$0.7 \sim 0.8$	4,922	13	1,588	6,523
	$0.8 \sim 0.9$	2,198	0	1,140	3,338
	$0.9 \sim 1.0$	3,700	0	1,051	4,751
$WebNLG$	$0.5 \sim 0.6$	4,023	3,186	6,028	13,237
	$0.6 \sim 0.7$	2,673	2,009	4,445	9,127
	$0.7 \sim 0.8$	968	1,345	1,123	3,436
	$0.8 \sim 0.9$	309	919	923	2,151
	$0.9 \sim 1.0$	3,019	444	6,935	10,398

Dataset We conduct our experiments on two representative English datasets, NYT (Sandhaus, 2008) and WebNLG (Gardent et al., 2017). Both types of datasets have two variations: fully annotated type (NYT, WebNLG) and partially annotated type (NYT*, WebNLG*). To evaluate robustness against semantic ambiguity, we also construct controlled evaluation datasets by injecting semantically ambiguous triples from SciER (Zhang et al., 2024) into all datasets at a 10% rate. These triples contain entities that appear semantically similar on the surface but differ in actual meaning (e.g., "Apple" as a company versus a fruit).

Evaluation and Selection of Thresholds Table 1 describes the number of augmented samples generated by SSDAU for different sets of semantic domains under various similarity thresholds. For the four datasets, we count the number of augmented data under different variable settings for different entities and relations. The results indicate that the number of augmented samples decreases as the threshold value increases.

4.2 Results

Comparison with Baselines Table 2 presents the effectiveness (Prec), performance (F1), and adaptability (IoU) results of SSDAU and seven baselines

Table 2: Comparison of SSDAU with baselines under normal conditions and with semantically ambiguous data.

Method	Quality	NYT*			WebNLG*			NYT			WebNLG		
		Prec.	F1	IoU	Prec.	F1	IoU	Prec.	F1	IoU	Prec.	F1	IoU
Original	–	90.17	91.45	84.24	90.62	90.25	82.23	92.83	92.17	85.47	90.66	89.08	80.31
WS	Clean	88.82	88.98	80.16	91.47	91.51	84.35	89.91	89.61	81.17	89.66	88.88	79.98
	With Error	75.60	76.30	61.80	77.10	75.20	60.30	78.20	71.40	55.60	75.80	69.50	53.20
BT	Clean	88.97	89.52	81.02	91.77	91.97	85.14	89.10	89.54	81.07	89.46	89.90	81.70
	With Error	64.40	69.80	53.60	72.40	69.50	53.30	71.10	56.30	39.20	66.60	37.70	23.20
NI	Clean	89.37	89.91	81.67	92.41	92.16	85.46	88.38	89.70	81.32	88.41	87.64	78.00
	With Error	74.80	75.20	60.20	72.60	69.70	53.50	77.90	69.30	53.00	73.50	65.20	48.40
SSN	Clean	89.03	89.55	81.08	91.89	92.44	85.94	88.25	89.77	81.44	84.77	85.93	75.34
	With Error	73.90	74.30	59.10	70.80	69.30	53.00	76.10	67.20	50.60	72.40	64.70	47.80
GM	Clean	88.30	89.38	80.79	91.84	92.41	85.89	88.60	89.35	80.75	90.82	89.15	80.42
	With Error	72.10	73.50	58.20	76.50	73.80	58.60	75.30	66.20	49.50	72.60	65.10	48.30
Mixup	Clean	90.56	90.06	81.92	91.29	92.22	85.56	91.36	90.16	82.08	90.35	88.50	79.37
	With Error	76.80	75.20	60.20	77.30	74.20	59.00	77.60	70.60	54.60	76.40	70.30	54.20
ChatIE	Clean	62.92	50.20	33.50	67.75	47.30	31.00	62.40	62.60	45.60	66.70	34.60	20.90
	With Error	59.30	42.30	26.80	63.40	32.10	19.10	64.40	69.80	53.60	69.20	55.70	38.60
SSDAU	Clean	92.00	92.05	85.27	92.80	92.95	86.83	91.74	92.90	86.74	91.58	89.94	81.77
	With Error	83.20	80.70	67.70	88.10	89.80	81.40	80.10	77.90	63.90	84.50	86.40	76.10

for different JERE tasks. The results demonstrate that SSDAU consistently outperforms all baseline in terms of the effectiveness of data augmentation for various JERE tasks. In terms of performance, SSDAU achieves the best F1 scores and generates positive outcomes, unlike the seven baselines that negatively impact JERE models. Regarding adaptability, the results of IoU for augmented data indicate that our method performs better across different JERE models. Notably, our method maintains stable performance even in challenging scenarios. When tested with semantically ambiguous data, SSDAU exhibits minimal performance degradation (less than 2.1% F1 drop on NYT), while baselines like BT suffer significant deterioration (up to 22.5% drop).

