PNL, Pré-processamento e BERT

Luiza Barros Reis Soezima

Motivações

- Dentro dos desafios de ML e NLP, a busca por similaridade de documentos é um problema crucial, sendo útil para muitas áreas (eg. documentos legais)
- Dois documentos são ditos similares se possuem semelhanças semânticas e possuem mesmo conceito ou se são idênticos (duplicados)
- Modelos de NLP em geral precisam de datasets MUITO grandes que demandam muito tempo de processamento para o treinamento. Com isso é custoso obter-se pesos com boa acurácia, com relação tempo x esforço muitas vezes ruim

Problema:

- Atualmente uma pessoa realiza a busca de documentos.
- Precisamos de uma forma eficiente de realizar a pesquisa de documentos legais num dataset fornecido, tais que esses documentos que queremos são similares ao especificado na solicitação
- As demandas de NLP hoje em dia circulam em corpora de língua inglesa, o que muitas vezes reduz a eficiência ou ainda a aplicabilidade de técnicas de NLP para a língua portuguesa.
- Queremos levar me conta o contexto para avaliar as representações de texto (vetores) - word2vec, BERT, n-gram, GloVe...)

Pipeline

Copus	Pré- Processamento	Information Retrieval	Avaliação
Definição do corpus a ser usado, no caso usamos um corpus legislativo em língua portuguesa	1.Remoção de stopwords 2.Remoção de acentuação 3.Stemming (redução da palavra ao radical) 4.Word n-gram(seleção de contexto) 5.Vector Space Model(termo x documento) 6. Word Embedding (palavras com o mesmo significado têm representação similar semântica) - BERT, SBERT	1.Agrupamento baseado em cluster (ICIR) 2.Best Match (BM25) - Okapi, F, L, Plus	Acurácia - verificamos se o documento motivação da pesquisa aparece nos k documentos mais relevantes recuperados

Best Match - BM25

- O BM25 é uma função que rankeia um conjunto de documentos baseados na query de termos que aparecem em cada documento, independente de proximidade entre documentos.
- Faz parte das funções que recuperam informação a partir de bag-of-words.
- https://en.wikipedia.org/wiki/Okapi_BM25
- Existem variações que fazem mudanças na função do score, etc.
- Okapi - Punir documentos longos
- F - divide o documento em partes e roda o Okapi para cada parte - média ponderada
- L - muda o cálculo do IDF - evitar “preferências” por documentos mais curtos
- Plus - muda o cálculo do IDF - evitar preferência por documentos mais curtos

Vamos falar de BERT

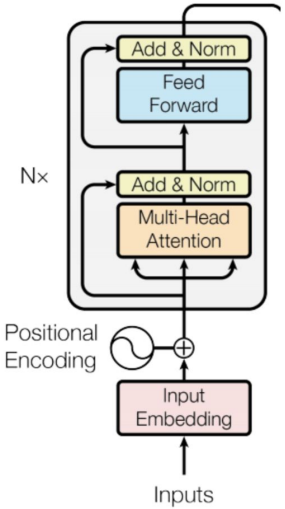
fonte:https://towardsdatascience.com/bert-based-cross-lingual-question-answering-with-deeppavlov-704242c2ac6f

BERT (Bidirectional Encoder Representations from Transformers)

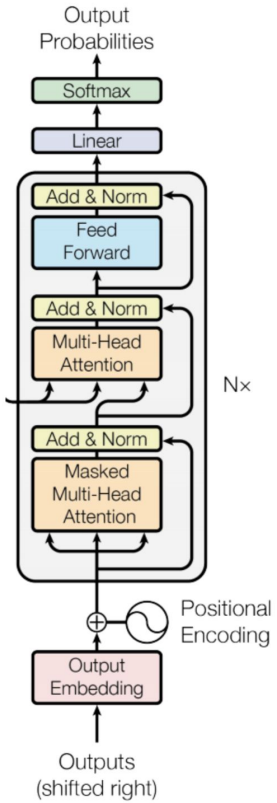
- É uma técnica para pré processamento em representações de linguagem que utilizou o Wiki corpus como treinamento. Isso gerou um modelo de pré treinamento que pode ser fine-tuned para tarefas específicas de NLP
- O BERT é um modelo de deep-learning pré treinado
- Em redes neurais profundas, tem-se camadas hierárquicas de representações que são ordenadas por complexidade.
- Para o BERT, ao invés de criar as camadas mais baixas do zero, as camadas mais baixas foram frutos de uma rede treinada e que pode ser utilizada posteriormente para tasks diferentes.

