TransBERT: A Synthetically Translated Language Model

Anonymous ACL submission

Abstract

The lack of non-English language data in specific fields greatly hinders the creation of Natural Language Processing (NLP) tools beneficial to professionals. This paper introduces TransBERT, a novel framework capable of pre-training a Language Model (LM) using solely synthetically translated text. This study focuses on the French language within the life sciences sector to evaluate the effectiveness of this approach. The research includes a comprehensive statistical approach based on an existing Domain-Specific (DS) benchmark. Alongside a vast corpus of 36.4GB of raw text, featuring 22M translated PubMed abstracts, both a Pre-trained Language Model (PLM) and a tokenizer were trained on the synthetically translated corpus. The model effectively addresses the shortage of DS PLMs for non-English languages, resulting in significant improvements that outperform previous State-of-the-Art (SOTA) models with statistical significance across various downstream tasks, potentially setting a new SOTA in multilingual and DS NLP solutions. The modular architecture of the framework further enables the demonstration of the impact of DS tokenizers in tasks such as NER. The results, corpus, code and models are publicly available to encourage further study in this area.

1 Introduction

PLMs have revolutionized the field of 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.

Despite the success of PLMs in English and other high-resource languages, there is a significant lack of DS PLMs for many languages other than English. This scarcity is primarily due to the limited availability of large-scale, high-quality corpora required for training such models. Additionally, the majority of NLP research and development has historically focused on English, leading to a disparity in resources and tools available for other languages. Consequently, many non-English languages, especially those with fewer speakers or less digital presence, do not benefit from the advancements in PLMs technology, hindering the development of robust NLP applications in these languages.

Recent advancements in Machine Translation (MT) have significantly improved both the quality and efficiency of translated text. SOTA models, such as Transformer-based (Vaswani et al., 2017) architectures, have demonstrated remarkable capabilities in capturing context and producing fluent translations. Techniques like back-translation, transfer learning, and multilingual training have further enhanced the performance of translation systems, enabling them to handle a wide range of languages and domains with greater accuracy. Moreover, the integration of large-scale parallel corpora and the use of pre-trained language models have reduced the need for extensive labeled data, making it feasible to generate high-quality translations even for low-resource languages with PLMs such as M2M-100 (Fan et al., 2020) which is able to handle 100 languages. These improvements have paved the way for more effective cross-lingual applications and have facilitated the development of DS language models in various languages.

In this paper, we present several key contributions to the field of NLP: (1) We introduce a novel framework that demonstrates the feasibility of training a PLM using solely synthetically translated data. This method capitalizes on MT advances to create high-quality training sets for low-resource languages and enhances DrBenchmark (Labrak et al., 2024) with important features such as Hyperparameter Optimization (HPO) and

5-fold cross-validations, adding robust statistical testing to a French life science benchmark. (2) We provide a substantial corpus consisting of 36.4GB comprising about 22M translated abstracts from PubMed, which serves as a valuable resource for training and evaluating PLMs in the biomedical domain in French. (3) We released both a PLM and a tokenizer specifically tailored for the French language in the context of Life Sciences. This model addresses the scarcity of DS PLMs for languages other than English and showcases significant improvements in various downstream tasks, thereby advancing the SOTA in DS NLP applications. (4) With our framework's modular design, we highlight the impact of DS tokenizers on tasks such as Named Entity Recognition (NER). (5) We make our results, corpus, code and models publicly available to encourage further research in this area and facilitate the development of NLP tools for low-resource language/domain pairs.

2 Related Work

In (Isbister et al., 2021), sentiment analysis is approached in four low-resource Scandinavian languages using three different methods. The first approach fine-tunes a native monolingual PLM on the original downstream task datasets, the second translates each sequence of the downstream task datasets into English and then fine-tunes an English PLM on the translated data, and finally the third fine-tunes a multilingual PLM directly on the native downstream task datasets. Generally, the results favor the third method, which employs the multilingual model. However, it is worth noting that fine-tuning the English model with translated data generally produces superior results compared to fine-tuning the low-resource language PLM.

