Ontology-Guided Synthetic Data Generation for Low-Resource Information Extraction: A Case Study in IT Heritage Domain

Nakanyseth Vuth[1]⋆ **Emrick Poncet**[1]⋆ **Benjamin Lecouteux**[1]
Caroline Djambian[2] **Didier Schwab**[1] **Gilles Sérasset**[1]
[1]Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG
[2]Univ. Grenoble Alpes, GRESEC
38000 Grenoble, France
{first.last}@univ-grenoble-alpes.fr

Abstract

Information Extraction (IE) in specialized domains often suffers from a severe cold-start problem due to the high cost of expert annotation. Recent Reverse-IE approaches leverage knowledge graphs to generate synthetic training corpora, but typically assume the availability of an existing knowledge base. In this work, we propose an ontology-driven pipeline for synthetic supervision that removes this requirement. Starting from a formal domain ontology, we introduce a stochastic motif sampling strategy that constructs schema-consistent Knowledge Graph structures with controllable topology, which are then verbalized into natural language. This ontology-first formulation also allows direct control over the data generation process, enabling oversampling of underrepresented entity types or relation patterns. Applied to the IT Heritage domain, our approach produces a fully labeled NER/RE corpus without large-scale manual annotation. Evaluation in a low-resource setting shows that while the synthetic corpus lacks the linguistic diversity of gold data, its scalability produces training sets large enough to alleviate the cold-start problem, making ontology-guided motif generation a practical strategy for domains where gold annotation is limited.

Keywords: synthetic data, ontology-guided data generation, knowledge-graph motifs, information extraction, named entity recognition, relation extraction, IT heritage

1. Introduction

Information Extraction (IE) tasks, including Named Entity Recognition (NER) and Relation Extraction (RE), have traditionally relied on large-scale human-annotated corpora such as TACRED (Zhang et al., 2017) and DocRED (Yao et al., 2019). However, in specialized or low-resource domains, such as cultural heritage, legal studies, or biomedicine. Acquiring expert annotations at scale is prohibitively expensive and time-consuming. This results in a severe cold-start problem, where modern deep learning models cannot be effectively trained due to insufficient supervised data.

To alleviate the burden of manual annotation, prior work has explored Distant Supervision (Mintz et al., 2009; Quirk and Poon, 2016; Ratner et al., 2017; Peng et al., 2019), which heuristically aligns structured Knowledge Bases (KBs) with raw text. Although scalable, this paradigm is well known to introduce substantial label noise (Meng et al., 2021; Zhou et al., 2022). More recently, the emergence of Large Language Models (LLMs) has shifted attention toward **synthetic data generation** (Josifoski et al., 2023; Vuth et al., 2024). This so-called Reverse IE paradigm inverts the traditional extraction process: instead of extracting structure from text, it begins with structured triples or Knowledge Graphs (KGs) and prompts an LLM to generate corresponding natural language documents.

In this paper, we conduct a preliminary study by adopting and extending the Reverse IE paradigm to settings where even a pre-existing KB is unavailable. Rather than relying on existing knowledge bases as a source of triples, we begin with a formal *ontology* that defines the domain schema and its semantic constraints. **This ontology-first approach provides a key advantage: structural controllability**. Because entity types and relation constraints are explicitly defined, we can stochastically sample KG motifs that are guaranteed to respect the schema while remaining structurally diverse. More importantly, the sampling process can be steered toward specific supervision needs. For instance, if certain entity types or relation patterns are underrepresented in the training data, the motif distribution can be adjusted to deliberately increase their frequency. We therefore introduce a stochastic sampling strategy to generate abstract and structurally diverse KG motifs directly from the ontology. These motifs are subsequently instantiated with domain-specific entities and verbalized into natural language using the KGAST framework (Vuth et al., 2024).

We apply this methodology to the specialized *IT Heritage* domain and conduct both topological and linguistic analyses of the resulting corpus. Finally, we evaluate its downstream effectiveness for

⋆ Equal contribution.

Proceedings of the Knowledge Graphs and Large Language Models Workshop 2026 (KG-LLM@LREC26), pages 73–81
May 16, 2026.

IE tasks, demonstrating that ontology-driven synthetic generation provides a scalable alternative for bootstrapping models in data-scarce domains.

2. IT Heritage Ontology

Our use case relies on *IT in Heritage* (Djambian et al., 2024), an ontology designed for the museum documentation of information-technology heritage. It models IT heritage as a combination of (1) **physical artefacts** (e.g., computers, peripherals, storage media, and other technical components), (2) **information objects** associated with these artefacts (e.g., manuals, catalogues, photographs, and other documentary resources), and (3) **context entities** used in cultural-heritage description such as *actors* (persons and organisations), *places*, *dates/time-spans*, *types/materials*, and *rights*. The ontology also represents **heritage events** (e.g., creation/production, modification, conservation states, and beginning/end of existence) to capture provenance and temporal aspects.

