Ensemble of Deep Masked Language Models for Effective Named Entity Recognition in Health and Life Science Corpora

Nona Naderi1,2*, Julien Knafou1,2, Jenny Copara4,1,2, Patrick Ruch1,2 and Douglas Teodoro4,1,2*

1Information Science Department, University of Applied Sciences and Arts of Western Switzerland (HES-SO), Geneva, Switzerland, 2Swiss Institute of Bioinformatics, Geneva, Switzerland, 3Computer Science Department, University of Geneva, Geneva, Switzerland, 4Department of Radiology and Medical Informatics, University of Geneva, Geneva, Switzerland

The health and life science domains are well known for their wealth of named entities found in large free text corpora, such as scientific literature and electronic health records. To unlock the value of such corpora, named entity recognition (NER) methods are proposed. Inspired by the success of transformer-based pretrained models for NER, we assess how individual and ensemble of deep masked language models perform across corpora of different health and life science domains—biology, chemistry, and medicine—available in different languages—English and French. Individual deep masked language models, pretrained on external corpora, are finetuned on task-specific domain and language corpora and ensembled using classical majority voting strategies. Experiments show statistically significant improvement of the ensemble models over an individual BERT-based baseline model, with an overall best performance of 77% macro F1-score. We further perform a detailed analysis of the ensemble results and show how their effectiveness changes according to entity properties, such as length, corpus frequency, and annotation consistency. The results suggest that the ensembles of deep masked language models are an effective strategy for tackling NER across corpora from the health and life science domains.

Keywords: named entity recognition, deep learning, patent text mining, transformers, clinical text mining, chemical patents, clinical NER, wet lab protocols

1 INTRODUCTION

In the health and life science domains, most of the information is encoded in unstructured reports. For example, it is estimated that around 90% of electronic health records (EHR) data are available as free text. While text format facilitates capturing information, it makes the secondary use of the data challenging. To support data structuring and to unlock the value of textual databases in secondary usage applications, named entity recognition (NER) methods have been proposed. NER is the task for detecting entities in text and assigning concept names, or categories, to them. The health and life science domains are notoriously known for their wealth of named entities and synonyms, such as microorganism taxonomies, drug brands, and gene names, to name a few. This richness of named entities together with the variety of formats, abbreviations, and (mis)spellings makes NER in health and life science corpora, like EHR, lab protocols, and scientific publications, a challenging task.
Basic NER approaches use the construction of dictionaries of named entities and the specification of tagging rules (Quimbaya et al., 2016). They normally require domain knowledge and feature engineering. While they are effective for simple and small corpora, their effectiveness is often limited when entities are complex and available in large numbers, as it is often the case in health and life sciences. Moreover, as the corpus evolves, it is hard to maintain the rules. More sophisticated methods are based on classical machine learning models, such as support vector machines, decision trees, hidden Markov models (Zhao, 2004), and conditional random fields (CRFs) (Li et al., 2008; Rocktäschel et al., 2012; Leaman et al., 2015). In these methods, annotated examples of text passages with entity classes are used to train the models. Textual features are combined with entity annotations to increase the model's performance. As the models are trained only on the annotated corpus, which is usually small, they struggle to generalize to out-of-sample data. Thus, they are currently mainly used to provide a baseline for more effective model evaluation or in combination with more powerful models.

More recently, deep masked language models trained on large corpora have achieved state-of-the-art in most NLP-related tasks, including NER. Bidirectional Encoder Representations from Transformers (BERTs) (Devlin et al., 2019) were the first to explore the transformer architecture as a general framework for NLP (Vaswani et al., 2017). Once the model is trained (or pretrained in the BERTology parlance) on a large corpus, it can be adapted and effectively fine-tuned on specialized downstream NLP tasks, such as question-answering, text classification, and NER by leveraging the feature representations learned by the model during the pretraining phase in combination with examples of the specific task. Since the advent of BERT, a myriad of transformer-based masked language models have been proposed (Alsentzer et al., 2019; Liu et al., 2019; Yang et al., 2019). These models vary mostly in the tokenization used, in how the masking is performed, and in the trained data used during the pretraining phase. Language models pretrained on a specialized corpus, such as Medline, often tend to outperform models trained on a generic corpus for biomedical-related tasks.

In this study, our goal is to perform an empirical evaluation of how individual BERT-like models perform in the NER task across different health and life science corpora available in different languages. More specifically, we aim to assess how individual models compare to ensemble strategies in such scenarios. To do so, we leverage deep language models pretrained on the external text and fine-tune them on specific health and life science corpora. Then, their predictions are combined to create ensembles of named entity recognizers. We evaluate our models in chemistry, clinical, and wet lab corpora provided in the context of the ChEMU (Cheminformatics Elsevier Melbourne University) (He et al., 2020b), DEFT (Dé Fournier de Textes) (Grabar et al., 2018), and WNUT (Workshop on Noisy User-generated Text) (Tabassum et al., 2020) challenges, respectively. Our results show that the ensembles of named entity recognizers based on masked language models can outperform individual language models and achieve effective NER performances in these different domains and languages. We further perform an analysis of certain entity properties, including entity length, corpus frequency, and annotation consistency, to have a better understanding of the model's performance.

