Análise da Robustez de Mecanismos de Busca quanto a Ruídos em Consultas

Leonardo Augusto da Silva Pacheco e Marcus Vinícius Borela de Castro

Dezembro 2021

Resumo

O trabalho presente analisa o impacto que ruídos inseridos em consultas podem gerar na efetividade de mecanismos de buscas. Foram gerados diversos tipos de ruídos sobre 54 consultas. Para cada mecanismo empregado, foram comparados os resultados para consulta com ruído em relação à consulta original, a partir da métrica proposta DG:nDCG@10 que traduz perdas e ganhos de efetividade. Foram encontrados variados impactos dependendo da query e do ruído gerado nos diversos mecanismos de pesquisa avaliados.

1 Introdução

Desde a introdução das redes neurais profundas e, em particular, dos novos modelos de transformers, diversos mecanismos novos de pesquisa vêm sendo propostos, entre eles, mecanismos de Rerank (MonoT5, MonoBert e DuoBert), mecanismos de busca densa (DPR [1]) e técnicas de enriquecimento de conteúdo (doc2query). Esses mecanismos têm promovido resultados positivos em *benchmarks[[1]](#footnote-0)* de Recuperação de Informação[2]. Nesse contexto, outros pontos devem ser considerados no processamento de buscas, como o tempo de resposta e o uso de recursos computacionais.

O objetivo do presente trabalho é analisar o impacto que ruídos inseridos em consultas (*queries*) podem gerar na efetividade de mecanismos de buscas. Um ruído é uma alteração no texto de uma *query*, como exclusão ou troca de palavras. Uma mudança pode impactar o resultado da pesquisa, por exemplo, deixando de trazer documentos relevantes nas primeiras posições. Buscou-se avaliar o impacto em termos de perda ou ganho em relação à efetividade da consulta original.

2 Conjunto de Dados

Utilizamos a base[[2]](#footnote-1) *TREC 2020 DL MsMarco Passage* contendo 54 queries sobre 8.841.823 passagens de texto e uma média de 210,9 julgamentos de relevância por query.

3 Metodologia

Como era objetivo replicar as análises para textos em português e por não haver GPU com fácil acesso, gerou-se uma base menor apenas com as 11.224 passagens que possuem julgamento (*TREC 2020 MsMarco Passage with judgment)* em uma instância local *Elastic Search*. Essa base menor foi utilizada para as buscas densas (DPR) tanto em português quanto em inglês. Para se gerar a base em português, foram traduzidos os 11.224 documentos usando-se *library* python *googletrans*.

Os dados[[3]](#footnote-2) usados e gerados durante o projeto estão publicados no repositório do projeto como dataframes Pandas/Python. Também estão disponíveis diagramas entidade x relacionamento (MER) dos conceitos representados[[4]](#footnote-3).

Sobre as 54 queries originais foram aplicados oito tipos de ruídos[[5]](#footnote-4):

* Permutação de palavras em determinadas posições da query: primeira e segunda, primeira e última e penúltima e última.
* Remoção de palavras: da segunda, da última e retirando palavras com probabilidade de 20% (pelo menos a palavra central).
* Back translation: tradução para outra língua (português/inglês) seguida de nova tradução de volta para o idioma original, gerando mudanças de palavras.
* Inserção de um erro tipográfico[[6]](#footnote-5), por exemplo, trocando “what” por “waht”.

Executaram-se buscas tanto com as queries com os ruídos[[7]](#footnote-6), como com as queries originais, usando as funções de ranqueamento[[8]](#footnote-7): BM25 [3], BM25 + Rerank (MonoT5) e busca densa (DPR). Usando os julgamentos de relevância (com as queries originais) das dez primeiras passagens retornadas, calculou-se as métricas DCG@10 (*Discounted cumulative gain at 10 documents*) e nDCG@10 (*Normalized Discounted cumulative gain at 10 documents)*.

Para confrontar a efetividade da query ruidosa com a query original, foi proposta a métrica DG:nDCG@10 (Discount or Gain on normalized Discounted cumulative gain at 10 documents):

onde eps=10e-6. Valores menores que zero representam desconto na efetividade, e maiores que zero representam ganhos na efetividade. No contexto deste relatório abreviamos DG:nDCG@10 como DG.

Foram retirados os cálculos realizados na base em português referente a quatro modelos devido à pouca eficácia encontrada. Critério usado: mais de 23% das pesquisas com nDCG@10 zerado e média menor do que a metade do valor alcançado pela função BM25 em mesma base. Assim, foram descartados os modelos de Rerank[[9]](#footnote-8): ptt5-small-portuguese-vocab, ptt5-base-portuguese-vocab e ptt5-base-t5-vocab. E na busca densa, o modelo voidful/dpr-question\_encoder-bert-base-multilingual.