The complete ontology contains **329 classes** and **86 object properties** (relations), reflecting a rich taxonomy of IT-related object types and documentation patterns.

For our experiments, we use a reduced schema consisting of **18 core classes** (entity types) and **15 properties** (relations). This reduction keeps the main upper-level classes needed to represent the domain and does not reduce the overall coverage, except for classes considered too specific or not relevant to our use case. In practice, specialised subclasses are replaced by their more general parent classes, which keeps the model simpler while preserving the essential concepts required by our pipeline. This reduced core therefore maintains the ability to represent the essential domain concepts required by our pipeline (*artefact/document/context/provenance*), while limiting modelling complexity and keeping extraction and validation steps tractable.

3. Methodology

We treat the ontology as a formal set of rules and semantic constraints that define a given domain. Because these constraints explicitly characterize how entities and relations interact, they can be leveraged to generate an unlimited number of synthetic Knowledge Graphs (KGs). Since each generated graph follows the ontology's internal logic, the resulting structures are inherently consistent and semantically meaningful. We refer to these structured graphs as KG Motifs, which act as blueprints for subsequent text generation. Building on this formulation, we frame synthetic data generation as a two-stage pipeline: (1) the synthesis of a structured KG Motif from the ontology, and (2) the verbalization of this motif into natural language.

3.1. Task Formalization

Let $\mathcal{O} = (\mathcal{T}, \mathcal{R}, \mathcal{C})$ be an Ontology, where $\mathcal{T}$ represents a set of entity types, $\mathcal{R}$ denotes the set of relation predicates, and $\mathcal{C}$ specifies the cardinality and domain constraints. A Knowledge Graph $\mathcal{G}$ is defined as a set of triples $\mathcal{E} = \{t_1, t_2, ..., t_n\}$, where each triple $t = (h, r, t)$ consists of a head entity h, a tail entity t, and a directed relation $r \in \mathcal{R}$ such that the types of h and t satisfy the constraints in $\mathcal{C}$.

In our method, we introduce the concept of a **KG Motif**. A Motif is a sub-graph structure generated from $\mathcal{O}$ that serves as a semantic blueprint for a document. Unlike disjoint triples, a motif captures topological patterns that reflect real-world information.

3.2. Ontology-to-KG Motif Generation

Let a Motif be defined as a directed graph $\mathcal{M} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is the set of entity nodes. The generation of a motif is a stochastic expansion process defined by the following steps:

1. **Anchor Initialization:** The process begins by initializing an empty node set $\mathcal{V} = \emptyset$ and triple set $\mathcal{E} = \emptyset$. We then sample an anchor entity type τ_{anchor} from the ontology. To ensure the Motif can grow, we prioritize growable types $\mathcal{T}_{grow} \subseteq \mathcal{T}$ that possess at least one outgoing relation in $\mathcal{R}$, such that $\tau_{anchor} \sim P(\tau \mid \tau \in \mathcal{T}_{grow})$. An initial node v_0 of type τ_{anchor} is generated and added to $\mathcal{V}$.

2. **Degree Sampling:** For each node $v \in \mathcal{V}$ of type τ_v, we determine the number of outgoing edges k_v by sampling from a Poisson distribution:
$$k_v \sim \text{Poisson}(\lambda) \quad (1)$$
This parameter λ controls the information density of the resulting motif.

3. **Probabilistic Relation Sampling:** For each of the k_v edges, we select a relation $r \in \mathcal{R}$ according to a conditional probability distribution $P(r \mid \tau_v)$. In a cold-start scenario, we assume a uniform distribution over the set of valid relations $\mathcal{R}_{\tau_v} = \{r \in \mathcal{R} \mid \text{head}(r) = \tau_v\}$. However, our method supports the integration of empirical priors derived from real-world datasets to enhance domain realism.

4. **Topology Control:** To prevent the generation of simplistic star graphs where all edges are sampled from a single center, we implement an entity re-use mechanism. Let $\tau_{tail} = \text{tail}(r)$ be the required target type, and $\mathcal{V}_{cand} = \{u \in$

$\mathcal{V} \mid \text{type}(u) = \tau_{tail}\}$ be the set of existing valid entities. When instantiating a tail entity u for a relation r, we select an entity according to the re-use probability α:

$$u \sim \begin{cases} \text{Uniform}(\mathcal{V}_{cand}) & \text{with probability } \alpha \\ u_{new} \notin \mathcal{V} & \text{with probability } 1-\alpha \end{cases} \quad (2)$$

This allows for the formation of cycles and multi-hop paths.

5. **Entity Injection:** Once the motif structure $\mathcal{M}$ is complete, we perform entity injection using a mapping function $\phi : \mathcal{V} \to \mathcal{S}$. Abstract IDs (e.g., `Person_0`) are mapped to real-world surface forms $s \in \mathcal{S}$ (e.g., *Alan Turing*) sampled from a pre-defined entity pool. This produces a realized KG, denoted as $\mathcal{M}^* = \{(\phi(h), r, \phi(t)) \mid (h, r, t) \in \mathcal{E}\}$, ready for verbalization.

