A Benchmark Evaluation of Clinical Named Entity Recognition in French

Nesrine Bannour, Christophe Servan, Aurélie Névéol, Xavier Tannier

A Benchmark Evaluation of Clinical Named Entity Recognition in French

Nesrine Bannour[1], **Christophe Servan**[1], **Aurélie Névéol**[1], **Xavier Tannier**[2]
[1]Université Paris-Saclay – CNRS – LISN,
[2]Sorbonne Université – Inserm – Université Sorbonne Paris Nord – LIMICS
Paris, France
[1]{firstname.lastname}@lisn.upsaclay.fr, [2]{firstname.lastname}@sorbonne-universite.fr

Abstract

Background: Transformer-based language models have shown strong performance on many Natural Language Processing (NLP) tasks. Masked Language Models (MLMs) attract sustained interest because they can be adapted to different languages and sub-domains through training or fine-tuning on specific corpora while remaining lighter than modern Large Language Models (LLMs). Recently, several MLMs have been released for the biomedical domain in French, and experiments suggest that they outperform standard French counterparts. However, no systematic evaluation comparing all models on the same corpora is available. **Objective:** This paper presents an evaluation of masked language models for biomedical French on the task of clinical named entity recognition. **Material and methods:** We evaluate biomedical models `CamemBERT-bio` and `DrBERT` and compare them to standard French models `CamemBERT`, `FlauBERT` and `FrALBERT` as well as multilingual `mBERT` using three publically available corpora for clinical named entity recognition in French. The evaluation set-up relies on gold-standard corpora as released by the corpus developers. **Results:** Results suggest that `CamemBERT-bio` outperforms `DrBERT` consistently while `FlauBERT` offers competitive performance and `FrAlBERT` achieves the lowest carbon footprint. **Conclusion:** This is the first benchmark evaluation of biomedical masked language models for French clinical entity recognition that compares model performance consistently on nested entity recognition using metrics covering performance and environmental impact.

Keywords: named entity recognition, domain adaptation, masked language models, clinical narratives

1. Introduction

The recent development of Clinical Data Warehouses in many French hospitals (Jannot et al., 2017; Madec et al., 2019; Pressat-Laffouilhère et al., 2022) is making unstructured clinical data, including narratives, available for secondary use. As a result, there is a growing need in the biomedical community for Natural Language Processing tools that facilitate the extraction of clinical information from text to support epidemiological studies. Many epidemiological indicators can be modeled as named entities to be extracted from the raw text of clinical reports.

For this reason, the task of Named Entity Recognition, or NER, has attracted a lot of attention in the past decades, in particular through shared tasks aiming at direct comparison of methods. Earlier challenges offered tasks for English (Uzuner et al., 2007, 2011) but more recently, other languages have also been addressed (Marimon et al., 2019; Intxaurrondo et al., 2018), including French (Névéol et al., 2015; Cardon et al., 2020).

The evaluation resources released in the shared tasks continue to be used for evaluating new methods and tools. However, individual efforts often come with adaptations of the data sets or metrics so that comparability is not possible across the board.

Herein, we address this issue by presenting a systematic evaluation that offers comparability across systems as well as with the literature introducing the reference corpora. The main contributions of this paper are:

- A benchmark evaluation of clinical named entity recognition in French based on original gold standard annotations, including nested entities
- A comparison of freely available masked language models for general and biomedical French on the NER task
- A comparison to strong symbolic baselines

2. Corpora

This section briefly presents the clinical French corpora used to train NER models and evaluate the systems considered in this benchmark.

- **DEFT** (Cardon et al., 2020) is a subset of 167 clinical cases from the CAS corpus (Grabar et al., 2018), introduced in the DEFT challenge in 2020[1]. This corpus is annotated with 13 types of clinical entities and five attributes. It

[1]`https://deft.limsi.fr/2020/index-en.html`.

