Syntactically Robust Training on Partially-Observed Data for Open Information Extraction

Ji Qi[1], **Yuxiang Chen**[2], **Lei Hou**[1], **Juanzi Li**[1], **Bin Xu**[1*]
[1]Department of Computer Science and Technology, BNRist,
Tsinghua University, Beijing, 100084, China
[2]University of California, San Diego
qj20@mails.tsinghua.edu.cn, yuc129@ucsd.edu

Abstract

Open Information Extraction models have shown promising results with sufficient supervision. However, these models face a fundamental challenge that the syntactic distribution of training data is partially observable in comparison to the real world. In this paper, we propose a syntactically robust training framework that enables models to be trained on a syntactic-abundant distribution based on diverse paraphrase generation. To tackle the intrinsic problem of knowledge deformation of paraphrasing, two algorithms based on semantic similarity matching and syntactic tree walking are used to restore the expressionally transformed knowledge. The training framework can be generally applied to other syntactic partial observable domains. Based on the proposed framework, we build a new evaluation set called CaRB-AutoPara, a syntactically diverse dataset consistent with the real-world setting for validating the robustness of the models. Experiments including a thorough analysis show that the performance of the model degrades with the increase of the difference in syntactic distribution, while our framework gives a robust boundary. The source code is publicly available at https://github.com/qijimrc/RobustOIE.

1 Introduction

Open Information Extraction (OpenIE) involves converting natural text to a set of n-ary structured tuples of the form (arg_1, predicate, arg_2, ..., arg_n), composed of a single predicate as well n arguments. With the advantages of domain independence and scalability, OpenIE serves as a backbone in natural language understanding and fosters many applications such as text summarization (Fan et al., 2019) and question answering (Yan et al., 2018).

Tremendous efforts have been devoted to build models that can better fit the extractions from texts (Michele et al., 2007; Angeli et al., 2015; Saha and Mausam, 2018; Kolluru et al., 2020b; Yu et al., 2021). However, a major issue remaining in OpenIE is the syntactic partial observability – the syntactic distribution on the existing training set is only based on partial observations, and it is far from covering the entire syntactic hypothesis space in the real world. This issue creates a challenge that the models rely heavily on the syntactic forms during training, and degrade significantly when the syntactic distribution changes in the real world.

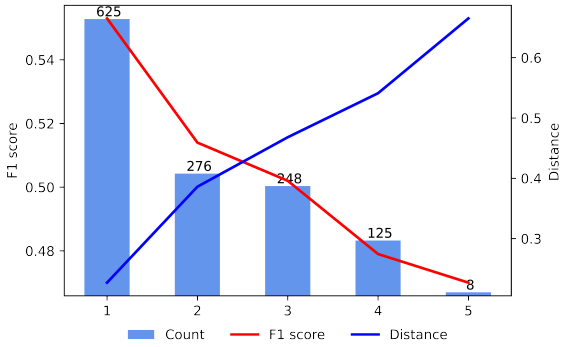


Figure 1: Cluster CaRB into 5 subsets based on the HW-Syntactic Distance and evaluate the IMOJIE model on them. The horizontal axis indicates the indices sorted by the number of samples (above the bars) in the subsets. The left and right vertical axes represent the F1 scores of the model and the distance between the training set and the clustering center of each subset, respectively.

An evaluation is shown in Figure 1. We cluster the CaRB (Bhardwaj et al., 2019) samples based on the *HW-Syntactic Distance* (introduced in Sec. 3.5), which is an effective metric that measures the syntactic difference between two sentences, and evaluate the state-of-the-art model trained on the OpenIE4 data (Kolluru et al., 2020b) on them. A frustrating result shows that the model performance exhibits a significant degradation as the syntactic similarity between the training set and clustering centers of subsets decreases. The biased performance comes from the inconsistency of the syn-

*Corresponding author: xubin@tsinghua.edu.cn

Findings of the Association for Computational Linguistics: EMNLP 2022, pages 6245–6257
December 7-11, 2022

tactic distributions among data. For example, in Figure 1, the model achieves a depressing F1 score of 0.47 on the subset 5 with the lowest average syntactic similarity to the training set. Therefore, to build robust OpenIE systems, we need to train the models on a sufficient syntactic distribution.

However, it is not trivial to obtain data that are both diverse and accurate to satisfy the distribution assumption. First, it is extremely expensive and almost impossible for human annotators to provide a large corpus with diverse syntactic expressions. Second, existing distant supervision-based methods are not applicable to OpenIE due to the uncertainties of both the type and form of arguments and predicates.

Humans learn syntactic grammar by paraphrasing the same meaning into different expressions. For example, the following two sentences convey the same meaning in different syntactic forms. The diverse paraphrases of normal-scale training data can guarantee sufficient syntactic distribution. However, an intrinsic problem that hinders the efficiency of this approach is the **Knowledge Deformation**. In the following example, it is difficult to reveal the source object *Earth* in the target paraphrase b as it has been transformed into the form of *the name of the planet* with different syntax.

- *a. After five years of searching, the Colonials found a new world and named it Earth.*
- *b. The colonials searched for five years until they discovered a new world and gave him the name of the planet.*

In this paper, we propose a syntactically robust training framework that enables OpenIE models to be trained on a syntactic-abundant distribution based on the diverse paraphrase generation. Specifically, we first generate a large-scale syntactically diverse paraphrase candidates set for the training data based on an off-the-shelf paraphrase generator. Then, we propose two adaptive algorithms to recover the deformed arguments of the original knowledge, a semantic similarity-based matching method to locate the disordered arguments and a syntactic tree walking-based method to complete the consecutive spans. We further employ the generative T5 (Raffel et al., 2020) model to restore the deformed predicates as there are potential tense and voice changes in the target paraphrase. Finally, a simple but effective denoising method is utilized to prevent the impact of false positives in training.

To exhaustively validate the syntactic robustness of OpenIE models in the real-world setting, an additional evaluation set including diverse paraphrases and knowledge triples has been built on the basis of CaRB. We conduct experiments on the standard and our proposed evaluation sets based on the division of different syntactic categories, and a comprehensive analysis shows that the model performance decreases with increasing the difference in the syntactic distributions, while our training framework gives a robust boundary.

