Synthetic Annotated Data for Named Entity Recognition in Computed Tomography Scan Reports

Alexander Platas[1], Elena Zotova[1,2], Paola Martínez-Arias[1], Karen López-Linares[1] and Montse Cuadros[1]

[1]Fundación Vicomtech, Basque Research and Technology Alliance (BRTA), Mikeletegi 57, 20009 Donostia-San Sebastián (Spain)
[2]Department of Languages and Computer Systems, University of the Basque Country, Spain

Abstract
It is widely acknowledged that clinical data, in general, is scarce, and this scarcity worsens when focusing on specific domains. Moreover, the challenge escalates when annotated data is required. In this paper, we propose an approach to create synthetic annotated datasets for Named Entity Recognition (NER) tasks in Computed Tomography Reports (CTR) by leveraging large language models (LLMs). We investigate the potential of LLMs to generate meaningful texts in the healthcare domain through a combination of text generation techniques and automatic annotation using LLMs. Additionally, we conducted a series of experiments to demonstrate the efficacy of using synthetic data compared to real data for solving NER tasks.

Keywords
Biomedical NER, text generation, data synthesis

1. Introduction

This work presents a method for creating synthetic annotated datasets for Named Entity Recognition (NER) in Computed Tomography Reports (CTR). We experiment with text generation and automatic annotation with large language models (LLMs), considering their capacity to produce meaningful texts on a given topic and zero-shot learning [1]. LLMs have already shown potential in extracting valuable information from unstructured data, such as electronic health records (EHRs) and digital medical data. Instead of applying LLMs in a zero-shot setting, we propose creating synthetic-labelled data using LLMs for further fine-tuning supervised NER models. Our research is motivated by the following challenges in Biomedical Natural Language Processing (BioNLP).

High-quality annotated corpora are essential to train and validate predictive models in healthcare. Manual annotation requires personnel time and preparation, and the challenge is even more difficult in BioNLP, as the cost of expertise for annotation is higher than in general-purpose NLP, which makes using crowd-sourcing platforms for annotations almost impossible. This scarcity of annotated clinical narratives poses a significant challenge for machine learning (ML) and deep learning (DL) techniques, as they rely on large supervised corpora for training models [2]. BioNLP also addresses sensitive information and privacy concerns, such as private information in electronic health records (EHR), so most datasets are not publicly available for research and development purposes. Concerns regarding patient privacy and lack of reliable de-identification techniques have made hospitals and clinics highly reluctant to allow researchers to access clinical data outside the association [3].

We explore the new possibilities of synthetic textual data to overcome the above-mentioned factors. Synthetic data, in general, according to The Alan Turing Institute, is "data that has been generated using a purpose-built mathematical model or algorithm, with the aim of solving a (set of) data science task(s)" [4]. This type of data can statistically replicate real-world data's underlying patterns and characteristics despite its artificial nature, so its defining feature is this ability to mimic real-world characteristics. Synthetic data can be classified into three broad categories: fully synthetic, partially synthetic, and hybrid. The fully synthetic data does not contain any original information; partially synthetic data replaces only the values of the sensitive attribute selected with synthetic values; and the hybrid synthetic data, which we have generated, uses both the original and synthetic data [5].

The contributions of this paper are the following:

- We propose a hybrid method for generating synthetic annotated corpus from real-world structured data using an existing dataset of Computed Tomography (CT) scans reports. This synthetic data is used as a training corpus for fine-tuning of language models for the biomedical NER task.

SEPLN-2024: 40th Conference of the Spanish Society for Natural Language Processing. Valladolid, Spain. 24-27 September 2024.
aplatas@vicomtech.org (A. Platas); ezotova@vicomtech.org (E. Zotova); pnmartinez@vicomtech.org (P. Martínez-Arias); klopez@vicomtech.org (K. López-Linares); mcuadros@vicomtech.org (M. Cuadros)
0009-0002-5501-017X (A. Platas); 0000-0002-8350-1331 (E. Zotova); 0000-0002-5952-1578 (P. Martínez-Arias); 0000-0002-4800-6052 (K. López-Linares); 0000-0002-3620-1053 (M. Cuadros)

Our method provides various prompting techniques for data generation with LLMs and the analysis of the effectiveness of synthetic data as data augmentation. Leveraging real-world data in the text synthesis helps get good quality training data. The synthetic annotated corpus will be publicly available[1].

- Experiments with the models fine-tuned for the NER task show that the synthetic data can help to improve the models' performance in the situation of annotated data scarcity.

This paper is organised as follows. In Section 2 we overview works related to synthetic data and methods to get augmented corpora, both in biomedical and general-purpose NLP. Section 3 describes the task and the corpus we created with LLMs and the corpus with the original data manually annotated. In Section 4 is dedicated to the methodology of creating new corpora and in Section 5 we explain the details of the experimentation with corpora. In Section 6 the results of the experiments are shown, and Section 7 concludes our paper and discusses future work.

