

Introduction

CONTEXT

- Significant **performance improvement** across many clinical and biomedical NLP tasks using domain specific **BERT-based** masked language models (MLM)
 - English:** PudMedBERT, BioBERT, ClinicalBERT
 - French:** CamemBERT-bio, DrBERT, ALIBERT
- Problem:** Inconsistent evaluation among articles (selected corpora, datasets split, metrics, etc.)

CONTRIBUTIONS

- DrBenchmark**, an original French NLP evaluation framework for the biomedical domain with a large set of 20 diversified, proven and challenging tasks
- A **quantitative study** using of a wide range of 8 MLM including varied architectures, data sources and training strategies
- A new open biomedical dataset of **clinical cases** manually annotated into **22 ICD-10 categories** (CC BY-SA 4.0 license)

DrBenchmark

- Data** loading and norm. scripts on Hugging Face
- Modular architecture**
 - Recipes for each task
 - Hyperparameters in YAML and BASH format
- Centralized** evaluation metrics
- Automated execution** of thousands of parallel executions
- Logging system** saving predictions, metrics, hyperparameters

Tasks & Metrics

6 tasks with different requirements and objectives:

- Part-of-Speech (POS) tagging
- Multi-class, Multi-label and Intent classification
- Named-Entity Recognition (NER)
- Multiple-Choice Question-Answering (MCQA)
- Semantic Textual Similarity (STS)

Highly varied data sources:

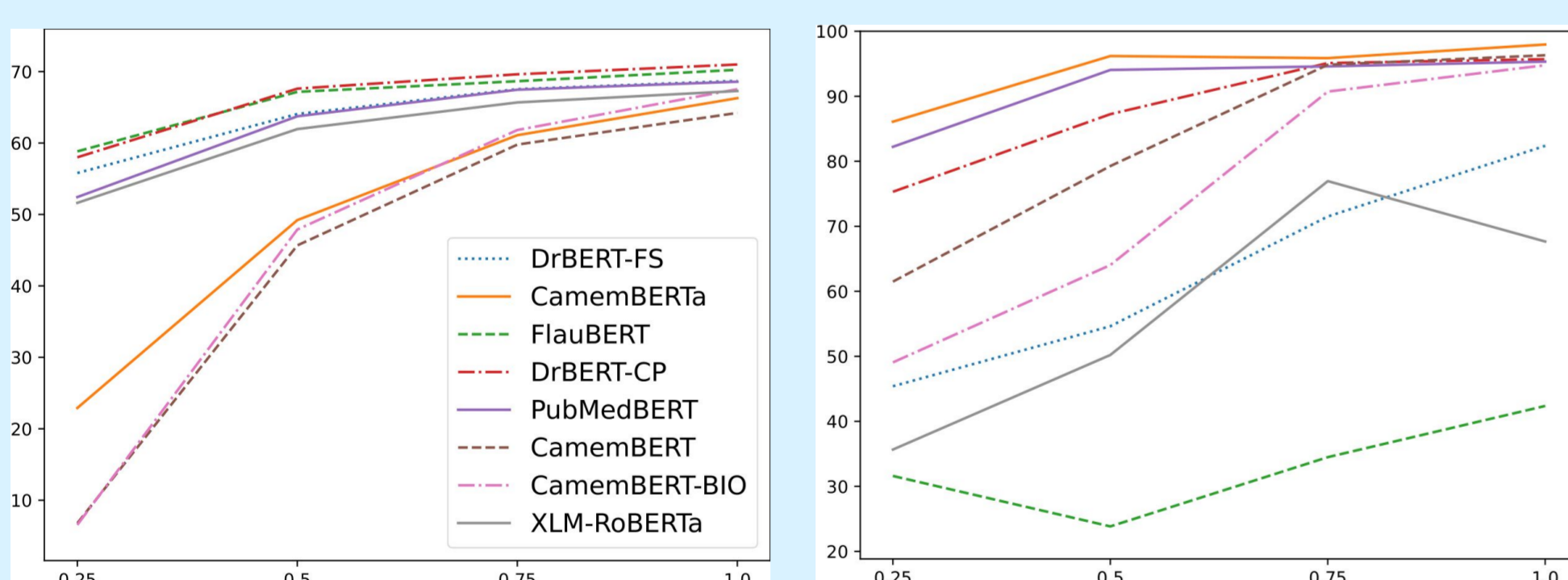
Clinical cases, Clinical trial protocols, Drug leaflets, Biomedical articles, Pharmacy Exam, Patent, Encyclopedia & Drug prescriptions transcripts