2 Syntactically Robust Training Framework for OpenIE

2.1 Overview

The task of OpenIE aims to build a model p_θ to automatically extract a set of n-ary tuples $\{r_i = (a_1, p_r, a_2, a_3, ..., a_n)\}_{i=1}^{m}$ for each sentence, where p_r indicates the predicate, a_1, a_2 indicate the subject and object, and $a_3, ..., a_n$ refer to the other argments such as time and location. Given a training set $\mathcal{D} = (s_1, s_2, ..., s_{|\mathcal{D}|})$ consisting of sentences samples, where each sentence exhibits a syntactic structure e^s. Our goal is to maximize the expectation of log-likelihood function $\log \mathrm{p}_\theta(r_1, ..., r_m, e^s|s)$ with respect to the data distribution $\mathrm{p}_\mathcal{D}$ as following:

$$\begin{aligned}\mathcal{L}(\theta) =& \mathbb{E}_{r_i, e^s \sim \mathrm{p}_\mathcal{D}}[\log \mathrm{p}_\theta(r_1, ..., r_m, e^s|s)] \\ =& \mathbb{E}_{r_i, e^s \sim \mathrm{p}_\mathcal{D}}[\log \mathrm{p}_\theta(r_1, ..., r_m|e^s, s)\mathrm{p}_\theta(e^s|s)]\end{aligned}$$

where different OpenIE models may adopt a distinct strategy to model the probability p_θ, such as the triples generating paradigm (Kolluru et al., 2020a) or sequence labeling paradigm (Zhan and Zhao, 2020), and the maximization process is performed by gradient ascent.

The syntactic distribution in training set $e^s \sim p_\mathcal{D}$ is far from covering the entire syntactic hypothesis space, and plays a fatal role in OpenIE modeling. In this research, we aim to expand the training with a sufficient syntactic distribution. The proposed framework is illustrated in Figure 2. We first generate a syntactically diverse paraphrase candidate set for the training data with an off-the-shelf paraphrase generation model. Then, we restore the deformed arguments using semantic similarity-based matching and syntactic tree walking algorithms, followed by a T5-based predicate restoration. Finally, a denoised training is adopted to optimize the model on the sufficient distribution.

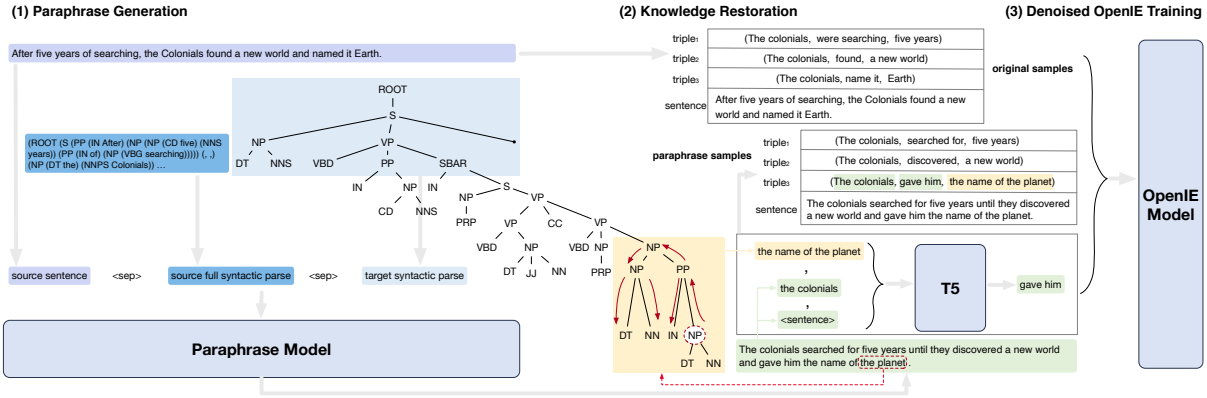


Figure 2: Overview of the proposed framework. Based on the diverse paraphrase candidates set generated by a syntactically controllable model, two algorithms, semantic similarity-based arguments localization and syntactic tree walking, are used to restore the deformed arguments. By taking the arguments as conditions, the predicates are generated with the T5 model.

2.2 Paraphrase Generation

To create syntactically diverse paraphrases candidates set on $\mathcal{D}$, we adopt AESOP (Sun et al., 2021), a syntactically controllable paraphrase generation model as our generator. As can be seen in Figure 2, by utilizing the BART (Lewis et al., 2020) model as a backbone, the model takes *source sentence*<sep>*source full syntactic parse*<sep>*target syntactic parse* as the input sequence, and outputs a sequence of the form *target syntactic parse*<sep>*paraphrase* in which the generated paraphrase conforms with the pruned target syntax.

The AESOP model used in our work is trained on a parallel annotated data with a two-level target syntactic tree. During generation given the training set $\mathcal{D}$, we first get their constituency parse trees[1] $\{T^{\mathcal{D}}_{s_1}, ..., T^{\mathcal{D}}_{s_{|\mathcal{D}|}}\}$ and linearize them into parentheses trees as the source full syntactic parses (A part is shown in Figure 2). Then, we collect a set of constituency parse pairs pruned at height 3 $\{(T^{\mathcal{P}}_{s_1}, T^{\mathcal{P}}_{t_1}), ..., (T^{\mathcal{P}}_{s_{|\mathcal{P}|}}, T^{\mathcal{P}}_{t_{|\mathcal{P}|}})\}$ from the ParaNMT-50M (Wieting and Gimpel, 2018) and count their frequencies. For each sentence in $\mathcal{D}$, following the original work we obtain m most similar parses $\{T^{\prime\mathcal{P}}_{s_1}, ..., T^{\prime\mathcal{P}}_{s_m}\}$ by calculating weighted ROUGE scores between parse strings, and select k top-ranked parses from $\{T^{\mathcal{P}}_{t_1}, ..., T^{\mathcal{P}}_{t|\mathcal{P}|}\}$ for each $T^{\prime\mathcal{P}}_{s_i}$ by a sampling with the distribution of:

$$T^{\mathcal{P}}_t \sim p(T^{\mathcal{P}}_t | T^{\prime\mathcal{P}}_{s_i}) = \frac{\#(T^{\prime\mathcal{P}}_{s_i}, T^{\mathcal{P}}_t)}{\sum_j \#(T^{\prime\mathcal{P}}_{s_i}, T^{\mathcal{P}}_{t_j})} \quad (1)$$

where $\#(T^{\prime\mathcal{P}}_{s_i}, T^{\mathcal{P}}_t)$ refers to the count of occurrence in the statistic data. In the end, we generate k paraphrases for each sentence in $\mathcal{D}$. For a tradeoff of quality and quantity, we set k and m to 5 and 2, respectively. As a result, we get the paraphrases candidates set $\mathcal{P}$, which is roughly five times the size of sentences in training set $\mathcal{D}$.

[1] We use Stanford CoreNLP (Manning et al., 2014).

2.3 Knowledge Restoration

As the paraphrases change the expression form of the original sentence, we need to recover the knowledge of transformed triples. The difficulty of knowledge restoration lies in two aspects: first, the OpenIE arguments are generally formed as a large span of words, which can be rearranged and rephrased in the target sentence. Second, the syntactic changes also lead to a transformation of tense or voice of verbs in the predicates. For example in Figure 2, the argument *the Earth* changes its expression and length to become *the name of the planet*, and the predicate *were searching* changes its tense to become *searched for*.

Therefore, we first locate the arguments with the contextualized semantic matching and complete it with syntactic tree walking. Then for each pair of recovered arguments, we restore the corresponding predicate with the T5 model (Raffel et al., 2020).

