CharBERT: Domain Adaptation and Multilingual QA Applications on Medical and Multilingual Datasets

Francesco Giuseppe Gillio
Department of Computer Science
Polytechnic University of Turin

March 26, 2024

Abstract

CharBERT [1], a character-aware version of BERT, enhances word embeddings by incorporating character-level information to catch morphological variations. This research explores CharBERT performance for Question Answering (QA) tasks, with an emphasis on multilingual and domain adaptation contexts. The exploration refines CharBERT on domain-specific datasets (BioASQ for QA in a medical framework), language-specific datasets (MLQA for QA in a multilingual environment), and measures the model's effectiveness in managing data with typos or morphological variations. The extensive experimentation reveals CharBERT's high performance in addressing state-of-the-art problems. The findings underscore the value of incorporating character-level information in QA tasks, leading to higher resilience across multiple languages and diverse contexts. In addition, the results reveal the effective adaptation of CharBERT to domain-specific and multilingual applications with morphological variations and typos in the data.

GitHub: https://github.com/305909/charbert

1 Introduction

Subword representations enhance the ability of language models to process complex morphology and diverse linguistic structures, as shown in 2013 by models such as Word2Vec [10] and FastText [11]. However, subwords methods often overlook critical information relating to entire words or individual characters. Even minor character variations may lead to substantial changes in subword combinations, which compromises the robustness of these models in practical applications. CharBERT [1], a pre-trained language model, merges character-level and word-level embeddings. This design addresses the limitations of traditional word embeddings methodologies in managing rare or unknown words and languages with intricate morphological structures. By embedding character-level features into models such as BERT or RoBERTa, CharBERT improves resilience against word form variations and spelling inconsistencies.

This research aims to assess the effectiveness and robustness of CharBERT [1] in domain adaptation and multilingual Question Answering (QA) tasks. The study progresses through three key stages. In the first stage, CharBERT, build upon bert-base-cased [3], undergoes fine-tuning on the English Wikipedia dataset [5] for language modeling and on the SQuAD dataset [6] for QA tasks. The second stage shifts to a domain-specific approach, fine-tuning the model on the PubMED dataset [7] for language modeling in a medical framework and on the BioASQ dataset [8] for medical QA. The third stage explores multilingual performance, fine-tuning CharBERT, build upon bert-base-multilingual-cased [4], on both English and German Wikipedia datasets [5] for language modeling in a multilingual framework and on the MLQA dataset [9] for multilingual QA.

