

Resolvendo tarefas de NLP com *few/one-shot learning*

MONIQUE
MONTEIRO -
[MONIQUELOUISE@
GMAIL.COM](mailto:MONIQUELOUISE@GMAIL.COM)

Conceitos importantes

- Segundo o ChatGPT (vamos testá-lo mais uma vez 😊):

- Reconhecimento de entidades nomeadas:

- “...(NER, do inglês "Named Entity Recognition") é uma técnica de processamento de linguagem natural que envolve a identificação de entidades nomeadas em um texto. As entidades nomeadas são objetos do mundo real que possuem um nome próprio, como pessoas, organizações, locais, datas, números de telefone, entre outros. (...)”*

- Few-shot learning:

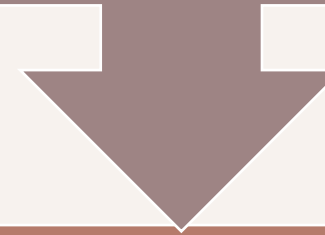
- “... é uma técnica de aprendizado de máquina que visa treinar um modelo para aprender a reconhecer novas classes com apenas algumas amostras de treinamento. (...)”*

- Zero-shot learning:

- “...é uma técnica de aprendizado de máquina que permite que um modelo aprenda a reconhecer novas classes sem a necessidade de exemplos de treinamento para essas classes específicas. (...)”*

Objetivo

Validar técnicas de few/one-shot learning para a tarefa de reconhecimento de entidades nomeadas com o uso da API do ChatGPT (GPT 3.5 Turbo)



Comparar desempenho obtido com métricas coletadas em fine-tuning anterior

Parâmetro de comparação

The screenshot shows the Hugging Face model card for `monilouise/ner_news_portuguese`. The card includes a header with the model name, a 'like' button, and a list of tags: Token Classification, PyTorch, JAX, Transformers, Portuguese, bert, ner, and AutoTrain Compatible. Below the header, there are tabs for 'Model card', 'Files and versions', and 'Community'. The main content area is divided into sections: 'RiskData Brazilian Portuguese NER', 'Model description', 'Intended uses & limitations', and 'How to use'. The 'Model description' section states that this is a finetuned version from [Neuralmind BERTimbau] and provides a link to the GitHub repository. The 'How to use' section contains a code snippet for using the model with the transformers library. On the right side of the card, there is a 'Hosted inference API' section with a 'Token Classification' dropdown, a text input field for a sentence, a 'Compute' button, and a 'JSON Output' toggle. A line graph shows the number of downloads last month, which is 595.

monilouise / **ner_news_portuguese** like 7

Token Classification PyTorch JAX Transformers Portuguese bert ner AutoTrain Compatible arxiv:1909.10649

Model card Files and versions Community

Edit model card

RiskData Brazilian Portuguese NER

Model description

This is a finetuned version from [Neuralmind BERTimbau]
(<https://github.com/neuralmind-ai/portuguese-bert/blob/master/README.md>) for Portuguese language.

Intended uses & limitations

How to use

```
from transformers import BertForTokenClassification, DistilBertTokenizerFast
model = BertForTokenClassification.from_pretrained('monilouise/ner_news_portuguese')
tokenizer = DistilBertTokenizerFast.from_pretrained('neuralmind/bertimbau-portuguese', model_max_length=512, do_lower_case=True)

nlp = pipeline('ner', model=model, tokenizer=tokenizer, grouped_entities=True)
result = nlp("O Tribunal de Contas da União é localizado em Brasília")
```

Downloads last month
595

Hosted inference API

Token Classification

Your sentence here...

Compute

This model can be loaded on the Inference API on-demand.

JSON Output

- `ner_news_portuguese`
- Fonte:
https://huggingface.co/monilouise/ner_news_portuguese
- Métricas:
 - *precision: 0.86*
 - *recall: 0.91*
 - *f1: 0.88*

Pré- processamento

Redução dos textos e do número de entidades para atender às restrições do modelo e da API

Em uma aplicação real:

- Quebra do texto em segmentos, sob pena de realização de várias chamadas

Conversão do dataset no formato esperado pelo modelo

- Doccano → texto + entidades agrupadas por tipo, sem índices
- Ref.: <https://sourajit16-02-93.medium.com/zero-shot-named-entity-recognition-using-openai-chatgpt-api-46738191f375>

Exemplo

- Seleção de 2 documentos (restrição de tamanho para exemplos)
- A API assume um formato que emula um "diálogo":

```
response = openai.ChatCompletion.create(  
    model="gpt-3.5-turbo",  
    messages=[  
        {"role": "system", "content": prompt},  
        {"role": "user", "content": "Entendeu?"},  
        {"role": "assistant", "content": "Entendi sim!"},  
        {"role": "user", "content": final_prompt}  
    ]  
)
```

- Few-shot com instrução:

```
HEADER_PROMPT = """Diga quais entidades são do tipo PESSOA, do tipo INSTITUIÇÃO  
PÚBLICA, do tipo ORGANIZAÇÃO PRIVADA ou do tipo LOCAL."""
```

Resultados

- Few-shot com instrução:

```
Recall = 0.6067663643049767  
Precision = 0.7745302713987474  
F1 = 0.6804605248044154
```

- Few-shot sem instrução:

```
Recall = 0.5741603334150527  
Precision = 0.7095980311730927  
F1 = 0.6347347809487763
```

- Zero-shot (apenas instrução)

```
Recall = 0.481000245158127
```

```
'Entidades do tipo INSTITUIÇÃO PÚBLICA: Ministério Público do Estado de Mato Gr  
osso (MPMT), Prefeitura Municipal de Rondonópolis. \n\nEntidades do tipo PESSOA  
: José Carlos Junqueira de Araújo (prefeito), Izalba Diva de Albuquerque (secre  
tária municipal de Saúde), outros dois servidores públicos, empresários proprie  
tários das empresas envolvidas no contrato para aquisição de 22 ventiladores pu  
lmonares, Jesus de Oliveira Vieira de Souza (suposto proprietário da empresa Li  
fe Med Comércio de Produtos Hospitalares e Medicamentos Eireli), Ramos de Faria  
e Silva Filho (real proprietário da empresa Life Med Comércio de Produtos Hospi  
tulares e Medicamentos Eireli).\n\nEntidades do tipo ORGANIZAÇÃO PRIVADA: empre  
sas envolvidas no contrato para aquisição de 22 ventiladores pulmonares (Life M  
ed Comércio de Produtos Hospitalares e Medicamentos Eireli e outras empresas co  
tadas na suposta cotação de preços apresentada). \n\nEntidades do tipo LOCAL: N  
ão há informações suficientes na matéria para determ... '
```

Exemplos de saída (*zero-shot learning*)

- O ChatGPT não somente retorna entidades, como também contextualiza.
 - Não é possível medir facilmente a precisão da saída:
 - Como identificar o que ele apontou como “entidade” dentro do texto de saída?
-

Análise Qualitativa

Resultado inesperado:

- O ChatGPT faz automaticamente correção ortográfica ao retornar as entidades!
- Impacto na coleta das métricas

Modo zero-shot demonstra mais “inteligência” (contextualização)

Tópico avançado



- Até que ponto ele apenas “prevê a próxima palavra”?
 - Haveria alguma etapa de parsing ou “pos-tagging”, por exemplo?
 - Múltiplas inferências pelo modelo etc.