Cross-Lingual NER for Financial Transaction Data in Low-Resource Languages

Sunisth Kumar
Genify
Gurugram, India
ksunisth@gmail.com

Davide Liu
Genify
Beijing, China
davide@genify.ai

Alexandre Boulenger
Genify
Abu Dhabi, UAE
alex@genify.ai

ABSTRACT

We propose an efficient modeling framework for cross-lingual named entity recognition in semi-structured text data. Our approach relies on both knowledge distillation and consistency training. The modeling framework leverages knowledge from a large language model (XLMRoBERTa) pre-trained on the source language, with a student-teacher relationship (knowledge distillation). The student model incorporates unsupervised consistency training (with KL divergence loss) on the low-resource, target language.

We employ two independent datasets of SMSs in English and Arabic, each carrying semi-structured banking transaction information, and focus on exhibiting the transfer of knowledge from English to Arabic. With access to only 30 labeled samples, our model can generalize the recognition of merchants, amounts, and other fields from English to Arabic. We show that our modeling approach, while efficient, performs best overall when compared to state-of-the-art approaches like DistilBERT pre-trained on the target language, or a supervised model directly trained on labeled data in the target language.

Our experiments show that it is enough to learn to recognize entities in English to reach reasonable performance on a low-resource language in the presence of a few labeled samples of semi-structured data. This has implications for developing multi-lingual applications, especially in geographies where digital endeavors rely on both English and one or more low-resource language(s), sometimes mixed with English or employed singly.

CCS CONCEPTS

• **Computing methodologies** → **Information extraction**; **Natural language processing**.

KEYWORDS

Named Entity Recognition, NLP, Knowledge Discovery, Cross-Lingual NER, Banking Transactions

ACM Reference Format:

Sunisth Kumar, Davide Liu, and Alexandre Boulenger. 2023. Cross-Lingual NER for Financial Transaction Data in Low-Resource Languages. In *Proceedings of The SIGIR '23 Workshop on Knowledge Discovery from Unstructured Data in Financial Services (KDF) (SIGIR'23).* ACM, New York, NY, USA, 4 pages. https://doi.org/X

SIGIR'23, July 23-27, 2023, Taipei, Taiwan

ACM ISBN X…$00.00
https://doi.org/X

1 INTRODUCTION

Named Entity Recognition (NER) has become an essential task in Natural Language Processing (NLP) within the finance domain, given the exponential growth of digital content and the need to extract meaningful insights from financial texts. NER involves identifying and classifying named entities such as organizations, currencies, financial instruments, and monetary values, which is crucial for various NLP applications in finance, including sentiment analysis, risk assessment, and investment recommendation systems. However, developing an accurate and efficient cross-lingual NER model poses significant challenges, such as the lack of labeled data in low-resource languages and the difficulty of capturing cross-lingual variations.

Cross-lingual NER has significant implications on industry, especially for companies that operate in numerous markets and need to analyze customer feedback, social media posts, and other types of unstructured data in multiple languages. Accurately identifying named entities in these languages can help companies extract valuable insights, identify trends, and make informed business decisions. However, developing accurate cross-lingual NER models requires a significant amount of resources, including labeled data, expertise in multiple languages, and computational resources. Therefore, there is a need for efficient and effective cross-lingual NER models that can transfer knowledge from high-resource languages to low-resource languages.

In this paper, we propose a novel framework to improve cross-lingual named entity recognition in semi-structured text data. The proposed framework leverages knowledge distillation to transfer knowledge from a teacher model pre-trained on a high-resource language (English) to a smaller student model. Next, we employ consistency training to fine-tune the student model to Arabic, a low-resource language [6]. Our experiments focus on recognizing entities in semi-structured SMSs that carry banking transaction information in English and Arabic, with access to only a few labeled examples in the target language.

The proposed model shows remarkable cross-lingual learning ability and outperforms state-of-the-art models directly trained on the target language. The main contributions of our work are as follows:

- We leverage knowledge distillation and consistency training to enhance cross-lingual NER in semi-structured text.
- This approach is efficient and effective, requiring only a few labeled examples in the target language.

