KGAST: From Knowledge Graphs to Annotated Synthetic Texts

Nakanyseth Vuth, Gilles Sérasset, Didier Schwab

KGAST: From Knowledge Graphs to Annotated Synthetic Texts

Nakanyseth Vuth and **Gilles Sérasset** and **Didier Schwab**
Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG
38000 Grenoble
France
first.last@univ-grenoble-alpes.fr

Abstract

In recent years, the use of synthetic data, either as a complement or a substitute for original data, has emerged as a solution to challenges such as data scarcity and security risks. This paper is an initial attempt to automatically generate such data for Information Extraction tasks. We accomplished this by developing a novel synthetic data generation framework called **KGAST**, which leverages Knowledge Graphs and Large Language Models. In our preliminary study, we conducted simple experiments to generate synthetic versions of two datasets—a French security defense dataset and an English general domain dataset, after which we evaluated them both intrinsically and extrinsically. The results indicated that synthetic data can effectively complement original data, improving the performance of models on classes with limited training samples. This highlights **KGAST**'s potential as a tool for generating synthetic data for Information Extraction tasks.

1 Introduction

Information Extraction (IE) models serve as crucial components across various domains, enabling us to make informed decisions based on complex data. However, the effectiveness of these models is dependent on the availability and quality of training data. In this context, we encounter two critical challenges: *1) Data Scarcity*: Frequently, despite having complex modeling techniques, researchers deal with datasets that are either insufficient in size or entirely unavailable. Without a sufficient number of labeled examples, IE models struggle to generalize effectively, compromising their predictive capabilities. *2) Privacy and Compliance Concerns*: In an era of heightened data privacy regulations, organizations must navigate the balance between model performance and safeguarding sensitive information. Certain datasets whether due to privacy risks or legal constraints cannot be openly shared, which further complicates training effective IE models. To address these issues, researchers often resort to manual data augmentation. This labor-intensive process involves collecting additional data points and meticulously annotating them. While effective, it is time-consuming and resource-intensive, making it less feasible for small organizations operating on limited budgets. Numerous studies have explored the expansion of training data by introducing additional synthetic data (Kobayashi, 2018; Wei and Zou, 2019; Zhang et al., 2020). These studies presented straightforward strategies, such as substituting certain words with their equivalent terms. These equivalents can be retrieved from external sources like WordNet (Miller, 1995), DBnary (Sérasset, 2015), or they can be calculated using word embedding models such as Word2Vec (Mikolov et al., 2013), FastText (Bojanowski et al., 2016), and Glove (Pennington et al., 2014). Although these techniques can indeed augment the initial training dataset, they fail to generate adequate diversity for the models to generalize effectively in subsequent tasks, owing to the minimal semantic variations from the original data. Back-translation is another recognized technique for augmenting the initial training data. With Machine Translation models (Xie et al., 2017; Fabbri et al., 2020), paraphrases of each sentence can be obtained through the back-translation process. While back-translation effectively amplifies the dataset size twofold, it introduces a notable challenge in terms of annotation. The text derived from back-translation diverges from the original annotation. Thus, either careful manual annotation or sophisticated annotation algorithms are required to update the annotations in alignment with the back-translated text, ensuring the precision of the dataset.

To address these issues of low data diversity and misalignment of the original annotation, we propose a novel synthetic data generation

Proceedings of the 1st Workshop on Knowledge Graphs and Large Language Models (KaLLM 2024), pages 43–55

framework called **KGAST**. This framework leverages **K**nowledge **G**raph to automatically generate **A**nnotated **S**ynthetic **T**exts that can be used for training IE models. In this study, we sought to answer questions similar to the ones raised in this paper (Claveau et al., 2021):

- Can synthetic data serve as supplementary data to improve the performance of classes with limited training samples?
- Can synthetic data be a viable alternative to gold standard data?

2 Related Works

2.1 Data Augmentation

Given the resource-intensive nature of manual data creation and annotation, a variety of data augmentation strategies have been employed to address the issue of data scarcity. One of the well-known approaches is the rule-based (Kobayashi, 2018; Wei and Zou, 2019; Zhang et al., 2020) word replacement. This method requires a heuristic for selecting and replacing words within a sentence. On the other hand, some research has approached this at sentence level by leveraging dependency tree (Coulombe, 2018; Dehouck and Gómez-Rodríguez, 2020). For example: "John did the math exercises." is replaced with "The math exercise was done by John". The data augmented through these methods often conveys information very similar to the original, thereby limiting semantic diversity. Back-translation (Xie et al., 2017; Fabbri et al., 2020) is another method to augment the original dataset. This straightforward technique involves translating text from the original language to another language, and then translating it back to the original language to produce a new version. However, this method presents its own challenges, as labels from the original text may no longer align with the new text due to changes in syntax or semantics.

2.2 Distant Supervision

Automatically generating new supervised data is a compelling alternative to manual annotation, especially when creating large-scale datasets for natural language processing tasks. One such approach is Distant Supervision (Roller et al., 2015; Deng and Sun, 2019), which leverages existing knowledge bases to construct and label new training samples. The core assumption of this method is that if two entities share a relation in the knowledge base, any sentence containing those two entities might express that relation. However, the automated annotation process introduces errors such as incorrectly assumed entity types or the relations between entity pairs.

2.3 In context Learning

In-context learning (ICL) has recently emerged as a new paradigm in the field of natural language processing. This approach allows Large Language Models (LLMs) to make predictions based solely on contexts that are augmented with a select number of examples. Often, ICL aids in refining the output of an LLM, enhancing the accuracy of the output even in the absence of fine-tuning. With ICL, the performance achieved by LLMs can rival that of previous supervised learning methods (Brown et al., 2020; Shin et al., 2021; Wan et al., 2023). This can be achieved by carefully crafting clear instructional prompts along with high-quality task-specific k-shot examples (Zhao et al., 2021; Liu et al., 2022).

