Get the Best out of 1B LLMs: Insights from Information Extraction on Clinical Documents

Saeed Farzi*, **Soumitra Ghosh***, **Alberto Lavelli** and **Bernardo Magnini**
Fondazione Bruno Kessler, Via Sommarive 18, Povo, Trento, Italy
{sfarzi, sghosh, lavelli, magnini}@fbk.eu

Abstract

While the popularity of large, versatile language models like ChatGPT continues to rise, the landscape shifts when considering open-source models tailored to specific domains. Moreover, many areas, such as clinical documents, suffer from a scarcity of training data, often amounting to only a few hundred instances. Additionally, in certain settings, such as hospitals, cloud-based solutions pose privacy concerns, necessitating the deployment of language models on traditional hardware, such as single GPUs or powerful CPUs. To address these complexities, we conduct extensive experiments on both clinical entity detection and relation extraction in clinical documents using *1B parameter* models. Our study delves into traditional fine-tuning, continuous pre-training in the medical domain, and instruction-tuning methods, providing valuable insights into their effectiveness in a multilingual setting. Our results underscore the importance of domain-specific models and pre-training for clinical natural language processing tasks. Furthermore, data augmentation using cross-lingual information improves performance in most cases, highlighting the potential for multilingual enhancements.

1 Introduction

In the last few years the deep learning revolution has produced significant changes in information extraction (IE) from clinical text. Pre-trained large language models (LLMs) based on attention and transformer architectures (e.g., BERT, T5, etc.) have become popular mainly due to their superior performance with respect to traditional machine learning approaches. Fine-tuning on downstream tasks has been the standard approach used to transfer general pre-trained knowledge to specific tasks of interest, including information extraction from clinical documents. In addition to fine-tuning, continuous pre-training (Gururangan et al., 2020) has shown to be effective to adapt a LLM to the medical domain or to a specific set of languages. Recently, very large language models have further increased both the amount of data used in the pre-training phase, and the complexity of the model parameters. The resulting models (e.g., GPT-4) achieve high performance with few-shot or even zero-shot in-context learning techniques (i.e., prompting) (Liu et al., 2023). Finally, instruction-tuning (Zhang et al., 2023) has emerged as a powerful approach to align pre-trained LLMs to human expectations for a number of natural language processing (NLP) and conversational tasks, further improving usability and performance of LLMs.

Although there is a clear trend towards large language models with general purpose conversational abilities (e.g., ChatGPT), when the choice is constrained to open source models for a domain-specific downstream task (more than often in a low-resource setting), the current landscape of solutions is rather restricted. In addition, there are good reasons to constraint application solutions to small models, particularly because they are computationally manageable, avoiding the need of expensive hardware or to move sensitive data on the cloud. Given the above considerations, there is a lack of consensus on what would be the best solution.

With the aim of shedding light in the current LLM landscape, in this paper we investigate how available small LLMs perform on fine-tuning, continuous pre-training and instruction-tuning on information extraction from clinical documents. We consider LLMs that are available open source, are within the range of 1B parameters, and that are available in several versions, allowing to investigate the impact of multilinguality and instruction-tuning. Our experiments encompass English and Italian datasets for clinical entity detection, and Italian and Spanish for relation extraction, addressing

*These authors contributed equally to this work.

Proceedings of the 23rd Workshop on Biomedical Language Processing, pages 266–276

two primary research questions:

- Is continuous pre-training, both on languages and domain, effective on our tasks and domains? Does it allow to reduce the need of fine-tuning data in our low-resource setting?
- Is general purpose instruction-tuning effective? Is it competitive with continuous pre-training, both on domain and languages?

In addition to *core experiments* on small LLMs, we conducted additional experiments aiming at assessing the role of data augmentation on the same models and tasks. Data augmentation is a common practice to boost performance, and we are interested in either merging or translating datasets of different languages, as they are becoming more and more available, although in limited amounts.

The primary contributions of the paper include: (i) comparing fine-tuning, continuous pre-training, and instruction-tuning of the same pre-trained model on two NLP tasks, a novel comparison to our knowledge; (ii) investigating the relations between continuous pre-training on languages and continuous pre-training on a specialized domain, suggesting a promising research direction; (iii) demonstrating that, through accurate parameter optimization, language models with 1B parameters remain competitive, although absolute performance was not the primary focus; and (iv) indicating that language-based data augmentation enhances performance in our low-resource setting.

2 Background

2.1 Pre-trained Language Models

In recent years, there has been extensive research on LLMs owing to their capacity for pre-training, allowing them to learn from vast amounts of data in a self-supervised manner. These models have demonstrated remarkable performance across various NLP tasks (Howard and Ruder, 2018; Radford et al., 2019). Scaling LLMs, either by increasing model size or training data, often enhances their capacity for downstream tasks. Several studies have explored the performance limits through scaling, primarily focusing on enlarging model size while maintaining similar architectures and pre-training tasks. Continuous pre-training has emerged as a method to enhance LLM performance in specific domains (Gururangan et al., 2020).