2. Related Work

2.1. General-purpose NLP

An upsurge in data synthesis and augmentation in general-purpose NLP began with rule-based approaches, such as grammar and lexicon replacement [6, 7, 8], and then adopted model-based approaches, such as sentence retrieval and backtranslation with machine learning techniques [9, 10, 11]. The interest in synthetic data generation is also related to the emergence of new architectures of deep neural networks and pre-trained language models. Various authors use BERT [12], BART [13], and GPT-2 [14] to generate data for classification and common sense reasoning tasks, experiment with conditioning on labels by prepending the label to training data during fine-tuning [15, 16, 17, 18]. [19] propose a task augmentation approach that utilises conditional generation to create in-domain synthetic data for an auxiliary Natural Language Inference (NLI) task, which then is employed to initialise the target task classifier. However, these works show better results with synthetic data, but observe that one needs to detect and discard low-quality labelled data or optionally re-label it. In the work of [20], the authors try to overcome these problems by knowledge distillation and self-training on domain-specific data.

The most recent works explore the capacity of Large Language Models (LLMs) to annotate corpora automatically. [21] report that the GPT-3.5-turbo[2] outperforms crowd-source workers for annotation such tasks as relevance, stance, topics, and frame detection. The authors provided the corpora collected from Twitter and news and the annotations guides to the LLM as a prompt. A similar approach [22] leverages LLMs to generate a few-shot prompt with explanations, which is then used to annotate unlabelled data query and keyword relevance assessment, question-answering task, disambiguating word senses through binary classification of sentence pairs. [23] and [24] use LLMs for annotation with noisy labels and an active learning loop to determine what to efficiently annotate.

In a multilingual setting, a fine-tuned 5-billion-parameter multilingual sequence-to-sequence model was used to generate annotated data for intent classification and slot tagging [25], and it was reported to perform better than the back-translation method.

2.2. Biomedical NLP

Synthetic data generation has also witnessed a marked increase in research publications in biomedical NLP, as well, suggesting a potential for broader adoption. The surveys carried out by [26, 27] provide evidence that synthetic is helpful in different aspects of healthcare care and has possibilities to bridge data access gaps in research and evidence-based policy making. [28], on the contrary, explore the problem of synthetic data in healthcare: although it promises various positive opportunities, synthetic data potential carries concerns such as the risk of bias amplification, low interpretability, and an absence of robust methods for examining data quality.

In [29], the authors tackle the task of generation of medical imaging reports using a hierarchical recurrent neural network decoder, which generates a sequence of topic representations conditioned on image information, and this then conditions the generation of respective sentences. [30] propose the approach based on encoder-decoder Transformer models [31] trained for the gap-filling task to generate discharge summaries from a large mental healthcare provider and an intensive care unit. The model learns a sequence-to-sequence task where the clinical information and the key phrases are in the input, and the full original EHR record is in the output. A classification model trained on synthetic data shows results comparable to the models trained on original data.

The methods for creating synthetic data with text generation models are explored by [32]: CharRNN [33], SeqGAN [34], GPT-2 [14], and CTRL [35]. Then, the authors annotated the resulting data manually for in Named Entity Recognition (NER) task. The best-performing generation model was GPT-2. [36] explores the ability of LLMs to extract structured information from unstructured healthcare texts, specifically for biological NER and relation extraction (RE) tasks, in a zero-shot setting.

[1] The corpus will be released when the paper is accepted
[2] https://platform.openai.com/docs/models/gpt-3-5-turbo

The quality of the synthetic corpora is evaluated by fine-tuning supervised models; the authors report improvements in the performance of downstream tasks, compared to the zero-shot scenario, but not in original data, although the performance is comparable.

We should note that most existing works experiment with corpora in English. There are only few attempts to create a multilingual datasets, for instance, a corpus for Health Question Answering and compare various LLMs [37], including T5 [38], BART [13] and GPT-3.5[3].

3. Task Definition and Corpora

Named Entity Recognition (NER) [39] in the biomedical domain is crucial to extracting concepts (in general-purpose NLP known as named entities), such as locations, treatment plans, medicines/drugs, diagnoses, etc, from clinical narratives. NER uses an IOB (Inside, Outside, Begin) tagging scheme, where each word is assigned a tag indicating whether it is the beginning of a named entity (B), inside a named entity (I), or outside a named entity (O). Formally, a sentence s in a medical text is denoted as a sequence of words $s = (w_1, w_2, \ldots, w_n)$, and the corresponding tags for each word in the sentence are denoted as $y = (y_1, y_2, \ldots, y_n)$, where tag y_i is an element of the tag set $\{B, I, O\}$.

Our goal is to train a NER model for detecting the following named entities in the Computed Tomography Scan Reports (CTSR): SEX (patient's sex), AGE (patient's age), HEPATOPATHY (type of hepatopathy found), TUMOR_SIZE (liver tumor size), and PROCEDURE (procedure performed). We consider two types of annotated corpora for the experimentation: (1) authentic data from liver cancer cases collected in a hospital and (2) synthetic dataset generated and annotated by an LLM.

The first type of data includes a private dataset in Spanish comprising 100 CTSRs performed on 66 patients. This corpus is manually annotated by experts and it is used as gold-standard for the systems. Additionally, we used six real samples as examples in instructions for LLMs, which are not included in training data and are used only to show report details such as structure, length and vocabulary. The second type of corpus consists of 197 reports, created and annotated by the LLM (see details of text generation and annotation in Section 4). The authentic corpus is split in train, development and test sets, as shown in Table 1, while synthetic dataset is used in training split only. The test set is used to evaluate the NER systems. Authentic reports are annotated with 635 entities and synthetic reports contain a total of 1311 entities, as we can observe in Table 2. We can point out that classes SEX and AGE are unbalanced, appearing only in one report in the authentic reports dataset.