3 Method

3.1 Overview

Our proposed methodology is built upon two primary elements: LLMs and ICL. Contrary to previous research that modified the original texts of the dataset, our approach involves the generation of new synthetic texts and their corresponding annotations using LLMs. A well-known reasoning prompt method, Chain-of-Thoughts (Wei et al., 2022), enables us to create complex ICL prompt templates to instruct LLM models for such tasks. The intuition behind this approach stems from the concept of distant supervision, where we make a naive assumption that if a pair of entities (e_{head}, e_{tail}) is present in both the text and the Knowledge Graph (KG), these two entities maintain the same relation as the one in the KG.

3.2 Task Formulation

We formalize the task of synthesizing annotated data as a natural language generation task. Consider a given gold text $t \in \mathcal{G}$, where $\mathcal{G} = \{t_1, \ldots, t_n\}$ represents the set of gold standard texts, $A_t = \{a_1, \ldots, a_n\}$ is the set of label annotations, $R_t = \{r_1, \ldots, r_n\}$ is the set of relations, K_t is the KG and $A_t, R_t \subseteq K_t$. The intuition is to construct a text generation prompt p by utilizing the KG as an input to get our intended output synthetic

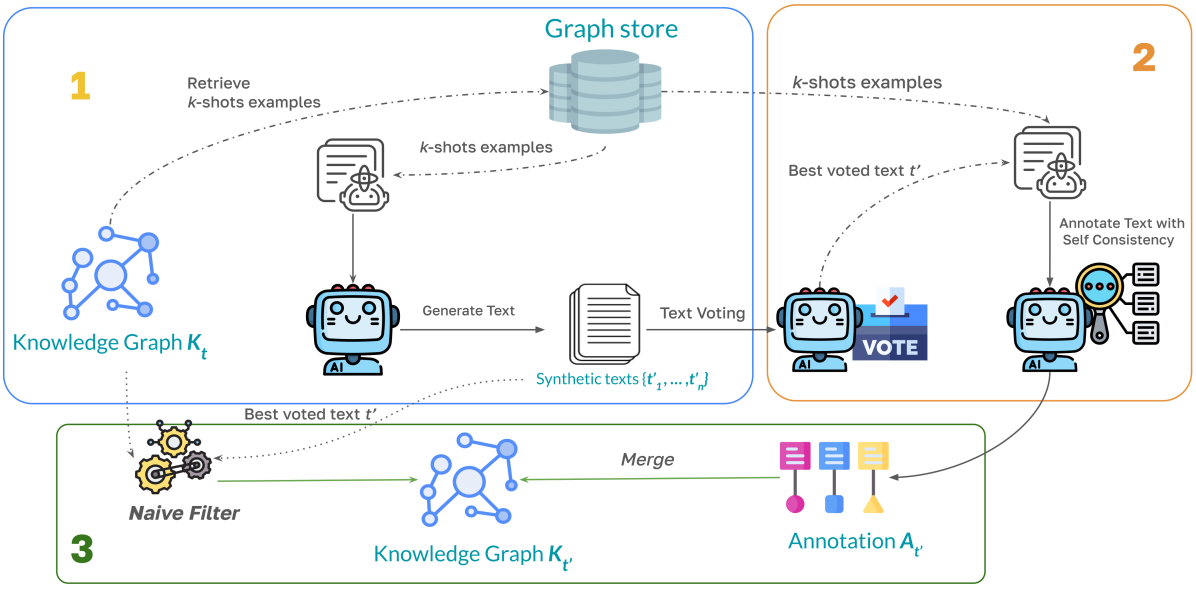

Figure 1: KGAST framework. (1) It begins by using a knowledge graph K_t as a starting point. Based on this knowledge graph, the framework retrieves a set of relevant k-shots examples to build prompt templates. We then prompt the LLM to generate a set of synthetic texts $\{t'_1, \ldots, t'_n\}$. (2) From these generated texts, we select the best t' through a voting prompt which will then be used as input in the annotation prompt template to prompt the model for text annotating. (3) Finally, annotations retrieved from the LLM outputs are merged with the filtered knowledge graph $K_{t'}$ to get the final annotation.

text t'. The KG K_t of a prompt p is constructed by extracting all the annotated relations R_t from the text t. Once the synthetic text t' is produced, we proceed to extract the set of label annotations $A_{t'}$ and the set of relations $R_{t'}$ by simply filtering out any triples of the KG K_t where either the head or the tail is not present in the text t'. This can be formulated as $A_{t'}, R_{t'} \subseteq K_{t'}$, where:

$$K_{t'} = naive\_filter(K_t, t') \quad (1)$$

3.3 Prompt Construction

Our prompt template is divided into three main sections for ease of understanding.

Instruction Serves as a clear directive for the LLM, which outlines its role and task. We clearly specify the role and task for which we want to generate output. For instance, in the case of text generation:

> "You are a *creative text writer*. Write me a text using the provided Knowledge Graph. Your objective is to write a coherent text that incorporates all the given triples (head, relation, tail) of the Knowledge Graph. You have the right to make the text creative and informative, but you must make sure that the text reflects the given Knowledge Graph."

