Extractive Q&A - Performance Comparison between Learning Methods: Context and Transfer

Leonardo Augusto da Silva Pacheco e Marcus Vinícius Borela de Castro

Julho 2022

Resumo

O trabalho presente compara o desempenho de métodos de aprendizagem de transformers, transfer learning ou context learning, em atividade extrativa de perguntas e respostas. Além disso, avalia o uso de perguntas em inglês e em português, a indicação ou não de exemplos no prompt para modelos de context learning, e a aplicação de ajustes na estrutura do prompt. Foram observados desempenhos superiores em transfer learning em relação a context learning, bem como para o uso de dados em inglês em relação a português. No caso específico de context learning, melhores desempenhos com modelos maiores e com few-shot, em relação a zero-shot.

1 Introdução

A atividade extrativa de perguntas e respostas (question answering - Q&A) é aquela em que a resposta para uma pergunta é um fragmento de um texto de documentos informados. Um sistema completo de resposta extrativa pode ser dividido em duas partes: um retriever que retorna um conjunto de documentos cujo conteúdo tem relação com uma pergunta informada, usando, por exemplo, um sistema de recuperação de informação baseado em BM25; e um reader, que seleciona trechos dos documentos recebidos do retrieve que correspondem à resposta da pergunta.

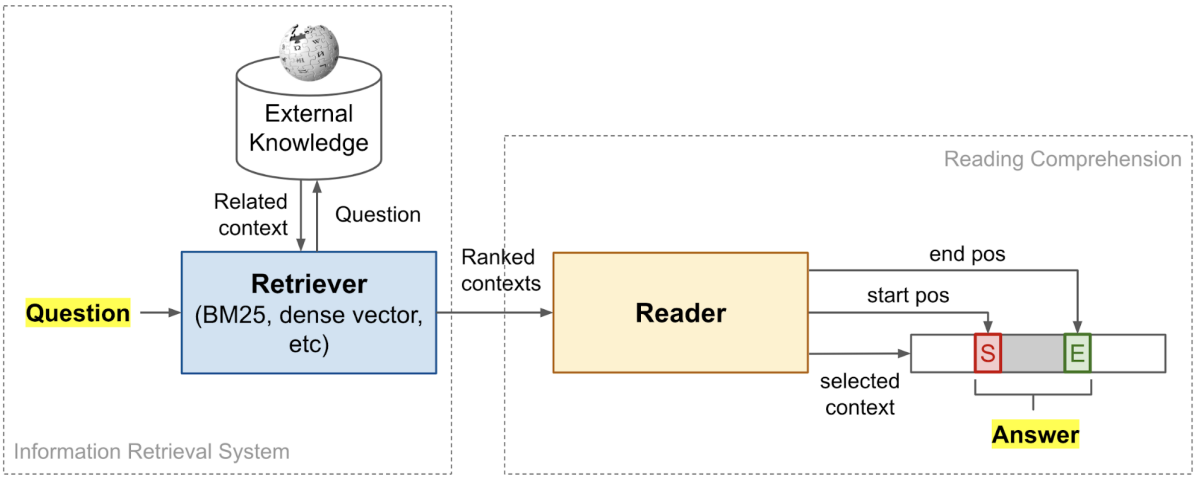

Figura 1 - O conjunto retriever-reader para Q&A. Fonte Weng(2020)

Um reader pode ser um modelo transformer que pode aprender a responder a perguntas de duas formas (learning methods): pelo ajuste de valores de seus parâmetros por meio de ajuste fino em dados de treinamento específicos para a tarefa Q&A (transfer learning) ou pelo aprendizado a partir do contexto da pergunta, com descrição e exemplos. Para context learning, empregam-se modelos de linguagem autorregressivos, treinados com grande massa de dados e com maior quantidade de parâmetros, como GPT-3 e similares.

O objetivo do presente trabalho é comparar esses dois métodos de aprendizagem de readers. Não é escopo do trabalho a etapa do retriever (Figura 1).

2 Conjunto de Dados

O conjunto de perguntas e respostas usado foi a base Stanford Question Answering Dataset (SQuAD), versão 1.1. Esta base é formada por 536 artigos e mais de 100 mil pares de perguntas e respostas, dos quais 10570 formam o conjunto de validação, ou dev set, usados na avaliação dos modelos (Rajpurkar *et al*, 2016). Os dados de treinamento correspondem a 80% da base e foram utilizados na construção dos modelos ajustados que empregamos. A versão original em inglês foi traduzida automaticamente e disponibilizada em português no repositório da Hugging Face com o nome squad_v1_pt..