Transformers Neural Network

- Inicialmente criado para resolver problemas de tradução de linguagem (Attention is all you need)
- Antes, LSTM eram usados para resolver esses problemas, mas são lentos e word sequentials, não bidirecionais.
- Transformer são mais rápidos pois palavras são processadas simultaneamente, e o contexto das palavras são melhor aprendidos, sendo bidirecionais (context based)
- **O Transformer encoder resolve sequence to sequence task lidando com baixas dependências. O conceito de self-attention permite que o modelo utilize o contexto das palavras do input para entender a próxima palavra.**
- Criar sistemas que entendem a linguagem/ estado da arte
- Representações unidirecionais são free of context(word2vec, glove), gerando os word embeddings, mas não captura polissemia/contexto.



(a) Encoder



(b) Decoder

Transformer Flow

Encoder

- Pega os inputs (palavras) simultaneamente e gera embedding(vetores que contém o sentido de cada palavra) para cada palavra simultaneamente
- Então, palavras parecidas tem embedding/ vetores parecidos
- Aprende o contexto
- No BERT, ocorre o empilhamento de encoders e utiliza-se os embedding das camadas mais do topo.

Decoder

- Utilizam os embeddings das palavras gerados pelo encoder e as palavras alvos e usam a correção para predizer a próxima palavra
- Aprende como as palavras se relacionam

Bidirectional Language Model

-LEFT -> RIGHT

> “All of the BERT results presented so far have used the fine tuning [MASK]”

-RIGHT ->LEFT

> “[MASK} of the BERT results presented so far have used the fine tuning approach”

-BIDIRECTIONAL - tokenização intermediária

> “All of the [MASK] results presented so [MASK] have used the fine tuning approach”