For text annotation, we draw inspiration from Tree-of-thought (Yao et al., 2023) prompting and construct the instruction as follows:

> "You are a *text annotator*. Your objectives are to: 1/ Analyze the given text in detail, 2/ Annotate possible entities based on these entity types: person(PER), location(LOC), organization(ORG), time(TIME), numbers(NUM), and miscellaneous entity names(MISC). Response with the annotation in this format: "Possible Entities: e_1 e_2,...., e_n", where e_1 to e_n are the extracted entities."

k-shots Examples Similar to the Retrieval-Augmented Generation (RAG) (Lewis et al., 2020), where all documents are embedded into latent space, we did the same for all examples retrieved from $\mathcal{G}$. We then use top-k retrieval to find examples that are close to the input KG K_t as k-shot examples for prompting.

Test Input We used KG K_t as an input in the form of a list of triples in natural language to prompt the LLM model. The structure of these input triples is as follows:

```
("Head":Type, "Relation", "Tail":Type)
```

The goal is to provide the LLM with as much information as possible so that it can generate a coherent text that corresponds to the input triples.

3.4 The Framework: KGAST

The process outlined in Section 3.2 yields both the text and its corresponding annotation. However, we identified two primary issues:

Incomplete Text Annotation Despite having annotations, we found that it is often incomplete as seen in Figure 2. The method's effectiveness is heavily reliant on the performance of the LLM used. Consequently, the likelihood of generating a text, t', that includes all the input triples of K_t is contingent on the LLM's performance for our given tasks.

Text Coherence and Validity Without a validation heuristic, the texts generated by our method may contain nonsensical phrases. This is because LLMs are known to produce hallucinations, such as incoherent texts with their input KGs, repetitive tokens, inclusion of parts of the prompt, and texts in different languages.

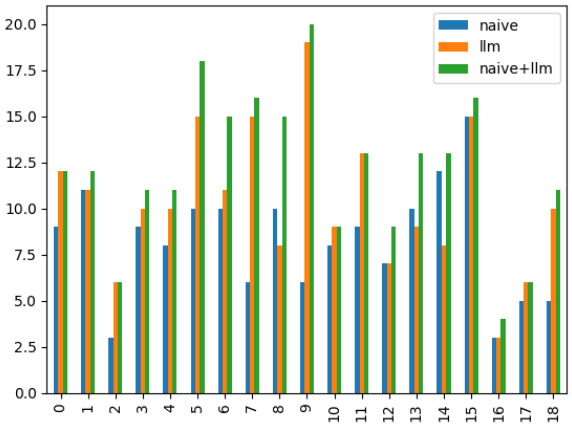

Figure 2: The graph displays entity coverage statistics for various texts. The Y-axis represents the number of entities, while the X-axis corresponds to the text.

To address these shortcomings, we integrated Self-consistency (Wang et al., 2022) into our framework, aiming to mimic real human annotation processes as closely as possible. The idea behind self-consistency is to prompt the LLM to generate a set of n outputs and select the most consistent one. For text generation, we prompted LLM to generate $n = 3$ outputs. We used the 3 output texts as input to prompt the LLM to vote $n = 5$ times, evaluating each based on creativity, coherence, and the text's capacity to include all the input triples of the KG. We then select the best text t' with the highest voting score. A similar approach is also applied for extracting annotations from text t'. We prompted the same model to generate $n = 5$ outputs and merged the most consistent annotation with a threshold of 0.5. This means that if an annotation appears in at least 50% of the n outputs, it is extracted. Subsequently, this annotation is merged with the annotation from our naive heuristic. The entire procedure is described in Figure 1.

4 Experiment Setup

4.1 Datasets

For simplicity, we will refer to gold standard data as $\mathcal{G}$ and synthetic data as $\mathcal{S}$.

DocRED (Yao et al., 2019), is a document-level relation extraction dataset constructed from Wikipedia and Wikidata. This dataset contains a total of 96 relations and 6 entity types for English general domain. Each relation is annotated along with its supporting evidence.

French Security and Defense for which we will refer as **FRSD**, is a document-level relation extraction dataset that covers the annotation for Event Extraction, Entity Recognition, Attribute Extraction, and Relation Extraction tasks for French intelligence service. FRSD contains 2,000 French documents, of which 800 are manually written and annotated by humans. It consists of 35 entity types, 20 attributes, and 49 relations.

4.2 Synthetic Data Generation

In this preliminary study of generating synthetic data, we did simple generation experiments by using the training set of $\mathcal{G}$ as a reference to generate the synthetic version $\mathcal{S}$. For DocRED, this resulted in a total of 3023 new documents along with their annotations. The LLM model used in the framework for this dataset was Zephyr-7B[1], a fine-tuned model of Mistral-7B (Jiang et al., 2023). The same approach was applied to FRSD but on 400 (train) documents. In FRSD, we observed a significant data imbalance among Event classes. With this in mind, we manually selected the top 10 event classes with the fewest samples and used their texts as a reference to generate 1200 new documents (Oversampling). We used Vigostral-7B[2], a chat-

[1] https://huggingface.co/HuggingFaceH4/zephyr-7b-beta

[2] https://huggingface.co/bofenghuang/vigostral-7b-chat

Pre-training Strategy	Events		Entities	
	F1 Macro	F1 Micro	F1 Macro	F1 Micro
No pre-training	43.28±1.32	58.58 ±1.90	66.78 ±1.59	81.81±0.33
$\mathcal{S}$ pre-training	45.17 ±0.83	58.81 ±0.46	67.45 ±1.33	81.61±0.12
$\mathcal{G}$ pre-training	43.60 ±3.46	58.56±1.95	67.72 ±2.13	81.95±0.36
$\mathcal{S} \cup \mathcal{G}$ pre-training	**45.19** ±2.07	**58.93**±0.90	**68.38** ±0.34	**82.03**±0.34

Table 1: Unified Model results for Event and Entity Extraction.

