Capítulo 25

E agora, PLN?

Maria das Graças Volpe Nunes

Neste último Capítulo elencamos alguns desafios e perspectivas para o PLN em língua portuguesa e finalizamos com uma discussão sobre os limites atuais do PLN.

25.1 Desafios e perspectivas para o PLN-Português

Por razões históricas e econômicas, os sistemas atuais de PLN "estado da arte" são muito mais comuns em inglês do que em qualquer outra língua. Enquanto que outras comunidades têm adaptado para suas línguas os sistemas originalmente criados para o inglês (por meio de novos treinamentos, mas com aproveitamento de parâmetros), comunidades linguísticas minoritárias e comunidades linguísticas de países menos desenvolvidos são invisibilizadas no mundo digital, com consequências negativas e diretas na sua economia e desenvolvimento.

Segundo o Instituto Camões, em 2022, a comunidade de falantes de português no mundo era estimada em cerca de 260 milhões de pessoas (3,7% da população mundial) sendo o quarto idioma mais usado, depois do mandarim, inglês e espanhol. Contudo, essa representatividade não é contemplada no estado da arte da ciência, que está majoritariamente nas mãos de instituições e organizações não falantes do português. Pesquisadores brasileiros e portugueses têm levantado a necessidade de unir forças para colocar o português no lugar de destaque que ele merece[1].

O processamento do português brasileiro tem avançado de maneira consistente desde meados da década de 1990, principalmente a partir do uso de AM e de abordagens *cross-language* e multilíngue, que facilitam a construção rápida de recursos e soluções, e permitem a geração de uma aplicação em uma língua a partir de uma aplicação em outra língua. Mas ainda é precária a união de esforços entre os países da Comunidade de Países de Língua Portuguesa (CPLP), que inclui Portugal, Angola, Moçambique, Cabo Verde, Guiné-Bissau, São Tomé e Príncipe, além do Brasil. Se as diferenças linguísticas entre os diferentes idiomas representam barreiras para a criação de sistemas comuns, não há dúvida de que a união de esforços trará benefícios para todos. Por ora, o esforço mais visível é aquele entre os mais fortes do grupo, Brasil e Portugal, que realizam um evento científico bianual comum, o PROPOR[2], e mantêm vínculos acadêmicos há várias décadas. Dois grandes repositórios de recursos e ferramentas linguístico-computacionais do Português, que pretendem abranger

[1] https://www.publico.pt/2023/02/09/opiniao/opiniao/lingua-portuguesa-tecnologia-futuro-2038078

[2] CE-PLN. PROPOR (*International Conference on Computational Processing of Portuguese Language*). Disponível em: https://sites.google.com/view/ce-pln/eventos/propor.

as diversas comunidades de língua portuguesa são a Linguateca[3] e o Portulan Clarin[4].

Em países extensos como o Brasil, onde há uma grande variedade linguística, a exemplo das diferentes línguas indígenas faladas em território nacional[5], das variações dialetais e sociais e dos sotaques regionais do português brasileiro, suas riquezas e diversidades linguísticas dificilmente são representadas nos *corpora*. Essa sub-representação nos dados de treinamento de modelos de aprendizado de máquina é um dos fatores que contribuem para aumentar a codificação de vieses por esses sistemas. Percebe-se, portanto, a importância de os dados linguísticos que alimentam tais sistemas serem coletados de forma responsável, buscando representar as variações linguísticas e idiomáticas das línguas faladas no país.

Um dos primeiros *corpora* em português brasileiro usado para treinar um modelo de língua é o BrWac (*Brazilian Portuguese Web as corpus*), composto por 3,53 milhões de documentos da web, totalizando 2,68 bilhões de *tokens*, com acesso público para pesquisadores[6]. Já o *corpus* Carolina, do Centro de IA, C4AI[7], é, de acordo com os autores, "um corpus com um volume robusto de textos em Português Brasileiro contemporâneo (1970-2021), com informações de procedência e tipologia. O corpus está disponível em acesso aberto, para download gratuito, desde 8 de março de 2022. A versão atual, Ada 1.2 (8 de março de 2023), tem 823 milhões de *tokens*, mais de dois milhões de textos e mais de 11 GBs"[8]. Esse *corpus* é um importante passo para o treinamento de LLM do português brasileiro, e tem o mérito de incluir uma grande variedade de gêneros (jornalismo, literatura, poesia, judiciário, wikis, mídia social, legislativo, acadêmico etc.), mas ainda não contempla as diversidades regionais e culturais dessa língua, meta perseguida pelo C4AI com a construção do *corpus* de fala (transcrições) TaRSila[9], previsto para contemplar os diferentes dialetos brasileiros. Todos esses *corpora* pretendem ser variados quanto a gênero textual e domínio.

