Named Entity Recognition using BERT, RoBERTa and ALBERT

Navya Sree Nutakki,[1] Sai Pranaswi Mullangi,[2] Sai Goutham Nelanuthula[3]

Spring 2023, NYU Tandon School of Engineering

nn2382,[1] sm11006,[2] sn3533[3]

Codelink:
https://github.com/SaiPranaswi23/Final_Project

Abstract

Named Entity Recognition (NER) is a crucial task in Natural Language Processing (NLP) that involves identifying and extracting entities from text, such as people, organizations, locations, and other types of named entities. Recent advances in deep learning have led to the development of powerful language models, such as BERT ,RoBERTa and ALBERT that have significantly improved the performance of NER systems. In this paper, we present a comparative study of the performance of BERT and RoBERTa models for NER tasks on a variety of datasets. We then describe our experimental setup, including the datasets used, the pre-processing steps, and the evaluation metrics.

Overview

Named Entity Recognition (NER) is a task in Natural Language Processing (NLP) that involves identifying and extracting entities from text. Entities are typically proper nouns that refer to specific people, places, organizations, or other named objects. The goal of NER is to automatically identify and classify these entities into pre-defined categories, such as person, organization, location, date, time, etc. NER is an important task in many NLP applications, such as information extraction, text summarization, question-answering systems, and sentiment analysis.

We have used 2 datasets for our models MIT movie Corpus, and MIT Restaurant Dataset. The MIT movie Corpus consists wide range of dialogues, and the Restaurant Dataset consists of the reviews obtained from customers. These datasets provide a diverse range of data for our model to optimize and work elegantly.

Our results show that both BERT, RoBERTa and ALBERT models achieve state-of-the-art performance on most of the datasets tested, with RoBERTa and ALBERT generally outperforming BERT. We also investigate the impact of fine-tuning and data augmentation on the performance of the models and find that both techniques can improve the performance significantly. In this paper, we propose a novel approach for Named Entity Recognition using the BERT, RoBERTa and ALBERT fine-tuned on the MIT Movie Corpus, MIT restaurant Dataset.

BERT, RoBERTa and ALBERT are pre-trained language models that have achieved state-of-the-art performance on various natural language processing tasks, including NER. These models are based on the transformer architecture and have been trained on massive amounts of text data to learn representations of natural language that capture the nuances and complexities of language. By incorporating these modes with the datasets, we are able to better understand the words, and classify them into correct entities, which allows us to more accurately classify and rank them. This ultimately leads to a more refined and effective word classification process, resulting in higher quality responses to user queries.

Related Work

There has been a significant amount of research on named entity recognition in the natural language processing literature. Previous approaches have used a variety of methods, including rule-based methods, template-based methods, and machine learning-based methods. It has demonstrated strong performance in various NLP benchmarks, including NER. BERT's architecture allows it to capture contextual information effectively, which is crucial for tasks like NER.

Previous research in the field of Named Entity Recognition (NER) has extensively explored the use of BERT, RoBERTa and ALBERT models. BERT provided a breakthrough by leveraging deep bidirectional transformers for language understanding. Although the focus was not NER, researchers quickly realized its potential for fine-tuning on NER tasks. RoBERTa an optimized variant of BERT, further enhanced performance by employing a robust pretrained approach and larger-scale training data. Both models have proven effective in capturing information, making them popular choices for NER. ALBERT is also an optimized variant of BERT that enhances performance by employing a different approach to training and model size. ALBERT improves upon BERT by using the parameter sharing and cross-layer parameter sharing to reduce model size and increase efficiency.

However, previous approaches to Named Entity Recognition (NER) typically relied on hand-crafted features and models that were trained on small, domain-specific datasets. These approaches often required significant engineering effort and domain expertise to develop and achieve good performance. Machine learning-based methods can be more flexible, but they may require large amounts of labeled data to perform well.

Our approach aims to address these limitations by using theBERT model and the RoBERTa models, which have demonstrated strong performance on a variety of natural language processing tasks.

About the datasets:

The MIT Movie Corpus is a dataset of English-language movie transcriptions, which has been widely used for training and evaluating natural language processing models. The corpus contains 25,000 dialogues from 617 movies, covering a wide range of genres and styles. The text in the corpus includes spoken dialogues, character names, scene descriptions, and other annotations.

