Transformer-Based Named Entity Recognition for French Using Adversarial Adaptation to Similar Domain Corpora (Student Abstract)

Arjun Choudhry[*1, 2], **Pankaj Gupta**[*1], **Inder Khatri** [1], **Aaryan Gupta**[1], **Maxime Nicol** [2], **Marie-Jean Meurs** [2], **Dinesh Kumar Vishwakarma** [1]

[1] Biometric Research Laboratory, Delhi Technological University, New Delhi, India
[2] IKB Lab, Université du Québec à Montréal, Montréal, QC, Canada
{choudhry.arjun, pankajgupta.dtu, inderkhatri999, aryan227227}@gmail.com, nicol.maxime@courrier.uqam.ca, meurs.marie-jean@uqam.ca, dinesh@dtu.ac.in

Abstract

Named Entity Recognition (NER) involves the identification and classification of named entities in unstructured text into predefined classes. NER in languages with limited resources, like French, is still an open problem due to the lack of large, robust, labelled datasets. In this paper, we propose a transformer-based NER approach for French using adversarial adaptation to similar domain or general corpora for improved feature extraction and better generalization. We evaluate our approach on three labelled datasets and show that our adaptation framework outperforms the corresponding non-adaptive models for various combinations of transformer models, source datasets and target corpora.

Introduction

Named Entity Recognition (NER) is an information extraction task where specific entities are extracted from unstructured text and labelled into predefined classes. While NER models for high-resource languages like English have seen notable performance gains due to improvements in model architectures and availability of large datasets, limited-resource languages like French still face a dearth of openly available, large, labelled datasets. Recent research works use adversarial adaptation frameworks for adapting NER models from high-resource domains to low-resource domains. These approaches have been used for high-resource languages, where robust language models are available. We utilize adversarial adaptation to enable models to learn better, generalized features by adapting them to large, unlabelled corpora for better performance on source test set.

We propose a Transformer-based NER approach for French using adversarial adaptation to counter the lack of large, labelled NER datasets in French. We train transformer-based NER models on labelled source datasets and use larger corpora from similar or mixed domains as target sets for improved feature learning. Our proposed approach helps outsource wider domain and general feature knowledge from easily-available large, unlabelled corpora. While we limit our evaluation to French datasets and corpora, our approach can be applied to other languages too.

[*]These authors contributed equally.

Proposed Methodology

Datasets and Preprocessing

We use WikiNER French (Nothman et al. 2012), WikiNeural French (Tedeschi et al. 2021), and Europeana French (Neudecker 2016) datasets as the labelled source datasets in our work. Europeana is extracted from historic European newspapers using Optical Character Recognition (OCR), and contains OCR errors, leading to a noisy dataset. For unlabelled target corpora, we use the WikiNER and WikiNeural datasets without their labels, and the Leipzig Mixed French (Mixed-Fr) corpus. These enable us to evaluate the impact of adapting models to similar domain, as well as generalized corpora. During preprocessing, we convert all NER tags to Inside-Outside-Beginning (IOB) format.

Adversarial Adaptation to Similar Domain Corpus

Adversarial adaptation helps select domain-invariant features transferable between source and target datasets (Ganin et al. 2016). Based on this premise, we propose that adversarially adapting NER models to large, unlabelled corpora from similar domain as the source helps enable the model to extract more generalizable features. This reduces overfitting on the intricate training set-specific features. We also test the same for the case where target dataset is a mixed-domain, large corpus. We test our approach for three conditions: source and target datasets are from the same domain; source and target datasets are from relatively different domains; and target dataset is a mixed-domain, large-scale, general corpus. Figure 1 illustrates our proposed framework. The domain classifier acts as a discriminator. NER classifier loss, adversarial loss, and total loss are defined as:

$$L_{NER} = \min_{\theta_f, \theta_n} \sum_{i=1}^{n_s} L_n^i \tag{1}$$

$$L_{adv} = \min_{\theta_d} (\max_{\theta_f} (\sum_{i=1}^{n_s} L_{ds}^i + \sum_{j=1}^{n_t} L_{dt}^j)) \tag{2}$$

$$L_{Total} = L_{NER} + \alpha(L_{adv}) \tag{3}$$

where n_s and n_t are number of samples in source and target sets, θ_d, θ_n and θ_f are number of parameters for domain classifier, NER classifier and transformer model, L_{d_s} and L_{d_t} represent the Negative log likelihood loss for source and target respectively, and α is ratio between L_{NER} and L_{adv}. We found $\alpha = 2$ to provide the best experimental results.

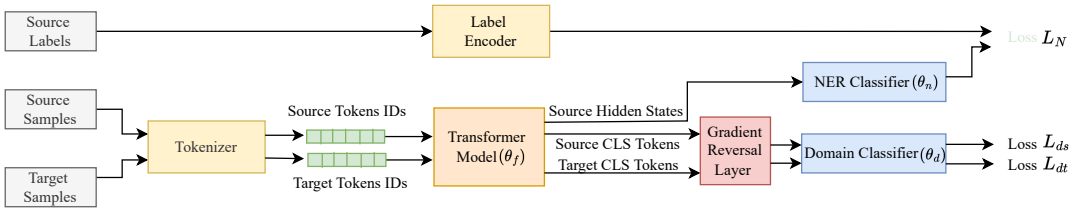

Figure 1: Graphical representation of our adversarial adaptation framework for training NER models on source and target sets.