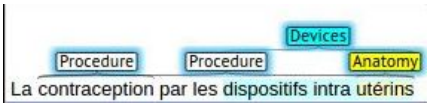

Figure 1: 3 layers of nested entities in an excerpt of the MEDLINE corpus

is divided into a training set of 85 documents, a validation set of 20 documents, and a test set of 62 documents.

- **E3C** (Magnini et al., 2021) is a European corpus of clinical cases. We use the French subcorpus, which comprises 1,615 clinical cases collected in the public domain. It is annotated with 6 types of named clinical entities, including `CLINENTITY`, which we disaggregate into subtypes in order to have an annotation scheme with a diversity approaching that of other corpora. Each entity of this type is associated with a Concept Unique Identifier from the UMLS (Unified Medical Language System) metathesaurus, which can be used to retrieve semantic groups (McCray et al., 2001). In our experiments, we use the gold-standard annotations in the first layer. We use 20% of the training set for validation in the NER models.

- **QUAERO French Med** (Névéol et al., 2014) comprises documents belonging to two text genres, which we treat separately. The EMEA subcorpus is a collection of 13 patient information leaflets supplied by the European Medicines Agency that describes drugs marketed in Europe. The MEDLINE subcorpus consists of 2,500 titles of scientific articles indexed in the MEDLINE database[2]. The entire corpus is annotated with 10 types of clinical entities derived from UMLS semantic groups.

Table 1 presents general descriptive statistics of the study corpora. Table 2 presents descriptive statistics of the distribution of entities in layers in the study corpora. Figure 1 present an excerpt of the MEDLINE corpus containing annotations over three layers: on Layer 1, "contraception par les dispositifs intra utérins" (*contraception with intra uterine devices*) is annotated with the entity type "PROCEDURE", on layer 2 "dispositifs intra utérins" (*intra uterine devices*) is annotated with the entity type "DEVICE" while "contraception" is annotated as a "PROCEDURE" and on layer 3, "utérins" (*uterine*) is annotated with the entity type "ANATOMY".

[2] `http://pubmed.ncbi.nlm.nih.gov/`

	train	dev	test
		DEFT	
Tokens	31,752	5,076	20,360
Entities (all)	7,584	1,432	5,140
Entities (Unique)	5,037	1,230	3,809
		E3C$_{FR}$	
Tokens	19,808	-	4,671
Entities (all)	3,406	-	706
Entities (Unique)	2,197	-	566
		EMEA	
Tokens	14,944	13,271	12,042
Entities (all)	2,695	2,260	2,204
Entities (Unique)	923	756	658
		MEDLINE	
Tokens	10,552	10,503	10,871
Entities (all)	2,994	2,977	3,103
Entities (Unique)	2,296	2,288	2,390

Table 1: Number of tokens and entity annotations in each split of the study corpora.

Layer 1	Layer 2	Layer 3	Layer 4
	DEFT		
65.13%	30.84%	4.02%	0%
	E3C$_{FR}$		
94.47%	5.36%	0.12%	0.06%
	EMEA		
85.73%	13.60%	0.64%	0.03%
	MEDLINE		
74.69%	23.75%	1.50%	0.06%

Table 2: Distribution of entity annotations in each layer of the study corpora.

3. Named Entity Recognition models

We trained named entity recognition (NER) models using the python library `NLstruct` (Wajsbürt, 2021)[3]. `NLstruct` NER models comprise a text encoder, a word tagger, and a bounds matcher. In addition to state-of-the-art performance on NER tasks, `NLstruct` features the ability to address nested entities, which can be found in QUAERO French Med. `NLstruct` also supports annotations in the BRAT standoff format[4], which is used by all three corpora in our study.

The text encoder component in `NLstruct` relies on embeddings produced by a BERT language model, a char-CNN encoder, and static French Fast-

[3] `https://github.com/percevalw/nlstruct`

[4] The Brat Rapid Annotation Tool - BRAT (Stenetorp et al., 2012) produces annotations in the so-called standoff format described in the BRAT online manual `https://brat.nlplab.org/standoff.html`

Text embeddings. To compare French MLMs, we train one NER model using each of the language models in our benchmark. These MLMs are available in the HuggingFace transformers library (Wolf et al., 2020). However, it can be noted that some level of adaptation was needed at the tokenization step to use some of the models within `NLstruct`, especially for `frALBERT` and `FlauBERT` models[5].

