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 avalia e compara o desempenho de métodos de aprendizagem de transformers, transfer learning ou context learning, em atividade extrativa de perguntas e respostas. Além disso, compara-se o uso de perguntas em inglês e português, a indicação ou não de exemplos no prompt para modelos de context learning, e ajustes na estrutura do prompt. Foram observados desempenhos superiores em transfer learning, e em inglês. 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 - QA) é aquela em que se parte de um documento que supostamente contém a resposta a uma dada pergunta, e dele é extraída a resposta final, que deverá ser um trecho do texto extraído. Um sistema completo de resposta extrativa pode ser dividido em duas partes: um retriever, um módulo que dados uma pergunta e uma base de pesquisa, fornece um conjunto de documentos cujo conteúdo tem relação com a pergunta, por exemplo usando um sistema de recuperação de informação baseado em BM25; e um reader, que dado um documento e uma pergunta, seleciona o trecho deste documento que corresponde à resposta da pergunta, possivelmente um modelo de rede neural baseado em transformer.

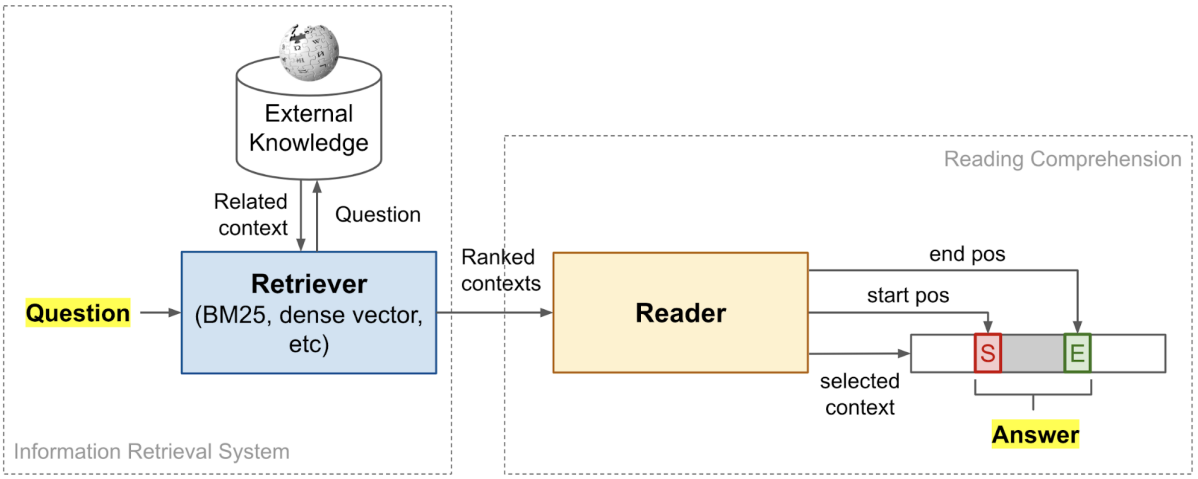

Figura 1 - O conjunto retriever-reader para QA. Fonte: https://lilianweng.github.io/posts/2020-10-29-odqa/.

O reader pode ser um modelo de transformer tradicional, por exemplo, Bert ou T5, que foi ajustado para essa tarefa de QA extrativa, por meio de transfer learning. Outra opção é empregar um modelo de linguagem autorregressivo, treinado com grande massa de dados e com maior quantidade de parâmetros, por exemplo, GPT-3 e similares, que pode ser instruído a responder a uma pergunta com base em um contexto (context learning), a partir da descrição da tarefa e de alguns exemplos (few-shot, pode-se prescindir da descrição da tarefa) ou não (zero-shot).

O objetivo do presente trabalho é comparar esses dois métodos de aprendizagem para utilização do reader: modelos de transfer learning, ajustados para QA, e modelos de context learning, não ajustados. Não é escopo do trabalho a etapa do retriever (Figura 1).

