Ensemble of deep masked language models for effective named entity recognition in multi-domain corpora

Nona Naderi^{1,2*} Julien Knafou^{1,2,3} Jenny Copara^{1,2,3}
Patrick Ruch^{1,2} Douglas Teodoro^{1,2*}

¹University of Applied Sciences and Arts of Western Switzerland (HES-SO)

²Swiss Institute of Bioinformatics, Geneva, Switzerland

³University of Geneva, Switzerland

{firstname.lastname}@hesge.ch

Abstract

The health and life science domains are well known for their wealth of entities. These entities are presented as free text in large corpora, such as biomedical scientific and electronic health records. To enable the secondary use of these corpora and unlock their value, named entity recognition (NER) methods are proposed. Inspired by the success of deep masked language models, we present an ensemble approach for NER using these models. Results show statistically significant improvement of the ensemble models over baselines based on individual models in multiple domains - chemical, clinical and wet lab - and languages - English and French. The ensemble model achieves an overall performance of 79.2% macro F₁-score, a 4.6 percentage point increase upon the baseline in multiple domains and languages. These results suggest that ensembles are a more effective strategy for tackling NER. We further perform a detailed analysis of their performance based on a set of entity properties.

1 Introduction

In the health and life science domains, most of the information is encoded in free text reports. For example, it is estimated that around 90% of electronic health records (EHR) data is available as free text. While text format facilitates capturing information, it makes the secondary use of the data challenging. To support data structuring and unlock the value of textual databases in secondary usage applications, named entity recognition (NER) methods have been proposed. NER is

the task of detecting entities in text and assigning concept names, or categories, to them. The health and life science domains (e.g., biology, chemistry, medicine, etc.) are notoriously known for the wealth of named entities and synonyms, for example, microorganism taxonomies, drug brands, and gene names, to name a few. This richness of named entities together with the variety of formats and (mis)spellings make NER in health and life sciences corpora (e.g., EHR, lab protocols, scientific publications, patents, etc.) a challenging task.

Traditional NER methods mainly involve dictionary and rule-based, machine learning, and hybrid methods that combine these approaches [Quimbaya et al., 2016]. They normally require domain knowledge and feature engineering. Fully supervised NER approaches include Support Vector Machines (SVM), decision trees, hidden Markov models [Zhao, 2004], and conditional fields (CRFs) [Li et al., 2008, Leaman et al., 2015, Rocktäschel et al., 2012]. They are often used to provide a baseline for model evaluation.

More recently, deep masked language models trained on large corpus have achieved state-of-the-art in most NLP related tasks, including NER. Bidirectional Encoder Representations from Transformers (BERT) [Devlin et al., 2019] was the first to explore the transformer architecture as a general framework for NLP [Vaswani et al., 2017]. Once the model is trained (or pre-trained in the BERTology parlance) on a large corpus, it can be adapted and effectively used in specialised downstream NLP tasks, such as question-answering, text classification and NER, leveraging the feature representations learned by the model during the pre-training phase.

Since the advent of BERT, a myriad of transformer-based masked language models have been proposed [Liu et al., 2019, Yang et al., 2019, Alsentzer et al., 2019]. These models vary mostly in the tokenization used, how the masking is performed, and the trained data used during the pre-training phase. In this paper, we assess BERT-like models for NER in cross-domain (chemistry, clinical and wet lab) and -language (English and French) corpora. Particularly, we leverage different pre-trained models available in the literature 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., 2020a], DEFT (Défi Fouille de Textes) [Grabar et al., 2018] and WNUT (Workshop on Noisy User-generated Text) [Tabassum et al., 2020] challenges, respectively. Our results show that ensembles of named entity recognizers based on masked language models can achieve effective NER performances in these different domains and languages. We further perform an analysis of certain entity properties, including entity lengths, entity frequencies, and consistencies for better understanding the performance of these models.

2 Related work

Deep learning approaches trained on large unstructured data have shown considerable success in NLP problems, including NER [Devlin et al., 2019, Liu et al., 2019, Lample et al., 2016, Beltagy et al., 2019, Jin et al., 2019]. These models use the learned representations over the large data and reuse them in a supervised setting for a downstream task. For domain-specific tasks, the models that are trained on large general text can be further trained on domain specific large data and then adapted for a downstream task [Lee et al., 2019, Gururangan et al., 2020, Alsentzer et al., 2019] or the models can be trained only on domain-specific data and then adapted for a specific task [Beltagy et al., 2019]. Furthermore, several models are proposed for cross-domain NER; however, only a few focus on biomedical or clinical NER [Jia et al., 2019]. One study is that of [Lee et al., 2018] that utilizes the idea of transfer learning to identify the named entities in i2b2 2014/2016 using the MIMIC dataset.