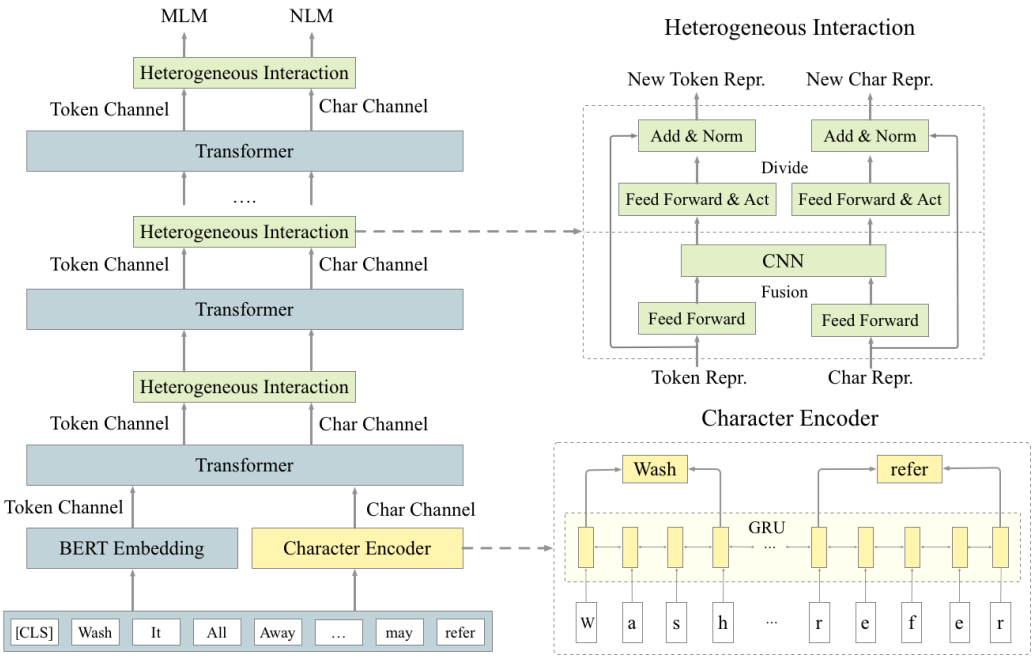

Figure 1: The Neural Architecture of CharBERT [1].

Through this research, CharBERT's performance across domain-specific and multilingual datasets undergoes rigorous assessment. This study offers an in-depth analysis of CharBERT's effectiveness across various linguistic environments and domain-specific demands.

2 Related Works

2.1 Character-Level Embeddings in NLP

CharBERT [1] represents an high-level language model built upon the foundations of BERT [2] and RoBERTa. The primary objective of the model focuses on refining performance and enhancing robustness in subword granularity representations. CharBERT incorporates a dual-channel architecture to model in a separate fashion information coming from subwords and characters. As shown in Figure 1, CharBERT engages pre-trained models such as BERT and RoBERTa as base framework, while integrating two innovative core modules to merge information from both subwords and characters.

The **Character Encoder** manages the encoding process of character sequences coming from input tokens. This framework transforms token sequences into characters, then embedding those characters into fixed-size vectors. Furthermore, a bidirectional GRU layer constructs contextual character embeddings, which then integrate with the original token vectors. The **Heterogeneous Interaction** module merges information from two distinct sources, outputting independent representations for each. Embeddings from both the character and token channels pass through the same transformer layers within the pre-trained models. Post-processing by each transformer layer, the heterogeneous interaction module merges and segregates token and character representations. In the end, a CNN layer performs concatenation and fusion of these representations.

2.2 Question Answering (QA)

Question Answering represents a core task in NLP. CharBERT [1] addresses the challenge through the dual-channel architecture to improve performance in this tasks. The architecture handles subwords via BERT tokenizer and enhances these with character-level embeddings to catch morphological nuances.

Data Preparation: The BERT tokenizer breaks down text into subwords, and a separate module outputs character sequences for each token to enhance character-level embedding.

Character Encoder: The Character Encoder converts character sequences into fixed-size vectors, and a bidirectional GRU layer processes these vectors to establish contextual character embeddings. This process enables the model to understands morphological nuances and subword variations.

Heterogeneous Interaction Module: The Heterogeneous Interaction Module merges the subword and character embeddings to construct separate representations for each token after each transformer layer, enabling both subword and character information to contribute to the final representation.

Transformer Layers: The Transformer Layers stack processes the subword-character embeddings via self-attention mechanisms to refine and improve the representations.

Final Fusion and Prediction: After the transformer layers, a CNN layer concatenates the embeddings and directs the final representation into a prediction layer.

This methodology ensures that CharBERT [1] harnesses both character-level and subword-level information to respond questions and oputput accurate answers across diverse languages and domains.

2.3 Domain Adaptation

Domain adaptation within Natural Language Processing (NLP) encompasses the transfer of knowledge from a source domain to a target domain. This approach proves crucial for enhancing model performance in context-specific domains without requiring a substantial increase in annotations. CharBERT's dual-channel architecture proves adept at transfer learning for addressing domain-specific QA challenges. The domain adaptation process with CharBERT involves several key steps:

Domain-Specific Pre-Training

1. Pre-Training Corpus: domain-specific corpus, comprising text pertinent to the target domain, to pre-train the model to assimilate the nuances of domain-specific terminology, syntax, and contextual subtleties.
2. Pre-Training Objectives: upon this phase, CharBERT learns representations that align with the unique characteristics of the domain, positioning the model to handle domain-specific questions and answers.

Task-Specific Fine-Tuning

1. Fine-Tuning Corpus: task-specific corpus, comprising domain-relevant questions and answers, to fine-tune the model to align with task-specific requirements.
2. Fine-Tuning Objectives: upon this stage, CharBERT tunes internal parameters to optimize the model's performance in managing on domain-specific QA tasks.