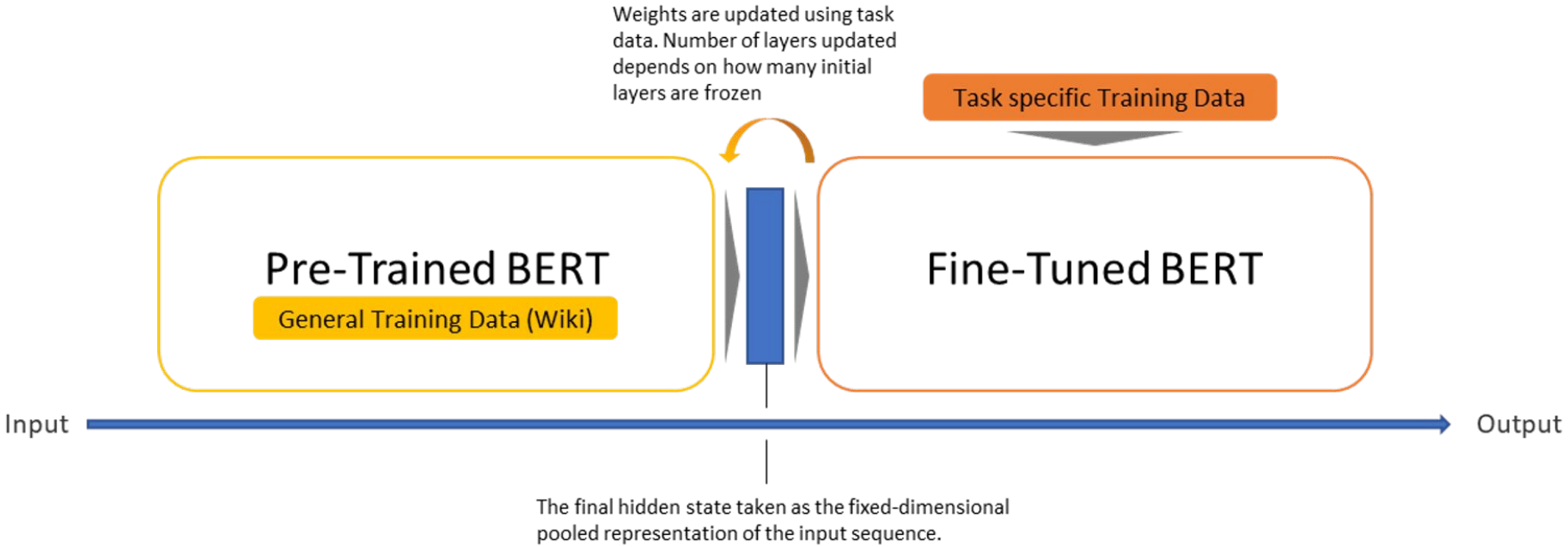
BERT Training

Pre-training

- Objetivo é entender a linguagem

Fine Tunning

- Aprender alguma task específica


Weights are updated using task data. Number of layers updated depends on how many initial layers are frozen
Task specific Training Data
Pre-Trained BERT
General Training Data (Wiki)
Fine-Tuned BERT
Input
Output
The final hidden state taken as the fixed-dimensional pooled representation of the input sequence.

BERT Pre-Training (Pass 1)

- O que é linguagem? O que é Contexto?
- BERT aprende a linguagem pelo treinamento em duas tasks não supervisionadas (MLM - Masked Language Model e NSP - Next sentence Prediction)
- No MLM pega-se uma sentença, remove-se palavras e nela colocamos máscaras [MASK] como tokens. O objetivo é dar um output para esses tokens chamados de [MASK]. Temos uma tarefa de preencher os gaps. Aqui ajuda o BERT a entender contexto bidirecional na sentença.
- No NSP o BERT pega duas sentenças e decide se a segunda sentença segue a primeira. Aqui ajuda o BERT a entender contexto em diferentes sentenças.

A sentença B segue a sentença A?

Language Model + Masked Language Model

- Language model é um modelo estatístico da probabilidade de uma sentença ou uma frase.

  P (Cachorro come maçã) > P (Maçã come cachorro)

- Mas o Masked Language Model é diferente, pois, ao invés de utilizar a probabilidade para uma frase inteira, treina-se o modelo para preencher os espaços em branco
- Essa é uma grande contribuição do paper BERT

Contextual Word Embedding

- São muito úteis pois MLM são um dos tipos de Contextual Word Embedding e podem ser utilizados como input embeddings
- No word embedding tradicional

  Os cachorros latem vs As árvores latem

  frases completamente diferentes possuem a mesma representação no embedding

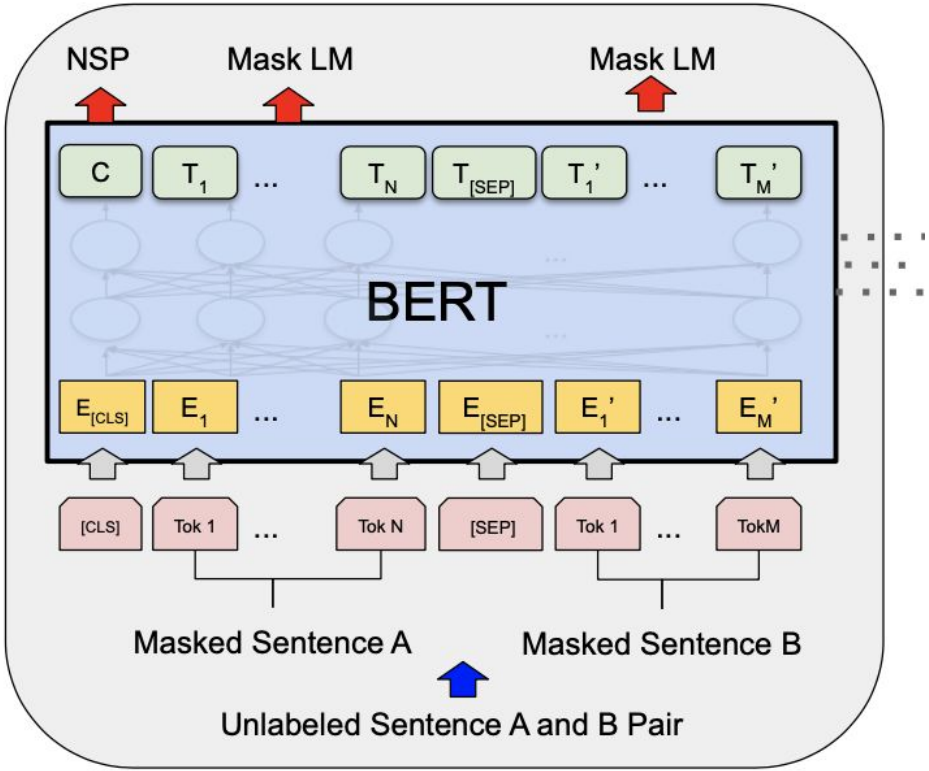
- Mas os masked models consegue criar representações diferentes para sentidos diferentes

