Predição da Próxima Palavra na Língua Portuguesa Utilizando Arquitetura Baseada no Transformer

Caio Pinho e David Araújo
Departamento de Informática
Universidade Federal do Espírito Santo
Vitória, Brasil
{caio.pinho & david.araujo}@edu.ufes.br

***Resumo*—Este artigo tem como objetivo apresentar o processo da predição da próxima palavra com textos na língua portuguesa através da arquitetura Transformer que utiliza redes neurais profundas.**

Palavras-chave—transformer, PLN, deep learning, BERT

I. Introdução

O Processamento de Linguagem Natural (PLN) vem sendo utilizado e aplicado em vários contextos, principalmente quando se trata das tarefas de tradução de texto, legenda de imagem, legendagem de vídeo, chatbots, sumarização de texto entre outras e o nos últimos anos o PLN teve uma grande evolução com a utilização da rede neural *Long Short-Term Memory* (LSTM). No entanto, no ano de 2018, a empresa Google, lançou uma nova tecnologia chamada Transformer[1] e essa tecnologia tem apresentado resultados eficientes nas tarefas do tipo sequência a sequência, ou seja, aquelas que recebem uma sequência de dados e produzem outra.

O Transformer em PNL é uma arquitetura recente que visa resolver tarefas do tipo sequência a sequência, enquanto lida com dependências de longo alcance com facilidade, pois, essa arquitetura trabalha com a sentença inteira, ao contrário de outras tecnologias que trabalham palavra por palavra. O transformer se baseia inteiramente na autoatenção para computar as representações de sua entrada e saída sem usar *Recurrent Neural Network* (RRN) alinhados em sequência ou convolução, a arquitetura transformer tem ganhado relevância no mundo de PNL devido sua eficiência e de acordo com a literatura aos poucos substituindo outras arquiteturas tradicionais nestas tarefas, ou seja, as RNNs.

Depois do lançamento da arquitetura transformer, surgiram novas arquiteturas derivadas dele e muitas delas já foram implementadas algumas funcionalidades necessárias para o devido funcionamento do transformer, uma dessas arquitetura é o *Bidirectional Encoder Representations from Transformers* (BERT). Ao longo deste artigo é apresentado um processo de predição de palavras utilizando textos na língua portuguesa como entrada de dados para uma rede neural baseada na arquitetura Transformer.

II. Trabalhos Correlatos

Existem diversos trabalhos correlatos utilizando Transformer para PLN, porém a maioria deles está relacionado à tarefa de tradução de texto entre 2 idiomas distintos. Em [1], os autores apresentam pela primeira vez a proposta de uma nova arquitetura de rede neural, o Transformer, baseada somente em mecanismos de atenção, dispensando completamente recorrências e convoluções. Em [2], é apresentado o modelo BERT. Enquanto o Transformer original é composto de um encoder (para codificar o texto de entrada) e um decoder (para realizar a predição), BERT possui apenas o encoder. Na prática, BERT é um modelo pré-treinado para PLN que possui várias camadas do Transformer. Após o modelo ser pré-treinado, BERT pode ser aplicado em diversas tarefas de PLN como classificação, responder perguntas, classificação de sentimentos, reconhecimento de entidades, etc. Um diferencial em relação ao BERT é que ele foi pré-treinado com um corpus de texto com mais de 33 milhões de itens.

III. Metodologia

Neste trabalho foi utilizado um modelo de rede neural profunda pré-treinada, em português do Brasil, conhecido como "BERTimbau" [3], disponível em uma versão mais simples e outra mais completa, ambas treinadas utilizando o Corpus BrWaC (Brazilian Web as Corpus) [4], que é composto por 2,7 bilhões de *tokens*. Utilizamos o modelo mais simples (BERT-Base), cuja rede neural possui 12 camadas, 768 para *embedding size* e 110 milhões de parâmetros.

O BERT é um modelo de caráter bidirecional, ou seja, ele utiliza o contexto da esquerda e da direita para o entendimento das palavras e no processo de modelar uma linguagem com essa arquitetura, é passado uma lista de palavras, um texto, como entrada, essas palavras são convertidas para números e logo após a arquitetura retorna a melhor representação matricial possível das palavras, ou seja, cada palavra é representada por uma matriz específica e isso basicamente seria o *embedding* tendo como retorno uma lista de vetores.

Durante o processo é inserida uma máscara ao final do texto original, na posição onde estaria a palavra a ser predita pelo modelo. Esse procedimento é necessário para o tokenizador preparar os textos inseridos e serem utilizados com o modelo pré-treinado. Nessa parte do processo de tokenização do texto, o mesmo é convertido em uma sequência de IDs (inteiros), usando a tokenização e o vocabulário do modelo. Na etapa de predição, é adicionada uma camada de classificação após a saída da etapa anterior. Em seguida, é calculada a probabilidade de cada palavra no vocabulário ser a palavra que foi omitida por uma máscara, através do uso do softmax. Durante a fase de desenvolvimento foi incluído um parâmetro na chamada da função que faz a predição da próxima palavra sendo possível selecionar quantos registros de maior probabilidade

devem ser retornados. O código do programa que foi construído pode ser consultado em https://github.com/caio-pinho/Next_Word_Prediction.

IV. Experimentos

Foram realizados diversos testes e experimentos após a consolidação do modelo. Os experimentos foram feitos utilizando fragmentos de frases e trechos de parágrafos e solicitando a geração das 5 mais prováveis próximas palavras. Foi então avaliado a coerência e o contexto das palavras sugeridas pelo algoritmo em comparação com os trechos utilizados como entrada no modelo.

Foram feitos também alguns experimentos que contemplaram o treinamento completo de uma rede, porém devido à restrição de recursos computacionais para treinamento da rede utilizando um conjunto de dados satisfatório, não foram obtidos bons resultados com os conjuntos de dados que foram possíveis utilizar para treinamento (conjuntos de dados extremamente menores do que os modelos pré-treinados).

V. Resultados

Os resultados obtidos foram surpreendentemente bons. Em todos os casos testados, as predições foram coerentes com o contexto inserido e com o idioma português do Brasil. Não houve casos em que o modelo retornasse algum tipo de erro ou que não conseguisse fazer uma predição. A critério de exemplo, foi utilizado o trecho “Dinheiro não traz” e a primeira palavra predita foi “felicidade”, em total coerência com os 99.000 resultados encontrados no Google para a frase “Dinheiro não traz felicidade”.

Utilizando outro exemplo, ao realizar a predição da próxima palavra para o trecho “Há males que vem para o”, a palavra com maior probabilidade retornada foi “bem”.

Ao solicitar a palavra com maior probabilidade de ser utilizada após o trecho “Água mole em pedra dura tanto bate até que” foi retornada a palavra “fura”.

Foram realizados diversos outros experimentos e os resultados em todos os casos foram satisfatórios, permitindo que o modelo BERT seja utilizado amplamente para o objetivo para o qual foi construído.

Bibliografia

[1] A. Vaswani, Ashish, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in Advances in neural information processing systems, pp. 5998-6008, 2017.

[2] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805, 2018.

[3] F. Souza, R. Nogueira and R. Lotufo, "BERTimbau: pretrained BERT models for brazilian portuguese," in 9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil,2020.

[4] J. Wagner, R. Wilkens, M. Idiart and A. Villavicencio, “The brWaC corpus: a new open resource for brazilian portuguese,” 2018.