- The resulting model outperforms other approaches, including DistilBERT pre-trained on the target language or a supervised model trained directly on labeled data in the target language, especially for multi-lingual entity recognition.

2 RELATED WORKS

The problem of cross-lingual Named Entity Recognition (NER) in low-resource languages presents unique challenges and has gathered significant attention in recent research. In recent years, a lot of focus has been on using transfer learning-based methods to address this problem. These methods leverage pre-trained models such as BERT, XLM-RoBERTa to improve the performance of NER models.

Knowledge Distillation has gained attention as a transfer learning technique for cross-lingual NER [3, 7]. By transferring the knowledge from a pre-trained model (teacher) to a smaller model (student), knowledge distillation enables the student model to benefit from the rich representations learned by the teacher model. This method allows for the effective utilization of large-scale pretraining while adapting to the specific cross-lingual NER task.

In addition to the knowledge distillation, we propose incorporating consistency training into the cross-lingual NER framework. Consistency training [5, 9] uses an unsupervised loss function to measure the consistency of the model's predictions on the same input with small random perturbations. By encouraging the model to produce consistent outputs under perturbations, the generalization capability of the model can be enhanced, leading to improved cross-lingual NER performance.

3 METHODS

In this section, we present the methodology employed for the problem of cross-lingual NER for semi-structured financial text data in low-resource languages. This section is structured into three subsections: Problem Formulation, Model Architecture, and Training.

3.1 Problem Formulation

We formulate the task of cross-lingual NER as follows:

Let an input text sequence $X = \{x_1, x_2, ..., x_n\}$, where n is the length of the sequence. Each token x_i is associated with a label y_i, representing its NER tag. The set of possible NER tags is denoted as $Y = \{y_1, y_2, ..., y_k\}$, where k is the total number of entity types.

Given a set of labeled data $D = \{(X_1, Y_1), (X_2, Y_2), ..., (X_m, Y_m)\}$, where each (X_i, Y_i) pair represents an input text sequence and its corresponding NER tag, the objective is to learn a model M that can accurately predict the NER tag Y_i for an unseen input sequence X_i in different languages. In this paper, we are using the Arabic language as the low-resource language.

The cross-lingual aspect of the problem arises from the scarcity of labeled data in low-resource languages. Therefore, the model M should be capable of transferring knowledge from high-resource languages such as English, to low-resource languages, such as Arabic, to improve the performance of NER in those languages.

3.2 Model Architecture

To address these challenges of cross-lingual NER task, we propose a novel framework based on knowledge distillation and consistency training. The model architecture is designed to leverage the benefits of both student-teacher knowledge distillation and consistency training.

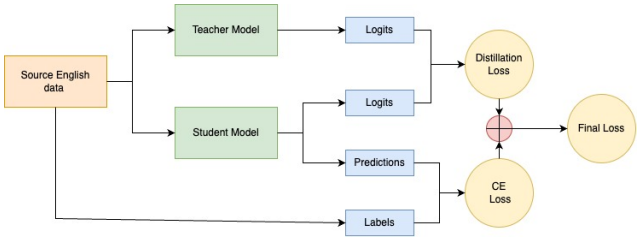

Figure 1: Overview of the student-teacher training framework (KD) with knowledge distillation and cross-entropy loss training on English data.

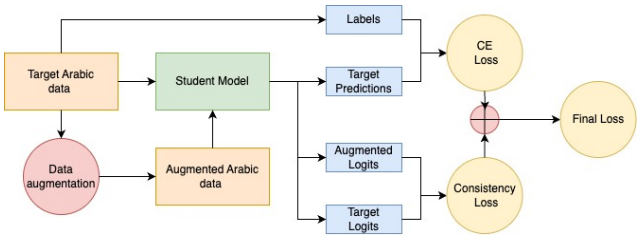

Figure 2: Overview of the knowledge distillation and consistency training framework (KD+CT) for training on Arabic data with consistency-training and cross-entropy loss.