Dataset	Task	Metric	Split	License
CAS	POS tagging	SeqEval F1	2,653 / 379 / 758	DUA
ESSAI	POS tagging	SeqEval F1	5,072 / 725 / 1,450	DUA
QUAERO	NER - EMEA NER - MEDLINE	SeqEval F1	429 / 389 / 348 833 / 832 / 833	GFDL 1.3
E3C	NER - Clinical NER - Temporal	SeqEval F1	969 / 140 / 293	CC BY-NC
MorFITT	Multi-label Classification	Weighted F1	1,514 / 1,022 / 1,088	CC BY-SA 4.0
FrenchMedMCQA	Question-Answering Multi-class Classification	Hamming / EMR Weighted F1	2,171 / 312 / 622	Apache 2.0
Mantra-GSC	NER - EMEA NER - Medline NER - Patents	SeqEval F1	70 / 10 / 20 70 / 10 / 20 35 / 5 / 10	CC BY 4.0
CLISTER	Semantic Textual Similarity	EDRM / Spearman	499 / 101 / 400	DUA
DEFT-2020	Semantic Textual Similarity Multi-class Classification	EDRM / Spearman Weighted F1	498 / 102 / 410 460 / 112 / 530	DUA
DEFT-2021	Multi-label Classification NER	Weighted F1 SeqEval F1	118 / 49 / 108 2,153 / 793 / 1,766	DUA
DiaMed	Multi-class Classification NER	Weighted F1 SeqEval F1	509 / 76 / 154	CC BY-SA 4.0
PxCorpus	Multi-class Classification	SeqEval F1 Weighted F1	1,386 / 198 / 397	CC BY 4.0

Table 2: Descriptions and statistics of DrBenchmark.