2 RELATED WORK

Deep learning approaches trained on large unstructured corpora have shown considerable success in NLP problems, including NER (Lample et al., 2016; Beltagy et al., 2019; Devlin et al., 2019; Jin et al., 2019; Liu et al., 2019; Andrioli de Souza et al., 2020). These models learn contextual token and sentence representations using often a self-supervised masked language algorithm, in which they attempt to predict masked tokens within sentences. This step is usually referred to as pretraining. The learned representations can then be reused in a supervised setting for downstream tasks, such as question-answering, NER, and text classification. For domain-specific tasks, models originally pretrained on general corpora, such as BERT, can be further pretrained or specialized on domain-specific corpora to improve the originally learned representations according to the domain specificity (Alsentzer et al., 2019; Lee et al., 2019; Gururangan et al., 2020). There exist also models pretrained only on domain-specific data (Beltagy et al., 2019; Gu et al., 2021), which reduces the overall training time as domain-specific corpora tend to be smaller in favor of lower generalization power. In both cases, in practice, those models are further trained or fine-tuned with task-specific examples. In this case, the model is no-longer trained to predict masked tokens but rather the actual NLP task, such as token classification in the case of NER.

Several models are proposed for cross-domain NER (Pan et al., 2013; Lin and Lu, 2018; Jia et al., 2019; Liu et al., 2020, 2021). These models are usually trained to leverage embeddings from the different domains via a transfer learning process to improve entity tagging. Only a few of these studies focus on health and life science NER. One study is that of Lee et al. (2018), in which the authors utilize the idea of transfer learning to identify named entities in the i2b2 2014/2016 corpus using a model trained on the MIMIC dataset. In this study, we adopt a different approach for the cross-domain problem. Instead of benefiting from joint named-entity learning, we investigate a methodology based on the ensemble of deep masked language models and show how it can be effectively applied across complex NER domains. Moreover, we believe this is the first work proposing a generic and robust approach for NER across chemical, clinical, and wet lab corpora available in English and French.

2.1 Chemical Named Entity Recognition

To further improve the performance of traditional approaches based on hand-crafted features for the extraction of chemical entities (Rocktäschel et al., 2012; Leaman et al., 2015; Habibi et al., 2016; Zhang et al., 2016; Akhondi et al., 2016), a number of studies leverage the power of word embeddings created using neural networks, such as word2vec (Mikolov et al., 2013), in combination with traditional approaches like CRF (Leaman et al., 2015; Rocktäschel et al., 2012) in a single recurrent network model, usually based on the long short-term memory (LSTM) architecture (Habibi et al., 2017; Corbett and Boyle, 2018; Zhai
et al., 2019; Hemati and Mehler, 2019). These methods have shown a significant improvement over the traditional methods on multiple datasets, such as CHEMDNER patent (Krallinger et al., 2015a,b) and BioSemantics (Akhondi et al., 2014). For example, on the chemical domain, Habibi et al. (2017) report about 5% improvement in F1-score using an LSTM-CRF model with word embeddings over a CRF with BANNER features (Leaman and Gonzalez, 2008), such as part-of-speech and character n-grams. Zhai et al. (2019) extended the Bidirectional LSTM-CRF (BiLSTM-CRF) model with contextualized word representations of Embeddings from Language Models (ELMo) (Peters et al., 2018) and reported an F1-score improvement of 3.7 percentage point over BiLSTM-CRF and LSTM character models.

Recently, the ChEMU evaluation lab (He et al., 2020b) organized an information extraction task from patent documents for the identification of chemical compounds and their specific roles in chemical reactions. The named entities in this task consist of four categories, including chemical compounds involving in a chemical reaction, conditions of the chemical reaction, yields for the final chemical product, and example labels. Teams participating in the task were evaluated based on both strict and relaxed span matching conditions. Various approaches have been proposed in the competition, including rule-based models (Dönmez et al., 2020; Wang et al., 2020), BiLSTM-CNN-CRF (Dao and Nguyen, 2020; Mahendran et al., 2020), and transformer-based models (Copara et al., 2020b; Dönmez et al., 2020; Ruas et al., 2020).

2.2 Clinical Named Entity Recognition

Various NER challenges and shared tasks, such as the i2b2 and n2c2 NLP challenges (Uzuner et al., 2010; Suominen et al., 2013; Kelly et al., 2014; Bethard et al., 2015; Névéol et al., 2015; Henry et al., 2020), fostered the development of NER methods (De Bruijn et al., 2011; Jiang et al., 2011; Kim et al., 2015; Van Mulligen et al., 2016; El Boukkouri et al., 2019) for the clinical domain in different languages (Lopes et al., 2019; Sun and Yang, 2019; Andrioli de Souza et al., 2020; Schneider et al., 2020). The DEFT challenge proposed an information extraction task for the French clinical corpus, with entities distributed across four categories: anatomy, clinical practices, treatments, and time (Cardon et al., 2020). Several teams participated in the challenge and the proposed approaches relied on rule-based models (Lemaître et al., 2020; Royan et al., 2020; Hiot et al., 2021), CRF-based models (Minard et al., 2020), and transformer-based models (Copara et al., 2020a; Nzali, 2020).