The MIT Restaurant Dataset is a dataset of customer reviews of restaurants, along with corresponding ratings and attribute tags. The dataset consists of over 1,000 reviews of 20 different restaurants in the Boston area. Each review includes the text of the review, the rating (on a scale from 1 to 5), and a set of attribute tags that describe various aspects of the restaurant, such as the food, service, ambiance, and price range.

We used these datasets to fine-tune and train BERT and RoBERTa models with improved features, for Named Entity Recognition and evaluate the performance of our approach. The inclusion of a diverse set of words in these datasets allows us to train and evaluate our models on a wide range of entities and test their generalization ability.

Data Processing

We obtained the MIT movie corpus and MIT restaurant corpus which contain annotated text with named entities. We split the data into training and testing. We used the tokenizer provided by the Hugging Face library, specifically designed for BERT and RoBERTa models to tokenize the input text. We used auto tokenizer to tokenize the input text for ALBERT model.Tokenization involves breaking down the text into individual tokens (words or sub words) to create sequences that the models can process.

For NER, we assigned labels to each token in the dataset to indicate whether it is a part of the named entity or not. The named entities of the MIT movie and MIT restaurant corpus are labelled using the IOB (Inside, Outside, Beginning). This scheme assigns tags to each token to indicate its role in a named entity. Common labels include B-PER (beginning of a person entity), I-LOC (inside a location entity), O (non-entity token). We introduced special tokens such as [CLS] (classification) and [SEP] (separator) to mark the beginning and separation of input sequences. These tokens provide additional contextual information to the models during training.

To ensure consistent input sizes, we applied padding or truncation techniques. Padding involved adding special padding tokens to sequences that were shorter than the maximum sequence length, while truncation involved removing tokens from sequences that exceeded the maximum length. We organized the processing data into batches for effective training. It involves grouping sequences with similar lengths together, minimizing the need for padding and maximizing the computational efficiency during training. The three models BERT, RoBERTa and ALBERT are used on the datasets.

Architecture

BERT (Bidirectional Encoder Representations from Transformers) is a transformer based model designed for various natural language processing tasks including NER. BERT architecture consists of multiple layers of self-attention and feed-forward neural networks. The model takes tokenized input text and learns contextual representations of each token by considering its surrounding context from both left and right. The architecture of BERT includes an embedding layer, followed by multiple transformer layers. The embedding layer converts the input tokens into continuous vectors, incorporating token, segment and position embeddings. These embeddings provide the information about token identity, sentence segmentation and token position respectively.

The transformer layers in BERT utilizes the self-attention mechanism to capture contextual dependencies within the input sequence. The attention mechanism helps capture long-range dependencies and enhances the model's abilities to understand the relationship between tokens, which is crucial for NER. The output of the transformer layers is then fed into a final classification layer, which predicts the probability of each token belonging to a named entity class. A linear layer with softmax function is used for the classification tasks.

RoBERTa, an optimized version of BERT, follows a similar architecture but employs some modifications in the training methodology. RoBERTa stands for Robustly Optimized BERT Pretraining approach. RoBERTa also consists of transformer layers with self-attention mechanisms. RoBERTa has larger-scale training data and an extensive pre-trained approach. It is trained on the massive corpus of unlabelled text, which helps it capture a broader range of language patterns and improve the performance on tasks like NER. The major difference in the architecture of RoBERTa lies in its training process, which involves more iterations and larger batch sizes compared to BERT.

ALBERT stands for A Lite BERT. It involves tokenizing the input text into subwords and mapping them to word embeddings. Multiple transformer encoder layers capture contextual dependencies between tokens. The model predicts the entity labels for each token using a token classification. During training, the model is fine-tuned on the NER task using labelled examples, and its performance is evaluated based on the metrics like F1-score.

Methodology

After preprocessing the dataset according to our requirement, we used three models to predict named entities of movie and restaurant review datasets. The models we used are BERT, RoBERTa and ALBERT.

We used F1 score metrics to evaluate the performance of our models. F1 score measures the hormonic mean of precision and recall.

Model 1:

Firstly, we started with fine-tuning BERT model. In this model,we used BertTokenizer to convert the input sentences into tokens that are suitable for input to BERT model. It provides some features like tokenization, vocabulary, special tokens and functionalities including encoding and decoding for input text and token ids. Here, WordPiece algorithm is used to split words into sub words allowing model to handle out of vocabulary words.

Our model first encodes the sentences/reviews to token ids with attention masks. Then the new labels for each sentence were created with the help of label map and null_label_id as -100. Here, label map is the mapping of original labels of dataset to numbers from 1 to n, n being the number of original labels and null_label_id is assigned to new labels when tokens are padding, separator and classification tokens.