Evaluation and Iteration

1. Performance Assessment: post fine-tuning, evaluation of the model's performance through domain-specific benchmarks ensures adherence to relevance standards.
2. Continuous Improvement: performance assessments require further fine-tuning and adjustments to enhance the model's effectiveness in managing domain-specific QA tasks.

3 Methodology

3.1 CharBERT Model for QA

The project explores domain-specific pre-training and task-specific tuning processes to enhance the model's effectiveness across domain-specific and multilingual Question Answering tasks. In the pre-training phase, the model processes domain-relevant corpora through Masked Language Modeling (MLM).

This technique involves masking a percentage of tokens in a sentence, with the model predicting the `[MASK]` tokens from the surrounding context. This approach enables the model to catch semantic and syntactic patterns pertinent to the domain, enhancing adaptation to domain-specific knowledge. In the task-specific tuning phase, the model identifies two critical tokens within a passage that represent the boundaries of the answer.

The system learns to locate these `[CLS]` and `[SEP]` tokens according to the input question and the corresponding context, refining the model's effectiveness on the task at hand. These sequential procedures - pre-training with MLM and task-specific fine-tuning - enhance the model's performance in handling complex domain-specific and multilingual Question Answering scenarios.

Baseline Evaluation

1. Fine-Tuning the pre-trained language model, bert-base-cased [3], on the English Wikipedia dataset [5] via Masked Language Modeling (MLM) approach to enhance the model's comprehension performance in English.
2. Fine-Tuning the model on the SQuAD dataset [6] via Question Answering approach to refine the model's performance in English QA tasks.

Subsequent assessments on the same dataset provide a reference point for evaluating the performance of domain-adaptation and multilingual versions of CharBERT.

Domain Adaptation

1. Fine-Tuning the pre-trained language model, bert-base-cased [3], on the PubMED dataset [7] via Masked Language Modeling (MLM) approach to enhance the model's comprehension performance in medical knowledge.
2. Fine-Tuning the model on the BioASQ dataset [8] via Question Answering approach to refine the model's performance in medical QA tasks.

Multilingual Context

1. Fine-Tuning the pre-trained multilingual language model, bert-base-multilingual-cased [4], on both English and German Wikipedia datasets [5] via Masked Language Modeling (MLM) approach to enhance the model's comprehension performance in English and German.
2. Fine-Tuning the model on the MLQA dataset [9] via Question Answering approach to refine the model's performance in English and German QA tasks.

4 Experiments

4.1 Dataset

- **Wikipedia Dataset** [5]: Wikipedia's vast corpus offers a wide source of text from diverse fields, providing a foundation for language model pre-training. This project samples subsets of English and German articles, branching into training, validation, and test sets. These subsets cover an array of topics, contributing to a broad understanding of both languages. The diverse and extensive nature of Wikipedia ensures that fine-tuning on this dataset aids in enhancing the model's general linguistic capabilities, including syntax, semantics, and contextual awareness.
- **SQuAD Dataset** [6]: The Stanford Question Answering Dataset (SQuAD) serves a crucial role in advancing QA model evaluation. Comprising questions that refer to passages from Wikipedia articles, SQuAD challenges models to extract precise spans of text as answers. The focus on understanding the nuances of context and the ability to accurately locate information within a passage makes this dataset indispensable for testing a model's performance in comprehension and information retrieval, particularly in English-language tasks.
- **PubMED Dataset** [7]: PubMED offers a domain-specific repository of biomedical texts, providing a wealth of domain-specific knowledge for models operating in the medical framework.

This dataset aids in training models to navigate complex medical terminology and domain-specific knowledge. By leveraging PubMED, models fine-tune their understanding of biomedical literature, enhancing more accurate responses to medical questions and enhancing performance in health-relevant applications.