3 Modelos avaliados

Para transfer learning em inglês, utilizamos o modelo distilbert-base-cased-distilled-squad, que é uma versão destilada do modelo Bert base, ou seja, um modelo reduzido alimentado por transferência de conhecimento do modelo original, ajustado para a tarefa de Q&A do SQuAD v1.1. Para transfer learning em português, utilizamos o modelo bert-large-cased-squad-v1.1-portuguese, que tem como base o BERTimbau Large, ajustado para a tarefa de Q&A do SQUAD v1.1 em português.

Para context learning, empregamos os modelos multilíngues GPT-J, com 6 bilhões de parâmetros, e GPT_Neo, com 1,3 e 2,7 bilhões de parâmetros.

4 Metodologia

Todos os modelos já haviam sido previamente treinados, e foram empregados no experimento apenas para se avaliar o desempenho nas predições de respostas. Os contextos e as questões do Squad 1.1 entram como entrada para os modelos, que geram as respostas. No caso do context learning, a entrada do modelo é o prompt, uma estrutura textual que se constitui no contexto de aprendizagem da tarefa para o modelo. No trabalho, todos prompts usados possuem uma instrução para a tarefa e, opcionalmente, alguns exemplos (shot). Se não há exemplos, chama-se de zero-shot, e few-shot caso contrário. A resposta dos modelos retornada pelos modelos é comparada com a lista de respostas esperadas para a pergunta do dataset, ambas limpas de pontuações, artigos e espaços desnecessários.

As métricas empregadas para a avaliação foram: Exact Matching (EM), que indica se a resposta gerada pelo modelo é idêntica à resposta esperada; e a F1, que se baseia na contagem de palavras comuns em relação ao total de palavras. Além disso, foram propostas e derivadas duas outras métricas: EM@3 e F1@3, que correspondem ao maior valor das anteriores calculadas sobre as três primeiras respostas diferentes geradas. A avaliação global é computada pela média aritmética das avaliações realizadas sobre as questões.

As medições calculadas foram persistidas em arquivos CSV, formando o banco de rastro, conforme sugerido por Castro e Balaniuk (2020). A estruturação das informações representa um amadurecimento da aprendizagem, além de facilitar a organização de processos e a estruturação de conceitos. A Figura 2 além de representar o conteúdo do rastro gerado apresenta os valores fixos usados para alguns parâmetros passados aos modelos (em azul). Todos os modelos foram executados em GPU NVIDIA GeForce RTX 3090, com 24GB de memória.

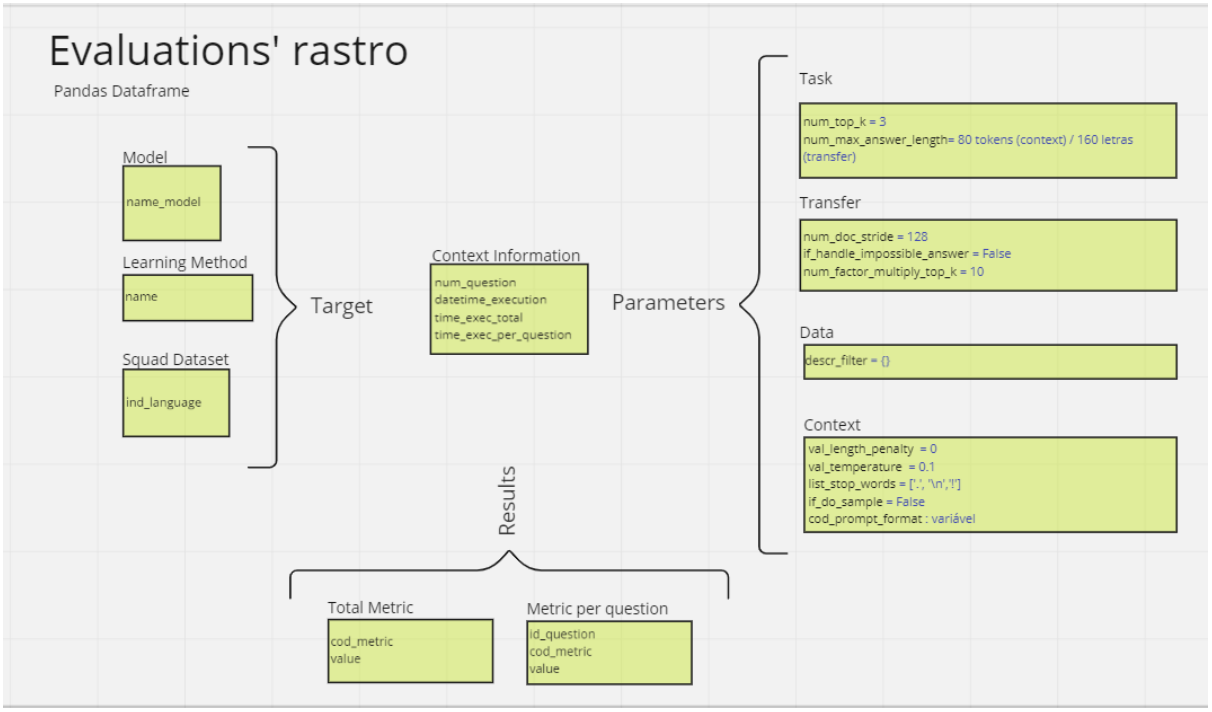

Figura 2 - conteúdo do rastro gerado e valores fixos de parâmetros passados aos modelos (em azul)

5 Resultados

Os cálculos[1] e os detalhes da análise[2] estão publicados no github do projeto. A tabela 1, a seguir, indica as médias obtidas para as métricas EM, EM@3, F1 e F1@3 em cada execução de predições dos modelos em todo o dev set.