The scarcity of data for Luxembourgish, a language with limited resources and a related to German, was tackled by partially translating unambiguous words from a high-resource auxiliary language to train a LM (Lothritz et al., 2022). The research evaluated four different models: mBERT, a Bidirectional Encoder Representations from Transformers (BERT) focused exclusively on Luxembourgish, a BERT combining Luxembourgish and German, and LuxemBERT, a model trained on mixed corpora with partial translations. LuxemBERT suggests superior performance compared to mBERT, however below statistical significance.

After the introduction of ElhBERTeu (Urbizu et al., 2022), a PLM trained on a corpus of 351M words in Basque, a strategy has been implemented using synthetic translated data to improve the corpus size of this low-resource language (Urbizu et al., 2023). Using Spanish as the auxiliary language, a MT Transformer Base model has been trained on 8.6M parallel sentences in order to translate a corpus from Spanish to Basque. Evaluated on BasqueGLUE (Urbizu et al., 2022), results show that the PLM trained solely on synthetic data is competitive, although it does not outperform the model trained only on a native Basque corpus. A final experiment that tweaks the native/translated data ratio suggests that the addition of synthetic data enhances the native PLM performance.

In their paper, (Phan et al., 2023) enhance Mtet (Ngo et al., 2022), the current SOTA MT model in the English-to-Vietnamese direction by injecting synthetic biomedical parallel text into its training corpus using a self-training approach (He et al., 2019). Although the MT model size is not disclosed, the fine-tuned MT system outperforms the models to which it is compared, that is, M2M-100, Google Translate and Mtet, in two translation test sets covering both general and biomedical domains. The resulting translation model is used to generate ViPubmed, a Vietnamese-translated corpus comprising 20M abstracts, as well as ViMedNLI, a benchmark dataset generated by translation of MedNLI (Romanov and Shivade, 2018) and refined with human experts. Subsequently, ViPubmed is used to keep pre-training ViT5 (Phan et al., 2022), the first pre-trained Text-to-Text Transfer Transformer (T5) for the generation of the Vietnamese language, while ViMedNLI is used for fine-tuning. ViPubMedT5, the final model, outperformed models including ViT5 in ViMedNLI and acrDrAid, an acronym disambiguation task, while being close second in a summarization task, showing that using artificially translated data can improve model performance.

3 TransCorpus

3.1 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 (Lee et al., 2019) 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.

3.2 Large Scale Translation Process

To select the MT model among different candidates both a qualitative and quantitative analysis were conducted. Figure 1 shows the quantitative analysis based on a 1000-abstracts sample, comparing (a) the input level by examining the number of tokens per sentence or abstract, (b) the translation time per abstract by both model sizes and translation methods (i.e. abstract or sentence-wise), and lastly, (c) the word distribution after translation for both translation methods and model sizes compared to the original distribution methods.

Quantitatively, Figure 1a clearly demonstrates that when translations are carried out by sentence, the distribution tends to favor parallelization. Figure 1b shows sentence-by-sentence translation consistently results in faster processing for any given model size, with the speed advantage becoming more pronounced as the model size increases. Finally, the distribution disparity observed in Figure 1c was reviewed qualitatively and appears to be partially attributed to a 'repetition' problem. Figure 5 shows an observed example. It is worth noting that M2M-100 was trained on sentence pairs and is probably aimed to be used the way it was trained.

Both quantitative and qualitative analyses led to the choice of sentence-wise translation. Following some extrapolations and using multiple V100 Graphics Processing Units (GPUs), the use of the 1.2B parameters model was deemed feasible, resulting in the complete translation of the entire corpus in approximately 11.52k GPU/hours. Figure 2 depicts the process deployed for translating the whole corpus. First, the 22M abstracts are divided and distributed across different machines to parallelize the translation process. Each abstract is then split into sentences with the Fairseq package handling the tokenization as shown in Appendix A.4. By grouping sentences of the same length, bucketing is employed to minimize padding, thereby avoiding computational inefficiency that results from juxtaposing long and short sentences. Though 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 to their respective abstracts and sentence numbers, and the entire corpus is reconciled. Appendix A.5 shows an abstract translation example.