The overall architecture consists of two components, namely (1) knowledge distillation with supervised cross-entropy loss and (2) consistency training.

3.2.1 Teacher Model. The teacher model T is a pretrained XLM-RoBERTa model [1], fine-tuned on the source language (English) dataset. It serves as the source of knowledge transfer and provides soft target distributions for the student model during training. The teacher model takes input tokens X and produces token-level predictions $P^T = \{p_1^T, p_2^T, ..., p_n^T\}$, where p_i^T represents the probability distribution over the NER tags for the token x_i.

3.2.2 Student Model. The student model S is a DistilBERT model [4]. It consists of a multi-layer transformer encoder, similar to the teacher model but with fewer layers and smaller hidden dimensions. The student model takes input tokens X and produces token-level predictions $P^S = \{p_1^S, p_2^S, ..., p_n^S\}$, where p_i^S represents the probability distribution over the NER tags for the token x_i.

3.2.3 Knowledge Distillation. We use knowledge distillation to transfer knowledge from the teacher model to the student model. The distillation loss combined with supervised cross-entropy loss (i.e., $\mathcal{L}_{\text{CE}}$) is defined as:

$$\mathcal{L}_{\text{distill}} = \alpha \mathcal{L}_{\text{CE}} + (1 - \alpha)\text{KL}\left(P^T \| P^S\right) \quad (1)$$

where α is the weight coefficient, and P^T and P^S represent the softened probability distributions obtained by applying the softmax function to the logits of the teacher model and the student model, respectively.

3.2.4 Consistency Training. After the knowledge distillation training, the student model is fine-tuned on the target language (Arabic) using consistency training. Consistency training encourages the model to produce consistent predictions when given different perturbations of the same input. We use a combination of supervised cross-entropy loss (i.e., $\mathcal{L}_{\text{CE}}$) and the unsupervised KL divergence as the consistency loss, comparing the predictions of the augmented data and the original data:

$$\mathcal{L}_{\text{consistency}} = \alpha \mathcal{L}_{\text{CE}} + (1-\alpha)\text{KL}\left(P^{\text{augmented}} \| P^{\text{original}}\right) \quad (2)$$

where α is the weight coefficient, and $P^{\text{augmented}}$ and P^{original} represent the softmax probabilities obtained from the augmented data and the original data, respectively.

During consistency training, we generate augmented versions of the target language data [8] using back translation, RandAugment, and TF-IDF word replacement. These augmented examples are used to compute the unsupervised consistency loss and update the student model parameters accordingly.

4 EXPERIMENTS

4.1 Dataset

We conduct our experiments on a financial transactions dataset consisting of semi-structured SMS data in English and Arabic. The dataset is sourced from Egypt. The English language dataset consists of 1730 sentences along with associated annotated NER tags. The Arabic language dataset consists of 30 sentences. Both language datasets were preprocessed to hide sensitive information and converted to the standard IOB format for NER before training. The Arabic language dataset is used unlabeled for the consistency loss and labeled for the supervised loss. The augmented dataset is generated from this original dataset in Arabic language.

4.2 Experimental Setup

We implement our NER model using the Transformers library and the BERT-based architecture for the teacher and student models. We use AdamW optimizer [2] with a learning rate $l_r = 2e-5$. We use a batch size of 28 and train the NER model for 20 epochs.

We run experiments on α over the range of $0, 0.2, 0.5, 0.8, 1$, as shown in figure **??**). We set $\alpha = 0.8$ based on the best performance. At $\alpha = 0$ (i.e., only unsupervised consistency training loss), the NER model does not learn, and the validation loss increases. At $\alpha = 1$ (i.e., only supervised cross-entropy loss), we observe overfitting on the limited target language data (Arabic), and the validation loss starts to increase after going down. However, at $\alpha = 0.8$ (combination of supervised and unsupervised losses), the NER model gives best performance on cross-lingual NER task for low-resource language.

