TransBERT: A Framework for Synthetic Translation in Domain-Specific Language Modeling

Anonymous ACL submission

Abstract

The scarcity of non-English language data in specialized domains significantly limits the development of effective Natural Language Processing (NLP) tools. We present TransBERT, a novel framework for pre-training language models using exclusively synthetically translated text, and introduce TransCorpus, a scalable translation toolkit. Focusing on the life sciences domain in French, our approach demonstrates that state-of-the-art performance on various downstream tasks can be achieved solely by leveraging synthetically translated data. We release the TransCorpus toolkit, the TransCorpus-bio-fr corpus (36.4GB of French life sciences text), TransBERT-bio-fr, its associated pre-trained language model-and reproducible code for both pre-training and fine-tuning. Our results highlight the viability of synthetic translation for building high-quality NLP resources in low-resource language/domain pairs.

1 Introduction

Pre-trained Language Models (PLMs) have revolutionized the field of Natural Language Processing (NLP) by leveraging large-scale datasets and powerful neural network architectures to learn rich linguistic representations. These models, such as BERT (Devlin et al., 2019), GPT (Radford et al., 2018), and T5 (Raffel et al., 2019), are pre-trained on vast amounts of text data in an unsupervised manner, enabling them to capture intricate patterns and nuances of human language. PLMs can be fine-tuned for specific tasks, such as text classification, Named Entity Recognition (NER), and Question Answering (QA), by training them on smaller labeled datasets. This transfer learning approach has significantly improved the performance of NLP models across various languages and domains.

Unfortunately, the success of PLMs has not been equally distributed across all languages. While high-resource languages like English, Chinese, and French have seen significant advancements in NLP applications, many low-resource languages still lack the necessary data and resources to develop effective models. This disparity is particularly evident in specialized domains such as life sciences, where the availability of high-quality datasets is crucial for training accurate models. For example, Hindi, which is spoken by over 600M people, has no available PLM for the life sciences domain. Although BioBERT (Lee et al., 2019), the first pre-trained language model for the life sciences, was released in 2019, significant efforts to gather sufficient Domain-Specific (DS) data for training life sciences models in other high-resource languages have begun to emerge in recent years. Since 2023, life sciences models have emerged for German (Bressem et al., 2024), Italian (Buonocore et al., 2023), and French (Labrak et al., 2023; Touchent et al., 2023). Life sciences is only an example of a domain where the lack of data is a significant barrier to the development of NLP tools. Other domains, such as legal, finance, and patent, also face similar challenges.

In this paper, we introduce TransCorpus, an open-source toolkit leveraging the fairseq translation framework (Ott et al., 2019) to generate extensive synthetic DS corpora in up to 100 languages, featuring a production-level API/CLI setup with multi-GPU and multi-processing capabilities, and reliable checkpoint recovery for scalable corpus management. We demonstrate that a Language Model (LM) trained on TransCorpus output can achieve state-of-the-art performance on various downstream tasks which leverage DrBenchmark, a French life sciences benchmark (Labrak et al., 2024b).

Our contributions are threefold with (1) the release of an open-source toolkit for scalable multilingual corpus translation, (2) a French life sciences corpus synthetically translated of 36.4GB along

with a tokenizer, and a PLM released on Hugging Face, and (3) reproducible code for the pre-training and fine-tuning of the French life sciences PLM made publicly available on GitHub.

2 Related Work

The paradigm of training LMs on massive datasets is relatively recent, gaining prominence after the introduction of BERT (Devlin et al., 2018), and as a result, there is still limited research leveraging translation for training such models, especially in low-resource settings.

In Isbister et al. (2021), sentiment analysis in four low-resource Scandinavian languages is explored using three strategies: fine-tuning a native monolingual PLM, translating the data into English and fine-tuning an English PLM, and fine-tuning a multilingual PLM on the native data. Results generally favor the multilingual approach, though fine-tuning an English model on translated data often outperforms using a monolingual low-resource PLM.

For Luxembourgish, Lothritz et al. (2022) tackle data scarcity by partially translating unambiguous words from a related high-resource language, evaluating several models including a Luxembourgish-only BERT, a Luxembourgish-German Bidirectional Encoder Representations from Transformers (BERT), and LuxemBERT, which is trained on mixed corpora. LuxemBERT shows improved performance over mBERT, though not to a statistically significant extent.

In the Basque context, following the introduction of ElhBERTeu (Urbizu et al., 2022), Urbizu et al. (2023) use synthetic translated data from Spanish to enlarge the Basque corpus, finding that while a PLM trained solely on synthetic data is competitive, it does not outperform one trained only on native data; however, supplementing native data with synthetic translations does enhance performance.

Finally, Phan et al. (2023) improve the English-to-Vietnamese Machine Translation (MT) model Mtet by injecting synthetic biomedical parallel text via self-training (He et al., 2019), resulting in a system that outperforms strong baselines and enables the creation of ViPubmed and ViMedNLI datasets. Continued pre-training and fine-tuning on these resources lead to ViPubMedT5, which achieves state-of-the-art results in several biomedical NLP tasks, further demonstrating the potential of synthetic translation data for advancing low-resource language modeling.

3 TransCorpus: A Scalable Translation Framework

In this section, we present TransCorpus, a framework designed to facilitate the translation of large-scale corpora into multiple languages. First, the selection of the MT toolkit along with its model will be presented, then the model size and context length will be discussed, and finally, the proposed translation workflow will be illustrated.

3.1 Machine Translation Framework & Model Selection

To achieve our goal of translating large volumes of text between any two languages, we selected M2M-100 (Fan et al., 2020) in conjunction with fairseq as a versatile tool for implementation. Fairseq offers broad support for multilingual tasks and facilitates rapid deployment with multi-GPU processing capabilities. Facebook AI's M2M-100 enables direct translation between languages without using English as an intermediary, making it perfect for converting text from any one of 100 languages to another. Moreover, fairseq's modular framework allows for easy model swapping and the integration of new or specialized translation models. This flexibility customizes the translation process to meet specific needs, such as domain adaptation or the incorporation of additional languages.

The M2M-100 model is available in three sizes: 418M, 1.2B, and 12B parameters. The smallest model is faster and uses less memory, while the largest requires multiple Graphics Processing Units (GPUs) for deployment. Because translation quality improvements come at a quadratic increase in computational cost, we did not consider the 12B model for our experiments. However, since the 418M and 1.2B models differ substantially in translation quality but not as much in computational cost, the next section will compare these two models, focusing primarily on computation time.

3.2 Model Size & Context Length

The context length relationship with model complexity is quadratic due to the way attention mechanisms operate in transformers. As context length increases, the number of interactions that the model must account for grows quadratically because every token attends to every other token in the sequence. This means that computational resources,

such as time and memory, increase significantly with longer sequences. Overlooking this relationship can lead to inefficient computation times and resource usage, particularly with large datasets or models, resulting in slower processing speeds and potentially prohibitive resource demands. Moreover, MT models such as M2M-100 typically trained on sentence pairs might exhibit unexpected behavior if used otherwise. Conversely, having no context would reduce translation to a word-by-word level, resulting in nonsensical outcomes. The following analysis explores document-based and sentence-based translation methods while considering both models sizes on a sample of 1000 life sciences abstracts.