Similar to the chemical domain, word embeddings helped improve the recognition of entities in clinical corpora. Roberts (2016) used the combination of a general domain and in-domain word2vec embeddings and showed improvement over only in-domain embeddings. Using the i2b2 NLP dataset (Uzuner et al., 2011), El Boukkouri et al. (2019) showed that the concatenation of off-the-shelf ELMo contextualized representations (Peters et al., 2018) and word2vec embeddings trained on i2b2 task outperformed ELMo embeddings alone. Contextualized embeddings provided by ELMo were also used by Zhu et al. (2018). The authors used an ELMo version trained on medical articles from Wikipedia and clinical notes and reported the state-of-the-art on MIMIC-III. Wei et al. (2020) used three approaches to identify entities on n2c2 dataset: a CRF, a BiLSTM, and a joint BiLSTM-CRF model. They investigated different ensemble strategies to combine those models and found that the best results were achieved using a majority voting.

As in other NLP tasks, recent studies to extract entities from clinical corpora focus mostly on the use of deep masked language models. Si et al. (2019) trained BERT on MIMIC-III and showed further improvement over the previous models on MIMIC-III. Alsentzer et al. 2019 trained BERT and BioBERT (Lee et al., 2019), on MIMIC notes, and showed that Bio + Clinical BERT performed better than BERT and BioBERT trained on MedNLI dataset and i2b2 2010 datasets. Similarly, Schneider et al. (2020) demonstrated that the fine-tuned BERT using Portuguese clinical notes outperformed BERT trained on general corpora.

2.3 Wet Lab Named Entity Recognition

NLP approaches have only been applied to experimental protocols relatively recently (Soldatova et al., 2014; Kulkarni et al., 2018). Luan et al. (2019) introduced a model based on a dynamic span graph to jointly extract named entities and relations on wet lab protocols and other corpora. Wadden et al. (2019) built upon Luan et al. (2019)’s model by combining BERT and dynamic span graph. Dai et al. (2019) computed the similarity of the pretrained data and the data of the target application to investigate the effectiveness of pretrained word vectors. Their results showed that the word

---

**FIGURE 1** An example of a patent passage of the ChEMU dataset with entity annotations. The annotations are color-coded, representing the different entities in the dataset.
vector’s effectiveness depends on the vocabulary overlap of the source and target domains.

In contrast to the chemical and clinical domains, challenges and shared tasks are not as common for wet lab protocol corpora. Recently, WNUT-2020 (Tabassum et al., 2020) introduced a NER task for analyzing Wet Lab protocols. The task covers entity types from five categories of Action, Constituents, Quantifiers, Specifiers, and Modifiers. More than a hundred manually annotated protocols were used to evaluate the submissions of 13 teams. Most of the participants used NER models based on contextualized word representations (Knafou et al., 2020; Singh and Wadhawan, 2020; Sohrab et al., 2020; Vaidhya and Kaushal, 2020; Zeng et al., 2020). A few participants used CRF-based models (Acharya, 2020).

### 3 MATERIAL AND METHODS

#### 3.1 Datasets

In this section, we present the datasets used to train and assess the individual and ensembles of masked languages models for the extraction of named entities in chemical, clinical, and wet lab domains. The first dataset, provided in the context of the ChEMU 2020 challenge, consists of a collection of English chemistry patents annotated with chemical reaction entities. The second dataset, provided in the context of the DEFT 2020 challenge, consists of a collection of French EHR notes annotated with clinical entities. Finally, the third dataset, provided in the context of the WNUT 2020 challenge, consists of English laboratory protocols annotated with wet lab entities.

#### 3.1.1 Benchmark for Chemical Entity Recognition—ChEMU 2020 Dataset

The ChEMU 2020 benchmark dataset contains snippets sampled from 170 English patents from the European Patent Office and the United States Patent and Trademark Office (He et al., 2020b,a, 2021; Verspoor et al., 2020). As shown in Figure 1, these snippets are annotated with several chemical reaction entities, including reaction_product, starting_material, and temperature. The training and test set of the ChEMU dataset contains a total of 1,500 snippets annotated with 26,857 entities using the BRAT standoff format (Stenetorp et al., 2012).

![Figure 1](image1.png)

**Figure 1**: An example of a clinical narrative of the DEFT dataset with entity annotations. The annotations are color-coded, representing the different entities in the dataset. Notice that some entities are nested.

| Entity               | Train | Test  |
|----------------------|-------|-------|
| EL                   | example_label | 1,104 | 5.5   |
| OC                   | other_compound | 5,720 | 28.3  |
| RP*                  | reaction_product | 2,558 | 12.7  |
| RC                   | reagent_catalyst | 1,570 | 7.8   |
| So                   | Solvent | 1,390 | 6.9   |
| SM                   | starting_material | 2,167 | 10.7  |
| Te                   | Temperature | 1,861 | 9.2   |
| Ti                   | Time | 1,311 | 6.5   |
| YO                   | yield_other | 1,322 | 6.5   |
| YP                   | yield_percent | 1,183 | 5.9   |
| Total                |       | 20,186 | 100.0 |

**Table 1**: Entity distribution in the official training and test sets of ChEMU benchmark dataset.
3.1.2 Benchmark for Clinical Entity Recognition—DEFT 2020 Dataset

The DEFT benchmark dataset is a subset of the CAS corpus (Grabar et al., 2018), containing 100 French clinical documents manually annotated with the 8,098 entities in the following categories: pathologie, sosy (symptoms and signs), anatomie, dose, examen, mode, moment, substance, traitement, and valeur. An example of a clinical note annotation is shown in Figure 2. We can notice that nested entities appear in the annotations.