- **BioASQ Dataset** [8]: BioASQ dataset contains a collection of biomedical questions and corresponding answers drawn from PubMED articles. This dataset serves as a benchmark for evaluating the model's effectiveness to handle domain-specific QA tasks. This dataset also supports the BioASQ challenge, which encourages the development of systems capable of semantic indexing and precise question answering within the biomedical domain.
- **MLQA Dataset** [9]: The Multilingual Question Answering (MLQA) dataset presents a crucial benchmark for assessing models in cross-lingual question-answering scenarios. With data covering multiple languages, including English and German, MLQA tests a model's ability to handle questions and passages in various linguistic contexts. This multilingual dimension provides valuable insights into model's effectiveness to generalize across languages, becoming essential for evaluating performance in global and cross-lingual applications.

4.2 Implementation

To broaden the scope of model performance beyond bert-base, this project explores various versions of BERT model, available through HuggingFace. These models offer high-level linguistic capabilities through prior training on domain-relevant corpora. This research runs the experiments on NVIDIA L4 GPUs, each offering 16 GB of GDDR6 memory, fine-tuning the language models for 2 training epochs. Pre-Trained Language Models:

- **BERT** [2] **Base Cased** [3]: Pre-Trained model on English language via Masked Language Modeling (MLM) objective. This model handles uppercase and lowercase distinctions, suitable for several natural language processing applications.
- **BERT** [2] **Multilingual Base Cased** [4]: Pre-Trained model on the top 104 languages with the largest Wikipedia via MLM objective. This model retains case sensitivity across different languages.

4.3 Evaluation Metrics

- **F1-score**: F1-score serves as a comprehensive metric that balances precision and recall by calculating their harmonic mean. In QA tasks, precision refers to the proportion of correct predictions out of the answers the model provides, while recall measures the proportion of correct predictions out of the correct answers the data provides. By integrating these two aspects, the F1-score offers an evaluation of the model's performance to catch as many relevant answers as possible.
- **Exact Match (EM)**: Exact Match assesses the proportion of instances where the model's prediction matches the ground-truth answer. This metric evaluates whether the model returns the correct answer without any deviation, including differences in phrasing, punctuation, or tokenization, thus providing a more stringent measure of accuracy. Although Exact Match offers a clear reflection of the model's precision, the metric does not account for nearly correct answers or semantically equivalent responses that differ in phrasing from the ground-truth.

Table 1: QA Evaluation Results on SQuAD Dataset

Model	F1-Score	EM
BERT	73.7	76.3
CharBERT	75.2	78.1

Table 2: QA Evaluation Results on BioASQ Dataset

Model	F1-Score	EM
BERT	65.3	69.7
CharBERT	69.1	75.2

Table 3: QA Evaluation Results on MLQA Dataset

Model	F1-Score	EM
BERT	71.5	73.0
CharBERT	73.9	76.4

5 Results

CharBERT [1] shows superior performance across multiple datasets with respect to the baseline BERT [2] model. The model's character-aware embeddings help improve managing of morphological variations and typographical errors. CharBERT achieves higher scores than BERT on SQuAD [6], with an F1-score of 75.2 and an Exact Match (EM) of 78.1, with respect to BERT's F1-score of 73.7 and EM of 76.3, as shown in Table 1. This indicates CharBERT's stronger expertise to catch exact answers in standard QA tasks. In the domain of biomedical literature of BioASQ [8], CharBERT showcases a significant improvement. With an F1-score of 69.1 and EM of 75.2, CharBERT outperforms BERT's F1-score of 65.3 and EM of 69.7, as shown in Table 2. The results suggest that CharBERT's character-level understanding aids in resolving domain-specific terminology and complex questions. For multilingual QA tasks on MLQA [9], CharBERT continues to outperform BERT, with an F1-score of 73.9 and EM of 76.4, with respect to BERT's F1-score of 71.5 and EM of 73.0, as shown in Table 3. This illustrates the model's effectiveness in cross-lingual settings, where character-level capabilities enhance understanding and resilience across different languages.