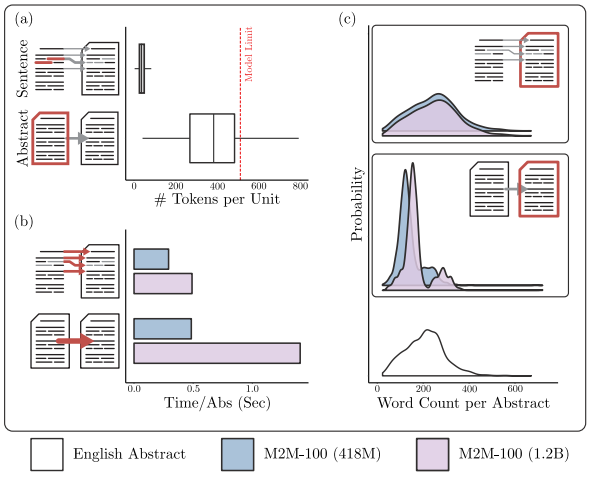

Figure 1: **Translation Method Analysis on a 1000-Abstracts Sample** - (a) Box plot comparing the number of tokens per sentence and abstract, with a red line at 512 tokens representing the maximum token limit that M2M-100 can handle. (b) The average time to translate each abstract using the 418M and 1.2B model versions, comparing sentence-based and document-based translation. (c) Distribution of word count per abstract for both model sizes, displayed with the original English abstract at the bottom when translating by abstract (middle) and by sentence (top). All distributions are normalized to the same scale, so their areas add up to 1.

Figure 1a clearly demonstrates that when translations are performed by sentence, the distribution tends to favor parallelization because larger differences in sequence length require more padding, leading to wasted computation. Figure 1b shows that sentence-based translation consistently results in faster processing for any given model size, with the speed advantage becoming more pronounced as the model size increases. It is important to note that the sentence-based approach will tend to scale linearly with the amount of data to be translated, while the document-based approach will highly depend on the document length distribution of a given domain. Finally, the sentence-based approach seems to mimic the original distribution of words per translated document, while the disparity observed in Figure 1c for the document-based approach was qualitatively reviewed and appears to be partially attributed to a'repetition' problem. Appendix A.3 shows an observed example. As already mentioned, M2M-100 was trained on sentences pairs, which might explain this behavior.

While maintaining translation consistency, the sentence-based strategy is adaptable and scalable to various types of documents. Regarding the size of the model, the difference in the translation time in sentence length is negligible compared to the gains reported in quality. This observation led to the decision to adopt the 1.2B model along with a sentence-based translation approach.

3.3 Framework Translation Workflow

Figure 2 depicts TransCorpus toolkit applied to an English life sciences corpus consisting of 22M abstracts. First, the corpus is divided and distributed among different machines to parallelize the translation process. Each abstract is then divided into sentences with fairseq handling tokenization as shown in Appendix A.4. By grouping sentences of the same length, bucketing is employed to minimize padding, thereby avoiding the computational inefficiency that results from juxtaposing long and short sentences. Although it may seem counterintuitive, there is a considerable increase in speed when translating sentences of the same length simultaneously. Once the sentences are translated, they are matched with their respective abstracts and sentence numbers, and the entire corpus is reconciled by concatenating each output of each subprocess. Appendix A.5 shows an abstract translation example. To avoid too short context issues, sentences that contain fewer than 10 characters are concatenated to the following or preceding sentence.

The TransCorpus toolkit includes a Command Line Interface (CLI) that features four primary commands: (1) `transcorpus download-corpus [domain]` allows users to fetch a corpus from specified domains, (2) `transcorpus preprocess [domain] [target-language] [num-splits]` allows users to optionally preprocess the corpus before translating, and (3) `transcorpus translate [domain] [target-language] [num-splits]` handles both preprocessing (if not previ-

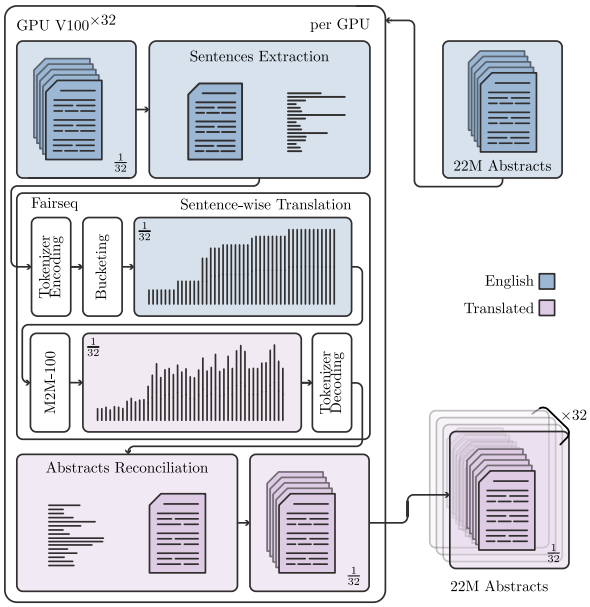

Figure 2: **TransCorpus Translation Workflow** - Illustration of the deployment of TransCorpus on a machine with 32 GPUs.

ously completed) and translating the corpus. Users can perform preprocessing and translation concurrently across numerous GPUs and processes. A checkpoint recovery feature enables users to resume preprocessing or translation from where it last stopped, which is especially beneficial for academics facing GPUs usage time limits. Lastly, (4) `transcorpus preview [domain] [language1] [language2]` provides a side-by-side document preview of the corpus in two languages. Currently, only the life sciences corpus is ready for download, but by updating the `domains.json` file on GitHub with a new corpus URL, additional corpus domains can be quickly integrated into the toolkit. For further details, please check the GitHub repository at *Anonymous*.

4 TransCorpus-bio-fr: A French Life Sciences Corpus

As already highlighted, the limited availability of DS PLMs for certain language/domain pairs presents a notable challenge to the progression of NLP tools. Evaluating our framework in a real-world context is complex for several reasons: (1) it is uncommon to find sufficient DS for low-resource language/domain pairs that also have datasets for model evaluation, and (2) even when such benchmarks exist, an appropriate PLM for comparison might not be available. Fortunately, in the domain of French life sciences, two important papers have recently been published. The first is DrBenchmark (Labrak et al., 2024b) a life sciences benchmark that includes multiple datasets supporting the evaluation of in-domain models. The second is DrBERT (Labrak et al., 2023), a French life sciences PLM. Comparing our model against DrBERT on life sciences tasks allows us to assess the practical effectiveness of our framework. Indeed, as DrBERT is pre-trained from scratch, it does not rely on an advanced general domain PLM such as CamemBERT (CmBERT) (Martin et al., 2020), which is also the case for most languages.