This data is trained on fine-tuned BertForTokenClassification model to predict the name of each entity. It includes token level classification layer on the top of BERT model architecture. This linear layer helps the model to map each encoded token in the sequence to a label. So, our model performs token level classification by applying a softmax activation function over the output logits for each token. This gives predicted probabilities for each token belonging to each entity.

The hyperparameters used during training include AdamW optimizer with learning rate of 5 × 0.00001 and batch size 32. The get_linear_schedule_with_warmup is used as learning rate scheduler. During the training process for 4 epochs, loss reduced from 0.24 to 0.12 for MIT restaurant dataset and from 0.42 to 0.09 for MIT movie dataset.

Model 2:

Then we experimented with RoBERTaForTokenClassification model, a robustly optimized variant of BERT model. Here, RoBERTaTokenizer was used for tokenization of reviews for both datasets. It has the same properties as BERTTokenizer with an additional tokenization step for lowercasing and handling special tokens. And the convergence during fine-tuning of RoBERTa model is faster when compared to BERT model. So, it has better generalization capabilities and handles diverse token classification scenarios.

We used the same hyperparameters as the previous model. During training process for 4 epochs, training loss was improved to 0.06 from 0.15 for MIT Movie dataset and for MIT Restaurant dataset. Training loss is monitored for every batch in the dataset.

Model 3:

To analyze the behavior of different models on our datasets, we then trained the MIT Movie and MIT Restaurant datasets on AlbertForTokenClassification model. Like previous models, this model also has a linear token classification layer on the top of ALBERT state-of-the art model. Here AutoTokenizer was used for tokenizing the input sentences. And similar steps were adapted for creating the labels. The number of parameters in the ALBERT model were reduced while maintaining the performance. So it is faster and efficient to train when compared to BERT model.

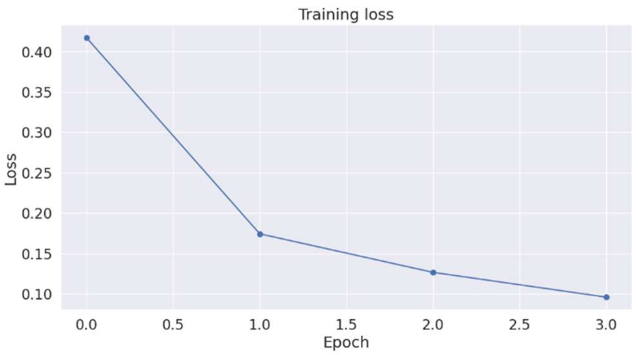
Hyperparameters were kept constant but the tokens and tokenization was different. During the training of ALBERT model for 4 epochs, loss reduced from 0.16 to 0.06 for both the MIT movie dataset and for MIT restaurant datasets.

Results

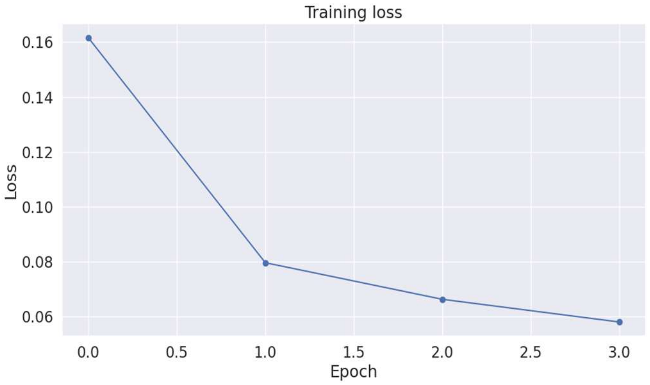
We evaluated our approach on MIT Movie and MIT restaurant datasets with F1 scores as metrics. We then compared the performances of fine-tuned BERT, RoBERTa and Albert models. Here, tokens of input sentences are considered as features for predicting different entities in the sentences.

After fine-tuning BERT, RoBERTa, Albert models for token classification, we compared the training performance for each of the models with each dataset. Both the datasets are from different areas with different entities or labels. We observed that the performance of both the datasets are similar with respect to models. This is because the models with their pre-trained language representations and fine-tuning have the ability to generalize well on named entity recognition tasks.