In comparison to Back Translation (Xie et al., 2020) and Generative Models (Hou et al., 2021), maintaining the semantic structure of the text proves to be more effective than preserving semantic continuity. Contrasted with Noise Introduction (Fanghua Ye, 2022) and Same-tag Semantic Noise (Yan et al., 2019), the method that maps discrete text by tags exhibits superior performance to adding noise directly. In contrast to Word Substitution (Wei and Zou, 2019) and Mixup (Cheng et al., 2020), labeled discrete texts demonstrate superior properties in JERE data augmentation tasks compared to unlabeled samples. When compared to ChatIE (Wei et al., 2023), SSDAU demonstrates substantially higher precision and F1 scores across all datasets. This highlights the superiority of our structured approach over LLM-based methods for precise extraction tasks. Based on these results, we conclude that SSDAU's approach of preserving structured semantics through contextualized embeddings and topic-aware consistency filtering is superior to existing data augmentation strategies, particularly in scenarios requiring high precision and effective disambiguation capabilities.

We evaluate method robustness using semantically ambiguous data constructed with SciER (Zhang et al., 2024). Table 2 compares SSDAU with seven baselines under normal and error conditions. Baselines show significant performance degradation with ambiguous data, with BT's F1 score dropping by 52.20% on WebNLG (from 89.90% to 37.70%). In contrast, SSDAU maintains stable performance across all datasets, with maximum F1 score reduction of only 1.80% on NYT*. These results confirm our topic-aware consistency filtering mechanism effectively mitigates error propagation, providing superior robustness against semantic ambiguity.

Performance on Different JERE Tasks Table 3 displays the effectiveness of SSDAU and baselines for various JERE models. The results indicate that the SSDAU-augmented dataset exhibits improvements across different types of JERE models,

Table 3: The precision of different models under different datasets. Each cell (A/B) represents the performance of training with the original dataset (A) and the data augmented by SSDAU (B). Values in bold indicate the improvement.

Model	NYT*	WebNLG*	NYT	WebNLG
SPN (Sui et al., 2020)	91.44/**91.95**	93.81/**96.84**	92.67/92.64	90.21/**90.88**
PRGC (Zheng et al., 2021)	93.33/**93.36**	94.00/**94.46**	93.54/**94.40**	89.92/**91.32**
CasRel (Wei et al., 2020)	88.97/**91.47**	91.77/**92.13**	89.10/**91.74**	89.46/**91.58**
OneRel (Shang et al., 2022)	90.17/**92.00**	90.62/**92.80**	92.83/**92.90**	90.66/**91.60**
TPLinker (Wang et al., 2020)	90.23/92.21	90.89/**91.34**	91.33/**92.27**	89.12/**89.93**

such as 3.03% improvement on precision for the $WebNLG_g^*$ dataset in SPN and a 0.94% improvement for the NYT_g dataset in the TPLinker model. These outcomes demonstrate the feasibility of our approach for augmenting unstructured texts into structured semantic data for JERE tasks. Moreover, we observe that SSDAU performs better on partially annotated type datasets than on fully annotated type datasets. Notably, our method achieves about a 3% improvement with the NYT* of the CasRel model and the WebNLG* dataset of the SPN model.

4.3 Ablation Study

We conduct an ablation study to evaluate SSDAU's important components. Throughout this process, we maintain consistent settings across all components.

Data Discretization and Reconstruction We evaluate performance after removing the pre-processing component by directly splitting data based on triad messages without semantic tags, and by applying conventional *no-split* and *full-split* schemes (Gao et al., 2020).