Table 2 shows the distribution of annotations among the entities in the training and test datasets. 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 pathologie represent together only 13% of the training dataset. In contrast to the ChEMU data, the distribution of the training and test sets varies significantly.

3.1.3 Benchmark for Wet Lab Entity Recognition—WNUT 2020 Dataset

The WNUT benchmark dataset is composed of 727 unique English wet lab protocols that describe experimental procedures (Kulkarni et al., 2018). The dataset was manually annotated with the 102,957 entities in the following categories: Action, Amount, Concentration, Device, Generic-Measure, Location, Measure-Type, Mention, Modifier, Numerical, Reagent, Seal, Size, Speed, Temperature, Time, and pH. An example of a lab protocol annotation is shown in Figure 3.

In Table 3, we see the distribution of the 18 entities by each subset. As it is commonly found in the health and life science domains, there is a significant class imbalance, with only two classes (Action and Reagent) representing more than 50% of annotations in the training set. Similar to the ChEMU dataset, the proportions of entities are fairly similar across the training and test subsets.

3.2 Proposed Methodology

Figure 4 shows a high-level view of our proposed ensemble model to recognize entities in health and life science corpora. In step 1 (data), documents are preprocessed to create small text units using a sentence-splitting algorithm. In step 2 (training), the resulting sentences with entity annotations are used to fine-tune the individual deep neural masked language models. In the training process, sentences are tokenized according to the specific language model tokenizer algorithm, and each token is assigned a label (entity class label or no-entity) based on the training annotations. Then, in step 3 (prediction), sentences are fed to the individual models previously fine-tuned, which split them into tokens and assign an entity class. Finally, in step 4 (ensemble), the predictions created for each token are aligned using a majority voting algorithm.

In the following, we describe the methodology to fine-tune a single deep masked language model to recognize named entities in the chemical, clinical, and wet lab domains in English and French corpora. Then, we detail how these different fine-tuned language models were combined to provide an ensemble NER model.

3.2.1 Single Deep Masked Language Model for Named Entity Recognition

To build the ensemble NER model, we fine-tuned different individual masked language models based on the transformers architecture (Vaswani et al., 2017). In the case of NER, masked language models are fine-tuned using a specialized training set—in our case, the chemical, clinical, and wet lab annotated...
To classify tokens according to the named entity classes, we used deep language models based on or derived from the BERT architecture. BERT was originally pretrained on a large corpus of English text extracted from BookCorpus (Zhu et al., 2015) and Wikipedia, with the different number of attention heads for the base and large types (12 and 24 transformer layers and hidden representations of 768 and 1,024 dimensions, respectively).

To fine-tune a particular masked language model for the NER task, we leverage the token representation created in its pretraining phase. A fully connected layer is added on top of the token representations and trained to classify whether a token belongs to a class or not. As transformers usually use tokenizers that work on word bits (or sub-tokens), during prediction, the entity label with the highest probability will be assigned to all sub-tokens of a word, and the sub-tokens will be then merged to build back the original word with the respectively assigned label. Finally, in a given sequence, if two adjacent words were given the same entity prediction, we would consider the two words as a phrase related to that entity.

Following this approach, the masked language model is then fine-tuned on the domain-specific data—chemical, clinical, and wet lab—using the training datasets previously discussed (ChEMU, DEFT, and WNUT). The fine-tuning is performed with the maximum sequence length of 265 tokens. The only preprocessing done was sentence-splitting. For the chemical and wet lab NER experiments, for which no nested entities were considered, we used a softmax function. Conversely, for the clinical NER, for which a token could be assigned to more than one entity, we used a sigmoid function to provide a multi-class classifier.
3.2.2 Ensemble of Deep Masked Language Models for Named Entity Recognition

Our ensemble method is based on a voting strategy, where each model votes with its predictions and a simple majority of votes is necessary to assign the predictions (Copara et al., 2020b,a; Knafou et al., 2020). In other words, for a given document, our models infer their predictions independently for each entity (as shown in Figure 4). Then, a set of passages (token or phrases) that received at least a vote for the named entities is taken into consideration for casting votes. This means that, for a given document and a given entity, we end up with multiple passages associated with a number of votes. Then, again for a given entity, the ensemble method will assign labels to all the passages that get the majority of votes. Note that each entity is predicted independently and that the voting strategy allows a passage to be labeled as positive for multiple entities at once. Thus, our ensemble strategy is also capable of assigning labels to nested entities.

3.3 Experimental Setup

3.3.1 Training Details

We conduct experiments using the three datasets listed in Tables 1, 2, and 3 for the individual models listed in Table 4. As shown in Table 5, we split the annotated collection into train, dev, and test sets and trained our models using subsets (train split) of the three datasets separately. The individual models of chemical, clinical, and wet lab NERs were fine-tuned on ChEMU, DEFT, WNUT train splits, respectively. The train, dev, and test sets were used to train the model weights, set the hyperparameters, and find the best ensemble configuration, respectively. The ensemble threshold for chemical and clinical NER was set to 3 and for wet lab NER to 4. More information about the fine-tuning of the models and the hyperparameter settings can be found in Copara et al. (2020b,a) and Knafou et al. (2020).

