Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Contextualized French Language Models for Biomedical Named Entity Recognition

Jenny Copara*[1,2,3] Julien Knafou*[1,2] Nona Naderi[1,2] Claudia Moro[4] Patrick Ruch[1,2] Douglas Teodoro[1,2]

(1) University of Applied Sciences and Arts of Western Switzerland, Rue de la Tambourine 17, 1227, Geneva, Switzerland
(2) Swiss Institute of Bioinformatics, Rue Michel-Servet 1, Geneva, Switzerland
(3) University of Geneva, Rue du Général-Dufour 24, 1211, Geneva, Switzerland
(4) Pontifical Catholic University of Paraná, Rua Imaculada Conceição 1155, 80215-901, Curitiba, Brazil
{jenny.copara,julien.knafou}@hesge.ch,
{nona.naderi,patrick.ruch,douglas.teodoro}@hesge.ch, c.moro@pucpr.br

RÉSUMÉ

Modèles contextualisés en langue française pour la reconnaissance des entités nommées dans le domaine biomédical

La reconnaissance des entités nommées (NER) est essentielle pour les applications biomédicales car elle permet la découverte de connaissances dans des données en texte libre. Comme les entités sont des phrases sémantiques, leur signification est conditionnée par le contexte pour éviter toute ambiguïté. Dans ce travail, nous explorons les modèles de langage contextualisés pour la NER dans les textes biomédicaux français dans le cadre du Défi Fouille de Textes. Notre meilleure approche a obtenu une mesure F1 de 66% pour les symptômes et les signes, et les catégories de pathologie, en étant dans le top 1 pour la sous-tâche 1. Pour les catégories anatomie, dose, examen, mode, moment, substance, traitement et valeur, elle a obtenu une mesure F1 de 75% (sous-tâche 2). Si l'on considère toutes les catégories, notre modèle a obtenu le meilleur résultat dans le cadre de ce défi, avec une mesure F1 de 72%. L'utilisation d'un ensemble de modèles de langages neuronaux s'est révélée très efficace, améliorant une base de référence du CRF de 28% et un modèle de langage spécialisé unique de 4%.

ABSTRACT

Named entity recognition (NER) is key for biomedical applications as it allows knowledge discovery in free text data. As entities are semantic phrases, their meaning is conditioned to the context to avoid ambiguity. In this work, we explore contextualized language models for NER in French biomedical text as part of the Défi Fouille de Textes challenge. Our best approach achieved an F_1-measure of 66% for symptoms and signs, and pathology categories, being top 1 for subtask 1. For anatomy, dose, exam, mode, moment, substance, treatment, and value categories, it achieved an F_1-measure of 75% (subtask 2). If considered all categories, our model achieved the best result in the challenge, with an F_1-measure of 72%. The use of an ensemble of neural language models proved to be very effective, improving a CRF baseline by up to 28% and a single specialised language model by 4%.

MOTS-CLÉS : reconnaissance d'entités nommées, encapsulation de mots contextualisés, CRF,

∗. Authors JC and JK contributed equally to this work.

Atelier DÉfi Fouille de Textes, pages 36–48.

BERT, CamemBERT.

KEYWORDS: named entity recognition, contextualized word embeddings, CRF, BERT, CamemBERT.

1 Introduction

The large amount of raw text data available in the biomedical domain enables to leverage the wealth of the content. Combined with manually curated data, it allows the development of automatic techniques to unlock the value of the raw resources for supporting healthcare and advance science. In particular, information extraction methods enable the extraction of specific data types from text data (e.g., entities). Information extraction fosters several applications from tracking of technologies (Teodoro *et al.*, 2010) in patents to clinical decision support (Liu *et al.*, 2016), biocuration assistance (Liu *et al.*, 2016; Teodoro *et al.*, 2020), and healthcare-associated infections detection (Tvardik *et al.*, 2018) in the biomedical domain. It has also important challenges associated with the application domain and the language in which the text is available. Indeed, information extraction systems for recognizing entities are mostly focused on English. However, it is widely recognised that it is crucial that research expands to other languages in the same scale (Dupont, 2017; Grabar *et al.*, 2019).

Named entity recognition (NER) is a key part in information extraction systems. Named entities are phrases that contain names of persons, organizations, locations (Tjong Kim Sang & De Meulder, 2003) to name but a few examples. There are many studies for NER in French language, for instance in i) journalistic data (Dupont, 2017; Martin *et al.*, 2020) with a set of entities like person, organization, company, location, point of interest, fiction character and product ; and ii) recognizing entities in tweets (Sileo *et al.*, 2017), including person, music artist, organisation, product and media, among others. In the biomedical domain, recognizing entities is mainly focused on semantic groups and concepts from Unified Medical Language System (UMLS) on The Quaero French Medical Corpus (Névéol *et al.*, 2014). The Quaero corpus contains annotated French Medline (titles and abstracts) and European Medicines Agency (EMEA) documents (drug labels).