4.1 MEDLINE/PubMed Abstracts Collection

For the building of this life sciences corpus, the 2021 MEDLINE/PubMed Baseline Repository (MBR), encompassing 31M citations, and updates up until April 2021 was downloaded. Then, each citation in the dataset that includes a PMID, a title, and an abstract is kept, subsequently, its raw text is modified by substituting any sequence of one or more whitespace characters with a single space. An example of a title and abstract after modification, as it would appear prior to translation can be found in Appendix A.1.

A considerable amount of citations lacks one of the three essential attributes, i.e. title, abstract, or PMID. Consequently, after filtering the complete dataset, our corpus comprises about 22M abstracts. Despite a few missing unknown values, a comprehensive comparison of our corpus statistics against several models can be found in Appendix A.2. Despite both BioBERT and PubMedBERT (Gu et al., 2020) have a version that also includes PubMed Central (PMC) full-text articles, only those that use PubMed are displayed for a better comparison. This juxtaposition is crucial for understanding the scale of data that similar models have been trained on, which directly impacts their performance and applicability in various NLP tasks.

4.2 Translated Corpus Statistics

The complete translation process was executed using 32 NVIDIA Tesla V100 GPUs with 32Go of memory each, taking roughly 15 days, which translates to approximately 11,520 GPU hours[1]. After translation, the resultant raw text file is 36.4GB, containing 221M sentences and 5.25B words. Table 1 compares TransCorpus with the only two

[1]These figures derive from an early version of TransCorpus; the latest version appears faster with the NVIDIA A100, and these numbers will probably be updated soon

French life sciences corpora leveraged for pre-training. The comparison reveals that DrBERT the State-of-the-Art (SOTA) life sciences LM in French, despite it utilizes the largest corpus until now, is about five times smaller than TransCorpus.

	TransCorpus	DrBERT Corpus	CmBERT Bio Corpus
Size	36.4GB	7.5GB	2.7GB
Sentences	221M	54M	-
Words	5.25B	1.1B	413M

Table 1: **Translated Corpus Statistics Compared to French Life Science Corpora** - '-': Unknown value

Even if the corpus size is important, its quality must also be closely monitored. While MBR is already considered a benchmark of quality in English as it is used for pre-training models such as BioBERT and PubMedBERT, it is crucial to assess the quality of our translations to make sure that everything has been conducted properly. As depicted in Figure 1c, a comparable density check of the entire translated corpus reveals a density profile similar to the original corpus. After manually reviewing a randomly chosen set of abstracts, no irregular translation events, such as repetitions, were detected. A few translated abstracts alongside their counterparts originally written in French can be found in Appendix A.6. The translated corpus is available on Hugging Face at *Anonymous*.

5 TransBERT Pre-Training

The reasons for pre-training a LM from scratch are twofold. Firstly, it enables us to project the usage of our framework on languages that might lack a general domain PLM. Secondly, it allows the use of our custom tokenizer, which typically provides enhanced performance for DS LMs.

5.1 TransTokenizer Training

Subword segmentation algorithms aim to split words optimally using probability. Considering the potential addition of more languages in future works, choosing a tokenizer capable of handling specific linguistic features could prove beneficial. In that context, SentencePiece treats whitespaces as regular characters rather than relying on them, which means that it is suited for all kinds of languages. The original SentencePiece implementation[2] (Kudo and Richardson, 2018) is used to train an Unigram tokenizer with a vocabulary size of 32k and a character coverage set to 0.9995 (default values).

5.2 Pre-training Hyperparameters

A BERT architecture i.e. a Transformer encoder with 12 hidden layers, each with 12 attention heads of dimension 768, is pre-trained on TransCorpus following Robustly optimized BERT approach (RoBERTa) (Liu et al., 2019) with an extensive batch size of 8k, an Adam Optimizer (Kingma and Ba, 2017), along with 24k warm-up steps and a learning rate of 6e-4. The model was updated for 500k steps on a Masked Language Model (MLM) objective function.

6 TransBERT-bio-fr: Application to Life Sciences in French

This section details the pre-training of TransBERT-bio-fr and compares it with other French PLMs.

6.1 TransTokenizer-bio-fr Training

The tokenizer training on TransCorpus-bio-fr took approximately 12 hours on a single machine. As SentencePiece tokenizers require a considerable amount of RAM, a cut-off at 10M translated abstracts were randomly selected in order to train a DS tokenizer based on our synthetic translated corpus. An example showcasing the difference between the tokenization of TransTokenizer (ours) and CmBERT's tokenizer can be found in Appendix A.7.

6.2 CmTransBERT: Tokenizer Ablation

To evaluate the impact of the tokenizer, TransBERT-bio-fr is pre-trained using TransTokenizer-bio-fr while CmTransBERT is combined using CmBERT's tokenizer. Both models are trained on the TransCorpus-bio-fr, with the same hyper-parameters. Prior to fine-tuning our models, the Pseudo-Perplexity (PPPL) (Salazar et al., 2020) per token and word for each model was computed on a 50 authentic French abstracts. This step confirms the success of the pre-training and provides the go-ahead for the experimental phase. For further details, the results are presented in Appendix A.8.

6.3 Pre-training Statistics

Both TransBERT-bio-fr and CmTransBERT were pre-trained for approximately three months using a machine with 3 NVIDIA A100 GPUs, each with 80GB of memory. To ensure a fair pre-training

[2] https://github.com/google/sentencepiece

comparison, TransBERT adopted RoBERTa's training methodology, processing 4B sequences over 500k steps with a batch size of 8k. In contrast, DrBERT was pre-trained on 310M sequences over 78k steps with a batch size of 4k. Although TransCorpus-bio-fr's corpus is about five times larger than DrBERT's, TransBERT-bio-fr's overall training data updates is thirteen times larger compared to DrBERT. CmBERT bio undergoes even less training update, but as it is based on CmBERT the comparison is not relevant. TransCorpus-bio-fr is available on Hugging Face at *Anonymous*.

7 Experimental Setup

To compare TransBERT-bio-fr with other PLMs, we leveraged DrBenchmark which is composed of multiple datasets and tasks. This section describes the experimental setup under which each model was evaluated.

7.1 Baseline Models

To evaluate our method against strong baselines, we selected the state-of-the-art French PLM, CmBERT as well as DrBERT the only PLM pre-trained from scratch in French. As previously noted, CmBERT bio was excluded from comparison because it is derived from CmBERT rather than being trained from scratch, whereas our framework is specifically designed for scenarios where DS data are available in one language, but resources in the target language are lacking.