2.1 Chemical NER

To further improve the performance of the traditional approaches using hand-crafted features [Leaman et al., 2015, Rocktäschel et al., 2012, Habibi et al., 2016, Zhang et al., 2016, Akhondi et al., 2016], a number of studies leveraged the power of word embeddings in addition to the hand-crafted features in a single model (LSTM-CRF) [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.

2.2 Clinical NER

Various NER challenges and shared tasks [Uzuner et al., 2010, Kelly et al., 2014, Névél et al., 2015, Suominen et al., 2013, Bethard et al., 2015] fostered the development of NER methods [Van Mulligen et al., 2016, Kim et al., 2015, Jiang et al., 2011, De Bruijn et al., 2011, El Boukkouri et al., 2019] for clinical domain in different languages [Lopes et al., 2019, Schneider et al., 2020, Sun and Yang, 2019]. Performance varies greatly across the different methods and corpora, with more modern methods achieving F_1 -score as high as 95%.

2.3 Wet lab NER

Luan et al. [2019] introduce a model based on dynamic span graph to jointly extract named entities and relations on wet lab protocols and other corpora. Wadden et al. [2019] build upon Luan et al. [2019]’s model by combining BERT and dynamic

span graph. Dai et al. [2019] compute the similarity of the pre-train data and the data of the target application to investigate the effectiveness of pre-trained word vectors. They found that the effectiveness of the pre-trained word vectors depends on the vocabulary overlap of the source and target domains.

3 Data

In this section, we present the datasets used to assess the 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 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 Benchmark data for chemical entity recognition - ChEMU 2020 dataset

The ChEMU 2020 benchmark dataset¹ [He et al., 2021] contains snippets sampled from 170 English patents from the European Patent Office and the United States Patent and Trademark Office [He et al., 2020a,b, 2021]. As shown in Fig. 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 1500 snippets (training: 1,125; test: 375) annotated with 26,857 entities (training: 20,186; test: 6,671) using the BRAT standoff format [Stenetorp et al., 2012].

Table 1 shows the entity distribution for the training and test sets. The majority of the annotations are provided for the *other_compound*, *reaction_product* and *starting_material* entities, covering 52% of the examples in the training and test datasets. In contrast, *example_label*, *yield_other* and *yield_percent* entities represent together only 18% of entities in the training and test sets.

3.2 Benchmark data for clinical entity recognition - DEFT 2020 dataset

The DEFT benchmark dataset², a subset of the CAS corpus [Grabar et al., 2018], is composed of 100 French clinical documents (training: 90; test: 10) manually

¹<https://data.mendeley.com/datasets/wy6745bjfj/2>

²<https://jep-taln2020.loria.fr/ateliers/deft/>

Example 15A

1-(2-Methoxyethyl)-5-methyl-2,4-dioxo-3-(2-phenylethyl)-1,2,3,4-tetrahydrothieno[2,3-d]pyrimidine-6-carboxylic acid

75 ml of trifluoroacetic acid were added to a solution of 5.0 g (11.2 mmol) of the compound from Ex. 10A in 225 ml of dichloromethane, and the mixture was stirred at RT for 2 h. The reaction mixture was then concentrated to dryness on a rotary evaporator. The remaining residue was stirred in diethyl ether and filtered off with suction, and the solid was dried under high vacuum. 4.1 g (92% of theory) of the title compound were obtained.

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

Entity	Train		Test	
	Count	%	Count	%
example_label	1,104	5.5	349	5.2
other_compound	5,720	28.3	1931	28.9
reaction_product	2,558	12.7	855	12.8
reagent_catalyst	1,570	7.8	504	7.6
solvent	1,390	6.9	428	6.4
starting_material	2,167	10.7	711	10.7
temperature	1,861	9.2	612	9.2
time	1,311	6.5	452	6.8
yield_other	1,322	6.5	440	6.6
yield_percent	1,183	5.9	389	5.8
Total	20,186	100.0	6,671	100.0

Table 1: Entity distribution in the training and test sets of ChEMU benchmark dataset.

Figure 2: 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.

annotated with the 8,098 entities (training: 7,421; test: 677) in the following categories: *pathologie*, *soso* (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 often 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 *soso*, *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 vary significantly.