Community challenges, such as CLEF eHealth, have been evaluating specific information extraction tasks for the clinical domain (Sankhavara & Majumder, 2019). Erasmus MC, one of the CLEF eHealth top scorers, is a dictionary-based NER for French UMLS and translations for non-French terms, achieving 74.9% of F_1-measure for EMEA and 69.8% of F_1-measure for Medline corpus in semantic groups annotation (Van Mulligen *et al.*, 2016). In Erasmus MC, semantics could be missed by the presence of compounded semantic groups or UMLS concepts. Similarly, SIFR annotator is a semantic annotator for French clinical narratives (Tchechmedjiev *et al.*, 2018). It relies on Mgrep, a concept recognizer based on label matching and heuristics, and achieves 62.6% of F_1-measure in EMEA and 58.9% of F_1-measure in Medline for semantic groups. The tool is limited due to the lack of word disambiguator and scarce French ontologies concerning English ontologies.

Deep neural language models have been recently leveraged to improve NER methods (Lee *et al.*, 2019). Deep neural language models are self-supervised models that take advantage of free text to learn word representations using their context (Turian *et al.*, 2010). With the advent of low-dimensional representation models supported by deep neural networks, such as word2vec (Mikolov *et al.*, 2013), the importance of word representations has become more evident. Further research has been taken to find out more accurate representations of words, such as in Global Vectors (GloVe) (Pennington *et al.*, 2014), and to the more recent contextualized representations, like ELMo (Peters

et al., 2018), UMLFiT (Howard & Ruder, 2018), and BERT (Devlin *et al.*, 2019). In particular, BERT is based on the transformers architecture, which uses an attention mechanism, via bidirectional pre-training from unlabeled text, conditioned in left and right contexts in all layers (Devlin *et al.*, 2019). BioBERT (Lee *et al.*, 2019), a BERT-based model trained on large-scale biomedical corpora has shown significant improvements in downstream tasks in the biomedical domain, including NER. Similarly, CamemBERT [1] is a contextualized language model trained and optimized specifically for French language (Martin *et al.*, 2020; Devlin *et al.*, 2019) based on RoBERTa model (Liu *et al.*, 2019).

In this paper, we investigate contextualized language models for French NER in clinical texts in the context of the Information Extraction task of the Défi Fouille de Textes (DEFT) challenge (Cardon *et al.*, 2020). This task is divided in two subtasks, which aim to identify *anatomie* (anatomy), *dose*, *examen*, *mode*, *moment*, *pathologie* (pathology), *sosy* (symptoms and signs), *substance*, *traitement* and *valeur* (value) entities in clinical narratives. Inasmuch as each language has its own peculiarities, our hypothesis is that it is worth designing a specific language model for French clinical corpora. Thus, we explore a CamemBERT-based model pre-trained on a biomedical corpus and fine-tuned on the DEFT information extraction task data. We compare its performance with multilingual BERT, CamemBERT and an ensemble of language models. In the following sections, we describe the design and results of the experiments.

2 Methods

In this work, we explore two perspectives for NER : as information extraction and as word representation. For the first, named entities are considered as a sequence classification problem, for which we propose a baseline method using the conditional random fields (CRF) framework. For the latter, our methodology is based on different deep neural language models derived from the BERT architecture. These methods were used to extract named entities in subtask 1 - symptoms and signs, and pathology - and subtask 2 - anatomy, dose, exam, mode, moment, substance, treatment, and value of the DEFT Information extraction task.

2.1 Conditional Random Fields

We used a linear chain CRF sequence classifier as a baseline and relied on the implementation of *CRFSuite* [2]. This probabilistic graphical model considers correlations between the neighborhood of words in a sentence and its features, jointly with the corresponding labels. Such correlation allows this model to learn the labels in a sequence (Lafferty *et al.*, 2001). In fact, linear chain CRF estimates the conditional probability of a label given a word sequence (Sutton, 2012). As shown in Table 1, our model relies on a set of NER standard features defined over a window of ± 2 tokens (Guo *et al.*, 2014; Copara *et al.*, 2016), including the word itself, lower-cased word, capitalization pattern, prefixes, suffixes, among others. Additionally, we used language-based features, such as lower-casing the words in the text, checking if the current token is a measure unit and whether the current token contains a French character. It is worth noting that we have not used gazetteers extensively, just a short list of units.