Pre-training Strategy	Attributes		Relations	
	F1 Macro	F1 Micro	F1 Macro	F1 Micro
No pre-training	**61.63**±2.15	81.87±0.52	44.53±1.45	56.74±0.78
$\mathcal{S}$ pre-training	60.05±1.97	**82.07**±0.62	43.26±1.22	55.85±0.75
$\mathcal{G}$ pre-training	55.98 ±8.18	81.33±0.73	43.25±4.78	**57.10**±0.70
$\mathcal{S} \cup \mathcal{G}$ pre-training	60.39±2.76	81.27±0.75	**46.49**±0.55	56.87±0.84

Table 2: Unified Model results for Attribute and Relation Extraction.

based model that has been fine-tuned on Mistral-7B (Jiang et al., 2023) for this dataset. Supporting evidence for each relation of $\mathcal{S}$ on both datasets was automatically extracted using a simple heuristic. This heuristic identifies the sentence index where the head or the tail entity appears and uses it as supporting evidence.

4.3 Tasks

To evaluate the effectiveness of our proposed framework, two types of evaluations were conducted:

Intrinsic Evaluation This evaluation aims to understand the accuracy of our framework's annotations. We evaluated the annotation accuracy of DocRED's synthetic documents. For Named Entity, this was done by using BERT-NER[3] as an inference model to predict the set of synthetic texts, and then comparing the output prediction with our framework's annotations. As for Relations, we used the best model of DREEAM (Ma et al., 2023) (RoBERTa)[4] as the inference model and followed the same procedure.

Extrinsic Evaluation The goal of this evaluation is to assess how $\mathcal{S}$ impacts the performance of downstream tasks. For DocRED, we used $\mathcal{S}$ to train on two tasks: Named Entity Recognition (NER), and Relation Extraction (RE). NER task was trained on the Flair framework[5] which used Bi-LSTM with flair embeddings. RE was trained using DREEAM (teacher) model, which is based on BERT with $\lambda = 0.05$. We trained a total of 5 times with different seeds, evaluated the models against the original test set, and computed the average to get the final results. For FRSD, we conducted experiments on two models: Boundary Smoothing (BS) (Zhu and Li, 2022): which is used for Event, Entity, and Attribution recognition tasks. Unified Model (UM) (Prieur et al., 2024): this model approaches the tasks jointly for NER and RE tasks. The architecture of this model includes a module for detecting entity spans and a second for predicting their interactions. In these experiments, we follow the same as the experiment conducted with DocRED and report the results in Section 5.3.

5 Experiment Results

5.1 Dataset Characteristic

A descriptive statistic of both datasets can be seen in Table 3. We observed that $\mathcal{G}$ tends to contain longer documents and possesses a greater number of unique entities and labeled tokens, indicating a higher semantic quality and more robust representation of specific objects, people, places, etc. We computed the Self-BLEU (trigram) for each dataset, revealing lower Self-BLEU scores for $\mathcal{G}$. The scores suggest that the gold datasets are more diverse and less prone to repetitive use of the same tokens/entities in the text. On the other hand, $\mathcal{S}$ can be seen to have a larger number of entity pairs (triples) due to the repetition of entities used in the texts. While this increases the raw count of entity

[3] https://huggingface.co/dslim/bert-base-NER
[4] https://github.com/YoumiMa/dreeam
[5] https://huggingface.co/flair/ner-english

pairs, it may not necessarily enhance the diversity or quality of these pairs.

In addition to the descriptive statistics of the datasets, we also studied the semantic and lexical difference between the set of gold texts t and synthetic text t' using Cosine Similarity as a measure. The sentence encoders, SimCSE[6] (for English) Camembert-large[7] (for French), were used for analyzing the semantic difference, while Bag-of-Words with TF-IDF was used for the lexical difference. Figure 3 shows the distribution of the score. Since t and t' describe the same knowledge graph K_t, despite their different writing styles, higher semantic similarity scores are expected. Lower lexical similarity scores indicate that different lexical properties and grammatical structures were used, even though both t and t' describe roughly the same K_t. A more in-depth study for producing more diverse texts needs to be done whether through parameter control, reworking the prompt template, or filtering out texts that will increase the Self-BLEU score.

	DocRED		FRSD	
	$\mathcal{G}$	$\mathcal{S}$	$\mathcal{G}$	$\mathcal{S}$
# Docs	3053	3023	400	1200
# Tokens	603468	493638	56128	184667
Toks/Doc	197.66	163.29	140.32	153.89
Sents/Doc	7.94	6.66	6.20	6.69
Sent Len	25.96	24.94	23.48	23.53
# Entity	79481	56766	11436	36825
# Triples	117712	157905	12940	41771
# Labels	147358	102558	15781	46912
Labels/Doc	48.27	33.93	39.45	39.09
Self-BLEU	0.53	0.63	0.58	0.76

Table 3: Descriptive statistics of $\mathcal{G}$ and $\mathcal{S}$ for both the DocRED and FRSD dataset. **# Labels** here represents the total number of labeled tokens in the dataset.

5.2 Intrinsic Evaluation

The performance on NER task can be observed in Table 4. We achieved high F1-scores of 0.93 and 0.92, demonstrating the effectiveness of our framework's annotation capacity. Among all the tags, we noticed that `MISC` was the only tag that scored the lowest. As for RE tasks, we considered two types of annotation accuracy: 1) Relation and 2) Evidence. We need to take into account that, originally the best DREEAM model only achieved a 67.53 F1-score on DocRED test set, thus the results reported might not be very accurate. Table 5 presents the accuracy results, showing that our naive assumption heuristic achieved a 0.63 F1-score for Relation and a 0.37 F1-score for Evidence. As $\mathcal{S}$'s evidence was solely based on a very naive heuristic, an over-prediction of the evidence is to be expected, leading to a low precision score and a high recall score.

