Exploring Domain Adaptation and Multilingual NER with CharBERT: Applications in News and Multilingual Datasets

Ahmad Sidani
Politecnico di Torino
Torino, Italia
s312919@studenti.polito.it

Fabio Rizzi
Politecnico di Torino
Torino, Italia
s308770@studenti.polito.it

Shaoyong Guo
Politecnico di Torino
Torino, Italia
s296966@studenti.polito.it

***Abstract*—Word embedding have been enhanced by CharBERT [4], a character-aware version of BERT, which incorporates character-level information to capture morphological variations. We implement CharBERT for Named Entity Recognition (NER) in this project, with an emphasis on multilingual and domain adaption features. We refine CharBERT on the AG News [9] dataset and address multilingual settings' NER issues by utilizing its advantages in situations with a dearth of language-specific annotated data.**

Our thorough experiments demonstrate that CharBERT performs well in for state-of-the-art problems. The findings demonstrate how adding character-level information to NER tasks can improve accuracy and resilience in a variety of languages and contexts. In addition to demonstrating CharBERT's successful adaption for specialized and multilingual applications.

***Code:*—https://github.com/developer-sidani/CharBERT**

I. Introduction

Subword granularity representations are widely adopted in many language models like Word2Vec and GloVe, because they enhance the models' capability to manage intricate morphology and diverse vocabularies. However, these methods construct representations based on smaller units (subwords), which can potentially omit information about the complete word and individual characters. Furthermore, minor alterations in characters can significantly alter the resulting subword combinations, impacting the model's robustness.

CharBERT is a pre-trained language model [4] that integrates character-level and word-level information, aimed at addressing the limitations of traditional word-level models in handling unknown words, rare words, and languages with complex morphology. By introducing character-level embeddings on top of BERT or RoBERTa. CharBERT enhances robustness to variations in word forms and spelling changes.

In the research, our objective is to assess the effectiveness and robustness of CharBERT [4] in tasks related to domain adaptation and multilingualism. The study is divided into three parts. Firstly, we replicated the evaluation process from Wentao et al. (2020) for Named Entity Recognition (NER) tasks by fine-tuning CharBERT based on bert and roberta pretrained language models, we took the best CharBERT based and trained a named entity recognition task on the CoNLL-2003 [6] dataset.

We also explore the robustness of CharBERT [4] with news [9] domain, so we finetuned CharBERT based on bert-base-cased, bert-base-cased-ag-news[5] [5], roberta-base, and roberta-base-ag-news [7], and using these finetuned versions of CharBERT on AG news Dataset, we evaluated named entity tasks from the finetuned CharBERT versions, and trained on NER task for CoNLL-2003 [6], and WNUT-17 [1] named entity recognition dataset.

On the multilingual, we evaluated the efficiency of CharBERT based on bert multilingual base cased pretrained language model on multilingual dataset. Firstly, fine-tuned it on a new dataset that merges the Wikipedia English and Wikipedia Italian datasets. Subsequently, CharBERT's performance in the NER task was compared on both CoNLL-2003[6] English and CoNLL-2003 Italian dataset.

II. Related Works

A. Character-Level Embeddings in NLP

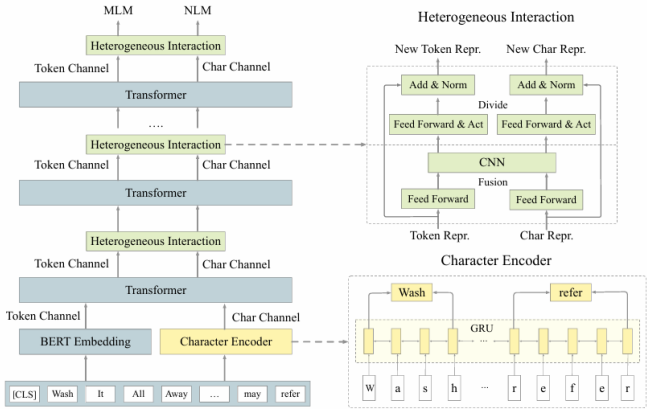

Fig. 1: The neural architecture of CharBERT.