Para diminuir impacto de outliers no cálculo da métrica DG, optou-se por excluir 29 registros referentes a queries com ruído que obtiveram nDCG@10 superior a zero enquanto que a query original alcançou zero. E por excluir, na sequência, outros 25 registros que tiveram z-score absoluto superior a três (três vezes o desvio padrão).

4 Resultados

Os cálculos[[10]](#footnote-9) e os detalhes da análise[[11]](#footnote-10) estão publicados no github do projeto.

O impacto de um determinado ruído em cada query varia. Por exemplo, no caso da remoção de uma palavra em uma determinada posição na query, se esta palavra for uma stopword, é possível que a resposta para a query ruidosa seja melhor do que a resposta para a query original. Por outro lado, se a palavra removida for um substantivo com significado chave para a intenção por trás da query, descaracteriza-se completamente a consulta. O Gráfico 1 ilustra um caso desses, em que há queries onde há desconto reduzido ou mesmo ganhos, até queries com desconto máximo (-1).

Os resultados das pesquisas de Rerank em português com 3 modelos diferentes obtiveram resultados próximos e por isso foram unificados pela média na análise que se segue.

O Gráfico 2 mostra os valores médios de DG para os diversos contextos de busca e tipos de ruído. Percebe-se maior robustez em relação a permutações e erros tipográficos, e maior sensibilidade a *back-translation* e remoções de palavras. A busca densa (DPR) apresentou maiores descontos em pesquisas realizadas na base em português em relação às outras funções de ranqueamento, ao contrário da base em inglês onde obteve menor desconto na média. A busca BM25 é plenamente robusta quanto às permutações, porém, em geral, é relativamente mais sensível às deleções. O Rerank mostrou-se mais sensível do que o BM25 para deleções e mais robusto em outros tipos de ruídos. Empataram com o maior ganho (0.55) o ruído de permutar a primeira e a segunda palavra na busca densa em inglês e o ruído de *back translation* na busca por BM25 na base em português. E os piores foram com o BM25 na deleção.

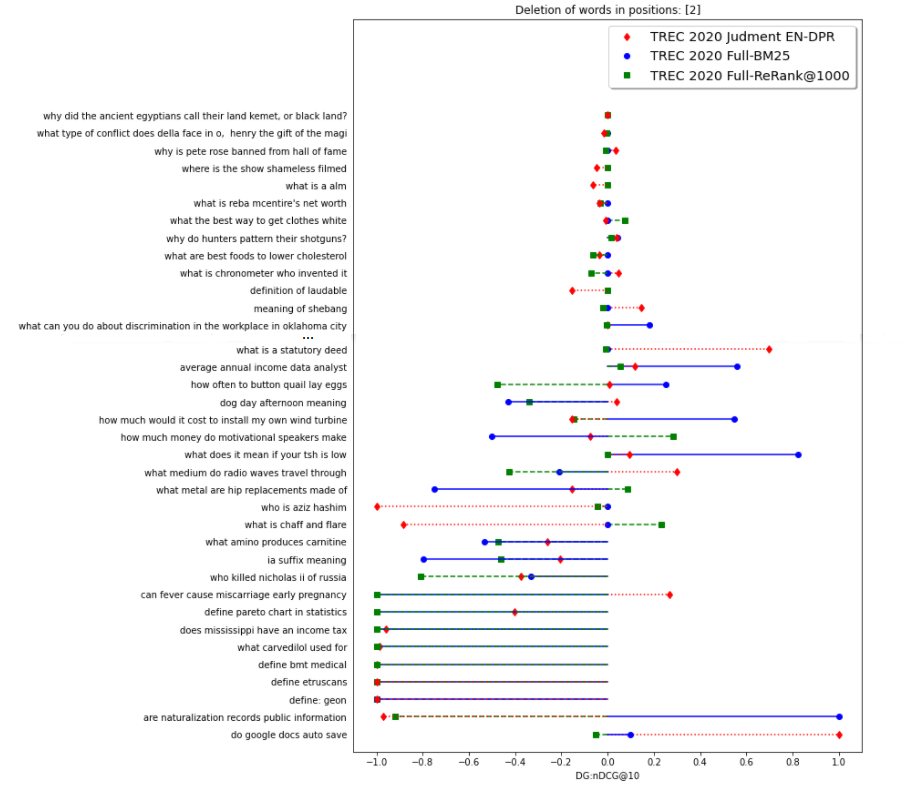


Gráfico 1 - Valores obtidos de DG:nDCG@10, quey por query (parcial)