	Precision	**Recall**	**F1-score**
B-LOC	0.95	0.98	0.96
B-MISC	0.92	0.74	0.82
B-ORG	0.91	0.89	0.90
B-PER	0.98	0.97	0.98
I-LOC	0.96	0.93	0.94
I-MISC	0.83	0.83	0.83
I-ORG	0.93	0.87	0.90
I-PER	0.98	0.99	0.98
Micro	0.94	0.91	0.93
Macro	0.94	0.91	0.92

Table 4: Intrinsic performance with BERT-NER's predictions as true labels.

	Precision	**Recall**	**F1-score**
Relation	0.66	0.61	0.63
Evidence	0.26	0.60	0.37

Table 5: Intrinsic performance with DREEAM's prediction as true labels.

5.3 Extrinsic Evaluation

While generating a large volume of new annotated synthetic texts can be accomplished with relative ease, the challenge lies in optimally utilizing this synthetic data. We conducted a series of experiments in order to address this.

DocRED In NER task, we conducted training under different scenarios. We trained the models separately on 1) the original training set $\mathcal{G}$, 2) the synthetic set $\mathcal{S}$, 3) a combination of $\mathcal{G} + \mathcal{S}$, 4) a subset of $\mathcal{G}$ by sampling documents that have $\geq$ 20% labeled tokens which left us with 1564 documents. As can be seen from Table 6, using only $\mathcal{G}$ generally yields better results. Although training solely on $\mathcal{S}$ produces acceptable results (79.14 on Weighted-F1), there is a significant performance gap between $\mathcal{G}$ and $\mathcal{S}$. Similar training strategies were applied for the RE task, except for scenario

[6] https://github.com/princeton-nlp/SimCSE

[7] https://huggingface.co/dangvantuan/sentence-camembert-large

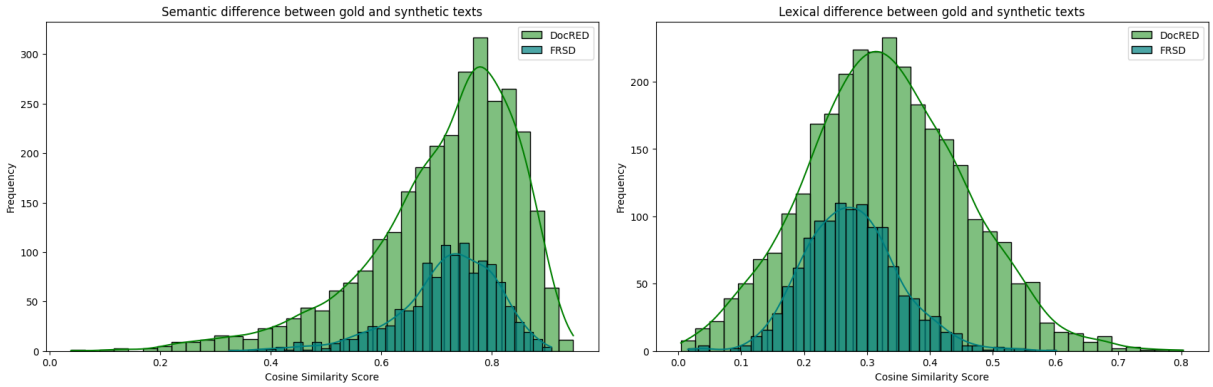

Figure 3: Distribution of the cosine similarity score between text t and t'. The left figure depicts semantic differences, while the right shows lexical differences. The text example can be found in the Appendix section A.

4, in which the model was trained using a random sample of 20% of the training set. We observed the same trend based on the scores shown in Table 7. From the evaluation results, it can be inferred that both tasks might benefit from more diverse training sets with higher semantic differences to generalize better and produce more robust performance.

FRSD For our first model BS, we randomly split $\mathcal{S}$ into three and used them as complement data to train the BS model. Results in Table 8 show the interest of using $\mathcal{S}$ as a complement training data. Notably, as shown in Tables 9 and 10, $\mathcal{S}$ enhances the performance of classes with fewer samples for Event and Attribute extraction tasks. As outlined in Section 3.4, annotations from $\mathcal{S}$ carry the risk of introducing a lot of noise due to incomplete annotation or wrongly assumed relations. Furthermore, the annotations also heavily rely on the capacity of the LLMs used. To address this issue, we tried to improve annotations by implementing a Teacher-Student learning strategy. This solution consists of training a Teacher model on $\mathcal{G}$. The Teacher model is then used to make predictions on $\mathcal{S}$. These predictions are used as annotations to pre-train a second model, the Student model. Finally, the training of the Student model is refined on the $\mathcal{G}$. For this experiment, we used only the batch of 400 texts from $\mathcal{S}$. Four training scenarios were explored: no pre-training, pre-training with $\mathcal{S}$'s annotation, pre-training on synthetic texts with annotations produced by the Teacher model, and finally, a pre-training with $\mathcal{S}$'s annotation together with those of the Teacher model. We discovered that there is an increase in the performance for classes with low samples, except for Attributes. The second observation is that using $\mathcal{S}$'s annotations alone is useful for low sample classes for Events and Entities. This significance grows when the annotations are combined with predictions from the Teacher model. The results are shown in Tables 1, 2.

