POPCORN: Fictional and Synthetic Intelligence Reports for Named Entity Recognition and Relation Extraction Tasks

Bastien Giordano, Maxime Prieur, Maxime Prieur, Nakanyseth Vuth, Sylvain Verdy, Kévin Couzot, Gilles Serasset, Guillaume Gadek, Didier Schwab, Cédric Lopez

Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 00 (2024) 000–000

Procedia Computer Science

www.elsevier.com/locate/procedia

28th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES 2024)

POPCORN: Fictional and Synthetic Intelligence Reports for Named Entity Recognition and Relation Extraction Tasks

Bastien Giordano*[a], Maxime Prieur*[b], Nakanyseth Vuth[c], Sylvain Verdy[a], Kévin Cousot[a], Gilles Sérasset[c], Guillaume Gadek[b], Didier Schwab[c], Cédric Lopez[a,**]

[a] *Emvista, 10, rue Louis Breguet, Jacou 34830, France*
[b] *Airbus Defence & Space, Elancourt 78990, France*
[c] *Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, Grenoble 38000, France*

Abstract

POPCORN is a research project aiming at maturing Information Extraction (IE) solutions for intelligence services. Due to defense security constraints, reports analyzed by intelligence services are not to be accessible to the scientific community. To address this challenge, we propose a dataset made of "fictional" (handcrafted) and "synthetic" (AI generated) French reports. Those synthetic reports are produced by an innovative approach that generates texts closely resembling real-world intelligence reports, facilitating the training and evaluation of IE tasks such as Entity and Relation Extraction. Experiments demonstrate the interest of synthetic reports to enhance the performance of IE models, showcasing their potential to augment real-world intelligence operations.

Keywords: Synthetic Data Generation; Dataset; Natural Language Processing; Large Language Models; Information Extraction

1. Introduction

POPCORN (Peuplement OPérationnel de COnnaissances et Réseaux Neuronaux) is a collaborative research project that aims to enhance Information Extraction (IE) technologies for intelligence services. These services receive numerous reports describing criminal activities worldwide. Analyzing all the information in its entirety is impossible for operational agents, and reports in natural language cannot be assimilated by intelligence information systems. Due to security concerns, these reports cannot be shared with the scientific community. Our project addresses this issue by creating datasets resembling real reports for training and evaluating IE models, focusing on Entity Recognition (ER) and Relation Extraction (RE).

* Equal contribution
** Corresponding author.
E-mail address: cedric.lopez@emvista.com

Our approach encompasses two distinct strategies. Firstly, we outline a meticulous process that consists in manually crafting and annotating "Fictional Reports". Secondly, we employ an automated technique to generate "Synthetic Reports" based on the previously crafted reports and their corresponding Knowledge Graph (KG). These KGs encapsulate the entities, attributes and relations present in the text and matching the POPCORN ontology. In our generation approach, KGs serve a dual purpose: as inputs for generating synthetic texts and as a means to ensure the inclusion of the required elements.

In this paper, we present the following key contributions:

- We introduce POPCORN, a new dataset comprising 2,000 French reports, of which 800 are manually written and annotated, while 1,200 reports are generated and annotated automatically. This dataset is specifically designed for training and evaluating IE models.
- We experiment with IE models with and without synthetic reports to evaluate the relevance of such a dataset.
- We publicly introduce the first models capable of processing several dozen classes of entities and relations for intelligence services.

2. Related works

Few datasets with a "Defense" thematic, designed for IE tasks, exist. Some of the resources available, such as DocRED [6] or DWIE [1] come close to such description. However, neither of them is suitable for our use case, the first containing Wikipedia abstracts with a restricted ontology and the second being a corpus of press articles. Since both are in English, some attempts to translate them have been made, such as the automatically translated dataset DWIE-FR [2] but at the expense of a decrease in the quality of annotations. Recent French IE challenges such as the EvalLLM2024 [1], demonstrate how topical this issue is.

Synthetic data refers to information generated through computer algorithms or AI. It proves particularly valuable in scenarios where training data is lacking or inaccessible due to confidentiality or compliance concerns. Numerous studies have explored the expansion of training data by introducing additional synthetic data [9, 10, 11]. These studies presented straightforward strategies, such as substituting words with their equivalent terms. These equivalents can be sourced from external repositories like WordNet [12], DBnary [13], or they can be calculated using word embedding models such as Word2Vec [14], and Glove [15]. 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. By employing Machine Translation models [16, 17], paraphrases of each sentence can be obtained through back-translation. While back-translation effectively amplifies the dataset size twofold, it introduces a notable challenge when token or character level annotation is required. 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.

3. Knowledge representation

Natural Language Processing (NLP) datasets [3, 1] have significantly contributed to the formalization of data representation and the introduction of modeling standards within the field. However, few proposals have concentrated on the security and defense topic, and are often limited to a small number of text samples, such as the Re3d dataset[2] made of 98 text samples. This section presents the ontology that shapes the annotation of POPCORN's 2000 documents.