As shown in Table 4, we evaluate the effectiveness of the pre-processing components both before and after removal using precision as a metric. Our results demonstrate that the ***Data Discretization and Reconstruction*** component outperforms the no-pre-processing approach, with an improvement of approximately 2.02%-3.20%. Furthermore, we find that incorporating semantic tagging prompts positively impacts discrete text data augmentation in low-resource JERE tasks.

Structured Semantic Data Augmentation We evaluate the augmentation component by measuring similarity between pre-processed texts using exact matching, generating augmented data by substituting labels in composed discrete texts. The augmented data is classified by triplet type, used for model training, and assessed after component removal.

Table 4: Ablation study for SSDAU. "No Split" denotes not splitting the text. "No Label Split" denotes splitting by semantics without semantic tag. "Full Split" denotes complete splitting of the words in the text.

Dataset	NYT*	WebNLG*	Avg.
CasRel Baseline	90.17	90.62	90.39
SSDAU	92.00	92.80	92.40
Ablation for Pre-processing			
No Split	89.32	90.17	89.75
No Label Split	90.33	90.42	90.38
Full Split	88.64	89.76	89.20
Ablation for Augmentation			
(h,t)	64.21	73.83	69.02
(r)	77.42	84.31	80.87
(h,r,t)	90.41	91.13	90.77
(h,r,h)	85.66	88.53	87.10
(t,r,t)	82.12	84.44	83.28
Ablation for Filtering			
No Filtering	89.92	90.84	90.38

As shown in Table 4, only the third group (h, r, t) of augmented text shows a slight positive effect (0.38%) on JERE tasks, while the other four types negatively impact precision. Removing the augmentation component eliminates threshold restrictions, introducing low-quality data that reduces model precision. The component's absence disrupts text extraction and semantic structure preservation, causing significant performance degradation and highlighting the importance of semantically structured augmentation.

Scoring-based Consistency Filtering We assess the impact of the consistency filtering component in SSDAU. Table 4 shows the precision of the JERE models with and without filtered data. The results demonstrate that the filtered data positively impacts the model's precision, whereas dthe precision de-

Table 5: Semantic consistency verification of augmented text. ν is the syntactic coherence.

Source	Text = South Africa, and the rest of Africa. Triple = [[Africa, /location/location/contains, South Africa]] Structured Semantic = location contain location
Syntax Matching	Text1 = South Africa is a part of Africa. $\nu = 0.516$ Text2 = North Africa, and the rest of Africa. $\nu = 0.923$ Triple = [[Africa, /location/location/contains, North Africa]] Structured Semantic = location contain location

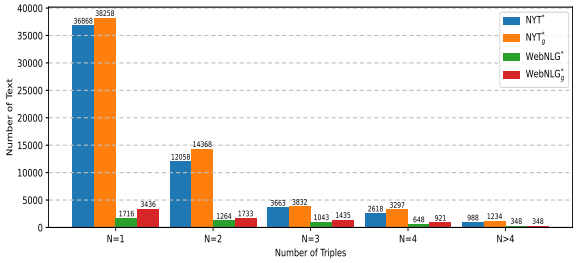
creases when low-quality augmented data are not removed. This highlights the importance of consistency filtering in maintaining the model's precision.

Parameter Initialization We investigate the impact of parameter initialization in our model by comparing three initialization methods: random initialization, zero initialization, and pretrained initialization (used in SSDAU). As shown in Table 6, the pretrained initialization method consistently outperforms others across all four datasets. Significance tests (t-test) confirm these improvements are statistically significant (p=0.012 on NYT, p=0.009 on WebNLG, p=0.016 on NYT*, p=0.008 on WebNLG*). These results validate our design choice of using HuggingFace's default pretrained parameters for the BERTTopic model.

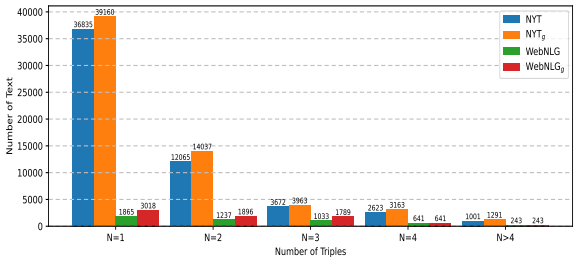
4.4 Analysis

Semantic coherence analysis. During the semantic coherence analysis of SSDAU, we follow a two-step process to ensure semantic consistency in the augmented text. First, we augment all texts by considering similarities between annotations of the same type and entity text, while preserving the semantic annotations (e.g., "location contains location"). Next, we use Biber Tagger (A. Bergman, 2022) to match triplet texts with identical tags. The high degree of syntactic agreement between $Text1$ and $Text2$ is demonstrated in Table 5. We filter out texts with low relevance (below 0.8) and incorporate the remaining data into the training set as augmented data, ensuring the semantic consistency of the augmented text.