The robustness of CharBERT became more evident in managing dataset with morphological variations and typos. CharBERT outperforms BERT on the adversarial version of the SQuAD dataset [6], achieving an F1-score of 66.8 and EM of 69.4, with respect to BERT's F1-score of 62.3 and EM of 64.8, as shown in Table 4. This demonstrates the model's expertise to recover with greater success with data inconsistencies. In the medical domain [8], CharBERT again shows an edge over BERT, scoring 66.5 in F1 and 72.8 in EM, while BERT manages an F1 of 61.9 and EM of 65.4, as shown in Table 5. This performance emphasizes CharBERT's resiliance in high-stakes environments such as healthcare. For multilingual data [9], CharBERT achieves a higher F1-score (70.5) and EM (72.9) than BERT (F1: 67.1, EM: 69.2), as shown in Table 6, further validating the character-aware effectiveness in multilingual environments.

Table 4: QA Evaluation Results on Attack SQuAD Dataset

Model	F1-Score	EM
BERT	62.3	64.8
CharBERT	66.8	69.4

6 Conclusion

This research evaluates CharBERT [1], a character-aware variant of BERT [2], in domain adaptation and multilingual Question Answering (QA) tasks. CharBERT demonstrates superior performance with respect to BERT by managing morphological variations, typos, and multilingual contexts. CharBERT's dual-channel architecture, integrating character-level and subword-level embeddings, shows notable improvements across various datasets.

Table 5: QA Evaluation Results on Attack BioASQ Dataset

Model	F1-Score	EM
BERT	61.9	65.4
CharBERT	66.5	72.8

Table 6: QA Evaluation Results on Attack MLQA Dataset

Model	F1-Score	EM
BERT	67.1	69.2
CharBERT	70.5	72.9

On the SQuAD dataset, CharBERT achieves higher F1-score (75.2 vs. 73.7) and Exact Match (78.1 vs. 76.3) with respect to BERT. In the medical domain, CharBERT surpasses BERT with an F1-score of 69.1 and an EM of 75.2, with respect to BERT's 65.3 and 69.7. For multilingual tasks, CharBERT outperforms BERT on the MLQA dataset, showing an F1-score of 73.9 and an EM of 76.4, with respect to BERT's 71.5 and 73.0. CharBERT also excels in managing data with noise, with significant improvements in F1-score and EM on adversarial versions of the SQuAD and BioASQ datasets. These results highlight CharBERT's robustness and resiliance across diverse domains and languages.

In conclusion, CharBERT's integration of character-level information enhances the model's performance and resilience in QA tasks, marking a valuable advancement in Natural Language Processing (NLP). Future research may investigate the application of this model in additional domains and emerging language technologies.

References

[1] Wentao Ma, Yiming Cui, Chenglei Si, Ting Liu, Shijin Wang, Guoping Hu. CharBERT: Character-Aware Pre-Trained Language Model. *arXiv preprint arXiv:2011.01513*, 2020.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*, 2019.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova. BERT Base Cased Model. *arXiv preprint arXiv:1810.04805*, 2019. Hugging Face: https://huggingface.co/google-bert/bert-base-cased.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova. BERT Base Multilingual Cased Model. *arXiv preprint arXiv:1810.04805*, 2019. Hugging Face: https://huggingface.co/google-bert/bert-base-multilingual-cased.

[5] Wikimedia Foundation. Wikimedia Downloads. Hugging Face: https://huggingface.co/datasets/legacy-datasets/wikipedia.

[6] Association for Computational Linguistics. Stanford Question Answering Dataset (SQuAD). Hugging Face: https://huggingface.co/datasets/rajpurkar/squad_v2.

[7] *PubMED*, Courtesy of the U.S. National Library of Medicine. (2018). Hugging Face: https://huggingface.co/datasets/ncbi/pubmed.

[8] BioASQ: Large-Scale Biomedical Semantic Indexing and Question Answering Competition. Hugging Face: https://huggingface.co/datasets/kroshan/BioASQ.

[9] MLQA: Evaluating Cross-lingual Extractive Question Answering. Hugging Face: https://huggingface.co/datasets/facebook/mlqa.

[10] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, Jeffrey Dean. Distributed Representations of Words and Phrases and their Compositionality. *arXiv preprint arXiv:1310.4546*, 2013.

[11] Piotr Bojanowski, Edouard Grave, Armand Joulin, Tomas Mikolov. Enriching Word Vectors with Subword Information. *arXiv preprint arXiv:1607.04606*, 2017.