7.2 DrBenchmark: An Adaptation

Common LM benchmarks in life sciences are predominantly biomedical or clinical, such as Biomedical Language Understanding & Reasoning Benchmark (BLURB) (Gu et al., 2021) and Biomedical Language Understanding Evaluation (BLUE) (Peng et al., 2019) in English. In French, only one option was recently published DrBenchmark. Available in our GitHub, an adaptation of the benchmark containing a few additions such as Hyperparameter Optimization (HPO) implementation instead of fixed hyperparameters setting, a few data cleaning steps to avoid duplicates, datasets merging to avoid unnecessary small datasets and the implementation of a k-fold cross-validation strategy with multiple iterations to allow for a more robust evaluation. Appendix A.9 shows the adapted benchmark datasets statistics, which includes 15 tasks, five of which are classification, six NER, two Part-Of-Speech (POS), and two Semantic Textual Similarity (STS).

7.3 Statistical Testing

Once a metric is computed for each label/class/entity/tag/regression, a statistical test is performed to assess if there is a significant difference between models (1) at the dataset level comparing labels performance across labels and folds and (2) at the task level comparing performances across labels, folds and datasets. For comparisons involving more than two models, the Friedman test is employed, followed by the Nemenyi test. When comparing two models, the Wilcoxon test is used. Appendix A.10 shows the statistical testing process following (Demšar, 2006) recommended practice for comparing metrics rankings to assess model difference for one or multiple datasets.

8 Results & Discussion

	Datasets	CmBERT	DrBERT	TransBERT
CLS	DEFT-2020/T2	**98.91** (1)	97.55 (1)	98.82 (**4**)
	DiaMed	64.70 (22)	68.89 (27)	**75.32***(**55**)
	FrMedMCQA	56.95 (**14**)	56.01 (9)	**57.25** (10)
	MorFITT	73.16 (14)	72.74 (8)	**75.36***(**38**)
	PxCorpus/T2	**96.31** (**11**)	95.34 (8)	95.34 (7)
NER	E3C/Clinical	74.88 (0)	75.44 (1)	**76.83** (**4**)
	E3C/Temporal	85.44 (**12**)	83.92 (2)	**85.73** (**12**)
	MantraGSC	60.56 (12)	57.80 (8)	**62.83** (**16**)
	PxCorpus/T1	92.86 (40)	92.56 (66)	**95.17***(**96**)
	QUAERO/EMEA	84.70 (12)	84.74 (13)	**85.67***(**26**)
	QUAERO/MdL	62.22 (17)	60.71 (5)	**64.06** (**29**)
POS	CAS	97.66 (74)	97.56 (50)	**97.74** (**75**)
	ESSAI	**98.66***(**107**)	98.53 (53)	98.64 (71)
STS	CLISTER	**82.80** (2)	75.44 (0)	82.62 (**3**)
	DEFT-2020/T1	**83.95** (**3**)	71.69 (0)	83.46 (2)

* Significant at 0.05 level (Friedman & Nemenyi test).

Table 2: **Performance Evaluation on the French Life Science Datasets** - Table compares the main metrics for each dataset for Classification, Named Entity Recognition, Part-of-Speech Tagging, and Semantic Textual Similarity tasks. F_1-score is used for each task as the main metric aside STS which uses R^2. In (parentheses) is the count of class/label/entity/tag across all the folds where a model achieved the highest metric. In **bold** is the highest metric/count while underlined text represents the second.

Table 2 presents models performances across all folds for each dataset with the weighted F_1-score for each task except STS, which utilizes the R^2 metric. Among the 15 datasets evaluated, TransBERT-bio-fr (TransBERT) outperforms the other models

in 10 cases, with statistical significance noted on four occasions. CmBERT ranks first in five cases, with one statistically significant result. DrBERT fails to achieve the top metric in any dataset and ranks lowest in 11 datasets. In parentheses are the highest labels metric count across all the folds. For instance, in DiaMed, TransBERT secures the highest F_1-score for 55 labels over five folds, whereas CmBERT and DrBERT attain the highest F_1-score for 22 and 27 labels, respectively.

In classification tasks, even though CmBERT achieves the top performance on two datasets, the differences in metrics and ranking between the models on these datasets are not significant. Conversely, on the DiaMed and MorFITT datasets where TransBERT outperforms, the distinction in metrics and ranking is notable and statistically significant.

In NER, TransBERT leads across all datasets in both metrics and rankings, achieving statistical significance in two instances. In POS tasks, the models demonstrate high and closely matched performances, with the lowest-scoring model achieving a weighted F_1-score of 97.56. Despite this narrow margin, CmBERT secures top results for one dataset, showing statistical significance and attaining the highest F_1-score across 107 tags in all five folds. In STS, CmBERT and TransBERT perform similarly, with minor differences, obtaining three and two top results, respectively. However, DrBERT performs poorly in this task, particularly with a margin exceeding 10 points in DEFT-2020/T1.

8.1 Aggregated Results by Task

Table 3 presents the weighted precision, recall, and F_1-score across each task, except for STS, which utilizes the R^2 metric. TransBERT achieves the best performance for both classification and NER, with statistically significant results at the 0.01 level for every metric. CmBERT secures second place in weighted recall for the NER task, also with statistical significance. The difference between CmBERT and TransBERT in the POS task is minimal; though TransBERT leads in terms of the three metrics, the margin between them is slight. In the STS task, both CmBERT and TransBERT do not show statistical significance, while DrBERT comes last with statistical significance.

With the second more precise classifier, DrBERT ends up having the poorest results in 9 of the 10 metrics. It is worth noting that despite DrBERT is pre-trained on a native French corpus, its sources are quite varied, which could lead to confusion during the pre-training stage for a LM. Specifically, it draws from 24 diverse sources such as disease and condition descriptions, clinical cases, meeting reports, health courses, or even optical character recognition data. Beyond this diversity factor, if a provided sequence is too short for the model to deduce a context helping it identify the kind of document it is receiving, this may cause confusion, potentially resulting in ineffective learning. As already mentioned, even if TransBERT corpus is made of synthetic data, it has already been proved that using MBR worked in English for pre-training of BioBERT and PubMedBERT.

8.2 Tokenizer Ablation Study

Although prior work by Labrak et al. (2024a) explored similar analyses, we identify methodological inconsistencies in their pre-training of 16 PLMs. Specifically, the use of a fixed time-based stopping criterion resulted in unequal training durations across models, potentially biasing outcomes. Furthermore, the justification for employing reduced batch sizes remains unclear, although computational constraints may have been a contributing factor. To address these limitations, we performed systematic replication under controlled experimental conditions. To our knowledge, this study is the first rigorous examination of how tokenizer impacts DS PLMs, offering insights for optimizing architecture decisions in resource-constrained scenarios.

Table 4 presents the comprehensive set of weighted main metrics for both models. The results indicate that TransBERT generally outperforms CmTransBERT in almost all tasks, with statistical significance achieved solely in NER. This implies that NER is more influenced by tokenization compared to other tasks, which seems trivial as NER is basically token-based.