Execução	EM	EM@3	F1	F1@3
transfer en distilbert-base-cased-distilled-squad 80 tokens	78,34%	88,00%	85,94%	91,98%
transfer en distilbert-base-cased-distilled-squad 160 tokens	78,32%	87,86%	85,92%	91,84%
transfer pt pierreguillou/bert-large-cased-squad-v1.1-portuguese 80 tokens	72,20%	84,55%	83,17%	89,74%
transfer pt pierreguillou/bert-large-cased-squad-v1.1-portuguese 160 tokens	72,14%	84,41%	83,10%	89,65%
context en EleutherAI/gpt-j-6B 80 tokens 2-shots prompt en format tpp	56,81%	64,49%	68,73%	77,83%
context en EleutherAI/gpt-j-6B 80 tokens 2-shots prompt en format tptp	56,23%	65,01%	67,67%	77,77%
context en EleutherAI/gpt-j-6B 80 tokens 1-shots prompt en	56,40%	64,70%	67,79%	77,31%
context en EleutherAI/gpt-j-6B 80 tokens 2-shots prompt en format tptp	53,51%	63,19%	64,94%	75,74%
context en EleutherAI/gpt-neo-2.7B 80 tokens 2-shots prompt en format tptp	46,19%	53,70%	58,52%	68,33%
context en EleutherAI/gpt-neo-2.7B 80 tokens 2-shots prompt en format tpp	45,07%	50,82%	58,79%	66,85%

[1] https://github.com/marcusborela/exqa-complearning/blob/main/data/vw_evaluation.csv
[2] https://github.com/marcusborela/exqa-complearning/tree/main/source/calculation/comparison

Execução	EM	EM@3	F1	F1@3
context en EleutherAI/gpt-neo-2.7B 80 tokens 1-shots prompt en	43,85%	51,25%	56,93%	66,31%
context pt EleutherAI/gpt-j-6B 80 tokens 2-shots prompt en format tptp	40,81%	45,90%	53,37%	61,37%
context en EleutherAI/gpt-neo-1.3B 80 tokens 2-shots prompt en format tptp	38,72%	45,72%	51,42%	60,68%
context en EleutherAI/gpt-neo-1.3B 80 tokens 1-shots prompt en	36,14%	42,77%	48,80%	57,94%
context en EleutherAI/gpt-neo-1.3B 80 tokens 2-shots prompt en format tpp	35,37%	40,43%	49,02%	56,69%
context pt EleutherAI/gpt-j-6B 80 tokens 2-shots prompt pt format tptp	36,62%	40,66%	48,83%	55,24%
context pt EleutherAI/gpt-j-6B 80 tokens 2-shots prompt en format tptp	29,91%	33,77%	42,37%	48,79%
context en EleutherAI/gpt-j-6B 80 tokens 0-shots prompt en	30,96%	36,38%	39,49%	47,89%
context pt EleutherAI/gpt-neo-1.3B 80 tokens 2-shots prompt en format tptp	20,05%	23,65%	30,34%	36,40%
context pt EleutherAI/gpt-neo-1.3B 80 tokens 2-shots prompt pt format tptp	19,23%	24,38%	27,17%	35,64%
context en EleutherAI/gpt-neo-2.7B 80 tokens 0-shots prompt en	16,97%	20,09%	29,38%	35,25%
context pt EleutherAI/gpt-neo-1.3B 80 tokens 2-shots prompt pt format tpp	17,51%	21,66%	25,51%	33,13%
context en EleutherAI/gpt-neo-1.3B 80 tokens 0-shots prompt en	14,42%	16,67%	27,38%	32,33%
context pt EleutherAI/gpt-neo-1.3B 80 tokens 1-shots prompt pt	14,66%	19,42%	20,97%	29,33%
context pt EleutherAI/gpt-neo-1.3B 80 tokens 0-shots prompt pt	2,55%	2,65%	15,68%	16,94%
context pt EleutherAI/gpt-neo-1.3B 80 tokens 0-shots prompt en	4,64%	5,44%	13,58%	15,98%

Tabela 1 - Lista de execuções sobre o dev set, com as médias obtidas. Estão indicados o tipo de aprendizagem, a linguagem da pergunta, o modelo empregado e o tamanho máximo da resposta. Para context-learning, indica ainda o número de exemplos (n-shots), a linguagem das instruções e o número de exemplos, e algumas indicações de formato do prompt (tpp/tptp)

Em média, as medidas de F1 são maiores que as de EM em 11%, e medições com 3 respostas são maiores do que 1 resposta em 7%. As respostas a perguntas em inglês apresentam melhor desempenho que em português, porém, em transfer learning, a vantagem é pequena, de 2% (F1@3) a 6% (EM), em média, conforme Gráfico 1.