General models. We used the major French MLMs freely available for our experiments and a multilingual model.

- CamemBERT (Martin et al., 2020): a language model for French based on the RoBERTa (Liu et al., 2019) model that was pretrained on the OSCAR French corpus (Suárez et al., 2019). We use the `camembert-base` model.
- FlauBERT (Le et al., 2020): a language model for French based on the BERT (Devlin et al., 2019) model that was pretrained on a large multiple-source French corpus. We use the `flaubert-base-uncased` model.
- FrALBERT (Cattan et al., 2021): a compact model based on the ALBERT (Lan et al., 2020) model that was pretrained on 4GB of French Wikipedia part. We use the `fralbert-base` model.
- Multilingual BERT (Devlin et al., 2019): a language model that was pretrained on the 102 languages with the largest Wikipedia, including French. We use the `bert-base-multilingual-uncased` model. We denote Multilingual BERT as mBERT.

Domain-specific models. We used the French MLMs dedicated to the biomedical domain that were freely available for our experiments.

- CamemBERT-bio (Touchent et al., 2023): an adapted CamemBERT model for the biomedical domain that was built using continual-pretraining from `camembert-base` model and trained on a created French biomedical corpus from three sources: The ISTEX database, the CLEAR corpus (Grabar and Cardon, 2018) and the third unannotated layer of the E3C corpus (Magnini et al., 2021). we use the `camembert-bio-base` model.
- DrBERT (Labrak et al., 2023): a RoBERTa-based French biomedical model that was trained from scratch on a web-based medical corpus. We use the `DrBERT-4GB` model.

We also considered `AliBERT` (Berhe et al., 2023) but did not find a publicly accessible version. Similarly, the model described by (Le Clercq de Lannoy et al., 2022) was not available to us.

Baseline. As a baseline, we use a symbolic method that builds a dictionary of entities found in the training and development splits of a corpus and simply matches these entities in the test split. In practice, we use the BRAT propagation tool introduced by (Grouin, 2016) to "propagate" BRAT annotations to the test sets[6].

4. Evaluation metrics

We evaluate the performance of our models at the entity level by measuring the micro Precision, Recall, and F-measure. Confidence intervals at 95% confidence level were computed using the empirical bootstrap method (Dekking et al., 2007, p. 275). Each test corpus is sampled with replacement 1000 times, and evaluation metrics are calculated for each sample. Baseline scores were computed using `brateval` (Verspoor et al., 2013). To measure the carbon footprint of training and testing our models, we use the `Carbon tracker` tool (Anthony et al., 2020). These estimates are approximative and are computed by using an average carbon intensity of 58.48 gCO_2/kWh corresponding to our location in France.

5. Results and discussion

Tables 3 to 6 present the results of our NER experiments on the study corpora.

5.1. NER performance

Overall entity extraction performance. The NER models trained using masked language models outperform the symbolic baseline by at least 10 points of F-measure (up to 40 points for DEFT), although the symbolic baseline can exhibit high precision (e.g., on EMEA, MEDLINE). Interestingly, the knowledge-based approach proposed by Van Mulligen et al. (2016) continues to achieve the best results on the MEDLINE and EMEA corpora, with an F-measure of 0.7 and 0.75 respectively.

