CamemBERT-bio

Leveraging Continual Pre-training for Cost-Effective Models on French Biomedical Data

Rian Touchent Laurent Romary Eric de la Clergerie

ALMAnaCH - Inria Paris

Motivation

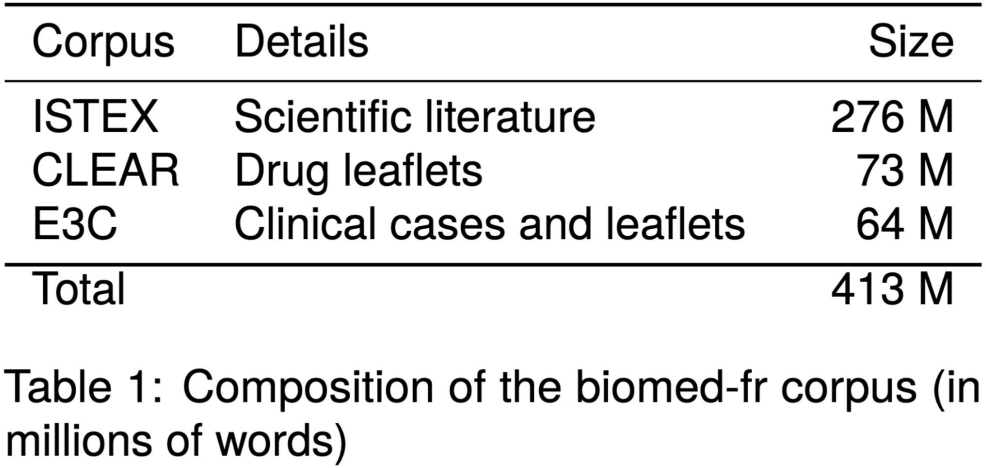
- Hospitals' **clinical data** is accessible but **unstructured**.
- For research, **information extraction** from clinical reports is needed; CamemBERT, while skilled, is **less effective for biomedical data**.
- At the beginning of the work, **no public French biomedical was available**

Contribution

- A **new public French biomedical dataset**.
- A **publicly available**[1] **adaptation of CamemBERT** for the **biomedical** domain, which demonstrates **improved performance on NER**
- **continual-pretraining** from a French model **is proven successful**, necessitating a reevaluation of previous works due to the **impact of evaluation methodology**

[1] hf.co/almanach/camembert-bio-base

A new Corpus: biomed-fr



Corpus	Details	Size
ISTEX	Scientific literature	276 M
CLEAR	Drug leaflets	73 M
E3C	Clinical cases and leaflets	64 M
Total		413 M

Table 1: Composition of the biomed-fr corpus (in millions of words)

Other sources considered for new versions:

- Scientific articles from HAL or PudMed
- Wikipedia
- …

biomed-fr-small :

- a 10% subset of biomed-fr from random documents.

Continual-pretraining

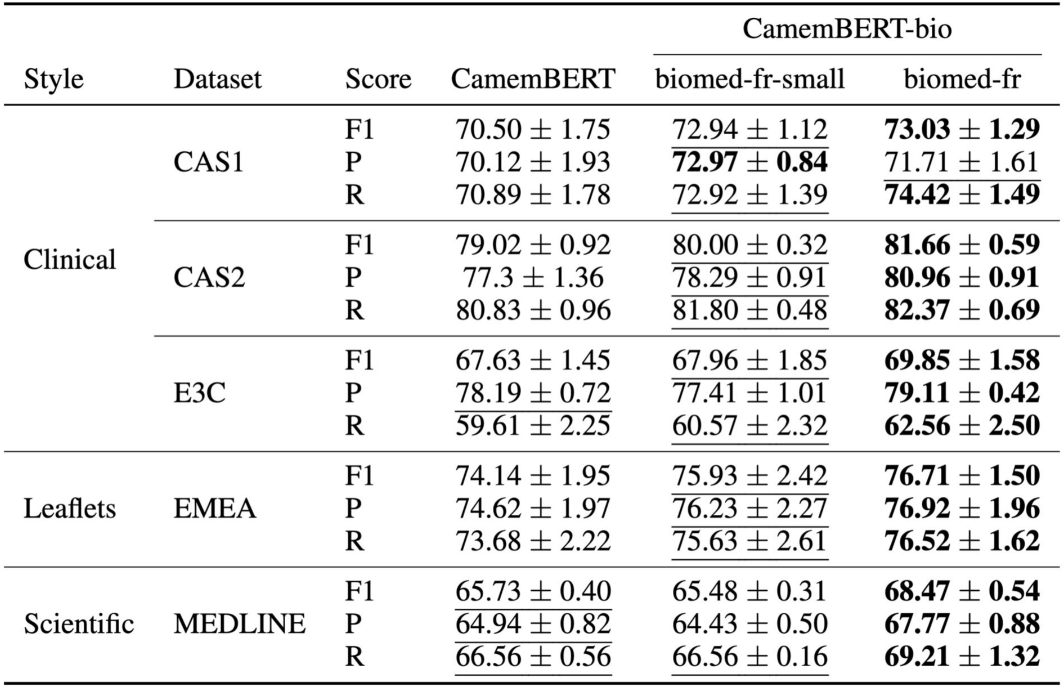
- We followed the methodology of Martin et al. (2020)[1] using the same hyperparameters.

