Digital Health and Informatics Innovations for Sustainable Health Care Systems
J. Mantas et al. (Eds.)

doi:10.3233/SHTI240502

Beyond Tokens: Fair Evaluation of French Large Language Models for Clinical Named Entity Recognition

Jamil ZAGHIR [a,b,1], Mina BJELOGRLIC [a,b], Jean-Philippe GOLDMAN [a,b], Adel BENSAHLA [a,b], Yuanyuan ZHENG [a,b] and Christian LOVIS [a,b]
[a] *Division of Medical Information Sciences, Geneva University Hospitals, Geneva, Switzerland*
[b] *Department of Radiology and Medical Informatics, University of Geneva, Geneva, Switzerland*
ORCiD ID: Jamil Zaghir https://orcid.org/0000-0002-8209-6098

Abstract. Named Entity Recognition (NER) models based on Transformers have gained prominence for their impressive performance in various languages and domains. This work delves into the often-overlooked aspect of entity-level metrics and exposes significant discrepancies between token and entity-level evaluations. The study utilizes a corpus of synthetic French oncological reports annotated with entities representing oncological morphologies. Four different French BERT-based models are fine-tuned for token classification, and their performance is rigorously assessed at both token and entity-level. In addition to fine-tuning, we evaluate ChatGPT's ability to perform NER through prompt engineering techniques. The findings reveal a notable disparity in model effectiveness when transitioning from token to entity-level metrics, highlighting the importance of comprehensive evaluation methodologies in NER tasks. Furthermore, in comparison to BERT, ChatGPT remains limited when it comes to detecting advanced entities in French.

Keywords. Clinical NLP, Named Entity Recognition, Medical Prompt Engineering

1. Introduction

The field of clinical Natural Language Processing (NLP) has witnessed significant advancements, even in languages other than English [1]. These progressions manifest in the form of increased accessibility to datasets and the development of Transformer-based models specifically tailored for biomedical applications. A substantial body of literature highlights the efficacy of fine-tuned BERT models for Named Entity Recognition (NER) in clinical NLP [2–4]. The reported performances of these BERT-based models, often assessed at the token-level, underscore their capacity to achieve state-of-the-art results.

While token-level evaluation is a common practice in assessing prediction-wise effectiveness, it is imperative to use evaluation metrics that are not specific to the tokenizer of the Large Language Models (LLM). To assess the model's performances fairly, there are entity-level metrics such as MUC (Machine Understanding Conference) [5], CoNLL (Computational Natural Language Learning) [6], and ACE (Automatic

[1] Corresponding Author: Jamil Zaghir; E-mail: jamil.zaghir@unige.ch.

Content Extraction) [7]. These metrics offer a generalizable and comparable evaluation of the NER system. This research aims to contribute to the ongoing discourse by exploring and discussing the implications of adopting entity-level metrics for a more robust evaluation of clinical NER systems.

While our focus is on evaluating clinical NER systems, it is essential to highlight ChatGPT's versatility. Through prompt design, it can perform various NLP tasks, including NER, making it a potential choice for low-resource and no-resource scenarios.

2. Methods

In this investigation, we utilized a subset of the FRASIMED corpus [8], specifically CANTEMIST-FR. This French counterpart to CANTEMIST [4] was generated through cross-lingual annotation projection. CANTEMIST itself is a Spanish dataset containing 1'301 synthetic clinical notes annotated with morphological oncology entities. Notably, the Spanish dataset has served as a benchmark for evaluating models in a challenge. Like its Spanish counterpart, CANTEMIST-FR contains 1'301 documents, encompassing a total of 15'978 annotated entities.

We employed four state-of-the-art fine-tuned French BERT models: CamemBERT [9], FlauBERT [10], CamemBERT-Bio [11], and DrBERT [12]. The latter two models are domain-specific, incorporating biomedical datasets during their pre-training phase.

The training data was split into three sets – training, validation, and testing – with a distribution of 80%, 10%, and 10%, respectively. As BERT-based models work with tokens, the IOB (Inside Outside Beginning) format was employed: entities are marked by "B-morphology" for the initial token, followed by "I-morphology" for subsequent tokens within the same entity, and "O" for tokens outside any entity. To fine-tune these models for NER, we chose the best fine-tuning learning rate among 5e-5, 4e-5, 3e-5, and 2e-5, a batch size of 32, and Adam as the optimizer.

Furthermore, with the same test set, we examined the performance of ChatGPT (GPT-3.5) with two prompt strategies: Zero-shot and Few-shot prompting (Figure 1). In both prompts, we provided the official definition provided by the International Classification of Diseases for Oncology. The Few-shot variant reduces the chance of incorrect entity matching by including the preceding token for each predicted entity.