9 Conclusion & Contributions

This work establishes a rigorous framework for assessing LMs on DS for non-English dataset. It builds upon prior research and extends it to a more comprehensive benchmark that includes a more robust way of evaluating the models by applying HPO, multiple training repetition, 5-folds cross-validation, and statistical testing on 15 datasets along with their aggregation by task. It illustrates that employing translated synthetic data for

	CmBERT			DrBERT			TransBERT		
	$\mathbf{P_w}$	$\mathbf{R_w^{(2)}}$	$\mathbf{F_w}$	$\mathbf{P_w}$	$\mathbf{R_w^{(2)}}$	$\mathbf{F_w}$	$\mathbf{P_w}$	$\mathbf{R_w^{(2)}}$	$\mathbf{F_w}$
Classification	74.65	75.54	74.17	74.81	73.42	73.73	**75.82****	**76.69****	**75.71****
Named Entity Recognition	81.23	82.13**	81.55	80.74	81.27**	80.88	**83.03****	**83.46****	**83.15****
Part-Of-Speech	98.31	98.29	98.29	98.20**	98.18	98.18**	**98.33**	**98.30**	**98.31**
Semantic Textual Similarity	-	**83.38**	-	-	73.56**	-	-	83.04	-

** Significant at 0.01 level (Friedman & Nemenyi test)

Table 3: **Performance Evaluation on the French Life Science by Task** - Weighted Precision, Recall, and F_1-scores for each task taking into account each class/label/entity/tag and weighted across all folds and datasets. For Semantic Textual Similarity, the weighted R^2 is reported. In **bold** is highest metric/count while underlined text represents the second.

	TransBERT			CmTransBERT		
	$\mathbf{P_w}$	$\mathbf{R_w^{(2)}}$	$\mathbf{F_w}$	$\mathbf{P_w}$	$\mathbf{R_w^{(2)}}$	$\mathbf{F_w}$
Classification	**75.82**	**76.69**	**75.71**	75.10	76.05	74.70
Named Entity Recognition	**83.03****	**83.46****	**83.15****	81.02**	82.13**	81.44**
Part-Of-Speech	**98.33**	**98.30**	**98.31**	98.31	98.29	98.29
Semantic Textual Similarity	-	83.04	-	-	**84.36**	-

** Significant at 0.01 level (Wilcoxon test)

Table 4: **Ablation study comparing TransBERT and CmTransBERT** - Weighted Precision, Recall, and F_1-scores for each task taking into account each class/label/entity/tag and weighted across all folds and datasets. For STS, the weighted R^2 is reported. In **bold** is the highest metric/count while underlined text represents the second.

training DS LMs is a viable approach to address the lack of native language data. Our proposed model, TransBERT-bio-fr, outperforms existing SOTA models in various life sciences tasks, including classification, NER, POS, and STS.

In addition to offering a viable methodology to address data scarcity, we release to the public TransCorpus, an adaptative toolkit designed to facilitate the translation of large-scale corpora into multiple languages. The resources generated from this work, including TransCorpus-bio-fr, TransBERT-bio-fr and the code for the pre-training and fine-tuning of the models are also made available on GitHub.

10 Future Work

One encouraging direction for future research is to expand our approach to encompass a wider array of languages, especially those that are underrepresented in the life sciences field. Applying our methodology across various linguistic settings will help us better understand its generalizability and any possible constraints. Additionally, creating multilingual models capable of managing several languages within the life sciences sector poses a fascinating challenge. These models might exploit cross-lingual knowledge transfer, allowing for a more efficient use of scarce data resources and promoting a more inclusive global scientific community. Exploring other domains via our toolkit could also yield valuable insights into the adaptability of our approach.

Another path for future research is an extensive comparison between our method and the latest generative Large Language Models (LLMs) on identical datasets. Such a comparison would yield valuable understanding of the trade-offs between specialized, domain-focused models and more general, resource-heavy models LLMs. Assessing performance, efficiency, and cost-effectiveness across different life sciences tasks would help researchers and practitioners in making informed decisions. Furthermore, this analysis could highlight the possibility of integrating the strengths of both approaches.

A promising direction for upcoming research involves exploring the use of generative LLMs to create synthetic data for training DS models, as an alternative to our translation-based method. This approach could yield more varied and nuanced datasets, encapsulating intricate DS knowledge and linguistic patterns. Assessing the quality, reliability, and possible biases of LMs-generated synthetic data in comparison to translated data could offer valuable insights into data augmentation strategies for low-resource domains and languages.

Limitations

10.1 Baseline Model

While TransBERT-bio-fr got better results than CmBERT, it would be interesting to see if DrBERT, the DS baseline PLM would have had better results if it had undergone a proper pre-training process (e.g., 500k steps, 8k batch size, etc.). Also, in order to extend our approach to languages where high qualityPLMs are available, it would have been interesting to compare a pre-training continuation of CmBERT on TransCorpus-bio-fr with CmBERT bio.

10.2 In-Domain/Language Generalization

Even though our benchmark includes a broad range of datasets and tasks, it is impossible to cover every potential application or future development in the field. The performance of our model, while impressive within the scope of our study, may not necessarily be consistent across all possible tasks or datasets in the life sciences domain. Additionally, the idea of a universally 'best' model is inherently flawed in the realm of NLP. Different models might excel in particular contexts or specific types of tasks, and their performance can be affected by factors such as domain specificity, data distribution, and the nuances of individual use cases. What works optimally in one scenario may not be the best choice in another, emphasizing the need for context-specific model evaluation and selection. It is also important to recognize that the fast-paced advancements in NLP research could lead to new architectures, pre-training techniques, or fine-tuning strategies that may surpass our current model in certain aspects. The dynamic nature of the field requires ongoing evaluation and comparison against new innovations.

10.3 Other Domains Generalization

Although our model, which was trained on translated synthetic data within the life sciences corpus, shows encouraging generalization towards other domains, it is important to recognize the constraints when extrapolating these results to other areas. The success of our method in addressing the lack of native language data in life sciences should not be automatically expected to apply to other specialized sectors such as finance, law, or engineering. Each field presents its own unique linguistic hurdles, specialized terminologies, and DS conceptual frameworks that general-purpose MT systems might not handle effectively. The quality and relevance of translated synthetic data can differ greatly between domains, possibly affecting the model's performance. Moreover, the subtleties of DS language use, such as idiomatic phrases, technical lingo, and context-dependent meanings, may not be accurately preserved in translated data, which could lead to misunderstandings or errors in other fields. Additionally, the success of our approach may depend on the degree to which translatable concepts are within a given domain, which can vary greatly. For example, concepts that are highly specific to a culture or legally bound in sectors like law or social sciences might pose particular difficulties for this approach. Hence, even if our results suggest a promising avenue for mitigating language resource shortages in specialized fields, further research is essential to confirm the broad applicability of this method across various domains, each with its own distinct linguistic and conceptual challenges.