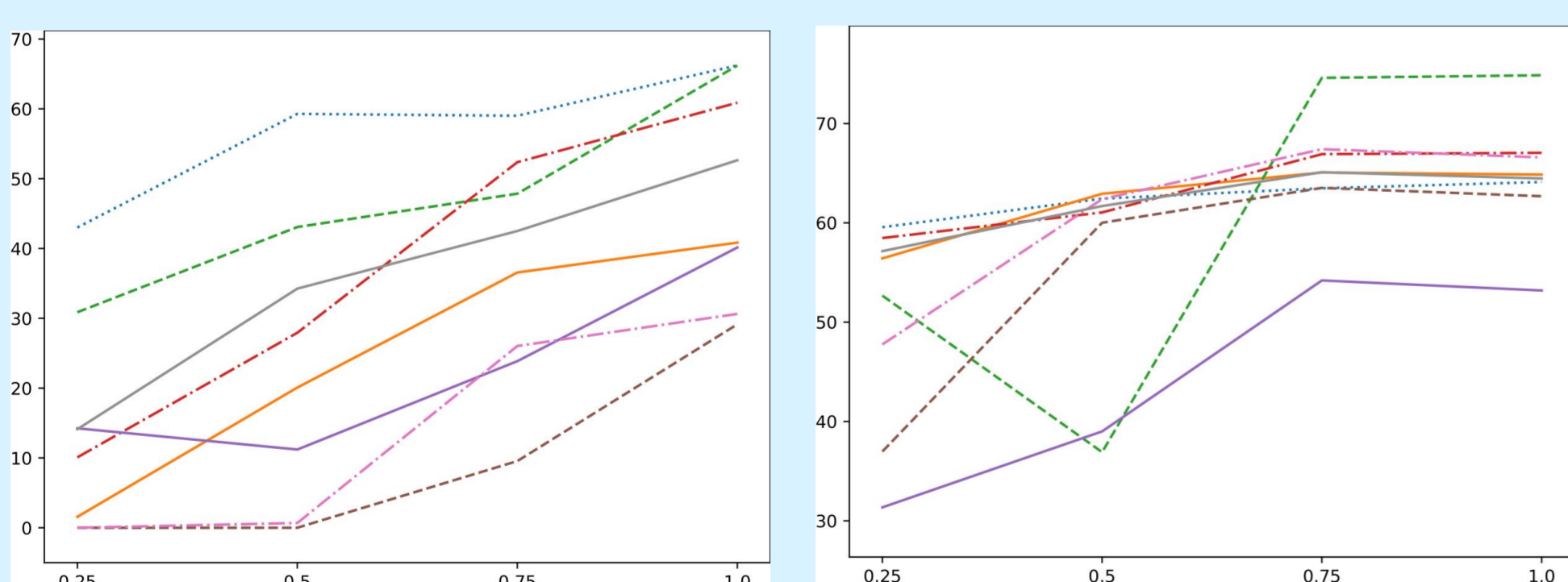
Analysis

- Some **datasets** and **models** capture information **faster than others**, as in Figures A, B and C.
- In NER tasks (Figures C and D), DrBERT-FS achieves the best performance in scenarios with very little data, indicating better model robustness
- For **intermediate corpus portions** (like 50% and 75%): **drop in performance** for certain models, such as FlauBERT, demonstrating **less resilience** to noise that may be present in its portions
- Surprisingly, FlauBERT most frequently obtains poor results despite the fact that it offers the **shortest segmentation**

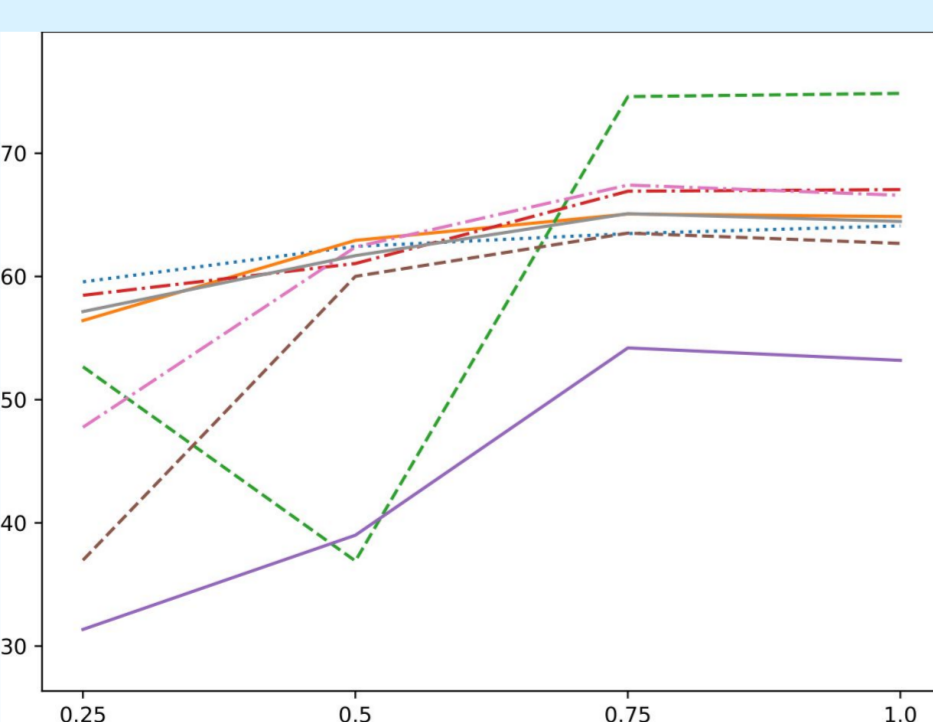


(A) MorFITT CLS

(B) DEFT 2020 CLS



(C) MantraGSC NER EMEA



(D) QUAERO NER EMEA

Table 3: Statistics based on the gold morphological segmentation from 150 biomedical terms.