3.3. Data Generation

To transform the realized KG Motifs $\mathcal{M}^*$ into natural language, we adopt the KGAST framework introduced by (Vuth et al., 2024). This process utilizes an LLM as a controlled verbalizer. The LLM is prompted in a few-shot manner to incorporate every triple from the KG into a fluid, coherent narrative T.

Crucially, this framework ensures the alignment between the input KG and the generated text. By mapping the original entities back to their span positions in the synthesized document, we produce a fully annotated dataset pair $(T, \mathcal{M}^*)$ without the need for manual intervention. This Reverse IE approach ensures that the labels are **consistent by construction**, as the target annotations are derived directly from the generative seed. Some examples of the generated data can be seen in Table 1.

4. Experimental Setup

4.1. Building the Baseline Data (Pseudo-Gold)

As this is a preliminary study in a niche domain, we lack a large-scale, expert-annotated gold dataset for baseline comparison. To address this, we constructed a **Pseudo-Gold** baseline by selecting 100 high-quality documents from the IT Heritage corpus using an LLM-as-judge reranking paradigm (Zheng et al., 2023) [1]. We then utilized an automated extractor to generate RDF-style triples for these documents, constrained by our defined schema (Section 2).

To ensure this baseline is of sufficient quality, we manually annotated 10 samples to validate the automated extractor. The extractor achieved a high recall (0.991) but a lower precision (0.745). However, with an overall micro-F_1 of 0.85, we consider this pseudo-gold baseline a functional starting point for evaluation.

4.2. Synthetic Data Generation

We generated three sets of 3,000 KG motifs to observe the effect of the re-use probability $\alpha \in \{0.3, 0.5, 0.7\}$. Each motif was generated with a target size of $N = 8$ and a density parameter $\lambda = 2$. Note that N represents a target size rather than a strict bound; the actual motif size may exceed 8, depending on the sampling process. Following the structural analysis in Section 5.1, we selected the $\alpha = 0.7$ set for verbalization, as it provided the most balanced connectivity for the target domain. We utilized `Mistral-Small-3.1-24B-Instruct` [2] as our verbalizer to produce the final synthetic corpus. To ensure the quality of the Reverse IE process, we evaluated the **verbalization fidelity**, the degree to which the LLM actually incorporated the provided triples and entities into the final narrative. By mapping the input KG motifs back to the synthesized text, we found that the corpus retained **94.63**% of the input entities and **93.45**% of the input triples. This high coverage confirms that the model successfully grounded the majority of the structural facts, though the ∼6.5% loss rate indicates a minor risk of lost triples that exist in the labels but were omitted or incorrectly phrased in the text. Table 2 provides a comparative analysis of the resulting datasets.

4.3. Downstream Tasks

We evaluate the utility of our synthetic data on two core Information Extraction tasks: Named Entity Recognition (NER) and Relation Extraction (RE). For both tasks, we fine-tune a `bert-base-uncased` model.

Our experimental setup follows a consistent data-splitting protocol across both tasks. The 100 gold samples are divided into 80 training samples and 20 held-out samples for evaluation. We compare four training scenarios: (1) **Gold**, using only the 80 gold training samples; (2) **Synthetic**, using the full synthetic dataset; (3) **Synthetic** (Random), a randomly sampled 50% subset of the synthetic data; and (4) **Synthetic** (Strat.), a stratified 50% subset constructed to better preserve entity distribution balance. All models are evaluated exclusively on the 20-sample gold test set. We report micro/macro

