Avaliando Modelos Encoder-Decoder em Tarefas de Múltipla Escolha sob o Viés de Forma Superficial

Breno Gabriel, Gabriel Henrique, Matheus Barney, Pedro Lima, Ricardo Bizerra
Centro de Informática
UFPE
Recife, Brasil
{bgml, ghv, mbmg, pbsl, rblf}@cin.ufpe.br

***Abstract*—Este relatório constitui parte integrante do projeto executado para a disciplina "Tópicos Avançados em Inteligência Artificial", que possui o artigo intitulado "Surface Form Competition: Why the Highest Probability Answer Isn't Always Right" como referência.**

***Index Terms*—Encoder-decoder models, Language models, Multiple-choice questions, Natural Language Inference, Surface form competition, Zero-shot learning**

I. Introdução

Os modelos de Linguagem Grande (em inglês, LLM's), representam um grande avanço na realização de diversas tarefas no campo da inteligência artificial. Utilizando a configuração **zero-shot**, ou seja, quando o modelo executa uma tarefa sem ter sido previamente ajustado com exemplos específicos daquela atividade, é possível realizar tarefas como classificação de texto, inferência textual, sumarização e resposta automática a perguntas. Entretanto, foi detectado que, em tarefas envolvendo a escolha de alternativas em perguntas de múltipla escolha, esses modelos podem selecionar uma resposta incorreta, mesmo possuindo conhecimento suficiente para apontar a resposta correta.

A explicação para esse fenômeno é denominado **surface form competition**, uma propriedade dos modelos generativos na qual a probabilidade de gerar uma resposta é diluída entre diferentes formas superficiais possíveis, mesmo que todas essas formas transmitam a mesma ideia. Em outras palavras, o modelo distribui a probabilidade entre várias versões equivalentes da resposta, o que pode levar a uma pontuação reduzida para a forma específica presente entre as alternativas oferecidas. Essa competição entre formas textuais prejudica a precisão do modelo e afeta negativamente o seu desempenho em tarefas de múltipla escolha.

Melhor explicando, podemos considerar a seguinte pergunta: "Uma pessoa quer se submergir na água. O que ela deve usar?" e as suas alternativas "banheira de hidromassagem", "poça", "copo" e "xícara". A resposta correta seria "banheira de hidromassagem", no entanto, o modelo pode atribuir maior probabilidade à palavra "banheira", por ser mais comum, mesmo que ela não esteja entre as opções listadas. Assim, ele pode acabar escolhendo uma alternativa errada, como "poça", simplesmente porque a resposta correta teve sua probabilidade "roubada" por uma forma equivalente mas ausente da lista.

Diante desse problema, o artigo [1] propõe o método Domain Conditional Pointwise Mutual Information (PMIDC) como solução. Esse método busca recalcular a pontuação das respostas candidatas, considerando não apenas a probabilidade da resposta dada a pergunta, mas também o quão provável aquela resposta seria em um contexto neutro da tarefa (o domínio). Dessa forma, o modelo penaliza respostas genéricas ou comuns, e favorece aquelas que de fato se tornam mais prováveis por causa da pergunta, corrigindo o viés introduzido pela surface form competition.

II. Metodologia

A. Arquiteturas de Modelos de Linguagem

A arquitetura Transformer constitui um progresso significativo no aprendizado profundo, transformando a maneira como tratamos a linguagem natural. A arquitetura Transformer emprega mecanismos de atenção em vez de redes recorrentes, permitindo o processamento simultâneo de sequências e a identificação mais eficiente de relações complexas a distância. Basicamente, ela é composta por duas partes: o encoder, encarregado de analisar a entrada, e o decoder, responsável por gerar a saída.

O encoder processa a sequência de entrada, convertendo cada elemento em um vetor contínuo por meio de embeddings e incorporando dados de posição para preservar a ordem original. Esses vetores passam por várias camadas, formadas por mecanismos de autoatenção multi-cabeça, que possibilitam que cada elemento analise todos os demais na sequência, e redes feed-forward, que melhoram as representações. Ao final do processo, o encoder produz vetores contextualizados que capturam o sentido completo da entrada.