2.1.1 Instruction Tuning

A significant issue with LLMs is the discrepancy between their training objective and users' needs: while LLMs are typically trained to minimize contextual word prediction errors on large datasets, users expect the model to "follow their instructions helpfully and safely". Instruction tuning (Khashabi et al., 2020; McCann et al., 2018) is proposed as a technique to enhance the capabilities and controllability of LLMs. It consists in training LLMs using (INSTRUCTION, OUTPUT) pairs, where INSTRUCTION denotes the human instruction for the model, and OUTPUT the desired output that follows the INSTRUCTION. Instruction tuning bridges the gap between the next-word prediction objective of LLMs and users' objectives of instruction following, thereby increasing controllability and predictability. Additionally, it is computationally efficient and aids LLM adaptation to specific domains.

2.2 LLMs and Information Extraction

Named Entity Recognition (NER) is a key NLP task involving the identification and classification of entities within text. Early methods relied on rule-based systems and manual dictionaries for entity identification (Petasis et al., 2001; Ruokolainen et al., 2020). A significant advancement in NER came with the introduction of Conditional Random Fields (CRFs), which effectively addressed sequence labeling tasks (Lafferty et al., 2001). The emergence of transformer-based models such as BERT (Devlin et al., 2019), GPT-3 (Brown et al., 2020), and their specialized variants has revolutionized NER by capturing contextual information adeptly. These models exhibit remarkable performance across various domains. Furthermore, NER extends to multilingual and cross-lingual scenarios (Zanoli et al., 2023), where models like XLM-R have proven effective (Conneau et al., 2020).

Relation extraction (RE) is concerned with the identification and categorization of relationships between entities mentioned in text (Mintz et al., 2009). In this paper our focus is on the extraction of relationships from clinical documents.

2.3 IE from Clinical Documents

Historically, medical concept extraction began with rule-based systems like MetaMap (Aronson and Lang, 2010) or hybrid systems (rule-based and ML) cTAKES (Savova et al., 2010) but faced challenges with the complexity of clinical text. Dictionary

Lookup, another popular approach (Doğan et al., 2014), relied on exact matching with predefined clinical terms but lacked robustness to term variability. BANNER, an open-source biomedical NER system (Leaman and Gonzalez, 2008), emerged as a domain-independent solution, surpassing baseline systems and serving as a benchmark. Deep learning models, notably CNN-based architectures (Zhu and Wang, 2019), have enhanced NER by leveraging neural networks' sequential insight.

Transformer-based models like BERT (Devlin et al., 2019) have further improved NER performance in clinical documents, adapted by researchers for biomedical NER (Michalopoulos et al., 2021; Lamproudis et al., 2022). XLM-RoBERTa model experimentation on the E3C corpus (Zanoli et al., 2023) showcased its efficacy in various setups. Recent trends show a shift towards employing transformer models in diverse roles, including pipeline systems and Seq2Seq models (Wang and Lu, 2020; Yamada et al., 2020).

The specialized nature of medical texts requires tailored research and model adaptation. Our work addresses this need by focusing on small generative language models, particularly T5 and its variants (Raffel et al., 2020), for NER and RE tasks within clinical documents. These models offer a resource-efficient alternative, aligning with the practical requirements of the medical domain.

3 Core Experiments

In this section, we present core experiments comparing four model versions across two information extraction tasks (NER and RE) in clinical documents with limited training data and across three different languages.

3.1 Experimental Design

We address the relations among fine-tuning, continuous pre-training and instruction-tuning using models with "1B parameters". The experimental design includes: (i) four versions of a "1B parameters" generative model: a base version (T5), a version with continuous pre-training on several languages (mT5), a version with continuous pre-training on the medical domain (MedMT5), and a version which has been instruction-tuned on general NLP tasks (FLAN-T5). Full fine-tuning approaches have been employed to train the language models to tackle NER and RE tasks whereas for Flan-T5, the prompt fine-tuning approach has been adopted. We run the four models on two IE tasks on clinical documents, NER and RE. For each task we provide results both on a dataset with low-resource data and on a dataset with high-resource data. Finally, experiments cover three languages: English, Italian and Spanish.

A core question behind our experiments on small models is the following: does instruction-tuning overcome the need for continuous pre-training (on languages and domain) on our core models applied to our experimental setting?

3.2 Task 1: Clinical Entity Detection

This task consists in identifying relevant clinical entities from clinical texts, such as patient records, medical reports, and clinical notes. Unlike scientific publications, which focus on research findings, clinical notes encompass documents that report various aspects of clinical practice, including the rationale for a clinical visit, descriptions of physical examinations, assessments of the patient's condition, diagnosis, and subsequent treatment plans. For instance, consider a clinical note:

"Patient John Doe, a 45-year-old male, was admitted on July 15, 2023, with complaints of chest pain. He has a medical history of hypertension and diabetes. During the examination, his blood pressure measured 150/90 mm Hg, and his blood glucose level was 180 mg/dL."