1. https://camembert-model.fr/
2. http://www.chokkan.org/software/crfsuite/

	Feature					
word	Une	première	dose	de	20	mg
lowercase word	une	première	dose	de	20	mg
capitalization pattern	ULL	LLLLLLLL	LLLL	LL	DD	LL
type	InitUpper	AllLetter	AllLetter	AllLetter	AllDigit	AllLetter
prefixes	u, un, une	p, pr, pre	d, do, dos	d, de	2, 20	m, mg
sufixes	e, ne, une	e, re, ère	e, se, ose	e, de	0, 20	g, mg
unit	no	no	no	no	no	yes
french char	no	yes	no	no	no	no

TABLE 1 – Example of features for the sentence *"Une première dose de 20 mg"*. *U → uppercase ; L → lowercase ; D → digit.

In our CRF model each entity is associated with one label (as usually in NER) and when there are nested entities, we keep entities that encompass other and dismiss nested entities.

2.2 Transformers with a token classification on top

For this experiment, we selected five BERT-based language models. The first, bert-base-multilingual-cased (Devlin *et al.*, 2019), is used as our transformer baseline as it was not trained specifically on a French corpus. The second and the third models, camembert-base and camembert-large, respectively, are based on the RoBERTa architecture (Liu *et al.*, 2019), a BERT-based model with some changes (tokenizer, training task, optimization, etc.) and trained on a large French corpus (Martin *et al.*, 2020). Models 4 and 5, so called, camembert-bio-base and camembert-bio-large, respectively, are CamemBERT-based models pre-trained on a french biomedical corpus containing 31k+ scientific publications extracted from PubMed. To further pre-train these models, we took CamemBERT weights as a starting point. Then, using an Adam optimizer (Kingma & Ba, 2014), we minimized a masked-language modeling loss. We trained it using 512 tokens during 5 epochs with a learning rate of 5e-5 and batch size of 24 [3].

As RoBERTa models are based on the BERT architecture, all our base and large models share hyper-parameters. For the base models, we have 12 layers (L), with 768 hidden units (H) and 12 attention heads (A). For the large architecture versions, we have L=24, H=1024 and A=16. The multilingual BERT model [4] uses WordPiece [5] as a tokenizer whereas the CamemBERT-based models use SentencePiece (Kudo & Richardson, 2018). [6] The tokenizer's choice was driven by the original model's tokenizer. Indeed, as we were fine-tuning BERT or CamemBERT models, we had to reuse the whole pipeline which includes the tokenizer (makes the link between a token and its trained representation possible). As explained in their paper (Martin *et al.*, 2020), SentencePiece does not require pre-tokenization which makes it a non-language specific tokenizer. Table 2 summarizes these architectural differences.

3. For the large model, as each step was too big for our 4 GPUs machine, we used gradient accumulation (i.e., the accumulation of 2 batches of 12 in order to get a batch of 24)

4. The multilingual BERT model uses a vocabulary size of 30K.

5. Comparison between WordPience and SentencePiece tokenizers : https://github.com/google/sentencepiece

6. CamemBERT uses a vocabulary size of 32K.

Tokenizer	WordPiece	SentencePiece			
Model	bert-base-multilingual-cased	camembert-base	camembert-bio-base	camembert-large	camembert-bio-large
layer (L)	12			24	
hidden (H)	768			1024	
heads (A)	12			16	

TABLE 2 – Architectural differences of BERT-based models.

For the fine-tuning of the NER models, we used the hugging face [7] framework, which basically standardizes the process for all the transformers. Each NER model is a BERT module with a fully connected layer on top of the hidden states of each token. As entities could overlap, we decided to use a binary or one-vs-all approach per entity instead of using a softmax which does not allow multi-labelling. All previously presented language models were fine-tuned on the DEFT task 3 dataset for 20 epochs, with a sequence length of maximum 256 tokens, a learning rate of 4e-5 and a warmup proportion of 0.1. As for the CRF baseline, we use one label for each entity, discarding nested entities.

2.3 Dataset

In DEFT task 3 - information extraction - there are two subtasks. Subtask 1 is focused on the *pathologie* and *sosy* (symptoms and signs) entities. Subtask 2 concerns the identification of *anatomie*, *dose*, *examen*, *mode*, *moment*, *substance*, *traitement*, and *valeur* entities. For assessing these subtasks, the challenge organisers provided a training dataset composed of 100 French clinical documents manually annotated with the 8098 entities (Grabar *et al.*, 2018). The annotated data include all the entities mentioned for each subtask, in addition to informational entities (e.g. *date*, *durée*, *frequence*) that have not been considered in our models. An example of annotation is shown in Figure 1. As we can notice, nested entities appear often in the annotations, sometimes within the same subtask and sometimes across subtasks.