10.4 Other Languages Generalization

While our study highlights the effectiveness of employing synthetic translated data for training LMs in the field of life sciences in French, caution is warranted when applying these findings to other languages, especially those with limited resources. We believe that the success of our method is highly dependent on the quality and availability of MT systems for the target language, which can differ greatly among various language pairs. Even if M2M-100 has a great potential to secure relatively great results in low-resource languages compared to other models, some language pairs often lack strong machine translation models, which can undermine the quality of the translated synthetic data. Additionally, the linguistic gap between the source language and the target language can greatly affect the effectiveness of the approach. Languages with different syntactic frameworks, morphological structures, or writing systems might pose additional difficulties in maintaining semantic subtleties and DS language during translation. Furthermore, the cultural and scientific context embedded in the original material might not always have direct counterparts in the target language or culture, which could result in meaning loss or the introduction of biases. Although our findings indicate a potential solution for addressing the deficit of scientific corpora in some languages, the method's suitability across different linguistic contexts requires thorough eval-

uation and additional investigation.

References

Keno K. Bressem, Jens-Michalis Papaioannou, Paul Grundmann, Florian Borchert, Lisa C. Adams, Leonhard Liu, Felix Busch, Lina Xu, Jan P. Loyen, Stefan M. Niehues, Moritz Augustin, Lennart Grosser, Marcus R. Makowski, Hugo J.W.L. Aerts, and Alexander Löser. 2024. medbert.de: A comprehensive german bert model for the medical domain. *Expert Systems with Applications*, 237:121598.

Tommaso Mario Buonocore, Claudio Crema, Alberto Redolfi, Riccardo Bellazzi, and Enea Parimbelli. 2023. Localizing in-domain adaptation of transformer-based biomedical language models. *Journal of Biomedical Informatics*, 144:104431.

Janez Demšar. 2006. Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.*, 7:1–30.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2020. Beyond english-centric multilingual machine translation. *CoRR*, abs/2010.11125.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2020. Domain-specific language model pretraining for biomedical natural language processing. *arXiv preprint arXiv:2007.15779*.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, 3(1):1–23.

Junxian He, Jiatao Gu, Jiajun Shen, and Marc'Aurelio Ranzato. 2019. Revisiting self-training for neural sequence generation. *CoRR*, abs/1909.13788.

Tim Isbister, Fredrik Carlsson, and Magnus Sahlgren. 2021. Should we stop training more monolingual models, and simply use machine translation instead? *Preprint*, arXiv:2104.10441.

Diederik P. Kingma and Jimmy Ba. 2017. Adam: A method for stochastic optimization. *Preprint*, arXiv:1412.6980.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Yanis Labrak, Adrien Bazoge, Béatrice Daille, Mickael Rouvier, and Richard Dufour. 2024a. How important is tokenization in French medical masked language models? In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8223–8234, Torino, Italia. ELRA and ICCL.

Yanis Labrak, Adrien Bazoge, Richard Dufour, Mickael Rouvier, Emmanuel Morin, Béatrice Daille, and Pierre-Antoine Gourraud. 2023. DrBERT: A robust pre-trained model in French for biomedical and clinical domains. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16207–16221, Toronto, Canada. Association for Computational Linguistics.

Yanis Labrak, Adrien Bazoge, Oumaima El Khettari, Mickaël Rouvier, Pacôme Constant Dit Beaufils, Natalia Grabar, Béatrice Daille, Solen Quiniou, Emmanuel Morin, Pierre-antoine Gourraud, and Richard Dufour. 2024b. DrBenchmark: A Large Language Understanding Evaluation Benchmark for French Biomedical Domain. In *Fourteenth Language Resources and Evaluation Conference (LREC-COLING 2024)*, Torino, Italy. Nicoletta Calzolari and Min-Yen Kan.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Cedric Lothritz, Bertrand Lebichot, Kevin Allix, Lisa Veiber, Tegawende Bissyande, Jacques Klein, Andrey Boytsov, Clément Lefebvre, and Anne Goujon. 2022. LuxemBERT: Simple and practical data augmentation in language model pre-training for Luxembourgish. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages

5080–5089, Marseille, France. European Language Resources Association.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online. Association for Computational Linguistics.

P McPhie. 1975. The origin of the alkaline inactivation of pepsinogen. *Biochemistry*, 14(24):5253—5256.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten benchmarking datasets. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 58–65, Florence, Italy. Association for Computational Linguistics.

Long Phan, Tai Dang, Hieu Tran, Trieu H. Trinh, Vy Phan, Lam D. Chau, and Minh-Thang Luong. 2023. Enriching biomedical knowledge for low-resource language through large-scale translation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3131–3142, Dubrovnik, Croatia. Association for Computational Linguistics.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. `https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf`.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *CoRR*, abs/1910.10683.

Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. 2020. Masked language model scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Rian Touchent, Laurent Romary, and Eric de la Clergerie. 2023. Camembert-bio: a tasty french language model better for your health. *Preprint*, arXiv:2306.15550.

Gorka Urbizu, Iñaki San Vicente, Xabier Saralegi, Rodrigo Agerri, and Aitor Soroa. 2022. BasqueGLUE: A natural language understanding benchmark for Basque. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1603–1612, Marseille, France. European Language Resources Association.

Gorka Urbizu, Iñaki San Vicente, Xabier Saralegi, and Ander Corral. 2023. Not enough data to pre-train your language model? MT to the rescue! In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3826–3836, Toronto, Canada. Association for Computational Linguistics.

A Appendix

A.1 Example of an English Abstract

PMID: 44

Title: The origin of the alkaline inactivation of pepsinogen.

Abstract: Above pH 8.5, pepsinogen is converted into a form which cannot be activated to pepsin on exposure to low pH. Intermediate exposure to neutral pH, however, returns the protein to a form which can be activated. Evidence is presented for a reversible, small conformational change in the molecule, distinct from the unfolding of the protein. At the same time, the molecule is converted to a form of limited solubility, which is precipitated at low pH, where activation is normally seen. The results are interpreted in terms of the peculiar structure of the pepsinogen molecule. Titration of the basic NH2-terminal region produced an open form, which can return to the native form at neutral pH, but which is maintained at low pH by neutralization of carboxylate groups in the pepsin portion.

Figure 3: **Example of a Citation From the MBR Database**

A.2 Corpus Statistics

	Corpus Before Translation	BERT	BioBERT	PubMedBERT
Abstracts	22M	N/A	-	14M
Size	30.2GB	16GB	-	21GB
Sentences	202M	-	-	-
Words	4.4B	3.3B	4.5B	3.1B
Tokens	6.7B	-	-	-

-: Unknown value
N/A: Not Applicable

Table 5: **Statistics of English Life Science Corpora Used to Pre-Train Different Models** - Tokens number is computed using a BERT cased tokenizer.

A.3 Example of a Translation with Repetition

Model Size: 418M

Translation Approach: By abstract

Abstract: Des modifications structurelles et fonctionnelles dans les ovaries de l'ovaire de contrôle des ovaries des ovaries de contrôle des ovaries.

Figure 4: **Example of a Translation: 418M, By Abstract (With Repetition)**

A.4 Example of a Tokenized Abstract

PMID: 44

Sentence 1: The origin of the alkaline inactivation of pepsinogen.
['_The', '_origin', '_of', '_the', '_alkal', 'ine', '_in', 'activ', 'ation', '_of', '_pep', 'sin', 'ogen', '.']