The performance of the general French and multilingual models is quite similar, including the multilingual model[7], and, on average, lower compared to the performance of the biomedical models, except for the E3C corpus, where the `FlauBERT model` performs better. The `CamemBERT-bio`

[5]The source code is available here: https://gitlab.lisn.upsaclay.fr/nlp/deep-learning/nlstruct

[6]https://github.com/grouin/propa

[7]Copara et al. (2020) found `mBERT` to be outperformed by `CamemBERT-large`

Models	**DEFT**			
	Precision	*Recall*	*F-measure*	*CO_2 eq (g.)*
CamemBERT	0.73 [0.71-0.75]	0.75 [0.73-0.77]	0.74 [0.73-0.76]	7.7
FlauBERT	0.73 [0.72-0.76]	0.75 [0.73-0.77]	0.74 [0.73-0.76]	8
mBERT	0.72 [0.70-0.74]	0.74 [0.72-0.76]	0.73 [0.71-0.75]	8.3
frALBERT	0.68 [0.66-0.69]	0.67 [0.65-0.69]	0.67 [0.66-0.69]	4.5
CamemBERT-bio	**0.75** [0.73-0.77]	**0.77** [0.75-0.78]	**0.76** [0.74-0.78]	8.7
DrBERT	0.71 [0.68-0.73]	0.72 [0.70-0.74]	0.71 [0.69-0.73]	5.5
Baseline	0.38	0.32	0.35	-
Copara et al. (2020)	-	-	0.73	-

Table 3: Performance of nested entity extraction on the DEFT test set.

Models	**E3C**			
	Precision	*Recall*	*F-measure*	*CO_2 eq (g.)*
CamemBERT	0.52 [0.42-0.63]	0.50 [0.45-0.56]	0.51 [0.46-0.56]	3.6
FlauBERT	**0.54** [0.49-0.60]	**0.53** [0.46-0.60]	**0.54** [0.51-0.57]	4.1
mBERT	0.51 [0.45-0.58]	0.52 [0.47-0.57]	0.52 [0.48-0.54]	4.8
frALBERT	0.50 [0.41-0.59]	0.55 [0.51-0.59]	0.52 [0.47-0.58]	2.5
CamemBERT-bio	0.52 [0.44-0.61]	0.52 [0.45-0.59]	0.52 [0.48-0.55]	3.6
DrBERT	0.47 [0.40-0.57]	0.52 [0.46-0.60]	0.49 [0.46-0.53]	3.6
Baseline	0.24	0.37	0.29	-

Table 4: Performance of nested entity extraction on the E3C test set.

Models	**MEDLINE**			
	Precision	*Recall*	*F-measure*	*CO_2 eq (g.)*
CamemBERT	0.64 [0.62-0.66]	0.66 [0.64-0.67]	0.65 [0.63-0.66]	1.9
FlauBERT	0.67 [0.65-0.68]	0.69 [0.67-0.71]	0.68 [0.66-0.69]	2.2
mBERT	0.63 [0.61-0.65]	0.67 [0.65-0.69]	0.65 [0.63-0.67]	3
frALBERT	0.53 [0.51-0.54]	0.52 [0.49-0.54]	0.52 [0.50-0.54]	1.1
CamemBERT-bio	0.66 [0.65-0.68]	0.70 [0.68-0.72]	0.68 [0.66-0.70]	2.2
DrBERT	0.63 [0.61-0.65]	0.65 [0.63-0.67]	0.64 [0.62-0.66]	2
Baseline	**0.73**	0.30	0.42	-
Van Mulligen et al. (2016)	0.68	**0.72**	**0.70**	-

Table 5: Performance of nested entity extraction on the MEDLINE test set.

Models	**EMEA**			
	Precision	*Recall*	*F-measure*	*CO_2 eq (g.)*
CamemBERT	0.66 [0.62-0.70]	0.65 [0.56-0.73]	0.65 [0.59-0.72]	3.9
FlauBERT	0.69 [0.67-0.72]	0.66 [0.59-0.73]	0.68 [0.63-0.71]	4.4
mBERT	0.67 [0.64-0.72]	0.67 [0.61-0.73]	0.67 [0.63-0.72]	4
frALBERT	0.62 [0.57-0.67]	0.65 [0.61-0.70]	0.63 [0.59-0.68]	2.4
CamemBERT-bio	0.70 [0.66-0.74]	0.68 [0.61-0.75]	0.69 [0.63-0.74]	5.2
DrBERT	0.69 [0.66-0.72]	0.64 [0.58-0.71]	0.66 [0.62-0.71]	3
Baseline	**0.73**	0.43	0.55	-
Van Mulligen et al. (2016)	0.72	**0.79**	**0.75**	-

Table 6: Performance of nested entity extraction on the EMEA test set.

biomedical model seems to perform better than the `DrBERT` biomedical model, suggesting that continual-pretraining from an existing French model on biomedical data might be beneficial in achieving good outcomes.