3.3 Benchmark data for wet lab entity recognition - WNUT 2020 dataset

The WNUT benchmark dataset [Kulkarni et al., 2018] is composed of 727 unique (English) wet lab protocols³ that describe experimental procedures. This dataset (training: 616; test: 111) was manually annotated with the 102,957 entities (training: 79,757; test: 23,200) 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 in the other chemical and clinical datasets, there is a significant class imbalance, with only two of them (*Action* and *Reagent*) representing more than 50% of annotations in the training set. This table also shows that, similar to the ChEMU dataset, the

³<https://www.protocols.io/>

Entity	Train		Test	
	Count	%	Count	%
anatomie	1,298	17.5	174	25.7
dose	342	4.6	5	0.7
examen	1,081	14.6	137	20.2
mode	238	3.2	11	1.6
moment	440	5.9	54	8.0
pathologie	351	4.7	184	27.2
sosy	1,647	22.2	33	4.9
substance	968	13.0	22	3.3
traitement	494	6.7	52	7.7
valeur	562	7.6	5	0.7
Total	7,421	100.0	677	100.0

Table 2: Entity distribution in the training and test sets of the DEFT benchmark dataset.

Figure 3: An example of a wet lab protocol of the WNUT dataset with entity annotations. The annotations are color coded, representing the different entities in the dataset.

Entity	Train		Test	
	Count	%	Count	%
Action	20,504	25.7	5,346	23.0
Amount	5,712	7.2	1,223	5.3
Concentration	2,287	2.9	701	3.0
Device	2,836	3.6	888	3.8
Generic-Measure	759	1.0	173	0.8
Location	6,643	8.3	1,657	7.1
Measure-Type	1,453	1.8	720	3.1
Mention	396	0.5	142	0.6
Method	2,716	3.4	1,059	4.6
Modifier	7,736	9.7	3,416	14.7
Numerical	1,322	1.7	513	2.2
Reagent	18,710	23.5	5,012	21.6
Seal	366	0.5	119	0.5
Size	498	0.6	232	1.0
Speed	1,032	1.3	238	1.0
Temperature	2,610	3.3	744	3.2
Time	4,011	5.0	951	4.1
pH	166	0.2	66	0.3
Total	79,757	100.0	23,200	100.0

Table 3: Entity distribution in the training and test sets of the WNUT benchmark dataset.

proportions of entities are fairly similar across the training and test subsets.

4 Method

In this section, we describe our 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.

4.1 Single deep masked language model for NER

To build the ensemble NER model, we fine-tuned different individual masked languages models based on the transformers architecture [Vaswani et al., 2017]. In the case of NER, masked language models are fine-tuned using a specialised training

set - in our case, the chemical, clinical and wet lab annotated corpora - to classify tokens according to the named entity classes. As shown in Table 4, we assessed single language models based on or derived from the BERT architecture. BERT was originally pretrained on a large corpus of English text extracted from Book-Corpus [Zhu et al., 2015] and Wikipedia, with different number of attention heads for the base and large types (12 and 24 transformer layers and hidden representations of 768 and 1024 dimensions, respectively).

To fine-tune a particular masked language model for the NER task, we leverage the token representation created in its pre-training 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 highest probable entity label 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 respective 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. More information about the fine-tuning of the models and the hyper-parameter settings can be found in [Copara et al., 2020b,a, Knafo et al., 2020].

4.2 Ensemble of deep masked language models for NER

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, Knafo et al., 2020]. In other words, for a given document, our models infer their predictions independently for each entity. 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, in case of nested entities. The individual

Dataset	Pretrained model	Model size	Corpus type
ChEMU	BERT [Devlin et al., 2019]	base	general
	CNN	large	
DEFT	BERT-base-multilingual-cased	base	general
	CamemBERT [Martin et al., 2020]	base	
	CamemBERT-bio [Copara et al., 2020a]	large	bio + medical
WNUT	RoBERTa [Liu et al., 2019]	base	general
	XLNet [Yang et al., 2019]	large	general
	BioBERT [Lee et al., 2019]	large	bio
	Bio+ClinicalBERT [Alsentzer et al., 2019]		bio + clinical
	PubMedBERT [Gu et al., 2020]		bio + medical
	BioMed RoBERTa [Gururangan et al., 2020]		bio + medical

Table 4: Pretrained models used for NER in the ChEMU, DEFT, and WNUT benchmark datasets.

models used in the ensemble models for the ChEMU, DEFT, and WNUT datasets are presented in Table 4.

4.3 Training and evaluation procedures

To train our models, we used the training subsets of the ChEMU, DEFT and WNUT. As shown in Table 5, we split this subset into train, dev and test sets to train the model weights, the hyper-parameters, and the best ensemble configuration, respectively. The ensemble threshold for chemical and clinical NER were set to 3 and, for wet lab NER, to 4. Then, a blind test set, which was provided as part of the official evaluation for the respective challenges, was used to evaluate the models. Table 5 shows the distribution of the splits for the different collections.

Results are reported in terms of the the micro and macro F_1 -score metrics and were computed using the BRAT eval tool⁴ against the blind test set split. The ensemble models created for the different domains are compared to a baseline based on BERT. 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.