[1] The LLM model we use: https://docs.mistral.ai/models/mistral-large-3-25-12

[2] https://huggingface.co/mistralai/Mistral-Small-3.1-24B-Instruct-2503

KG Motif	Realized KG	Generated Text
(VideogameConsole_0, in custody of, Museum_0) (Museum_0, resides in, Country_0) (Maker_0, resides in, City_0) (VideogameConsole_0, produced by, Maker_0) (Maker_0, resides in, Country_0) (VideogameConsole_0, in custody of, Maker_0) (VideogameConsole_0, owned by, Museum_0) (VideogameConsole_0, start date, Date_0)	(Famicom (Family computer) console, in custody of, Heinz Nixdorf MuseumsForum) (Heinz Nixdorf MuseumsForum, resides in, United Kingdom) (Government Code and Ciphers School, resides in, Bletchley Park) (Famicom (Family computer) console, produced by, Government Code and Ciphers School) (Government Code and Ciphers School, resides in, United Kingdom) (Famicom (Family computer) console, in custody of, Government Code and Ciphers School) (Famicom (Family computer) console, owned by, Heinz Nixdorf MuseumsForum) (Famicom (Family computer) console, start date, 2018)	The Famicom (Family computer) console, a pioneering videogame console, has an unusual custodial history. Although it was produced by the Government Code and Ciphers School, an entity based in Bletchley Park, United Kingdom, it is not currently in their custody. Instead, the console is owned and held by the Heinz Nixdorf MuseumsForum, which is located in the United Kingdom. The console's journey into the museum's collection began in 2018, marking the start of its tenure under the museum's care. Despite its origins with the Government Code and Ciphers School, the Famicom console's story is now intertwined with that of the Heinz Nixdorf MuseumsForum, which preserves and showcases its historical significance.
(Computer_0, located in, Country_0) (CipherMachine_0, has identifier, Identifier_0) (Computer_0, is related to, CipherMachine_0) (Computer_0, located in, City_0) (Maker_0, resides in, City_0) (CipherMachine_0, owned by, Museum_0) (CipherMachine_0, produced by, Maker_0) (Computer_0, located in, Country_1) (CipherMachine_0, has title, Title_0) (Computer_0, has identifier, Identifier_0)	(IBM 1410, located in, British) (Swiss-K, has identifier, arXiv:1706.03762) (IBM 1410, is related to, Swiss-K) (IBM 1410, located in, Paderborn) (Domark, resides in, Paderborn) (Swiss-K, owned by, The Seattle Computer Museum) (Swiss-K, produced by, Domark) (IBM 1410, located in, Cheshire) (Swiss-K, has title, Golf) (IBM 1410, has identifier, arXiv:1706.03762)	The IBM 1410, a notable transistors and batch systems computer, has a complex history of locations and associations. Initially situated in Britain, it later found its way to Paderborn, a city where the maker Domark also resides. Interestingly, the IBM 1410 shares a connection with the Swiss-K cipher machine, which bears the same identifier, arXiv:1706.03762. The Swiss-K, known by the title "Golf," was produced by Domark and is currently owned by The Seattle Computer Museum. Despite its ties to Paderborn, the IBM 1410 has also been located in Cheshire, adding another layer to its intriguing journey.

Table 1: Examples mapping abstract KG Motifs to Realized KG instances and their corresponding generated texts. **The entity combinations in the Realized KGs are not necessarily factually accurate**. The objective of this generation process is to sample structurally valid and complex graphs for synthetic data generation, rather than to strictly reflect real-world facts.

Dataset	Tokens	Total Ent.	Total Triples	Avg Len	Density	Self-BLEU
Gold	15,413	854	815	192.7	10.68	0.408
Synthetic	476,139	36,143	24434	158.7	12.05	0.866
Synthetic (Random)	238,933	18,154	12293	159.2	12.09	0.825
Synthetic (Strat.)	238,198	18,016	12190	158.8	12.01	0.828

Table 2: Descriptive statistics of the gold and synthetic datasets. Density denotes the average number of entities per document. Self-BLEU (4-grams) here measures intra-dataset similarity (lower indicates higher diversity). For reference, the Self-BLEU of DocRED is 0.355.

F_1 scores across three random seeds. For RE, we employ strict triplet matching, requiring exact subject, predicate, and object correspondence.

5. Results and Analysis

5.1. Data Analysis

Topological Analysis We conduct a topological evaluation to determine if our generated KG motifs structurally resemble real-world knowledge representations. We compare our synthetic motifs against the domain-specific **IT Heritage** samples and the general-domain **DocRED** dataset (Yao et al., 2019). DocRED is included as a benchmark to see how our logic performs against a high-density, web-like dataset from the general domain.

The results in Table 3 highlight a significant structural gap in terms of average **clustering coefficient**, which measures the frequency of interconnected triangles between entities. While DocRED exhibits high clustering (0.1786), our synthetic motifs remain near zero (0.0148 at $\alpha = 0.3$). This difference is largely driven by the relational diversity of the underlying ontology. DocRED utilizes 96 distinct relations, which naturally allows entities to form many interconnected loops. In contrast, our simplified IT Heritage ontology contains only 15 relations, which restricts our method's ability to create these dense triangles and leads to a more linear structure of graphs. Despite the low clustering, our motifs show strong alignment with the gold IT Heritage data in terms of average degree, which represents the average number of relations linked to a single entity. At $\alpha = 0.7$, our motifs show 1.80 relations per entity, which closely aligns with the 1.94 found in the Gold IT samples. This indicates that while the overall shape of our graphs is more simplistic than human data, the volume of information per entity is realistic for the target domain. For comparison, DocRED's much higher average degree (3.01) reflects its higher complexity and larger number of triples per document.

It can also be seen that the synthetic motifs at $\alpha = 0.3$ use 10.71 nodes to describe only 8.80 triples, whereas the Gold data uses nearly a 1:1 ratio (10.50 nodes for 10.17 triples). Furthermore, the results show that increasing α from 0.3 to 0.7 increases the graph density, the ratio of actual connections to the maximum possible, from 0.0891 to 0.1078. However, even at $\alpha = 0.7$, the synthetic graphs still fail to match the density (0.1311) and clustering (0.0587) of the Gold IT samples. This confirms that while our method can control the volume of connections, the current sampling logic is

still more sparse and linear than real-world data.