BERT Fine Tuning (Pass 1)

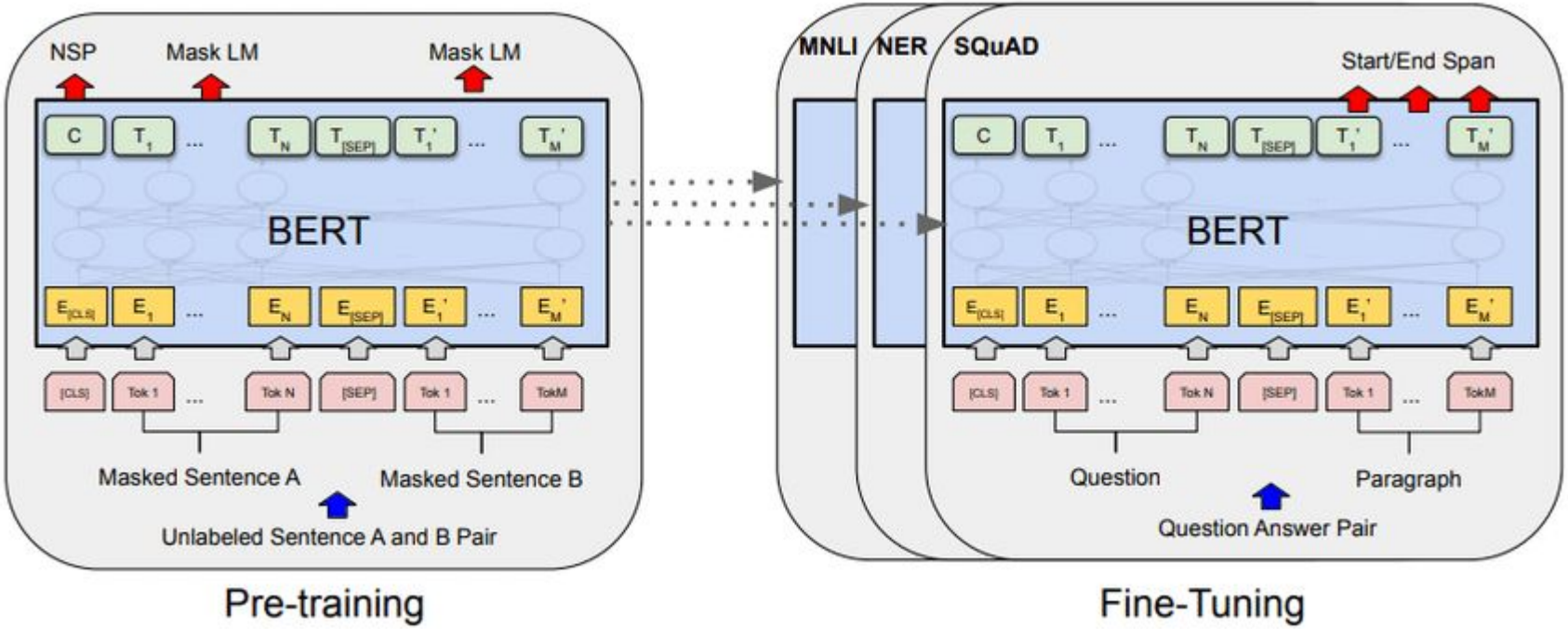
- Como usamos a linguagem para uma task específica (NLP tasks)
- Exemplo : Question & Answer
- Substituímos o output com resultados possíveis para o input, executando treinamento supervisionado usando um dataset.
- Aqui temos vantagem de tempo pois apenas o output que é aprendido do zero, o resto é sutilmente fine tuned