3.1. Data model

We divide the elements of interest in the texts into the following three groups:

[1] https://evalllm2024.sciencesconf.org/
[2] https://github.com/dstl/re3d

- **Entities**: In our modeling, entities are the central elements. The dataset is based on the intelligence pentagram structure outlined in [4] (NATO STANAG-2433) and use five primary classes to form the basis of our ontology: Person, Organization, Event, Material and Place. The classes Person, Organization and Event are divided into sub-classes which, themselves, can have child classes.
- **Attributes**: Attributes provide additional information to Entities such as their age, size or job. They are values that can be shared by several entities. We distinguish Attributes from entities because, in our modelling, attributes are used to describe entities (similarly to properties graphs) when stored in a database. Attributes target, among others, temporal information, measurements, colors, quantities, or nationalities. A distinction is made between "Fuzzy", "Exact", "Min" or "Max" quantities and temporal information.
- **Relations**: Relations encompass all interactions expressed in a document, whether they imply two entities or an entity and an attribute. For each predicate, the set of subject entities, target entities or attribute types is defined.

The complete ontology includes a total of 35, 20 and 49 fine-grained entity, attribute and relation types. The POPCORN ontology thus extends the recurring classes of popular IE datasets with entity (MILITARY, CRIMINAL, etc) and relation (Has Consequence, Injured Number, etc) types specific to the security theme. The annotation of fine-grained classes introduces a high degree of imbalance (we measure an imbalance ratios of 4536 and 902.5 for the entity and relation types). For the sake of completeness, annotations on all the classes in the ontology have been kept, even classes with a low occurrence rate. Depending on the use case, users are free to discard low support classes. In addition, although only the child classes have been annotated, the hierarchical nature of the ontology means that child classes can be re-assigned the parent type in order to return to a sufficient occurrence level.The list, definitions and distribution of all classes can be found on the git repository of the dataset[3].

3.2. Proposed knowledge representation

Textual Knowledge base Population is a complex challenge made of various sub-tasks [5]. It involves processing raw documents to extract structured typed elements and their interactions. Within a text, entities are represented as clusters of co-referencing mentions (pronouns, noun phrases and named mentions referring the same "real word" entity) and is assigned one or more types from a given ontology. In some cases, such as for dates, attributes can also have a formatted value. A mention is defined by its textual value and offsets. All mentions within a cluster refer to the same entity. For instance, "Paris" and "City of light" may appear within a text, referring to the entity *Paris*.

In representing textual information, our approach aligns with the format introduced in [6]. A first list is composed of the textual entities, each of them being defined by its type, an index, and a list of co-referencing mentions. Mentions are composed of their textual value and offsets. A second list captures interactions between all entity pairs through triples, comprising the index of the subject entity, a predicate, and the index of the target entity. Unlike recent datasets, we have chosen to extract both proper (Barack Obama, World War II) and common (he, the man) nouns. This choice was made in order to improve co-reference chain detection and relation extraction. Regarding attributes, some types, such as time information, have their occurrences annotated with a formatted value in addition to their textual one.

4. Data

One of the main identified obstacles is the lack of annotated data regarding intelligence, particularly for French. To overcome this shortcoming, we considered two solutions.

The first one was to call on a service provider to create a dataset of 2,000 dummy intelligence reports by hand, from draft to annotation, with our ontology. The second one involved enriching this initial dataset with synthetic data to achieve a more balanced representation between classes.

[3] https://github.com/Emvista/popcorn-dataset

4.1. Manual corpus

To build a dataset of 2,000 dummy intelligence reports from scratch, we called on Isahit[4], an ethical service provider specializing in data labeling for artificial intelligence systems. The dataset creation process was structured to ensure accuracy and coherence. Authors were tasked with crafting fictional scenarios using provided key elements (the central event, the people and organizations involved, the locations where the action is set, the materials used, etc). Despite this support, the initial editorial quality required a post-editing phase to remove spelling mistakes, grammatical errors and inconsistencies. Moreover, we drew specifications up. These contained relatively minimalist annotation guidelines accompanied by an exhaustive description of the ontology. We illustrated each concept of the ontology with examples.

Before starting the annotation process, we carried out a test phase. Kili Technology[5] was chosen as the annotation platform. The complexity of the task and the richness of the ontology prompted us to create an initial 7-step annotation protocol with several objectives: to maximize the completeness of the annotation and to make the process more straightforward, intuitive, and precise. The provider annotated 285 texts using this protocol, with moderate control and regular feedback. Although the outcome yielded mixed results, this test phase proved invaluable for both the annotators and ourselves, offering insightful lessons and opportunities for refinement.