Synthetic Data Analysis To examine the realism of the generated data, we conduct a linguistic analysis of the synthetic corpus. Specifically, we evaluate three complementary dimensions: intra-corpus diversity, information density, and syntactic complexity.

First, we observe in Table 2 a significantly high Self-BLEU score (0.866) in the synthetic data compared to the Gold IT (0.408) and DocRED (0.355). This indicates a high level of intra-dataset repetition. Consistent with the findings of Vuth et al. (2024), we find that strict grounding in a knowledge graph encourages LLMs to generate repetitive or structurally predictable syntactic patterns. However, the primary issue here is our **limited entity pool**. Because our pool is derived from the 100 gold samples, for example, the pool contains only 30 unique *City* and 49 *Videogame console* instances; the same entities appear across nearly 3,000 documents, naturally inflating the lexical overlap in the corpus. Future work must prioritize expanding the entity pool to break this repetitive cycle.

Second, the **Information Compression Ratio** (#triples / #sentences) reveals that the synthetic text is less information-dense, reflecting a tendency to verbalize triples in a more explicit and decomposed manner. As shown in Figure 1, DocRED averages 2.56 triples per sentence, while Gold IT reaches 1.91. Our synthetic data (Silver) lags behind at 1.41. The distribution for Silver is heavily peaked around 1.0, suggesting that the LLM often defaults to a **one fact per sentence** structure. This confirms that the synthetic corpus is **bland** compared to human writing; it fails to utilize the complex narrative structures (like relative clauses) that humans use to pack multiple facts into a single sentence.

Third, we measured **Syntactic Complexity** using Average Dependency Distance (ADD). While the gap is narrower here, the trend persists: DocRED (3.35) and Gold (3.22) both outperform the Synthetic data (3.07). While an ADD of 3.07 indicates that the LLM is producing grammatically correct and somewhat complex English, the leftward shift in the distribution confirms that the sentences are structurally simpler than those in the Gold baseline.

These results suggest a direct causal link between the input graph and the resulting text. **The structural sparsity we found in the topological analysis (low clustering) directly limits the LLM's ability to integrate information.** Because the KG Motifs are primarily linear chains rather than interconnected webs, the LLM is forced to write a sequence of simple, disjointed sentences. Consequently, the blandness of the corpus is not just a failure of the LLM's creativity, but a direct reflection of the **structural simplicity of the knowledge graph motifs**.

5.2. Downstream Task Results

5.2.1. Named Entity Recognition

The NER results, presented in Table 4, demonstrate the practical utility of our synthetic generation pipeline in a low-resource setting. The **Full Synthetic** paradigm ($N = 3,000$) achieved a Micro-F1 of 0.4691 and a Macro-F1 of 0.4724, significantly outperforming the **Gold** baseline ($N = 80$), which severely struggled with a Macro-F1 of just 0.1871.

However, this performance gap must be interpreted by the vast difference in training volume. The primary advantage demonstrated here is not that synthetic data is intrinsically superior to gold annotation per sample, but rather that our Ontology-to-KG pipeline provides a highly scalable mechanism to overcome the cold-start problem. The poor performance of the Gold baseline (Micro-F1 0.2772) highlights that 80 gold samples are simply insufficient to train a deep learning model on a complex, 18-class domain schema.

The subset experiments further clarify this scaling behavior. The **Synthetic Random** subset ($N = 1{,}500$) attains a Micro-F1 of 0.4314, preserving most of the performance gains while using only half the data. This suggests diminishing returns with increasing synthetic scale: the transition from 80 to 1,500 samples enables the model to learn core entity boundaries and class structure, whereas doubling the corpus to 3,000 samples yields comparatively modest improvements ($+0.03$ Micro-F1).

Moreover, the **Synthetic Stratified** subset ($N = 1{,}500$) slightly improves Macro-F1 over the random subset (0.3996 vs. 0.3834), which indicates that class-balanced generation is particularly beneficial for rare-entity recognition.

Overall, despite the syntactic regularity and elevated Self-BLEU discussed in Section 4.2, the structural grounding induced by our KG motifs provides a sufficiently strong supervision signal to bootstrap NER performance in domains where expert annotation is costly or scarce.

5.2.2. Relation Extraction

The RE results, presented in Table 5, demonstrate that the synthetic pipeline is also effective for relation extraction in a low-resource setting. Under strict triplet matching (case-insensitive), the **Gold** baseline ($N = 80$ train items) reaches only 0.1989 Micro-F1 and 0.2125 Macro-F1, while all synthetic regimes improve over this baseline. With **Full Synthetic** ($N = 3{,}000$, $\texttt{neg\_ratio} = 1.0$), performance rises to 0.2974 Micro-F1 and 0.3621 Macro-

Dataset	KG	Triples	Nodes	Clustering Coef.	Density	Avg Deg
Gold	100	10.17	10.50	0.0587	0.1311	1.94
Synthetic 0.3	3000	8.80	10.71	0.0148	0.0891	1.66
Synthetic 0.5	3000	8.72	10.19	0.0259	0.0984	1.73
Synthetic 0.7	3000	8.60	9.72	0.0406	0.1078	1.80
DocRED	3027	17.41	10.99	0.1786	0.1751	3.01

Table 3: Topological comparison between our synthetic KG motifs and existing datasets.