Entity detection here targets essential information, including:

- *Patient Information:* "John Doe", "45-year-old male"
- *Admission Date:* "July 15, 2023"
- *Chief Complaint:* "Chest pain"
- *Medical History:* "hypertension", "diabetes"
- *Vital Signs:* "blood pressure 150/90 mm Hg", "blood glucose 180 mg/dL"

We frame the Clinical Entity Detection task as a text-to-text generation task, emphasizing the identification and labeling of textual spans as named entities within a context. We use the following two datasets.

European Clinical Case Corpus (E3C). This is a dataset of clinical cases already published in journals, covering Spanish, Basque, English, French, and Italian (Magnini et al., 2021, 2022). The annotations focus on both clinical entities, specifically disorders, as classified in UMLS taxonomy, and temporal expressions following the THYME standard. For our experiments, we utilize the pre-processed E3C corpus from (Zanoli et al., 2023),

conducting experiments on the English (E3C_En) and Italian (E3C_It) datasets, as outlined in Table 1. We acknowledge that the E3C clinical notes are an idealized version of real-world notes, but they offer a privacy-compliant alternative.

NCBI Disease Corpus. NCBI (Doğan et al., 2014) includes 6,892 mentions of disease names and their corresponding identifiers in 793 PubMed abstracts. Categorized mentions allow flexible matching to MeSH and OMIM concepts, while preserving intended meaning. High inter-annotator agreement and low ambiguity make NCBI a strong foundation for machine learning systems, benefiting biomedical knowledge discovery. Table 3 presents the distribution of disease mentions in the NCBI dataset over the training, dev and test sets.

3.3 Task 2: Test-Result Relation Extraction

This task consists of identifying the relations between laboratory tests and their measurements within a clinical note. Building on recent advancements in the field, we approach test-result relation extraction as a form of text summarization, leveraging text-to-text transformer-based models. The key idea is to represent relations as summarized text of a given input as illustrated in Table 2. For the given clinical note, the summarized text is "<EVENT>creatininemia<RESULT> pari o inferiori a 1.5 mg/dl and <EVENT>ipercolesterolemia<RESULT> *280 mg/dl*". We use the following datasets.

CLinkaRT and TESTLINK datasets. We rely on data sourced from the Italian and Spanish sections of the E3C Corpus, respectively, CLinkaRT (Altuna et al., 2023b) and TESTLINK (Altuna et al., 2023a), which introduce relation extraction in the context of clinical cases. Table 2 reports an example extracted from the CLinkaRT dataset (Altuna et al., 2023b), along with its associated relations between medical laboratory tests and their respective results. Each relation comprises an event, the associated result, and their corresponding positions within the text. For example, in the first relation, "creatininemia" is found within positions [286 - 281], with value "pari o inferiori a 1.5 mg/dl." Table 4 reports some statistics about the CLinkaRT and TESTLINK datasets.

	Training		Test	
Language	**Gold**	**Pre-processed**	**Gold**	**Pre-processed**
English	463	437	561	516
French	596	569	731	695
Italian	361	345	508	461*
Spanish	525	509	820	800
Basque	846	835	1064	1054

Table 1: Entity distribution over E3C languages. [*] We found 460 entities in the GitHub link instead of 481 as reported in (Zanoli et al., 2023).

3.4 Models

We focus on "1B parameters" models, because: (i) fine-tuning is manageable with limited computational infrastructure, often available in industry and academy; (ii) inference can be performed without need of dedicated hardware, which is a great advantage when data can not be transferred on the cloud (e.g., hospitals). Although there might be several options (e.g., BERT models), for our experiments we used T5 models (Raffel et al., 2020), because there are several versions available and they show competitive performance. We report the main characteristics of the T5 models in Table 5.

3.4.1 T5

T5 (Text-to-Text Transfer Transformer) (Raffel et al., 2020) employs a transformer architecture with shared encoder-decoder parameters and undergoes pretraining on extensive text data followed by fine-tuning for specific tasks, ensuring versatility across NLP tasks. Unlike BERT (Devlin et al., 2019), which predicts masked words, T5 formulates tasks as text-to-text problems, leading to superior performance across various benchmarks.

3.4.2 mT5

mT5, or "Multilingual Text-to-Text Transfer Transformer" (Xue et al., 2021), is a multilingual variant of the T5 model pretrained in an unsupervised manner on a diverse multilingual corpus, supporting 101 languages. Demonstrating impressive performance on tasks like translation (Patel et al., 2022), lemmatization (Ulčar and Robnik-Šikonja, 2023), and text simplification (Gonzalez-Dios et al., 2022), mT5 showcases its versatility and effectiveness across different language tasks.