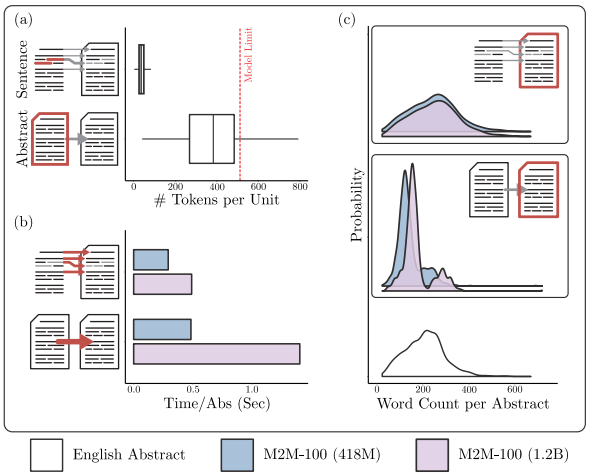

Figure 1: **Abstract Translation Method Analysis on a 1000 Abstracts Sample** - (a) Box plot showing the number of tokens per sentence and abstract, with a red line at 512 tokens representing the maximum token limit that M2M-100 can handle. (b) The average time in seconds to translate each abstract using the 418M and 1.2B model versions, comparing sentence-level and abstract-level 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.

3.3 TransCorpus Comparison with Others

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 French life science corpora leveraged for

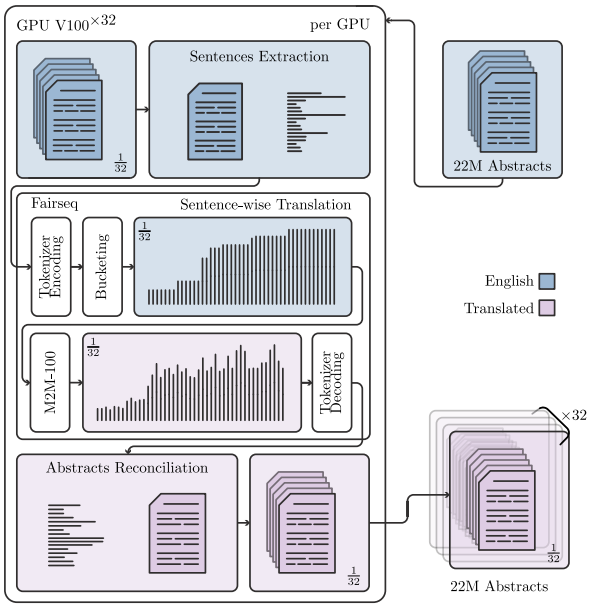

Figure 2: **Large Scale Translation Workflow** - Following the extraction of 22 million abstracts from JSON files, the corpus was shuffled to reduce length biases, then divided and allocated across 32×GPUs. Before translating $\frac{1}{32}$ of the corpus, each abstract was broken down into sentences. The Fairseq toolkit encoded each sentence with the model's tokenizer and translated them into batches using bucketing to optimize the process. Once translation was finished, sentences were decoded back into strings and reassembled into abstracts. Finally, all pieces of the translated abstracts were concatenated, completing the translation of the entire corpus.

pre-training. The comparison reveals that DrBERT (Labrak et al., 2023), the SOTA life science 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	∼4GB*
Sentences	221M	54M	-
Words	5.25B	1.1B	413M

Table 1: **Translated Corpus Statistics Compared to French Life Science Corpora** - '-': Unknown value, '*': Number obtained by linear extrapolation because only the size in GB for a given proportion is disclosed in their paper.

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.

4 TransBERT

4.1 TransTokenizer

Evaluations of BERT-like models have been extensive, yet comprehensive studies and consensus on the best tokenizer remain limited. 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. 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. The original SentencePiece implementation[1] (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). An example showcasing the difference between the tokenization of TransTokenizer and CamemBERT (CmBERT)'s (Martin et al., 2020) tokenizer can be found in Appendix A.7.

4.2 Pre-training Hyperparameters

A BERT architecture (Devlin et al., 2018), i.e. a Transformer decoder 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.

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

4.2.1 TransBERT Vs. CmTransBERT

To evaluate the impact of the tokenizer, TransBERT is pre-trained using TransTokenizer while CmTransBERT is combined using CmBERT's tokenizer. Both models are trained on the same corpus, with the same hyperparameters, and evaluated on the same test set. 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.9.