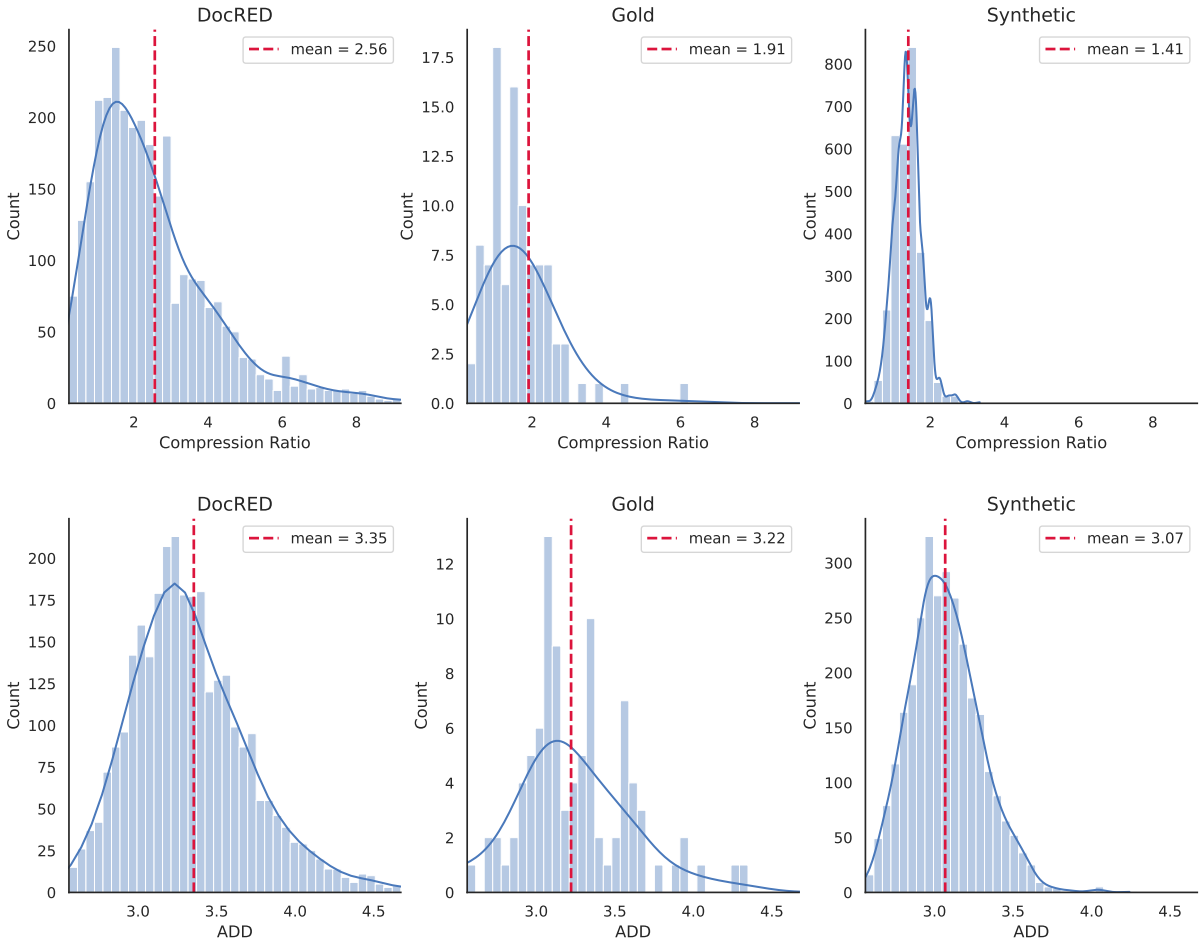

Figure 1: Distributions of Information Compression Ratio (top) and Average Dependency Distance (bottom) across the DocRED, Gold, and Silver datasets.

Dataset	Micro-F1	Macro-F1
Gold	0.2772 ± 0.01	0.1871 ± 0.01
Synthetic	**0.4691 ± 0.04**	**0.4724 ± 0.04**
Synthetic Random	0.4314 ± 0.02	0.3834 ± 0.02
Synthetic Stratified	0.4397 ± 0.02	0.3996 ± 0.04

Table 4: Named Entity Recognition results across different datasets.

F1, confirming a substantial gain in relational generalization.