3.3.2 Evaluation Details

A blind test set (blind test split in Table 5), provided as part of the official evaluation for the respective challenges, was used to evaluate our models. Results are reported using the competition official metrics—precision, recall, and F1-score—considering the exact span matching, that is, both the starting and the end offsets of the text spans of the predicted and gold standard reference entities must match. They were computed using the BRAT eval tool\(^1\), and the evaluation code was provided by WNUT organizers against the blind test set split. The ensemble models created for the different domains are compared to the respective individual language models participating in the ensemble. The Student’s t-test is used to assess the significance of the results. Results are considered statistically significant for \(p\)-values smaller than 0.05.

\(^1\)https://bitbucket.org/nicta_biomed/brateval/
TABLE 6 | Comparison of the ensemble model with the individual models on three datasets (ChEMU, DEFT, and WNUT). “Individual model was taken as reference for the model’s baseline.” The official test set of ChEMU is not publicly available, so we report the results on the development set that was used as a test set (the training set was split into training and development sets). The comparison of individual models and ensemble for the DEFT and WNUT challenges are on the official test sets.

| Model          | P   | R   | F1   |
|----------------|-----|-----|------|
| **ChEMU**      |     |     |      |
| BERT-base-cased* | 90.83 | 91.14 | 90.98 |
| Ensemble (t = 3) | 93.78 | 90.87 | 92.30 |
| **DEFT**       |     |     |      |
| BERT-base-multilingual-cased* | 68.62 | 69.27 | 68.94 |
| CamemBERT-base | 71.93 | 69.72 | 70.81 |
| CamemBERT-large | 74.12 | 74.70 | 74.41 |
| CamemBERT-bio-base | 68.61 | 71.05 | 69.91 |
| CamemBERT-bio-large | 73.74 | 73.67 | 73.70 |
| Ensemble (t = 3) | 78.75 | 72.46 | 75.47 |
| **WNUT**       |     |     |      |
| BioBERT        | 78.45 | 72.66 | 75.44 |
| Bio + Clinical BERT* | 77.09 | 71.44 | 74.16 |
| PubMedBERT     | 79.12 | 73.70 | 76.32 |
| RoBERTa base   | 76.66 | 70.69 | 73.55 |
| RoBERTa large  | 77.57 | 71.75 | 74.55 |
| BioMed RoBERTa | 76.92 | 71.78 | 74.26 |
| XLNet          | 79.51 | 71.53 | 75.31 |
| Ensemble (t = 4) | 84.73 | 72.25 | 77.99 |

Bold in shows the best results.

4 RESULTS

4.1 Individual vs. Ensemble Models

Table 6 presents the NER results for the chemical, clinical, and web lab corpora obtained using the official blind evaluation set from the ChEMU, DEFT, and WNUT challenges, respectively. Individual model performance is compared with their respective ensembles for each corpus using the official challenge metrics. As the official test set of ChEMU is not yet publicly available, we also provide the performance of the ensemble and all its respective individual models on the official development set, taken as a blind test set, so that the chemical NER ensemble performance can be compared against all its individual models. In this case, the official training set was split into train and dev sets (as shown in Table 5). As we can notice, the ensemble models consistently outperform the individual models across the different domains and languages (English: ChEMU and WNUT; French: DEFT), with performance varying between 75.47 and 92.30% (considering only the official blind test evaluation). These results suggest that the ensemble strategy is a robust methodology for NER in the health and life science domains.

Considering each domain, the ensemble model on the chemical corpus outperforms its respective individual models, achieving 92.30% of the exact F1-score on the official blind test set and yielding 1.3 percentage point improvement over the BERT-base-cased baseline \((p = 0.005)\). The ensemble model on the clinical corpus achieves an F1-score of 75.47%, outperforming the multilingual BERT baseline by 6.5 percentage point \((p = 0.025)\). The best performance among the participating individual model in the clinical NER ensemble is achieved by CamemBERT-large, around 1.1 percentage point below the ensemble. The clinical NER has the worse performance among the different domains assessed. We believe it could be due to two factors. First, clinical corpora are notoriously complex, with many abbreviations and heterogeneous writing style, particularly compared to patents and protocols, in which documents are expected to follow a more formal structure and writing standard. Second, the clinical NER might suffer from the known problem of lack of resources for non-English languages (worsened in the case of clinical corpora). Similar to the other domains, the ensemble model on the web lab corpus outperforms its respective individual models \((p = 0.05)\), achieving an overall F1-score of 77.99%. Among the individual models, the best performance is achieved by the fine-tuned PubMedBERT followed by the fine-tuned BioBERT.

Interestingly, the best recall is achieved by individual models in all tasks assessed, though not consistently across individual models. We believe that by combining the individual models in the ensemble, we restrict the predictions, taking only the ones that are more likely true, having thus a significant positive impact on the precision with an eventual negative impact on the recall for some models. For the particular case of wet lab protocols, the best recall is achieved by the fine-tuned PubMedBERT. Among all the contextualized models, PubMedBERT is the only model trained on biomedical text from scratch, and consequently, it has a more specific vocabulary set (Gu et al., 2021), whereas the other models...
are first trained on the general text and then further pretrained on biomedical, medical, and clinical texts.