Results

- None of the models surpasses the others** across all tasks
- The **data sources and quantities strongly impact performance** on certain tasks such as NER, MLC, and STS, where discrepancies can more easily widen
- Cross-lingual continual pretraining** from an English-language biomedical language model appears to be a more effective strategy than starting from the weights of a French-language model
- The English model PubMedBERT achieves very good performance on all tasks except for NER. It achieves, among other things, the best average on semantic similarity tasks
- Masked language models are not suited for MCQA tasks**

Dataset	Task	Baseline	French Generalist			French Biomedical			English Biomedical	Cross-lingual Generalist
			CamemBERT	CamemBERTa	FlauBERT	DrBERT-FS	DrBERT-CP	CamemBERT-bio	PubMedBERT	XLM-RoBERTa
CAS	POS	23.50	95.53	96.56	95.22	96.93	96.46	95.22	94.82	96.91
ESSAI	POS	26.31	97.38	98.08	97.05	98.41	98.01	97.39	97.42	98.34
QUAERO	NER EMEA	8.37	62.68	64.86	74.86	64.11	67.05	66.59	53.19	64.47
	NER MEDLINE	4.92	55.25	55.60	48.98	55.82	60.10	58.94	53.26	51.12
E3C	NER Clinical	4.47	54.70	55.53	47.61	54.45	56.55	56.96	38.34	52.87
	NER Temporal	21.74	83.45	83.22	61.64	81.48	83.43	83.44	80.86	82.6
MorFITT	Multi-Label CLS	3.24	64.21	66.28	70.25	68.70	70.99	67.53	68.58	67.28
FrenchMedMCQA	MCQA	21.83 / 11.57	28.53 / 2.25	29.77 / 2.57	27.88 / 2.09	31.07 / 3.22	32.41 / 2.89	35.3 / 1.45	32.90 / 1.61	34.74 / 2.09
	CLS	8.37	66.21	64.44	61.88	65.38	66.22	65.79	65.41	64.69
MantraGSC	NER FR EMEA	0.00	29.14	40.84	66.20	66.23	60.88	30.63	40.14	52.64
	NER FR Medline	7.78	23.20	22.55	20.69	42.38	35.52	23.66	27.53	18.73
	NER FR Patents	6.20	00.00	44.16	31.47	57.34	39.68	00.00	4.51	8.58
CLISTER	STS	0.44 / 0.00	0.55 / 0.33	0.56 / 0.47	0.50 / 0.29	0.62 / 0.57	0.60 / 0.49	0.54 / 0.26	0.70 / 0.78	0.49 / 0.23
	STS	0.49 / 0.00	0.59 / 0.58	0.59 / 0.43	0.58 / 0.51	0.72 / 0.81	0.73 / 0.86	0.58 / 0.32	0.78 / 0.86	0.60 / 0.26
DEFT-2020	CLS	14.00	96.31	97.96	42.37	82.38	95.71	94.78	95.33	67.66
DEFT-2021	Multi-Label CLS	24.49	18.04	18.04	39.21	34.15	30.04	17.82	25.53	24.46
	NER	0.00	62.76	62.61	33.51	60.44	63.43	64.36	60.27	60.32
DiaMED	CLS	15.36	30.40	24.05	34.08	60.45	54.43	39.57	54.96	26.69
PxCorpus	NER	10.00	92.89	95.05	47.57	95.88	71.38	93.08	94.66	95.80
	CLS	84.78	94.41	93.95	93.45	94.43	94.52	94.49	93.12	93.91

Table 4: Performance of the studied models over 4 runs. Best model in bold and second is underlined. Statistical significance is computed using Student's t-test: * stands for $p < 0.05$, ** stands for $p < 0.01$.

Distribution of tools and corpus

github.com/qanastek/DrBenchmark

huggingface.co/DrBenchmark



Conclusion

- Limitations of generalist models in tackling complex biomedical tasks
- No single model emerges as universally superior
- Certain out-of-domain models or models trained in different languages exhibit superior performance in specific tasks
- Future work:** exploring generative models along with their instruction-tuned versions, to find better solutions for specific tasks.