O decoder, por outro lado, utiliza as representações produzidas pelo codificador, combinando-as com o que já foi criado antes, para gerar a sequência de saída. Ele emprega autoatenção mascarada, garantindo que cada posição considere apenas os elementos já gerados, mantendo a ordem de criação, e um mecanismo de atenção cruzada que conecta as representações do codificador à saída em desenvolvimento. Desta forma, o decoder consegue gerar texto de forma lógica e consistente, elemento por elemento, até finalizar a tarefa pretendida.

Vale ressaltar que um modelo não precisa necessariamente utilizar os dois componentes. Existem modelos como GPT e

LLaMA que utilizam somente o decoder, recebendo o texto de entrada combinado ao que já foi gerado e prevendo o próximo token. Esses modelos são mais adequados para tarefas de geração de texto livre, como escrita criativa, conclusão de frases, desenvolvimento de código, chatbots e contação de histórias. Por outro lado, existem os modelos que utilizam tanto o encoder para receber e processar a entrada quanto o decoder para gerar a saída, sendo eficientes para tarefas que demandam um texto de entrada e um texto de saída diferentes. T5, BART e mBART são exemplos desse fenômeno.

B. Artigo original

Os seguintes modelos foram utilizados, no artigo original, para avaliação dos resultados dos experimentos, principalmente, em contexto de zero-shot learning e, em alguns casos, realizando few-shot learning (4 shots):

TABLE I
MODELOS USADOS PELO ARTIGO ORIGINAL

Família	Modelo	Parâmetros
GPT-2	GPT-2	125 milhões
GPT-2	GPT-2-medium	350 milhões
GPT-2	GPT-2-large	760 milhões
GPT-2	GPT-2-xl	1,6 bilhões
GPT-3	ada	2,7 bilhões
GPT-3	babbage	6,7 bilhões
GPT-3	curie	13 bilhões
GPT-3	davinci	176 bilhões

C. Nosso projeto

TABLE II
MODELOS USADOS NO PROJETO

Modelo	Parâmetros	Observação	Arquitetura
mT0-large	1,2 bilhões	Fine-tuning do mT5 em uma mistura de tarefas em várias línguas	Encoder-Decoder
FLAN-T5 large	780 milhões	Treinado sob uma técnica de transferência de conhecimento, pré-treinado e posteriormente realizado tunagem de instrução	Encoder-Decoder

Este projeto tem como objetivo realizar uma análise comparativa do impacto de diferentes arquiteturas de modelos de linguagem em tarefas de múltipla escolha. Para isso, propõe-se a aplicação de dois modelos com arquitetura encoder-decoder, com o intuito de estender o conteúdo do artigo original nesse quesito, dado que no mesmo foram apenas testados modelos decoder-only. A avaliação, que inicialmente tinha a meta de ser feita nos 10 datasets do artigo, acabou sendo realizada em apenas 5 deles devido a limitações computacionais. Os datasets foram: COPA, COPA-Reversed, StoryCloze, HellaSwag e RACE-M. A partir da análise das métricas de acurácia, busca-se extrair insights e promover uma discussão aprofundada sobre as vantagens e desvantagens de cada arquitetura na resolução de problemas que exigem compreensão de texto e raciocínio inferencial.

Considerando que o surface form competition é um fenômeno intrinsecamente ligado à maneira como os modelos calculam a probabilidade de sequências de texto, a arquitetura do modelo torna-se um fator crítico de análise. Modelos decoder-only, que tratam a tarefa como um problema de completar uma única sequência, podem ser particularmente sensíveis a esse viés. Em contrapartida, a arquitetura encoder-decoder, ao criar uma representação de contexto separada da entrada, pode possuir mecanismos de atenção distintos que potencialmente mitigam ou alteram o efeito deste fenômeno. Investigar essa diferença é, portanto, o objetivo central deste trabalho.