2.3.1 Argument Restoration

As the expressional transformations, it is difficult to get the corresponding arguments in the target paraphrase sentence based on methods like pattern matching. Therefore, we utilize the semantic similarity with BERT (Devlin et al., 2019) to locate the arguments. We first compute the embeddings $\mathbf{h}^s \in \mathbb{R}^{|s| \times d}$ and $\mathbf{h}^t \in \mathbb{R}^{|t| \times d}$ for the

source sentence s and target paraphrase sentence t, respectively. Then, for a triple (a_1^s, p_r^s, a_2^s) where $a_i^s \rightarrow (l_i^s, r_i^s), p_r^s \rightarrow (l_p^s, r_p^s)$ [2] in the source sentence, we calculate the semantic similarity scores $\mathbf{c}^{a_i}, \mathbf{c}^r \in \mathbb{R}^{|t|}$ by summing the cosine similarities between each word in a_i^s, p_r^s and target words of t:

$$\mathbf{c}^{a_i} = \sum_{j=l_i^s}^{r_i^s} \cos(\mathbf{h}_j^s, \mathbf{h}^t), \mathbf{c}^r = \sum_{j=l_p^s}^{r_p^s} \cos(\mathbf{h}_j^s, \mathbf{h}^t) \quad (2)$$

Next, we merge the consecutive indices of target words whose semantic similarity scores are greater than a threshold τ to get the resulting candidate spans $\{(l_{i1}^t, r_{i1}^t), ..., (l_{im}^t, r_{im}^t)\}$ and $\{(l_{p1}^t, r_{p1}^t), ..., (l_{pm}^t, r_{pm}^t)\}$ for a_i^s and p_r^s, and the final triplets are obtained by selecting a set of spans with the highest total score and no overlap. By applying this algorithm on $\mathcal{P}$, we get dataset $\mathcal{D}^{\mathcal{P}}$. We refer to the set expanded with this newly built set as $\mathcal{D}^{\Phi} = \mathcal{D} \cup \mathcal{D}^{\mathcal{P}}$.

Though the resulting spans based on semantic similarity matching are accurate in position, we find it incomplete due to the fact that words such as prepositions or adverbs can not be matched effectively by the contextualized embedding. On the other hand, a subtree with NP, QP or NX as the root in the constituency parses represents a continuous phrase fragment. Therefore, we propose to use the syntactic tree walking to further complete the target arguments. Specifically, for each word in span (l_{ij}^t, r_{ij}^t), we perform a post-order traversal for the target syntactic tree to effectively find the subtree with NP, QP or NX as the root and containing the the word as a node. We obtain the refined span $(l_{ij}^{t'}, r_{ij}^{t'})$ by replacing the original span (if it covers the original span, otherwise the original span is retained) with the corresponding words of the subtree. Finally, we select the optimal target spans $\{(l_1^{t*}, r_1^{t*}), ..., (l_n^{t*}, r_n^{t*})\}$ of all arguments from the refined spans set of each argument by a simple optimality criterion that maintains n spans with the highest similarity without overlaps. We retain the argument restoration as Algorithm 1 in detailed.

2.3.2 Predicate Restoration

As the paraphrase may change the voice and tense of the predicate in the original sentence, it is not applicable to recover the predicate using the same algorithm as the arguments restoration. We adopt the T5 model (Raffel et al., 2020) to restore the predicate in the target paraphrase sentence, as there are a lot of predicates that can not be found from the continuous span of the original sentence. Specifically, we build a new dataset on $\mathcal{D}$ with the same corpus size. For each data sample in the new dataset, the input is of the form of *source sentence, argument*$_1$, *argument*$_2$ <\s>, and the output is a generated sequence referring to the predicate. We train the basic T5 model on the new dataset. Then, we restore the predicate for each pair of arguments obtained from the algorithm 1 to get a final refined set $\mathcal{D}^{\mathcal{P}'}$. We refer to the refined final expanded set as $\mathcal{D}^{\Psi} = \mathcal{D} \cup \mathcal{D}^{\mathcal{P}'}$.

[2] For convenient, we use l_i^s and r_i^s to denote the indices of start word and end word of argument a_i in the sentence s.

Algorithm 1 Arguments Restoration

Input: Source/target sentence embeddings $\mathbf{h}^s/\mathbf{h}^t$, source tuple $(a_1^s, p_r^s, ..., a_n^s), a_i^s \rightarrow (l_i^s, r_i^s)$
Output: target n-tuple $(a_1^t, a_2^t, ..., a_n^t)$

get target constituency parse tree T^t
subtree roots $\mathcal{T} = \{NP, QP, NX\}$
threshold $\tau = 0.7$
for each argument $a_i^s \in (a_1^s, ..., a_n^s)$ **do**
    calculate scores $\mathbf{c}^{a_i} = \sum_{j=l_i^s}^{r_i^s} \cos(\mathbf{h}_j^s, \mathbf{h}^t)$
    get candidate spans $csp_i = \{(l_{i1}^t, r_{i1}^t), ...\}$ by merging the consecutive indices with values greater than τ in $\mathbf{c}^{a_i}$
    for $sp_{ij} = (l_{ij}^t, r_{ij}^t) \in csp_i$ **do**
        for $tok_k \in sp_{ij}$ **do**
            $traverse(T^t)$ to find subtree T_k^t that satisfies: $T_k^t.root \in \mathcal{T}$ && $tok_j \in T_k^t$
            $T_j^t \leftarrow T_j^t + T_k^t$
        end for
        $sp'_{ij} = (l_{ij}^{t'}, r_{ij}^{t'}) \Leftarrow T_j^t$
    end for
    $csp'_i = \{sp'_{i1}, sp'_{i2}, ...\}$
end for
return $\{sp_i^* | sp_i^* \in csp'_i, i = 1, ..., n\}$ with highest score without overlaps

2.4 Denoised Training

During the training, we aim to maximize the expectation of log-likelihood function with respect to the data distribution:

$$\mathcal{L}(\theta) = \mathbb{E}_{(r_1,...,r_m)\sim p_d}[\log p_\theta(r_1, ..., r_m | s)] \quad (3)$$

where p_d refers to a training set, and p_θ is a neural network model with learnable parameters θ, which either employs the sequence labeling

paradigm to predict classification labels on the input sequence, or leverages the generative paradigm to generate target triples each token at a time. In this paper, we validate our proposed training framework on IMOJIE (Kolluru et al., 2020b), a strong generative model that predicts triples conditioned on the previous generation.

As the rephrasing in large argument spans may introduce false-positive word noises, we employ a simple but effective masking strategy to ignore the impact of negative words while retaining the contribution of valuable correct words in the span. For a triple (a_1, p_r, a_2), we calculate the importance of each word in an argument a_i based on its semantic matching score obtained from the arguments restoration algorithm. For those words which are recovered from the syntactic tree, we set them to the average value of other words. We finally normalize the reciprocals of these importance scores and randomly select 15% of all words according to the probabilities distribution. These sampled words will be masked to not calculate their gradients in training. Note that we only mask the words in arguments as the predicate is short and less noisy.