Table 1. Revised 7-step annotation protocol with pairwise IAA and F1 scores

Step	Revised protocol	PW Cohen's κ	PW F1 against gold
1	Events	0.25	0.56
2	Entities of interest linked to events	0.50	0.76
3	Named entities not yet annotated		
4	Event and entity coreference	0.67	0.75
5	Entities' attributes	0.55	0.78
6	Space/Time/Quantity attributes and relations	0.15	0.54
7	Relations between entities	N/A	N/A

The assessment of annotation quality was conducted through objective measures. Building upon insights gained from previous experiments, we refined the initial protocol (see Table 1), subdividing the first step into two distinct phases and consolidating the two coreference steps into a single phase. A team comprising three annotators (now A1, A2, and A3) alongside a consortium expert undertook the annotation of 200 texts. Both the annotators and the expert are native French speakers. To avoid any bias that would be reflected throughout the protocol, each step, from the second one onwards, started for all annotators from the expert's annotations. This way, inter-annotator agreement (IAA) scores and F1 scores could be calculated for each step independently (see Table 1 for IAA and F1 scores).

First, as was expected, some steps were more challenging than others for the three annotators; being able to formally identify them (e.g., steps 1 and 6) was an observation we could capitalize on. Second, annotators' performances against gold were not uniform, one of them largely outperforming the others. Third, good agreement between two annotators did not necessarily mean relevant annotations: A1 is twice among the pair with the lowest Cohen's κ when paired with A2 and A3 (steps 2-3 and 6), though being closest to the gold annotations (i.e., the expert's annotations) in both cases. Overall, we measured these scores with strict expectations, but in a fair number of cases, the annotations regarded as faulty are acceptable. This phase enabled us to deem the annotation quality correct, although a more thorough control seemed necessary.

For the annotation process of the rest of the dataset, i.e., the remaining 1,800 reports, we implemented an increased level of control over the provider's annotations to remedy the problems and difficulties identified during the testing and quality control phases. It involved several elements. First, we reviewed the first step entirely, which laid the foundations for the rest of the protocol. Second, we partially reviewed the rest of the steps. Third, we validated every annotation schema used for each step and did collective annotation sessions before the annotators started. Finally, we used automatic checks listing detectable errors for the annotators to remedy.

[4] https://fr.isahit.com/
[5] https://kili-technology.com/

Despite the ontology and the guidelines, the distribution of classes is unbalanced. As far as events and entities are concerned, elements not supplied in the guidelines were naturally introduced for the needs of the editors' scenarios. On the other hand, a few concepts were semantically too close to differentiate them while annotating (e.g., HooliganismTroublemaking and AgitatingTroublemaking). These elements made total control of the information virtually impossible at the time. However, it is still possible to merge classes that are considered to have a close semantic. These elements were instructive, and the consortium is working on a second version of the ontology and the annotation guidelines.

Although this dataset remains proprietary for the most part, we release 800 reports, most of which we have reviewed, along with all the automatically generated data. Each one of the 400 texts with gold standards annotation were annotated by 2 experts with discussions in case of disagreements. The average text consists of 707 characters (or 143 tokens). We describe the generation process in the next section.

4.2. Synthetic Data

4.2.1. Generating synthetic data

To address the issues of low data diversity and misalignment of the original annotations described in Section 2, we decided on an alternative strategy for synthesizing data. The foundation of our proposed method consists of two main components: Large Language Model (LLM) and In-Context Learning (ICL). ICL is a newly created paradigm in NLP that enables LLMs to make predictions based only on contexts supplemented with a limited set of examples. ICL often helps to refine the output of an LLM, improving its accuracy even in the absence of fine-tuning. Contrary to previous research that modified the original dataset texts, our approach involves the generation of new synthetic texts and their corresponding annotations. This is achieved by using a fine-tuned LLM model, Vigostral-7B[6], which is a French chat-based model that has been fine-tuned (specialised) from Mistral-7B [8] using a variety of datasets, including distillation, translated, and open-source datasets[7]. A well-known reasoning prompt method, Chain-of-Thoughts [18], 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 [19, 20], where we make a naive assumption that if a pair of entities (e_{head}, e_{tail}) is present in both the text and the template Knowledge Graph (KG), these two entities maintain the same relation as the one in the KG.

4.2.2. Our approach

We formalize the task of synthesizing annotated data as a natural language generation task. For a given text t, consider $A = \{a_1, \ldots, a_n\}$ is the set of annotations, $R = \{r_1, \ldots, r_n\}$ is the set of relations, K is the KG and $A, R \subseteq K$. Annotation a_n could be an entity, attribute, or event. In our initial experiment, we constructed our text generation prompt p by utilizing the KG as an input to get our intended output synthetic text t'. The KG K of a prompt p is constructed by extracting all the annotated relations R_t from the text t of the manual corpus. Once the synthetic text t' is produced, we proceed to extract the set of annotations $A_{t'}$ and the set of relations $R_{t'}$ by simply filtering out any triples of the KG K where either the head or the tail is not present in the text. We observed that this method falls short of two problems:

- *Incomplete Text Annotation*: Despite having annotations, we found that it is often incomplete. The method's effectiveness heavily depends on the performance of the LLM used. Consequently, the likelihood of generating a text, t', that includes all the input triples of K 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 hallucination problems, such as incoherent texts with their input KGs, repetitive tokens, inclusion of parts of the prompt, and texts in different languages.