3.4.3 MedMT5

MedMT5 (García-Ferrero et al., 2024) is an encoder-decoder model developed by continuing the training of the mT5 (Xue et al., 2021) checkpoints on a medical domain corpus that includes 3B words in four languages (English, Spanish, French,

Example Clinical Note: Il decorso clinico era stato caratterizzato da un rigetto acuto nel primo mese post-trapianto e da alcuni episodi di tachicardia parossistica sopraventricolare negli anni successivi. La funzionalità renale, dopo l'episodio di rigetto, si era stabilizzata su valori di creatininemia pari 0 inferiori a 1.5 mg/dl. L'esame delle urine non aveva mai evidenziato proteinuria. Era presente da anni ipercolesterolemia (280 mg/dl).				
	Position of Result	**Result**	**Position of Event**	**Event**
Relation 1	282-310	pari o inferiori a 1.5 mg/dl	286-281	creatininemia
Relation 2	414-422	280 mg/dl	393-411	ipercolesterolemia

Table 2: An example from the CLinkaRT dataset.

Training	**Development**	**Test**	**Total**
5145	787	960	6892

Table 3: Disease mentions distribution in NCBI.

Datasets	**Docs**	**Relations**	**Unique Events**	**Unique Results**
CLinkaRT-Train (IT)	83	619	344	410
CLinkaRT-Test (IT)	80	612	332	407
TESTLINK-Train (SP)	81	597	317	332
TESTLINK-Test (SP)	80	668	340	421

Table 4: Statistics about the CLinkaRT dataset. IT: Italian, SP: Spanish

and Italian). It is the first open-source text-to-text multilingual model for the medical domain.

3.4.4 FLAN-T5

FLAN-T5 (Chung et al., 2022) is an instruction-tuned language model that excels in NLP tasks by training on diverse instructions, enabling it to handle a wide range of tasks. Mixing zero-shot, few-shot, and chain of thought prompts during training enhances FLAN-T5's performance, even on tasks not seen during fine-tuning, making it excel in both held-in and held-out tasks.

3.5 Experimental Setup

Both for Named Entity Recognition and Relation Extraction the core approach is based on text-to-text generation using the T5 models. The loss functions utilized for both tasks are the standard cross-entropy losses associated with T5 models.

3.5.1 NER Task

We maintained consistent hyperparameters for all models, including a batch size of 4, a maximum token length of 256 for input and output, epochs 30, 0.05 dropout, a warmup ratio of 0.06, and an epsilon of 1e-8 for Adam optimization. The remaining parameters used default values from the SimpleTransformers[1] library, and a seed[2] value of 32 ensured result reproducibility. While hyperparameter tuning was not exhaustive, we explored varying learning rates (1e-4, 2e-4, 3e-4, 2e-5, and, 3e-5). The most suitable learning rate (for all models) was observed to be 1e-4.

3.5.2 RE Task

We adhered to consistent hyper-parameters across all models during training, including a batch size of 2, a maximum token length of 128 for both input and output sequences, a training duration of up to 100 epochs with early stopping, a learning rate of 4e-5, a gradient accumulation step of 4, a dropout rate of 0.1, a warm-up step count of 500, and an epsilon value of 1e-8 for the Adam optimization. Default values from the Hugging Face library were used for the remaining parameters, and a seed value of 42 was employed to ensure result reproducibility.

All models were trained on an NVIDIA A40 GPU with 48 GB GDDR6 memory.

4 Results and Discussion

In this section we present the results of the core experiments on the two tasks, NER and RE.

4.1 Results on NER

Table 6 illustrated the results of our experiments with various T5 models for the NER task on the E3C and NCBI datasets. The performance of T5, mT5, MedMT5, and FLAN-T5 models on various datasets highlights key insights. The base T5 model shows a relatively high recall on E3C-English, indicating it can identify a large number of relevant entities, but its precision is lower, leading to a moderate F1 score. For E3C-Italian, the precision is higher than recall, but the F1 score remains balanced. On the NCBI dataset, T5 achieves strong precision and recall, resulting in a high F1 score. The multilingual mT5 model performs slightly better than T5 in terms of precision on E3C-English, but its recall is lower, leading to a slightly lower

[1] https://simpletransformers.ai/

[2] We chose a single seed to ensure consistent results and simplify model comparisons. We plan to conduct additional experiments with different random seeds.

Model	Architecture	Parameters	# languages	Data Source
T5 (Raffel et al., 2020)	encoder-decoder	770M	1 (English)	Colossal Clean Crawled Corpus (C4)
mT5 (Xue et al., 2021)	encoder-decoder	1.2B	101	Multilingual C4 (mC4)
MedMT5 (García-Ferrero et al., 2024)	encoder-decoder	738M	4	Multilingual
FLAN-T5 (Chung et al., 2022)	encoder-decoder	780M	60	473 datasets (SQuAD, MNLI, WMT-16, etc.)

Table 5: Comparison of T5 models.