3 Experiment

This work proposes a syntactically robust training framework including two knowledge restoration strategies. Therefore, our experiments are intended to demonstrate the effectiveness as well as the robustness of the proposed framework on test sets.

3.1 Datasets

We use the standard training set OpenIE4 (Kolluru et al., 2020b), and the constructed sets $\mathcal{D}^{\Phi}, \mathcal{D}^{\Psi}$ for model training. During evaluation, in addition to the benchmark dataset CaRB (Bhardwaj et al., 2019), we build a syntactically diverse evaluation set to validate the robustness of OpenIE model.

3.1.1 Training set

Data	# samples	Fact-level accuracy	Span-level accuracy
$\mathcal{D}$	215,356	/	/
$\mathcal{D}^{\Phi}$	429,171	87%	34%
$\mathcal{D}^{\Psi}$	382,752	91%	71%

Table 1: Train set statistics and the human verification results. We randomly sample 100 samples for each dataset and evaluate two fine-grained metrics.

We use the dataset OpenIE4 as the basic set $\mathcal{D}$ in our experiment, which is published by (Kolluru et al., 2020b) and prep-processed by (Kolluru et al., 2020a). The data is automatically built by running OpenIE-4, ClausIE, and RnnOIE on the sentences obtained from Wikipedia.

To estimate the quality of the generated samples of $\mathcal{D}^{\Phi}$ and $\mathcal{D}^{\Psi}$, we conduct fine-grained human verification by randomly samplling 100 data samples from each set. For a fair comparison, taking the triples from the human-annotated dataset CaRB as the reference criteria, we evaluate the generated samples on fact-level and span-level, respectively. Specifically, a triple is fact-level correct if all elements in the triple conform with the definition of arguments or predicate. A triple is span-level correct only if all arguments and predicates contain the complete words span in the sample sentence. The overall statistics are shown in Table 1. We can see that though the fact-level accuracy shows the useable for $\mathcal{D}^{\Phi}$, the spans of arguments and predicate are extremely inaccurate with the accuracy of 34%. By further performing the algorithms of syntactic tree walking-based arguments restoration and predicate restoration, we improve both the fact-level and span-level accuracy to 91% and 71%, suggesting the satisfaction of the generated data.

3.1.2 Evaluation set

Data	# sent.	arg$_{.len}$	pre$_{.len}$
CaRB	1282	14.9	2.7
CaRB-AutoPara	2269	17.3	2.3

Table 2: Evaluation set statistics. The # sent. refers to the total number of sentences, and arg.$_{len}$/pre.$_{len}$ are the average lengths of argument/predicate of all samples in corresponding data, respectively.

We use the standard benchmark CaRB (Bhardwaj et al., 2019) to evaluate the proposed framework, which is a high-quality crowdsourced dataset with 1282 sentences and each sentence has manually annotated about 4 n-tuples.

In order to evaluate the syntactic robustness of OpenIE models, we build a syntactically diverse dataset based on CaRB with the proposed framework. We generate 5 paraphrases for each sentence from CaRB, and get 2269 high-quality sentences after performing the knowledge restoration. We refer to this automatically generated dataset as CaRB-AutoPara. The statistics of both datasets are shown in Table 2. We can see that the newly built dataset is

twice as large in scale and the lengths of arguments and predicates conform with the CaRB.

3.2 Evaluation Metrics

We use the scoring system proposed by (Bhardwaj et al., 2019) to evaluate the OpenIE models on two test sets. The system first creates an all-pair matching table, with each column as a prediction tuple and each row as a gold tuple. It then computes single-match precision and multi-match recall by considering the number of common tokens in (gold, perdition) pair for each element of the fact.

Based on the confidence with each output triple, we report three important metrics: (1) Optimal F1: the largest F1 value in the P-R curve, (2) AUC: the area under the P-R curve, and (3) Last F1: the F1 score computed at the point of zero confidence.

3.3 Experimental Settings

We follow the original work to train a BART-based paraphrase model (Sun et al., 2021) on ParaNMT-small (Chen et al., 2019), and the syntactic mapping set is collected from (Wieting and Gimpel, 2018). For knowledge restoration, we use the pre-trained BERT (Devlin et al., 2019) model to calculate the cosine similarity, and fine-tune the T5 model (Raffel et al., 2020) with a language model head on it for the predicate restoration. The threshold τ and maintaining number of spans k are empirically set to 0.7 and 5, respectively.

We train two implementations of our proposed framework based on the baseline model IMOJIE (Kolluru et al., 2020b) to investigate the effectiveness and syntactically robustness. IMOJIE$^{\Phi}$ is trained on $\mathcal{D}^{\Phi}$ that adopts the semantic similarity matching as the knowledge restoration method only. IMOJIE$^{\Psi}$ is trained on $\mathcal{D}^{\Psi}$ that uses the entire knowledge restoration algorithms. All models followed the original implementations by using BERT as encoder and LSTM with the CopyAttention mechanism (Cui et al., 2018a) as the decoder. The detained parameters setting are shown in Appendix A.

3.4 Results on Different Datasets

How does the proposed framework perform on the syntactic identically distributed data?

In comparison with the baseline model, we find that the proposed syntactically robust training framework generally enhances the OpenIE model to achieve better performance on identically distributed data. As shown in Table 3, we compare three models on the evaluation set CaRB, a minor scale dataset including 1282 human-annotated sentences. We can see that with the simple contextual similarity-based knowledge restoration, our model IMOJIE$^{\Phi}$ achieves better performance than the basic model on F1 and optimal F1 metrics. By training model with the entire knowledge restoration algorithms, the model IMOJIE$^{\Psi}$ outperforms the basic model by 1.4 F1 pts, 0.7 pts of AUC, and 1.5 pts of optimal F1. The results suggest that the OpenIE model is syntactic sensitive and can benefit from more syntactically sufficient training.

Model	CaRB		
	F1	AUC	Opt.F1
IMoJIE	53.3	33.3	53.5
IMoJIE$^{\Phi}$	53.6	32.4	54.0
IMoJIE$^{\Psi}$	**54.7**	**34.0**	**55.0**

Table 3: Experimental results on CaRB.

We argue that the CaRB data is the **syntactic identically distributed evaluation set** with the training set OpenIE4, as they are both sampled from a specific domain of Wikipedia, making them hold similar writing styles. For example, one sentence describes the fact of "sb. won sth.", and there are two sentences *Murray Rothbard died in 1995 in Manhattan of a heart attack.* and *Burnham died of heart failure at the age of 86, on September 1, 1947.* in the train and evaluation set respectively, where both sentences can extract triples with the same syntactic structure.

How does the proposed framework perform on a non-identically distributed datasets?