No mesmo C4AI, o projeto PROINDL[10] promete usar a IA em parceria com comunidades indígenas para o desenvolvimento de ferramentas que promovam a preservação, revitalização e disseminação de línguas indígenas do Brasil. Um dos objetivos é explorar as técnicas que utilizam poucos dados para criar tradutores automáticos tanto para texto como para fala, além de outras aplicações.

Mesmo com a limitação de variedade e tamanho de *corpora* em português para treinamento de LLMs, grandes modelos de língua para o português já são encontrados, quer sejam modelos com capacidade multilíngue (ex. os modelos PALM da Google), quer sejam treinados apenas em português (ex. BERTimbau(Souza; Nogueira; Lotufo, 2020), Sabiá (Pires et al., 2023), Albertina[11]). Dessa forma, são claros os avanços em direção a produtos para a língua portuguesa. No entanto, o que pode parecer simples (*corpus* + redes neurais e Transformers + *fine-tuning* = LLM) pode ser, de fato, inviável. O custo de se produzir um LLM de qualidade é extremamente alto. Um ótimo LLM, como

[3] https://www.linguateca.pt/
[4] https://portulanclarin.net/
[5] https://www.gov.br/funai/pt-br/assuntos/noticias/2022-02/brasil-registra-274-linguas-indigenas-diferentes-faladas-por-305-etnias
[6] https://www.inf.ufrgs.br/pln/wiki/index.php?title=BrWaC
[7] https://c4ai.inova.usp.br/pt/sobre/
[8] https://sites.usp.br/corpuscarolina/
[9] TaRSila. Disponível em: https://sites.google.com/view/tarsila-c4ai.
[10] https://c4ai.inova.usp.br/pt/pesquisas/#PROINDL_port
[11] Família de modelos treinados para as variantes européia e brasileira do português disponível em: https://huggingface.co/PORTULAN/albertina-ptbr-nobrwac.

o LLaMA-65B, por exemplo, foi pré-treinado com 1.4 trilhão de palavras, em 40 mil GPU[12]-horas, consumindo energia equivalente ao consumo de cerca de 10 casas brasileiras em um ano[13].

De um lado, são necessárias muitas GPUs para treinar modelos competitivos: quanto maior o número de GPUs, mais parâmetros podem ser usados no modelo, aumentando sua eficácia numa tarefa. Atualmente, poucas instituições públicas ou privadas dispõem de infraestrutura para tal e, ainda assim, com número de GPUs bastante inferior (de 2 a 100) àquela disponível em nuvem (clusters de TPUs[14]) com preços de aluguel que podem chegar a um milhão de dólares. Pesquisadores costumam recorrer a recursos gratuitos e temporários oferecidos pelas gigantes internacionais (ex. Google Cloud). Essa dependência externa por recursos essenciais ao desenvolvimento tecnológico só pode ser minimizada por meio de ações e investimentos governamentais (p.ex. centralizados pelo CNPq) ou por iniciativas coletivas dos detentores de recursos no sentido de juntá-los para incrementar o poder computacional e compartilhá-lo com toda a comunidade. De outro lado, independentemente do fator financeiro, temos o custo energético, com efeito na emissão de carbono, que, como vimos, não é desprezível.

Essas questões nos fazem refletir sobre os próximos caminhos a seguir. Nem tudo se resolve com grandes modelos de língua, assim como há muitas aplicações interessantes que podem ser desenvolvidas ou com modelos mais modestos ou por meios distintos dos modelos de língua. Considerando tarefas e domínios de conhecimento particulares, é possível construir soluções a partir de modelos treinados apenas nesse domínio. De fato, os resultados tendem a ser melhores do que com o uso de modelos mais genéricos. Além disso, considerar uma tarefa mais específica pode levar a uma solução - qualquer que seja a abordagem - mais eficaz.

As limitações para a academia não impedem, no entanto, que o PLN seja cada vez mais usado por empresas e startups da área, cujo número vem crescendo muito em nosso país. Certamente isso é fruto da alta demanda por sistemas dessa natureza, mas também do investimento das universidades públicas na formação de recursos humanos nessa área. Estamos vivendo um momento de grande absorção dos profissionais de PLN pelo mercado. Mais um motivo para refletirmos sobre a formação desses profissionais frente aos grandes desafios que essa área (e a IA de modo geral) nos coloca.

Além de todas as questões levantadas anteriormente, vale ressaltar a relevância de se adequar os critérios de avaliação tradicionalmente usados para sistemas de IA e, em particular, de PLN, à nova realidade das aplicações oferecidas à sociedade. A cultura acadêmica sugere uma avaliação em cenários rigidamente controlados, usando apenas métricas objetivas (numéricas), visando quase que exclusivamente a comparação com outros sistemas. Assim é a ciência e assim ela evolui. No entanto, tendo em vista o alcance que as novas tecnologias têm na sociedade, é urgente que os métodos de avaliação considerem critérios de outras naturezas, critérios que ajudem a prever o comportamento do sistema em situações, de fato, reais, sabidamente complexas, onde a imprevisibilidade é um fator relevante.