[3]https://platform.openai.com/docs/models/gpt-3-5-turbo

Table 1
Corpora statistics. Number of reports and tokens in Authentic and Synthetic datasets.

Dataset		Authentic	Synthetic
Train	report	75	197
	token	20328	44272
Dev	report	15	
	token	4454	
Test	report	10	
	token	2675	

Table 2
Distribution of entities in medical reports

Entities	Number of entities	Avg. entities per report	Avg. tokens per entity
Synthetic data			
SEX	195	0.99	1.24
AGE	199	1.01	2
HEPATOPATHY	433	2.20	2.09
TUMOR_SIZE	286	1.45	3.52
PROCEDURE	198	1.01	3.06
Total	**1311**	**6.65**	**2.41**
Authentic data			
SEX	1	0.01	1
AGE	1	0.01	2
HEPATOPATHY	237	2.37	1.59
TUMOR_SIZE	249	2.49	2.61
PROCEDURE	147	1.47	3.31
Total	**635**	**6.35**	**2.39**

4. CT Reports Generation

In this Section, we describe how we create synthetic CT reports. In our case, synthetic data generation aims to create realistic clinical narratives similar to real reports while making them as diverse as possible. We reduce the probability of an error or hallucinations by incorporating information from real-world structured data.

The generated data were semi-automatically annotated by GPT-3.5-turbo model under human supervision to correct any potential annotation errors, such as entities left unlabeled or the annotation of words that were not entities. Our choice is explained by the model's state-of-the-art capabilities of coherent text generation with a given prompt, which is an instruction or an example of how to complete a task. Given that this dataset consists solely of 197 reports, we manually verified these annotations.

However, unlike other experiments carried out recently [40, 36], we compose prompts for an LLM instruction with real-world data from "*Colorectal-Liver-Metastases*" dataset [41]. This dataset contains CT images from 197 patients with liver cancer. It also includes structured data in a tabular format, as we can observe in Table 3, with 36 attributes for each patient, mostly numerical,

covering demographic, pathological, and survival data.

Table 3
Example of "*Colorectal-Liver-Metastases*" dataset [41].

Patient-ID	age	sex	...	fibrosis percent
CRLM-CT-1001	65	2	...	60
...	...	...	...	...
CRLM-CT-1198	46	1	...	0

Table 4
Comparison of results with different prompts (original in Spanish and in English)

Prompt 1	Escribe un informe médico del paciente <Patient-ID>
Informe Médico - Radiografía Preoperatoria - Paciente: CRLM-CT-1001 - Edad: 65 - Índice de Masa Corporal: 21.9 - Tipo de Tumor: Cáncer de Hígado - Enfermedad: Metástasis Colorrectales en el Hígado - Tamaño del tumor: 1.1 - [...]	
Prompt 2	Escribe una redacción de ámbito médico del paciente <Patient-ID>
El paciente con el código CRLM-CT-1001 es una mujer de 65 años de edad que ha sido diagnosticada con cáncer de hígado. Presenta metástasis hepáticas colorrectales como enfermedad principal. El paciente presenta un tumor de 1.1 cm de tamaño [...]	

Prompt 1	Write a medical report for patient <Patient-ID>.
Medical Report - Preoperative Radiograph - Patient: CRLM-CT-1001 - Age: 65 - Body Mass Index: 21.9 - Tumor Type: Liver Cancer - Condition: Colorectal Metastases in the Liver - Tumor Size: 1.1 - [...]	
Prompt 2	Write a medical writing for patient <Patient-ID>
The patient with the code CRLM-CT-1001 is a 65 year-old woman who has been diagnosed with liver cancer. She has colorectal liver metastases as the primary disease. The patient has a tumor of 1.1 cm in size [...]	

To create a prompt for the model, the role "system" is described as an expert oncologist, and the patient ID is provided to retrieve information from a structured dataset. For each column that must be included in the text, we wrote a brief description to help ChatGPT understand the meaning of each column. Then, the model is instructed to generate a medical report. We observed a significant difference during the initial text generations when we changed the type of text requested in the prompt. As we can see in Table 4, using the term "informe" (report, in English) we obtain a much more schematic generation, while with the term "redacción" (writing, in English) we obtain an output more similar to the required one.

Table 5
Text structure example given to GPT-3.5 (original in Spanish and in English)

INFORMACIÓN CLÍNICA:
[Frase corta sobre la descripción y los antecedentes]

TÉCNICA DE ESTUDIO:
[Frase corta sobre procedimiento realizado]

INFORME:
[Explicación extensa sobre los resultados obtenidos]

CONCLUSIONES
[Conclusión de los resultados]

CLINICAL INFORMATION:
[Short sentence about the description and background]

STUDY TECHNIQUE:
[Short sentence about the procedure performed]

REPORT:
[Detailed explanation of the results obtained]

CONCLUSIONS:
[Conclusion of the results]

Once the desired text style is obtained, we provide the model with a real sample as an example to generate a report with a similar structure. Providing real samples may result in the inclusion of information from those samples in the generated data. Therefore, instead of providing a real sample, we only show the structure of the report and a description of the content it should include in each section, as we can see in Table 5. When using structured data for report generation, the model creates identical reports by only changing the provided data. Furthermore, we add various synonyms, making annotated entities richer in vocabulary, as evidenced in Table 6. To achieve vocabulary variety, we employed high randomness in report generation and automatically replaced repeated phrases with a list of synonyms.