2 Conjunto de Dados

Utilizamos dados de validação da 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 10% formam o conjunto de validação ou dev set. Originalmente, disponibilizada em inglês, também está disponível em português, traduzida. Os dados de treinamento correspondem a 80% da base e foram utilizados na construção dos modelos ajustados que empregamos. Os demais 10% correspondem a dados de testes e são escondidos. [1]

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 QA 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 QA 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 fazer predições. 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 construída para consulta. Na variedade zero-shot, é formada por uma instrução, o contexto, a questão e o espaço para receber a resposta. Nas variedades few-shot, acrescenta exemplos de texto, pergunta e resposta. A resposta dos modelos, decodificada em texto, é limpa de pontuações, artigos, espaços desnecessários, e comparada à lista de respostas esperadas.

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 relativa ao total de palavras da resposta. Além disso, foram propostas e derivadas duas outras métricas: EM@3 e F1@3, que funcionam de forma similar às anteriores, mas calculadas sobre cada uma das três respostas geradas, e registrado o maior valor de F1 e Em dos três. A avaliação global é computada pela média aritmética das avaliações realizadas sobre as questões.

Todos os modelos foram executados em GPU NVIDIA GeForce RTX 3090, com 24GB de VRAM. No caso de modelos de transfer learning, foram produzidas dez vezes mais respostas, e foram eliminadas as respostas repetidas, e consideradas as três primeiras. Foram gerados no máximo 160 letras, valor superior ao número ao de todas as respostas. No caso de textos muito longos, são divididos em passagens, e 128 caracteres são repetidos entre eles (parâmetro num_doc_stride na library do pipeline do transformer).

Nos modelos de context learning, foi empregada temperatura de 0,1. Respostas de até 80 tokens, valor superior ao número ao de todas as respostas, e uso de sequências de parada.

As medições obtidas são persistidas em arquivos CSV, formando o banco de rastro. O modelo do banco representa um amadurecimento da aprendizagem, além de facilitar a organização de processos e a estruturação de conceitos

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%
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%

[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 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 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 que 1 resposta em 7%. As respostas a perguntas em inglês apresentam melhor desempenho que em português, porém, em transfer learning, a vantagem é discreta, de 2% (F1@3) a 6% (EM), em média. 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 figura ao lado.

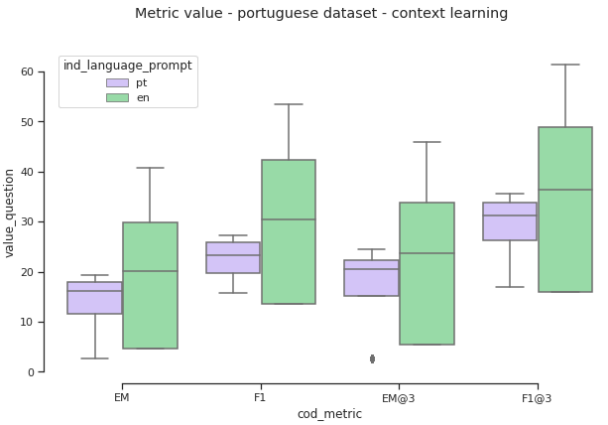

Figura 2 - Context learning em português, comparando instrução e exemplo em português com inglês

Modelos ajustados para o SQuAD apresentam melhor desempenho em relação a modelos não ajustados, 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.

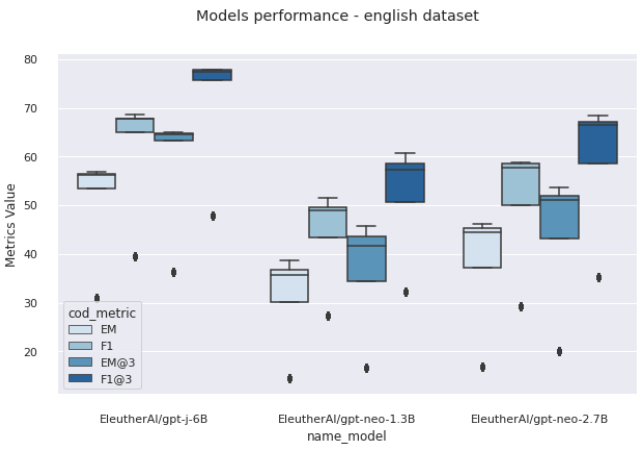

Figura 3 - Context learning em inglês, comparando cada modelo GPT empregado

Figura 4 - Context learning em inglês, comparando prompts zero-shot, 1-shot e 2-shot em cada modelo

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 Figura 4. Em prompts 2-shots, quando os exemplos são apresentados em trincas contexto-questão-resposta, o resultado é levemente melhor que com um único contexto para todos os exemplos de questão e resposta.