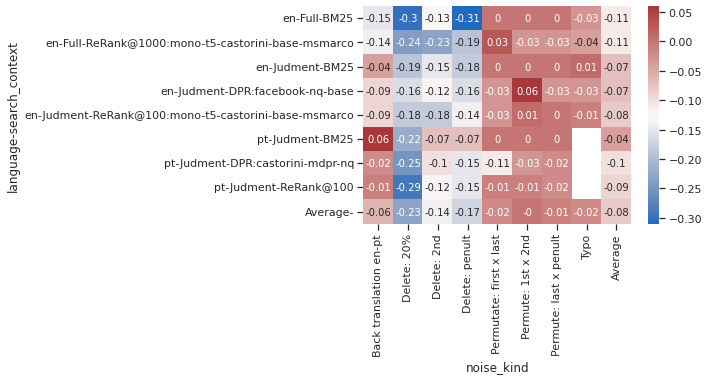


Gráfico 2 - Valores médios da métrica DG:nDCG@10 para vários mecanismos de busca e tipos de ruído

O Gráfico 3 demonstra o desvio padrão encontrado na métrica DG nas pesquisas. Percebe-se que os ruídos de permutação obtiveram uma variação menor. E que a busca com Rerank na base em português também teve valores em média menores em relação aos outros ruídos.

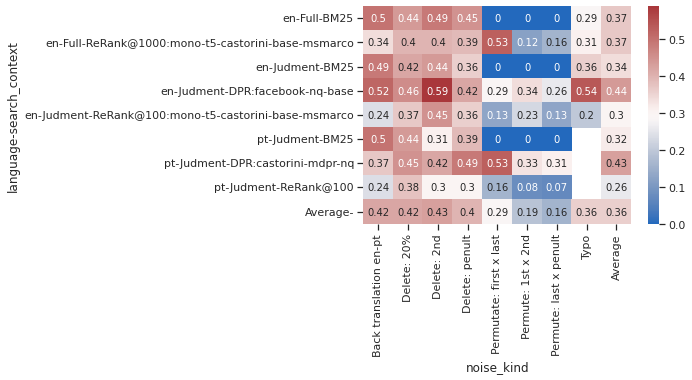


Gráfico 3 - Desvio padrão da métrica DG:nDCG@10 para vários mecanismos de busca e tipos de ruído

Os Gráficos 4 demonstra a distribuição dos valores da métrica DG por tipo de ruído. Percebe-se uma maior variação nas deleções, em que perdas maiores são identificadas. Em particular na deleção com 20% de probabilidade, em que mais palavras são retiradas.

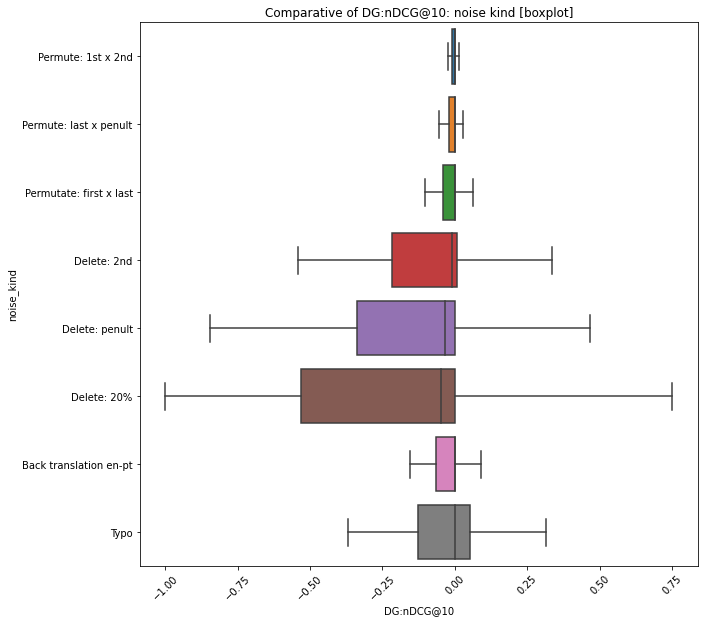


Gráfico 4 - Distribuições de DG:nDCG@10 por tipo de ruído

O Gráfico 5 traz as distribuições por contexto de pesquisa.