4.3 Performance Comparison

To evaluate the performance of our proposed cross-lingual NER model, we compare it with that of several baseline models and existing state-of-the-art approaches. The baselines include:

(1) Teacher Model: A pretrained large language model (XLM-RoBERTa) fine-tuned on the English language dataset.
(2) Student Model: A DistilBERT-based student model trained using knowledge distillation from the teacher model.
(3) Naive Benchmark Model: A pretrained DistilBERT model fine-tuned on the target language (Arabic) dataset.

We report the performance comparison in terms of F1 score and accuracy for NER on both the source (English) and the target (Arabic) datasets.

4.4 Results and Analysis

	English		Arabic	
Model	**F1**	**Acc**	**F1**	**Acc**
Teacher	0.9887	0.9888	0.5929	0.6543
Student (only KD)	**0.9957**	**0.9957**	0.5693	0.6852
Student (KD+CT)	0.9768	0.9782	**0.6540**	**0.7407**
DistilBERT	0.6263	0.7377	0.6065	0.7099

Table 1: Comparison of the NER performance of the models on English and Arabic datasets. The accuracies and F1 scores are shown for both English and Arabic datasets.

We compare the performance of our NER model with the Teacher model, the Student model, and the Naive Benchmark model on both the source (English) and the target (Arabic) datasets.

On the English dataset, our model achieves an F1 score of 0.9768, outperforming the Naive Benchmark model by a significant margin. Although the Teacher and Student models exhibit higher F1 scores, we note that our model achieves comparable performance while being smaller than the Teacher model, with F1 score of 0.9887. On the Arabic dataset, our model significantly outperforms the Teacher and the Student models, reaching an F1 score of 0.6540 and an accuracy of 0.7407. Furthermore, our model performs better than the Naive Benchmark model having an F1 score of 0.6065.

These results show that our model achieves competitive performance on both the English (source) and the Arabic (target) datasets. Despite its smaller size and the limited data available in the target language, our model demonstrates remarkable cross-lingual generalization capabilities. It effectively leverages the knowledge distilled from the Teacher model and further enhances its performance through consistency training on the limited target language data.

The overall superior performance of our model can be attributed to its ability to capture and transfer the underlying patterns learned by the Teacher model, leveraging the knowledge distilled during the training process. By incorporating consistency training, our model achieves more robust predictions by ensuring consistency across augmented versions of the input sequences. This training mechanism enhances the model's ability to adapt to cross-lingual contexts and improve performance. The successful combination of knowledge distillation and consistency training contributes to the model's superior performance in capturing both the general patterns and specific language characteristics required for effective cross-lingual named entity recognition.

Overall, our proposed cross-lingual NER model emerges as a promising approach for low-resource languages. Its ability to achieve competitive performance with a smaller model size makes it a practical and efficient solution for real-world applications.

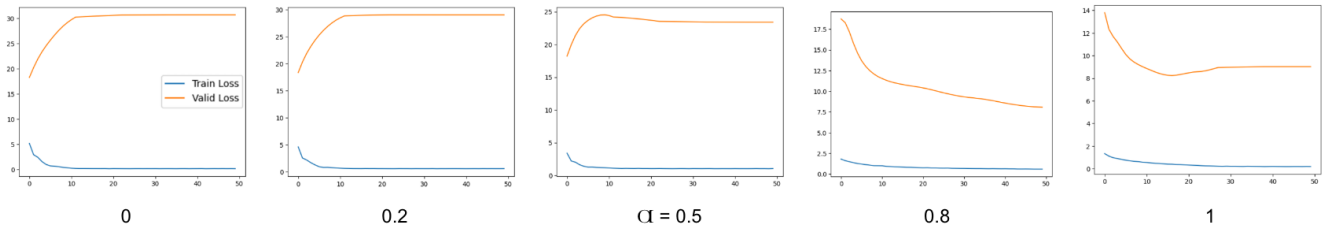

Figure 3: The training and validation loss of KD+CT model (our model) over different values of α in $[0, 0.2, 0.5, 0.8, 1]$ when training it on Arabic data. The loss value and the number of epochs are on the y-axis and x-axis, respectively. The results indicate that when setting the value of α to 0 and 0.2, the model exhibits overfitting behavior on the validation data, as evidenced by an increase in validation loss while the training loss continues to decrease. For α equal to 0.5 and 1 overfitting is still present but not so severe as it was for smaller α. Finally, we empirically found that $\alpha = 0.8$ shows the most desirable learning behavior for validation loss, which almost linearly decreases for the duration of the training.