To assess the models thoroughly, we initially employ token-level Precision, Recall, and F1-Score. Subsequently, at the entity level, we employ MUC metrics: "CORRECT" denotes entities accurately identified with matching indices, "PARTIAL" for partial matches, "MISSING" for instances where the system fails to identify expected entities, and "SPURIOUS" for predictions not found in the gold standard. Additionally, we utilized entity-level Precision, Recall, and F1-Score, including a Relaxed variant that treats PARTIALs as true predictions alongside CORRECTs (opposed to Strict).

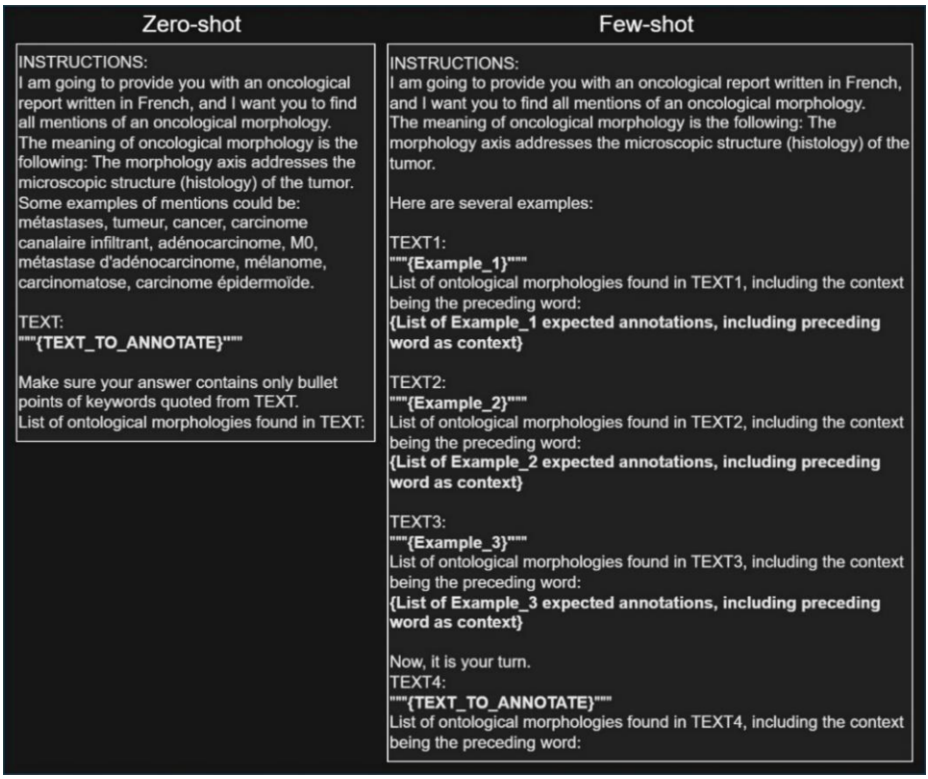

Figure 1. Prompt templates used for Zero-shot and Few-shot predictions on ChatGPT.

3. Results

Table 1 reveals that all four models demonstrate a commendable performance at the token-level, with CamemBERT slightly outperforming others in terms of F1-Score. Despite DrBERT ranking last overall, it excels in Recall. Shifting the focus to Table 2, which evaluates these models from an entity-level perspective using MUC metrics, the surprising finding is that FlauBERT performed the best. Unlike Table 1, DrBERT now ranks third in terms of Recall and exhibits a propensity for predicting PARTIAL more frequently than the other models, thereby impacting its Strict F1 score negatively.

Table 1. Token-level performances of LLMs

Model	Precision	Recall	F1-Score
CamemBERT (Base)	**88.1**	86.1	**87.1**
FlauBERT (Base)	87.8	85.4	86.6
CamemBERT-Bio (Base)	86.4	86.9	86.7
DrBERT (Base)	76.2	**87.3**	81.4

Table 2. Entity-level performances of LLMs using MUC metrics (COR = CORRECT, PAR = PARTIAL, SPU = SPURIOUS, MIS = MISSING, Rel = Relaxed, Str = Strict)

Model	COR	PAR	SPU	MIS	P/R/F1 (Rel)	P/R/F1 (Str)
CamemBERT (Base)	1140	224	**123**	142	**91.7**/90.6/91.1	**76.7**/75.7/76.2
FlauBERT (Base)	**1166**	226	132	**107**	91.3/**92.9**/**92.1**	76.5/**77.8**/**77.1**
CamemBERT-Bio (Base)	1157	221	159	114	89.7/92.3/91.0	75.3/77.6/76.4
DrBERT (Base)	1069	309	136	138	91.0/90.9/91.0	70.6/70.5/70.6
ChatGPT (Zero-Shot)	170	222	171	1093	69.6/26.4/38.2	30.2/11.4/16.6
ChatGPT (Few-Shot)	145	233	143	1104	72.5/25.5/37.8	27.8/9.8/14.5

Through the MUC results, ChatGPT makes relatively fewer predictions in both zero-shot and few-shot scenarios, explaining the poor Recall in Table 2. Despite this, the Relaxed Precision reaching 72.5% in few-shot prompting indicates that, in a no-resource scenario, ChatGPT could be a promising tool for pre-annotating texts to aid the manual annotation process.