Models	E3C						NCBI		
	English			Italian			English		
	P	R	F_1	P	R	F_1	P	R	F_1
Baselines									
DLU (Zanoli et al., 2023; Doğan et al., 2014)	37.08	60.08	45.86	48.46	61.52	54.21	21.3	71.8	31.6
CRF (Lafferty et al., 2001)	51.81	30.43	38.34	65.88	42.39	51.59	-	-	-
Inference Method (Doğan et al., 2014)	-	-	-	-	-	-	59.7	73.1	63.7
State-of-the-art									
BANNER (Leaman and Gonzalez, 2008)	-	-	-	-	-	-	83.8	80.00	81.8
XLM-RoBERTa-PL (Zanoli et al., 2023)	45.67	60.66	52.12	60.31	67.39	63.65	-	-	-
XLM-RoBERTa-CL (Zanoli et al., 2023)	40.81	60.27	48.67	43.28	70.00	53.49	-	-	-
T5-Large Family *(ours)*									
T5	51.50	66.47	58.04	63.22	58.91	61.04	85.65	82.88	84.24
mT5	53.68	57.95	55.73	62.53	60.22	61.35	83.72	78.39	80.97
MedMT5	53.94	66.28	59.48	64.74	70.00	**67.29**	86.70	82.99	84.80
FLAN-T5	53.44	69.19	**60.30**	60.79	63.70	62.21	86.80	83.09	**84.91**
mT5 (data augmented)	54.94	51.74	53.29	58.68	61.74	60.17	-	-	-
MedMT5 (data augmented)	55.65	63.95	**59.51**	64.24	71.09	**67.49**	-	-	-

Table 6: Results for Entity Recognition task. DLU: Dictionary look-up. Highest obtained scores among the T5 variants are highlighted in bold.

Models	CLinkaRT			TESTLINK		
	Italian			Spanish		
	P	R	F_1	P	R	F_1
Baselines						
voc. tran. (Altuna et al., 2023b,a)	29.95	31.86	30.88	17.41	30.24	22.10
GPT (Altuna et al., 2023b,a)	29.55	48.73	36.79	25.24	38.29	30.43
mBERT (Altuna et al., 2023b,a)	61.37	64.37	62.83	61.13	60.03	60.57
State-of-the-art						
ExtremITA-T5 (Hromei et al., 2023)	46.82	26.47	33.82	-	-	-
Simple Ideas-BERT (Micluta-Campeanu and Dinu, 2023a)	65.55	60.62	62.99	-	-	-
LinkMed6 (Muñoz-Castro et al., 2023)	-	-	-	46.99	43.26	45.05
Simple Ideas (Micluta-Campeanu and Dinu, 2023b)	-	-	-	71.45	65.57	68.38
T5-Large Family *(ours)*						
T5	53.20	48.03	50.51	58.66	36.97	45.36
mT5	65.72	53.26	58.84	55.96	50.59	53.14
MedMT5	65.22	59.15	**62.03**	62.28	54.64	**58.21**
FLAN-T5	52.99	49.18	51.01	58.03	41.61	48.47
mT5 (data augmented)	69.56	49.67	57.95	66.22	51.94	58.22
MedMT5 (data augmented)	71.72	56.37	**63.12**	70.76	52.54	**60.30**

Table 7: Results on the Relation Extraction task on clinical data, both core and augmented models.

F1 score. For E3C-Italian, mT5 has a more balanced precision and recall, resulting in a similar F1 score to T5. On the NCBI dataset, mT5's performance is slightly lower than T5 in all metrics. The MedMT5 model, pre-trained on medical data, shows improvements over both T5 and mT5. On E3C-English, it achieves higher recall and F1 scores. For E3C-Italian, MedMT5 significantly improves both precision and recall, resulting in the highest F1 score among the models. The performance on the NCBI dataset is also slightly better than T5. The instruction-tuned FLAN-T5 model achieves the highest F1 score on E3C-English, suggesting its effectiveness for this dataset. For E3C-Italian, its performance is slightly lower than MedMT5 but still strong. On the NCBI dataset, FLAN-T5 performs similarly to MedMT5, maintaining high precision and recall.

4.1.1 Discussion

Dataset size is indeed the primary reason behind the observed performance gap between the E3C and NCBI datasets. Additionally, considering our understanding of the two corpora and their annotation strategies, we comprehend that the E3C corpus is much more complex than the NCBI dataset. This complexity arises from the diverse range of medical concepts annotated in the E3C corpus, including disorders such as diseases or syndromes, findings, injuries or poisoning, and signs or symptoms, whereas the NCBI dataset primarily focuses on disease terms.

Fine-tuning, as traditionally done with models like T5 and mT5, continues to be a reliable approach and may yield superior results when abundant data and computing resources are available. The lagging performance of mT5 compared to T5 on English-only datasets can be attributed to T5's specialization and optimization specifically for the English language, which provides it with a distinct advantage.

MedMT5, designed for the medical domain, performs competitively in general NER tasks, suggesting the potential for domain-specific pre-training in specialized areas. Additionally, model versatility, as seen in FLAN-T5 and specialized models like MedMT5, is a key consideration in selecting the most suitable LLM for a particular NLP task.