Nested entity extraction performance. A layer-by-layer evaluation would be difficult to perform, due to the difficulty of aligning system outputs and reference annotations at the layer level. However, we can report that the system outputs contain annotations with depth 3 or 4 depending on the specific models and corpora and exhibit a distribution

of annotations across layers that is similar to that of reference annotations. This suggests that the nesting of annotations by NLStruct is successful. Moreover, F-measure for the DEFT corpus exceed .65, which would be the ceiling score for a system performing flat-entity extraction of layer 1 entities. This also suggests that the nested entity extraction is performed successfully.

Entity extraction performance per entity type. Due to space constraints, we are not providing detailed performance per entity type over the study corpora. Nonetheless, we can notice that the performance of the models tends to vary following similar trends, with highest performance reached for entity types with either high support in terms of training instances and/or high regularity in their occurrence patterns (e.g., temporal entities).

5.2. Comparability of models and experiments

To evaluate their biomedical `CamemBERT-bio` model, Touchent et al. (2023) fine-tuned a NER model on the semi-annotated layer 2 of the E3C corpus and evaluated it on the first layer of this corpus, yielding an F-measure of 69.85. A direct comparison with our results is not possible since we train and evaluate our model using only the first layer containing the gold standard annotations. However, it suggests that silver standard annotations can be useful for training an NER model.

Touchent et al. (2023) and Labrak et al. (2023) evaluated their biomedical MLMs on the two subcorpora of QUAERO French Med and compared them to general French models. However, the task was cast as direct token classification and did not address the nested named entities. Indeed, Touchent et al. (2023) removed nested entities by keeping only the coarse entities, whereas Labrak et al. (2023) concatenated the names of the nested entities to produce new entities and evaluated their results at the token level. These experiments can be seen as intrinsic evaluations of the masked language models.

In contrast, our experiments aim to address the entity recognition task in an extrinsic setting. Our results suggest that the size training data available to train the NER models had more impact on NER performance than the language models used: performance of all approaches is generally lower for E3C vs. other corpora. Similarly, when looking at performance on individual entity types, we generally note that entities with the highest prevalence in the training sets yield higher performance.

5.3. Carbon footprint

Tables 3 to 6 show the carbon footprint of our NER experiments in terms of CO_2 equivalent measure in grams. The highest CO_2 carbon emissions are observed when training and testing the DEFT NER models. This is partly due to the fact that this corpus has more tokens than the other corpora, as illustrated in Table 1. Overall, the `frALBERT`-based models have the lowest carbon footprint. These models offer a decrease of carbon emission between 20% and 63% compared to other models, depending on models and corpora. Note that `Carbon tracker` does not consider the execution environment or energy production. As a result, the obtained measures in our experiments remain approximative. Touchent et al. (2023) reported that the carbon emissions for pre-training their `CamemBERT-bio` model is estimated to 0.84 kg CO_2 eq. Labrak et al. (2023) reported the overall carbon emissions of their 7 `DrBERT`-based models, which is 376.45 kg CO_2 eq.

6. Conclusion

This is the first benchmark evaluation of masked language models for the biomedical domain on the clinical French NER task, using three publicly available clinical French corpora. The evaluation is based on released gold standard annotations, including nested entities. `CamemBERT-bio` consistently outperforms `DrBERT`, while `FlauBERT` offers competitive results. Overall, `frALBERT` offers a fair compromise between *F-measure* and carbon impact, with performance that exceeds the baseline consistently by at least 10 points, and a carbon impact that consistently represents a fraction of the impact of other models. On the QUAERO French Med corpus, MLMs fail to outperform the knowledge-based approach proposed by Van Mulligen et al. (2016). This systematic evaluation compares model performance using metrics covering both performance and environmental impact.