BERT Pre-training (Pass 2)

- No pre-training executamos MLM e NSP
- Na prática, os 2 são treinados simultaneamente
- No input temos um conjunto de 2 sentenças com palavras mascaradas {MASK], cada palavra é um token, e convertemos essas palavras em embeddings usando pre-trained embeddings
- No output temos o C como output de NSP(booleano) e word vectors que correspondem aos outputs esperados de MLM



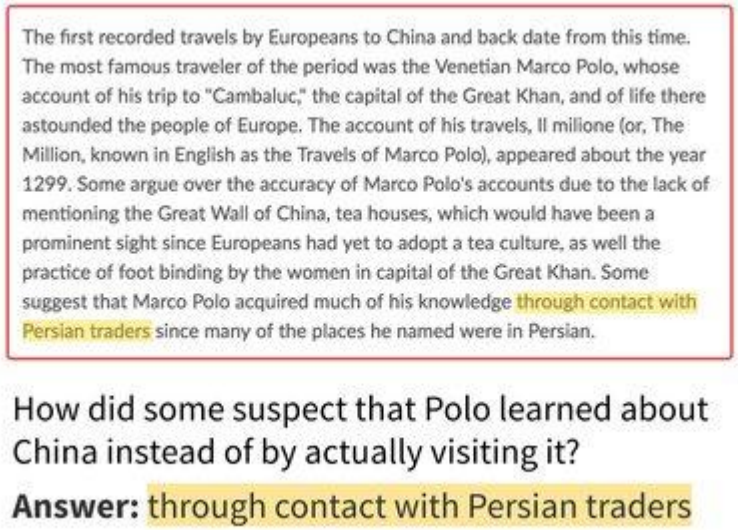
Pre-training


NSP
Mask LM
Mask LM
BERT
Masked Sentence A
Masked Sentence B
Unlabeled Sentence A and B Pair
Pre-training
MNLI
NER
SQuAD
Start/End Span
BERT
Question
Paragraph
Question Answer Pair
Fine-Tuning

BERT Fine tuning (Pass 2)

- Para o exemplo de Q&A mudamos o output para ter a resposta e o input para perguntas
- Para o contexto da IC nosso input são sentenças legislativas e o output são os documentos legislativos alvos

SQuAD - Stanford Q&A Dataset - BERT



Passage Sentence

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under gravity.

Question

What causes precipitation to fall?

Answer Candidate

gravity

- Between question and answer

cause---gravity
precipitation---gravity
fall---gravity
what---gravity

Passage Sentence

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under gravity.

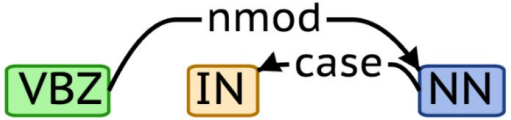
Question

What causes precipitation to fall?

Answer Candidate

gravity

- Path from passage sentence words (that also occur in question) to answer



- Combined with path from wh-word to question word.

BERT Pre - training (Pass 3)

- Como geramos os embedding a partir dos token de input
- O embedding inicial é construído a partir de 3 vetores: Token Embeddings, Segment Embedding, Position Embedding

