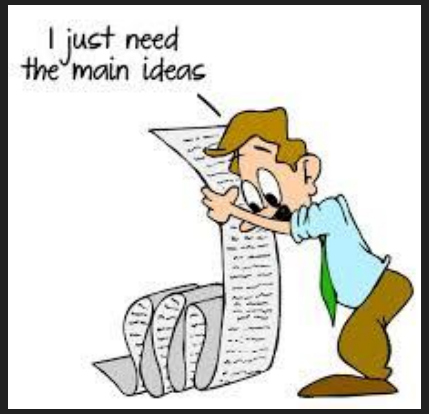
Projeto: Sumarizador e Tradutor

Insper 2020
Arthur Rizzo e Vitor Liu

O projeto: problema abordado

Fontes: youmatter.world e pngwig



Fonte: medium.com

Fonte: endpoint

O projeto: aplicação e motivação

Motivação

- Avanços de técnicas de deep-learning em PLN.
- Ganho de eficiência na análise de documentos, livros, artigos, notícias, etc.

Principais aplicações

- Seleção de informação a ser traduzida/consumida.
- Categorização de documentos em língua estrangeira.

Objetivo do projeto

Realizar a sumarização de documentos e traduzir a sumarização. Com isso é possível:

- Gerar representações resumidas de documentos (melhorando o projeto anterior)
- Disponibilizar prévias de documentos em diferentes línguas
- Ganhar eficiência no mundo da tradução e sumarização