The observed difference in performance between instruction-tuning and domain-specific pre-training may indeed stem from the size of the inherent pre-training or instruction-tuning datasets used for the models. Specifically, the domain-specific pretrained model (MedMT5) is trained on a larger corpus of Italian data compared to English. In contrast, the instruction-tuned model (FLAN-T5) may have a higher representation of English. This discrepancy in dataset composition could explain the superior performance of the domain-specific pretrained model on the Italian dataset, while the instruction-tuned model excels on the English datasets.

4.1.2 Error Analysis

Our error analysis revealed two key observations: Firstly, models struggle with interpreting abbreviations like "TAO", "NSIAD", "MIC", etc., often mislabeling them as entities or non-entities, indicating challenges in accurately recognizing and interpreting abbreviations. Secondly, the model erroneously labels "metastasis from adenocarcinoma" as a single entity, failing to recognize "metastasis" and "adenocarcinoma" as separate entities, suggesting a lack of contextual understanding and a tendency to group consecutive tokens into a single entity. This tendency to include stop words within entities contributes to a decrement in overall precision. To address these issues, we propose fine-tuning the model on a larger dataset to enhance abbreviation recognition and contextual understanding, along with improving the accuracy of identifying entity boundaries for enhanced precision.

4.2 Results on RE

As reported in Table 7, within the T5 family, MedMT5 models demonstrate a clear superiority over other family members in both languages, with the exception of Italian, where mT5 exhibits a slightly advantage in terms of precision.

4.2.1 Discussion

When comparing MedMT5 with other models, including baselines and state-of-the-art approaches, it is evident that MedMT5 achieves comparable results in terms of F_1 score across both languages. Notably, most models employ data augmentation, including the mBERT-based approach by Altuna et al. (2023b), which utilizes oversampling techniques for relation classification. In contrast, MedMT5 does not employ additional data. ExtremITA-T5, based on IT5 trained on Italian text from the public domain, performs well in certain NLP tasks (Hromei et al., 2023), but falls short compared to MedMT5, especially in the medical domain.

In terms of system complexity, all models follow a dual-model approach, with one model dedicated to named entity recognition and the other to relation classification in a pipeline manner. In contrast, MedMT5 functions as a generative model in an end-to-end manner. Notably, for mention-level relation extraction tasks such as CLinkaRT and TESTLINK, pipeline approaches do not necessitate position determination during post-processing, underscoring a strength of pipeline systems over generative models in these scenarios.

4.2.2 Error Analysis

Analysis of errors in the relation extraction system reveals two primary sources of errors. Firstly, errors stem from relation positioning, where event and result positions are calculated during post-processing. This involves gathering all input sentences and corresponding model-generated responses, determining sentence length, locating event and result positions, and selecting the closest occurrences if multiple exist. Finally, we compute the precise positions of the events and results.

Secondly, errors arise from partially accurate relations, wherein accurate relations contain one or two erroneously generated letters by the model in either the result or event. For instance, in a generated relation: "7,1 mg/dl**e** <– bilirrubina", the letter "e" is generated unnecessarily (the correct relation is "7,1 mg/dl <– bilirrubina"). These types of errors, rooted in data sparsity, significantly impact on system performance. Partially accurate relations are typically encountered in the context of infrequent or rare events or results. As we compare the MedMT5 and mT5 models outputs, it is evident that MedMT5 exhibits a notably lower count of partially accurate relations compared to mT5 models. This implies that the unsupervised learning approach employed by MedMT5 equips the model with certain in-domain lexicons.

Another notable observation in the outputs concerns the impact of input length on performance. Longer input sentences containing numerous relations tend to result in poor performance for most models, with MedMT5 notably outperforming the others. Our experiments involved exploring both longer sentences and sentences split into shorter ones, revealing a significant enhancement in results with shorter sentences. To mitigate this challenge, a potential solution is to implement a sliding window approach on the input to reduce its length. However, the choice of window size becomes a crucial factor, which we plan to investigate in future research efforts.

5 Data Augmentation Experiments

Here we present additional results on two tasks obtained through data augmentation on the core models discussed in section 3. Our aim is to explore potential correlations between the core and augmented models.

5.1 Data Augmentation on NER

To examine how the T5 model's performance is influenced by cross-lingual data augmentation, we trained the mT5 and MedMT5 models on datasets that included both the English and Italian E3C training sets and subsequently evaluated their performance on the English and Italian E3C test sets. We present the results in Table 6.

Data augmentation for mT5 increases precision on E3C-English but reduces recall, resulting in a lower F1 score compared to the non-augmented mT5. For E3C-Italian, the recall improves, and the F1 score remains comparable. Data augmentation enhances precision for MedMT5 on E3C-English and improves recall on E3C-Italian. However, the overall improvement in F1 scores is marginal in both E3C datasets. While data augmentation can improve certain metrics, its impact is mixed and dataset-dependent. It is most beneficial when it enhances recall without significantly compromising precision, as seen with MedMT5 on E3C-Italian.