Model	CaRB-AutoPara		
	F1	AUC	Opt.F1
IMoJIE	51.1	31.4	51.2
IMoJIE$^{\Phi}$	52.6	32.1	52.8
IMoJIE$^{\Psi}$	**53.4**	**33.9**	**53.4**

Table 4: Experimental results on CaRB-AutoPara.

To investigate the effectiveness as well as syntactic robustness on open world setting, we evaluate models on the syntactically diverse set CaRB-AutoPara. We find that the proposed training framework comprehensively improves the syntactic robustness of the existing model, making it exhibit consistent better performance on no-identically distributed data. As shown in Table 4, the best performing model significantly outperforms the base-

Data	CaRB-C1		CaRB-C2		CaRB-C3		CaRB-C4		CaRB-C5	
Distance	0.227		0.386		0.468		0.541		0.665	
Performance	AUC	Opt.F1	AUC	Opt.F1	AUC	Opt.F1	AUC	Opt.F1	AUC	Opt.F1
IMoJIE	34.2	55.3	31.5	51.4	25.7	50.0	30.7	47.9	24.7	47.0
IMoJIE$^{\Psi}$	**34.4**	**55.7**	27.9	**51.7**	**34.4**	**54.6**	**31.0**	**51.1**	**31.2**	**50.6**

Table 5: Experimental results on different subjects of syntactic categories.

line by 2.3 F1 pts, 2.5 pts of AUC, and 2.2 pts of optimal F1. In contrast, the basic model shows a large degradation on this dataset compared to the original CaRB. The results suggest that our proposed syntactically robust training is more compatible with the open-world scenarios, and it is necessary to train and evaluate models on a non-identically distributed dataset.

The proposed evaluation set CaRB-AutoPara is more challenging for OpenIE models that are trained on existing general datasets. The syntactic structures are varied with respect to the training set. By taking the same example mentioned above, there are sentences with a different voice and tense in the proposed CaRB-AutoPara, such as a question sentence *Isn't it possible that he died of a heart attack?*.

3.5 Analysis

We further explore the performance of the model on different subsets representing prototypical syntactic categories, and analyze the trend of the model effect as the syntactic differences between the training set and the subset changed.

How to effectively measure the syntactic difference between sentences? As the training data is massive, we need an efficient metric of the syntactic differences between sentences to divide the test set and calculate the syntactic distance between the training set and test set.

We propose a simple but effective syntactic distance algorithm called Hierarchical Weighted Syntactic Distance (HW-Syntactic Distance), to measure the differences. Intuitively, the more similar the skeleton of two sentences is, the less syntactic difference they have, i.e., the less syntactic distance. We use a hierarchical weighted matching strategy on the constituency parse trees to calculate the syntactic distance between two sentences. As shown in Figure 3, given two sentences with their constituency parse trees T_1, T_2 prune at height 3, we first transform the tree nodes in T_1, T_2 to sequences q_1, q_2 based on the level-order traversal. Then, we use the longest substring matching algorithm to accumulate the total matching length l^{tot} of two sequences, where the length of i-th matched substring is multiplied by a sequentially discounting weight w_i. The final distance is a normalized value based on the minimum sequence length of q_1, q_2, and its value domain is [0, 1]. The detailed algorithm of HW-Syntactic distance is available in Appendix B.1.

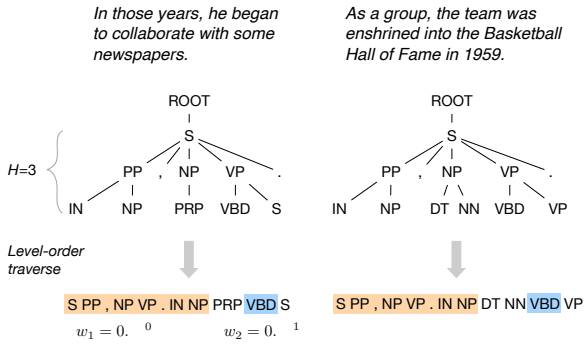


Figure 3: Illustration of HW-Syntactic Distance.

How does the models trained on partial syntactic distribution perform on syntactic-specific data? Based on this syntactic difference metric, we further analyze the performance of models trained on partially observed syntactic data $\mathcal{D}$ on different syntactic-specified datasets.

To this end, we first cluster the CaRB sentences into k subsets with the metric of HW-Syntactic Distance[3]. Then, we randomly sample 300 sentences in the training set, and calculate the distance between the training set and each subset by averaging the distances among sampled training sentences and each cluster center. We empirically clustered the CaRB sentences into 5 subsets with the optimal distance costs, and partial clustering results are available in Appendix B.3.

We find that the performance of the model on

[3]We use the K-means cluster algorithm, and cluster the samples with at most 300 epochs until convergence.

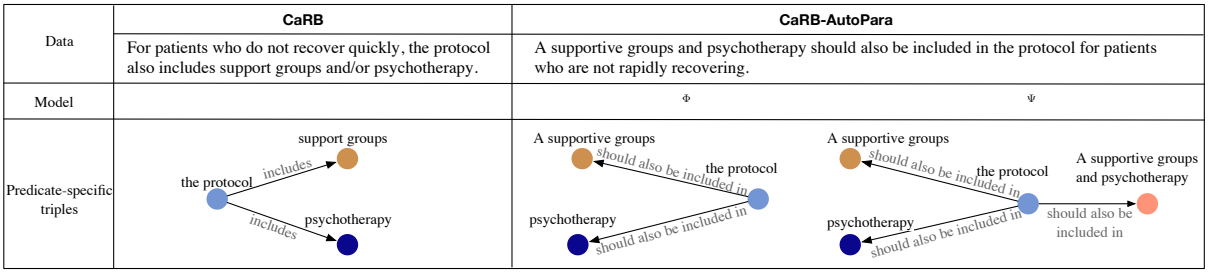


Figure 4: A case study shows the partial predictions of model trained on the proposed framework.

the subsets gradually increases as the syntactic distance between the training and test subsets decreases. As shown in Table 5, compared to the best performance of 55.3 obtained on the subset CaRB-C5 with a distance of 0.227, the basic model only achieved an optimal F1 score of 47.0 on the subset CaRB-C1. In addition, we find that our fully enhanced model is consistently better than the basic model trained on partial syntactic distribution, suggesting that the proposed training framework improves the syntactic robustness of the OpenIE model comprehensively. We remain more analysis and results of syntactic distribution in Appendix B.2.

3.6 Case Study

Figure 4 shows the case study of our proposed framework with different implementations. As is shown, compared to the original training sample, the generated sample exhibit a syntactically different structure. The model trained on the extended dataset with the semantic similarity-based knowledge restoration can only extract two separate triples around the predicate *should also be included in*. By using the full knowledge restoration algorithms, the trained model can extract all related triples for the predicate. A part of generated samples based on the proposed syntactic robust training framework are shown in Appendix C.