7. Ethical considerations and limitations

Limitations. We did not consider all versions of models: for example, `camembert-large` and `flaubert-base-cased` are not covered in our experiments. While a full evaluation could cover more models, it would incur a higher carbon footprint and we decided to select representative models of the categories that we aimed to cover: general and domain-specific models that had been recently evaluated on similar corpora without direct comparability.

8. Acknowledgements

We thank Dr. Bastien Rance for fruitful discussions on the content of this manuscript. This work was supported by ITMO-Cancer and ANR under grant CODEINE ANR-20-CE23-0026-01.

9. Bibliographical References

Lasse F. Wolff Anthony, Benjamin Kanding, and Raghavendra Selvan. 2020. Carbontracker: Tracking and predicting the carbon footprint of training deep learning models. In *ICML Workshop on "Challenges in Deploying and monitoring Machine Learning Systems"*.

Aman Berhe, Guillaume Draznieks, Vincent Martenot, Valentin Masdeu, Lucas Davy, and Jean-Daniel Zucker. 2023. AliBERT: A pre-trained language model for French biomedical text. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 223–236, Toronto, Canada. Association for Computational Linguistics.

Rémi Cardon, Natalia Grabar, Cyril Grouin, and Thierry Hamon. 2020. Présentation de la campagne d'évaluation DEFT 2020 : similarité textuelle en domaine ouvert et extraction d'information précise dans des cas cliniques (presentation of the DEFT 2020 challenge : open domain textual similarity and precise information extraction from clinical cases). In *Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Atelier DÉfi Fouille de Textes*, pages 1–13, Nancy, France. ATALA et AFCP.

Oralie Cattan, Christophe Servan, and Sophie Rosset. 2021. On the usability of transformers-based models for a French question-answering task. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 244–255, Held Online. INCOMA Ltd.

Jenny Copara, Julien Knafou, Nona Naderi, Claudia Moro, Patrick Ruch, and Douglas Teodoro. 2020. Contextualized French language models for biomedical named entity recognition. In *Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Atelier DÉfi Fouille de Textes*, pages 36–48, Nancy, France. ATALA et AFCP.

F. M. Dekking, C. Kraaikamp, H. P. Lopuhaä, and L. E. Meester. 2007. *A Modern Introduction to Probability and Statistics: Understanding Why and How*. SPRINGER NATURE.

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.

Cyril Grouin. 2016. Controlled propagation of concept annotations in textual corpora. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4075–4079, Portorož, Slovenia. European Language Resources Association (ELRA).

Ander Intxaurrondo, Montserrat Marimon, Aitor Gonzalez-Agirre, Jose Antonio Lopez-Martin, Heidy Rodriguez, Jesus Santamaria, Marta Villegas, and Martin Krallinger. 2018. Finding mentions of abbreviations and their definitions in spanish clinical cases: The barr2 shared task evaluation results. *IberEval@ SEPLN*, 2150:280–289.

Anne-Sophie Jannot, Eric Zapletal, Paul Avillach, Marie-France Mamzer, Anita Burgun, and Patrice Degoulet. 2017. The georges pompidou university hospital clinical data warehouse: a 8-years follow-up experience. *International journal of medical informatics*, 102:21–28.

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.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoit Crabbé, Laurent Besacier, and Didier Schwab. 2020. FlauBERT: Unsupervised language model pre-training for French. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2479–2490, Marseille, France. European Language Resources Association.

Tiphaine Le Clercq de Lannoy, Romaric Besançon, Olivier Ferret, Julien Tourille, Frédérique Brin-Henry, and Bianca Vieru. 2022. Stratégies d'adaptation pour la reconnaissance d'entités médicales en français (adaptation strategies for biomedical named entity recognition in French). In *Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1 : conférence principale*, pages 215–225, Avignon, France. ATALA.

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. *arXiv preprint arXiv:1907.11692*.