5.2 Data Augmentation on RE

Using translation data augmentation is indeed a common technique to leverage cross-lingual information and improve the performance of NLP models, including those used in specific domains such as medical. This approach allows models to generalize across languages and learn from diverse multilingual datasets. To expand the training dataset through translation-based data augmentation, the Spanish training data is translated into Italian using Google Translate then it is utilized as augmented data for an Italian task, and vice versa. Table 7 demonstrates a substantial performance boost in terms of precision for both languages. Specifically, there is an enhancement of almost 6 points for Italian and around 8 points for Spanish. This is influenced by two crucial aspects of the datasets. Firstly, the presence of numerous identical relations in both datasets enhances precision. Secondly, the introduction of translation errors in the training data

hampers the model's capability to generate rare relations. In summary, the abundance of similar relations contributes to improved precision, while translation errors negatively impact the model's ability to produce less common relations

6 Conclusions

In a world dominated by large language models, this work delves into the efficacy of smaller, domain-specific models in the context of fine-tuning, continuous pre-training, and instruction-tuning on clinical information extraction tasks. Our findings suggest that, while general-purpose instruction-tuning offers versatility, it may not always be as effective as continuous pre-training in domain-specific tasks. We observed instances where instruction-tuning (FLAN-T5) yielded competitive results, but its performance varied across languages and domains. While models like MedMT5 designed for the medical domain outperform general-purpose counterparts in NER and RE, we find that instruction-tuning varies in effectiveness across languages and domains, emphasizing the importance of domain-specific continuous pre-training. This highlights the need for careful consideration when selecting the most suitable approach for a particular NLP task, weighting factors such as data availability, domain specificity, and computational resources. In a landscape where bigger is often seen as better, our work emphasizes the value of smaller, versatile models in scenarios prioritizing data privacy and traditional hardware.

In our future work, we plan to assess parameter-optimized strategies such as PEFT, LORA, QLORA, and LLAMA-Adapter for training larger models on traditional hardware efficiently. This exploration aims to advance model scalability while considering computational constraints, particularly in resource-limited environments.

Limitations of the Study

Concerning relation extraction, our focus was on the CLinkaRT and TESTLINK tasks, which involve identifying test results and measurements and linking them to corresponding textual mentions of clinical laboratory tests. We specifically concentrated on discovering relations between clinical laboratory tests and their results. To the best of our knowledge, there is currently no relation extraction dataset derived from E3C that encompasses a wider variety of relation types.

Regarding the entity detection results, it is important to note that our experimental datasets contain only one type of entity, and thus the reported scores pertain specifically to that entity type. We acknowledge the value of providing results for individual entity types and will consider incorporating this in future iterations of our work.

Furthermore, we agree that variations in T5 models, such as instruction-tuned and domain-specific pre-trained versions, could potentially influence results due to differences in language coverage. While our current evaluation focuses on overall entity detection performance, we acknowledge the potential impact of these variations on the results. In our future work, we plan to conduct a more comprehensive analysis to explore how different T5 models, including instruction-tuned and domain-specific pre-trained variants, perform across various entity types.

Ethical Considerations

The datasets employed in this study, while residing within the clinical domain, do not contain sensitive or personally identifiable information. These datasets are publicly accessible and openly available for research purposes.

Acknowledgments

This work has been partially funded by the European Union under the Horizon Europe eCREAM Project (Grant Agreement No.101057726) and IDEA4RC Project (Grant Agreement No.101057048). Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the European Health and Digital Executive Agency (HADEA). Neither the European Union nor the granting authority can be held responsible for them. This work was supported by the CHISTERA grant of the Call XAI 2019 of the ANR with the grant number Project-ANR-21-CHR4-0002.

References

Begoña Altuna, Rodrigo Agerri, Lidia Salas-Espejo, José Javier Saiz, Alberto Lavelli, Bernardo Magnini, Manuela Speranza, Roberto Zanoli, and Goutham Karunakaran. 2023a. Overview of TESTLINK at IberLEF 2023: Linking results to clinical laboratory tests and measurements. *Procesamiento del Lenguaje Natural*, 71:313–320.

Begoña Altuna, Goutham Karunakaran, Alberto Lavelli, Bernardo Magnini, Manuela Speranza, and Roberto Zanoli. 2023b. CLinkaRT at EVALITA 2023: Overview of the task on linking a lab result to its test event in the clinical domain. In *Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR. org, Parma, Italy*.

Alan R Aronson and François-Michel Lang. 2010. An overview of MetaMap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229–236.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Y. Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *CoRR*, abs/2210.11416.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. 2014. Ncbi disease corpus: a resource for disease name recognition and concept normalization. *Journal of biomedical informatics*, 47:1–10.

Iker García-Ferrero, Rodrigo Agerri, Aitziber Atutxa Salazar, Elena Cabrio, Iker de la Iglesia, Alberto Lavelli, Bernardo Magnini, Benjamin Molinet, Johana Ramirez-Romero, German Rigau, Jose Maria Villa-Gonzalez, Serena Villata, and Andrea Zaninello. 2024. MedMT5: An open-source multilingual text-to-text LLM for the medical domain. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*.