4 Related Work

Open Information Extraction is a fundamental NLP task with a long research history (Niklaus et al., 2018). Traditional models adopt rule-based or statistical methods incorporating syntactic or semantic parsers to extract knowledge tuples (Michele et al., 2007; Fader et al., 2011; Angeli et al., 2015; Del Corro and Gemulla, 2013; Pal et al., 2016; Saha and Mausam, 2018; Stanovsky et al., 2015; Gashteovski et al., 2017). Recently, neural models that either adopt sequence labeling strategies (Stanovsky et al., 2018; Roy et al., 2019; Zhan and Zhao, 2020; Kolluru et al., 2020a; Yu et al., 2021), or leverage sequence generative paradigms (Cui et al., 2018b; Sun et al., 2018; Kolluru et al., 2020b) have achieved promising result. To alleviate the problem that neural models rely heavily on labor-intensive annotated data, (Tang et al., 2020) proposes an unsupervised method that pretrains the model on synthetic data automatically labeled by patterns and then refines it using the RL process.

Paraphrase Generation has proven to be useful for adversarial training and data augmentation (Zhou and Bhat, 2021). Early methods adopt hand-crafted rules (McKeown, 1983), synonym substitution (Bolshakov and Gelbukh, 2004), machine translation (Quirk et al., 2004), and deep learning (Gupta et al., 2018; Liu et al., 2020) to improve the quality of generated sentences. To acquire syntactic diverse samples, recent studies involve reinforcement learning (Qian et al., 2019) or syntactic constrains (Iyyer et al., 2018; Goyal and Durrett, 2020; Sun et al., 2021) into the models.

5 Conclusion

In this paper, we focus on solving the problem of partially observable of syntactic distribution on training data, and propose a syntactically robust training framework that enables OpenIE models to be trained on a syntactic-abundant distribution based on diverse paraphrase generation. We propose a knowledge restoration algorithm to recover the deformed triples in syntactically transformed sentences based on semantic similarity-based matching and syntactic tree walking. To investigate the syntactic robustness of models, we build a syntactically diverse evaluation set that is consistent with the real-world setting. The experimental result with extensive analysis demonstrated the efficiency of our framework.

Acknowledgement

We thank all reviewers for their work and suggestions. We thank Xiaozhi Wang for his help with insightful comments during this work. This work is supported by the Key-Area Research and Development Program of Guangdong Province (2019B010153002), the NSFC Youth Project (62006136) and a grant from the Institute for Guo Qiang, Tsinghua University (2019GQB0003).

Limitations

Although we have extensively studied different paraphrase generation models with diverse syntactic, it is difficult to guarantee the quality of the generated sentences in a specific domain. In this paper, some poorly generated sentences can cause errors to propagate into knowledge restoration and further lead to omitted triples. We built a syntactically diverse dataset to evaluate the robustness of the OpenIE models. However, researchers willing to use this dataset need to be aware of the inevitable noises due to the automatic generation process.

References

Gabor Angeli, Melvin Jose Johnson Premkumar, and Christopher D. Manning. 2015. Leveraging linguistic structure for open domain information extraction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*.

Sangnie Bhardwaj, Samarth Aggarwal, and Mausam Mausam. 2019. Carb: A crowdsourced benchmark for open ie. 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 6262–6267.

Igor A Bolshakov and Alexander Gelbukh. 2004. Synonymous paraphrasing using wordnet and internet. In *International Conference on Application of Natural Language to Information Systems*.

Mingda Chen, Qingming Tang, Sam Wiseman, and Kevin Gimpel. 2019. A multi-task approach for disentangling syntax and semantics in sentence representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*.

Lei Cui, Furu Wei, and Ming Zhou. 2018a. Neural open information extraction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*.

Lei Cui, Furu Wei, and Ming Zhou. 2018b. Neural open information extraction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*.

Luciano Del Corro and Rainer Gemulla. 2013. Clausie: clause-based open information extraction. In *Proceedings of the 22nd international conference on World Wide Web*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*.

Anthony Fader, Stephen Soderland, and Oren Etzioni. 2011. Identifying relations for open information extraction. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*.

Angela Fan, Claire Gardent, Chloé Braud, and Antoine Bordes. 2019. Using local knowledge graph construction to scale Seq2Seq models to multi-document inputs. 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)*. Association for Computational Linguistics.

Kiril Gashteovski, Rainer Gemulla, and Luciano del Corro. 2017. MinIE: Minimizing facts in open information extraction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*.

Tanya Goyal and Greg Durrett. 2020. Neural syntactic preordering for controlled paraphrase generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Ankush Gupta, Arvind Agarwal, Prawaan Singh, and Piyush Rai. 2018. A deep generative framework for paraphrase generation. In *Proceedings of the aaai conference on artificial intelligence*.

Mohit Iyyer, John Wieting, Kevin Gimpel, and Luke Zettlemoyer. 2018. Adversarial example generation with syntactically controlled paraphrase networks. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics*.

Keshav Kolluru, Vaibhav Adlakha, Samarth Aggarwal, Mausam, and Soumen Chakrabarti. 2020a. OpenIE6: Iterative Grid Labeling and Coordination Analysis for Open Information Extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Keshav Kolluru, Samarth Aggarwal, Vipul Rathore, Mausam, and Soumen Chakrabarti. 2020b. IMoJIE: Iterative memory-based joint open information extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Xianggen Liu, Lili Mou, Fandong Meng, Hao Zhou, Jie Zhou, and Sen Song. 2020. Unsupervised paraphrasing by simulated annealing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60.

Kathleen McKeown. 1983. Paraphrasing questions using given and new information. *American Journal of Computational Linguistics*.

Banko Michele, J Cafarella Michael, Soderland Stephen, Broadhead Matthew, and Etzioni Oren. 2007. Open information extraction from the web. *International Joint Conference on Artificial Intelligence (IJCAI)*.

Christina Niklaus, Matthias Cetto, André Freitas, and Siegfried Handschuh. 2018. A survey on open information extraction. In *Proceedings of the 27th International Conference on Computational Linguistics*.

Harinder Pal et al. 2016. Demonyms and compound relational nouns in nominal open ie. In *Proceedings of the 5th Workshop on Automated Knowledge Base Construction*.

Lihua Qian, Lin Qiu, Weinan Zhang, Xin Jiang, and Yong Yu. 2019. Exploring diverse expressions for paraphrase generation. 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)*.

Chris Quirk, Chris Brockett, and William Dolan. 2004. Monolingual machine translation for paraphrase generation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*

Arpita Roy, Youngja Park, Taesung Lee, and Shimei Pan. 2019. Supervising unsupervised open information extraction models. 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)*.

Swarnadeep Saha and Mausam. 2018. Open information extraction from conjunctive sentences. In *Proceedings of the 27th International Conference on Computational Linguistics*.

Gabriel Stanovsky, Ido Dagan, et al. 2015. Open ie as an intermediate structure for semantic tasks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*.