Finally, we obtained the optimal prompt as shown in Table 7, where we specified the type of text, the report structure, and the patient ID which is used as an index to get patients' information from the structured dataset.

Figure 1: An example of annotated synthetic report.

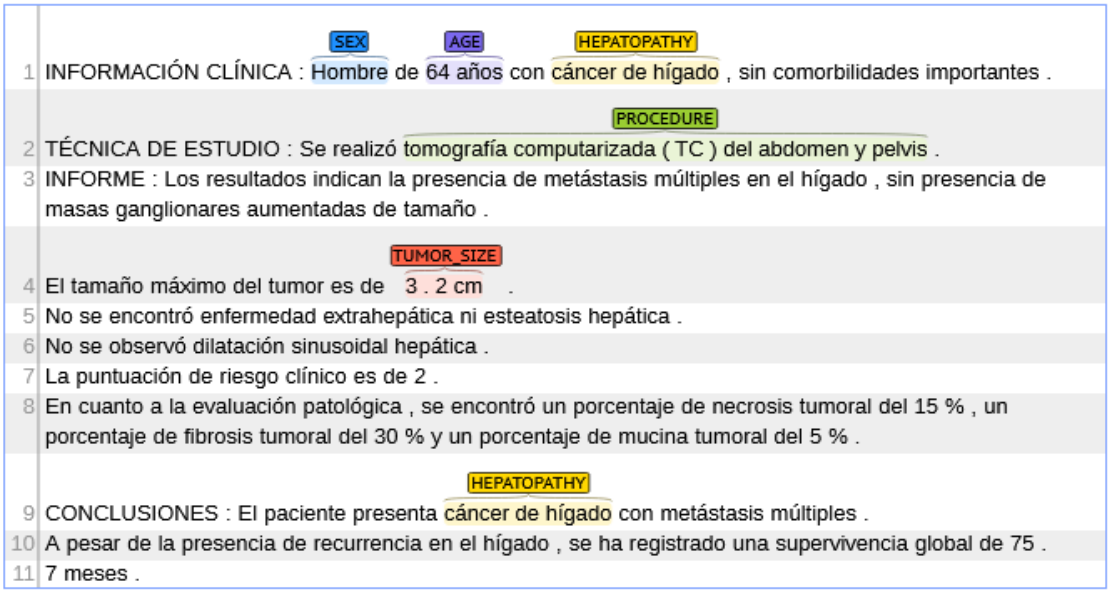

Table 6
Ways to describe the procedure carried out in the authentic and generated data. (original in Spanish and in English)

Synthetic data	Authentic data
TC del abdomen y pelvis	TC helicoidal de abdomen
TC del abdomen y pelvis	TC abdominopélvico
TC del abdomen y pelvis	TAC abdominal y pélvico
Abdomen and Pelvis CT	Helical Abdomen CT
Abdomen and Pelvis CT	Abdominopelvic CT
Abdomen and Pelvis CT	Abdominal and Pelvic CT

An example of a CT report generated using these same prompt is visualised in Figure 1. We can see a coherent grammatically correct text with required entities annotated.

Comparing the generated texts among themselves, we have observed that due to the high randomness used, the reports vary significantly. For instance, the lengths of the reports differ, the order in which the data is provided varies, and some reports repeat information in different sections of the text. However, all of them maintain the same structure provided in the prompt, as expected. Therefore, we consider that the generated reports have the expected quality and are suitable for use in NER.

On the other hand, when comparing the synthetic reports with the authentic ones, although the generated reports are coherent and grammatically correct, they can still be distinguished from each other.

Table 7
Optimal prompt used for report generation (original in Spanish and in English)

Rol	Prompt óptimo
System	"Eres un médico experto especialidado en oncología"
User	"Escribe una redacción de ámbito médico con frases cortas y concretas sobre el paciente <Patient-ID>. El texto generado debe tener la siguiente estructura de texto: <Text-Structure>. No incluir el ID del paciente en el informe."

Role	Optimal prompt
System	"You are an expert oncologist"
User	"Write a medical narrative with short and concise sentences for patient <Patinent-ID>. The generated text should have the following text structure: <Text-Structure>. Do not include the patient ID in the report."

For example, medical professionals use shorter sentences, employ more acronyms, and provide more detailed information than what we could extract from the dataset. However, both the generated and synthetic reports are annotated with the same entities and follow the same structure.

5. Experiments

In this section, we overview the experiments with fine-tuning the language models for the named entity recognition (NER) tasks. First, we explore how different combinations of real and synthetic data in the training corpus affect the models' performance.