D. Dataset

Neste projeto, foram utilizados cinco base de dados apresentados pelo artigo. O motivo pelo qual não foram utilizados todos os datasets está associado a limitações computacionais, decorrentes do uso de GPU. Os datasets escolhidos foram:

- **Choice of Plausible Alternatives (COPA):** é um conjunto de 1000 questões de causa e efeito. Cada questão é composta por uma premissa ao modelo e duas sentenças representando o efeito da mesma. O objetivo é fazer com que o modelo escolha qual sentença representa o efeito é mais provável para a premissa, permitindo avaliar a capacidade do modelo de raciocinar sobre relações causais no mundo real.
- **Choice of Plausible Alternatives Reversed (COPA-flipped):** é uma variação do COPA onde o modelo deve definir qual a premissa é mais provável a partir das hipóteses presentes em cada questão. Basicamente, ocorre uma inversão entre causa e efeito, com a devida adaptação nos conectores because e so para garantir coerência semântica. Essa inversão muda a forma de avaliação: em vez de escolher qual continuação tem maior probabilidade (P(y—x)), mede-se qual contexto torna a mesma continuação mais provável (P(x—y)). Isso remove a surface form competition, pois agora todos os candidatos compartilham o mesmo texto de continuação, e o modelo só precisa avaliar qual contexto é mais plausível para essa continuação.
- **StoryCloze (SC):** uma base de dados que fornece duas opções de final para uma história composta por cinco sentenças.
- **HellaSwag (HS):** um dataset que, assim como as bases de dados anteriores, é voltado para para avaliar modelos em tarefas de escolha de continuidade de história. Cada instância desse conjunto é composta por um contexto, que pode ser uma descrição curta ou narrativa parcial, e quatro possíveis continuações sendo que apenas uma é correta e foi escrita por humanos enquanto as outras três são distratores gerados de forma automática, mas cuidadosamente filtrados para parecerem plausíveis.

- **RACE-M:** é uma base de dados de testes de interpretação de texto em inglês para alunos do ensino médio chinês. Cada instância desse conjunto é composta por um texto, um enunciado e quatro alternativas.

E. Métricas no artigo original

Para entender a performance obtida pela nossa implementação, primeiramente é importante entender as métricas utilizadas no artigo motivador. As métricas que o artigo mostra são obtidas a partir da aplicação de diferentes métodos de pontuação aplicados às saídas dos modelos. Da aplicação desses diferentes métodos se obtiveram as acurácias que são mostradas nas tabelas.

Os métodos utilizados foram:

- **LM (Language Model)** é a abordagem padrão que simplesmente seleciona a opção de resposta com a maior probabilidade.
- Uma variação comum é a **AVG (Average Log-Likelihood)**, que normaliza os logaritmos das probabilidades pelo comprimento da resposta para mitigar o viés de modelos em favor de sequências mais curtas.
- **PMI (Pointwise Mutual Information)** mede o quanto a pergunta torna a resposta mais provável em geral, ao dividir a probabilidade condicional pela probabilidade incondicional da resposta.
- A métrica central do estudo, **PMIdc (Domain Conditional PMI)**, adapta a PMI para compensar a "competição de forma superficial"; ela faz isso ao normalizar a probabilidade da resposta por sua probabilidade dentro de um contexto de domínio específico da tarefa, em vez de uma probabilidade geral.
- Por fim, a **UNC (Unconditional)** serve como uma linha de base que ignora a pergunta e pontua a resposta apenas com base em sua probabilidade dentro do domínio, verificando se o modelo está de fato utilizando a informação da pergunta.