Gabriel Stanovsky, Julian Michael, Luke Zettlemoyer, and Ido Dagan. 2018. Supervised open information extraction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*.

Jiao Sun, Xuezhe Ma, and Nanyun Peng. 2021. Aesop: Paraphrase generation with adaptive syntactic control. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*.

Mingming Sun, Xu Li, Xin Wang, Miao Fan, Yue Feng, and Ping Li. 2018. Logician: a unified end-to-end neural approach for open-domain information extraction. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*.

Jialong Tang, Yaojie Lu, Hongyu Lin, Xianpei Han, Le Sun, Xinyan Xiao, and Hua Wu. 2020. Syntactic and semantic-driven learning for open information extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2020*.

John Wieting and Kevin Gimpel. 2018. ParaNMT-50M: Pushing the limits of paraphrastic sentence embeddings with millions of machine translations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Zhao Yan, Duyu Tang, Nan Duan, Shujie Liu, Wendi Wang, Daxin Jiang, Ming Zhou, and Zhoujun Li. 2018. Assertion-based qa with question-aware open information extraction. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Bowen Yu, Yucheng Wang, Tingwen Liu, Hongsong Zhu, Limin Sun, and Bin Wang. 2021. Maximal clique based non-autoregressive open information extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*.

Junlang Zhan and Hai Zhao. 2020. Span model for open information extraction on accurate corpus. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Jianing Zhou and Suma Bhat. 2021. Paraphrase generation: A survey of the state of the art. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*.

A Model Parameters settings

We train all models on an NVIDIA Tesla V100 with 32GB GPU ARM. Hyperparameter settings for the paraphrase generation, knowledge restoration and OpenIE are listed in Table 6, 7 and 8, respectively.

Hyperparameter	Value
Backbone Model	BART_{base}
Model Dimension	768
Learning Rate	3e-5
Target Tree Height	2
Optimizer	Adam

Table 6: Settings for paraphrase generation model.

Hyperparameter	Value
Contextual Similarity Model	BERT_{base}
Threshold τ	0.7
Maintaining Spans k	5
Predicate Restoration Model	T5_{base}
Model Dimension	768
Learning Rate	1e-3
Optimizer	Adafactor

Table 7: Settings for knowledge restoration model.

Hyperparameter	Value
Backbone Model	BERT_{small}
Model Dimension	768
Learning Rate	2e-5
LSTM Hidden Dimension	256
LSTM Word Embedding	100
Optimizer	Adam

Table 8: Settings for OpenIE model.

B Syntactic Distribution Analysis

B.1 Hierarchical Weighted Syntactic Distance

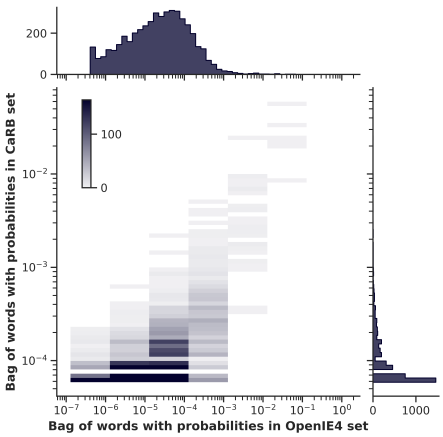
The proposed Hierarchical Weighted Syntactic Distance (HW-Syntactic Distance) is shown in algorithm 2. Given two sentences with their constituency parse trees T_1, T_2, the algorithm outputs their syntactic distance in [0, 1], where a smaller value means a closer distance. We first get their level-order traversal sequences q_1, q_2. Then we calculate their discounting weighted optimal matching length based on dynamic programming effectively. The final distance is a normalized value based on the minimum sequence length of q_1, q_2.

Algorithm 2 HW-Syntactic Distance

Input: Constituency parses T_1, T_2 of sentences s_1, s_2, pruning height h, discount factor α
Output: Syntactic distance d between s_1, s_2

Get trees T_1^h, T_2^h pruned at height h, and their level-order traversal sequences q_1, q_2
Initialize total length and count $l = 0; m = 0$
A[i][0]=1 if $q_1[i] == q_2[0], i = 1, ..., q_{1 \cdot len}$
A[0][j]=1 if $q_1[0] == q_2[j], j = 1, ..., q_{2 \cdot len}$
for $i = 2 \rightarrow q_1.len$ **do**
    for $j = 2 \rightarrow q_2.len$ **do**
        if $q_1[i] == q_2[j]$ **then**
            $\text{A}[i][j] = \text{A}[i-1][j-1] + 1$
        else
            $\text{A}[i][j] = 0$
            if $\text{A}[i-1][j-1] > 1$ **then**
                $l = \text{A}[i-1][j-1] \times \alpha^m$
                $m++$
            end if
        end if
    end for
end for
if $\text{A}[i-1][j-1] > 1$ **then**
    $l = \text{A}[i-1][j-1] \times \alpha^m$
end if
Return $1 - l/min(q_{1 \cdot len}, q_{2 \cdot len})$

B.2 Joint Words Distributions



We analyze the joint probability distribution of distinct words between training data and CaRB data based on the vocabulary built on CaRB. As shown above, we find that there is a word-level distribution difference between the two datasets.

B.3 Clustered Syntactic Samples

We cluster the CaRB data with the HW-Syntactic Distance. Partial examples are shown in Table 9.