We fine-tune four transformer [42] language models, each of them with different pretraining characteristics:

- Multilingual BERT [12]: A versatile pre-trained LM known for its multilingual support and robust performance across different domains.
- XLM-RoBERTa [43]: A multilingual version of RoBERTa [44] and enhances the capabilities of BERT in diverse tasks.
- Biomedical-Clinical RoBERTa [45]: A domain-specific model for this task in Spanish.
- BETO [46]: An extension of BERT model exclusively trained for the Spanish language.

To evaluate the effectiveness of the generated data, we used different combinations of authentic and synthetic data in the training set. These experiments can be divided into 2 types based on their objective, so many of the experiments can belong to both types, as shown in Table 8. All experiments have been evaluated with the same authentic test set, as shown in Table 1.

Table 8
Different combinations of authentic and synthetic data used

Training sets	Authentic reports		Synthetic reports
Trial 1	75	+	0
	75	+	25
	75	+	50
	75	+	100
	75	+	197
Trial 2	25	+	0
	50	+	0
	75	+	0
	0	+	197
	25	+	197
	50	+	197
	75	+	197

The first trial, composed of five experiments, used the entire training set of authentic reports and introduced different amounts of randomly selected generated data. The objective is to determine if synthetic reports can provide any improvement and how much data would be necessary.

In the second trial, composed of 7 experiments, we compared the metrics obtained using different amounts of authentic reports with and without the addition of synthetic reports with the aim of verifying their effectiveness across different corpus sizes.

6. Results

In this section, the results obtained by each experiment and model used are shown.

The metrics displayed in Figure 2 and in Table 9 represent the averages for all entities, with a strict evaluation of the entity span. Additionally, they are the average obtained after doing 3 trials for each experiment using 3 different random seeds. For the training of all models, early stopping with a patience of 10 epochs has been used, except Biomedical-Clinical RoBERTa [45], which required a patience of 15 as it did not achieve good results until after epoch 12.

It is worth nothing that with this early stopping, no model has exceeded 60 epochs in its training in any of the experiments.

6.1. Increasing Synthetic Data

The results of the first trial can be observed in Figure 2, where the amount of synthetic data in the training set has been progressively increased. As we can see, all models achieve better results when synthetic data is introduced into the authentic dataset, especially models based on RoBERTa [44], which show an increase in F1 score of between 8 and 10 points. On the other hand, the improvement achieved in BERT models is much lower, between 2 and 3 points. We can highlight that the mBERT F-Score drops considerably when adding the entire set of synthetic data (+197), which might indicate potential overfitting. However, none of the experiments show a decrease in performance compared to the baseline results (+0).

From the first insertion, where we introduced 25 reports or about 33% of the original data, the metrics stabilise, meaning that despite this data improving the results, the quantity added after 25 examples becomes irrelevant. The high lexical and stylistic similarity between synthetic reports could cause this; synthetic data could lead to greater improvement if we had generated more diverse reports using more samples as a reference.

6.2. Increasing Authentic Data

In this second trial, the effectiveness of synthetic data across different amounts of authentic data was tested. The average micro F1 score obtained, and the standard deviation for each experiment are presented in Table 9.

We observe a significant improvement when introducing synthetic data into a small training set (25 real

Figure 2: Micro-F1 obtained by adding different amounts of synthetic reports (trial 1).

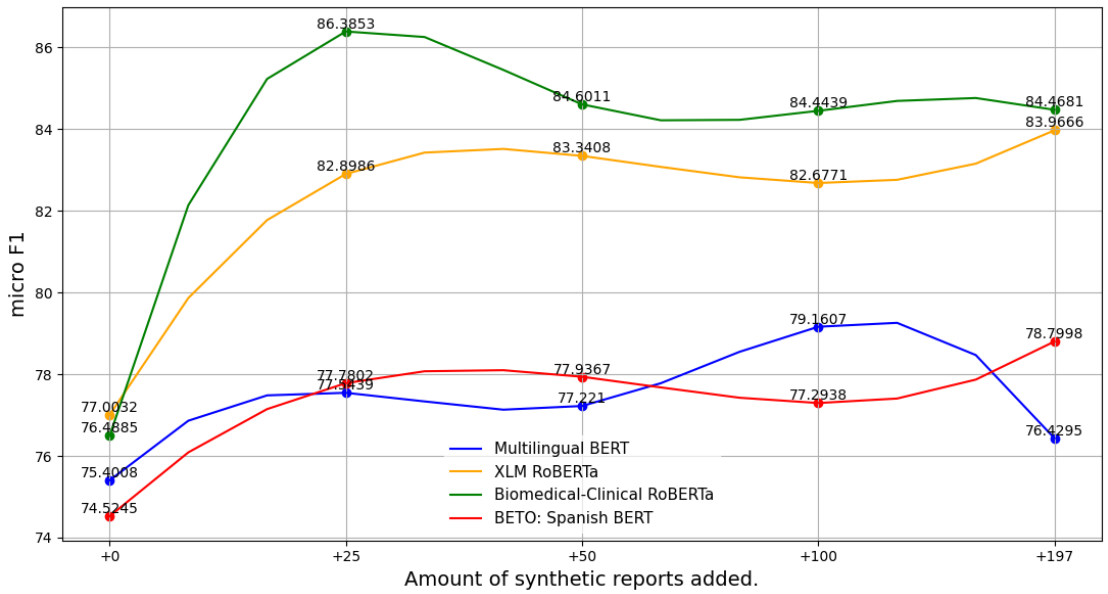

Table 9
Micro F1 score obtained by adding synthetic reports to different amounts of authentic data (trial 2).