Itziar Gonzalez-Dios, Iker Gutiérrez-Fandiño, Oscar m. Cumbicus-Pineda, and Aitor Soroa. 2022. IrekiaLFes: a new open benchmark and baseline systems for Spanish automatic text simplification. In *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339.

Claudiu D Hromei, Danilo Croce, Valerio Basile, and Roberto Basili. 2023. ExtremITA at EVALITA 2023: Multi-task sustainable scaling to large language models at its extreme. In *Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR. org, Parma, Italy*.

Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. 2020. UNIFIEDQA: Crossing format boundaries with a single QA system. In *Findings of the Association for Computational Linguistics: EMNLP 2020*.

John D Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pages 282–289.

Anastasios Lamproudis, Aron Henriksson, and Hercules Dalianis. 2022. Evaluating pretraining strategies for clinical BERT models. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 410–416.

Robert Leaman and Graciela Gonzalez. 2008. BANNER: an executable survey of advances in biomedical named entity recognition. In *Biocomputing 2008*, pages 652–663. World Scientific.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

Bernardo Magnini, Begoña Altuna, Alberto Lavelli, Anne-Lyse Minard, Manuela Speranza, and Roberto Zanoli. 2022. European clinical case corpus. In *European Language Grid: A Language Technology Platform for Multilingual Europe*, pages 283–288. Springer.

Bernardo Magnini, Begoña Altuna, Alberto Lavelli, Manuela Speranza, and Roberto Zanoli. 2021. The E3C project: European clinical case corpus. In *Proceedings of the Annual Conference of the Spanish Association for Natural Language Processing: Projects and Demonstrations (SEPLN-PD 2021) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2021), Málaga, Spain, September, 2021*, volume 2968 of *CEUR Workshop Proceedings*, pages 17–20. CEUR-WS.org.

Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. 2018. The natural language decathlon: Multitask learning as question answering. *Preprint*, arXiv:1806.08730.

George Michalopoulos, Yuanxin Wang, Hussam Kaka, Helen Chen, and Alexander Wong. 2021. UmlsBERT: Clinical domain knowledge augmentation of contextual embeddings using the Unified Medical Language System Metathesaurus. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Marius Micluta-Campeanu and Liviu P Dinu. 2023a. Simple ideas at CLinkaRT: LeaNER and MeaNER relation extraction. In *Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR. org, Parma, Italy*.

Marius Micluta-Campeanu and Liviu Petrișor Dinu. 2023b. Simple ideas@ TESTLINK: Relying on finer models. *Procesamiento del Lenguaje Natural*.

Mike Mintz, Steven Bills, Rion Snow, and Dan 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.

Carlos Muñoz-Castro, Andrés Carvallo, Matías Rojas, Claudio Aracena, Rodrigo Guerra, Benjamín Pizarro, and Jocelyn Dunstan. 2023. LinkMed: Entity recognition and relation extraction from clinical notes in Spanish. *Procesamiento del Lenguaje Natural*.

Ajay Patel, Bryan Li, Mohammad Sadegh Rasooli, Noah Constant, Colin Raffel, and Chris Callison-Burch. 2022. Bidirectional language models are also few-shot learners. In *The Eleventh International Conference on Learning Representations*.

Georgios Petasis, Frantz Vichot, Francis Wolinski, Georgios Paliouras, Vangelis Karkaletsis, and Constantine D Spyropoulos. 2001. Using machine learning to maintain rule-based named-entity recognition and classification systems. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, pages 426–433.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Teemu Ruokolainen, Pekka Kauppinen, Miikka Silfverberg, and Krister Lindén. 2020. A Finnish news corpus for named entity recognition. *Language Resources and Evaluation*, 54:247–272.

Guergana K Savova, James J Masanz, Philip V Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C Kipper-Schuler, and Christopher G Chute. 2010. Mayo clinical text analysis and knowledge extraction system (cTAKES): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513.

Matej Ulčar and Marko Robnik-Šikonja. 2023. Sequence-to-sequence pretraining for a less-resourced Slovenian language. *Frontiers in Artificial Intelligence*, 6:932519.

Jue Wang and Wei Lu. 2020. Two are better than one: Joint entity and relation extraction with table-sequence encoders. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1706–1721.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. 2020. Luke: Deep contextualized entity representations with entity-aware self-attention. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6442–6454.

Roberto Zanoli, Alberto Lavelli, Daniel Verdi do Amarante, and Daniele Toti. 2023. Assessment of the e3c corpus for the recognition of disorders in clinical texts. *Natural Language Engineering*, pages 1–19.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, and Guoyin Wang. 2023. Instruction tuning for large language models: A survey. *Preprint*, arXiv:2308.10792.

Yuying Zhu and Guoxin Wang. 2019. CAN-NER: Convolutional Attention Network for Chinese Named Entity Recognition. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*.