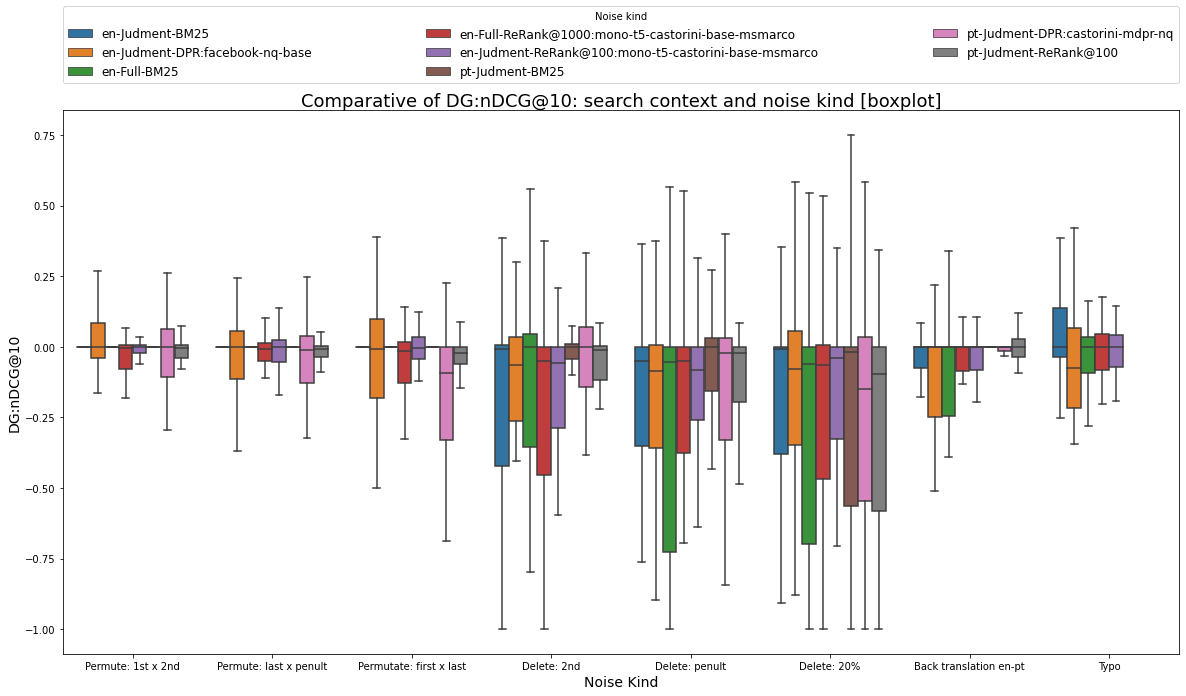


Gráfico 5 - Distribuições de DG:nDCG@10 por tipo de ruído para vários mecanismos de busca

Como tendência geral, observou-se que textos maiores de consulta têm menores variações de DG, sendo mais robustos aos ruídos experimentados, enquanto textos mais curtos apresentam resultados mais variados. O Gráfico 6 demonstra essa tendência.

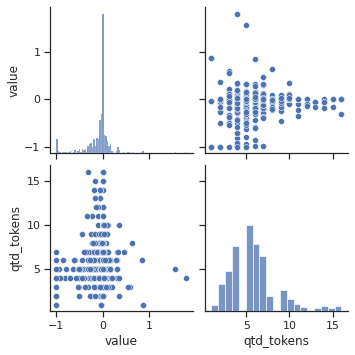


Gráfico 6 -Correlação = 0.166 entre quantidade de tokens e DG para contexto en-judment-DPR:facebook-nq-base

5 Conclusão

A robustez das funções de ranqueamento variou com os diversos contextos: linguagem e base utilizada. Exceção a permutações de palavras em que o BM25, por sua forma de cálculo, é plenamente robusto.

Espera-se que funções de ranqueamento baseadas em transformers, por capturar significado das palavras e relação com o contexto, fosse mais robustas em no caso de back-translation, já que elas não seriam muito impactados quando uma palavra se transforma em um sinônimo, o que prejudicaria buscas em mecanismos estatísticos tradicionais como o BM25. Mas isso não foi constatado nos dados na base em português. Uma possível explicação seria um possível afunilamento de termos nas traduções. Mais de uma palavra em uma língua é traduzida para uma única palavra. E tanto as consultas e as passagens foram traduzidas pelo mesmo método (python/googletrans).

O ruído de erro tipográfico (um por frase) não gerou muito impacto nos contextos de pesquisa. Uma possível explicação seria o fato de transformers tratarem melhor partes de palavras (pelo tokenizador) ou por serem treinados com dados com ruídos. Ou até pela existência do próprio ruído nos documentos pesquisados.

Observou-se que consultas com textos maiores são mais robustas aos ruídos experimentados, com menores variações de DG. Enquanto textos mais curtos apresentam resultados mais variados. A possível explicação para isso é que em textos maiores há uma menor chance de se perder palavras mais relevantes.