⁴https://bitbucket.org/nicta_biomed/brateval/

Dataset	Split	# patent snippets	# EHR notes	# wet lab protocols
Training	Train	800	80	370
Training	Dev	100	10	123
Training	Test	225	10	123
Test	Blind test	375	67	111

Table 5: Distribution of samples in the train, dev and test collections for the different NER tasks. Train: collection used to train model parameters. Dev: collection used to tune model hyperparameters. Test: collection used to define the ensemble models. Blind test: collection used to evaluate models.

5 Results

In this section, we present the performance assessment of BERT as a NER baseline and the ensemble model over the test collection with the parameters identified in the training phase for the chemical, clinical and wet lab corpora.

5.1 Chemical NER results

Table 6 shows the performance of the baseline BERT and our ensemble models for the chemical NER using the F_1 -score metric in terms of exact and relaxed span matching. In the exact match evaluation, both the starting and the end offsets of the text spans of the predicted and gold standard reference entities match, whereas in the relaxed-match evaluation, the text spans of two entities overlap. The ensemble model achieves 92.30% of exact micro F_1 -score, yielding 1.3 percentage point improvement over the baseline (BERT-base-cased) ($p = 0.005$). It outperforms the BERT model for all the entities in exact match. If we consider the relaxed metric, the difference between the individual and ensemble model is minimal and not statistically significant ($p = 0.07$). These results indicate that the individual model is able to detect the passages containing the entities (or part of them) while the ensemble, by combining the power of different models, is able to predict the exact spans of the entities.

The top-5 best performing entities identified by our models are *example_label*, *temperature*, *time*, *yield_other*, *yield_percent*. The entity with the lowest performance is *starting_material*, achieving 84.13% and 87.01% of exact F_1 -score for the baseline and ensemble models, respectively. Overall, the language models were effective to recognised chemical entities in patents, with a lower limit performance of 90.98% of micro F_1 -score for the single BERT model and as high as 96.56% of macro F_1 -score for the ensemble model.

Entity	BERT		Ensemble	
	exact	relaxed	exact	relaxed
example_label	96.17	97.30	96.69	97.84
other_compound	87.80	96.08	89.20	96.53
reaction_product	85.93	93.78	87.66	93.22
reagent_catalyst	87.91	90.82	90.22	91.76
solvent	94.44	94.91	95.41	95.41
starting_material	84.13	93.43	87.01	93.94
temperature	96.92	99.02	97.29	98.77
time	98.68	99.67	98.79	99.78
yield_other	97.99	98.21	98.42	98.65
yield_percent	99.36	99.62	99.74	99.74
all (micro)	90.98	95.96	92.30	96.24
all (macro)	92.93	96.28	94.04	96.56

Table 6: Test phase results using the F_1 -score metric for the chemical patents NER (ChEMU dataset). BERT: BERT base-cased model. Ensemble threshold set to 3.

5.2 Clinical NER results

The performance of BERT and the ensemble model for the clinical NER is summarised in Table 7. The ensemble model achieves 72.62% of overall micro F_1 -score, outperforming the baseline model by 8.82 percentage point ($p < 0.001$). Similar to the results obtained in the training set, the highest F_1 -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 *examen* (14.6% of the annotations) and *substance* (13.0% of the annotations) categories, we notice an important drop in performance (64.42% and 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 *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 litre*), measure units (e.g., *1mg/kg/j*) or simply words that could be easily associated with a non-entity word (e.g., *24 paquets/année* or *02*). *Mode* entities are mostly words without abbreviations or numbers (e.g., ‘*voie parasternale droite*’, ‘*voie centrale intraveineuse*’). Hence, they contain less variety for their values, which could lead to an easier way for the NER models to learn their patterns and make correct predictions. Nevertheless, results for the *Mode* categories are close to the median (64.86%). The lack of examples in the training set

Entity	BERT	Ensemble
anatomie	76.46	80.69
dose	36.04	52.17
examen	68.42	73.33
mode	59.35	64.86
moment	67.48	78.69
pathologie	36.28	56.44
sosy	55.74	67.33
substance	55.86	63.79
traitement	45.98	60.76
valeur	81.60	85.61
all (micro)	63.80	72.62
all (macro)	58.32	68.37

Table 7: Test phase results using the F_1 -score metric for the clinical notes NER (DEFT dataset). BERT: BERT base-multilingual-cased model. Ensemble threshold set to 3.

(3.2% of the annotations) could have impacted its performance negatively.

5.3 Wet lab NER results

The performance of BERT and the ensemble model for the wet lab NER is summarised in Table 8. As for the other domains, the ensemble model outperform BERT for all entities ($p = 0.003$), achieving an overall micro F_1 -score of 81.67%. The gain in performance is more relevant for the macro F_1 -score metrics, for which there is an increase of 3.28 percentage point between the baseline model and the ensemble model. In this case, it seems that the diversity brought by the ensemble enables the correct detection of more entity types.