### 4.2 Comparison With State-of-the-Art

Table 7 shows the comparative results of our ensemble models against the teams participating in the ChEMU, DEFT, and WNUT competitions ranked by exact F1-score. The best results in the ChEMU competition were achieved by Wang et al. (2020), whose models were based on BioBERT fine-tuned on ChEMU data and BiLSTM-CRF. Their predictions were further post-processed using hand-written rules, a step missing in our pipeline. Dao and Nguyen (2020) used BiLSTM-CNN-CRF with Word2Vec and Elmo embeddings trained on patent data. Ruas et al. (2020) also used BioBERT fine-tuned on ChEMU data to extract the entities. BioBERT is trained on PubMed and PMC, and these datasets provide a better pretraining dataset for the chemical domain than the General book and Wikipedia datasets. The competition baseline model is presented as BANNER (He et al., 2020b). Our ensemble model presented as Ours outperforms the BANNER baseline by 3.37% in terms of exact F1-score.

On the clinical dataset, our ensemble model achieved the second place in terms of F1-score. The best performing model in this corpus relied on a BiLSTM-CRF model and features provided by contextualized embeddings of the CamemBERT model (Wajsbürt et al., 2020). Finally, our ensemble model on the wet lab dataset achieved the best performance among the participants in terms of F1-score. The next model was based on BiLSTM-CRF architecture and features provided by the contextualized word embeddings of PubMedBERT (Singh and Wadhawan, 2020). As we can notice, the addition of a BiLSTM-CRF layer also provides a consistently high-performing strategy in such domains.

### 4.3 Entity Type Performance

Table 8 shows the performance of our ensemble models for all classes in the chemical, clinical, and wet lab NER tasks. In the chemical NER, the performance of the ensemble model ranges between 87% for starting_material and 99.74% for yield_percent. Error analysis on the training data shows that the starting_material entity is often confused with the reagent_catalyst entity. From the chemistry point of view, both starting material (reactants) and catalysts (reagents) entities are present at the start of the reaction, with the difference that the latter is not altered by the reaction. These terms are often used interchangeably though, which could be the reason for the confusion.

In the clinical NER, the highest F1-score in the blind test set is achieved for the valeur entity (85.61%). This entity represents 7.6% of the annotations in the training collection. One could assume that entities with annotation examples above this threshold would perform well; however, when looking at the results for the substance (13.0% of the annotations) category, we notice an important drop in performance (63.79%). Thus, it seems that the number of training data examples alone is not sufficient to learn an entity automatically. The lowest performance for the ensemble method is found for the dose entity. 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 L), measure units (e.g., 1 mg/kg/j), or simply words that could be easily associated with a nonentity word (e.g., 24 paquets/année or 02).

The performance of the ensemble model for the classes of wet lab NER ranges between 30.72 and 95.31%. Surprisingly, the entity with the highest F1-score, pH (95.31%), has only 0.2% of the annotations in the training sample. Again, the number of examples is not associated with the performance of the test set.
Indeed, the best performing entities for the wet lab NER—Temperature, Time, and pH—are responsible together for only 8.5% of the annotation examples. The performance of the ensemble model is low for the Generic-Measure, which is similar to dose in clinical NER task, getting various forms, such as measure units (volume), measurements (30 kDa, 2.5 bars, ~250–500 bp), and ratios (1:2, 1/500 to 1/1,000), which could also justify its low score.

4.4 Entity Property Analyses
To better understand our results across the different corpora, we performed a deeper analysis of the reference individual baseline and the ensemble model using different entity properties: frequency, length, and label consistency. Figure 5 shows the comparison of the BERT baselines and the ensemble models based on the entity frequency and length. For both DEFT and WNUT collections, on average, the highest performance gains over the individual model happen for the less frequent entities, whereas the opposite happens for the ChEMU collection. Concerning the entity length property, we notice that the average length is shorter in the WNUT dataset. The ChEMU dataset, as expected, includes the longest average entity lengths, necessary to represent molecules. For all datasets, as the entity length increases, the performance of the ensemble models improves over the individual models. Finally, Figure 6 shows the frequency of passages that were assigned more than one label for the evaluated datasets. Here, we consider “passage” as a token or a sequence of tokens that
were assigned a label, for example, “triethylamine” annotated as `reagent_catalyst` and `other_compound` and “sodium hydrogen carbonate” annotated as `reagent_catalyst` and `other_compound` in ChEMU dataset. As more than one class is assigned to the same passage, we expect that they would be more ambiguous and therefore harder for the models to recognize. After the analyses of the training set, we notice that the ChEMU and WNUT corpora include passages that were assigned two or more labels for almost 10% of the examples. This happens for around only 1% of the annotations in the DEFT corpus. Hence, we would expect a better performance for the latter compared with the former. As it is not the case, it seems that the deep masked language models might actually be able to recognize those passages correctly using contextual information.

5 DISCUSSION

We compared the effectiveness of individual masked language models and ensemble models based on the majority vote strategy for the NER task in multiple health and life science domains and languages. The ensemble model showed a robust performance across the assessed domains and languages, achieving an overall macro F1-score of 76.94% and improving the individual models by 6.0 percentage point (considering the BERT-based-cased, BERT-base-multilingual-cased, and Bio + Clinical BERT as reference for the individual models in the ChEMU, DEFT, and WNUT datasets, respectively) ($p = 0.005$). Out of the 38 entity classes assessed, 50% had an F1-score equal or higher than 85% for the ensemble model (compare to 34% for the individual BERT model).