Como visto, os modelos Rerank em português foram unificados nas análises pela média dados resutlados não muito díspares.. Mas cabe ressaltar que o modelo unicamp-dl/mt5-base-multi-msmarco alcançou menor DG (-0.068) em relação aos outros unicamp-dl/mt5-base-en-pt-msmarco (-0.100) e unicamp-dl/ptt5-base-pt-msmarco-100k (-0.091). Mas este último, se mostrou mais eficiente por consumir 63% menos memória (810 mb contra 2.17 gb) e ser 30% mais rápido.

6 Trabalhos Futuros

Como trabalhos futuros, vislumbram-se algumas possibilidades.

Quanto à análise, podem ser realizadas análises mais detalhadas por tipo de função de ranqueamento, por modelo usado, por base de texto pesquisada ou por linguagem, podendo-se ampliar o escopo desses. E serem investigadas hipóteses levantadas na conclusão.

Quanto às técnicas envolvidas na busca, poderia ser ampliado o leque de técnicas avaliadas, incluindo doc2query, serem usados outros tipos de busca densa, entre outros.

Quanto aos tipos de ruídos, sugere-se aumentar o número de ruídos, por variação dos sugeridos ou por junção dos mesmos. Ou mesmo com uso de novos tipos.

Quanto às bases de documentos, havia uma restrição de uso de GPU o que levou ao uso de uma base menor na maioria dos contextos de pesquisa. Sugere-se repetir os experimentos com os modelos na base completa TREC 2020 MsMarco Passage.

Quanto à métrica base de avaliação, sugere-se avaliar o comportamento da métrica proposta DG sobre outras métricas sem outros domínios. Por exemplo: DG:MRR (Mean Reciprocal Rank) e DG:MAP (Mean Average Precision).

Pode-se também avaliar:

* o uso de ruídos como mecanismo de enriquecimento de dados no treinamento de modelos e avaliar como isso impacta a robustez de pesquisas.
* os casos de ruídos que levaram a ganhos inesperados. Talvez floresçam novas técnicas de aprimoramento para consultas com a aplicação desses ruídos.

Por fim, espera-se que cada vez mais maiores ganhos sejam alcançados em detrimento de menores descontos na área de recuperação da informação, sendo ainda mais robusta mesmo com os ruídos!

7 Referências

[1] [Dense Passage Retrieval for Open-Domain Question Answering. Vladimir Karpukhin and Barlas Oğuz and Sewon Min and Patrick Lewis and Ledell Wu and Sergey Edunov and Danqi Chen and Wen-tau Yih.2020.](https://arxiv.org/pdf/2004.04906)

[2] Pretrained Transformers for Text Ranking: BERT and Beyond. Jimmy Lin, Rodrigo Nogueira, and Andrew Yates (anotações)

[3] [The Probabilistic Relevance Framework: BM25 and Beyond.Stephen Robertson & Hugo Zaragoza. 2009](https://www.researchgate.net/publication/220613776_The_Probabilistic_Relevance_Framework_BM25_and_Beyond)

1. MS MARCO Passage Ranking Leaderboard em <https://microsoft.github.io/msmarco/> [↑](#footnote-ref-0)
2. <https://microsoft.github.io/msmarco/TREC-Deep-Learning-2020> [↑](#footnote-ref-1)
3. <https://github.com/leonardo3108/robustez-query/tree/main/data> [↑](#footnote-ref-2)
4. <https://github.com/leonardo3108/robustez-query/tree/main/model> [↑](#footnote-ref-3)
5. <https://github.com/leonardo3108/robustez-query/blob/main/data/tab_noise_kind.csv> [↑](#footnote-ref-4)
6. Apenas esse ruído não foi gerado para pesquisas em português (restrição de tempo no projeto). [↑](#footnote-ref-5)
7. <https://github.com/leonardo3108/robustez-query/blob/main/data/tab_noisy_query.csv> [↑](#footnote-ref-6)
8. <https://github.com/leonardo3108/robustez-query/blob/main/data/tab_ranking_function.csv> [↑](#footnote-ref-7)
9. <https://huggingface.co/unicamp-dl> [↑](#footnote-ref-8)
10. <https://github.com/leonardo3108/robustez-query/blob/main/data/tab_calculated_metric.csv> [↑](#footnote-ref-9)
11. [https://github.com/leonardo3108/robustez-query/blob/main/code/analysis/Analysis\_Discount\_Gain.ipynb](https://github.com/leonardo3108/robustez-query/blob/main/code/analysis/Analysis_Discount_Gain_pt.ipynb) [↑](#footnote-ref-10)