Training Cost and Convergence. Figure 3 provides details about the original and augmented texts containing varying numbers of triplets. We focus specifically on scenarios where an entity appears in multiple triplet relations and categorize the texts based on the number of triplets to evaluate the effectiveness of SSDAU for such texts. By classifying the augmented data according to triplet counts and



(a) The Partially Data



(b) The Exactly Data

Figure 3: The comparison between the number of text after augmentation with SSDAU and the initial one for different types of datasets.

incorporating it into the training set, we assess the performance of different JERE models using the same test set. The results demonstrate the effectiveness of SSDAU for texts with different triplet counts. Our method proves valuable across texts with varying numbers of triplets, showing that as the number of triplets in the training set decreases, the availability of augmented data increases, leading to improved model precision.

4.5 Case Study

Table 7 presents three cases of SSDAU applied to JERE tasks. In the first case, we replace the head entity "$Mitch\ Mustain$" with "$Amy\ Grant$" while preserving the semantic label and other text intact. In the second case, we substitute the tail entity "$Arkansas$" with "$Nashville$" while maintaining the original semantic labels and other texts. In the third case, we modify all the text except for the entity and change the se-

Table 6: Ablation study on parameter initialization across four datasets.

Init Type	Partial Match						Exact Match					
	NYT*			WebNLG*			NYT			WebNLG		
	Prec.	F1	IoU	Prec.	F1	IoU	Prec.	F1	IoU	Prec.	F1	IoU
Pretrained	90.17	91.45	84.24	90.62	90.25	82.23	92.83	92.17	85.47	90.66	89.08	80.31
Random	73.92	52.30	35.41	76.54	65.72	48.97	67.35	48.12	31.69	90.26	62.45	45.36
Zero	78.42	70.83	54.85	83.78	54.97	37.90	78.22	70.43	54.42	87.65	51.78	34.94

Table 7: Augmented data generated by SSDAU. Black texts are the original examples. Red texts are the discrete text. Blue texts are the precondition for text segmentation and augmentation. ε_1 is the entity similarity threshold and ε_2 is the relation similarity threshold.

Source	Text: At Arkansas , the freshman Mitch Mustain led the Razorbacks in a 24-23 double-overtime upset of Alabama. Triples: Mitch Mustain(people)|Arkansas(place)|place_lived Razorbacks(group)|Mitch Mustain(people)|contain
Head → Head	Condition: $Tag_h = people, Tag_t = place, Tag_r = place\_lived, \Theta_h \geq \varepsilon_1$. Text: At Arkansas, the freshman Amy Grant led the Razorbacks in a 24-23 double-overtime upset of Alabama. Triples: Amy Grant(people)|Arkansas(place)|place_lived Razorbacks(group)|Amy Grant(people)|contain
Tail → Tail	Condition: $Tag_h = people, Tag_t = place, Tag_r = place\_lived, \Theta_t \geq \varepsilon_1$. Text: At Nashville, the freshman Mitch Mustain led the Razorbacks in a 24-23 double-overtime upset of Alabama. Triples: Mitch Mustain(people)|Nashville(place)|place_lived Razorbacks(group)|Mitch Mustain(people)|contain
Relation → Relation	Condition: $Tag_h = people, Tag_t = place, \Theta_r \geq \varepsilon_2$. Text: At Arkansas, the freshman Mitch Mustain led the Razorbacks in a 24-23 double-overtime upset of Alabama. Triples: Mitch Mustain(people)|Arkansas(place)|location Razorbacks(group)|Mitch Mustain(people)|contain

mantic label from "$people|people|place_lived$" to "$people|people|location$." Our data augmentation approach can expand texts without introducing additional noise, resulting in natural and diverse augmentations. Compared to existing methods, SSDAU's augmented data resolves diversity and quality issues more effectively.