To address these shortcomings, we integrated Self-consistency [21] in our text generation and annotation workflow. The idea behind self-consistency is to prompt the LLM to generate a set of n outputs and select the most consistent

[6] https://huggingface.co/bofenghuang/vigostral-7b-chat
[7] https://github.com/bofenghuang/vigogne/blob/main/docs/data.md

one. For text generation, we prompted Vigostral to generate $n = 3$ outputs and used a voting prompt to select the best text $n = 5$ times based on creativity, coherence, and the text's capacity to include all the input triples of the KG. We then select the text according to the Borda count method[8]. 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. In our study of synthetic data utilization, we seek to answer two questions:

- *Can synthetic data serve as supplementary data to improve the performance of classes with low support?*
- *Given the ability to generate an abundance of new texts for model training, how can we optimally utilize synthetic data, considering the challenge of selecting the best texts?*

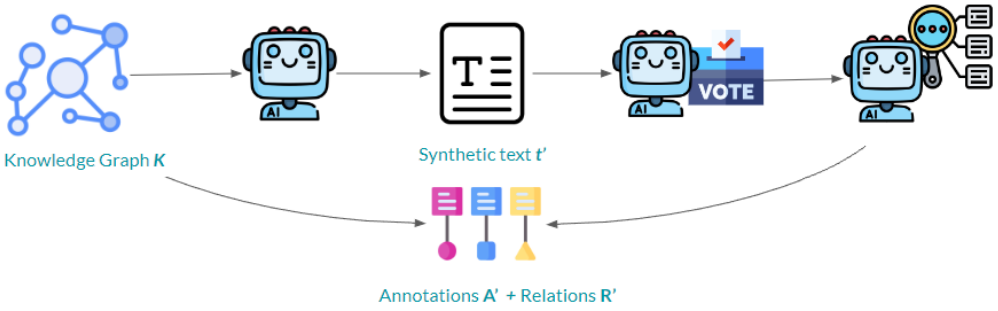

Fig. 1. Synthetic data generation workflow. It begins with a knowledge graph K as an input, which is used to generate a set of synthetic texts t'. Afterwards, the LLM is prompted to select the most suitable t' and to annotate the text. These annotations are then combined with the filtered knowledge graph $K' = \text{NAIVE\_FILTER}(K, t')$ to get the set of annotations A' and the set of relations R'.

As mentioned in section 3, the POPCORN dataset has an imbalanced class distribution linked to its large number of fine-grained classes. Generating coherent documents to balance the whole ontology remains complex with our current solution (our approach does not yet supports the generation of texts with modified KGs). In order to study the benefits of synthetic text generation, we therefore focused on events. Events are generally limited to one type per text and are therefore easier to balance. We manually selected the ten least represented Event classes and used the texts mentioning them as a reference to generate 1200 annotated synthetic texts (thus reducing the imbalance ratio of Events from 51 in the hand-crafted dataset to 27 with the addition of the synthetic texts) to answer the above questions. In order to study the second question, we experimented with different strategies for selecting a subset of synthetic texts in order to study the impact of these filtering strategies on model training. The proposed filtering mechanism is based on the semantic/lexical difference of the synthetic texts and the hand-crafted texts from which they are produced. We used the Cosine Similarity to measure these differences. We explored two methods for this mechanism: 1) A sentence encoder, Camembert-large [9], and 2) Bag-of-Words with TF-IDF. The distribution of cosine similarity scores obtained from both methods is illustrated in Figure 2 and 3. The different selection strategies and their results are described further in section 5. Since t and t' describe the same knowledge graph K_t, despite their different writing styles, higher semantic similarity scores are expected. Low lexical similarity scores indicate that although t and t' roughly describe the same K_t, different lexical properties were used. Samples of the original text t and the synthetic text t' are provided in Table 2 for reference.

[8] https://mathbooks.unl.edu/Contemporary/sec-5-2-borda.html
[9] https://huggingface.co/dangvantuan/sentence-camembert-large

Original Text	Synthetic Text
Suite à une plainte de Madame Oxanna Hernandez concernant des nuisances sonores, la police a effectué une descente à Porto, dans l'ancien entrepôt de l'usine portugaise de fabrication de plastique. L'usine est régulièrement prise d'assaut par des voyous. De plus, l'association citoyenne "Segurança" a demandé aux autorités de prendre des mesures pour empêcher désormais l'accès à cet entrepôt.	Dans la ville portugaise de Porto, une usine non gouvernementale, spécialisée dans la production de plastique, a commencé à causer des nuisances sonores. Des habitants et des autorités locales ont exprimé leur mécontentement face à cette situation, qui met à mal le climat de vie de la ville. L'usine est dirigée par un groupe de voyous qui semble avoir la main mise sur l'usine.

Table 2. Sample of an original text t, and synthetic text t' that is generated from knowledge graph K.