Cluster	Syntactic Parse	Score
C1	In those years, he began to collaborate with some newspapers. – (ROOT (S (PP (IN) (NP)) (,) (NP (PRP)) (VP (VBD) (S)) (.)))	0.245
	In Canada, there are two organizations that regulate university and collegiate athletics. – (ROOT (S (PP (IN) (NP)) (,) (NP (EX)) (VP (VBP) (NP)) (.)))	0.333
	As a result, it becomes clear that the microbe can not survive outside a narrow pH range. – (ROOT (S (PP (IN) (NP)) (,) (NP (PRP)) (VP (VBZ) (ADJP) (SBAR)) (.)))	0.346
	However, during his rehearsal, Knievel lost control of the motorcycle and crashed into a cameraman. – (ROOT (S (ADVP (RB)) (,) (PP (IN) (NP)) (,) (NP (NNP)) (VP (VP) (CC) (VP)) (.)))	0.474
	If given this data, the Germans would be able to adjust their aim and correct any shortfall. – (ROOT (FRAG (SBAR (IN) (S)) (.)))	0.711
C2	HTB 's aim is for an Alpha course to be accessible to anyone who would like to attend the course , and in this way HTB seeks to spread the teachings of Christianity . – (ROOT (S (S (NP) (VP)) (,) (CC) (S (PP) (VP)) (.)))	0.019
	" Business across the country is spending more time addressing this issue , " says Sen. Edward Kennedy (D. , Mass .) . – (ROOT (SINV (") (S (NP) (VP)) (,) ("") (VP (VBZ)) (NP (NP) (PRN)) (.)))	0.069
	Returning home , Ballard delivers her report , which her superiors refuse to believe . – (ROOT (S (S (VP)) (,) (NP (NNP)) (VP (VBZ) (NP)) (.)))	0.131
	" It 's really bizarre , " says Albert Lerman , creative director at the Wells Rich Greene ad agency . – (ROOT (SINV (") (S (NP) (VP)) (,) ("") (VP (VBZ)) (NP (NP) (,) (NP)) (.)))	0.136
	Feeling the naggings of a culture imperative , I promptly signed up . – (ROOT (S (S (VP)) (,) (NP (PRP)) (VP (ADVP) (VBD) (PRT)) (.)))	0.210
C3	Historically , Aiseau was a village dedicated to agriculture , logging , but also to the industry . – (ROOT (S (NP (NNP)) (,) (NP (NNP)) (VP (VBD) (NP)) (.)))	0.015
	They beat Milligan 1-0 , Grand View 3-0 , Webber International 1-0 and Azusa Pacific 0-0 to win the NAIA National Championships . – (ROOT (FRAG (S (NP) (VP)) (,) (NP (NP) (,) (NP) (CC) (NP)) (.)))	0.051
	For the record , Jeffrey Kaufman , an attorney for Fireman 's Fund , said he was " rattled – both literally and figuratively . " – (ROOT (S (PP (IN) (NP)) (,) (NP (NP) (,) (NP) (,)) (VP (VBD) (SBAR)) (.) ("")))	0.154
	Crouched at shortstop , Bert Campaneris , once Oakland 's master thief , effortlessly scoops up a groundball and flips it to second . – (ROOT (S (S (VP)) (,) (NP (NP) (,) (ADVP) (NP) (,)) (VP (ADVP) (VP) (CC) (VP)) (.)))	0.162
	Now Mr. Broberg , a lawyer , claims he 'd play for free . – (ROOT (S (ADVP (RB)) (NP (NP) (,) (NP) (,)) (VP (VBZ) (SBAR)) (.)))	0.178
C4	In the U.S. , more than half the PC software sold is either for spreadsheets or for database analysis , according to Lotus . – (ROOT (S (PP (IN) (NP)) (,) (NP (NP) (VP)) (VP (VBZ) (PP) (,) (PP)) (.)))	0.101
	It is part of the Surrey Hills Area of Outstanding Beauty and situated on the Green Sand Way . – (ROOT (S (NP (PRP)) (VP (VBZ) (NP)) (.)))	0.018
	This is the U.N. group that managed to traduce its own charter of promoting education , science and culture . – (ROOT (S (NP (DT)) (VP (VBZ) (NP)) (.)))	0.138
	Of the self - starting vacuum cleaner , he says : " Could be Cuddles , Mrs. Stinnett 's dog . " – (ROOT (S (PP (IN) (NP)) (:) (S (VP)) (,) (NP (PRP)) (VP (VBZ) (:) (") (S)) (.) ("")))	0.210
	According to the 2010 census , the population of the town is 2,310 . – (ROOT (S (PP (VBG) (PP)) (,) (NP (NP) (PP)) (VP (VBZ) (NP)) (.)))	0.271
C5	Sen. Mitchell is confident he has sufficient votes to block such a measure with procedural actions . – (ROOT (S (NP (NNP) (NNP)) (VP (VBZ) (ADJP)) (.)))	0.003
	Dr. Pim played for Ireland against England in 1892 , 1893 , 1894 and 1896 . – (ROOT (S (NP (NNP) (NNP)) (VP (VBD) (PP) (PP)) (.)))	0.051
	From 1909 to 1912 , the Miami Canal was dug , bypassing the rapids at the head of the North Fork . – (ROOT (S (PP (PP) (PP)) (,) (NP (DT) (NNP) (NNP)) (VP (VBD) (VP)) (.)))	0.112
	Mrs. Marcos has n't admitted that she filed any documents such as those sought by the government . – (ROOT (S (NP (NNP) (NNP)) (VP (VBZ) (RB) (VP)) (.)))	0.114
	Hapoel Lod played in the top division during the 1960s and 1980s , and won the State Cup in 1984 . – (ROOT (S (NP (NNP) (NNP)) (VP (VP) (,) (CC) (VP)) (.)))	0.203

Table 9: A part of samples in different clusters of CaRB data based on the HW-Syantactic distance. The score refers to the distance between current sentence and corresponding clustering center.

Original sample	Generated sample
This finding indicated that organic compounds could carry current. *(This finding, indicated that, organic compounds could carry current)*	According to these results, organic compounds can carry the current. *(organic compounds, can carry, the current)*
	This finding has shown that organic compounds are capable of transmitting impulses. *(This finding, has shown, that organic compounds are capable of transmitting impulses)* *(organic compounds, are capable, transmitting impulses)*
	That this finding has shown that organic compounds can be operated. *(this finding, has shown, that organic compounds can be operated)* *(organic compounds, can be operated,)*
Regulations meant that the original sixth lap would be deleted and the race would be restarted from the beginning of said lap. *(Regulations meant that, would be deleted, the original sixth lap)* *(Regulations meant that, would be deleted, the race)*	According to the rules, the original sixth round will be removed and the race will be re started at the beginning of the round. *(the race, will be re started, at the beginning of the round)* *(the original sixth round, will be removed,)*
	The rules have made it possible to cancel the original sixth round and restart the race at the start of the round. *(The rules, have made it possible, to cancel the original sixth round)* *(The rules, have made it possible, restart the race at the start of the round)*
	But the rules stipulated that the original sixth round would be removed and the race to be re-started at the beginning of the round. *(The rules, stipulated, that the original sixth round would be removed)* *(the race, to be re started, at the beginning of the round)* *(the original sixth round, would be removed)*
Maduveya Vayasu song from nanjundi kalyana was a track played during marriages for many many years in Kannada. *(Maduveya Vayasu, is, a song)* *(Maduveya Vayasu song, is from, anjundi kalyana)* *(Maduveya Vayasu song, was a track played during, marriages)*	The song of maduveya vayasu from nanjundi kalyana has been played in the marriage of many years in kannada. *(The song of maduveya vayasu from nanjundi kalyana, has been played, in the marriage of many years in kannada)* *(The song of maduveya vayasu, is from, nanjundi kalyana)*
	Maduveya vayasu, the song of nanjundi kalyana has been played in many marriages throughout the country. *(the song of nanjundi kalyana, has been played, in many marriages throughout the country)*
	When they were married, they played the song of maduveya vayasu from nanjundi kalyana. *(they, played, the song of maduveya vayasu from nanjundi kalyana)* *(they, were married)* *(the song, is from, nanjundi kalyana)*

Table 10: A part of generated syntactically robust data samples based the proposed framework.

C Syntactically Robust Samples

Base on the proposed framework, a part of generated samples are shown in Table 10.