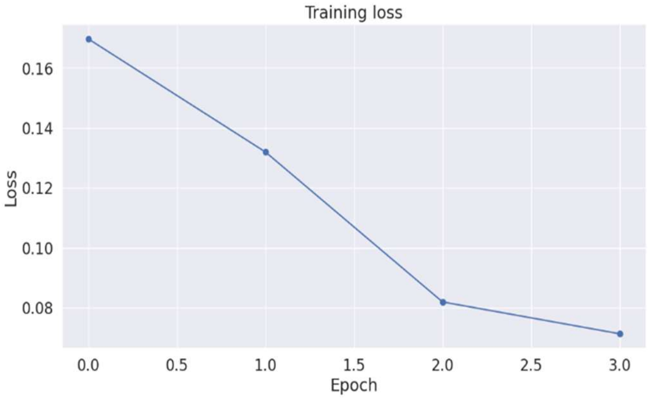
BERT on MIT Movie dataset:



RoBERTa on MIT Movie dataset:



Albert on MIT Movie dataset:



From the above figures, it can be said that RoBERTa and ALBERT performed similarly and better than the BERT model. At the fourth epoch , training loss of BERT, RoBERTa and Albert models on MIT movie dataset are 0.10, 0.05 and 0.05. The training loses for the other dataset are 0.12,0.05,0.05.

Finally, we were able to achieve good F1 scores for both the datasets with different models. With our first dataset, we found that used the ALBERT and RoBERTa models for named entity recognition resulted in an F1 scores of 98.63 , 98.72 .This was the significant improvement compared to using BERT model which only achieved an F1 score of 94.47. MIT Restaurant dataset also followed the same trend for BERT, RoBERTa, Albert models with F1 scores 91.90, 99.20,98.91. The RoBERTa and Albert models were able to remove the next sentence prediction objective used in BERT which focused solely on masked language modelling. This led to achieving higher parameter efficiency. And the broader pre-training data of the two models enhanced their ability to handle more complex data and to give generalized results. On the other hand, BERT model gave less generalized labels and had slower training time.

Overall, Albert and RoBERTa models with their optimization techniques demonstrated better performance when compared to BERT model.

Conclusion

In this paper, we compared the performance of BERT, RoBERTa and ALBERT on two different datasets. We used MIT movie dataset and MIT restaurant dataset to compare the three models. These models with their powerful language representation capabilities, have significantly enhanced the performance of Named Entity Recognition(NER) tasks on the datasets.

Further work could include incorporating the CoNLL-2003 dataset and evaluate the performance of the models on this dataset. Though, we tried implementing it due to the license constraints we were unable to work on the dataset.

References

ArXiv Paper

Devlin, J.; Chang, M.; Lee, K.; Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805

ArXiv Paper

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Veselin, S. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv: 1907.11692

ArXiv Paper

Lan, Z.; Chen, M.; Goodman, S.; Gimpel, K.; Sharma, P.; Soricut, R. 2019. ALBERT: A Lite BERT for Self-Supervised Learning of Language Representations. arXiv: 1909.11942

Website or online resource

HuggingFace. 2016. BERT Tokenizer. https://huggingface.co/docs/transformers/model_doc/bert.

Website or online resource

HuggingFace. 2016. RoBERTa Tokenizer. https://huggingface.co/docs/transformers/model_doc/roberta.

Website or online resource

HuggingFace. 2016. Albert-base-v2. https://huggingface.co/albert-base-v2.

Website or online resource

Github. 2023. Named Entity Recognition Using BERT with Pytorch. https://github.com/Kanishkparganiha/Named-Entity-Recognition-using-BERT-with-PyTorch.

Website or online resource

pytorch. 2022. Runtime error: CUDA error. https://discuss.pytorch.org/t/runtimeerror-cuda-error

Website or online resource

Github. 2022. Comparison of BERT and ALBERT architectures. https://github.com/christianversloot/machine-learning-articles/blob/main/albert-explained-a-lite-bert.md.

Website or online resource

Medium. 2022.Everything you need to know about ALBERT, RoBERTa, DistilBERT. https://towardsdatascience.com/everything-you-need-to-know-about-albert-roberta-and-distilbert-11a74334b2da

Website or online resource

Medium. 2021.Exploring BERT variants (Part 1): ALBERT, RoBERTa, ELECTRA. https://towardsdatascience.com/exploring-bert-variants-albert-roberta-electra-642dfe51bc23

Website or online resource

SKIM AI. 2021.How to fine-tune BERT for NER. https://skimai.com/how-to-fine-tune-bert-for-named-entity-recognition-ner/