- ***Continual-pretraining*** on biomed-fr **from camembert-base**.

- 50k *steps* during 39 hours on 2 Tesla V100.

[1]: CamemBERT: a Tasty French Language Model (Martin et al., ACL 2020)

Results

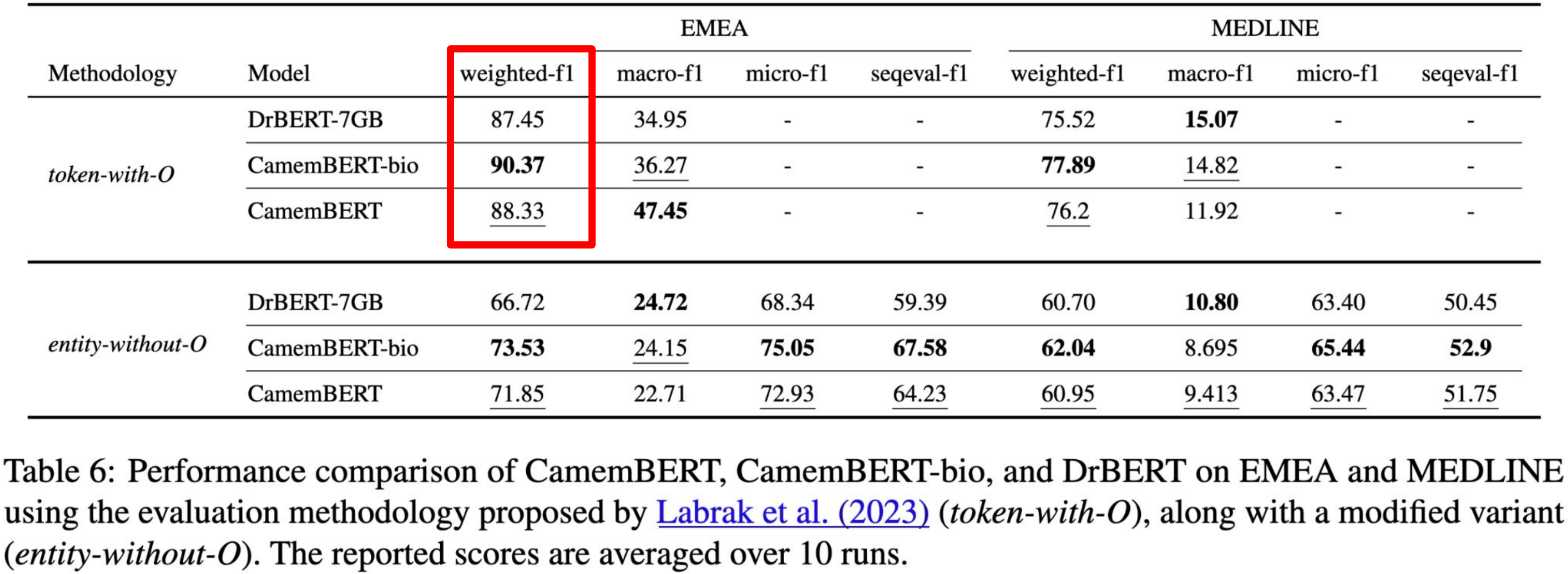


Style	Dataset	Score	CamemBERT	CamemBERT-bio biomed-fr-small	 biomed-fr
Clinical	CAS1	F1	70.50 ± 1.75	72.94 ± 1.12	**73.03 ± 1.29**
		P	70.12 ± 1.93	**72.97 ± 0.84**	71.71 ± 1.61
		R	70.89 ± 1.78	72.92 ± 1.39	**74.42 ± 1.49**
	CAS2	F1	79.02 ± 0.92	80.00 ± 0.32	**81.66 ± 0.59**
		P	77.3 ± 1.36	78.29 ± 0.91	**80.96 ± 0.91**
		R	80.83 ± 0.96	81.80 ± 0.48	**82.37 ± 0.69**
	E3C	F1	67.63 ± 1.45	67.96 ± 1.85	**69.85 ± 1.58**
		P	78.19 ± 0.72	77.41 ± 1.01	**79.11 ± 0.42**
		R	59.61 ± 2.25	60.57 ± 2.32	**62.56 ± 2.50**
Leaflets	EMEA	F1	74.14 ± 1.95	75.93 ± 2.42	**76.71 ± 1.50**
		P	74.62 ± 1.97	76.23 ± 2.27	**76.92 ± 1.96**
		R	73.68 ± 2.22	75.63 ± 2.61	**76.52 ± 1.62**
Scientific	MEDLINE	F1	65.73 ± 0.40	65.48 ± 0.31	**68.47 ± 0.54**
		P	64.94 ± 0.82	64.43 ± 0.50	**67.77 ± 0.88**
		R	66.56 ± 0.56	66.56 ± 0.16	**69.21 ± 1.32**

F-scores on different biomedical named entity recognition tasks

- CamemBERT-bio **improves** upon CamemBERT by **2.54 F-score**
- The improvement is seen **across all biomedical styles** evaluated
- Despite its reduced size, biomed-fr-small still surpasses CamemBERT, emphasizing the **positive impact of corpus size**.