The performance of the models on the French clinical corpus is lower than on the chemical and wet lab corpora. We believe this is likely due to the known issue of reduced French language resources compared to English, both in terms of the corpora to pretrain the masked language models and also to fine-tune for the clinical NER. As seen in entity distribution tables (Tables 1, 2, and 3), the training data for chemical and wet lab NER are larger, which results in better performance for the individual language models and consequently for the ensemble models. Additionally, the clinical dataset includes nested entities, which are known to be recognized more effectively using graph-based models (Yu et al., 2020). Nevertheless, it is for the clinical dataset that we notice the highest relative gain in performance for the ensemble model (9.5% of F1-score).

Our analysis shows that specialized language models achieve the best performance across the health and life science domains. Moreover, in terms of model architecture, BiLSTM-CRF-based models with contextualized language models for feature extraction achieve competitive results. These results are aligned with the current knowledge available in the literature (Fu et al., 2020b; Hahn and Oleynik, 2020). That said, existing methods for chemical, clinical, and wet lab NER focus mostly on a single domain and language. Here, we introduced a novel and generic NER methodology for diverse and complex corpora in multiple domains and languages. We believe that such an approach can be expanded to other domains and languages with similar effectiveness.

The detailed analysis of entity types shows that the models have often difficulties recognizing infrequent entities, such as `dose` (clinical corpus) and `Generic-Measure` (wet lab corpus), which is in-line with previous work (Fu et al., 2020a). However, we notice that for some entities, particularly in the wet lab corpora, the highest scores were provided by infrequent entities. Indeed, as shown by Fu et al. (2020a), a single holistic measure of F1-score cannot tell the details of the performance of different models. Diverse entity attributes, such as `length`, `frequency`, `sentence length`, and `out-of-vocabulary (OOV) density`, are important for further model analyses. Thus, we further examined three meta-features: entity frequency and length, and label consistency. There was a consistent performance gain brought by the ensemble as the entity length increased. As deep masked language models work at the sub-word level, the longer the entity size, the more correct classifications are needed by the individual model to provide an exact match NER. By combining the different models, the ensemble seems to be able to leverage the correct classifications among the models. Moreover, despite a relatively frequent multi-labeling for passages (2 or more) in the chemical and wet lab corpora compared to the clinical corpus, their performance was significantly higher than the latter. This result suggests that, as expected, the deep masked language models were able to distinguish the homographs by their context.

The main limitation of our results comes from the heterogeneity of both corpora and models used. We used different baseline models across domains, partly due to the nature of the datasets (different languages). Additionally, the distribution of entities differs significantly across the datasets. All of this hinders the comparison of the results. Nevertheless, we believe the overall methodology gives a strong indication of the robustness of the ensemble of deep language models for NER in multi-domain and -lingual corpora.

6 CONCLUSION

In this work, we propose a generic and robust approach for named entity recognition in the health and life science domain based on deep masked language models combined in a majority voting strategy. We compared the performance of individual BERT models and their siblings against the proposed ensemble models for three types of corpora—chemical, clinical, and wet lab—available in English and French languages. We show a significant performance improvement of 6.0 percentage point ($p = 0.005$) using the ensemble models compared to a strong baseline based on individual BERT models, with the ensemble models having 50% of entities assessed with an F1-score of 85% or more. We further performed a detailed analysis of the performance of the models based on a set of entity properties.
We found that ensemble models can be more beneficial for longer entities.

**DATA AVAILABILITY STATEMENT**

Publicly available datasets were analyzed in this study. The data used for chemical NER can be found at: http://chem2020.eng.unimelb.edu.au/, https://deft.limsi.fr/2020/, http://noisy-text.github.io/2020/wlp-task.html.

**AUTHOR CONTRIBUTIONS**

NN drafted the manuscript, implemented the models, and analyzed the results. JK designed and implemented the models, and analyzed the results. JC implemented the models and analyzed the results. PR analyzed the results. DT drafted the manuscript and analyzed the results. All authors reviewed and contributed to the writing.

**REFERENCES**

Acharya, K. (2020). "WNUT 2020 Shared Task-1: Conditional Random Field(CRF) Based Named Entity Recognition(NER) for Wet Lab Protocols," in Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020) (Online: Association for Computational Linguistics), 286–289. doi:10.18653/v1/wnut-1.37

Akhondi, S. A., Klemner, A. G., Tyrchan, C., Manchala, A. K., Bobpana, K., Lowe, D., et al. (2014). Annotated Chemical Patent Corpus: A Gold Standard for Text Mining. PLoS ONE 9, e107477. doi:10.1371/journal.pone.0107477

Akhondi, S. A., Pons, E., Afzal, Z., van Haagen, H., Becker, B. F., Hettne, K. M., et al. (2016). Chemical Entity Recognition in Patents by Combining Dictionary-Based and Statistical Approaches. Database 2016, baw061. doi:10.1093/database/baw061

Alsentzer, J., Murphy, J., Boag, W., Weng, W., Jindi, D., Naumann, T., et al. (2019). "Publicly Available Clinical BERT Embeddings," in Proceedings of the 2nd Clinical Natural Language Processing Workshop Minneapolis, Minnesota, United States: Association for Computational Linguistics, 72–78.