Metodologia

1 - Pré processamento

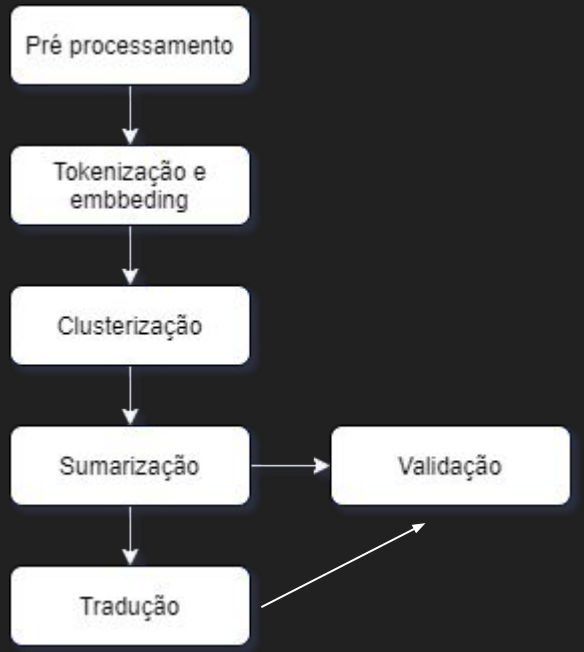
2 - Tokenização e embedding

3 - Clusterização

4 - Sumarização

5 - Resultados da sumarização

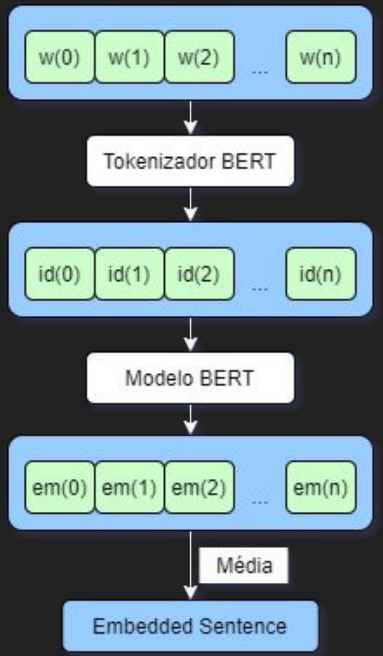
6 - Tradução



Fonte: O autor

Pré processamento, Tokenização e Embedding

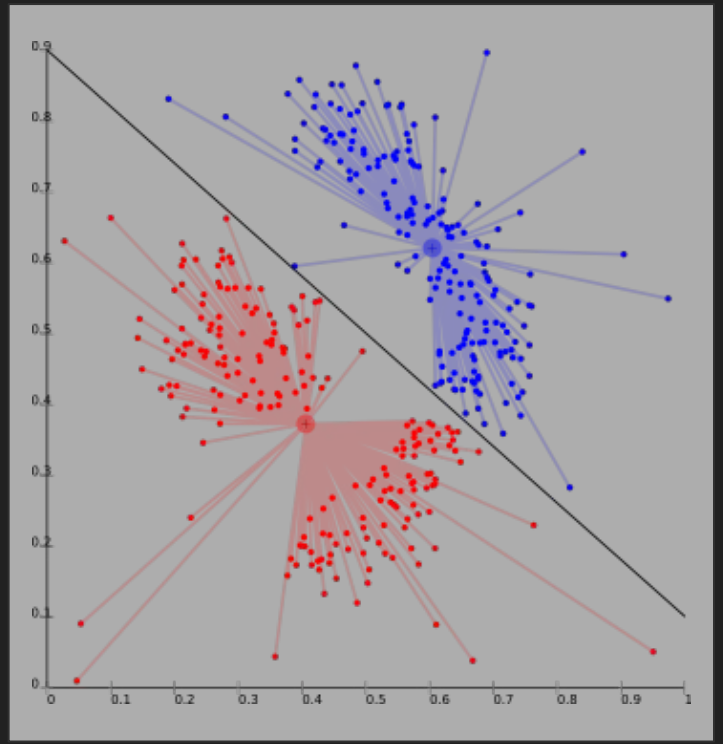
- Dataset usado: CNN stories tokenized
- Tokenizador: BERT base uncased (transformers)
- Modelo: BERT base uncased (transformers)



Fonte: O autor

Clusterização e Sumarização

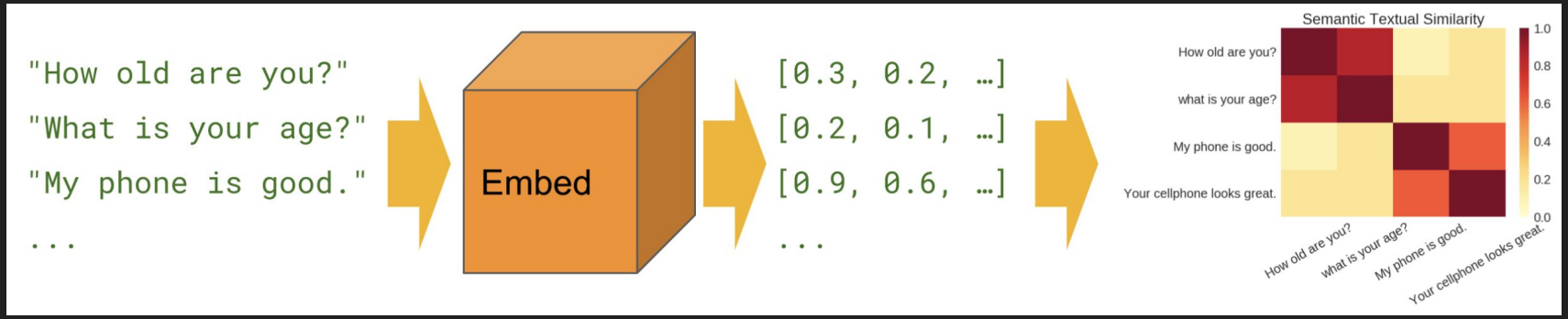
- Clusterização: MiniBatchKMeans (sklearn)
- Sumarização: Sentença mais perto do centro do cluster



Fonte: PngEgg

Validação da sumarização

- Medição da similaridade semântica entre o documento sumarizado e o “gabarito”
- Modelo: Universal Sentence Encoder (USE) - Boa performance



Fonte: TFHUB.DEV

Resultados da sumarização

Modelo usado: Universal Sentence Encode

Similaridade calculada por produto escalar ‘numpy.inner()’ (de 0 a 1)