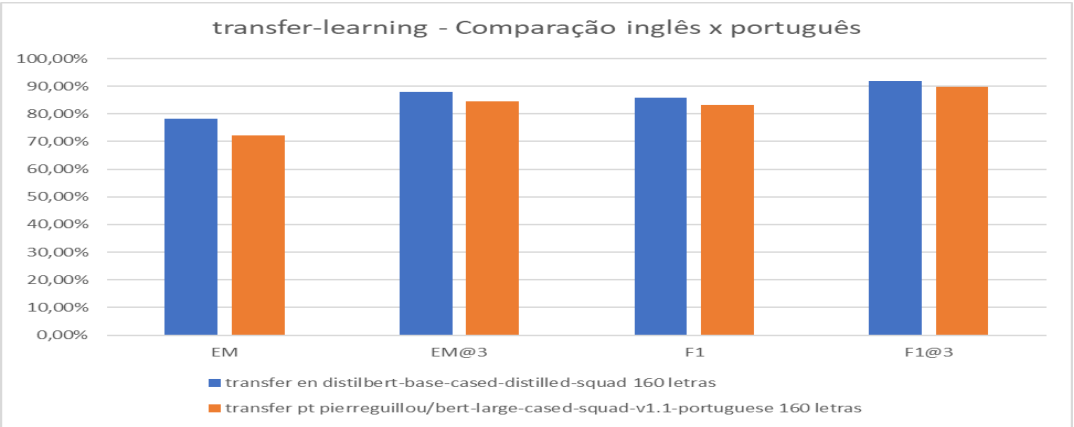

Gráfico 1 - Comparação de modelos de transfer learning em inglês e português

Em context learning, por outro lado, a vantagem é de mais de 20%, em média. Se, para as perguntas em português, as instruções e exemplos são mantidos em inglês, observamos uma melhoria média de 3 a 4% nas medições, porém, com maior desvio, conforme indicado na Gráfico 2.

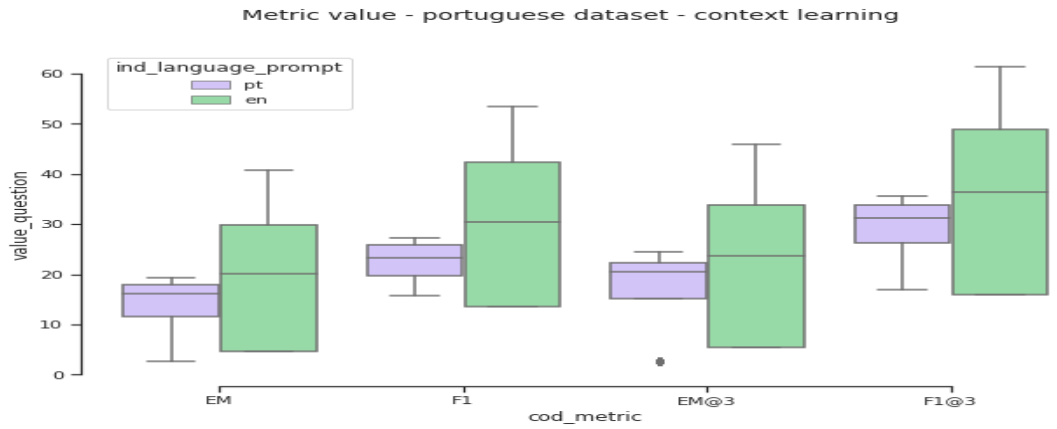

Gráfico 2 - Context learning em português, comparando instrução e exemplo em português com inglês

Modelos ajustados para o SQuAD apresentam melhor desempenho em relação a modelos em que se aplicou context learning, porém alguns modelos GPT maiores alcançaram diferenças menores: 21,52% para EM, 23,44% para EM@3, 17,20% para F1 e 14,08% para F1@3. Em português, porém, as diferenças foram bem maiores: 31,36% para EM, 38,58% para EM@3, 29,77% para F1 e 28,33% para F1@3.

Em context learning, a apresentação de exemplos (few shot) apresenta vantagens. O pior desempenho ocorreu com zero-shot, onde o modelo GPT-J alcançou EM@3 de cerca de 36% em prompts em inglês, enquanto prompts com 1-shot ou 2-shots alcançaram desempenhos próximos entre si, de cerca de 65%, conforme indicado na Gráfico 3.