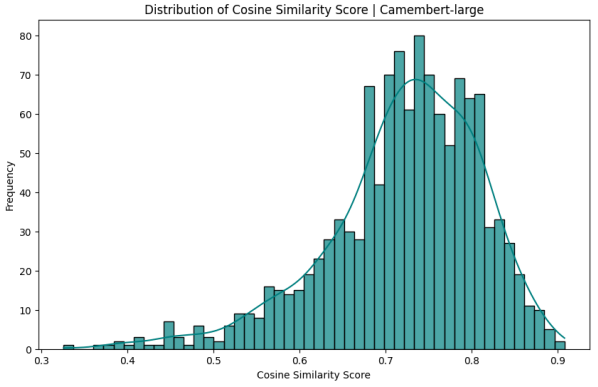

Fig. 2. Distribution of cosine similarity scores between text t and its synthetic version t' based on Camembert-large encoder.

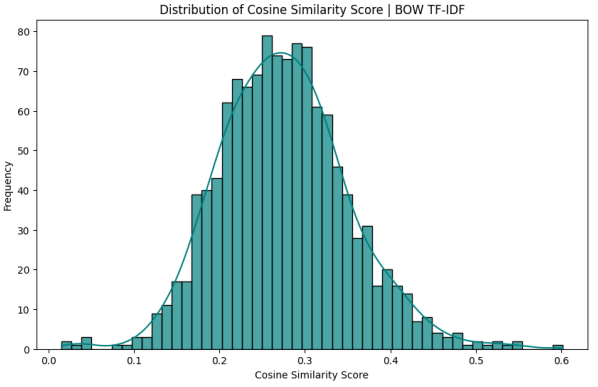

Fig. 3. Distribution of cosine similarity score between text t and its synthetic version t' based on Bag-of-Words with TF-IDF.

5. Experiments

We trained IE models using several strategies on the manual and synthetic datasets previously described. For all the experiments, we conducted the evaluation on a set of 400 reports annotated by experts and separated from the texts used to generate the synthetic reports or train the models. We conducted experiments on the following models:

- Boundary Smoothing (BS) [7]: This model reassigns entity probabilities from spans to the surrounding ones during loss computation as a regularization technique for span-based ER models in order to reduce the impact of badly annotated borders and improve processing of ambiguous ones.
- Unified Model (UM) [22]: Used to tackle the ER and RE tasks, this model features two classification modules that use representations produced by a shared encoder. The first module detects entity spans while the second is trained to predict interactions between them.

Experiment 1. The first experiment involved training the BS model with different proportions of the synthetic text corpus. To do this, 4 training sets were used, their compositions are as follows : 400 hand-crafted texts, 400 hand-crafted and 400 synthetic texts, 400 hand-crafted and 800 synthetic texts and finally, 400 hand-crafted and 1200 synthetic texts. In this experiment, only the annotations produced by the synthetic data generation are used for the synthetic texts.

Experiment 2. The second experiment used the UM architecture, with CamemBERT-base model [10] weights to initialize the backbone encoder, and explored several ways of taking advantage of synthetic texts. As mentioned in section 4.2.2, annotations from the synthetic text generation method carry a risk of being incomplete, or of wrongly assuming the presence of entities or relations. To address this issue, we sought to improve annotations by implementing a Teacher-Student learning strategy. This solution consists in training a Teacher model on a training set of 400 hand-

[10] https://huggingface.co/almanach/camembert-base

crafted reports. The Teacher model is then used to make predictions on an additional text set (the Synthetic corpus). These predictions serves as annotations to pre-train a second model, the Student model. Finally, the training of the Student model is refined on the 400 handcrafted texts. We use this solution both as an alternative to synthetic annotations and as a complement to it, by keeping the union of the Teacher predictions on the synthetic set and its initial synthetic annotations. As the focus is on how to select synthetic data, this experiment only involves for training 400 hand-crafted texts and 400 synthetic reports filtered out of the 1200 synthetic corpus.

5.1. Implementation details

The UM was implemented using the Pytorch and Lightning libraries and will be made public at the same time as the dataset. As for the implementation of the BS model, we have followed the one made available by the authors [7].

Both models were trained on 5 seeds with 50 epochs. The batch size was of 4 and 1, the pre-training learning rate of 2e−3 and 1e−4, additional learning rates of 2e−5 and 5e−6 for the BS and UM models respectively.

5.2. Results and Analysis

Experiment 1. Table 3 shows the main results of adding synthetic data to manual data. Synthetic data demonstrate their usefulness to improve the overall performance of the model, in particular the Macro F1. Event extraction yields better results when training with 1,200 texts, with +2.37% on Macro F1 and +0.89% on Micro F1. However, the performances seem to decrease for ER, as illustrated by a lower Macro F1 and an improvement too small to be significant on Micro F1. Unlike Entities, Attribute Extraction seems to benefit more from synthetic data. The model improves its performance with 1,200 synthetic texts up to +3.93% in Macro and +0.44% in Micro.

Table 3. Boundary Smoothing results for Event, Entity and Attribute Extraction with BS + flaubert_base_uncased.