[12]Graphics Processing Unit – unidade de processamento gráfico.

[13]https://www.youtube.com/watch?v=prJrQ8XL-AY&ab_channel=BrasileirasemPLN

[14]TPUs (Unidades de Processamento de Tensor) são aceleradores de treinamento e geração de modelos de machine learning.

25.2 Há limites para o PLN?

A língua é frequentemente citada como sinal de inteligência e, por isso, nos distinguiria de outros animais. É por essa razão, aliás, que o PLN sempre esteve ligado à área de Inteligência Artificial. Sistemas dotados de habilidades linguísticas estariam entre os (artificialmente) inteligentes. No entanto, inteligência é algo difícil de se definir. Apenas parecer inteligente nos faz inteligentes? Essa questão sempre esteve presente na IA. Como definir um sistema inteligente? É necessário que ele raciocine como os humanos (seja bioinspirado), que tenha conhecimento explicitamente representado em seus algoritmos, ou basta que suas respostas sejam similares às de um humano nas mesmas situações? Não há acordo sobre isso, até porque sequer conseguimos concordar com os critérios de classificação de inteligência humana.

No caso do PLN, isso se traduz na seguinte questão: aos sistemas que mostram habilidade linguística pode-se atribuir inteligência? Ainda: eles de fato dominam o conhecimento total sobre a língua e todos os fenômenos que a língua em uso nos apresenta?

A língua tem sido objeto de estudo, análise e fascínio nas mais variadas áreas do conhecimento: filosofia, literatura, linguística, psicologia, psicanálise, ciências cognitivas, comunicação social, entre outras, e, recentemente, do PLN. Isso revela que a língua é um objeto de estudo bastante rico e complexo, e, portanto, não é possível abordá-lo segundo uma única disciplina.

O PLN tem sido apresentado como uma área comum a duas disciplinas, Computação e Linguística. No passado, isso parecia suficiente, pois apenas a porção formal, estrutural da língua era tratada computacionalmente[15]. Com o passar do tempo, a evolução das máquinas e as redes sociais, isso mudou. Essa língua em uso no cenário digital atual só pode ser tratada de forma transdisciplinar. Não é um caminho simples, nem cômodo, nem garantidor de que o PLN terá sucesso. Pelo contrário, não é improvável que, ao tratar a língua em toda sua complexidade, concluamos que há um limite para o PLN que independe de avanços tecnológicos.

Os capítulos anteriores evidenciam que PLN é uma área de grande potencial, porém repleta de desafios, sobre os quais é difícil fazer previsões. Várias tarefas de IA têm sido solucionadas pelas tecnologias atuais (Redes Neurais, Aprendizado de Máquina) que não são ideias novas; elas ficaram adormecidas até que o hardware das máquinas pudesse processá-las eficientemente. Em se tratando de PLN, no entanto, não é razoável prever que avanços de hardware, ou mesmo de métodos, garantam a solução completa para todos os sistemas que envolvem a língua. A demanda por sistemas que processam a língua não para de crescer. Vale notar que demandas e métodos são interdependentes: enquanto as demandas provocam novos métodos, estes últimos abrem caminho para novas demandas antes não possíveis.

Este livro também evidenciou que o desempenho linguístico dos sistemas atuais de PLN espelham aquilo que aprendem a partir dos dados de treinamento dos algoritmos de aprendizado: língua na norma culta, língua mal formada, discursos de ódio, misoginia ou racismo; o que quer que tenha sido oferecido ao algoritmo de aprendizado a título de exemplo eventualmente será reproduzido pelo sistema gerado. Como o conhecimento (a língua) adquirido nesses sistemas não é explicitamente representado (ele está imerso em

[15] A rigor, somente a parte formal da língua é passível de processamento pela máquina. Toda tentativa de alcançar o extralinguístico trata-se apenas de uma aproximação.

valores probabilísticos ou parâmetros numéricos das redes neurais), não há um controle de quando e como ele será usado. Todos esses efeitos colaterais dessa tecnologia preocupam a sociedade e trazem para a comunidade de PLN desafios e responsabilidades não existentes antes. As trajetórias da IA e do PLN têm nos ensinado que o alcance de metas mais modestas e realistas, ao longo do tempo, tem nos levado a patamares cada vez mais surpreendentes.

Convidamos você a esperar para ver, ou fazer para acontecer.

Referências

PIRES, R. et al. **Sabiá: Portuguese Large Language Models**. Anais da XII Brazilian Conference on Intelligent Systems - BRACIS 2023. **Anais**...2023. Disponível em: <https://arxiv.org/abs/2304.07880>

SOUZA, F.; NOGUEIRA, R.; LOTUFO, R. **BERTimbau: pretrained BERT models for Brazilian Portuguese**. (R. Cerri, R. C. Prati, Eds.)Proceedings of the 2020 Brazilian Conference on Intelligent Systems. **Anais**...Springer International Publishing, 2020.