Média de todos documentos: 0,3036

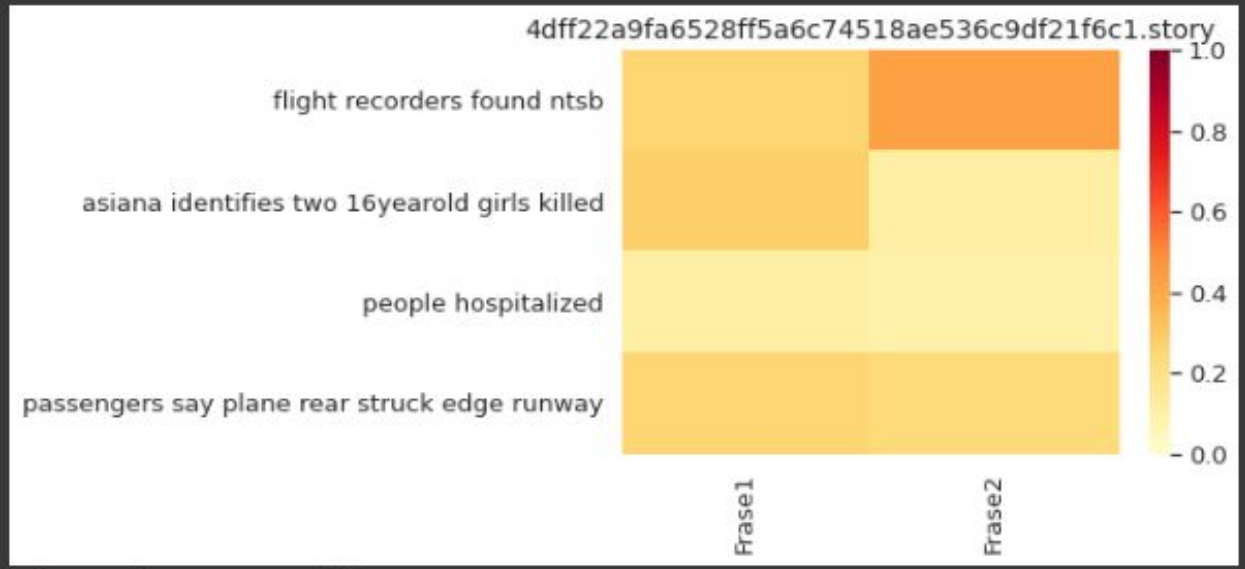
Média da maior correlação de todos documentos: 0,5571

```
@highlight

Syrian official : Obama climbed to the top of the tree , `` does n't know how to get down ''

@highlight

Obama sends a letter to the heads of the House and Senate

@highlight

Obama to seek congressional approval on military action against Syria

@highlight

Aim is to determine whether CW were used , not by whom , says U.N. spokesman
```

Exemplo highlights CNN stories tokenized, o “gabarito”

Resultados da sumarização

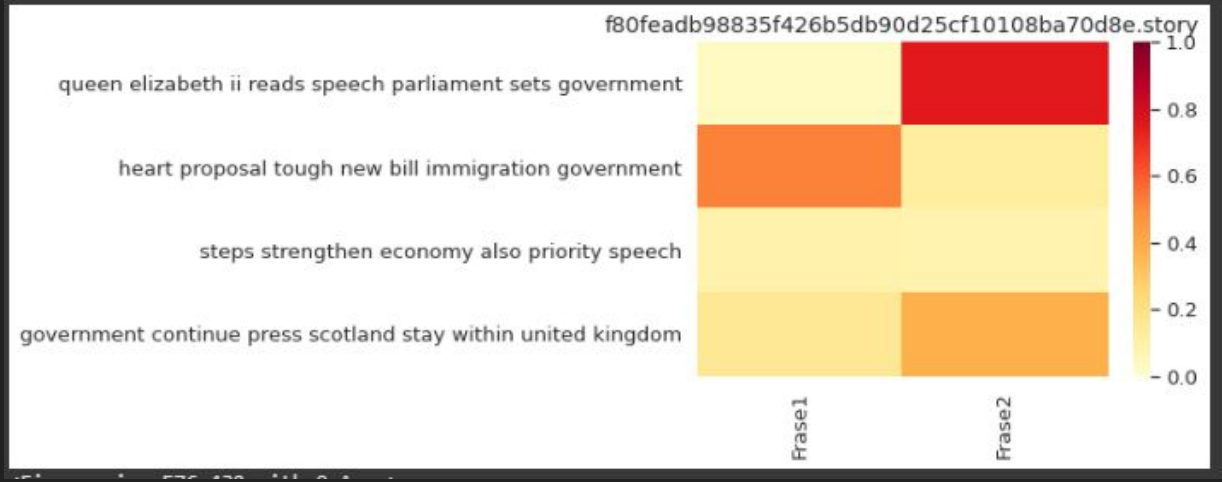
```
4dff22a9fa6528ff5a6c74518ae536c9df21f6c1.story
        Frase 1: prior saturday disaster asiana airlines endured two deadly crashes past years
        Frase 2: expert flight nt close
```