	Events		Entities		Attributes	
Model	F1 macro	F1 micro	F1 macro	F1 micro	F1 macro	F1 micro
BS (baseline)	41.83 ±0.79	55.56 ±0.63	65.41 ±1.04	**81.60** ±0.26	56.74 ±0.86	80.02 ±0.26
BS +400 synthetic texts	43.92 ±1.14	55.94 ±0.80	**65.82** ±1.04	81.14 ±0.14	59.17 ±1.11	**80.86** ±0.32
BS +800 synthetic texts	43.61 ±0.96	56.00 ±0.47	64.33 ±0.94	79.91 ±0.34	59.96 ±1.82	80.61 ±0.69
BS +1200 synthetic texts	**44.20** ±1.08	**56.45** ±0.85	63.56 ±0.82	80.06 ±0.38	**60.67** ±0.95	80.46 ±0.35

Low-support classes. The F1 Macro score provides a comprehensive assessment of model performance by considering the F1 score of each class, thereby offering better insights into classes with limited support. The significant improvements seen on this metric led us to study the variation occurring on infrequent classes. The results on low-support classes of the same model trained with and without synthetic data can be found in Tables 4, 5 and 6. The variations for Events are mostly positive (see Table 4). Only 4 out of 12 classes under-perform. The results are mostly negative for Entities (see Table 5). Notably, Attributes showcase the most promising performance overall, with only two classes showing minor dips in recognition accuracy. Quantifying the improvements across these categories, we observe a notable enhancement of +3.89% in Events, a slight decrease of −2.40% in Entities, and a significant boost of +8.50% in Attributes. These findings underscore the efficacy of synthetic data augmentation, particularly in enhancing the recognition capabilities for less represented classes, ultimately contributing to the overall robustness of the model.

Experiment 2. For the second experiment, results of the different model versions are shown in Table 7 for Event and Entity recognition and in Table 8 for Attribute Extraction and RE. Four training strategies are featured:

- No pre-training : the model is trained only on the set of 400 hand-crafted reports.
- Synthetic pre-training : the model is pre-trained on 400 synthetic texts with their corresponding synthetic annotations and is then fine-tuned on the set of 400 hand-crafted reports.
- Teacher pre-training : the model is pre-trained on 400 synthetic texts with annotations produced by a Teacher-model and is then fine-tuned on the set of 400 hand-crafted reports.
- Synthetic ∪ Teacher pre-training: the model is pre-trained on 400 synthetic texts with synthetic annotations together with those of a Teacher model and is then fine-tuned on the set of 400 hand-crafted reports.

Table 4. Results of the models on half of the event classes with the lowest support with and without the 1,200 synthetic texts.

	BS (baseline)				BS + 1200 synthetic texts			
Classes	Support train&dev	Precision	Recall	F1	Support train&dev	Precision	Recall	F1
AGITATING_TROUBLE.	16	2.85	3.32	**2.97**	846	0.00	0.00	0.00
CIVIL_WAR_OUTBREAK	19	52.24	65.00	**57.45**	440	51.40	67.21	54.83
COUP_D_ETAT	24	42.73	46.40	44.23	198	56.05	38.04	**45.22**
DEMONSTRATION	38	4.10	4.13	4.11	1215	9.69	17.93	**12.44**
DRUG_OPERATION	13	18.63	20.00	18.48	242	37.22	46.66	**41.30**
ELECTION	27	53.47	86.66	65.73	197	75.64	90.66	**82.45**
HOOLIGANISM_TROUBLE.	9	0.00	0.00	0.00	1008	0.00	0.00	0.00
ILLEGAL_CIVIL_DEMO.	29	26.46	28.38	**27.31**	30	23.98	18.06	20.48
NATURAL_CAUSES_DEATH	9	35.48	47.74	40.25	15	47.70	40.00	**43.02**
POLITICAL_VIOLENCE	29	9.15	13.10	**10.72**	137	4.92	2.75	3.51
POLLUTION	31	47.78	81.08	60.11	141	64.17	72.43	**68.01**
SUICIDE	22	35.33	45.45	39.27	22	38.99	45.45	**41.92**
TRAFFICKING	38	22.24	57.14	31.85	381	38.13	57.14	**45.27**

Table 5. Results of the models on half of the entity classes with the lowest support with and without the 1,200 synthetic texts.

	BS (baseline)				BS + 1200 synthetic texts			
Classes	Support train&dev	Precision	Recall	F1	Support train&dev	Precision	Recall	F1
INTERGOV._ORG.	14	36.50	50.00	41.14	67	49.64	50.00	**49.03**
MILITARY	15	5.49	3.80	**4.42**	142	0.00	0.00	0.00
MILITARY_ORG.	38	63.59	65.56	**64.42**	175	67.97	46.36	55.07
NON_MILITARY_GOV._ORG.	412	70.12	83.32	**76.15**	2249	77.49	74.27	75.84
TERRORIST_OR_CRIMINAL	129	58.31	57.04	**57.56**	1168	53.88	54.78	54.23

Table 6. Results of the models on half of the attribute classes with the lowest support with and without the 1,200 synthetic texts.