7. https://huggingface.co/transformers/

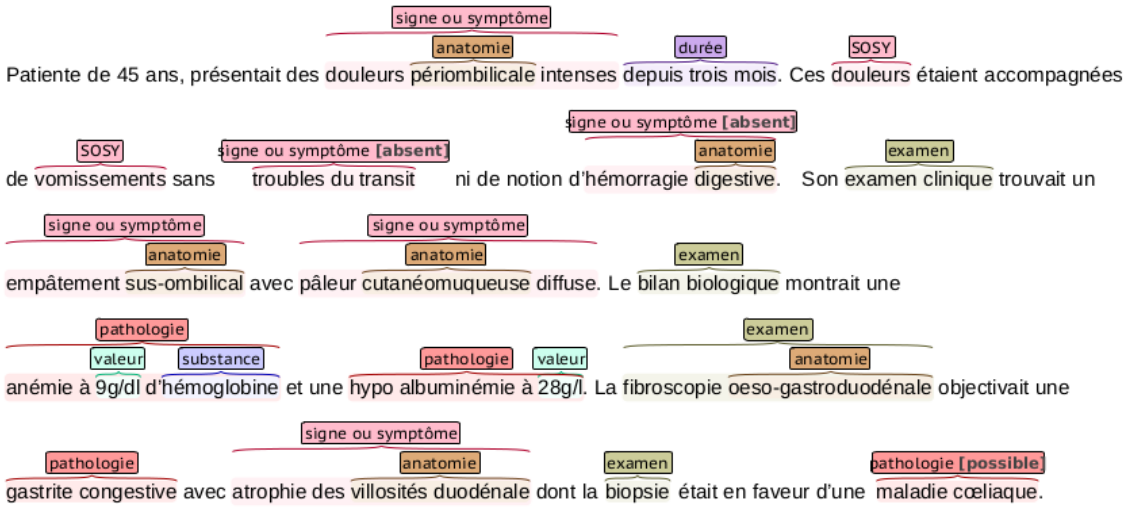

FIGURE 1 – An example of clinical narrative with entity annotations for subtasks 1 and 2. The annotations are color coded.

Table 3 shows the distribution of annotations among the entities in the training data. The majority of annotations come from the *sosy*, *anatomie* and *examen* entities, which compose together 54% of the training data. On the other hand, *mode*, *dose* and *moment* represent 13% of the dataset. To train and validate our models in the training phase, this dataset was split into train (80%), dev (10%) and test (10%) sets. The hyper-parameters of the models were selected for the test phase based on their performance on the dev set.

Entity	Train (count / %)	Dev (count / %)	Test (count / %)	All (count / %)
anatomie	1241/ 19	57 / 6	174 / 26	1472 / 18
dose	302 / 5	40 / 4	5 / 1	347 / 4
examen	962 / 15	119 / 12	137 / 20	1218 / 15
mode	214 / 3	24 / 2	11 / 2	249 / 3
moment	363 / 6	77 / 8	54 / 8	494 / 6
pathologie	260 / 4	91 / 9	184 / 27	535 / 7
sosy	1451 / 23	196 / 20	33 / 5	1680 / 21
substance	883 / 14	85 / 8	22 / 3	990 / 12
traitement	301 / 5	193 / 19	52 / 8	546 / 7
valeur	443 / 7	119 / 12	5 / 1	567 / 7
Total	6420 / 100	1001 / 100	677 / 100	8098 / 100

TABLE 3 – Entity distribution for the training phase collection.

3 Results and Discussion

In this section, we present the results of our models for the training and test phases for both subtasks. During the training phase, we used only the training collection provided in the challenge in order to

develop and tune our models. During the test phase, we evaluated over the test collection with the parameters identified in the training phase.

3.1 Training phase

Table 4 shows the results of our models in the training phase. The baseline CRF model achieved 0.4641 of overall micro F_1-measure, having a highest F_1-measure for the *valeur* entity (0.7708) and the lowest for the *pathologie* entity (0.1967). For the transformer-based models, the camembert-bio-base model outperforms both base models (BERT and CamemBERT) for the overall micro and macro F_1-measures, demonstrating the effectiveness of the specific biomedical corpus for the clinical NER task. For the large transformer models, CamemBERT achieves the highest micro F_1-measure (0.6826) and camembert-bio achieves the highest macro F_1-measure (0.5790). All contextualized language models outperform the baseline CRF model significantly, showing the outstanding performance of these architectures for NER in biomedical French texts.