	Weighted-F1	Macro-F1	Micro-F1
$\mathcal{G}$	**88.02**	**86.45**	**88.04**
$\mathcal{S}$	79.14	76.87	79.21
$\mathcal{G} + \mathcal{S}$	87.89	86.25	87.90
$\mathcal{G}_f$	**86.26**	**84.43**	**86.30**
$\mathcal{G}_f + \mathcal{S}$	85.87	84.06	85.89

Table 6: Evaluation results on DocRED's named entity recognition task.

	F1	Ign-F1	Evi-F1
$\mathcal{G}$	**61.51**±0.19	**59.7**±0.19	**52.09**±0.22
$\mathcal{S}$	44.41±0.48	43.51±0.46	31.17±0.46
$\mathcal{G} + \mathcal{S}$	60.04±0.34	58.27±0.36	50.67±0.36
$\mathcal{G}_f$	**56.12±0.28**	**55.46±0.28**	**46.99±0.28**
$\mathcal{G}_f + \mathcal{S}$	54.79±0.26	53.5±0.25	44.86±0.31

Table 7: Evaluation results on DocRED's relation extraction task. We used the same metrics that were proposed in DocRED's paper (Yao et al., 2019).

6 Conclusion

In this paper, we introduced a novel framework that leverages Knowledge Graphs and Large Language Models to generate annotated synthetic data for Information Extraction tasks. Our preliminary experiments demonstrated that while the data generated by this framework can enhance the performance of classes with limited training samples, it cannot yet serve as a substitute for the original data. Theoretically, within this framework, data anonymization and bias mitigation can be easily accomplished by modifying the input Knowledge Graphs. However,

	Events		Entities		Attributes	
Data	F1 macro	F1 micro	F1 macro	F1 micro	F1 macro	F1 micro
$\mathcal{G}$	41.83 ±0.79	55.56 ±0.63	65.41 ±1.04	**81.60** ±0.26	56.74 ±0.86	80.02 ±0.26
$\mathcal{G} + \mathcal{S}_{400}$	43.92 ±1.14	55.94 ±0.80	**65.82** ±1.04	81.14 ±0.14	59.17 ±1.11	**80.86** ±0.32
$\mathcal{G} + \mathcal{S}_{800}$	43.61 ±0.96	56.00 ±0.47	64.33 ±0.94	79.91 ±0.34	59.96 ±1.82	80.61 ±0.69
$\mathcal{G} + \mathcal{S}_{1200}$	**44.20** ±1.08	**56.45** ±0.85	63.56 ±0.82	80.06 ±0.38	**60.67** ±0.95	80.46 ±0.35

Table 8: Evaluation results for Event/Entity/Attribute extraction using BS. {400, 800, 1200} are dataset's sizes.

	$\mathcal{G}$		$\mathcal{G} + \mathcal{S}_{1200}$	
Classes	#Samples	F1-score	#Samples	F1-score
CIVIL_WAR_OUTBREAK	19	**57.45**	440	54.83
COUP_D_ETAT	24	44.23	198	**45.22**
DEMONSTRATION	38	4.11	1215	**12.44**
DRUG_OPERATION	13	18.48	242	**41.30**
ELECTION	27	65.73	197	**82.45**
ILLEGAL_CIVIL_DEMO.	29	**27.31**	30	20.48
NATURAL_CAUSES_DEATH	9	40.25	15	**43.02**
POLITICAL_VIOLENCE	29	**10.72**	137	3.51
POLLUTION	31	60.11	141	**68.01**
SUICIDE	22	39.27	22	**41.92**
TRAFFICKING	38	31.85	381	**45.27**

Table 9: Evaluation results on some of the Event classes with the lowest data samples based on BS.

	$\mathcal{G}$		$\mathcal{G} + \mathcal{S}_{1200}$	
Classes	#Samples	F1-score	#Samples	F1-score
HEIGHT	4	26.81	14	**45.56**
LATITUDE	3	41.83	4	**53.66**
LENGTH	4	23.11	13	**47.79**
LONGITUDE	5	41.83	5	**42.15**
MATERIAL_REFERENCE	14	47.36	31	**54.94**
QUANTITY_MIN	20	**46.72**	76	42.77
TIME_MAX	11	**42.81**	12	41.46
TIME_MIN	28	25.43	33	**30.20**
WEIGHT	15	74.42	24	**84.96**
WIDTH	4	4.66	11	**11.39**

Table 10: Evaluation results on some of the Attribute classes with the lowest data samples based on BS.

further research and experimentation are required to fully realize and validate these possibilities.

7 Limitation

One of the limitations of this study is that we only generated new data based on the original data's Knowledge Graphs, which led to low diversity in the dataset. Future work could involve experimenting with modified Knowledge Graphs to enhance diversity. We acknowledge that the annotations produced by our framework are far from perfect and require further enhancements. One potential improvement could be the use of a dependency tree to identify co-references and annotate them. It could also be used to extract relations between entities. Another path for improvement could be the use of attention weights from the generated texts. This could help identify the evidence of relations by pinpointing where the head and tail entities most attentively interact within the texts.

References

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *ArXiv*, abs/2005.14165.

V. Claveau, Antoine Chaffin, and Ewa Kijak. 2021. Generating artificial texts as substitution or complement of training data. *ArXiv*, abs/2110.13016.

Claude Coulombe. 2018. Text data augmentation made simple by leveraging nlp cloud apis. *ArXiv*, abs/1812.04718.

Mathieu Dehouck and Carlos Gómez-Rodríguez. 2020. Data augmentation via subtree swapping for dependency parsing of low-resource languages. In *International Conference on Computational Linguistics*.