Exemplo caso médio

Resultados da sumarização

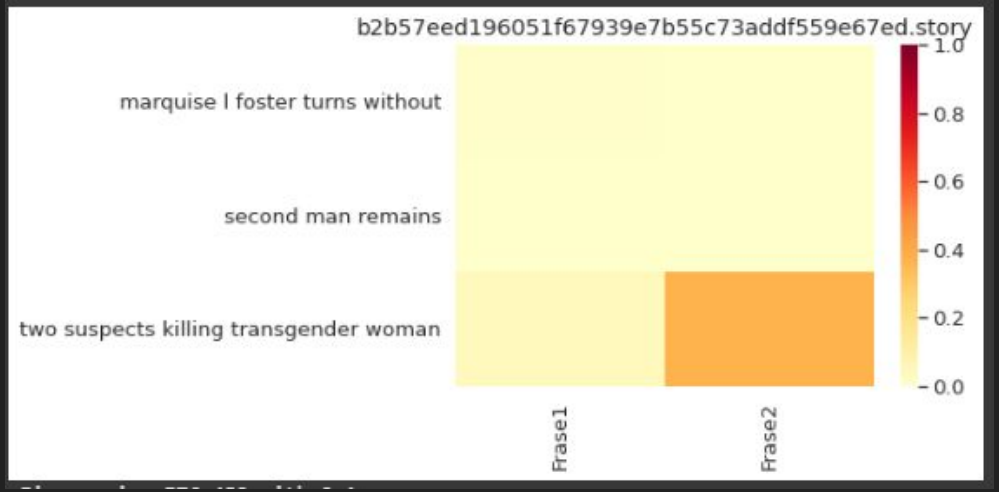
```
f80feadb98835f426b5db90d25cf10108ba70d8e.story
        Frase 1: center government legislative agenda new bill tough new measures continue immigration reform prevent illegal immigrants accessing services entitled home office said
        Frase 2: address queen elizabeth ii ceremonial state opening parliament written government although read monarch
```



Exemplo caso bom

Resultados da sumarização

```
b2b57eed196051f67939e7b55c73addf559e67ed.story
        Frase 1: special unit looks cases determine whether legal basis bring biascrime charges unit within office reviewing cases
        Frase 2: two suspects killing victoria carmen white newark fatally shot september apartment maplewood police said statement maplewood halfhour drive west new york city
```