As described in Section 2.2, we used a one-vs-all approach to predict the overlapped (nested) entities for the transformer models. This approach was not as effective for the CRF baseline, reducing the overall (micro) F_1-measure performance to 0.4464. Using this approach, no *dose* entity was correctly recognised and F_1-measure of *traitement* decreased to 0.1538. Nonetheless, the performance for recognising *anatomie* improved to 0.5300 of F_1-measure. These results suggest that in order to predict *dose* or *traitement* correctly, it is necessary to observe near entities, given the nature of the CRF learning model. We believe that the *anatomie* entity increased its performance mainly due to the fact that it appears usually nested in the *sosy*, *pathologie* or *examen* entities. However, in a one-vs-all approach, this entity will not be nested.

F_1-measure	baseline	bert-base-multilingual-cased	camembert-base	camembert-large	camembert-bio-base	camembert-bio-large
anatomie	0.3673	0.7170	0.7675	**0.8022**	0.7921	0.7751
dose	0.2500	0.1111	**0.4286**	0.1538	0.2857	0.1538
examen	0.5727	0.6618	0.6957	0.6667	0.6926	**0.7011**
mode	0.2857	0.2857	0.3333	0.2857	**0.4444**	0.2500
moment	0.4000	0.6957	**0.7273**	0.6364	0.6667	**0.7273**
pathologie	0.1967	0.3725	0.4956	**0.5714**	0.4248	0.5474
sosy	0.4356	0.6139	0.5838	**0.6961**	0.6563	0.6772
substance	0.4878	0.4400	0.3902	0.5854	0.5909	**0.6341**
traitement	0.4255	0.4590	0.3810	0.4848	0.3582	**0.5161**
valeur	0.7708	0.7961	0.7885	**0.8302**	0.8182	0.8077
Overall (micro)	0.4641	0.6135	0.6311	**0.6826**	0.6569	0.6791
Overall (macro)	0.4192	0.5153	0.5592	0.5713	0.5730	**0.5790**

TABLE 4 – Evaluation of different models in the training phase.

We also assessed a voting strategy, or ensemble, between the transformers models, where all 5 BERT

models vote with their predictions. For example, when the voting threshold $v = 1$, we use the set of positive predictions coming from all the models, whereas when $v = 3$, we only use the positive predictions when the majority of the models agree on an annotation. As shown in Figure 2, the precision increases proportionally to the voting threshold, whereas the recall decreases. This clearly points out the fact that the predictions of those models are different, otherwise recall would keep constant as we increase the number of votes to validate a positive prediction. For the test phase, we used the ensemble threshold $v = 3$, which resulted in the best overall (micro) F_1-score in the training phase.

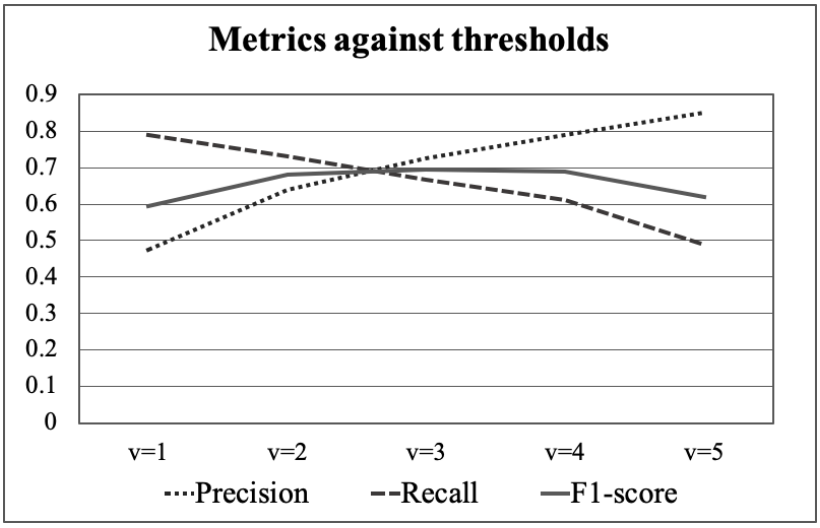

FIGURE 2 – Validation of the voting strategy.

3.2 Test phase

In the test phase, we evaluated 3 runs for a dataset of 67 clinical narratives. For run 1, we used the baseline model based on CRF. For run 2, we used the camembert-bio-large model. Finally, for run 3, we used an ensemble based on a voting threshold of 3. The performance of our models is summarised in Table 5. The ensemble model achieves 0.7262 of overall micro F_1-measure, surpassing in 2.6% the camembert-bio-large and in 14.8% the baseline. Taking into account camembert-bio-large and baseline models, the former is better by 12.18%. This clearly shows that transformer methods in biomedical French NER reach outstanding performance by only leveraging wealth in unstructured data and without the necessity to design handcrafted features. Concerning the ensemble model, it achieved the best overall F_1-measure for both subtasks among our models, being the highest score for subtask 1 and for the "all categories" evaluations among all models in the competition.