As in NER, this gap should be interpreted primarily as a *scalability* effect rather than a per-sample superiority claim. The ontology-to-KG generation process provides much larger supervision volume than the limited gold set, which helps overcome cold-start learning for a multi-relation schema. In other words, the gold-only setting is data-starved for this task, and synthetic expansion provides the coverage needed to learn relation patterns beyond a small annotated subset.

The subset experiments further support this trend. **Synthetic Random** ($N = 1{,}500$) obtains 0.2565 Micro-F1 / 0.3508 Macro-F1, and **Synthetic Stratified** ($N = 1{,}500$) obtains 0.2640 Micro-F1 / 0.3247 Macro-F1. Thus, moving from gold-only to 1,500 synthetic examples already yields strong

gains, while increasing from 1,500 to 3,000 examples brings additional but smaller improvements in Micro-F1 (roughly $+0.03$ to $+0.04$, depending on the 1,500 subset), indicating diminishing returns with scale.

A key RE-specific finding is that performance is highly sensitive to false-positive calibration. Increasing negative sampling in the full synthetic regime ($\texttt{neg\_ratio} = 2.0$) gives the best overall Micro-F1 (0.3506 ± 0.04), compared with 0.2974 ± 0.03 at $\texttt{neg\_ratio} = 1.0$. This improvement is driven by a large precision increase (average FP reduced from 332.3 to 153.0), at the cost of lower recall. Hence, the final RE behavior is governed by a precision–recall trade-off: $\texttt{neg\_ratio} = 1.0$ favors recall and broader relation coverage, while $\texttt{neg\_ratio} = 2.0$ yields better calibrated triplet extraction and stronger strict Micro-F1.

Dataset	Micro-F1	Macro-F1
Gold	0.1989 ± 0.03	0.2125 ± 0.02
Synthetic	$\mathbf{0.3506 \pm 0.04}$	0.3505 ± 0.01
Synthetic Random	0.2565 ± 0.04	0.3508 ± 0.03
Synthetic Stratified	0.2640 ± 0.02	0.3247 ± 0.02

Table 5: Relation Extraction results across different datasets.

6. Related Work

Named Entity Recognition and Relation Extraction are foundational tasks in Information Extraction. Early approaches relied on feature-based statistical models (Lample et al., 2016), while recent systems leverage pretrained language models such as BERT for joint or pipeline-based extraction (Devlin et al., 2019). Large annotated corpora, including TACRED (Zhang et al., 2017) and DocRED (Yao et al., 2019), have enabled substantial progress in supervised settings.

However, these models are highly data-dependent and struggle in low-resource or domain-specific scenarios. To address annotation scarcity, distant supervision (Mintz et al., 2009; Quirk and Poon, 2016; Ratner et al., 2017; Peng et al., 2019) aligns existing knowledge bases with raw text to generate weak labels. Although scalable, distant supervision introduces substantial noise and often requires complex denoising mechanisms (Meng et al., 2021; Zhou et al., 2022). Our work targets the same data-scarcity problem but adopts a generative strategy rather than aligning existing corpora with structured knowledge.

Ontologies have long served as formal representations of domain knowledge, defining entity types, relation schemas, and logical constraints within the semantic web and knowledge engineering communities. In Information Extraction, ontologies are typically used as target schemas for annotation or as background constraints during inference. In contrast, our approach uses the ontology as a *generative prior*. Rather than using it solely as a constraint on extracted outputs, we sample abstract Knowledge Graph motifs directly from ontological rules before any text is generated. This shifts the ontology from a passive representational artifact to an active structural design mechanism, enabling controllable generation of schema-consistent supervision.

Recent work has explored the use of LLMs to generate synthetic training data for Information Extraction through a Reverse-IE paradigm: rather than extracting structured information from text, the process is inverted, structured knowledge graph triples are first sampled from an existing knowledge base and then verbalized into natural language to construct synthetic corpora (Josifoski et al., 2023; Vuth et al., 2024). While these approaches demonstrate the viability of synthetic supervision in low-resource settings, they rely on the availability of a pre-existing knowledge base, an assumption that does not hold in highly specialized or new domains where such resources are absent.

Our work builds on this Reverse-IE paradigm but differs in one key aspect: we do not assume the existence of a pre-constructed knowledge base as the source of triples. Instead, we generate abstract KG motifs directly from an ontology via stochastic sampling. This ontology-driven motif construction provides structural controllability, allowing the generation process to be explicitly steered toward underrepresented entity types or relation patterns. In this sense, our method complements prior LLM-based synthetic generation approaches by introducing a schema-grounded and topology-aware sampling stage prior to text realization.

7. Conclusion

We presented an ontology-driven pipeline for generating synthetic data in low-resource Information Extraction. Starting from a formal domain ontology, we introduced a stochastic sampling strategy to construct structured Knowledge Graph Motifs and subsequently verbalized them into natural language through a Reverse-IE framework. This process enables the automatic creation of schema-consistent Named Entity Recognition and Relation Extraction training data without relying on an existing knowledge base or large-scale manual annotation.