5 Experiments

5.1 Downstream Tasks

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 (Labrak et al., 2024). Available in our GitHub, an adaptation of the benchmark containing a few additions such as 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 to allow for a more robust evaluation. Appendix A.8 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).

5.2 Baseline Models

To evaluate our method with competitive baselines, we chose the top performing models of each kind, a general French model, to see at least how our model compares with a general model and a DS model. Given that the general French models produced similar results in DrBenchmark, we selected the most downloaded one, CmBERT. For the DS model, the highest performing one in DrBenchmark, DrBERT, was picked.

5.2.1 Multiple Training Repetition

After a model has completed a training iteration with HPO on all tasks, it will undergo four additional rounds of retraining using the previously optimized hyperparameter sets on a freshly initialized model. This extra process, as initially introduced in the original DrBenchmark paper, helps to prevent a fortunate initialization from unfairly enhancing a model's performance for a specific dataset or task. Consequently, each model will be trained and evaluated over five folds, five times, totaling 25 runs per task or dataset. This will only serve to modify training randomness and will not enhance statistical power during testing. Hence, the iterations of models must be aggregated at the prediction phase, prior to evaluation. The key concept is that if one model misses a classification decision, for instance, while the other four rounds capture it, the combined predictions will consider these minor errors and adjust them to reflect what a particular PLM would typically predict.

5.3 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. Figure 3 shows the statistical testing process following (Demšar, 2006) recommended practice for comparing metrics rankings to assess model difference for one or multiple datasets.

6 Results & Discussion

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 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.

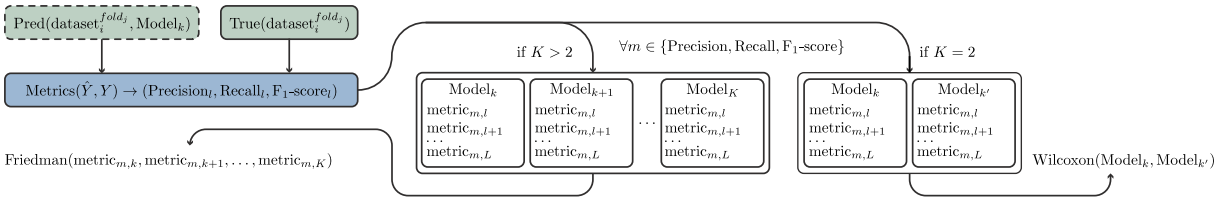

Figure 3: **Statistical Testing** - In order to compare more than two models, the Friedman test is used to determine if there is a significant difference between models, if so, the Nemenyi post-hoc test is used to determine which models are significantly different. For two models, the Wilcoxon test is used.

	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.

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 only 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.

6.1 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 only 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

	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	<u>75.54</u>	<u>74.17</u>	<u>74.81</u>	73.42	73.73	**75.82****	**76.69****	**75.71****
Named Entity Recognition	<u>81.23</u>	<u>82.13</u>**	<u>81.55</u>	80.74	81.27**	80.88	**83.03****	**83.46****	**83.15****
Part-Of-Speech	<u>98.31</u>	<u>98.29</u>	<u>98.29</u>	98.20**	98.18	98.18**	**98.33**	**98.30**	**98.31**
Semantic Textual Similarity	-	**83.38**	-	-	73.56**	-	-	<u>83.04</u>	-

** 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 <u>underlined</u> text represents the second.

of BioBERT and PubMedBERT.

6.2 Tokenizer Ablation Study

One approach to mitigating the impact of tokenizers on downstream applications is to conduct the same experiment twice from scratch. This entails replicating the pre-training process with different tokenizers. To our knowledge, no studies of this nature exist yet, since pre-training two PLMs on the same corpus is quite labor-intensive. Typically, researchers pre-train a model for comparison with others to evaluate the overall method's improvement. As mentioned earlier, to achieve this goal, a LM has been pre-trained on the same machine-translated corpus using the CmBERT tokenizer, which was trained on a non-DS corpus.

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 aligns with the fact that NER is basically token-based, involving the classification of tokens into specific categories.