5 Conclusion

We propose SSDAU, a data augmentation paradigm designed to perform instance augmentation for low-resource JERE tasks by labeling the semantic segmentation of entity texts and assessing similarity within neighboring semantic regions. Our approach integrates contextualized embeddings with traditional similarity scores to effectively distinguish semantically similar but distinct entities, while employing topic-aware consistency filtering with pre-trained initialization to mitigate error propagation. Compared to traditional methods, SSDAU effectively addresses the challenge of data scarcity in low-resource scenarios and mitigates issues such as reduced textual relevance and overlapping relations. These findings suggest that preserving the semantic structure of texts through structured semantic tags can be a promising approach for text data augmentation.

Limitations

Although the proposed SSDAU outperforms all baseline methods, it still has some limitations. Firstly, while we alleviate the need for high-quality data in SSDAU by filtering low-quality data, incorporating more high-quality data may further improve SSDAU's performance. Secondly, we improve *Similarities* for structured semantic matching of long texts through pre-processing The efficiency of our approach can be enhanced by utilizing a more efficient semantic text-matching component. In future work, it would be interesting to validate our approach in real-time using newly acquired high-quality data and explore the development of semantic text matching components that deliver superior results for long texts.

References

Mona Diab A. Bergman. 2022. Towards responsible natural language annotation for the varieties of arabic. In *Findings of the Association for Computational Linguistics: ACL 2022*, page 364–371, Dublin, Ireland. Association for Computational Linguistics.

Ibrahim Abdelaziz, Srinivas Ravishankar, Pavan Kapanipathi, Salim Roukos, and Alexander Gray. 2021. A semantic parsing and reasoning-based approach to knowledge base question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 15985–15987.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Dylan Cashman, Shenyu Xu, Subhajit Das, Florian Heimerl, Cong Liu, Shah Rukh Humayoun, Michael Gleicher, Alex Endert, and Remco Chang. 2020. Cava: A visual analytics system for exploratory columnar data augmentation using knowledge graphs. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):1731–1741.

Yong Cheng, Lu Jiang, Wolfgang Macherey, and Jacob Eisenstein. 2020. Advaug: Robust adversarial augmentation for neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5961–5970.

Emine Yilmaz Fanghua Ye, Yue Feng. 2022. Assist: Towards label noise-robust dialogue state tracking. In *Findings of the Association for Computational Linguistics: ACL 2022*, page 2719–2731, Dublin, Ireland. Association for Computational Linguistics.

Pan Gao, Qi Wan, and Linlin Shen. 2020. Split and merge: Component based segmentation network for text detection. In *International Conference on Pattern Recognition and Artificial Intelligence*, pages 14–27. Springer.

Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017. Creating training corpora for nlg micro-planning. In *55th Annual Meeting of the Association for Computational Linguistics, ACL 2017*, pages 179–188. Association for Computational Linguistics (ACL).

Issam Laradji Gaurav Sahu, Pau Rodriguez. 2022. Data augmentation for intent classification with off-the-shelf large language models. In *Proceedings of the 4th Workshop on NLP for Conversational AI*, pages 47–57, Dublin, Ireland. Association for Computational Linguistics.

Yutai Hou, Sanyuan Chen, Wanxiang Che, Cheng Chen, and Ting Liu. 2021. C2c-genda: Cluster-to-cluster generation for data augmentation of slot filling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13027–13035.

Yutai Hou, Yijia Liu, Wanxiang Che, and Ting Liu. 2018. Sequence-to-sequence data augmentation for dialogue language understanding. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1234–1245.

Zhiting Hu, Bowen Tan, Ruslan Salakhutdinov, Tom Mitchell, and Eric P Xing. 2019. Learning data manipulation for augmentation and weighting. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 15764–15775.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2020. Tinybert: Distilling bert for natural language understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, page 4163–4174, Online. Association for Computational Lingustics.

Nishant Kambhatla, Logan Born, and Anoop Sarkar. 2022. Cipherdaug: Ciphertext based data augmentation for neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, page 201–218, Dublin, Ireland. Association for Computational Linguistics.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Ying Lin, Heng Ji, Fei Huang, and Lingfei Wu. 2020. A joint neural model for information extraction with global features. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, page 7999–8009, Online. Association for Computational Linguistics.