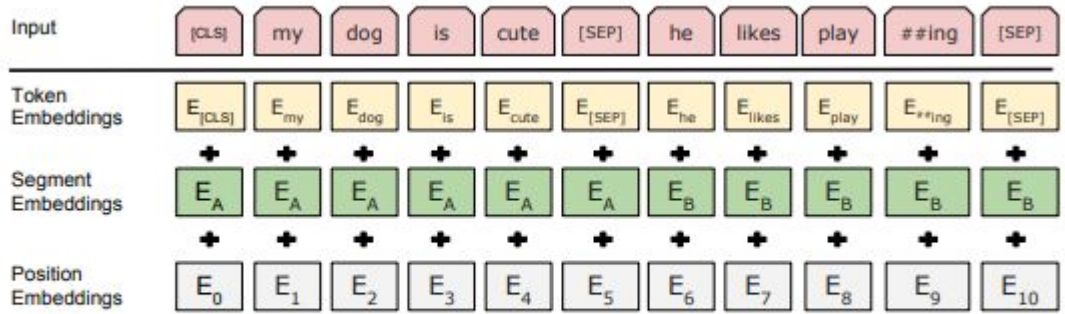
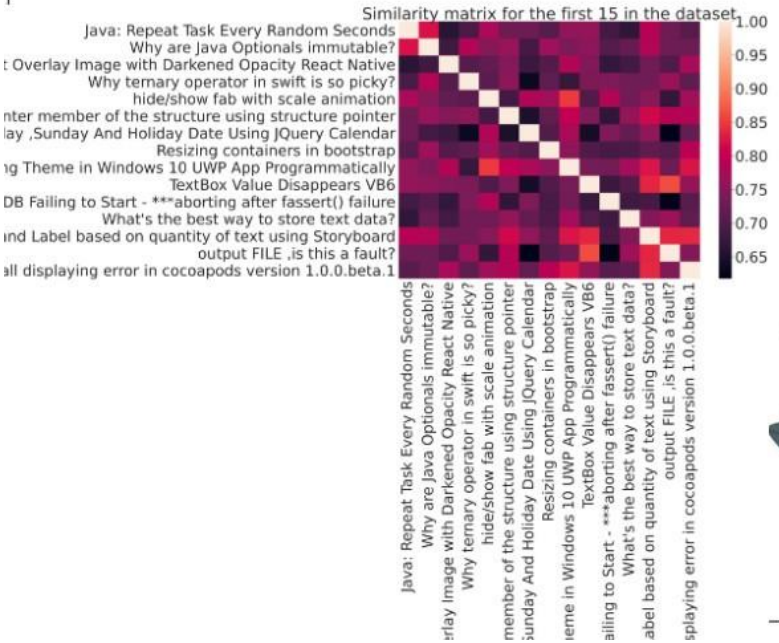


Figure 2: BERT input representation. The input embeddings is the sum of the token embeddings, the segmentation