Exemplo caso ruim

Tradução - Modelo XLM

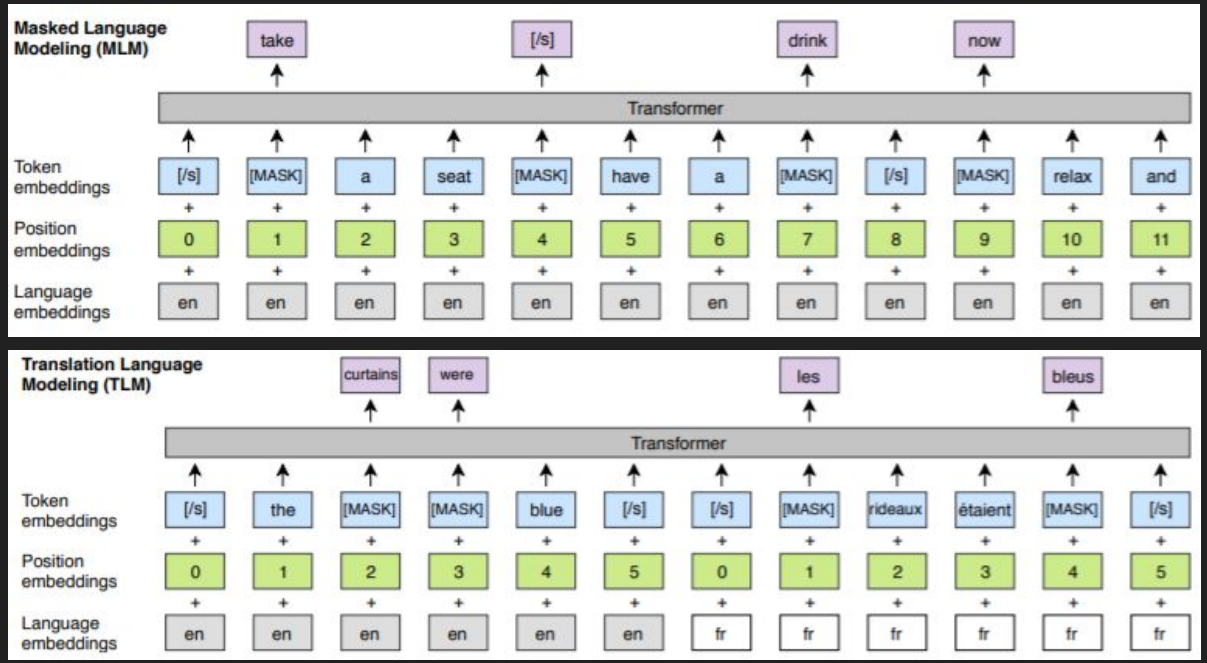
Implementação do modelo Cross-lingual Language Model Pre Training

Funcionalidades:

- Language model pre training:
  - Casual Language Model (CLM)
  - Masked Language Model (MLM)
  - Translation Language Model (TLM)
- GLUE e XNLI fine-tuning
- Supervised / Unsupervised MT training

Tradução - Modelo XLM

Treinamento MLM e TLM:



Fonte: facebookreaserch/XLM

Tradução - Modelo XLM

Vantagens:

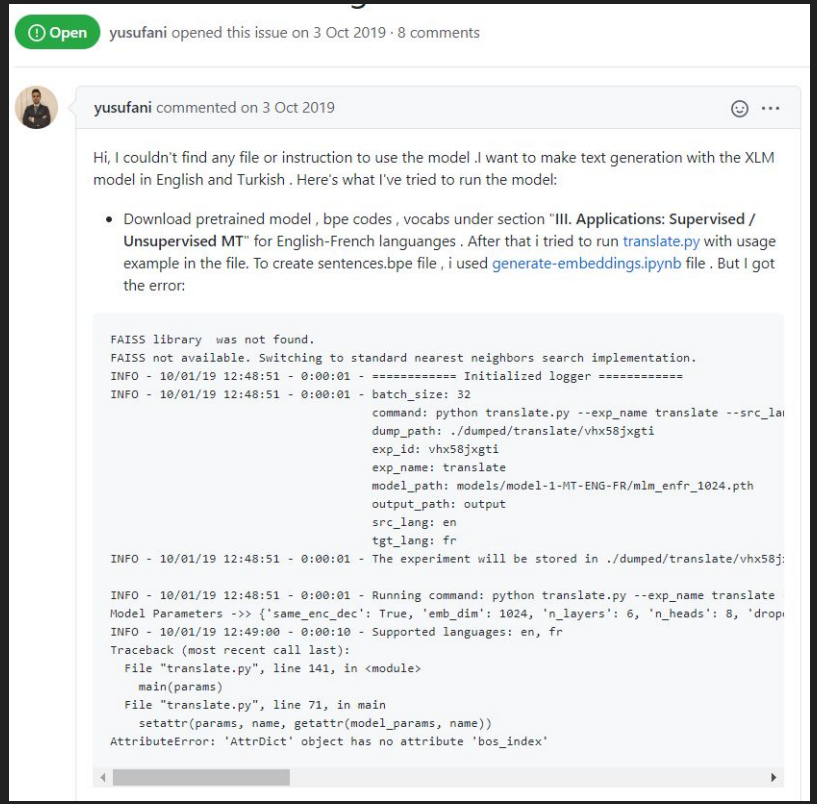
- Melhor desempenho na tradução
- Cross-lingual Language Model sem supervisão
- Possibilidade de fine-tuning

