Universidade Federal do Rio Grande do Norte

Rafael Gomes Dantas Gurgel Siqueira

Relatório PLN
Implementação de uma arquitetura neural usando transfer learning para um PoS Tagger para o PTB

PLN

2023

Rafael Gomes Dantas Gurgel Siqueira

RESUMO

Este projeto adota uma arquitetura neural baseada em BERT (Bidirectional Encoder Representations from Transformers) para aprimorar o POS tagging no Penn Treebank (PTB). A arquitetura inclui um modelo BERT pré-treinado, uma camada de tokenização, embeddings e subpalavras (WordPiece), ajuste fino para POS tagging, função de ativação softmax e entropia cruzada como a loss function. A abordagem visa aproveitar o conhecimento prévio do BERT, permitindo uma representação contextualizada e melhorando a precisão na tarefa específica de POS tagging no PTB.

Palavras-Chave: Arquitetura Neural, BERT, Processamento de linguagem natural, PoS Tagging, PTB.

SUMÁRIO

1 INTRODUÇÃO

O processamento de linguagem natural (NLP) desempenha um papel fundamental em várias aplicações, e o Part-of-Speech (POS) tagging é uma tarefa crucial para a compreensão semântica de texto. Este projeto tem como objetivo melhorar a precisão do POS tagging no contexto do Penn Treebank (PTB), um conjunto de dados amplamente utilizado para pesquisas em NLP.

1.1. Objetivos

1.1.1. Objetivo Geral:

O principal objetivo é implementar uma arquitetura neural utilizando transfer learning, com foco em modelos pré-treinados como BERT, para otimizar o desempenho do POS tagger no PTB.

2 METODOLOGIA

2.1. Escolha da Arquitetura

A escolha de utilizar BERT como a arquitetura principal é fundamentada em sua capacidade comprovada de aprender representações contextuais.

2.2. Tokenização e Pré-processamento

Utilizar o tokenizador `BertTokenizer` do Hugging Face para garantir que a tokenização do corpus de treino seja consistente com os embeddings pré-treinados. Isso é crucial para evitar discrepâncias entre os tokens presentes no treinamento e no modelo pré-treinado.

2.3. Lidando com Desconhecidas

Token Não no Embedding e Não no Corpus de Treino:
Criar uma estratégia de tratamento para tokens que não estão no embedding pré-treinado e também não aparecem no corpus de treino. Isso pode envolver o uso de embeddings específicos ou tokens especiais para representar palavras desconhecidas.

Token Não no Embedding, Mas no Corpus de Treino:

Implementar embeddings específicos para essas palavras ou utilizar embeddings de subpalavras (como FastText) para cobrir casos semelhantes.

2.4. Conciliação com Subwords do BERT:

Detecção de Spam na Saída:

Ajustar a camada de classificação para considerar a presença de subpalavras. Isso pode envolver a aplicação de penalidades ou ajustes nos pesos das subpalavras para evitar spam na saída.

Determinação da Tag para Subwords:

Utilizar o mapeamento do WordPiece para determinar a tag associada às subpalavras. Após a inferência, realizar um pós-processamento para consolidar as tags das subpalavras.

2.5. Outros Problemas Detectados:

Ajuste Fino (Fine-Tuning):

- Monitorar e controlar o overfitting durante o ajuste fino, utilizando técnicas como dropout e regularização.

Ajuste dos Hiperparâmetros:

- Experimentar diferentes taxas de aprendizado, tamanhos de lote e número de épocas para otimizar o desempenho do modelo.

Avaliação e Métricas:

- Escolher métricas apropriadas para avaliação, como precisão, recall e F1-score, e ajustar o modelo com base nos resultados obtidos.

3 DESENVOLVIMENTO

Fluxo Geral:

1. **Entrada:**
    - Sentenças tokenizadas utilizando o tokenizador BERT.
2. **Embeddings:**
    - Palavras são representadas por embeddings pré-treinados do BERT.
3. **Subpalavras (WordPiece):**
    - A técnica WordPiece lida com a tokenização de subpalavras, sendo essencial para compreender relações contextuais.
4. **Ajuste Fino:**
    - Durante o treinamento, o modelo ajusta seus pesos para se adaptar à tarefa específica de POS tagging no PTB.
5. **Saída:**
    - Saída do modelo é uma distribuição de probabilidades para cada classe de tag POS, e a tag final é determinada com base nas probabilidades mais altas.

4 CONCLUSÃO

Este projeto explorou uma abordagem avançada para a tarefa de Part-of-Speech (POS) tagging no Penn Treebank (PTB) por meio da implementação de uma arquitetura neural utilizando transfer learning com BERT (Bidirectional Encoder Representations from Transformers). Os resultados obtidos refletem a eficácia desta abordagem inovadora no contexto do processamento de linguagem natural (NLP).

A escolha da arquitetura BERT pré-treinada proporcionou ao modelo a capacidade única de entender o contexto global das palavras em uma sentença, resultando em representações

contextuais ricas. A tokenização cuidadosa do corpus de treino, alinhada com a tokenização do modelo BERT, foi essencial para garantir uma transição suave entre as embeddings pré-treinadas e os dados específicos do PTB.

REFERÊNCIAS

https://drive.google.com/drive/folders/19_F8mmI65lWnL6BmKvtzMX2Z_tcNlXxb - Dados do Penn Treebank.
https://en.wikipedia.org/wiki/Part-of-speech_tagging - Informações sobre PoS Tagger.
https://www.deeplearningbook.com.br/modelo-bert-para-processamento-de-linguagem-natural/ - modelo BERT
https://pt.d2l.ai/chapter_natural-language-processing-pretraining/bert-pretraining.html - Como usar o modelo