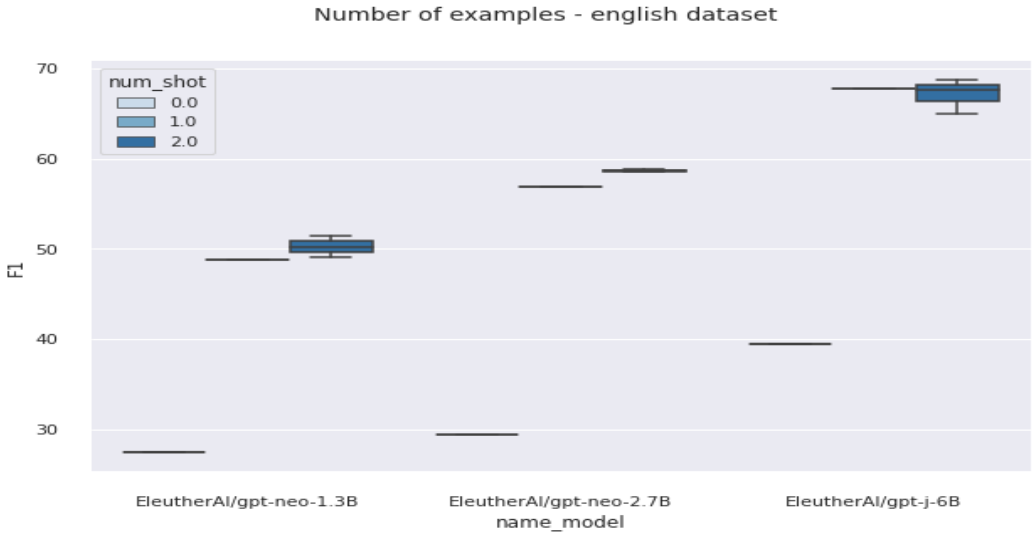

Gráfico 3 - Context learning em inglês, comparando prompts zero-shot, 1-shot e 2-shot em cada modelo

Em prompts 2-shots, quando os exemplos são apresentados em trincas contexto-questão-resposta (tptp), o resultado é levemente melhor que com um único contexto para pares de questão e resposta (tpp), conforme Gráfico 4.

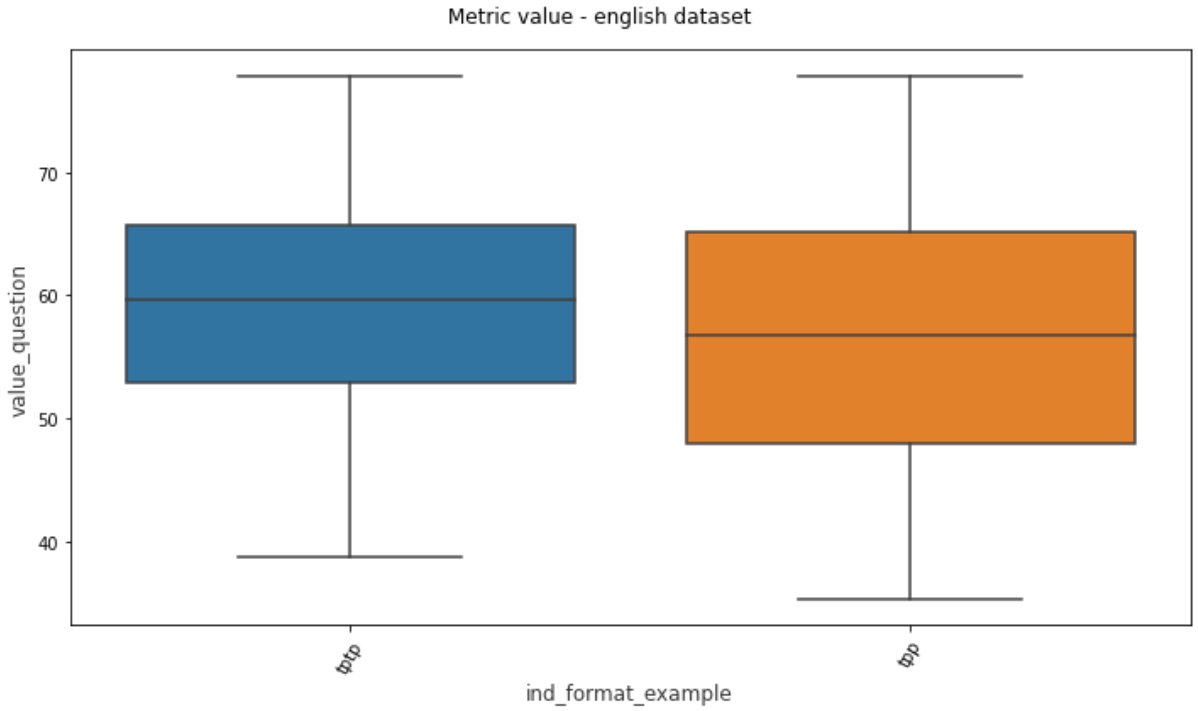

Gráfico 4 - Comparação de formatos de agrupamentos de exemplos.