Julia Madec, Guillaume Bouzillé, Christine Riou, Pascal Van Hille, Christian Merour, Marie-Lisen Artigny, Denis Delamarre, Veronique Raimbert, Pierre Lemordant, and Marc Cuggia. 2019. ehop clinical data warehouse: From a prototype to the creation of an inter-regional clinical data centers network. *Studies in health technology and informatics*, 264:1536–1537.

Montserrat Marimon, Aitor Gonzalez-Agirre, Ander Intxaurrondo, Heidy Rodriguez, Jose Lopez Martin, Marta Villegas, and Martin Krallinger. 2019. Automatic de-identification of medical texts in spanish: the meddocan track, corpus, guidelines, methods and evaluation of results. In *IberLEF@ SEPLN*, pages 618–638.

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.

Alexa T McCray, Anita Burgun, and Olivier Bodenreider. 2001. Aggregating umls semantic types for reducing conceptual complexity. *Studies in health technology and informatics*, 84(0 1):216.

Aurélie Névéol, Cyril Grouin, Xavier Tannier, Thierry Hamon, Liadh Kelly, Lorraine Goeuriot, and Pierre Zweigenbaum. 2015. Clef ehealth evaluation lab 2015 task 1b: Clinical named entity recognition. In *CLEF (Working Notes)*.

Thibaut Pressat-Laffouilhère, Pierre Balayé, Badisse Dahamna, Romain Lelong, Kévin Billey, Stéfan J Darmoni, and Julien Grosjean. 2022. Evaluation of doc'eds: a french semantic search tool to query health documents from a clinical data warehouse. *BMC medical informatics and decision making*, 22(1):34.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. Brat: A web-based tool for nlp-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, page 102–107, USA. Association for Computational Linguistics.

Rian Touchent, Laurent Romary, and Eric Villemonte de La Clergerie. 2023. Camembert-bio: Un modèle de langue français savoureux et meilleur pour la santé.

Özlem Uzuner, Yuan Luo, and Peter Szolovits. 2007. Evaluating the state-of-the-art in automatic de-identification. *Journal of the American Medical Informatics Association*, 14(5):550–563.

Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2011. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556.

Erik M. Van Mulligen, Zubair Afzal, Saber Akhondi, Dang Vo, and Jan Kors. 2016. Erasmus mc at clef ehealth 2016: Concept recognition and coding in french texts. In *CEUR Workshop Proceedings*, pages 171–178.

Karin Verspoor, Antonio Jimeno Yepes, Lawrence Cavedon, Tara McIntosh, Asha Herten-Crabb, Zoë Thomas, and John-Paul Plazzer. 2013. Annotating the biomedical literature for the human variome. *Database*, 2013:bat019.

Perceval Wajsbürt. 2021. *Extraction and normalization of simple and structured entities in medical documents*. Theses, Sorbonne Université.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of*

the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 38–45, Online. Association for Computational Linguistics.

10. Language Resource References

Natalia Grabar and Rémi Cardon. 2018. CLEAR – simple corpus for medical French. In *Proceedings of the 1st Workshop on Automatic Text Adaptation (ATA)*, pages 3–9, Tilburg, the Netherlands. Association for Computational Linguistics.

Natalia Grabar, Vincent Claveau, and Clément Dalloux. 2018. CAS: French corpus with clinical cases. In *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis*, pages 122–128, Brussels, Belgium. Association for Computational Linguistics.

Bernardo Magnini, Begona Altuna, Alberto Lavelli, Manuela Speranza, and Roberto Zanoli. 2021. The e3c project: European clinical case corpus. *Language*, 1(L2):L3.

Aurélie Névéol, Cyril Grouin, Jeremy Leixa, Sophie Rosset, and Pierre Zweigenbaum. 2014. The quaero french medical corpus: A ressource for medical entity recognition and normalization. *Proc of BioTextMining Work*, pages 24–30.

Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous pipeline for processing huge corpora on medium to low resource infrastructures. In *7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7)*. Leibniz-Institut für Deutsche Sprache.