Training sets		Language Models			
Authentic reports	Synthetic reports	mBERT	XLM-RoBERTa	Clinical RoBERTa	BETO
25	0	61.37 ± 7.07	30.60 ± 8.19	16.67 ± 5.06	49.47 ± 5.83
50	0	71.44 ± 3.24	71.04 ± 3.77	51.23 ± 1.78	70.19 ± 6.27
75	0	**75.40** ± 1.98	**77.00** ± 0.68	**76.49** ± 1.42	**74.52** ± 2.19
0	197	36.81 ± 1.42	46.31 ± 5.85	41.43 ± 1.18	39.78 ± 4.59
25	197	76.00 ± 1.32	81.95 ± 1.14	80.46 ± 0.32	78.40 ± 0.38
50	197	74.17 ± 0.78	78.83 ± 0.68	81.28 ± 1.69	**82.63** ± 3.79
75	197	**76.43** ± 0.71	**83.97** ± 0.19	**84.47** ± 1.46	78.80 ± 1.45

reports) in any of the 4 models tested. However, as in the previous trial, we can see a notable difference in the improvement obtained between the models based on RoBERTa and those based on BERT. Both XLM-RoBERTa [43] and Biomedical-Clinical RoBERTa [45] reach 80 F-Score points after the addition of synthetic reports, more than 50 points than without using them, representing the greatest improvement achieved in this trial.

On the other hand, the models mBERT [12] and BETO [46] are more robust, as although significant improvements are achieved on small datasets, we observe that using 50 reports, the F1 score already reaches 70 points without using synthetic data. Therefore, the difference between using them or not is smaller (improvement between 12 and 2 points of F1 score).

In the experiment with only synthetic data, we can observe that the obtained metrics are very low, comparable to using only 25 real reports. Therefore, we can deduce that synthetic reports are effective only when combined with real data. We can also observe that the results are less stable when training with smaller datasets, as the standard deviation exceeds 5 points in many experiments, which are only real reports. This deviation is considerably reduced when introducing synthetic data (less than 2 points on average) as the size of the training set increases significantly.

7. Conclusion and Future Work

Through the methods of transforming structured data into medical reports using a generative LLM, we have explored the benefits that such synthetic data can offer in fine-tuning of the pre-trained language models for NER tasks. We have developed a new synthetic NER corpus of 197 CT scan reports in Spanish, each from different patient. We used a structured and numerical data originating from an image dataset and took 6 samples of real reports as references.

During the experiments, we have demonstrated that the addition of synthetic data to the training set can lead to considerable improvements in the results of all tested models, especially those based on RoBERTa, one of them likely due to being trained on data from the same domain, and the other due to its large number of parameters, thus enhancing its capabilities in this type of tasks.

Our research leads to two valuable conclusions, which reveal some keys to generating effective reports. On the one hand, achieving the closest possible similarity to real data. Authentic reports typically contain a rich vocabulary, so this can be achieved by using high randomness during generation or by inserting or replacing synonyms in the text. On the other hand, maintaining minimal similarity between the generated texts so that each one contains relevant information to contribute while also avoiding overfitting. In this case, different text structures could be used in generation or even different generative models, apart of GPT-3.5-turbo.

It is worth noting that even though we apply the best techniques and models to create synthetic data, due to the textual complexity of the medical domain, there is still no technology capable of generating data that perfectly simulates real data. However, this synthetic data can be very useful when combined with authentic data.

We believe that the proposed methods can be useful for generating new datasets from information extracted from structured data, especially for languages such as Spanish, where more datasets are needed to improve the performance of Language Models in these languages.

8. Acknowledgments

This work is partially funded by the STEER project, a Multi-Area Internal initiative from Vicomtech, and the EMPHASIS project (ZE-2021/00039), supported by the Basque Business Development Agency, SPRI.

References

[1] R. Patil, V. Gudivada, A Review of Current Trends, Techniques, and Challenges in Large Language Models (LLMs), Applied Sciences 14 (2024). URL: https://www.mdpi.com/2076-3417/14/5/2074. doi:10.3390/app14052074.

[2] S. Kim, N. Fiorini, W. J. Wilbur, Z. Lu, Bridging the Gap: Incorporating a Semantic Similarity Measure for Effectively Mapping PubMed Queries to Documents, Journal of Biomedical Informatics 75 (2017) 122–127.

[3] W. W. Chapman, P. M. Nadkarni, L. Hirschman, L. W. D'avolio, G. K. Savova, O. Uzuner, Overcoming barriers to nlp for clinical text: the role of shared tasks and the need for additional creative solutions, 2011.

[4] J. Jordon, L. Szpruch, F. Houssiau, M. Bottarelli, G. Cherubin, C. Maple, S. N. Cohen, A. Weller, Synthetic Data – what, why and how?, 2022. arXiv:2205.03257.

[5] H. Surendra, H. Mohan, A review of synthetic data generation methods for privacy preserving data publishing, International Journal of Scientific & Technology Research 6 (2017) 95–101.