.

Modelos de context learning maiores, com mais parâmetros, apresentam melhores desempenho. Modelos GPT-J, com 6 bilhões de parâmetros, alcançaram 78% fe F1@3, enquanto modelos GPT-Neo, com 2,7 e 1,3 bilhões de parâmetros, alcançaram 68% e 61%, respectivamente (Gráfico 5).

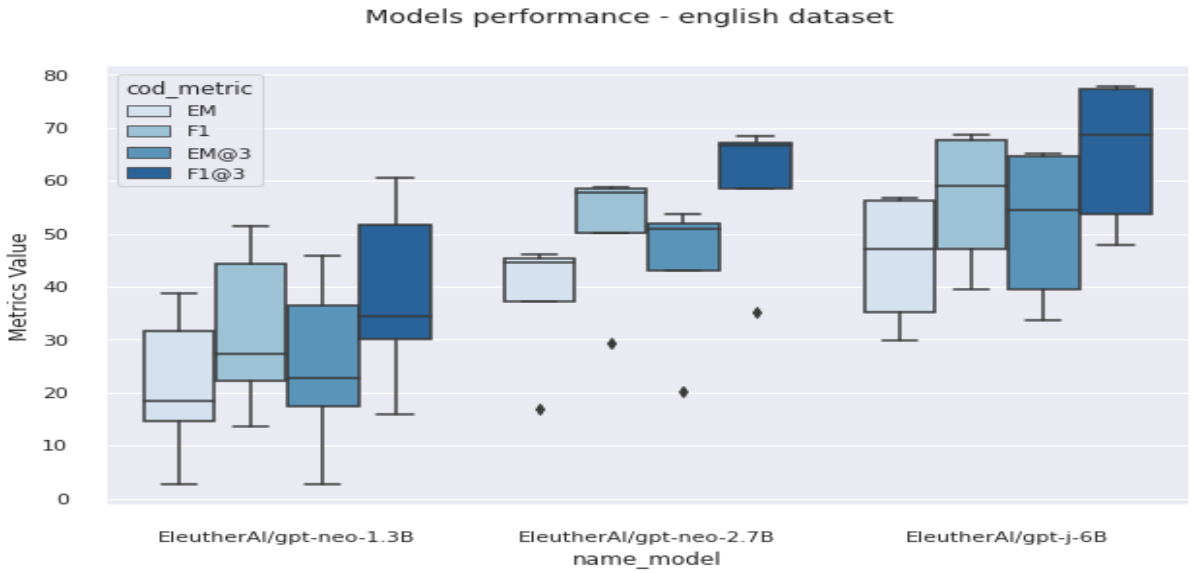

Gráfico 5 - Context learning em inglês, comparando cada modelo GPT empregado

De forma resumida, como esboçado na Gráfico 6, o método Transfer Learning alcançou resultados melhores do que o Context Learning.