Surprisingly, the entity with the highest F_1 -score, *pH* (95.31%), has only 0.2% of the annotations in the training sample. Again, the number of examples is not associate with the performance on 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 lowest performance was found for the *Size* entity (32.03%), more than 2-folds worse than the average wet lab entity (77.99% for the ensemble model). The performance of the models are also low for the *Generic-Measure* and *Numerical* entities. Both have a small number of annotations in the training set (1.0% and 1.7% respectively). The *Generic-Measure* entity is similar to *dose* in clinical NER task and get various forms, such as measure units (*volume*), measurements (*30 kDa*, *2.5 bars*, *~250–500 bp*), and ratios (*1:2*,

Entity	BERT	Ensemble
Action	87.35	88.35
Amount	84.47	86.76
Concentration	85.45	90.34
Device	69.66	71.46
Generic-Measure	40.00	50.51
Location	78.92	80.49
Measure-Type	68.61	68.07
Mention	70.10	73.09
Method	60.05	57.95
Modifier	55.82	58.11
Numerical	51.03	50.12
Reagent	89.19	92.03
Seal	86.84	90.13
Size	28.84	32.03
Speed	89.21	89.77
Temperature	90.12	94.15
Time	92.64	94.28
pH	94.74	95.31
all (micro)	79.34	81.67
all (macro)	74.71	77.99

Table 8: Test phase results using the F_1 -score metric for the wet lab protocols NER (WNUT dataset). BERT: Bio+Clinical BERT. Ensemble threshold set to 4.

1/500 to 1/1000), which could also justify its low score.

6 Discussion

We compared the effectiveness of individual masked deep language models and ensemble models for the NER task for multiple domains and languages. Overall, the ensemble model improves the baseline BERT model by 4.6 percentage point, achieving an overall macro F_1 -score of 79.20% across all entities in the chemical, clinical and wet lab corpora. The ensemble model was relatively effective to recognised named entities in those corpora. Out of the 38 entity classes assessed, 45% had an F_1 -score higher 90% for the ensemble model compare to 24% for the individual BERT model. However, we notice an important performance reduction in the French corpus compared to the English corpora. This is likely due to the known issue of reduced resources compared to English, both in terms of the corpora to pre-train the masked language models but also to fine-tune for the clinical NER.

Figure 4 shows the comparison of BERT and the ensemble model for each dataset. The performance of the models on the clinical corpus is lower than on the chemical and wet lab corpora. This can be due to a smaller dataset available for fine-tuning the models. As seen in entity distribution tables (Tables 1, 2 and 3), the

Figure 4: The performance of the ensemble model vs. BERT for different domains. Micro F_1 -score is used. The exact match is used for ChEMU dataset.

training data for chemical NER and wet lab NER are larger, which results in better performance of both the ensemble models and the BERT baseline. 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 (14%) for the ensemble model.

Our results show that the models have often difficulties recognizing infrequent entities, such as *dose* (clinical) and *Generic-Measure* (wet lab), which is inline with previous work [Fu et al., 2020]. However, we notice that for some entities, particularly in the wet lab corpora, the highest score were provided by infrequent entities. Indeed, as shown by Fu et al. [2020], a single holistic measure of F_1 -score cannot tell the details of 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.

To better understand our results across the different corpora, we performed a deeper analysis of the baseline and ensemble models using different entity properties: frequency, length, and label consistency. Figure 5 shows the comparison of the BERT baseline and ensemble models based on the frequency of the entities. For some infrequent entities, the ensemble model improves more over the BERT baseline. However, we also see performance gain for the ensemble model for frequent entities, such as *starting_material* (performance improvement of +2.9 point using the ensemble model). *starting_material* covers about 11% of ChEMU dataset and

Figure 5: Performance of the BERT model vs. the ensemble model based on the entity frequency on the training data. For the clinical NER, the exact match scores are used.

is more frequent than entities, such as *time* and *yield_percent*, for which the performance gains of the ensemble model are +0.11 point and +0.38 point, respectively.

Figure 6 shows the comparison of BERT and the ensemble model based on the entity length. The average entity length is shorter in the WNUT dataset. The ChEMU dataset, as expected, includes the longest average entity lengths. In general in all datasets, as the entity length increases, the performance of the ensemble model improves over BERT. In the ChEMU data, there is a direct link between the length and the performance of the models: the shorter the entity, the higher the performance.

Finally, Figure 7 shows the frequency of passages that were assigned more than one label for the three 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, each was considered as one instance. The ChEMU and WNUT annotated corpora include passages that were assigned more than 2 labels, which makes them more ambiguous for the models. This can also explain the more modest performance of the models on WNUT dataset compared to the ChEMU dataset despite its larger data size.