CharBERT [4] is a pre-trained language model that builds upon models like BERT and RoBERTa, aiming to enhance performance and robustness in subword granularity representations. It introduces a dual-channel architecture to separately model information from subwords and characters. As shown in Figure 1, CharBERT uses original pre-trained models, such as BERT and RoBERTa, as its base module. Additionally, it incorporates two new core modules designed to merge information from subwords and characters.

The character encoder, is responsible for encoding the character sequences from the input tokens. It first converts the sequences of tokens into characters and embeds them into fixed-size vectors. Then, it applies a bidirectional GRU layer to construct the contextual character embeddings, which are integrated with the original token vectors.

Heterogeneous interaction, fuses information from two sources and constructs new independent representations for each. The embeddings from the character and original token channels are fed into the same transformer layers in the pre-trained models. After each transformer layer, the heterogeneous interaction module fuses and splits the token and character representations. Finally, these representations are concatenated and fused by a CNN layer.

B. Named Entity Recognition (NER)

NER is a crucial task in NLP [2], and various models have been proposed to improve its accuracy. In CharBERT for NER, we detail the methodology employed for implementing Named Entity Recognition (NER) using CharBERT [4]. Building upon the robust architecture described earlier, CharBERT's dual-channel architecture is leveraged to enhance NER performance by integrating subword and character-level information. Data Preparation:

- Text data is tokenized into subwords using the BERT tokenizer.
- Concurrently, character sequences for each token are generated for character-level embedding.

Character Encoder:

- Character sequences are embedded into fixed-size vectors.
- A bidirectional GRU layer processes these vectors to create contextual character embeddings. These embeddings capture morphological features and subword variations that are essential for recognizing entities.

Heterogeneous Interaction Module:

- This module fuses the contextual character embeddings with subword embeddings from the pre-trained BERT or RoBERTa models.
- The module constructs new independent representations for each token by merging and then splitting the token and character representations after each transformer layer.

Transformer Layers:

- The combined embeddings are processed through multiple transformer layers, where self-attention mechanisms refine the representations.

Final Fusion and Classification:

- After passing through the transformer layers, the representations are concatenated and further processed by a CNN layer.
- The final fused representation is then passed to a classification layer that predicts the entity labels for each token.

This methodology ensures that CharBERT [4] effectively utilizes both character-level and subword-level information, enhancing its ability to recognize and classify named entities accurately across various languages and domains.

C. Domain Adaptation

In natural language processing (NLP), domain adaptation [3] is the process of moving knowledge from a target domain with little labeled data to a source domain with a large amount of labeled data. This procedure is essential for enhancing model performance in specialized or understudied domains without necessitating a large amount of additional annotations. Several methods were implemented using CharBERT for NER [2] tasks in the context of domain adaptation. Fine-Tuning BERT: BERT is especially well-suited for transfer learning because of its architecture, which is built for deep bidirectional representation. Optimizing BERT using domain-specific datasets has been demonstrated to improve performance on a number of tasks, including NER [2]. To achieve notable improvements in model performance, Gururangan et al. (2020) [3], for example, stressed the significance of ongoing pre-training on domain-specific corpora prior to task-specific fine-tuning.

III. Methodology

A. Finetuning CharBERT Model

We assess CharBERT's performance across various domains and multilingual [8] scenarios using different pretrained language models. This evaluation helps understand CharBERT's [4] effectiveness when tailored for specific datasets and tasks. Our studies include baseline evaluations, domain adaptation, and multilingual [8] evaluations.

As a baseline, we fine-tuned RoBERTa-base and BERT-base-cased on the Wikipedia English dataset for three epochs. These models were then tested on the same dataset to establish a performance benchmark for comparison with domain-adapted and multilingual CharBERT [4] versions.