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 (Figura 3).

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 (Figura 5).

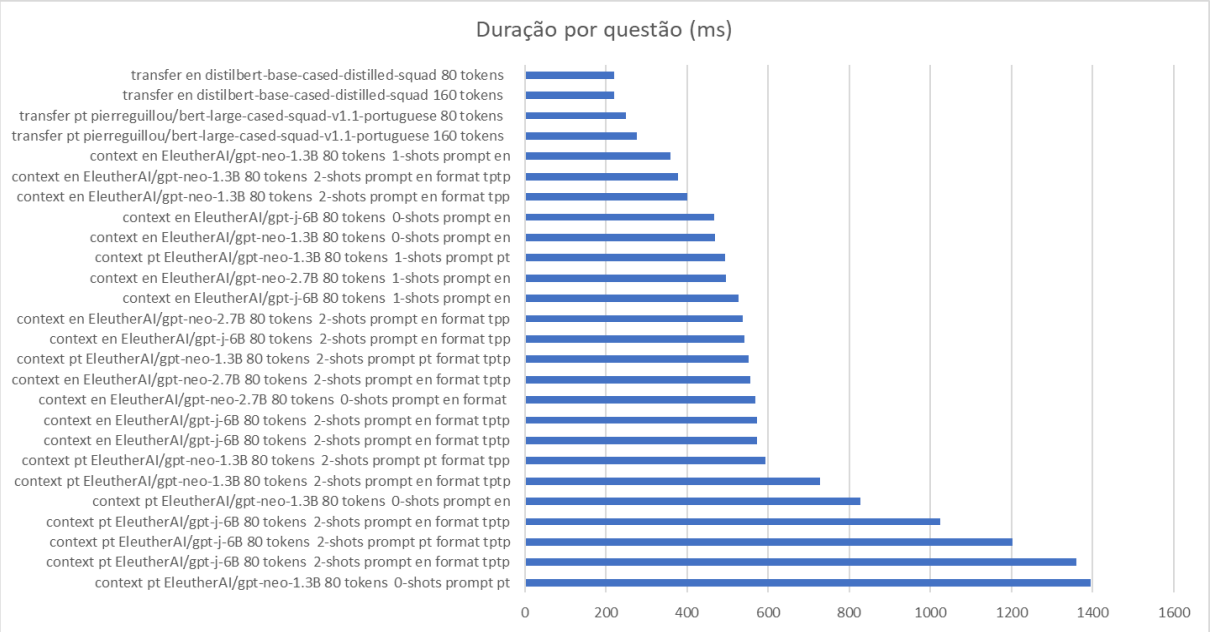

Figura 5 - Duração das execuções por questões, em milissegundos

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 de textos mais massivos, deverão se aproximar ainda mais.

Os resultados com perguntas em português foram inferiores aos obtidos com perguntas em inglês em todos os modelos. Isso seria esperado, uma vez que a base de dados em português foi obtida por tradução da base em inglês, o que traz 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.

Experimentamos no context learning o parâmetro min_length = 2, usado no pipe, que significa resposta com mínimo de 2 tokens. Havíamos achado que correspondiam a 2 caracteres, o que, posteriormente, se provou indevido, reduzindo o desempenho dos resultados. Seria recomendável repetir algumas execuções, retirando esse parâmetro.

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

6 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. Mas um outro trabalho, restrito ao inglês, é possível com a versão 2.0. Havendo acesso a um modelo ajustado para o SQuAD 2.0 em português, ou englobando o treinamento de tal modelo, tal trabalho com 2.0 em português seria viável.

Comparações com outros modelos também é possível, por exemplo, modelos T5 ajustados para o SQuAD, modelos para context-learning como o GPT-3 ou sucessores.

Também podem ser feitas comparações similares em outros tipos de atividades de NLP, como classificação de textos, NER, tradução, geração de texto, entre outros. Dentro do domínio de QA, podem ser utilizados outros datasets.

Estudos mais aprofundados sobre a variação da estrutura do prompt e mesclagem de línguas também podem ser efetuados.

7 Referências

[1] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, “SQuAD: 100,000+ Questions for Machine Comprehension of Text,” ArXiv e-prints, Jun. 2016.