[6] W. Y. Wang, D. Yang, That's so annoying!!!: A lexical and frame-semantic embedding based data augmentation approach to automatic categorization of annoying behaviors using# petpeeve tweets, in: Proceedings of the 2015 conference on empirical methods in natural language processing, 2015, pp. 2557–2563.

[7] X. Zhang, J. Zhao, Y. LeCun, Character-level convolutional networks for text classification, Advances in neural information processing systems 28 (2015).

[8] A. Marzoev, S. Madden, M. F. Kaashoek, M. J. Cafarella, J. Andreas, Unnatural Language Processing: Bridging the Gap Between Synthetic and Natural Language Data, ArXiv abs/2004.13645 (2020). URL: https://api.semanticscholar.org/CorpusID:216562596.

[9] S. Y. Feng, V. Gangal, J. Wei, S. Chandar, S. Vosoughi, T. Mitamura, E. Hovy, A survey of data augmentation approaches for nlp, arXiv preprint arXiv:2105.03075 (2021).

[10] S. Kobayashi, Contextual augmentation: Data augmentation by words with paradigmatic relations, in: M. Walker, H. Ji, A. Stent (Eds.), Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 452–457. URL: https://aclanthology.org/N18-2072.

[11] J. Lichtarge, C. Alberti, S. Kumar, N. Shazeer, N. Parmar, S. Tong, Corpora Generation for Grammatical Error Correction, in: J. Burstein, C. Doran, T. Solorio (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Lan-

guage Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 3291–3301.

[12] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, CoRR abs/1810.04805 (2018). URL: http://arxiv.org/abs/1810.04805. arXiv:1810.04805.

[13] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer, BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, in: D. Jurafsky, J. Chai, N. Schluter, J. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 7871–7880. doi:10.18653/v1/2020.acl-main.703.

[14] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al., Language models are unsupervised multitask learners, OpenAI blog 1 (2019) 9.

[15] V. Kumar, A. Choudhary, E. Cho, Data augmentation using pre-trained transformer models, 2021. arXiv:2003.02245.

[16] A. Anaby-Tavor, B. Carmeli, E. Goldbraich, A. Kantor, G. Kour, S. Shlomov, N. Tepper, N. Zwerdling, Do not have enough data? deep learning to the rescue!, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, 2020, pp. 7383–7390.

[17] Y. Yang, C. Malaviya, J. Fernandez, S. Swayamdipta, R. Le Bras, J.-P. Wang, C. Bhagavatula, Y. Choi, D. Downey, Generative data augmentation for commonsense reasoning, in: T. Cohn, Y. He, Y. Liu (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2020, Association for Computational Linguistics, Online, 2020, pp. 1008–1025.

[18] Y. Meng, J. Huang, Y. Zhang, J. Han, Generating Training Data with Language Models: Towards Zero-Shot Language Understanding, in: A. H. Oh, A. Agarwal, D. Belgrave, K. Cho (Eds.), Advances in Neural Information Processing Systems, 2022. URL: https://openreview.net/forum?id=4G1Sfp_1sz7.

[19] T. Vu, M.-T. Luong, Q. Le, G. Simon, M. Iyyer, STraTA: Self-Training with Task Augmentation for Better Few-shot Learning, in: M.-F. Moens, X. Huang, L. Specia, S. W.-t. Yih (Eds.), Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 5715–5731. URL: https://aclanthology.org/2021.emnlp-main.462. doi:10.18653/v1/2021.emnlp-main.462.

[20] X. He, I. Nassar, J. Kiros, G. Haffari, M. Norouzi, Generate, Annotate, and Learn: NLP with Synthetic Text, Transactions of the Association for Computational Linguistics 10 (2022) 826–842. doi:10.1162/tacl_a_00492.

[21] F. Gilardi, M. Alizadeh, M. Kubli, Chatgpt outperforms crowd workers for text-annotation tasks, Proceedings of the National Academy of Sciences 120 (2023) e2305016120. doi:10.1073/pnas.2305016120.

[22] X. He, Z.-W. Lin, Y. Gong, A. Jin, H. Zhang, C. Lin, J. Jiao, S. M. Yiu, N. Duan, W. Chen, AnnoLLM: Making Large Language Models to Be Better Crowdsourced Annotators, ArXiv abs/2303.16854 (2023). URL: https://api.semanticscholar.org/CorpusID:257805087.

[23] P. Bansal, A. Sharma, Large language models as annotators: Enhancing generalization of nlp models at minimal cost, arXiv preprint arXiv:2306.15766 (2023).

[24] R. Zhang, Y. Li, Y. Ma, M. Zhou, L. Zou1, LLMAAA: Making Large Language Models as Active Annotators, in: Findings of the Association for Computational Linguistics: EMNLP 2023, 2023, p. 13088–13103.

[25] A. Rosenbaum, S. Soltan, W. Hamza, Y. Versley, M. Boese, Linguist: Language model instruction tuning to generate annotated utterances for intent classification and slot tagging, in: COLING 2022, 2022. URL: https://arxiv.org/abs/2209.09900.

[26] A. Gonzales, G. Guruswamy, S. R. Smith, Synthetic data in health care: A narrative review, PLOS Digital Health 2 (2023) 1–16. doi:10.1371/journal.pdig.0000082.