5 CONCLUSION

In this paper, we introduced a novel framework that uses knowledge distillation and consistency training to enhance cross-lingual named entity recognition in semi-structured text data. Knowledge is transferred from a teacher model pre-trained on English to a smaller student model, which is then fine-tuned for Arabic. We validated the performance of our model (KD+CT) on semi-structured banking transaction data in both English and Arabic, showing competitive performance on both datasets.

Our experiments highlight the potential of our modeling approach, to combine knowledge distillation with consistency training, to address the significant challenges of developing accurate and efficient cross-lingual NER models in low-resource languages.

Our model significantly outperforms the naive benchmark, the student, and the teacher models in entity recognition on the target language dataset (Arabic) and achieves performance comparable to the larger teacher model, while being smaller in size on the source language dataset (English). This demonstrates the remarkable cross-lingual generalization capabilities of our model.

Our model demonstrates remarkable performance in entity recognition on the target language dataset (Arabic), outperforming the naive benchmark, the student, and the teacher models. Additionally, our model achieves comparable performance to the larger teacher model on the source language dataset (English) while maintaining a smaller size. These results highlight the exceptional cross-lingual generalization capabilities of our model.

We believe that our proposed cross-lingual NER model can contribute to the development of multi-lingual applications and enable companies to extract insights, identify trends, and make informed business decisions in multiple languages. We hope our work inspires further research in this field and facilitates the development of efficient and effective cross-lingual NER models, for low-resource languages and beyond.

REFERENCES

[1] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 8440–8451. https://doi.org/10.18653/v1/2020.acl-main.747

[2] Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*. https://openreview.net/forum?id=Bkg6RiCqY7

[3] Jun-Yu Ma, Beiduo Chen, Jia-Chen Gu, Zhenhua Ling, Wu Guo, Quan Liu, Zhigang Chen, and Cong Liu. 2022. Wider & Closer: Mixture of Short-channel Distillers for Zero-shot Cross-lingual Named Entity Recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 5171–5183. https://aclanthology.org/2022.emnlp-main.345

[4] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *ArXiv* abs/1910.01108 (2019).

[5] Rui Wang and Ricardo Henao. 2021. Unsupervised Paraphrasing Consistency Training for Low Resource Named Entity Recognition. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 5303–5308. https://doi.org/10.18653/v1/2021.emnlp-main.430

[6] Xiaobo Wang, Tianyu Fu, Shengcai Liao, Shuo Wang, Zhen Lei, and Tao Mei. 2020. Exclusivity-Consistency Regularized Knowledge Distillation for Face Recognition. In *Computer Vision – ECCV 2020*, Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (Eds.). Springer International Publishing, Cham, 325–342.

[7] Qianhui Wu, Zijia Lin, Börje Karlsson, Jian-Guang Lou, and Biqing Huang. 2020. Single-/Multi-Source Cross-Lingual NER via Teacher-Student Learning on Unlabeled Data in Target Language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 6505–6514. https://doi.org/10.18653/v1/2020.acl-main.581

[8] Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V Le. 2019. Unsupervised Data Augmentation for Consistency Training. *arXiv preprint arXiv:1904.12848* (2019).

[9] Ran Zhou, Xin Li, Lidong Bing, Erik Cambria, Luo Si, and Chunyan Miao. 2022. ConNER: Consistency Training for Cross-lingual Named Entity Recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 8438–8449. https://aclanthology.org/2022.emnlp-main.577

Received 21 May 2023