For domain adaptation, we used the AG News [9] dataset. CharBERT [4] was fine-tuned on RoBERTa-base, BERT-base-cased, and their AG News-adapted versions for one epoch. This evaluated CharBERT's adaptability to domain-specific data and improved language modeling.

In the multilingual context [8], we used BERT-multilingual-base-cased, fine-tuning CharBERT on a combined Wikipedia English and Italian dataset for three epochs. This assessment focused on CharBERT's ability to handle multilingual data and maintain high performance across languages, demonstrating its robustness and versatility.

B. Applying CharBERT for Named Entity Recognition (NER)

We investigate CharBERT for Named Entity Recognition (NER) [2] tasks across multilingual [8] and diverse domains, using top-performing BERT-based versions from earlier experiments. All NER models are trained for three epochs.

For the baseline, CharBERT was fine-tuned on the Wikipedia English dataset using BERT-base-cased and then trained on the CoNLL-2003 [6] NER dataset to establish a performance benchmark.

In domain adaptation, we assessed CharBERT's effectiveness on the AG News dataset using BERT-base-cased and BERT-base-cased-AG[5] versions. These were then trained on CoNLL-2003 and WNUT-17 [1] NER datasets, evaluating adaptability to domain-specific tasks.

For the multilingual evaluation, CharBERT was fine-tuned on BERT-multilingual-base-cased with English and Italian Wikipedia data, then trained on English and Italian CoNLL-2003 NER datasets [6]. This assessed CharBERT's robustness and effectiveness in multilingual [8] contexts.

IV. Experiments

A. Datasets

- **The CoNLL-2003 dataset** is a widely used benchmark dataset in the field of Natural Language Processing (NLP), specifically for Named Entity Recognition (NER) [2]. It was introduced during the Conference on Computational Natural Language Learning (CoNLL) in 2003 as part of a shared task on language-independent named entity recognition [6].
- **AG News dataset** is a popular benchmark dataset used for text classification tasks, particularly for classifying news articles into various categories. It was originally collected by the academic and research institution AG's corpus of news articles [9].
- **Wikipedia** is an extracted a subset of the Wikipedia dataset for both English and Italian languages. We divided the dataset into training, validation, and test sets. The training data was further shuffled to create a combined dataset of English and Italian text.
- **WNUT 2017 dataset** is a dataset designed for the Named Entity Recognition (NER) task, specifically focusing on emerging and rare entities. It was introduced as part of the 2017 Workshop on Noisy User-generated Text (WNUT).

B. Implementation Details

To expand the set of comparisons in addiction to bert-base and roberta-base, we also consider versions of bert and roberta available on huggingface that have already been pretrained with specific datasets. In our experiment, we utilized NVIDIA L4 GPUs. These GPUs, each equipped with 16 GB of GDDR6 memory. The executions of the finetuning for all were implemented for 3 epochs of finetuning followed by evaluation, except for news domain, the finetuning was done over 1 epoch. As for named entity recognition tasks, all were done on 3 epochs for training and predicting.

Pretrained Language Models:

- **BERT Base Cased** is a transformer-based neural network model trained on large amounts of English text. It is able to distinguish between uppercase and lowercase letters, making it suitable for various natural language processing tasks.
- **RoBERTa Base** is an enhanced version of BERT, designed to boost performance by fine-tuning training hyperparameters.
- **BERT Multilingual Base Cased** is a variant of BERT that supports multiple languages [8] as it has been trained on text from a wide range of languages. It maintains case sensitivity across all supported languages.
- **Bert AG base cased** is a BERT model fine-tuned on AG News classification dataset that can be found on huggingface [5].
- **Roberta AG base** is a Roberta model fine-tuned on AG News [9] classification dataset that can be found on huggingface [7].