4. Discussion

Our results in CANTEMIST-FR, akin to its Spanish counterpart, aligns with existing trends in token-level F1-scores for BERT-based models in clinical NER [4]. Surprisingly, general-domain BERT models exhibit superior performance compared to their biomedical-domain counterparts. However, a pivotal revelation arises in the entity-level evaluation, where FlauBERT unexpectedly emerges as the top performer. This discrepancy underscores the importance of comprehensive assessment methodologies beyond token-level metrics. The notable prevalence of PARTIAL predictions by DrBERT explains the decline in Strict F1-score during entity-level evaluations, a trend that is entirely absent in token-level assessments.

Notably, ChatGPT exhibits nuanced performance, showcasing suboptimal Recall but showing adequate Relaxed Precision during entity-level evaluation. While its entity-level metrics may not match fine-tuned BERT-based models, ChatGPT's no-requirement for training data positions it as a possible tool for preliminary text annotation.

5. Conclusions

In conclusion, our findings underscore the necessity of reporting entity-level metrics in NER evaluations, highlighting performance nuances not discernible at the token-level. Future clinical NER practices should prioritize this entity-level evaluation for a fairer assessment, especially if the objective is to deploy the model in real-world conditions.

Acknowledgments

This research has been funded by “NCCR Evolving Language, Swiss National Science Foundation Agreement #51NF40_180888”.

References

[1] Shaitarova A, Zaghir J, Lavelli A, Krauthammer M, Rinaldi F. Exploring the Latest Highlights in Medical Natural Language Processing across Multiple Languages: A Survey. Yearb Med Inform. août 2023;32(01):230-43.

[2] Agrawal M, Hegselmann S, Lang H, Kim Y, Sontag D. Large language models are few-shot clinical information extractors. In: Proceedings of the 2022 Conference on EMNLP. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics; 2022. p. 1998-2022.

[3] Cardon R, Grabar N, Grouin C, Hamon T. Presentation of the DEFT 2020 Challenge: open domain textual similarity and precise information extraction from clinical cases. In: TALN, 27e édition, RÉCITAL, 22e édition DÉfi Fouille de Textes. 2020. p. 1-13.

[4] Miranda-Escalada A, Farré E, Krallinger M. Named Entity Recognition, Concept Normalization and Clinical Coding: Overview of the Cantemist Track for Cancer Text Mining in Spanish, Corpus, Guidelines, Methods and Results. IberLEF@ SEPLN. 2020;303-23.

[5] Chinchor N, Sundheim B. MUC-5 Evaluation Metrics. In: Fifth Message Understanding Conference (MUC-5): Proceedings of a Conference Held in Baltimore, Maryland, August 25-27. 1993.

[6] Tjong Kim Sang EF, De Meulder F. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In: Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL. 2003. p. 142-7.

[7] Doddington GR, Mitchell A, Przybocki MA, Ramshaw LA, Strassel SM, Weischedel RM. The Automatic Content Extraction (ACE) Program Tasks, Data, and Evaluation. In: LREC. Lisbon; 2004. p. 837-40.

[8] Zaghir J, Bjelogrlic M, Goldman JP, Aananou S, Gaudet-Blavignac C, Lovis C. FRASIMED: a Clinical French Annotated Resource Produced through Crosslingual BERT-Based Annotation Projection. In: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). Torino, Italia; 2024. p. 7450–7460.

[9] Martin L, Muller B, Suárez PJO, Dupont Y, Romary L, de la Clergerie ÉV, Seddah D, Sagot B. CamemBERT: a Tasty French Language Model. Proceedings of the 58th Annual Meeting of the ACL. 2020;7203-19.

[10] Le H, Vial L, Frej J, Segonne V, Coavoux M, Lecouteux B, Allauzen A, Crabbé B, Besacier L, Schwab D. FlauBERT : Unsupervised Language Model Pre-training for French. In: Actes de la 6e conférence conjointe JEP, 33e édition, TALN, 27e édition, RÉCITAL, 22e édition. Nancy, France: ATALA et AFCP; 2020. p. 268-78.

[11] Touchent R, Romary L, De La Clergerie E. CamemBERT-bio : Un modèle de langue français savoureux et meilleur pour la santé. In: CORIA-TALN 2023 Actes de la 30e Conférence TALN, volume 1. Paris, France: ATALA; 2023. p. 323-34.

[12] Labrak Y, Bazoge A, Dufour R, Rouvier M, Morin E, Daille B, Gourraud PA. DrBERT: A robust pre-trained model in french for biomedical and clinical domains. bioRxiv. 2023;2023-04.