Figure 6: Performance of the BERT model vs. the ensemble model based on the entity length on the training data. For the clinical NER, the exact match scores are used.

Figure 7: The number of labels assigned to each passage for the training set of the three datasets (ChEMU, DEFT, WNUT).

7 Conclusion

In this work, we compared the performance of the BERT model and its siblings for name entity recognition for three domains of chemical, clinical and wet lab, and English, and French languages. In all the domains and languages that we analyzed for NER, we show a significant improvement of performance by combining different deep masked language models compared to a strong baseline based on BERT single models. Overall, the ensemble model outperformed the baseline by 4.6 percentage point ($p < 0.001$), having 45% of the 38 entities assessed across the domains with a F_1 -score of 90% 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.

Conflict of Interest Statement

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.

Author Contributions

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

Funding

Funding for this work is provided by the CINECA project (H2020 No 825775).

Data Availability Statement

The data used for chemical NER can be found at: <http://chemu2020.eng.unimelb.edu.au/>. The data used for clinical NER can be found at: <https://deft.limsi.fr/2020/>. The data used for wet lab NER can be found at: <http://noisy-text.github.io/2020/wlp-task.html>.

References

- Saber A. Akhondi, Ewoud Pons, Zubair Afzal, Herman van Haagen, Benedikt F.H. Becker, Kristina M. Hettne, Erik M. van Mulligen, and Jan A. Kors. Chemical entity recognition in patents by combining dictionary-based and statistical approaches. *Database*, 2016, 2016.
- Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. Publicly Available Clinical BERT Embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, 2019.
- Iz Beltagy, Kyle Lo, and Arman Cohan. 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)*, pages 3606–3611, 2019.
- Steven Bethard, Leon Derczynski, Guergana Savova, James Pustejovsky, and Marc Verhagen. Semeval-2015 task 6: Clinical tempeval. In *proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 806–814, 2015.
- Jenny Copara, Julien Knafou, Nona Naderi, Claudia Moro, Patrick Ruch, and Douglas Teodoro. Contextualized French Language Models for Biomedical Named Entity Recognition. In Rémi Cardon, Natalia Grabar, Cyril Grouin, and Thierry Hamon, editors, *6e conférence conjointe Journées d’Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Atelier DÉfi Fouille de Textes*, pages 36–48, Nancy, France, 2020a. ATALA.
- Jenny Copara, Nona Naderi, Julien Knafou, Patrick Ruch, and Douglas Teodoro. Named entity recognition in chemical patents using ensemble of contextual language models. In *Working notes of the CLEF 2020*, number CONFERENCE. 22-25 September 2020, 2020b.
- Peter Corbett and John Boyle. Chemlistem: chemical named entity recognition using recurrent neural networks. *Journal of cheminformatics*, 10(1):1–9, 2018.
- Xiang Dai, Sarvnaz Karimi, Ben Hachey, and Cecile Paris. 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, Volume 1 (Long and Short*

Papers), pages 1460–1470, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1149. URL <https://www.aclweb.org/anthology/N19-1149>.

Berry De Bruijn, Colin Cherry, Svetlana Kiritchenko, Joel Martin, and Xiaodan Zhu. Machine-learned solutions for three stages of clinical information extraction: the state of the art at i2b2 2010. *Journal of the American Medical Informatics Association*, 18(5):557–562, 2011.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, 2019.

Hicham El Boukkouri, Olivier Ferret, Thomas Lavergne, and Pierre Zweigenbaum. 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*, pages 295–301, 2019.

Jinlan Fu, Pengfei Liu, and Graham Neubig. Interpretable multi-dataset evaluation for named entity recognition. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6058–6069, 2020.

Natalia Grabar, Vincent Claveau, and Clément Dalloux. CAS: French corpus with clinical cases. In *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis*, pages 122–128, Brussels, Belgium, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5614. URL <https://www.aclweb.org/anthology/W18-5614>.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing, 2020.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. Don’t stop pretraining: Adapt language models to domains and tasks. In *Proceedings of ACL*, 2020.

Maryam Habibi, David Luis Wiegandt, Florian Schmedding, and Ulf Leser. Recognizing chemicals in patents: a comparative analysis. *Journal of cheminformatics*, 8(1):1–15, 2016.

Maryam Habibi, Leon Weber, Mariana Neves, David Luis Wiegandt, and Ulf Leser. Deep learning with word embeddings improves biomedical named entity recognition. *Bioinformatics*, 33(14):i37–i48, 2017.

