Processamento de Linguagem Natural com Redes Neurais Artificiais
PCS5029
Programa de Pós-Graduação em Engenharia Elétrica
Universidade de São Paulo

Etiquetador morfossintático com Transformer BERT

Fernando Leandro dos Santos

29 de dezembro de 2020

1 Introdução

O problema tratado neste trabalho é a etiquetagem morfossintática de sentenças em português, ou como é mais conhecido, *part-of-speech tagging* ou *POS tagging*. É apresentada uma solução utilizando utilizando o BERT [2], uma arquitetura de rede neural para modelos linguísticos baseada em *Transformers*. Em outras palavras, essa rede neural aqui apresentada é capaz de classificar palavras do português em determinadas classes morfossintáticas. Neste trabalho, utilizamos um modelo BERT pré treinado na língua portuguesa[1] e é realizado o ajuste da última camada (fine-tunning) para a tarefa específica de *POS tagging*.

Para treinamento da última camada foi necessária a utilização de um *corpus* de sentenças etiquetadas, isto é, uma grande variedade de sentenças do português com as respectivas classificações morfossintáticas para cada palavra. O *CETENFolha-1.0_jan2014.cg.gz*, utilizado neste trabalho, está disponível em `https://linguateca.pt/`. A escolha deste corpus se deu pela realização prévia de outro trabalho[3] de *POS tagging* deste autor com esse mesmo corpus. Neste outro trabalho, foram desenvolvidos etiquetadores morfossintáticos com outras metodologias: através de uma implementação do algoritmo de Viterbi e também através de uma rede neural de arquitetura bidirecional LSTM. Os detalhes dessas duas implementações estão disponíveis em outro trabalho e por isso não serão repetidas aqui. Entretanto, o intuito de utilizar a mesma base de dados é poder comparar a performance do modelo proposto aqui, o BERT com ajuste fino para POS-tagging com os modelos anteriormente propostos.

2 Implementação

O corpus CETENFolha utilizado, foi inicialmente separado e preparado em um total de 1693101 sentenças do português. Cada sentença é composta por uma sequência de palavras e a sua etiqueta morfossintática correspondente. As etiquetas possíveis de classificação são descritas a seguir, com adição de 3 tokens utilizados pela arquitetura da rede BERT, <CLS>,

<PAD>, e <UNK>, que representam inicio da sentença, preenchimento do final da sentença e tokens desconhecidos, respectivamente. Mais detalhes deste procedimento de limpeza e pré-processamento estão presentes no código [4] e no trabalho anterior a este[3].

- Etiquetas morfossintáticas: 'N', 'DET' ,'PRP', 'V', 'PROP', 'ADJ' , 'ADV', 'NUM', 'KC','SPEC','PERS', 'KS', 'IN', 'EC', 'PRON', '__PUNCT__'

A arquitetura da rede utilizada nesta solução é basicamente o modelo BERT, concatenado à uma camada linear final(com dropout) responsável por fazer a predição final para uma etiqueta. O modelo BERT é utilizado como uma camada de *embedding* bastante complexa, que interpreta cada palavra visualizando o seu contexto de palavras vizinhas (de forma bidirecional). Um detalhe importante nesse modelo é que modelo BERT pré-treinado que foi utilizado usa um tokenizador a nível de pedaços de palavras, isto é, o tokenizador pode quebrar uma palavra em mais de um token. Essa tokenização não faz sentido para a tarefa de *POS tagging* porém isso é resolvido atribuindo a etiqueta do primeiro token para os demais tokens da palavra.

Para treinamento (ajuste fino) da rede, o corpus total foi divido em treino (60%), validação (20%) e teste (20%). A avaliação final do etiquetador foi avaliada na base de teste e também em uma rodada de *5-fold cross validation* sobre o corpus completo, assim como foi realizado nos trabalhos anteriores, com o algoritmo de Viterbi e com a rede LSTM.

3 Resultados

O ajuste fino da rede foi feito sobre a base de treinamento e validação e teve um tempo de execução de aproximadamente 103 minutos por epoch. A rede foi treinada por 5 epochs somente, visto que já no último epoch a performance na base de validação começava a decair enquanto na de treino continuava crescendo, indicando um início de *overfitting*.

A performance final da tarefa na base de testes apresentou uma acurácia de 91.79%, isto é, 91.79% das etiquetas foram corretamente identificadas. Apresentamos também na tabela 1, a acurácia percentual do modelo em cada conjunto da etapa de validação cruzada, juntamente com a performance do modelo de Viterbi e da rede LSTM. Apesar de 91.79% ser uma acurácia bastante alta para um modelo de *POS tagging*, concluímos que a performance do modelo BERT aqui proposto foi significativamente inferior aos modelos anteriores. Imagina-se que esta diferença é devido à forma como o BERT tokeniza as palavras. Ao quebrar tokens em pedaços de palavras, o espaço dimensional de tokens aumenta significativamente. Além disso, tokens iguais (provenientes de palavras e etiquetas diferentes) acabam sendo atribuídos com etiquetas morfossintáticas diferentes, o que pode estar dificultando o aprendizado da rede. A confirmação desta hipótese fica como uma proposta de trabalho futuro.

4 Conclusão

Foi apresentado nesse artigo o desenvolvimento de uma solução para o problema de etiquetamento morfossintático para sentenças do português. Foi desenvolvido um modelo de linguagem utilizando uma rede neural com arquitetura BERT. A performance desse modelo foi comparada com outros dois modelos desenvolvidos anteriormente, um modelo seguindo a implementação do algoritmo de Viterbi, que utiliza um modelo oculto de Markov de ordem

Rodada	*Acurácia LSTM*	*Acurácia Viterbi*	*Acurácia BERT*	*Total de palavras*
$k = 1$	96.64%	96.27%	94.61%	5.502.802
$k = 2$	96.83%	96.27%	92.76%	5.499.002
$k = 3$	96.65%	96.28%	92.77%	5.523.073
$k = 4$	96.65%	96.29%	92.78%	5.517.860
$k = 5$	96.64%	96.26%	92.72%	5.497.557
Média	**96.682%**	**96.275%**	**93.528%**	**27.540.294**

Tabela 1: Acurácia dos etiquetadores

2 para encontrar a sequência de etiquetas mais provável para uma determinada sentença e também um modelo com arquitetura LSTM. O modelo BERT não superou a performance dos outros dois modelos, mesmo sendo treinado com a mesma quantidade de dados utilizada nos modelos anteriores. Fica comprovada aqui a boa performance de redes LSTM para a tarefa de etiquetagem morfossintática, visto que mesmo um modelo de altíssima complexidade como o BERT não superou a performance do modelo LSTM para esta tarefa, neste corpus de sentenças em português.

Referências

[1] `https://github.com/neuralmind-ai/portuguese-bert`. [Online; acessado em Dezembo de 2020].

[2] Jacob Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". Em: *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)*. Minneapolis, Minnesota: Association for Computational Linguistics, jun. de 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: https://www.aclweb.org/anthology/N19-1423.

[3] F. Santos. `https://github.com/flsantos/pos_tagging_lstm_pytorch/blob/master/Ling_Computacional_Fernando_Trabalho2.pdf`. [Online; acessado em Dezembo de 2020].

[4] F. Santos. `https://github.com/flsantos/pos_tagging_bert_fine_tunning`. [Online; acessado em Dezembo de 2020].