EneBERT: A State-of-the-art Language Model Trained on a Corpus of Texts Generated from the Set of DSO Activities

Eunice Akani, Romain Gemignani, Rim Abrougui

ENEBERT: A STATE-OF-THE-ART LANGUAGE MODEL TRAINED ON A CORPUS OF TEXTS GENERATED FROM THE SET OF DSO ACTIVITIES

Eunice AKANI
Enedis, Aix-Marseille Univ – France
adueni-adjoba-eunice.akani@enedis.fr

Romain GEMIGNANI
Enedis – France
romain.gemignani@enedis.fr

Rim ABROUGUI[1]
Aix-Marseille Univ – France
r.abrougui@gmail.com

ABSTRACT

Pre-trained language model based on the Transformers architecture has significantly advanced natural language processing. This allowed for better results on several tasks, such as text classification or named entity recognition. Most existing models are general, and their use in a particular domain is sometimes tedious. Indeed, some domains have terms that are outside of commonly written languages. In these conditions, the implementation of a domain-specific model is necessary.
In this paper, we have introduced a language model for the DSO activities of Enedis. This model is based on state-of-the-art language models but has been trained on domain-specific data. We also show that using the model for specific tasks, such as text classification, increases the performance obtained compared to a general language model.

INTRODUCTION

Natural language processing has seen a spectacular advance thanks to the arrival of pre-trained language models based on the Transformers architecture [1], like BERT [2] and CamemBERT [3]. Building a language model means using statistical and probabilistic techniques to obtain the probability of a sequence of words occurring in a sentence. It allows having a base of word predictions from a large mass of textual data previously analyzed. BERT was trained on Wikipedia and BookCorpus [4]. The language model trained on a large corpus such as Wikipedia is sometimes limited when using it in a particular domain. Studies in biomedicine [5] showed that in-domain text could give better results than general-domain language models. Thus, several industries have created language models adapted to their application domain to have a system that understands the technical words specific to the domain.
In this paper, we introduce EneBERT, an in-domain language model. It is based on CamemBERT architecture and was trained on the DSO activities dataset. We show the benefit of this language model by using it for diverse tasks like text classification. Our contribution can be summarized as follows:

- EneBERT, a DSO language model.
- The importance of having an in-domain language model for DSO activities.

RELATED WORK

Language models. Word2vec [11], GloVe [12], and fastText [13] were the first notable neural word embedding introduced. Due to the lack of context representation of the word, the use was limited to a specific task. That is why contextualized word embedding, such as ElMo [14], was introduced to represent the context of the word in a sentence. From this approach, the language model was trained on large text data and applied to downstream tasks. It was an LSTM-based architecture that gave promising results until the arrival of Transformers-based architecture BERT [2], RoBERTa [6], GPT2 [10], ALBERT, and T5. These models based on transformers use the same principle as those with LSTM.
BERT gave promising results for multi-downstream NLP tasks like … Since BERT [2], several Transformers-based language models have emerged. Most of them were trained for English application. Non-English contextualized models were introduced first into a multilingual model like mBERT [2]. In the same period, a few monolingual models (in German, Japanese or Portuguese) that were not in English were made based on the language model trained in English. Models like CamemBERT [3] and FlauBERT [14] have been developed to remedy this in French. These are two language models trained on a French corpus. While FlauBERT is based on the BERT [2] architecture, CamemBERT is based on RoBERTa [6]. Our approach is based on CamemBERT.
In-domain language model. This was introduced because general models, such as domain-specific term extraction, could not always be used in specific domains for certain tasks. To address this problem, bioBERT [5], a trained language model for biomedical-specific tasks, was introduced. BioBERT was initialized with the BERT model weights and then pre-trained on large domain-specific corpora. We used the same approach to train our model for DSO activities. Other paper like [16] shows that domain-specific pretraining can outperform mixed-domain pretraining for biomedical NLP applications.

ENEBERT: A FRENCH DSO LANGUAGE MODEL

EneBERT has the same architecture as the French state-of-the-art language model CamemBERT [3]. CamemBERT is based on the architecture of RoBERTa [6], which is a robust optimized BERT [2]. In this section, we will present BERT and RoBERTa architecture, then CamemBERT, and finally introduce a DSO French language model

1 Was at ENEDIS during the works

EneBERT.