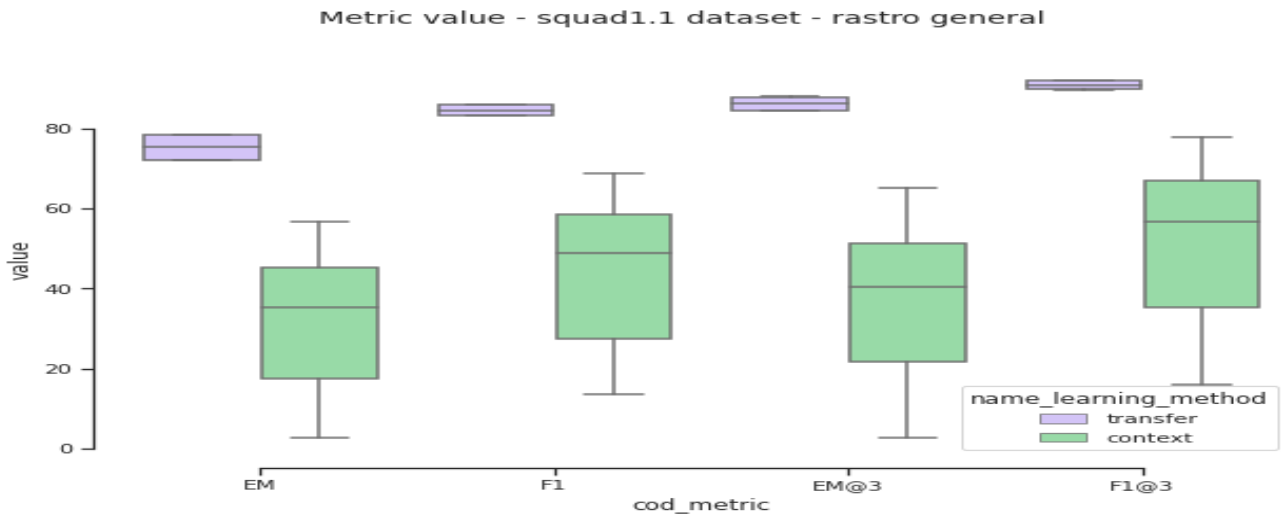

Gráfico 6 - Comparação geral do context learning e transfer learning

Em termos de velocidade, modelos ajustados apresentaram menores tempos, em média 220 milissegundos por questão para perguntas em inglês, e 262 ms em português. Para processar prompts em inglês, modelos GPT-Neo-1.3, 2.7 e GPT-J-6 levam em média 402, 539 e 537 milissegundos, respectivamente. Prompts em português levam mais tempo de processamento, desde 494, chegando até 1395 milissegundos (Gráfico 7).

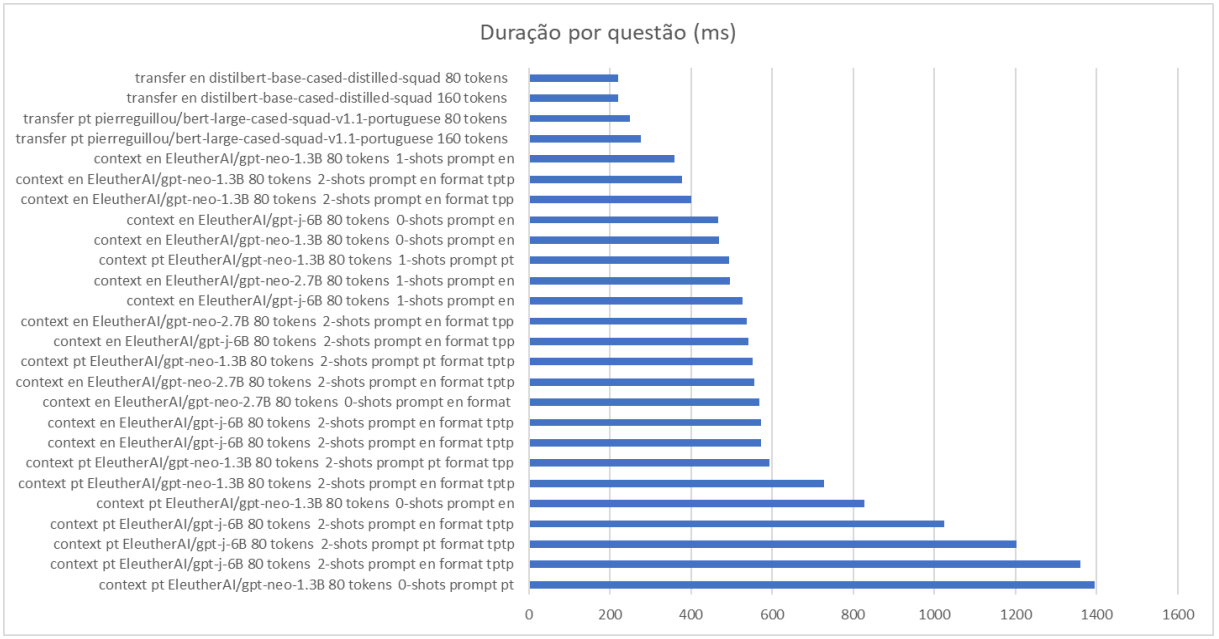

Gráfico 7 - Duração das execuções por questões, em milissegundos

As perguntas do dataset em inglês foram classificadas pelo seu tipo (what, who, etc). O Gráfico 8 traz os resultados da métrica F1.