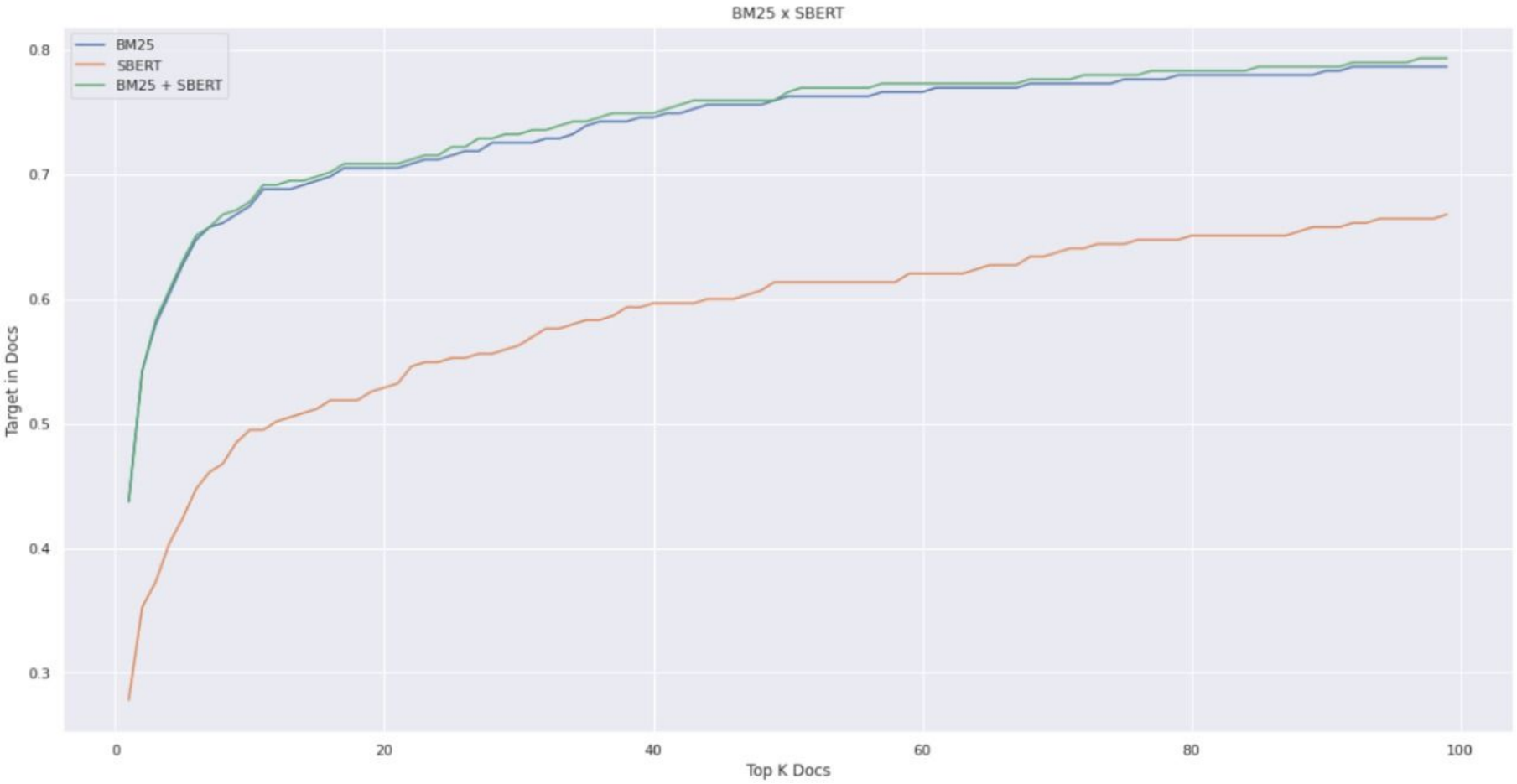
- Word Vector Ti tem mesmo tamanho
- Word Vector Ti são gerados simultaneamente
- Cada word vector é então convertido para uma distribuição (label real dessa distribuição pe um “vetorzão” codificado para a palavra) para treinar usando cross entropy loss, com o treinamento completo, BERT ganha noção da linguagem
- Stanford Q&A (squad) - 30 minutos de treinamento para fine-tune, com 91% de performance
- Modelos : BERT Base (110M parameters) vs. BERT large (340M parameters)
- AlBERTa, RoBERTa, mBERT