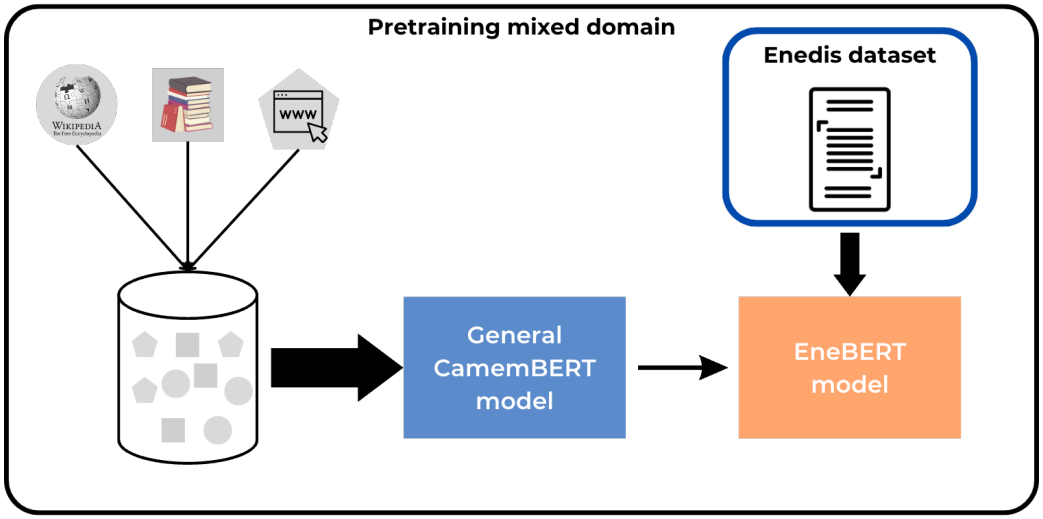

Figure 1 - EneBERT pretrain schema.

BERT and RoBERTa

BERT [2], Bidirectional Encoder Representations from Transformers, is a pretrained language model based on the original Transformer [1] architecture. It is a multi-layer bidirectional Transformer encoder that was trained following two objectives: the mask language model (MLM) and the next sentence prediction (NSP). The MLM pre-trained objective consists of randomly masking some tokens in the input sentence, and the goal is to predict these mask words regarding the context of the sentence. The NSP objective enables the model to capture and understand the relation between two sentences. It consists of a binarized task that predicts whether the sentence is next.

RoBERTa [6] is a robust implementation of BERT better to improve the model's performance in downstream tasks. The main difference between RoBERTa and BERT is the training objective and parameters. RoBERTa used only the MLM pretrained objective and was trained longer than BERT with a bigger batch and over more data.

CamemBERT

CamemBERT [3] is a state-of-the-art language model for French based on the RoBERTa architecture. The main difference between the two is that RoBERTa uses WordPiece tokenization [7], while CamemBERT uses SentencePiece tokenization [8] and whole-word masking (WWM). CamemBERT was pretrained on the French part of OCSAR [9], a multilingual corpus. There are two versions of CamemBERT as it was for BERT: $\text{CamemBERT}_{\text{BASE}}$, which consists of 12 layers, 768 hidden dimensions, 12 attention heads, 110M parameters, and $\text{CamemBERT}_{\text{LARGE}}$, which consists of 24 layers, 1024 hidden dimensions, 12 attention heads, 110M parameters. We do not detail the components of the transformers, and we refer to [1].

EneBERT

Language model like CamemBERT is trained on general data. However, DSO activities have domain-specific terms and proper nouns which can be understood only by the domain specialist. Thus, we design EneBERT, an in-domain French language model trained on a DSO corpus from the domain. EneBERT is based on the architecture $\text{CamemBERT}_{\text{BASE}}$. We initialized EneBERT with the weights of the pre-trained model CamemBERT provided by [3]. These have been taken on the hugging face website. For tokenization, we used the same tokenizer as CamemBERT, but we added some tokens proper to the domain. Figure 1 shows the pretrain schema of EneBERT.

Training data

We used a dataset created from the DSO activities to train our language model. We concatenated multiple datasets from the database. We increased the size of the dataset by using both spelling mistakes or acronyms and adequate expressions. We create a new line for each combination of acronyms that we extract from a sentence.

Figure 2 gives an example. In the example, from the

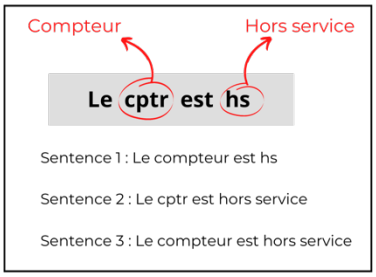

Figure 2 Example of a sentence containing terms specific to the domain. In red the correction of the terms in French. Sentences 1, 2 and 3 have been obtained by playing with the correct term in French and the one specific to the domain.

sentence “Le cptr est hs” which is understandable only by professional agents, we obtained three other sentences that are variations of our starting sentence. These three sentences were added to the corpus. Using a combinatorial library, we generated multiple lines of text, each containing N acronyms or spelling mistakes. By doing this, we were able to consider all possible combinations of words, acronyms or spelling mistakes, and corrections. This helped us explore all the different ways that spelling mistakes and corrections could be formed. We obtained a file of 3.5Gb to train our language model.