Similarly to the training phase, the highest F_1-measure in the test phase is achieved for the *valeur* entity (0.8561). This entity represents 9% of the annotations in the test collection, while in the whole data collection it represents 16%. Thus, it seems that the training data is sufficiently characterized to learn this entity automatically. The lowest performance for the ensemble method is found for *dose* entity, as well we can confirm the lowest performance for this entity in Table 4 (during training phase). This can be due to the variety of values in the annotated data, combining numbers and words (e.g. *de 0,5 à 0,75 litre*), measure units (e.g. *1mg/kg/j*) or simply words that easily could be associated with a non-entity word (e.g. *'24 paquets/année'* or *'02'*). *Mode* entities mostly are words without

abbreviations neither numbers (e.g. *'voie parasternale droite'* or *'voie centrale intraveineuse'*) i.e. it contains less variety in the kind of values, this could come with an easier way to learn patterns and make predictions.

Task 3	F_1-measure	baseline	bert-base-multilingual-cased*	camembert-large*	camembert-bio-large	ensemble (t=3)
Subtask 1	pathologie	0.3984	0.3628	0.5617	0.5344	0.5644
	sosy	0.5091	0.5574	0.6318	0.6268	0.6733
	Overall	0.4984	0.5303	0.6225	0.6153	**0.6603**
Subtask 2	anatomie	0.5561	0.7646	0.8024	0.7978	0.8069
	dose	0.3684	0.3604	0.5385	0.4118	0.5217
	examen	0.6787	0.6842	0.7169	0.7149	0.7333
	mode	0.3423	0.5935	0.6543	0.6386	0.6486
	moment	0.6273	0.6748	0.7219	0.7576	0.7869
	substance	0.578	0.5586	0.6667	0.6702	0.6379
	traitement	0.4756	0.4598	0.5724	0.5573	0.6076
	valeur	0.7969	0.8160	0.8637	0.8393	0.8561
	Overall	0.6151	0.6894	0.7441	0.7370	**0.7547**
All categories	**Overall** (micro)	0.5778	0.6380	0.7073	0.6996	**0.7262**
	Overall (macro)	0.5331	0.5832	0.6730	0.6549	**0.6837**

TABLE 5 – Test phase results. *Non-official runs.

Table 6 shows the statistics of the official results for all participants in the challenge. For subtask 1, our voting approach resulted in a F_1-score of 0.6603, which is the max reported for the overall result of the competition. In subtask 2, our best model was 1% lower than the top score, which achieved a F_1-score of 0.7626 (against our 0.7547). Considering both tasks across all categories, our voting model achieved the highest score in the competition. This results is shown in the "Non official" of Table 6 provided by the challenge organisers. The "Non official" results takes into account even entities that were counted as informational (e.g. *date*, *durée*, *frequence*). As we did not predict any of those informational entities, our F_1-measure for those are 0.0000 and we end up with a diminished overall F_1-measure of 0.7152, which is the max reported in the non official row. However, without taking into account those entities, our overall F_1-measure is 0.7262 as reported in Table 5.

Task 3	Min	Max	Median	Mean
Subtask 1	0.0645	0.6603	0.4557	0.4347
Subtask 2	0.1352	0.7626	0.6151	0.6012
Non official	0.1297	0.7152	0.5679	0.5533

TABLE 6 – Official summarize results over DEFT task 3.

3.3 CamemBERT *vs.* CamemBERT bio

If we compare the results of the camembert-base model against the camembert-bio-base model, we notice a significant improvement in performance for the latter. However, this result is unexpectedly not translated to the large version of the camembert model, as locally trained models tend to have superior performance (Lee *et al.*, 2019). We believe that this is due to the size of the biomedical corpus (31k French abstracts from PubMed) used to pre-train the CamemBERT models, which is relatively small compared to the size of the original CamemBERT corpus. For a comparison, BioBERT (Lee *et al.*, 2019) was pre-trained on 18B words corpora extracted from PubMed and PMC ; while Clinical BERT (Alsentzer *et al.*, 2019) was trained on clinical text from approximately 2 million notes in the MIMIC-III v1.4 database. While the biomedical French corpus works well for the smaller model, it was limited to improve the original camembert-large model weights for the specificities of the biomedical language as this model contains much more parameters than the base version (335M vs. 110M parameters).