Jiayuan He, Dat Quoc Nguyen, Saber A. Akhondi, Christian Druckenbrodt, Camilo Thorne, Ralph Hoessel, Zubair Afzal, Zenan Zhai, Biaoyan Fang, Hiyori Yoshikawa, Ameer Albahem, Lawrence Cavedon, Trevor Cohn, Timothy Baldwin, and Karin Verspoor. Overview of chemu 2020: Named entity recognition and event extraction of chemical reactions from patents. In Avi Arampatzis, Evangelos Kanoulas, Theodora Tsikrika, Stefanos Vrochidis, Hideo Joho, Christina Lioma, Carsten Eickhoff, Aurélie Névél, Linda Cappellato, and Nicola Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Eleventh International Conference of the CLEF Association (CLEF 2020)*, volume 12260. Lecture Notes in Computer Science, 2020a.

Jiayuan He, Dat Quoc Nguyen, Saber A. Akhondi, Christian Druckenbrodt, Camilo Thorne, Ralph Hoessel, Zubair Afzal, Zenan Zhai, Biaoyan Fang, Hiyori Yoshikawa, Ameer Albahem, Jingqi Wang, Yuankai Ren, Zhi Zhang, Yaoyun Zhang, Mai Hoang Dao, Pedro Ruas, Andre Lamurias, Francisco M. Couto, Jenny Copara, Nona Naderi, Julien Knafo, Patrick Ruch, Douglas Teodoro, Daniel Lowe, John Mayfield, Abdullatif Köksal, Hilal Dönmez, Elif Özkırmılı, Arzucan Özgür, Darshini Mahendran, Gabrielle Gurdin, Nastassja Lewinski, Christina Tang, Malarkodi T.McInnes, Bridget C.S., Pattabhi RK Rao., Sobha Lalitha Devi, Lawrence Cavedon, Trevor Cohn, Timothy Baldwin, and Karin Verspoor. An extended overview of the clef 2020 chemu lab: Information extraction of chemical reactions from patents. In *Proceedings of the Eleventh International Conference of the CLEF Association (CLEF 2020)*. 2020b.

Jiayuan He, Dat Quoc Nguyen, Saber A. Akhondi, Christian Druckenbrodt, Camilo Thorne, Ralph Hoessel, Zubair Afzal, Zenan Zhai, Biaoyan Fang, Hiyori Yoshikawa, Ameer Albahem, Lawrence Cavedon, Trevor Cohn, Timothy Baldwin, and Karin Verspoor. Chemu 2020: Natural language processing methods are effective for information extraction from chemical patents. *Frontiers in Research Metrics and Analytics*, 6:12, 2021. ISSN 2504-0537. doi: 10.3389/frma.2021.654438.

Wahed Hemati and Alexander Mehler. Lstm voter: chemical named entity recognition using a conglomerate of sequence labeling tools. *Journal of cheminformatics*, 11(1):1–7, 2019.

- Chen Jia, Xiaobo Liang, and Yue Zhang. Cross-domain ner using cross-domain language modeling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2464–2474, 2019.
- Min Jiang, Yukun Chen, Mei Liu, S Trent Rosenbloom, Subramani Mani, Joshua C Denny, and Hua Xu. A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries. *Journal of the American Medical Informatics Association*, 18(5):601–606, 2011.
- Qiao Jin, Bhuwan Dhingra, William Cohen, and Xinghua Lu. Probing biomedical embeddings from language models. In *Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP*, pages 82–89, 2019.
- Liadh Kelly, Lorraine Goeuriot, Hanna Suominen, Tobias Schreck, GONDY Leroy, Danielle L Mowery, Sumithra Velupillai, Wendy W Chapman, David Martinez, Guido Zuccon, et al. Overview of the share/clef ehealth evaluation lab 2014. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 172–191. Springer, 2014.
- Youngjun Kim, Ellen Riloff, and John F Hurdle. A study of concept extraction across different types of clinical notes. In *AMIA Annual Symposium Proceedings*, volume 2015, page 737. American Medical Informatics Association, 2015.
- Julien Knafo, Nona Naderi, Jenny Copara, Douglas Teodoro, and Patrick Ruch. Bitem at wnut 2020 shared task-1: Named entity recognition over wet lab protocols using an ensemble of contextual language models. In *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, Online, 2020. Association for Computational Linguistics, Association for Computational Linguistics. URL <https://www.aclweb.org/anthology/2020.wnut-1.40>.
- Chaitanya Kulkarni, Wei Xu, Alan Ritter, and Raghu Machiraju. An annotated corpus for machine reading of instructions in wet lab protocols. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 97–106, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.
- Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, 2016.