Sentence 2: Above pH 8.5, pepsinogen is converted into a form which cannot be activated to pepsin on exposure to low pH.
['_Ab', 'ove', '_pH', '_8.', '5,', '_pep', 'sin', 'ogen', '_is', '_convert', 'ed', '_into', '_a', '_form', '_which', '_cannot', '_be', '_activ', 'ated', '_to', '_pep', 'sin', '_on', '_expos', 'ure', '_to', '_low', '_pH', '.']

Sentence 3: Intermediate exposure to neutral pH, however, returns the protein to a form which can be activated.
['_Inter', 'medi', 'ate', '_expos', 'ure', '_to', '_neutral', '_pH', ',', '_however', ',', '_retur', 'ns', '_the', '_protein', '_to', '_a', '_form', '_which', '_can', '_be', '_activ', 'ated', '.']

Sentence 4: Evidence is presented for a reversible, small conformational change in the molecule, distinct from the unfolding of the protein.
['_Ev', 'idence', '_is', '_present', 'ed', '_for', '_a', '_re', 'vers', 'ible', ',', '_small', '_conform', 'ational', '_change', '_in', '_the', '_mol', 'ec', 'ule', ',', '_distin', 'ct', '_from', '_the', '_un', 'fold', 'ing', '_of', '_the', '_protein', '.']

Sentence 5: At the same time, the molecule is converted to a form of limited solubility, which is precipitated at low pH, where activation is normally seen.
['_At', '_the', '_same', '_time', ',', '_the', '_mol', 'ec', 'ule', '_is', '_convert', 'ed', '_to', '_a', '_form', '_of', '_limited', '_sol', 'ub', 'ility', ',', '_which', '_is', '_precip', 'itat', 'ed', '_at', '_low', '_pH', ',', '_where', '_activ', 'ation', '_is', '_norm', 'ally', '_seen', '.']

Sentence 6: The results are interpreted in terms of the peculiar structure of the pepsinogen molecule.
['_The', '_results', '_are', '_interpret', 'ed', '_in', '_terms', '_of', '_the', '_pec', 'uliar', '_structure', '_of', '_the', '_pep', 'sin', 'ogen', '_mol', 'ec', 'ule', '.']

Sentence 7: Titration of the basic NH2-terminal region produced an open form, which can return to the native form at neutral pH, but which is maintained at low pH by neutralization of carboxylate groups in the pepsin portion.
['_T', 'itr', 'ation', '_of', '_the', '_basic', '_NH', '2-', 'termin', 'al', '_region', '_produc', 'ed', '_an', '_open', '_form', ',', '_which', '_can', '_return', '_to', '_the', '_n', 'ative', '_form', '_at', '_neutral', '_pH', ',', '_but', '_which', '_is', '_mainta', 'ined', '_at', '_low', '_pH', '_by', '_neutr', 'aliz', 'ation', '_of', '_car', 'box', 'yl', 'ate', '_groups', '_in', '_the', '_pep', 'sin', '_por', 'tion', '.']

Figure 5: **Example of Sentence & Word Tokenization**

A.5 Example of a Translated Citation

PMID: 44

Title: L'origine de l'inactivation alcaline du pepsinogène.

Abstract: Au-dessus du pH de 8,5, le pepsinogène est converti en une forme qui ne peut pas être activée en pepsine en cas d'exposition à un pH bas. L'exposition intermédiaire au pH neutre, cependant, renvoie la protéine à une forme qui peut être activée. Des preuves sont présentées pour un changement réversible, de petite conformation dans la molécule, distinct du déploiement de la protéine. Dans le même temps, la molécule est convertie en une forme de solubilité limitée, qui est précipitée à faible pH, où l'activation est normalement observée. Les résultats sont interprétés en termes de la structure particulière de la molécule de pepsinogène. La titration de la région terminale de base NH2 produit une forme ouverte, qui peut revenir à la forme native à pH neutre, mais qui est maintenue à un pH bas par la neutralisation des groupes carboxylés dans la portion de pepsine.

Figure 6: **Example of Title and Abstract Citation From the MBR Database Translated in French (McPhie, 1975)**

A.6 Translation Examples Compared to True French Abstracts

Original (PMID:33739270)
Le foie assure une grande partie du métabolisme des xénobiotiques. Ses particularités en font pourtant une cible privilégiée pour des composés toxiques. Les hépatotoxicités des xénobiotiques, ces molécules étrangères à notre organisme, constituent un vrai défi pour les cliniciens, l'industrie pharmaceutique, et les agences de santé. à la différence des hépatotoxicités intrinsèques, prévisibles et reproductibles, les hépatotoxicités idiosyncrasiques surviennent de manière non prévisible. La physiopathologie des hépatotoxicités idiosyncrasiques à médiation immune reste la moins bien connue. Le développement d'outils qui permettent désormais d'améliorer la prédiction et la compréhension de ces atteintes hépatiques paraît être une approche prometteuse pour identifier des facteurs de risque, et de nouveaux mécanismes de toxicité.

Translated (PMID:33739270)
Le foie assure une grande partie du métabolisme des xénobiotiques grâce à son équipement enzymatique considérable, à sa localisation anatomique et à sa vascularisation abondante. Cependant, ces différentes caractéristiques en font également une cible privilégiée pour les composés toxiques, en particulier dans le cas d'un métabolisme toxique. L'hépatotoxicité induite par les xénobiotiques est une cause majeure de lésions hépatiques et un véritable défi pour les cliniciens, l'industrie pharmaceutique et les agences de santé. Les hépatotoxicités intrinsèques, c'est-à-dire les hépatotoxicités prévisibles et reproductibles qui se produisent à des doses limites, sont distinguées des hépatotoxicités idiosyncratiques, qui se produisent de manière imprévisible chez les personnes présentant des sensibilités individuelles. Parmi eux, la pathophysiologie de l'hépatotoxicité immunomédiée idiosyncratique n'est toujours pas claire. Cependant, le développement d'outils visant à améliorer la prévision et la compréhension de ces troubles peut ouvrir des voies pour l'identification de facteurs de risque et de nouveaux mécanismes de toxicité.

Original (PMID:32334967)
La tuberculose est due au complexe M. tuberculosis, dont la croissance lente entraîne un long délai de rendu des tests phénotypiques utilisés pour le diagnostic bactériologique. La biologie moléculaire a réduit considérablement ce délai, notamment grâce au déploiement de la méthode Xpert® MTB/RIF (Cepheid) qui permet de détecter le complexe M. tuberculosis et la résistance à la rifampicine en 2 heures. D'autres tests détectant en plus la résistance à l'isoniazide et aux antituberculeux de seconde ligne ont été développés. Cependant, les performances de ces tests sont nettement moins bonnes si l'examen microscopique est négatif. Il est donc crucial de restreindre leur indication aux fortes suspicions cliniques. Les tests de détection de la résistance n'explorent que certaines positions caractérisées ; or, toutes les mutations responsables de l'acquisition de résistance ne sont pas connues. De plus, les performances sont variables pour les différents antituberculeux. L'avènement du séquençage génomique est une perspective prometteuse. La faisabilité en routine doit encore être évaluée et l'analyse des données reste à standardiser. L'essor des techniques de biologie moléculaire a révolutionné le diagnostic de la tuberculose et de la résistance. Cependant, elles restent des tests de dépistage dont les résultats doivent être confrontés aux méthodes phénotypiques de référence.