Xiang Deng and Huan Sun. 2019. Leveraging 2-hop distant supervision from table entity pairs for relation extraction. *ArXiv*, abs/1909.06007.

A. R. Fabbri, Simeng Han, Haoyuan Li, Haoran Li, Marjan Ghazvininejad, Shafiq R. Joty, Dragomir R. Radev, and Yashar Mehdad. 2020. Improving zero and few-shot abstractive summarization with intermediate fine-tuning and data augmentation. In *North American Chapter of the Association for Computational Linguistics*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Sosuke Kobayashi. 2018. Contextual augmentation: Data augmentation by words with paradigmatic relations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 452–457, New Orleans, Louisiana. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandara Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Kuttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *ArXiv*, abs/2005.11401.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. What makes good in-context examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, Dublin, Ireland and Online. Association for Computational Linguistics.

Youmi Ma, An Wang, and Naoaki Okazaki. 2023. DREEAM: Guiding attention with evidence for improving document-level relation extraction. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1971–1983, Dubrovnik, Croatia. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *International Conference on Learning Representations*.

George A. Miller. 1995. Wordnet: a lexical database for english. *Commun. ACM*, 38(11):39–41.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Conference on Empirical Methods in Natural Language Processing*.

Maxime Prieur, Cédric du Mouza, Guillaume Gadek, and Bruno Grilheres. 2024. Shadowfax: Harnessing textual knowledge base population. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024, Wshington, USA, July 23-27, 2023*, Washington, USA. ACM.

Roland Roller, Eneko Agirre, Aitor Soroa Etxabe, and Mark Stevenson. 2015. Improving distant supervision using inference learning. *ArXiv*, abs/1509.03739.

Gilles Sérasset. 2015. Dbnary: Wiktionary as a lemon-based multilingual lexical resource in rdf. *Semantic Web*, 6:355–361.

Richard Shin, Christopher Lin, Sam Thomson, Charles Chen, Subhro Roy, Emmanouil Antonios Platanios, Adam Pauls, Dan Klein, Jason Eisner, and Benjamin Van Durme. 2021. Constrained language models yield few-shot semantic parsers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7699–7715, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zhen Wan, Fei Cheng, Zhuoyuan Mao, Qianying Liu, Haiyue Song, Jiwei Li, and Sadao Kurohashi. 2023. Gpt-re: In-context learning for relation extraction using large language models. *ArXiv*, abs/2305.02105.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Huai hsin Chi, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *ArXiv*, abs/2203.11171.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Huai hsin Chi, F. Xia, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *ArXiv*, abs/2201.11903.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388, Hong Kong, China. Association for Computational Linguistics.

Ziang Xie, Sida I. Wang, Jiwei Li, Daniel Lévy, Allen Nie, Dan Jurafsky, and A. Ng. 2017. Data noising as smoothing in neural network language models. *ArXiv*, abs/1703.02573.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *ArXiv*, abs/2305.10601.

Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. 2019. DocRED: A large-scale document-level relation extraction dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 764–777, Florence, Italy. Association for Computational Linguistics.

Danqing Zhang, Tao Li, Haiyang Zhang, and Bing Yin. 2020. On data augmentation for extreme multi-label classification.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12697–12706. PMLR.

Enwei Zhu and Jinpeng Li. 2022. Boundary smoothing for named entity recognition. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7096–7108, Dublin, Ireland. Association for Computational Linguistics.

A Example Appendix

Examples of text comparison between $\mathcal{G}$ and $\mathcal{S}$ are provided in Table 11 for English and Table 12 for French. Figure 4 illustrates a sample of Knowledge Graph K_t along with its corresponding gold text t and synthetic text t'. An example of an entity extraction prompt can be seen in Figure 5.

Figure 4: Sample of a knowledge graph K_t with its t and t'.

Figure 5: Example of prompt used for entity extraction.