Nevertheless, what makes our approach powerful is the dissimilarities of the respective model predictions. Indeed, if camembert-bio and CamemBERT models were to predict the same entities for a given text, the voting would not have made any sense. Our hypothesis is that, by creating different models, we were able to start our fine-tuning with a language model that has different perspectives. Then, by allowing each model to vote, we were able to outperform the camembert-large model by two basis points. To verify this hypothesis, it would be interesting to see if fine-tuning the same model 5 times (number of models we used for voting) would have improved its performances. For example, would the camembert-large have improved if we had fine-tuned it 5 times and used those 5 models as an ensemble ? The only randomness in such experiment would be the order of the documents during the training phase.

3.4 Language specific *vs.* multi-language model

Bert-base-multilingual-cased model was trained in 104 languages, including French, however camembert-base model was trained and optimized specifically for French language. The camembert-base model (not part of the official evaluation) shows slightly better performance for most entities (*pathologie*, *sosy*, *anatomie*, *dose*, *examen*, *moment*, *traitement*, *valeur*) compared to the bert-base-multilingual-cased model.

For subtask 1, camembert-base achieves 0.5802 of overall F_1-measure *vs* 0.5303 of bert-base-multilingual-cased, i.e. almost 5% of improvement. In subtask 2, an overall F_1-measure of 0.7081 of camembert-base *vs* 0.6894 of overall F_1measure of bert-base-multilingual-cased, makes camembert-base 1.87% better. These improvements in camembert-base highlight a direct relationship between language and performance. In addition, these differences show that subtask 1 is more dependent on the language than subtask 2.

4 Conclusion

Among the experiments we have done, we can conclude that recognizing entities in biomedical domain is not a straightforward task and adding the complexity of language makes this task more difficult. For DEFT task 3 challenge, we proposed mainly two families of learning methods: a baseline

based on CRF and a set of transformers language models. We focused on exploring the performance of our NER models with the contextualized language models enriched with local text. Our best results were given by the ensemble method based on a voting strategy between the BERT based models, including CamemBERT (trained specifically for French) and CamemBERT-bio (trained on French biomedical texts), achieving 66% F_1-measure in subtask 1 and 75% F_1-measure in subtask 2. The ensemble of deep neural language models proved to be the most effective method for biomedical information extraction in French texts. As next steps, we will investigate whether fine-tuning the same model a number of times would improve performance. We are also interested to investigate nested entities approaches over DEFT task 3 data.

Références

ALSENTZER E., MURPHY J., BOAG W., WENG W.-H., JINDI D., NAUMANN T. & MCDERMOTT M. (2019). Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, p. 72–78, Minneapolis, Minnesota, USA : Association for Computational Linguistics. DOI : 10.18653/v1/W19-1909.

CARDON R., GRABAR N., GROUIN C. & HAMON T. (2020). Présentation de la campagne d'évaluation DEFT 2020 : similarité textuelle en domaine ouvert et extraction d'information précise dans des cas cliniques. In *DEFT 2020 - Défi fouille de texte*, France.

COPARA J., OCHOA LUNA J. E., THORNE C. & GLAVAŠ G. (2016). Spanish NER with word representations and conditional random fields. In *Proceedings of the Sixth Named Entity Workshop*, p. 34–40, Berlin, Germany : Association for Computational Linguistics. DOI : 10.18653/v1/W16-2705.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (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)*, p. 4171–4186, Minneapolis, Minnesota : Association for Computational Linguistics. DOI : 10.18653/v1/N19-1423.

DIAS G., Éd. (2015). *Actes de TALN 2015 (Traitement automatique des langues naturelles)*, Caen. ATALA, HULTECH.

DUPONT Y. (2017). Exploration de traits pour la reconnaissance d'entités nommées du Français par apprentissage automatique. In *TALN 2017*, Orléans, France. HAL : hal-02448614.

GRABAR N., CLAVEAU V. & DALLOUX C. (2018). CAS : French corpus with clinical cases. In *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis*, p. 122–128, Brussels, Belgium : Association for Computational Linguistics. DOI : 10.18653/v1/W18-5614.

GRABAR N., CLAVEAU V. & DALLOUX C. (2019). **28**(01), 218–222. DOI : 10.1055/s-0039-1677937.

GUO J., CHE W., WANG H. & LIU T. (2014). Revisiting embedding features for simple semi-supervised learning. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 110–120, Doha, Qatar : Association for Computational Linguistics. DOI : 10.3115/v1/D14-1012.

HOWARD J. & RUDER S. (2018). Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume*

1 : Long Papers), p. 328–339, Melbourne, Australia : Association for Computational Linguistics. DOI : 10.18653/v1/P18-1031.

KINGMA D. P. & BA J. (2014). Adam : A method for stochastic optimization. *arXiv preprint arXiv :1412.6980*.