It is essential to highlight the particular configuration of our experiment. Despite integrating a DS tokenizer prior to pre-training with a DS corpus shows improvements, it does not ensure the same enhancement when training on a non-DS corpus, despite evidence suggesting this potential. Indeed, even if no experiment directly supports this hypothesis, it can be deduced from the fact that TransBERT significantly outperforms CmBERT in the NER task while CmTransBERT performs only on the same level as CmBERT for that task. In other words, it implies that tokenization has a significant impact in NER task as TransBERT significantly outperforms CmTransBERT. Therefore, examining a model pre-trained on the CmBERT corpus with the TransTokenizer would likely yield better results in at least DS NER datasets. The question remains if it would be competitive with a model pre-trained on a DS corpus as it could be a compound effect. Although this might seem a bit trivial, the data and computational power required to train a tokenizer are very low, and these findings could allow better pre-training of DS LM by only pre-training a LM on a generic corpus using the DS tokenizer. It is worth noting that CmTransBERT outperforms CmBERT with statistical significance in the classification task which shows that pre-training continuation is also a viable approach to improve an existing model with DS data.

7 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.

This framework illustrates that employing translated synthetic data for training LMs within the life sciences domain is a viable approach to address the lack of native language data. Our proposed model, TransBERT, outperforms existing SOTA models in various life science tasks, including classification, NER, POS, and STS. By making this framework available, it facilitates future research in determining the required data volume or translation quality needed to attain optimal results or break even.

	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.

In fact, thanks to the modular nature of this framework, a minor adjustment in variables enabled the tokenizer ablation study. This study demonstrates that tokenization significantly influences model performance, particularly in NER tasks within the life sciences domain. Although it would be interesting to determine if similar results would occur in other domains or language, our results indicate that utilizing a DS tokenizer can additionally improve the performance of models pre-trained on a DS corpus.

In addition to offering a framework for addressing data scarcity in certain domains, TransCorpus, TransBERT, and TransTokenizer are accessible to the public and can be used by the life sciences community to enhance various NLP applications. By providing a competitive model that is specifically tailored to the life sciences domain, we aim to facilitate research, innovation, and collaboration within the community.

8 Future Work

Although our work has provided important information on the use of translated synthetic data for training LMs within the field of life sciences, it has generated more research questions than definitive answers. This result emphasizes the intricate and dynamic nature of NLP in specialized areas. The issues prompted by our study span several aspects of machine translation, domain adaptation, and the interaction between artificial and natural language data. These emerging research paths underline the necessity for ongoing exploration into the subtleties of cross-lingual and cross-domain knowledge transfer in language models. One focal area is the evaluation of both the quantity and quality of the translated data required to outperform a SOTA model's performance.

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.

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 science 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 domain-specific 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

8.1 In-Domain/Language Generalization

While our benchmark study presents strong evidence for the effectiveness of our proposed model across various datasets, it is important to note the limitations in generalizing these findings. Although our benchmark was meticulously designed to cover a wide array of tasks within the life sciences domain, it cannot comprehensively represent every possible scenario or use case. One major limitation lies in the wide variety of NLP tasks and the continually evolving nature of scientific language. 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.

8.2 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 machine translation 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.

8.3 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 machine translation 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 evaluation and additional investigation.

References

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, 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. 2024. 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.

Chinh Ngo, Trieu H. Trinh, Long Phan, Hieu Tran, Tai Dang, Hieu Nguyen, Minh Nguyen, and Minh-Thang Luong. 2022. Mtet: Multi-domain translation for english and vietnamese. *Preprint*, arXiv:2210.05610.

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.

Long Phan, Hieu Tran, Hieu Nguyen, and Trieu H. Trinh. 2022. Vit5: Pretrained text-to-text transformer for vietnamese language generation. *Preprint*, arXiv:2205.06457.

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.

Alexey Romanov and Chaitanya Shivade. 2018. Lessons from natural language inference in the clinical domain. *Preprint*, arXiv:1808.06752.

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.

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

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 4: **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 5: **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 6: **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 7: **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 8: **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 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 7: **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.

A.9 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 8: **Pseudo-Perplexity Comparison Across Models** - Pseudo-Perplexity across models, with the highest uncertainty highlighted in bold.