III. Resultados

TABLE III
RESULTADOS NO DATASET COPA

Métrica	Flan-T5-Large	mT0-large
UNC	**0.560000**	0.558000
LM	**0.746000**	0.584000
AVG	**0.748000**	0.598000
PMIdc	**0.772000**	0.618000
PMI	**0.772000**	0.614000

TABLE IV
RESULTADOS NO DATASET COPA-REV

Métrica	Flan-T5-Large	mT0-large
UNC	0.500000	0.500000
LM	**0.780000**	0.620000
AVG	**0.780000**	0.620000
PMIdc	**0.780000**	0.620000
PMI	**0.780000**	0.620000

TABLE V
RESULTADOS NO DATASET STORYCLOZE

Métrica	Flan-T5-Large	mT0-large
UNC	**0.511491**	0.508819
LM	**0.706574**	0.558525
AVG	**0.739711**	0.570283
PMIdc	**0.753608**	0.574025
PMI	**0.764832**	0.580438

TABLE VI
RESULTADOS NO DATASET HELLASWAG

Métrica	Flan-T5-Large	mT0-large
UNC	**0.299442**	0.265286
LM	**0.379406**	0.284804
AVG	**0.467038**	0.293866
PMIdc	**0.447421**	0.315973
PMI	**0.449213**	0.316570

IV. Discussões

Os resultados obtidos nos experimentos mostram um padrão consistente de superioridade do FLAN-T5 large em relação ao mT0-large na maioria dos datasets. Essa diferença pode ser explicada por fatores como os focos de linguagem diferentes e o impacto diferenciado do surface form competition em cada arquitetura.

No dataset COPA, o FLAN-T5 obteve 0,748 em AVG, superando o mT0-large (0,598). Esse desempenho é inclusive superior ao registrado pelos modelos no artigo original, evidenciando que, mesmo com menor número de parâmetros, a arquitetura encoder-decoder do FLAN-T5, aliada ao ajuste fino em instruções, foi capaz de capturar melhor as relações causais exigidas. O mT0-large, embora também encoder-decoder, apresentou menor acurácia, possivelmente devido à interferência de padrões adquiridos em diferentes idiomas devido a sua natureza multilíngue, que podem introduzir ruído na interpretação causal.

No COPA-REV, o FLAN-T5 atingiu o melhor resultado geral de todo o conjunto de experimentos (0,780 em AVG), superando tanto o mT0-large (0,620) quanto os resultados originais dos modelos GPT-2-large e GPT-2-xl. Esse resultado reforça o efeito mitigador que a inversão de causa e efeito tem sobre o surface form competition: como todos os candidatos compartilham o mesmo texto de continuação, o viés de formas superficiais é neutralizado, permitindo que o modelo se concentre na avaliação de plausibilidade contextual. A vantagem do FLAN-T5 aqui sugere que seus mecanismos de atenção e codificação contextual são mais eficientes quando o ruído introduzido pelo surface form competition é removido.

No StoryCloze, o FLAN-T5 manteve alta performance (0,7397), muito próxima à obtida em COPA, enquanto o mT0-large apresentou queda significativa (0,5703). Essa diferença pode ser atribuída ao fato de que a tarefa exige modelagem de coerência narrativa em inglês, beneficiando um modelo treinado prioritariamente nesse idioma.

No HellaSwag, ambos os modelos tiveram desempenho reduzido (FLAN-T5: 0,4670; mT0-large: 0,2939), o que é consistente com a dificuldade intrínseca do dataset, projetado para

TABLE VII
RESULTADOS NO DATASET RACE-M

Métrica	Flan-T5-Large	mT0-large
UNC	0.212396	**0.215181**
LM	**0.561978**	0.459610
AVG	**0.591922**	0.490251
PMIdc	**0.580780**	0.493036
PMI	**0.581476**	0.487465

TABLE VIII
COMPARATIVO ARTIGO ORIGINAL COPA DATASET

Métrica	GPT-2-large	GPT-2-xl	Flan-T5-Large	mT0-large
UNC	0.556000	**0.560000**	**0.560000**	0.558000
LM	0.698000	**0.690000**	**0.746000**	0.584000
AVG	0.676000	**0.684000**	**0.748000**	0.598000
PMIdc	0.694000	**0.716000**	**0.772000**	0.618000

enganar modelos que dependem de plausibilidade superficial. As métricas PMI e PMIdc não mostraram ganhos expressivos, indicando que a dificuldade não se deve apenas ao surface form competition, mas à própria complexidade semântica e necessidade de conhecimento de mundo.