KUDO T. & RICHARDSON J. (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* : Association for Computational Linguistics. DOI : 10.18653/v1/d18-2012.

LAFFERTY J. D., MCCALLUM A. & PEREIRA F. C. N. (2001). Conditional random fields : Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, p. 282–289, San Francisco, CA, USA : Morgan Kaufmann Publishers Inc.

LEE J., YOON W., KIM S., KIM D., KIM S., SO C. H. & KANG J. (2019). BioBERT : a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*. DOI : 10.1093/bioinformatics/btz682.

LIU F., CHEN J., JAGANNATHA A. & YU H. (2016). Learning for biomedical information extraction : Methodological review of recent advances. *CoRR*, **abs/1606.07993**.

LIU Y., OTT M., GOYAL N., DU J., JOSHI M., CHEN D., LEVY O., LEWIS M., ZETTLEMOYER L. & STOYANOV V. (2019). Roberta : A robustly optimized BERT pretraining approach. *CoRR*, **abs/1907.11692**.

MARTIN L., MULLER B., ORTIZ SUÁREZ P. J., DUPONT Y., ROMARY L., VILLEMONTE DE LA CLERGERIE É. & SAGOT B. (2020). CamemBERT : a Tasty French Language Model. In *The 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*, Seattle, Washington, United States.

MIKOLOV T., SUTSKEVER I., CHEN K., CORRADO G. & DEAN J. (2013). Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, p. 3111–3119, Red Hook, NY, USA : Curran Associates Inc.

NÉVÉOL A., GROUIN C., LEIXA J., ROSSET S. & ZWEIGENBAUM P. (2014). The QUAERO French medical corpus : A ressource for medical entity recognition and normalization. In *Proc of BioTextMining Work*, p. 24–30.

PENNINGTON J., SOCHER R. & MANNING C. (2014). Glove : Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* : Association for Computational Linguistics. DOI : 10.3115/v1/d14-1162.

PETERS M., NEUMANN M., IYYER M., GARDNER M., CLARK C., LEE K. & ZETTLEMOYER L. (2018). Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long Papers)* : Association for Computational Linguistics. DOI : 10.18653/v1/n18-1202.

SANKHAVARA J. & MAJUMDER P. (2019). Advances in biomedical entity identification : A survey. In *Biotechnology and Biological Sciences*, p. 114–120. CRC Press. DOI : 10.1201/9781003001614-19.

SILEO D., PRADEL C., MULLER P. & DE CRUYS T. V. (2017). Synapse at cap 2017 NER challenge : Fasttext CRF. *CoRR*, **abs/1709.04820**.

SUTTON C. (2012). An introduction to conditional random fields. *Foundations and Trends® in Machine Learning*, **4**(4), 267–373. DOI : 10.1561/2200000013.

TCHECHMEDJIEV A., ABDAOUI A., EMONET V., ZEVIO S. & JONQUET C. (2018). SIFR annotator : ontology-based semantic annotation of french biomedical text and clinical notes. *BMC Bioinformatics*, **19**(1). DOI : 10.1186/s12859-018-2429-2.

TEODORO D., GOBEILL J., PASCHE E., RUCH P., VISHNYAKOVA D. & LOVIS C. (2010). Automatic ipc encoding and novelty tracking for effective patent mining. In *The 8th NTCIR Workshop Meeting on Evaluation of Information Access Technologies : Information Retrieval, Question Answering, and Cross-Lingual Information Access*, p. 309–317, Tokyo, Japan.

TEODORO D., KNAFOU J., NADERI N., PASCHE E., GOBEILL J., ARIGHI C. N. & RUCH P. (2020). UPCLASS : a deep learning-based classifier for UniProtKB entry publications. *Database*, **2020**. DOI : 10.1093/database/baaa026.

TJONG KIM SANG E. F. & DE MEULDER F. (2003). 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. 142–147.

TURIAN J., RATINOV L.-A. & BENGIO Y. (2010). Word representations : A simple and general method for semi-supervised learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, p. 384–394, Uppsala, Sweden : Association for Computational Linguistics.

TVARDIK N., KERGOURLAY I., BITTAR A., SEGOND F., DARMONI S. & METZGER M.-H. (2018). Accuracy of using natural language processing methods for identifying healthcare-associated infections. *International Journal of Medical Informatics*, **117**, 96–102. DOI : 10.1016/j.ijmedinf.2018.06.002.

VAN MULLIGEN E. M., AFZAL Z., AKHONDI S., VO D. & KORS J. (2016). Erasmus mc at clef ehealth 2016 : Concept recognition and coding in french texts. In *CEUR Workshop Proceedings*, p. 171–178.