Training parameters

We load CamemBERT weight from the hugging face library and finetuned this on our data. We trained the model during 3 epochs with a small learning rate and a batch of 16. The model was trained on 3 NVIDIA RTX A6000 of 48Go each.

DOWSTREAM TASKS

We fine-tuned EneBERT for the text classification tasks.
Enedis have many types of requests as a DSO company. Some types are categorized into clusters. There is also another type of request that has no categories. They are called "miscellaneous requests". Their treatment is, therefore, complex. The way requests are processed is not optimal because requests are processed randomly from the list, not according to their urgency or importance. Indeed, the agents analyze each task, process the ones they are competent at, or leave them in the processing bin with the risk of taking them up again at another time.
We created a text classification model to classify them automatically. To show the contribution of the DSO language model, we finetuned EneBERT on the miscellaneous requests’ dataset. We used the dataset of miscellaneous requests to perform text classification. It consists of 22391 examples for 14 classes.

Class id	Dataset	Train data	Validation data	Train data augmented
class_0	597	478	119	1434
class_1	563	450	113	1350
class_2	1246	997	249	1994
class_3	6652	5321	1331	5321
class_4	280	224	56	672
class_5	452	362	90	1086
class_6	960	768	192	1536
class_7	2766	2213	553	2213
class_8	897	717	180	1434
class_9	4078	3262	816	3262
class_10	202	162	40	486
class_11	1422	1137	285	1137
class_12	656	525	131	1050
class_13	1620	1296	324	1296
Total	**22391**	17912	**4479**	**24271**

Table 1 - Data distribution into train, validation and train data augmented. In gray the classes that have been augmented.

There are two difficulties in using a classification model on these data. The first one is the imbalanced data in each class. To solve the problem of imbalanced data, we split our dataset into train and validation splits. Then, we augmented the class with a small number of lines in the training data by duplicating some lines proportionally. Table 1 shows the data distribution of classes into the dataset, after splitting into train and validation, and into the train data augmented. The second difficulty is the in-domain data, this means that the data contains terms that are specific to the specialists of the domain. Thus, we used the language model described in the previous section to train our request classification model. This was done using the transformer library of Hugging Face which offers functions already defined for the task we wanted to perform. We trained the model for 10 epochs by keeping the best epoch that maximizes the score and minimizes the loss function.

Next section describes the results we obtained for the different tasks.

RESULTS

To show the performance of our language model, we compared it with CamemBERT$_{BASE}$ following the same tasks: next-word prediction and text classification.

Next word prediction

We used the FillMaskPipeline provided by Hugging Face to check whether EneBERT learns interesting things. We gave the model an incomplete sentence and asked it to predict the 4 next probable words. The sentence is “Le cptr est <mask>”. The pipeline predicts the masked token “<mask>” using the model. We compared these outputs to the one obtained using CamemBERT. The results are shown in Table 2. We can show that CamemBERT doesn't know what “cptr” means from his predictions, but EneBERT knows very well what it means because he has been trained on Enedis business data. It is therefore used to see this kind of abbreviation for an "electricity meter".

Model	Possible predictions
CamemBERT	Le cptr est *disponible*. Le cptr est *terminé*. Le cptr est *ouvert*. Le cptr est *gratuit*.
EneBERT	Le cptr est *accessible*. Le cptr est *alimenté*. Le cptr est *défectueux*. Le cptr est *communicant*.

Table 2 - Comparison of CamemBERT and EneBERT outputs for the next word prediction task

Text classification

We evaluated our EneBERT-based demand classification model on the test set. To show the effectiveness of EneBERT, we compared the results obtained with those of CamemBERT trained and evaluated on the same data as EneBERT. The results are recorded in Table 3. We have calculated the F1 score of the 14 classes and the accuracy of the different models. The F1 score is the harmonic mean

between precision and recall. It is, therefore, a good indicator of the model's performance, especially when the data need to be balanced, as in our case. Although the accuracy is not the most indicative measure, we still calculated it to compare the two models.
We can see that EneBERT has better results for the classification of business requests. More specifically, when we can give for a class, CamemBERT has the isolation of the others. These results show that EneBERT has distinguished several classes from the Enedis professions.