Por fim, no RACE-M, o FLAN-T5 (0,5919) novamente superou o mT0-large (0,4903), embora ambos tenham ficado abaixo de seu desempenho em COPA e COPA-REV. Isso sugere que a compreensão de textos mais longos e formais impõe desafios adicionais à arquitetura encoder-decoder em configuração zero-shot, possivelmente por exigir maior retenção de informações ao longo do contexto.

De modo geral, os resultados sugerem que:

1. A inversão de causa e efeito (COPA-REV) reduz substancialmente o impacto do surface form competition e permite que modelos encoder-decoder atinjam desempenho comparável ou superior aos modelos decoder-only do artigo original, mesmo com menos parâmetros.

2. O FLAN-T5 large, especializado em inglês, demonstra vantagem consistente sobre o mT0-large em tarefas monolíngues, especialmente quando há baixa interferência de ruído linguístico.

3. Datasets com forte viés adversarial (HellaSwag) ou textos longos e formais (RACE-M) permanecem como desafios significativos para ambas as arquiteturas na configuração zero-shot.

V. CONCLUSÃO

Os experimentos realizados foram importantes para expandirmos o entedimento sobre as arquiteturas de modelos de linguagem e sua influência no desempenho da tarefa desempenhada por eles. Similar ao que foi observado no artigo original, pudemos observar que a técnica PMIdc, no geral, é capaz de melhorar o desempenho dos modelos nas tarefas de múltipla escolha, apesar de algumas exceções. Observamos esse fato comparando as métricas LM e AVG com a PMIdc para cada dataset que avaliamos os nossos modelos.

À título de comparação trouxemos as métricas obtidas pelos modelos GPT-2-large e GPT-2-xl, os quais são os modelos mais próximos em números de parâmetros dos que utilizamos,

TABLE IX
COMPARATIVO ARTIGO ORIGINAL COPA-REV DATASET

Métrica	GPT-2-large	GPT-2-xl	Flan-T5-Large	mT0-large
UNC	0.5	0.5	0.5	0.5
LM	0.708000	**0.730000**	**0.780000**	0.620000
AVG	0.708000	**0.730000**	**0.780000**	0.620000
PMIdc	0.708000	**0.730000**	**0.780000**	0.620000

Um destaque interessante foi o Flan-T5, que obteve ótimas métricas que não só superaram o seu concorrente com número de parâmetros mais próximo, como também superou o GPT-2-xl, que possui 1,3 bilhões de parâmetros.

Apesar dos resultados positivos obtidos pelo Flan-T5, o mT0 não se destacou da mesma maneira, mesmo tendo um número de parâmetros maior. Esse fato enfraquece a formação da hipótese de que a arquitetura encoder-decoder se adapta melhor a uma situação de múltipla escolha, tendo uma menor vulnerabilidade à **surface form competition**. Apesar disso, acreditamos que mais testes são necessários para ampliar o conhecimento nessa área, visto que o desempenho de apenas 2 modelos não pode gerar uma conclusão sobre toda a classe de modelos que possuem a mesma arquitetura. Portanto, como extensão deste trabalho, sugerimos a continuação da aplicação de testes em modelos encoder-decoder, possuindo mais parâmetros, testados na base completa, e construídos de diferentes formas, a fim de coletar informações mais robustas sobre essa classe de modelos e seu desempenho neste tipo de tarefa.

REFERENCES

[1] Ari Holtzman, Peter West, Vered Shwartz, Yejin Choi, and Luke Zettlemoyer. Surface form competition: Why the highest probability answer isn't always right. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7038–7051, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.564. URL https://aclanthology.org/2021.emnlp-main.564/.