Applied to the IT Heritage domain, our ontology-driven pipeline produces a fully labeled synthetic corpus with controllable structural properties. While the generated text exhibits reduced lexical diver-

sity and structural regularity compared to human-authored corpora, it nonetheless provides a strong supervision signal in cold-start NER experiments, substantially outperforming a sparse gold-only baseline. We view this approach as a step toward principled synthetic supervision for niche domains, where expert knowledge can be encoded structurally before it is ever expressed linguistically.

Limitations

Despite the performance gains observed in our downstream tasks, several factors constrain the strength of our current claims. A primary issue with validity is our **pseudo-gold baseline**. Because our evaluation relies on a resource built via automatic extraction and validated on only 10 manual samples, the results may not fully reflect the complexities of expert-level annotation.

Furthermore, we observe a significant **lack of diversity** in the generated corpus. The high Self-BLEU scores indicate that the synthetic text is highly repetitive, a problem exacerbated by a small entity pool and motif structures that remain structurally sparse compared to human data. This topological simplicity, characterized by a lack of interconnected loops, leads the LLM toward a one fact per sentence realization style. Consequently, the synthetic data exhibits lower information compression and less varied syntax than natural documentation.

Future work will focus on addressing these issues along three axes. (*1*) **Stronger evaluation:** Building a larger, genuinely gold benchmark with broader entity/relation coverage and expert validation, for more reliable comparisons and error analysis. (*2*) **Stronger ontology constraints and sampling:** Enforcing richer ontological constraints and introducing motif-level objectives that explicitly target higher density and clustering to better approximate real-world discourse graphs. (*3*) **Diversity at scale:** Expanding the entity pool (including curated domain lexicons and larger catalog sources) and adopting distribution-aware sampling over types and predicates to reduce repetition and improve long-tail coverage.

Ethical Statement

The Knowledge Graphs and texts produced by our pipeline are entirely synthetic and are not guaranteed to be factually accurate. Although real-world entity names may be instantiated during generation, the relations expressed between them may not reflect true historical or factual associations.

The intended purpose of this framework is to provide structured supervision for Information Extraction models under a predefined domain schema. It is not designed to function as a factual knowledge base or authoritative historical resource. We emphasize that human oversight remains necessary to ensure factual reliability in specialized domains.

8. Bibliographical References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*.

Caroline Djambian, Micaela Rossi, Giada D'Ippolito, Emrick Poncet, and Pierre Maret. 2024. New terminological approaches for new heritages and corpora: The ITinHeritage project. In *Proceedings of the 3rd International Conference on Multilingual Digital Terminology Today (MDTT 2024), Granada, Spain, June 27-28, 2024*, volume 3703 of *CEUR Workshop Proceedings*. CEUR-WS.org. HAL Id: hal-04604833. Open-access version: https://hal.univ-grenoble-alpes.fr/hal-04604833/document.

Martin Josifoski, Marija Sakota, Maxime Peyrard, and Robert West. 2023. Exploiting asymmetry for synthetic training data generation: SynthIE and the case of information extraction. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1555–1574, Singapore. Association for Computational Linguistics.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *North American Chapter of the Association for Computational Linguistics*.

Yu Meng, Yunyi Zhang, Jiaxin Huang, Xuan Wang, Yu Zhang, Heng Ji, and Jiawei Han. 2021. Distantly-supervised named entity recognition with noise-robust learning and language model augmented self-training. *ArXiv*, abs/2109.05003.

Mike Mintz, Steven Bills, Rion Snow, and Daniel Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1003–1011, Suntec, Singapore. Association for Computational Linguistics.

Minlong Peng, Xiaoyu Xing, Qi Zhang, Jinlan Fu, and Xuanjing Huang. 2019. Distantly supervised named entity recognition using positive-unlabeled learning. In *Proceedings of the 57th*

Annual Meeting of the Association for Computational Linguistics, pages 2409–2419, Florence, Italy. Association for Computational Linguistics.

Chris Quirk and Hoifung Poon. 2016. Distant supervision for relation extraction beyond the sentence boundary. In *Conference of the European Chapter of the Association for Computational Linguistics*.

Alexander Ratner, Stephen H. Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. 2017. Snorkel: rapid training data creation with weak supervision. *Proceedings of the VLDB Endowment*, 11(3):269–282.

Nakanyseth Vuth, Gilles Sérasset, and Didier Schwab. 2024. KGAST: From knowledge graphs to annotated synthetic texts. In *Proceedings of the 1st Workshop on Knowledge Graphs and Large Language Models (KaLLM 2024)*, pages 43–55, Bangkok, Thailand. Association for Computational Linguistics.

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.

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. 2017. Position-aware attention and supervised data improve slot filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 35–45, Copenhagen, Denmark. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena.

Kang Zhou, Yuepei Li, and Qi Li. 2022. Distantly supervised named entity recognition via confidence-based multi-class positive and unlabeled learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7198–7211, Dublin, Ireland. Association for Computational Linguistics.