Texts from $\mathcal{G}$	Texts from $\mathcal{S}$
Pacific Fair is a major shopping centre in Broadbeach Waters on the Gold Coast, Queensland, Australia. It was Queensland's largest regional shopping centre until 2006. Pacific Fair was developed by Hooker Retail Developments and opened in 1977 on what was swampland with 96 specialty stores and two anchor tenants. Since then, Pacific Fair has undergone numerous expansions and has grown to have more than 300 specialty stores and four anchor tenants. In January 2014, work began on a major redevelopment project to meet the predicted regional growth on the Gold Coast. Prior to the redevelopment, the shopping centre had four main major stores including a four-level Myer, Kmart, Target, Coles and Toys 'R' Us. Daimaru operated in the centre before its Australian withdrawal, albeit briefly. It also had a 12-screen Birch Carroll and Coyle Cinema (re-opened as Event Cinemas in late 2015). Pacific Fair is a major public transport interchange on the Gold Coast, serviced by Surfside Buslines, and the Broadbeach South station located not far from the shopping centre. Nearby is The Star Gold Coast and Gold Coast Convention & Exhibition Centre. Pacific Fair fronts Little Tallebudgera Creek and is the southern end of the Surfers Riverwalk. This shopping centre was featured on the third season of The Mole.	Located in the state of Queensland, Australia, Surfers Riverwalk is a scenic walkway along the banks of the Nerang River in the city of Gold Coast. This administrative territorial entity falls under the larger jurisdiction of Queensland. The Gold Coast, also situated in Queensland, is a popular tourist destination and is known for its stunning beaches, theme parks, and shopping complexes like Pacific Fair, which was established in 1977. Some well-known retail chains operating in the area include Toys 'R' Us, Myer, Kmart, Coles, and Daimaru. Transportation options include Surfside Buslines and the Broadbeach South train station. Nearby attractions include Little Tallebudgera Creek and the Gold Coast Convention & Exhibition Centre. Surfers Riverwalk offers views of both the river and the beachfront, making it a popular spot for locals and tourists alike. Event Cinemas, another prominent establishment in the area, provides entertainment options for moviegoers. Overall, the Gold Coast boasts a diverse range of businesses and amenities catering to the needs of its residents and visitors.
Lark Force was an Australian Army formation established in March 1941 during World War II for service in New Britain and New Ireland. Under the command of Lieutenant Colonel John Scanlan, it was raised in Australia and deployed to Rabaul and Kavieng, aboard SS Katoomba, MV Neptuna and HMAT Zealandia, to defend their strategically important harbours and airfields . The objective of the force, was to maintain a forward air observation line as long as possible and to make the enemy fight for this line rather than abandon it at the first threat as the force was considered too small to withstand any invasion. Most of Lark Force was captured by the Imperial Japanese Army after Rabaul and Kavieng were captured in January 1942. The officers of Lark Force were transported to Japan, however the NCOs and men were unfortunately torpedoed by the USS Sturgeon while being transported aboard the Montevideo Maru. Only a handful of the Japanese crew were rescued, with none of the between 1,050 and 1,053 prisoners aboard surviving as they were still locked below deck.	During World War II, the Australian Army played a significant role, with over 330,000 men serving overseas. Among them was John Scanlan, a soldier who served in various campaigns in the Pacific theater. Born in Australia, Scanlan enlisted in the army in 1940 and was assigned to the Lark Force, a specialized unit tasked with defending Singapore against Japanese invasion. In March 1941, the Lark Force was formally established, and Scanlan was among its founding members. The force, composed primarily of Australian infantry, artillery, and engineering units, operated mainly in Malaya and Singapore. Its primary mission was to provide defensive cover against any potential Japanese threat in the region. As the war intensified, the Lark Force found itself facing increasingly formidable opponents. The Imperial Japanese Army, which had already conquered much of Southeast Asia, launched a massive assault on Singapore in February 1942. The Lark Force, along with other Allied forces, put up a fierce resistance, but ultimately they were overwhelmed. Thousands of soldiers, including Scanlan, were captured and sent to prisoner-of-war camps in Japan. Many never returned home.

Table 11: Samples of English Gold and Synthetic texts. Highlighted words are the entities and words in Red are the entities that did not appear in the original text t.

Texts from $\mathcal{G}$	**Texts from $\mathcal{S}$**
Le 19 septembre 2020, un mouvement de grève a perturbé les cours au sein de l'Université du Michigan. En effet, les enseignants de l'université ont décidé de fermer les portes de toutes les salles d'étude. Les enseignants reprochent au gouvernement d'avoir ignoré leurs relances concernant le non-respect des termes du contrat signé entre les deux parties. Après avoir été annoncée sur les réseaux sociaux, la grève a entraîné une confrontation entre les enseignants et les forces de l'ordre. Monsieur Mehmet-akhif Hassan est mort après avoir été brutalement frappé sur la tête avec une matraque durant la confrontation. Il a été évacué en urgence dans une ambulance pour être conduit à l'hôpital, mais les secouristes ont déclaré son décès en route.	Le 19 septembre 2020, une grève des enseignants au Michigan a provoqué une confrontation entre ces derniers et les forces de l'ordre, entraînant également une crise dans le gouvernement. La confrontation a débouché sur un accident de la route, où Mehmet-akhif Hassan, un citoyen présent à l'endroit des événements, a été grièvement blessé. L'accident a également causé divers dégâts matériels, notamment la perte d'un ambulance. Les secouristes ont rapidement pris en charge Mehmet-akhif Hassan et l'ont emmené à l'hôpital. Les enseignants, qui avaient par ailleurs bloqué les portes de l'université, ont également protesté devant cet incident, affirmant qu'il était lié à leur conflit avec la police et le gouvernement. Ces incidents ont également affecté les relations entre les différents groupes impliqués, en particulier celles entre les enseignants et les forces de l'ordre, ainsi que celles entre les enseignants et le gouvernement.
Naomi Zhang, la multimilliardaire et fondatrice de l'entreprise "BOMO", était à la tête de l'entreprise numéro 1 dans la distribution d'alcool en Argentine. Un de ses employés, Monsieur Thibault Alekseï, d'origine serbe, était le chef d'un gang de vente d'armes à feu. Il utilisait sa voiture de fonction pour livrer de la drogue et de la cocaïne à plusieurs entreprises. Il avait livré des dizaines déntreprises, dont "POLIS". Lorsque la nouvelle avait été annoncée dans la presse, l'entreprise de Naomi Zhang avait été au cœur dún scandale médiatique et avait vu son nombre de commandes diminuer.	Naomi Zhang est une femme d'affaires influente basée en Argentine. Elle est fondatrice de BOMO, une organisation non gouvernementale qui travaille activement pour améliorer la situation de nombreuses personnes au sein du pays. Thibault Alekseï, quant à lui, est un criminel serbe vivant également en Argentine. Il est impliqué dans la distribution illégale de drogues et fait partie d'une bande organisée appelée gang. Naomi Zhang et Thibault Alekseï sont tous deux présents en Argentine, où ils exercent leur influence respective sur divers secteurs de la vie quotidienne.

Table 12: Samples of French Gold and Synthetic texts. Highlighted words are the entities and words in Red are the entities that did not appear in the original text t. Certain entities may not have been included in K_t because they have no relationship to any other entity, which would explain why they were present in t but absent from t'.