	BS (baseline)				BS + 1200 synthetic texts			
Classes	Support train&dev	Precision	Recall	F1	Support train&dev	Precision	Recall	F1
HEIGHT	4	17.76	55.55	26.81	14	39.02	55.55	**45.56**
LATITUDE	3	35.24	54.28	41.83	4	57.85	51.42	**53.66**
LENGTH	4	31.33	20.00	23.11	13	47.46	49.33	**47.79**
LONGITUDE	5	27.92	54.28	41.83	5	34.80	54.28	**42.15**
MATERIAL_REFERENCE	14	36.33	68.42	47.36	31	51.75	58.94	**54.94**
QUANTITY_MIN	20	37.41	62.50	**46.72**	76	34.36	56.66	42.77
TIME_MAX	11	48.53	38.46	**42.81**	12	47.38	38.46	41.46
TIME_MIN	28	30.07	22.50	25.43	33	39.02	25.00	**30.20**
WEIGHT	15	61.91	93.91	74.42	24	75.24	98.26	**84.96**
WIDTH	4	4.00	6.00	4.66	11	14.16	10.00	**11.39**

First of all, we observe that the tendency of the previous experiment, i.e., an increase in the performance for classes with low support, still holds, except for Attribute Extraction. The second observation is that using synthetic annotations alone is useful for classes with low support (for Event and Entity Extraction). Their interest is all the more significant when they are coupled with the predictions of a Teacher model, this time also including the RE task.

Table 7. Unified Model results for Event and Entity Extraction.

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
Synthetic pre-training	45.17 ±0.83	58.81 ±0.46	67.45 ±1.33	81.61±0.12
Teacher pre-training	43.60 ±3.46	58.56±1.95	67.72 ±2.13	81.95±0.36
Synthetic ∪ Teacher pre-training	**45.19** ±2.07	**58.93**±0.90	**68.38** ±0.34	**82.03**±0.34

Table 8. Unified Model results for Attribute and Relation 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
Synthetic pre-training	60.05±1.97	**82.07**±0.62	43.26±1.22	55.85±0.75
Teacher pre-training	55.98 ±8.18	81.33±0.73	43.25±4.78	**57.10**±0.70
Synthetic ∪ Teacher pre-training	60.39±2.76	81.27±0.75	**46.49**±0.55	56.87±0.84

Tables 9 and 10 show the results of the UM with pre-training when we try to select the 400 synthetic texts used for pre-training from the 1,200 synthetic texts available. Annotations for the pre-training texts are the union of synthetic annotations and teacher predictions. Although weaker than without filtering, the evaluation shows an advantage in keeping only the texts with the lowest semantic similarity scores (particularly for Event Extraction) compared to filtering texts with the highest scores. Strong lexical similarity with model texts seems preferable when considering lexical filtering. Minimum semantic similarity appears to be preferable for Relations and Events, maximum semantic similarity for Attributes, while maximum lexical similarity is more interesting for Entities. A more in-depth study of training or pre-training text selection strategies should be carried out, whether by selecting from a larger set of texts or using other text generation approaches, for instance.

Table 9. Unified Model results with filtering strategies for Event and Entity Extraction.

Filtering Strategy	Events		Entities	
	F1 Macro	F1 Micro	F1 Macro	F1 Micro
Min Semantic	**45.16** ±2.11	59.01 ±0.66	67.91 ±1.24	81.46±0.37
Max Semantic	42.96±0.86	58.54 ±0.91	67.64 ±2.57	**81.80**±0.25
Min Lexical	43.22 ±2.66	57.82±1.14	66.95 ±0.31	81.55±0.37
Max Lexical	44.91±2.01	**59.17**±1.59	**70.24** ±1.41	**81.80**±0.53

Table 10. Unified Model results with filtering strategies for Attribute and Relation Extraction.

Filtering Strategy	Attributes		Relations	
	F1 Macro	F1 Micro	F1 Macro	F1 Micro
Min Semantic	60.11±1.19	81.58 ±0.63	**46.25** ±0.51	**57.20**±0.44
Max Semantic	**60.94** ±1.44	**81.66** ±1.11	45.54 ±2.67	56.57±0.48
Min Lexical	60.08±3.33	81.31±0.72	44.50 ±1.78	56.86±0.60
Max Lexical	60.35 ±2.75	81.61±1.04	45.51 ±1.62	56.86±0.61

6. Conclusion and future works

This paper presented POPCORN, the first French defence-themed textual dataset, manually designed and annotated to serve as benchmark for training and evaluating NLP models. By augmenting this dataset with synthetic texts generated through a sophisticated process involving LLMs, we showcased the potential of synthetic data, especially in addressing low-support classes. All resources used in this study, including data, prompt templates, models and code are readily accessible via the provided Git link[11]. The effort made on the creation of an synthetic but realistic dataset makes it possible to publicly introduce the first models capable of processing several dozen classes of entities and relations for intelligence services. We continue to work on improving data quality and we are currently experimenting with original IE models such as models integrating the task of Entity Linking to a structured Knowledge Base.