Translated (PMID:32334967)
La tuberculose est causée par le complexe M. tuberculosis. Sa croissance lente retarde le diagnostic bactériologique basé sur des tests phénotypiques. La biologie moléculaire a considérablement réduit ce retard, notamment grâce au déploiement du test Xpert® MTB/RIF (Cepheid), qui détecte le complexe de M. tuberculose et la résistance à la rifampicine en 2 heures. D'autres tests détectant la résistance à l'isoniazide et aux médicaments antituberculeux de deuxième ligne ont été développés. Cependant, les performances des tests moléculaires sont considérablement réduites si le dépistage de la microscopie de bacille acide rapide est négatif. Il est donc crucial de limiter leur indication à de fortes suspicions cliniques. Les tests de détection de la résistance n'explorent que certaines positions caractérisées; cependant, toutes les mutations de résistance aux médicaments ne sont pas connues. En outre, les performances varient pour différents médicaments antituberculeux. L'avènement de la séquençage génomique est prometteur. Son intégration dans le flux de travail de routine doit encore être évaluée et l'analyse des données doit encore être normalisée. La montée des techniques de biologie moléculaire a révolutionné le diagnostic de la tuberculose et de la résistance aux médicaments. Cependant, ils restent des tests de dépistage; les résultats doivent encore être confirmés par des méthodes de référence phénotypiques.

Original (PMID: 33742585)
Dans un souci d'amélioration de la qualité de vie des personnes atteintes de maladie chronique, les pratiques de soins se sont enrichies de l'éducation thérapeutique du patient (ETP). Celle-ci vise l'acquisition de savoirs et de compétences plurielles par les malades pour favoriser une gestion optimale de la pathologie au quotidien et des changements qui en découlent, en limitant les répercussions négatives sur leur autonomie et leur bien-être. Le sujet est placé au cœur de son dispositif, en position de décision et de responsabilité, et collabore activement avec les différents acteurs de soins. L'ETP implique donc la prise en compte de la dimension psychique du patient, en s'appuyant sur la psychologie et des concepts fondamentaux pour sa mise en œuvre.

Translated (PMID: 33742585)
Dans un effort pour améliorer la qualité de vie des personnes atteintes de maladies chroniques, les pratiques de soins ont été enrichis par l'éducation thérapeutique des patients (TPE). Cela vise à l'acquisition de connaissances et de compétences plurielles par les patients, ce qui favorise une gestion optimale de la maladie sur une base quotidienne et des changements qui en découlent, en limitant leurs répercussions négatives sur leur autonomie et leur bien-être. Le sujet est placé au cœur du système, dans une position de décision et de responsabilité, et collabore activement avec les différents acteurs de la santé. Le TPE implique donc la prise en compte de la dimension psychologique du patient, en utilisant la psychologie et les concepts fondamentaux pour sa mise en œuvre.

A.7 Tokenizers Comparison Example

Entity: ['infarctus', 'du', 'myocarde,'] (3 words)
TransTokenizer: ['__infarctus', '__du', '__myocarde', ',''] (4 tokens)
CamemBERT: ['__inf', 'arc', 'tu', 's', '__du', '__my', 'oc', 'arde', ',''] (Δ+5)

Figure 7: **CamemBERT Vs TransTokenizer Sample** - An example of tokenization shows that the tokenizer of TransBERT (i.e., TransTokenizer) requires less tokens than the tokenizer of CamemBERT to encode the same sequence.

A.8 Pseudo-Perplexity Comparison Across Models

	TransBERT	CmTransBERT	CmBERT	DrBERT
PPPL_{token}	6.00	4.14	**174.42**	8.30
PPPL_{word}	11.71	8.59	**2474.88**	17.55
$n_{sentence}$	376			
n_{word}	9204			
n_{token}	12 640	13 934	13 934	12 459

Table 7: **Pseudo-Perplexity Comparison Across Models** - Pseudo-Perplexity across models, with the highest uncertainty highlighted in bold.

A.10 Statistical Testing

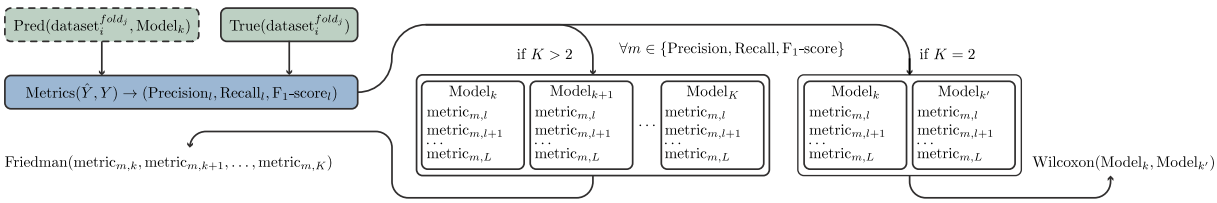

Figure 8: **Statistical Testing** - In order to compare more than two models, the Friedman test is used to determine if there is a significant difference between models, if so, the Nemenyi post-hoc test is used to determine which models are significantly different. For two models, the Wilcoxon test is used.

A.9 Downstream Tasks Summary

Name	Task	Instance	Label	Source
CAS	POS	86 805	30T	CC
CLISTER	STS	1000	0 to 5	CC
DEFT-2020	STS	1009	0 to 5	CC, encyclopedia & drug
	CLS	1100	3C	
DiaMed	CLS	726	15C	CC
E3C/Clinical	NER	3270	1E	CC
E3C/Temporal		5756	5E	
ESSAI	POS	150 269	29T	Clinical Trial Protocols
FrenchMedMCQA	CLS	3102	5C	Pharmacy Exam
MantraGSC	NER	879	7E	Biomedical, Drug & Patent
MorFITT	CLS	5115	12L	Biomedical
PxCorpus	NER	11 465	30E	Drug
	CLS	1727	4C	
QUAERO/EMEA	NER	6001	10E	Drug & Biomedical
QUAERO/Medline		6765		

Table 8: **DrBenchmark Adaptation: Data & Tasks Summary** - By alphabetical order - Overall, every model tested will be evaluated using cross-validation on 15 distinct datasets covering a broad range of tasks. In the Label column, `C` indicates a class within a multi-class framework, while `L` denotes the count of potential labels in a multi-label classification, `T` tag and `E` entity. The instance count reflects the number of positive `C`, `L`, `T` or `E`. In the source column `CC` stands for Clinical Cases.