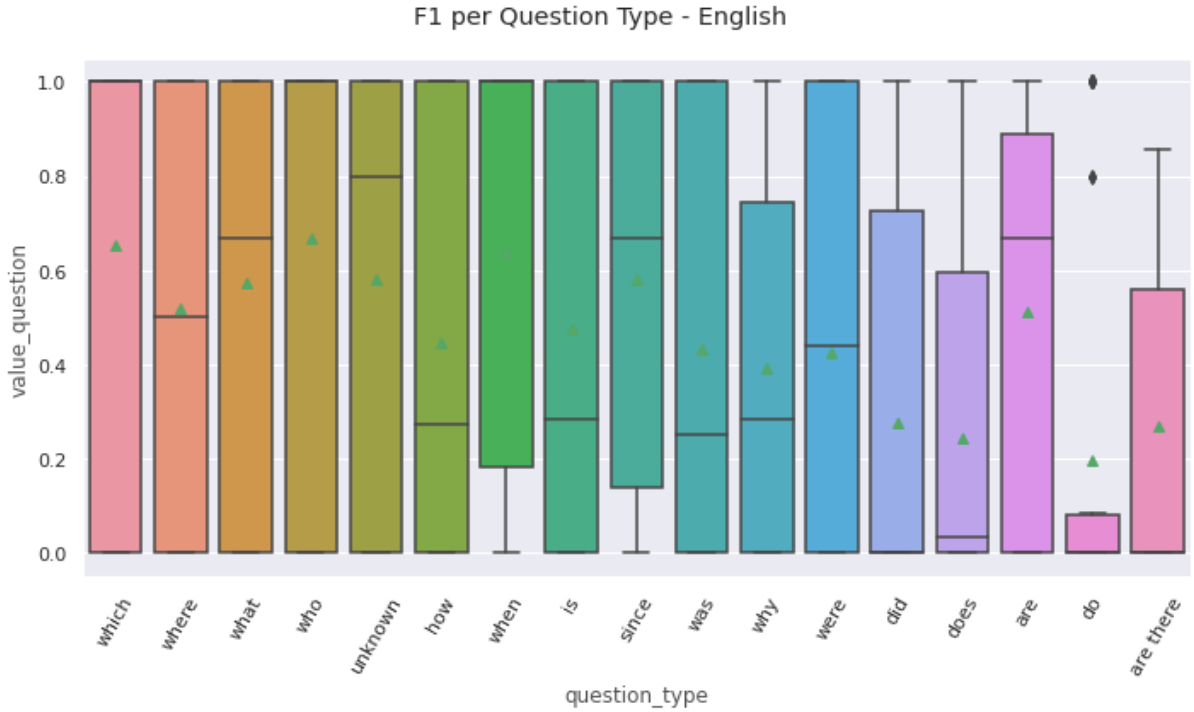

Gráfico 8 - Desempenho por tipo de pergunta (triângulo verde representa média)

6 Conclusão

Modelos similares ao GPT-3, que prescindem de ajuste fino, apresentaram desempenho inferior aos modelos ajustados, da ordem de 15% menor no caso de medida de F1, ou pouco mais de 20% no caso de EM. Modelos GPT maiores, com mais parâmetros, apresentam melhor desempenho, o que nos indica que modelos ainda maiores em parâmetros, como o GPT-3, com 175 bilhões de parâmetros, e treinados com conjuntos mais massivos de dados, deverão se aproximar ainda mais ou mesmo ultrapassar.

Os resultados com perguntas em português foram inferiores aos obtidos com perguntas em inglês em todos os modelos. Contribui para essa diferença o fato de a base de dados em português ter sido obtida por tradução automática da base em inglês, o que pode trazer erros. Nos modelos ajustados, porém, essa diferença foi bem reduzida, da ordem de 2% no F1, o que indica uma boa qualidade do procedimento de ajuste em português. Nos modelos GPT, por outro lado, encontramos uma penalidade bem maior, entre 30 e 36%. Uma possível causa dessa disparidade é a qualidade do treinamento do modelo: apesar desses modelos serem multilíngues, a massa de textos lidos em inglês é muito maior que em português. Por exemplo, o Common Crawl, que é o maior dataset empregando no treinamento do GPT-3, possui 46% dos textos em inglês e 2% em português[3]. Uma possível solução seria obtermos um modelo GPT-3 ajustado com maior massa de textos em português, o que não é possível, no momento.

7 Trabalhos Futuros

Como trabalhos futuros, vislumbram-se algumas possibilidades.

Utilização da base de dados SQuAD 2.0, que adiciona às 100 mil perguntas do SQuAD 1.1 cerca de 50 mil questões que não possuem resposta. Optamos pela versão 1.1 em função da existência da base traduzida para o português e de modelo ajustado nessa base. Mas um novo trabalho com a versão 2.0 é possível se for restrito ao inglês.

[3] https://commoncrawl.github.io/cc-crawl-statistics/plots/languages.html

Comparações com outros modelos também é possível, por exemplo, modelos T5 ajustados para o SQuAD, modelos para context-learning como o GPT-3, GPT-NeoX ou sucessores. Dentro do domínio de Q&A, podem ser utilizados também outros datasets e se ampliar para abstractive Q&A, em que a resposta pode não estar no texto, área promissora.

Estudos mais aprofundados sobre a variação da estrutura do prompt e mesclagem de línguas também podem ser efetuados. Bem como variações nos parâmetros passados aos modelos (Figura 2).

8 Referências

Castro, Marcus. Balaniuk, Remis. Rastro-DM: Mineração de Dados com Rastro. Revista do TCU. edição n. 145. 2020.

Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P., “SQuAD: 100,000+ Questions for Machine Comprehension of Text,”, arXiv:1606.05250, 2016.

Weng, Lilian, "How to Build an Open-Domain Question Answering System?", 2020.