Impact of the evaluation methodology



		EMEA				MEDLINE			
Methodology	Model	weighted-f1	macro-f1	micro-f1	seqeval-f1	weighted-f1	macro-f1	micro-f1	seqeval-f1
token-with-O	DrBERT-7GB	87.45	34.95	-	-	75.52	**15.07**	-	-
	CamemBERT-bio	**90.37**	36.27	-	-	**77.89**	14.82	-	-
	CamemBERT	88.33	**47.45**	-	-	76.2	11.92	-	-
entity-without-O	DrBERT-7GB	66.72	**24.72**	68.34	59.39	60.70	**10.80**	63.40	50.45
	CamemBERT-bio	**73.53**	24.15	**75.05**	**67.58**	**62.04**	8.695	**65.44**	**52.9**
	CamemBERT	71.85	22.71	72.93	64.23	60.95	9.413	63.47	51.75

Table 6: Performance comparison of CamemBERT, CamemBERT-bio, and DrBERT on EMEA and MEDLINE using the evaluation methodology proposed by Labrak et al. (2023) (*token-with-O*), along with a modified variant (*entity-without-O*). The reported scores are averaged over 10 runs.

AliBERT: A Pre-trained Language Model for French Biomedical Text (Berhe et al., BioNLP 2023)
DrBERT: A Robust Pre-trained Model in French for Biomedical and Clinical domains (Labrak et al., ACL 2023)

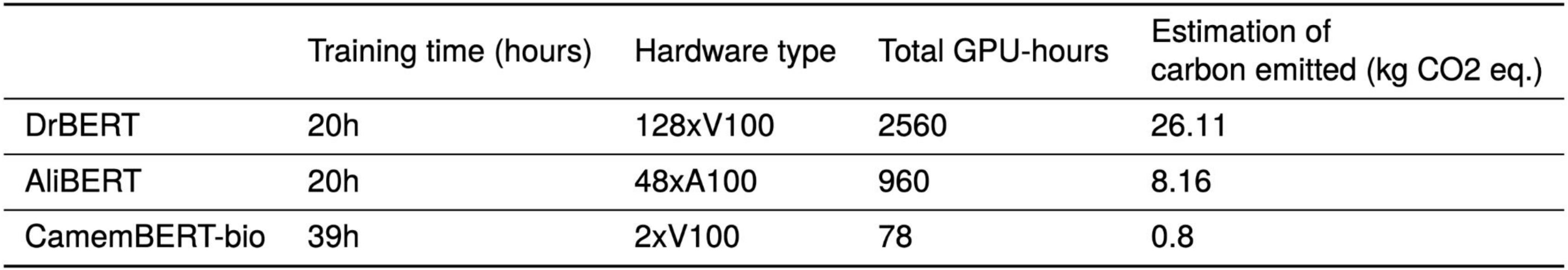
Impact of the evaluation methodology

- On the same evaluation datasets, DrBERT (Labrak et al., 2023) used a token classification metric, while we used an entity classification metric based on seqeval.

- We observe **significant change in performances** between the two methodologies, which underscores the need for a standard unified benchmark to facilitate fair comparison.

AliBERT: A Pre-trained Language Model for French Biomedical Text (Berhe et al., BioNLP 2023)
DrBERT: A Robust Pre-trained Model in French for Biomedical and Clinical domains (Labrak et al., ACL 2023)

Environmental impact



	Training time (hours)	Hardware type	Total GPU-hours	Estimation of carbon emitted (kg CO2 eq.)
DrBERT	20h	128xV100	2560	26.11
AliBERT	20h	48xA100	960	8.16
CamemBERT-bio	39h	2xV100	78	0.8

Carbon emitted estimation based on hardware and training time for different French biomedical models

- **Continual-pretraining requires less energy** consumption, while offering **equal or better performances**. This leads us to advocate for continual-pretraining as the **preferred adaptation method**.
- Other **from-scratch** approaches are estimated **to emit 10 to 32 times more**.

AliBERT: A Pre-trained Language Model for French Biomedical Text (Berhe et al., BioNLP 2023)
DrBERT: A Robust Pre-trained Model in French for Biomedical and Clinical domains (Labrak et al., ACL 2023)

Conclusion

- **We introduce CamemBERT-bio**, a biomedical adaptation of CamemBERT, with a **2.54 F-score point increase** across our NER evaluation datasets.

- Considering the **performances** and the **environmental impact**, we advocate for **continual-pretraining** as the **preferred approach**.