Sisi Liu, Kyungmi Lee, and Ickjai Lee. 2020. Document-level multi-topic sentiment classification of email data with bilstm and data augmentation. *Knowledge-Based Systems*, 197:105918.

Junghyun Min, R Thomas McCoy, Dipanjan Das, Emily Pitler, and Tal Linzen. 2020. Syntactic data augmentation increases robustness to inference heuristics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, page 2339–2352, Online. Association for Computational Linguistics.

Jonas Mueller and Aditya Thyagarajan. 2016. Siamese recurrent architectures for learning sentence similarity. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, pages 2786–2792.

Evan Sandhaus. 2008. The new york times annotated corpus. *Linguistic Data Consortium, Philadelphia*, 6(12):e26752.

Yu-Ming Shang, Heyan Huang, and Xianling Mao. 2022. Onerel: Joint entity and relation extraction with one module in one step. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11285–11293.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15:1929–1958.

Dianbo Sui, Yubo Chen, Kang Liu, Jun Zhao, Xiangrong Zeng, and Shengping Liu. 2020. Joint entity and relation extraction with set prediction networks. *arXiv preprint arXiv:2011.01675*.

William Yang Wang and Diyi Yang. 2015. That's so annoying!!!: A lexical and frame-semantic embedding based data augmentation approach to automatic categorization of annoying behaviors using# petpeeve tweets. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 2557–2563.

Yucheng Wang, Bowen Yu, Yueyang Zhang, Tingwen Liu, Hongsong Zhu, and Limin Sun. 2020. Tplinker: Single-stage joint extraction of entities and relations through token pair linking. In *Proceedings of the 28th International Conference on Computational Linguistics)*, page 1572–1582, Barcelona, Spain (Online). Association for Computational Linguistics.

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)*, page 6382–6388, Hong Kong, China. Association for Computational Linguistics.

Xiang Wei, Xingyu Cui, Ning Cheng, Xiaobin Wang, Xin Zhang, Shen Huang, Pengjun Xie, Jinan Xu, Yufeng Chen, Meishan Zhang, et al. 2023. Chatie: Zero-shot information extraction via chatting with chatgpt. *arXiv preprint arXiv:2302.10205*.

Zhepei Wei, Jianlin Su, Yue Wang, Yuan Tian, and Yi Chang. 2020. A novel cascade binary tagging framework for relational triple extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, page 1476–1488, Online. Association for Computational Linguistics.

Xinyi Wu, Zhenyao Wu, Yuhang Lu, Lili Ju, and Song Wang. 2022. Style mixing and patchwise prototypical matching for one-shot unsupervised domain adaptive semantic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2740–2749.

Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V Le. 2020. Unsupervised data augmentation for consistency training. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, pages 6256–6268.

Ge Yan, Yu Li, Shu Zhang, and Zhenyu Chen. 2019. Data augmentation for deep learning of judgment documents. In *International Conference on Intelligent Science and Big Data Engineering*, pages 232–242. Springer.

Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. 2017. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*.

Qi Zhang, Zhijia Chen, Huitong Pan, Cornelia Caragea, Longin Jan Latecki, and Eduard Dragut. 2024. Scier: An entity and relation extraction dataset for datasets, methods, and tasks in scientific documents. *arXiv preprint arXiv:2410.21155*.

Yi Zhang, Tao Ge, and Xu Sun. 2020. Parallel data augmentation for formality style transfer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, page 3221–3228, Online. Association for Computational Linguistics.

Nikolay Arefyev Zhang Bingyu. 2022. The document vectors using cosine similarity revisited. In *Proceedings of the Third Workshop on Insights from Negative Results in NLP*, page 129–133, Dublin, Ireland. Association for Computational Linguistics.

Hengyi Zheng, Rui Wen, Xi Chen, Yifan Yang, Yunyan Zhang, Ziheng Zhang, Ningyu Zhang, Bin Qin, Xu Ming, and Yefeng Zheng. 2021. Prgc: Potential relation and global correspondence based joint relational triple extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, page 6225–6235, Online. Association for Computational Linguistics.

Ming Zhong, Pengfei Liu, Yiran Chen, Danqing Wang, Xipeng Qiu, and Xuan-Jing Huang. 2020. Extractive summarization as text matching. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, page 6197–6208, Online. Association for Computational Linguistics.