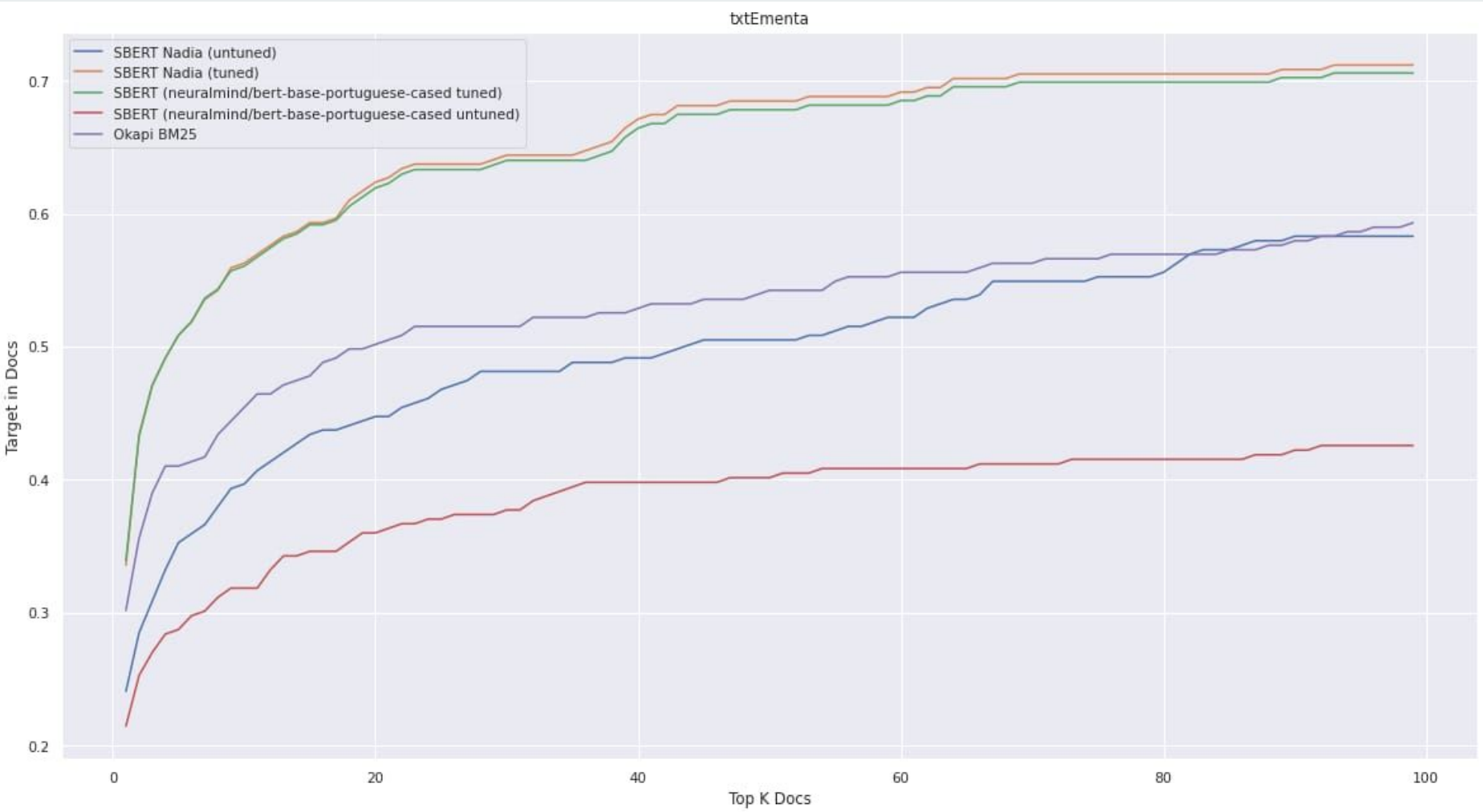
Sentence Similarity with BERT Sentence Embeddings



John Snow LABS
NLU

Alguns resultados no dataset da IC:


BM25 x SBERT
BM25
SBERT
BM25 + SBERT
Target in Docs
0.8
0.7
0.6
0.5
0.4
0.3
0
20
40
60
80
100
Top K Docs


txtEmenta
SBERT Nadia (untuned)
SBERT Nadia (tuned)
SBERT (neuralmind/bert-base-portuguese-cased tuned)
SBERT (neuralmind/bert-base-portuguese-cased untuned)
Okapi BM25
Target in Docs
0.2
0.3
0.4
0.5
0.6
0.7
0
20
40
60
80
100
Top K Docs

Próximos Passos

- Aplicação de NER (Named Entity Recognition) como forma de avaliar o corpora de forma mais eficiente e ter uma busca de documentos com maior taxa de equivalência
- Busca de técnicas de NER aplicáveis em língua portuguesa
- Testar técnicas como OpenNLP, NLKT, CRF, algoritmos de redes neurais(elmo, BERT, SBERT), algoritmos de deep learning no nosso dataset

Referências

-https://medium.com/spark-nlp/easy-sentence-similarity-with-bert-sentence-embeddings-using-john-snow-labs-nlu-ea078deb6ebf

-https://medium.com/analytics-vidhya/semantic-similarity-in-sentences-and-bert-e8d34f5a4677

- Devlin, J. et al., BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

- Sutskever, I. et. al, Sequence to Sequence Learning with Neural Networks

- Vaswani, A. et al., Attention Is All You Need

- https://www.analyticsvidhya.com/blog/2021/05/measuring-text-similarity-using-bert/

-https://www.youtube.com/watch?v=xI0HHN5XKDo&t=397s

- https://cloud.google.com/architecture/overview-extracting-and-serving-feature-embeddings-for-machine-learning?hl=pt-br

- https://rajpurkar.github.io/SQuAD-explorer/
- https://towardsdatascience.com/the-quick-guide-to-squad-cae08047ebee
- https://www.youtube.com/watch?v=-9vVhYEXeyQ&t=457s