Andrioli de Souza, J. V., Terumi Rubel Schneider, E., Oliveira Cezar, J., Silva e Oliveira, L. E., Boneski Gumiel, Y., Cabrera Paraiso, E., et al. (2020). "A Multilabel Approach to Portuguese Clinical Named Entity Recognition," in Proceedings of the XVII Congresso Brasileiro de Informática em Saúde (CBIS 2020), published in Journal of health informatics, 7-11 December 2020.

Beltagy, I., Lo, K., and Cohan, A. (2019). "SciBERT: A Pretrained Language Model for Scientific Text," in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 3606–3611. doi:10.18653/v1/d19-1371

Bethard, S., Derczynski, L., Savova, G., Pustejovsky, J., and Verhagen, M. (2015). "SemEval-2015 Task 6: Clinical TempEval," in proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), 806–814. doi:10.18653/v1/s15-2041

Cao, D., Benamar, A., Boumghar, M., Bethua, M., Ouadi, L. O., and Suignard, P. (2020). "Participation d’EDF R&D à DEFT 2020," in 6e conference conjointe Journées d’Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27é édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Atelier DÉF Fouille de Textes (ATALA: AFCP), 26–35.

Cardon, R., Grabar, N., Grouin, C., and 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 (Presentation of the DEFT 2020 Challenge: open domain textual similarity and precise information extraction from clinical cases)," in Actes de la 6e conference conjointe Journées d’Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Information pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Atelier DÉF Fouille de Textes (ATALA: AFCP), 26–35.

Copara, J., Knafou, J., Naderi, N., Moro, C., Ruch, P., and Teodoro, D. (2020a). "Contextualized French Language Models for Biomedical Named Entity Recognition," in 6e conference conjointe Journées d’Études sur la Parole (CLEF, 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. Editors R. Cardon, N. Grabar, C. Grouin, and T. Hamon (Nancy, France: ATALA), 36–48.

Corbett, P., and Boyle, J. (2018). Chemilsent: Chemical Named Entity Recognition Using Recurrent Neural Networks. J. Cheminf. 10, 1–9. doi:10.1186/s13321-018-0313-8

Dai, X., Karimi, S., Hachey, B., and Paris, C. (2019). "Using Similarity Measures to Select Pretraining Data for NER," in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Long and Short Papers. (Minneapolis, Minnesota: Association for Computational Linguistics), 1460–1470. doi:10.18653/v1/N19-1149

Dao, M. H., and Nguyen, D. Q. (2020). "VinAI at ChEMU 2020: An Accurate System for Named Entity Recognition in Chemical Reactions from Patents," in Empirically Grounded Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 3606–3611. doi:10.18653/v1/d19-1371

Devlin, J., Chang, M.-W., Lee, K., and 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. Long and Short Papers, 4171–4186.

Dinneen, H., Köksal, A., Özkirimli, E., and Orgür, A. (2020). "BOUN-REX at CLEF-2020 ChEMU Task 2: Evaluating Pretrained Transformers for Event Extraction," in CLEF (Working Notes).

El Boukkouri, H., Ferret, O., Lavergne, T., and Zweigenbaum, P. (2019). “Embedding Strategies for Specialized Domains: Application to Clinical Entity Recognition,” in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, 295–301. doi:10.18653/v1/p19-2041

Fu, J., Liu, P., and Neubig, G. (2020a). "Interpretable Multi-Dataset Evaluation for Named Entity Recognition," in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 6058–6069. doi:10.18653/v1/2020.emnlp-main.489

**ACKNOWLEDGMENTS**

The authors would like to thank the reviewers for their valuable comments and suggestions.
Zeng, Q., Fang, X., Liang, Z., and Meng, H. (2020). "Fancy Man Launches Zippo at WNUT 2020 Shared Task-1: A Bert Case Model for Wet Lab Entity Extraction," in Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020) (Online: Association for Computational Linguistics), 299–304. doi:10.18653/v1/2020.wnut-1.39

Zhai, Z., Nguyen, D. Q., Akhondi, S. A., Thorne, C., Druckenbrodt, C., Cohn, T., et al. (2019). Improving Chemical Named Entity Recognition in Patents with Contextualized Word Embeddings. BioNLP 2019, 328. doi:10.18653/v1/w19-5035

Zhang, Y., Xu, J., Chen, H., Wang, J., Wu, Y., Prakasam, M., et al. (2016). Chemical Named Entity Recognition in Patents by Domain Knowledge and Unsupervised Feature Learning. Database 2016, baw049. doi:10.1093/database/baw049

Zhao, S. (2004). "Named Entity Recognition in Biomedical Texts Using an Hmm Model," in Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA/BioNLP), 87–90. doi:10.3115/1567594.1567613

Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., et al. (2015). "Aligning Books and Movies: Towards story-like Visual Explanations by Watching Movies and reading Books," in Proceedings of the IEEE International Conference on Computer Vision (ICCV) (USA: IEEE Computer Society), 19–27. doi:10.1109/iccv.2015.11

Zhu, H., Paschalidis, I. C., and Tahmasebi, A. (2018). "Clinical Concept Extraction With Contextual Word Embedding," In NIPS Machine Learning for Health Workshop.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Naderi, Knafou, Copara, Ruch and Teodoro. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.