[11] https://github.com/Emvista/popcorn-dataset

Acknowledgements

This research was partially supported by Agence Innovation Défense and Direction Générale de l'Armement.

References

[1] Zaporojets, Klim, Deleu, Johannes, Develder, Chris and Demeester, Thomas. (2021) "DWIE: An entity-centric dataset for multi-task document-level information extraction". *Information Processing and Management*. **58**(4): 102563

[2] Verdy, S., Prieur, M., Gadek, G. & Lopez, C. DWIE-FR : Un nouveau jeu de données en français annoté en entités nommées. *Actes De CORIA-TALN 2023. Actes De La 30e Conférence Sur Le Traitement Automatique Des Langues Naturelles, TALN 2023 - Volume 2 : Travaux De Recherche Originaux - Articles Courts, Paris, France, June 5-9, 2023*. pp. 63-72 (2023)

[3] Sang, E. and Meulder, F. (2003) "Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition". *Proceedings Of The Seventh Conference On Natural Language Learning, CoNLL 2003, Held In Cooperation With HLT-NAACL 2003, Edmonton, Canada, May 31 - June 1, 2003*. pp. 142-147.

[4] Serrano, L., Bouzid, M., Charnois, T., Brunessaux, S. & Grilheres, B. Extraction et agrégation automatique d'événements pour la veille en sources ouvertes: du texte à la connaissance. (2013) *IC-24èmes Journées Francophones D'Ingénierie Des Connaissances*.

[5] Prieur, M., Mouza, C., Gadek, G. & Grilhères, B. Evaluating and Improving End-to-End Systems for Knowledge Base Population. *Proceedings Of The 15th International Conference On Agents And Artificial Intelligence, ICAART 2023, Volume 3, Lisbon, Portugal, February 22-24, 2023*. pp. 641-649 (2023)

[6] Yao, Y., Ye, D., Li, P., Han, X., Lin, Y., Liu, Z., Liu, Z., Huang, L., Zhou, J. & Sun, M. DocRED: A Large-Scale Document-Level Relation Extraction Dataset. *Proceedings Of The 57th Conference Of The Association For Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*. pp. 764-777 (2019)

[7] Zhu, Li (2022) "Boundary Smoothing for Named Entity Recognition" *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*

[8] Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., Casas, D.D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L.R., Lachaux, M., Stock, P., Scao, T.L., Lavril, T., Wang, T., Lacroix, T., & Sayed, W.E. (2023). Mistral 7B. ArXiv, abs/2310.06825.

[9] 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), pp. 452–457

[10] 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), pp. 6382–6388

[11] Zhang, Danqing & Li, Tao & Zhang, Haiyang & Yin, Bing. (2020). On Data Augmentation for Extreme Multi-label Classification.

[12] George A. Miller. 1994. WordNet: A Lexical Database for English. In Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994.

[13] Sérasset, Gilles. 'DBnary: Wiktionary as a Lemon-based Multilingual Lexical Resource in RDF'. 1 Jan. 2015 : 355 – 361.

[14] Mikolov, Tomas & Chen, Kai & Corrado, G.s & Dean, Jeffrey. (2013). Efficient Estimation of Word Representations in Vector Space. Proceedings of Workshop at ICLR. 2013.

[15] Jeffrey Pennington, Richard Socher, and Christopher Manning. (2014) GloVe: Global Vectors for Word Representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543

[16] Xie, Z., Wang, S. I., Li, J., Lévy, D., Nie, A., Jurafsky, D., & Ng, A. Y. (2019). Data noising as smoothing in neural network language models. Paper presented at 5th International Conference on Learning Representations, ICLR 2017, Toulon, France.

[17] Alexander Fabbri, Simeng Han, Haoyuan Li, Haoran Li, Marjan Ghazvininejad, Shafiq Joty, Dragomir Radev, and Yashar Mehdad. 2021. Improving Zero and Few-Shot Abstractive Summarization with Intermediate Fine-tuning and Data Augmentation. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 704–717

[18] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Chi, E.H., Xia, F., Le, Q., & Zhou, D. (2022). Chain of Thought Prompting Elicits Reasoning in Large Language Models. ArXiv, abs/2201.11903.

[19] Roland Roller, Eneko Agirre, Aitor Soroa, and Mark Stevenson. 2015. Improving distant supervision using inference learning. 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), pp. 273–278, Beijing, China. ACL.

[20] Xiang Deng and Huan Sun. 2019. Leveraging 2-hop Distant Supervision from Table Entity Pairs for Relation Extraction. 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), pp. 410–420, Hong Kong, China. Association for Computational Linguistics.

[21] Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E.H., & Zhou, D. (2022). Self-Consistency Improves Chain of Thought Reasoning in Language Models. ArXiv, abs/2203.11171.

[22] Prieur, M., Du Mouza, C., Gadek, G. & Grilheres, B. Shadowfax: Harnessing Textual Knowledge Base Population. *Proceedings Of The 47th International ACM SIGIR Conference On Research And Development In Information Retrieval, SIGIR 2024, Washington, USA, July 14-18, 2024*.