	en	fr	es	de	el	bg	ru	tr	ar	vi	th	zh	hi	sw	ur	Δ
Machine translation baselines (TRANSLATE-TRAIN)																
Devlin et al. (2018)	81.9	-	77.8	75.9	-	-	-	-	70.7	-	-	76.6	-	-	61.6	-
XLM (MLM+TLM)	85.0	80.2	80.8	80.3	78.1	79.3	78.1	74.7	76.5	76.6	75.5	78.6	72.3	70.9	63.2	76.7
Machine translation baselines (TRANSLATE-TEST)																
Devlin et al. (2018)	81.4	-	74.9	74.4	-	-	-	-	70.4	-	-	70.1	-	-	62.1	-
XLM (MLM+TLM)	85.0	79.0	79.5	78.1	77.8	77.6	75.5	73.7	73.7	70.8	70.4	73.6	69.0	64.7	65.1	74.2

Fonte: facebookreaserch/XLM

Resultados da Tradução

Não foi possível realizar a tradução devido a um problema no código disponibilizado.



Open **yusufani** opened this issue on 3 Oct 2019 · 8 comments

yusufani commented on 3 Oct 2019

Hi, I couldn't find any file or instruction to use the model .I want to make text generation with the XLM model in English and Turkish . Here's what I've tried to run the model:

- Download pretrained model , bpe codes , vocabs under section "**III. Applications: Supervised / Unsupervised MT**" for English-French languanges . After that i tried to run translate.py with usage example in the file. To create sentences.bpe file , i used generate-embeddings.ipynb file . But I got the error:

```
FAISS library  was not found.
FAISS not available. Switching to standard nearest neighbors search implementation.
INFO - 10/01/19 12:48:51 - 0:00:01 - ============ Initialized logger ============
INFO - 10/01/19 12:48:51 - 0:00:01 - batch_size: 32
                                     command: python translate.py --exp_name translate --src_la
                                     dump_path: ./dumped/translate/vhx58jxgti
                                     exp_id: vhx58jxgti
                                     exp_name: translate
                                     model_path: models/model-1-MT-ENG-FR/mlm_enfr_1024.pth
                                     output_path: output
                                     src_lang: en
                                     tgt_lang: fr
INFO - 10/01/19 12:48:51 - 0:00:01 - The experiment will be stored in ./dumped/translate/vhx58j

INFO - 10/01/19 12:48:51 - 0:00:01 - Running command: python translate.py --exp_name translate
Model Parameters ->> {'same_enc_dec': True, 'emb_dim': 1024, 'n_layers': 6, 'n_heads': 8, 'drop
INFO - 10/01/19 12:49:00 - 0:00:10 - Supported languages: en, fr
Traceback (most recent call last):
  File "translate.py", line 141, in <module>
    main(params)
  File "translate.py", line 71, in main
    setattr(params, name, getattr(model_params, name))
AttributeError: 'AttrDict' object has no attribute 'bos_index'
```

Fonte da imagem: XLM/issues/208

Considerações finais

Resultados resumidos

- Bons resultados na sumarização
- Falha na implementação da tradução

Para o futuro

- Implementar modelo de tradução com sucesso
- Melhorias na sumarização: experimentar outras técnicas

Referências

https://github.com/facebookresearch/XLM

https://arxiv.org/abs/1901.07291

https://tfhub.dev/google/universal-sentence-encoder/4 (USE)

https://github.com/google-research/bert (BERT)

https://arxiv.org/abs/1810.04805 (BERT)

https://huggingface.co/transformers/model_doc/bert.html (BERT)

https://github.com/facebookresearch/XLM/issues/208

https://www.pngegg.com/en/png-dsgwd

Perguntas??