[27] H. Murtaza, M. Ahmed, N. F. Khan, G. Murtaza, S. Zafar, A. Bano, Synthetic data generation: State of the art in health care domain, Computer Science Review 48 (2023) 100546. doi:https://doi.org/10.1016/j.cosrev.2023.100546.

[28] M. Giuffrè, D. Shung, Harnessing the power of synthetic data in healthcare: innovation, application, and privacy, npj Digital Medicine 6 (2023). doi:10.1038/s41746-023-00927-3.

[29] B. Jing, P. Xie, E. Xing, On the automatic generation of medical imaging reports, in: I. Gurevych, Y. Miyao (Eds.), Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 2577–2586. URL: https://aclanthology.org/P18-1240. doi:10.18653/v1/P18-1240.

[30] J. Ive, N. Viani, J. Kam, L. Yin, S. Verma, S. Puntis, R. N. Cardinal, A. Roberts, R. Stewart, S. Velupillai, Generation and evaluation of artificial mental health records for natural language processing, NPJ digital medicine 3 (2020) 69.

[31] I. Sutskever, O. Vinyals, Q. V. Le, Sequence to sequence learning with neural networks, in: Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, K. Weinberger (Eds.), Advances in Neural Information Processing Systems, volume 27, Curran Associates, Inc., 2014. URL: https://proceedings.neurips.cc/paper_files/paper/2014/file/a14ac55a4f27472c5d894ec1c3c743d2-Paper.pdf.

[32] J. Li, Y. Zhou, X. Jiang, K. Natarajan, S. V. Pakhomov, H. Liu, H. Xu, Are synthetic clinical notes useful for real natural language processing tasks: A case study on clinical entity recognition, Journal of the American Medical Informatics Association 28 (2021) 2193–2201.

[33] W. Ling, C. Dyer, A. W. Black, I. Trancoso, R. Fermandez, S. Amir, L. Marujo, T. Luís, Finding function in form: Compositional character models for open vocabulary word representation, in: L. Màrquez, C. Callison-Burch, J. Su (Eds.), Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Lisbon, Portugal, 2015, pp. 1520–1530. URL: https://aclanthology.org/D15-1176. doi:10.18653/v1/D15-1176.

[34] X. Chen, Y. Li, P. Jin, J. Zhang, X. Dai, J. Chen, G. Song, Adversarial sub-sequence for text generation, 2019. arXiv:1905.12835.

[35] N. S. Keskar, B. McCann, L. Varshney, C. Xiong, R. Socher, CTRL - A Conditional Transformer Language Model for Controllable Generation, arXiv preprint arXiv:1909.05858 (2019).

[36] R. Tang, X. Han, X. Jiang, X. Hu, Does synthetic data generation of llms help clinical text mining?, 2023. arXiv:2303.04360.

[37] R. Ackerman, R. Balyan, Automatic multilingual question generation for health data using llms, in: International Conference on AI-generated Content, Springer, 2023, pp. 1–11.

[38] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, Journal of Machine Learning Research 21 (2020) 1–67. URL: http://jmlr.org/papers/v21/20-074.html.

[39] B. Jehangir, S. Radhakrishnan, R. Agarwal, A survey on Named Entity Recognition — Datasets, Tools, and Methodologies, Natural Language Processing Journal 3 (2023) 100017.

[40] B. Kaabachi, J. Despraz, T. Meurers, K. Otte, M. Halilovic, F. Prasser, J. L. Raisaro, Can we trust synthetic data in medicine? a scoping review of privacy and utility metrics, medRxiv (2023). URL: https://www.medrxiv.org/content/early/2023/11/28/2023.11.28.23299124. doi:10.1101/2023.11.28.23299124.

[41] A. L. Simpson, J. Peoples, J. M. Creasy, G. Fichtinger, N. Gangai, A. Lasso, K. N. Keshava Murthy, J. Shia, M. I. D'Angelica, R. K. G. Do, Preoperative ct and survival data for patients undergoing resection of colorectal liver metastases (colorectal-liver-metastases), 2023. URL: https://www.cancerimagingarchive.net/collection/colorectal-liver-metastases/. doi:10.7937/QXK2-QG03.

[42] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is All You Need, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17, Curran Associates Inc., Red Hook, NY, USA, 2017, p. 6000–6010.

[43] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, CoRR abs/1911.02116 (2019). URL: http://arxiv.org/abs/1911.02116. arXiv:1911.02116.

[44] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized BERT pretraining approach, CoRR abs/1907.11692 (2019). URL: http://arxiv.org/abs/1907.11692. arXiv:1907.11692.

[45] C. P. Carrino, J. Llop, M. Pàmies, A. Gutiérrez-Fandiño, J. Armengol-Estapé, J. Silveira-Ocampo, A. Valencia, A. Gonzalez-Agirre, M. Villegas, Pretrained biomedical language models for clinical NLP in Spanish, in: Proceedings of the 21st Workshop on Biomedical Language Processing, Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 193–199. URL: https://aclanthology.org/2022.bionlp-1.19. doi:10.18653/v1/2022.bionlp-1.19.

[46] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, J. Pérez, Spanish pre-trained bert model and evaluation data, in: PML4DC at ICLR 2020, 2020.