- Robert Leaman, Chih-Hsuan Wei, and Zhiyong Lu. tmchem: a high performance approach for chemical named entity recognition and normalization. *Journal of cheminformatics*, 7(1):1–10, 2015.
- Ji Young Lee, Franck Dernoncourt, and Peter Szolovits. Transfer learning for named-entity recognition with neural networks. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.
- Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 09 2019.
- Dingcheng Li, Guergana Savova, and Karin Kipper. Conditional random fields and support vector machines for disorder named entity recognition in clinical texts. In *Proceedings of the workshop on current trends in biomedical natural language processing*, pages 94–95, 2008.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019.
- Fábio Lopes, César Teixeira, and Hugo Gonçalo Oliveira. Contributions to clinical named entity recognition in portuguese. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 223–233, 2019.
- Yi Luan, Dave Wadden, Luheng He, Amy Shah, Mari Ostendorf, and Hannaneh Hajishirzi. A general framework for information extraction using dynamic span graphs. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3036–3046, 2019.
- Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de La Clergerie, and Benoît Sagot. CamemBERT: a Tasty French Language Model. In *The 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*, Seattle, Washington, United States, 2020.
- Aurélié Névél, Cyril Grouin, Xavier Tannier, Thierry Hamon, Liadh Kelly, Lorraine Goeuriot, and Pierre Zweigenbaum. Clef ehealth evaluation lab 2015 task 1b: Clinical named entity recognition. In *CLEF (Working Notes)*, 2015.

- Alexandra Pomares Quimbaya, Alejandro Sierra Múnera, Rafael Andrés González Rivera, Julián Camilo Daza Rodríguez, Oscar Mauricio Muñoz Velandia, Angel Alberto Garcia Peña, and Cyril Labbé. Named entity recognition over electronic health records through a combined dictionary-based approach. *Procedia Computer Science*, 100:55–61, 2016.
- Tim Rocktäschel, Michael Weidlich, and Ulf Leser. Chemsport: a hybrid system for chemical named entity recognition. *Bioinformatics*, 28(12):1633–1640, 2012.
- Elisa Terumi Rubel Schneider, Joao Vitor Andrioli de Souza, Julien Knafou, Lucas Emanuel Silva e Oliveira, Jenny Copara, Yohan Bonescki Gumiel, Lucas Ferro Antunes de Oliveira, Emerson Cabrera Paraiso, Douglas Teodoro, and Cláudia Maria Cabral Moro Barra. Biobertpt-a portuguese neural language model for clinical named entity recognition. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 65–72, 2020.
- Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun’ichi Tsujii. Brat: a web-based tool for nlp-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107, 2012.
- Cong Sun and Zhihao Yang. Transfer learning in biomedical named entity recognition: An evaluation of bert in the pharmaconer task. In *Proceedings of The 5th Workshop on BioNLP Open Shared Tasks*, pages 100–104, 2019.
- Hanna Suominen, Sanna Salanterä, Sumithra Velupillai, Wendy W Chapman, Guergana Savova, Noemie Elhadad, Sameer Pradhan, Brett R South, Danielle L Mowery, Gareth JF Jones, et al. Overview of the share/clef ehealth evaluation lab 2013. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 212–231. Springer, 2013.
- Jeniya Tabassum, Sydney Lee, Wei Xu, and Alan Ritter. WNUT-2020 Task 1 Overview: Extracting Entities and Relations from Wet Lab Protocols. In *Proceedings of EMNLP 2020 Workshop on Noisy User-generated Text (WNUT)*, 2020.
- Özlem Uzuner, Imre Solti, and Eithon Cadag. Extracting medication information from clinical text. *Journal of the American Medical Informatics Association*, 17(5):514–518, 2010.
- Erik M Van Mulligen, Zubair Afzal, Saber Akhondi, Dang Vo, and Jan Kors. Erasmus mc at clef ehealth 2016: Concept recognition and coding in french texts. 2016.

- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6000–6010, 2017.
- David Wadden, Ulme Wennberg, Yi Luan, and Hannaneh Hajishirzi. Entity, relation, and event extraction with contextualized span representations. 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)*, pages 5788–5793, 2019.
- Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. XLNet: Generalized Autoregressive Pretraining for Language Understanding. In *Advances in neural information processing systems*, pages 5753–5763, 2019.
- Juntao Yu, Bernd Bohnet, and Massimo Poesio. Named entity recognition as dependency parsing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6470–6476, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.577. URL <https://www.aclweb.org/anthology/2020.acl-main.577>.
- Zenan Zhai, Dat Quoc Nguyen, Saber A Akhondi, Camilo Thorne, Christian Druckenbrodt, Trevor Cohn, Michelle Gregory, and Karin Verspoor. Improving chemical named entity recognition in patents with contextualized word embeddings. *BioNLP 2019*, page 328, 2019.
- Yaoyun Zhang, Jun Xu, Hui Chen, Jingqi Wang, Yonghui Wu, Manu Prakasam, and Hua Xu. Chemical named entity recognition in patents by domain knowledge and unsupervised feature learning. *Database*, 2016, 2016.
- Shaojun Zhao. 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)*, pages 87–90, 2004.
- Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 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)*, page 19–27, USA, 2015. IEEE Computer Society.