C. Evaluation Metrics

- **AdvAll** This metric represent the total number of adversarial examples on which the model was evaluated. A higher value implies that the model was tested on a larger number of this examples, so providing a more robust assessment of its ability to handle adversarial scenarios.
- **AdvHit1** This metric measures the percentage of times the model correctly identifies the target as the most probable response in the presence of adversarial examples. An high percentage means that the model is highly accurate.
- **AdvHit5** This metric measures the percentage of times the correct target was among the top 5 most probable response predicted by the model in the presence of adversarial examples. An high percentage means that the model is highly accurate.
- **Perplexity** This metric is a measure of how well a language model predicts a sequence of words. It represent, on average, how many words the model considers as possible continuation of a given sequence. Thus, a lower value indicates that the model generate more precise and less uncertain prediction.
- **Precision** measures the accuracy of a model's predictions by evaluating the proportion of true positives (correctly predicted instances) among all instances predicted as positive. A higher precision indicates a better performance.
- **Recall** assesses the accuracy of a model's predictions by calculating the proportion of true positives identified from all actual positive instances in the dataset. A higher recall signifies a better performance.
- **F1-score** combines precision and recall into a single metric using the harmonic mean. It provides a balanced measure of a model's overall performance.

V. Results

This section showcases the diversity of CharBERT [4] with a thorough investigation of its performance in a range of experimental scenarios. Three primary categories comprise

the results: multilingual [8] performance, domain adaptability, and baseline evaluation. Every category offers valuable perspectives on the performance of CharBERT on named entity recognition (NER) and language modeling tasks once it has been refined using several pretrained language models.

PLM	AdvAll	AdvHit1	AdvHit5	Perplexity
bert-base	**75810**	**0.735**	**0.790**	**12.8730**
roberta-base	73538	0.137	0.297	3728.2451

TABLE I: Evaluation Results of CharBERT on English Wikipedia Dataset

PLM	F1-score	Precision	Recall
bert-base	0.908	0.914	0.901

TABLE II: Test results of NER on finetuned CharBERT on CoNLL-2003 english dataset

Significant performance insights are shown by the baseline evaluation results for CharBERT, which were refined on the BERT-base-cased model using the English Wikipedia dataset and then tested on the CoNLL-2003 [6] NER dataset. In contrast to the RoBERTa-base model, which dramatically underperformed with adversarial hit rates of 0.137 and 0.297, CharBERT fine-tuned with BERT-base-cased obtained higher adversarial hit rates (AdvHit1 and AdvHit5) of 0.735 and 0.790, respectively, as shown in Table I . Furthermore, the BERT-base-cased model's perplexity score (12.8730) was significantly lower than the RoBERTa-base model's (3728.2451), suggesting that the former has a stronger language modeling capacity. Table II provides additional evidence of the efficacy of CharBERT fine-tuned with BERT-base-cased, as demonstrated by the NER performance, which on the CoNLL-2003 [6] English dataset yielded an F1-score of 0.908, recall of 0.914, and precision of 0.901 Table II. These results demonstrate CharBERT's robustness and high performance when fine-tuned on BERT-base-cased for both language modeling and NER tasks.

PLM	AdvAll	AdvHit1	AdvHit5	Perplexity
roberta-base-ag	815628	0.133	0.243	2454.03
roberta-base	815628	0.133	0.243	2460.06
bert-base-ag	830013	0.796	0.860	18.3319
bert-base	831387	**0.937**	0.968	**6.991**

TABLE III: Evaluation Results of CharBERT on AG News dataset

Dataset	PLM	F1-score	Precision	Recall
CONLL-2003	bert-base-ag	0.902	0.896	0.909
	bert-base	**0.910**	**0.902**	**0.918**
WNUT-17	bert-base-ag	0.536	0.664	0.449
	bert-base	**0.558**	**0.692**	**0.468**

TABLE IV: Test Results of NER finetuned CharBERT for different pretrained language models on WNUT-17 dataset and CoNLL-2003 English Dataset