	Class id	CamemBERT	EneBERT
F1 score	class_0	0.87	**0.91**
	class_1	0.99	0.99
	class_2	0.94	**0.97**
	class_3	0.97	0.97
	class_4	0.65	**0.68**
	class_5	0.00	**0.56**
	class_6	0.88	**0.91**
	class_7	0.77	**0.84**
	class_8	0.74	**0.81**
	class_9	0.94	0.94
	class_10	0.49	**0.50**
	class_11	0.87	**0.91**
	class_12	0.55	**0.71**
	class_13	0.85	**0.92**
Accuracy		88%	**91%**

Table 3 - F1 and accuracy of CamemBERT and EneBERT. In gray the classes that have been augmented.

DISCUSSION

Having a domain-specific language model is something that is an added value for DSO activities. Indeed, several NLP models have been created on this basis. We continuously improve this model with new data generated by various processes in our company. We have three industrial applications for classification purposes based on EneBERT, zero-shot learning classification, and named entity recognition (such as addresses, phone numbers, and subscribed power...).
Using a personalized state-of-the-art transformer-based language model allowed Enedis to build powerful applications used daily by our operators in various finality scopes such as grid connection, customer relation, and intervention planning.

CONCLUSION

In this paper, we presented EneBERT, a French language model for DSO activities. We showed that the in-domain language model for DSO activities outperformed the general domain language model. One of the challenges of having a specific language model is the implementation of an adapted corpus and the cost of training in the model. However, once implemented, it allows for better performance on domain-specific tasks such as term extraction.

REFERENCES

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, and A. N. Gomez and L. K. and I. Polosukhin, 2017, "Attention Is All You need", *arXiv.*

[2] J. Devlin, M. Chang, K. Lee, and K. Toutanova, 2019, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", *Proceedings Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, ACL, vol.1, 4171-4186.

[3] M. Louis, B. Muller, P. J. Ortiz Suarez, Y. Dupont, L. Romary, E. de la Clergerie, D. Seddah, and B. Sagot, 2020, "CamemBERT: a Tasty French Language Model", *Proceedings Association for Computational Linguistics (ACL)*, 7203-7219.

[4] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler, 2015, "Aligning books and movies: Towards story-like visual explanations by watching movies and reading books", *Proceedings ICCV.*

[5] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, Chan Ho So, and J. Kang, 2019, "BioBERT: a pre-trained biomedical language representation model for biomedical text mining", *Proceedings Bioinformatics,* 3615–3620.

[6] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. 2019. "Roberta: A robustly optimized BERT pre-training approach.", *ArXiv.*

[7] T. Kudo, 2018 "Subword regularization: Improving neural network translation models with multiple subword candidates", *Proceedings Association for Computational Linguistics, ACL*, vol. 1, 66-75.

[8] T. Kudo and J. Richardson, 2018 "Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing", *Proceedings Conference on Empirical Methods in Natural Language Processing, EMNLP,* ACL, 66-71.

[9] P. J. O. Suárez, B. Sagot, and L. Romary, 2019. "Asynchronous pipeline for processing huge corpora on medium to low resource infrastructures ", *Challenges in the Management of Large Corpora (CMLC-7),* pages 9.

[10] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, 2019. "Language models are unsupervised multitask learners", *Preprint*

[11] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, 2013. "Distributed representations of words and phrases and their compositionality", *Proceedings Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems*, 3111–3119.

[12] J. Pennington, R. Socher, and C. D. Manning, 2014. "Glove: Global vectors for word representation.", *Proceedings Conference on Empirical Methods in Natural Language Processing, EMNLP, ACL*, 1532-1543.

[13] T. Mikolov, E. Grave, P. Bojanowski, C. Puhrsch, and A. Joulin, 2018. "Advances in pre-training distributed word representations", *Proceedings Conference on Language Resources and Evaluation, LREC, ELRA*.

[14] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, 2018. "Deep contextualized word representations.", *Proceedings Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, ACL*, vol. 1, 2227–2237.

[15] H. Le, L. Vial, J. Frej, V. Segonne, M. Coavoux, B. Lecouteux, A. Allauzen, B. Crabbé, L.Besacier, D. Schwab, 2020, "FlauBERT: Unsupervised Language Model Pre-training for French", *Proceedings Conference on Language Resources and Evaluation, LREC, ELRA*, 2479-2490.

[16] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, H. Poon, 2020, " Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing", *ArXiv.*