Language Models for NER

Recent NER research has incorporated large language models due to their contextual knowledge learnt during pretraining. We use three French language models for evaluating our proposed approach: CamemBERT-base (Martin et al. 2020), CamemBERT-Wiki-4GB (a variant of CamemBERT pretrained on only 4GB of Wikipedia corpus), and FlauBERT-base (Le et al. 2020). Comparing CamemBERT-base and CamemBERT-Wiki-4GB helps us analyse if we can replace large language models with smaller ones adapted to unlabelled corpora during fine-tuning on a downstream task.

Experimental Results and Discussion

We evaluated our approach on various combinations of language models, source and target datasets. Each model was evaluated on the test set of source dataset. Table 1 illustrates our results. Some findings observed are described hereafter.

Adversarial adaptation models outperform their non-adaptive counterparts: We observed that the adaptation models consistently outperformed their non-adaptive counterparts across almost all combinations of datasets and language models on precision, recall and F1-score.

Adversarial adaptation can help alleviate performance loss on using smaller models: Fine-tuning CamemBERT-Wiki-4GB using our adversarial approach helped achieve similar performance to non-adapted CamemBERT-base for certain datasets. CamemBERT-Wiki-4GB adapted to WikiNeural corpus even outperformed unadapted CamemBERT-base for WikiNER dataset. Thus, adversarial adaptation during fine-tuning could act as a substitute for using larger language models.

Adapting models to same domain target corpora leads to slightly better performance than adapting to a mixed corpus: We observed that models adapted to corpora from same domain as source dataset (like for WikiNER and WikiNeural as source and target datasets, or vice versa) showed equal or slightly better performance than models adapted to general domain.

Adapting models to mixed-domain target corpus leads to better performance than adapting to a corpus from a different domain: We observed that models adapted to mixed-domain corpora (Europeana to Mixed-Fr) showed noticeably better performance than models adapted to corpora from different domains (Europeana to WikiNER).

Model	Source	Target	Precision	Recall	F1
CamemBERT-Wiki-4GB	WikiNER	-	0.911	0.925	0.918
		WikiNeural	**0.966**	**0.963**	**0.969**
		Mixed-Fr	0.956	0.962	0.959
	WikiNeural	-	0.859	0.872	0.866
		WikiNER	**0.872**	**0.891**	**0.881**
		Mixed-Fr	0.870	0.879	0.875
	Europeana	-	0.728	0.642	0.682
		WikiNER	0.738	**0.691**	**0.714**
		Mixed-Fr	**0.774**	0.640	0.701
CamemBERT-base	WikiNER	-	0.960	0.968	0.964
		WikiNeural	**0.973**	0.976	**0.975**
		Mixed-Fr	0.972	**0.978**	0.974
	WikiNeural	-	0.943	0.950	0.946
		WikiNER	0.943	**0.953**	**0.948**
		Mixed-Fr	**0.946**	0.950	**0.948**
	Europeana	-	0.927	0.933	0.930
		WikiNER	0.911	0.927	0.920
		Mixed-Fr	**0.942**	**0.943**	**0.943**
FlauBERT-base	WikiNER	-	0.963	0.964	0.963
		WikiNeural	0.964	0.968	0.966
		Mixed-Fr	**0.974**	**0.972**	**0.973**
	WikiNeural	-	0.934	0.946	0.940
		WikiNER	0.935	**0.950**	**0.942**
		Mixed-Fr	**0.941**	0.943	**0.942**
	Europeana	-	0.835	0.863	0.849
		WikiNER	0.855	**0.865**	0.860
		Mixed-Fr	**0.882**	0.854	**0.867**

Table 1: Performance evaluation of our proposed approaches

Acknowledgments

Calcul Québec, The Alliance and MITACS.

References

Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; Marchand, M.; and Lempitsky, V. 2016. Domain-Adversarial Training of Neural Networks. *The Journal of Machine Learning Research*, 17(1).

Le, H.; Vial, L.; Frej, J.; Segonne, V.; Coavoux, M.; Lecouteux, B.; Allauzen, A.; Crabbé, B.; Besacier, L.; and Schwab, D. 2020. FlauBERT: Unsupervised Language Model Pre-training for French. In *LREC*.

Martin, L.; Muller, B.; Ortiz Suárez, P. J.; Dupont, Y.; Romary, L.; de la Clergerie, É.; Seddah, D.; and Sagot, B. 2020. CamemBERT: a Tasty French Language Model. In *ACL*.

Neudecker, C. 2016. An Open Corpus for Named Entity Recognition in Historic Newspapers. In *LREC*.

Nothman, J.; Ringland, N.; Radford, W.; Murphy, T.; and Curran, J. R. 2012. Learning Multilingual Named Entity Recognition from Wikipedia. *Artificial Intelligence*, 194.

Tedeschi, S.; Maiorca, V.; Campolungo, N.; Cecconi, F.; and Navigli, R. 2021. WikiNEuRal: Combined Neural and Knowledge-based Silver Data Creation for Multilingual NER. In *EMNLP*.