The AG News dataset's domain adaptation results for CharBERT fine-tuned show how its performance is improved in particular domains. Table III shows that CharBERT fine-tuned on BERT-base-cased (bert-base) performed better than any other model. It achieved the lowest perplexity score of 6.991, indicating superior language modeling capabilities in the AG News domain, and the highest adversarial hit rates (AdvHit1 and AdvHit5) of 0.937 and 0.968, respectively. The efficacy of CharBERT fine-tuned on BERT-base-cased is further demonstrated by the NER test results, which are shown in Table IV. The BERT-base-cased model outperformed the BERT-base-cased-AG[5] model for the CoNLL-2003 [6] NER dataset, achieving an F1-score of 0.910 with precision and recall scores of 0.902 and 0.918, respectively. Comparing the BERT-base-cased model against the BERT-base-cased-AG model, the former performed better on the WNUT-17 [1] NER dataset, with an F1-score of 0.558, precision of 0.692, and recall of 0.468 IV. These findings demonstrate how well CharBERT fine-tuned on BERT-base-cased adapts to domain-specific data, improving its performance in the AG News domain for language modeling and NER tasks.

PLM	AdvAll	AdvHit1	AdvHit5	Perplexity
bert-base-multilingual	156285	0.802	0.841	28.5877

TABLE V: Evaluation Results of CharBERT on Multilingual Wikipedia Dataset

Dataset	F1-score	Precision	Recall
conll-2003-Eng	0.903	0.909	0.897
conll-2003-Ita	**0.924**	0.925	**0.923**

TABLE VI: Test Results of NER on finetuned CharBERT

On the multilingual Wikipedia dataset, it obtained an AdvHit1 of 0.802, an AdvHit5 of 0.841, and a perplexity of 28.5877, as shown in Table V. For CoNLL-2003 English, the NER findings in Table VI display precision, recall, and F1-score of 0.897, 0.909, and 0.903, and for CoNLL-2003 Italian, 0.923, 0.925, and 0.924, respectively. These outcomes demonstrate CharBERT's outstanding performance and resilience in multilingual environments.

CONCLUSION

Our exploration demonstrates the effectiveness of leveraging CharBERT as a baseline and applying domain adaptation techniques within the context of news and different languages. The successful training showcases the robustness and versatility of BERT-based models in handling diverse challenges and specific domain contexts. With that being said, exploring learning techniques will be crucial to keep improving the model's usefulness and robustness in real-world media analysis in different languages and different domains. Future work will involve integrating more advanced domain adaptation methods, exploring few-shot learning, and extending evaluations to underrepresented languages and diverse datasets to further enhance CharBERT's capabilities.

REFERENCES

[1] Leon Derczynski et al. “Results of the WNUT2017 Shared Task on Novel and Emerging Entity Recognition”. In: *Proceedings of the 3rd Workshop on Noisy User-generated Text*. 2017, pp. 140–147.

[2] Jacob Devlin et al. “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)*. 2019, pp. 4171–4186.

[3] Suchin Gururangan et al. “Don’t Stop Pretraining: Adapt Language Models to Domains and Tasks”. In: *arXiv preprint arXiv:2004.10964* (2020).

[4] Wentao Ma et al. *CharBERT: Character-aware Pre-trained Language Model*. 2020. arXiv: 2011 . 01513 [cs.CL].

[5] Lucas Resck. *BERT Base Cased AG News*. https : / / huggingface . co / lucasresck / bert - base - cased - ag - news. Accessed: 2024-07-17. 2020.

[6] Erik F. Tjong Kim Sang and Fien De Meulder. “Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition”. In: *Proceedings of CoNLL-2003*. 2003, pp. 142–147.

[7] textattack. *ROBERT Base Cased AG News*. https : / / huggingface.co/textattack/roberta-base-ag-news.

[8] Shijie Wu and Mark Dredze. “Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT”. In: *arXiv preprint arXiv:1904.09077* (2019).

[9] Xiang Zhang, Junbo Zhao, and Yann LeCun. *Character-level Convolutional Networks for Text Classification*. https://arxiv.org/abs/1509.01626. 2015.