Capítulo 36

E agora, PLN?

Maria das Graças Volpe Nunes

Publicado em: 26/09/2023
Atualizado em: 20/11/2024

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.

36.1 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.

A comunidade de falantes do português no mundo é estimada, em 2024, em cerca de 300 milhões de pessoas (3,7% da população mundial) sendo o quinto idioma mais usado, depois do inglês, mandarim, indi 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. No Brasil, os recursos de PLN compartilhados pela comunidade distribuem-se pelos centros de pesquisa,

[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.

sendo dois exemplos o NILC[3] e o C4AI[4]. Em Portugal, dois importantes repositórios de recursos e ferramentas para português europeu e brasileiro são a Linguateca[5] e o Portulan Clarin[6].

Em países extensos como o Brasil, onde há uma grande variedade linguística, a exemplo das diferentes línguas indígenas faladas em território nacional[7], 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[8]. Já o *corpus* Carolina, do Centro de IA, C4AI[9], 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". 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.). Já na área de língua falada, o projeto TaRSila[10] tem como meta a construção de *datasets* robustos para o alcance de resultados no estado da arte para tarefas de reconhecimento e síntese de fala, identificação do falante e clonagem de voz. Destacamos o grande e multipropósito *corpus* de áudios, CORAA[11], alinhado com transcrições e manualmente validado para o treinamento de modelos de reconhecimento e síntese, bem como de análise de sentimentos usando características acústicas.

No mesmo C4AI, o projeto PROINDL[12], em parceria com comunidades indígenas, investe no uso da IA para o desenvolvimento de ferramentas que promovam a preservação, revitalização e disseminação de línguas indígenas do Brasil. Entre as técnicas usadas estão as de AM 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 podem ser encontrados. São modelos com capacidade multilíngue (ex. os modelos PALM da Google) ou treinados apenas em português (ex. BERTimbau (Souza et al., 2020a), Sabiá (Pires et al., 2023b), Albertina[13]). Além desses, de caráter geral, vários outros *corpora* de domínios específicos

[3]https://sites.google.com/view/nilc-usp/resources-and-tools
[4]https://c4ai.inova.usp.br/pt/pesquisas_2/#NLP2_B_eng
[5]https://www.linguateca.pt/
[6]https://portulanclarin.net/
[7]https://www.gov.br/funai/pt-br/assuntos/noticias/2022-02/brasil-registra-274-linguas-indigenas-diferentes-faladas-por-305-etnias
[8]https://www.inf.ufrgs.br/pln/wiki/index.php?title=BrWaC
[9]https://sites.usp.br/corpuscarolina/
[10]TaRSila. Disponível em: https://sites.google.com/view/tarsila-c4ai.
[11]https://sites.google.com/view/tarsila-c4ai/coraa-versions
[12]https://conexoesoriginarias.github.io/
[13]Família de modelos treinados para as variantes européia e brasileira do português disponível em: https://huggingface.co/PORTULAN.

(nas áreas jurídica, médica, científica, etc.) têm sido criados visando aplicações mais direcionadas a essas áreas, como os capítulos anteriores evidenciam.

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 o LLaMA-65B, por exemplo, foi pré-treinado com 1.4 trilhão de palavras, em 40 mil GPU[14]-horas, consumindo energia equivalente ao consumo de cerca de 10 casas brasileiras em um ano[15]. Ainda, um estudo realizado em 2020 estimou que o custo de treinar um LLM com 1,5 bilhão de parâmetros seria da ordem de US$ 1,6 milhão.

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[16]) 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).

No ano de 2024, o Governo brasileiro anunciou um plano de incentivo para mitigar essa dependência externa por recursos essenciais ao desenvolvimento tecnológico. O Plano Brasileiro de Inteligência Artificial (Pbia) foi proposto por uma equipe multidisciplinar do Conselho Nacional de Ciência e Tecnologia (CNPq) e tem, como objetivos, "equipar o Brasil com infraestrutura tecnológica avançada com alta capacidade de processamento, incluindo um dos cinco supercomputadores mais potentes do mundo, alimentada por energias renováveis; desenvolver modelos avançados de linguagem em português, com dados nacionais que abarcam nossas características culturais, sociais e linguísticas, para fortalecer a soberania em IA e promover a liderança global do Brasil em IA por meio do desenvolvimento tecnológico nacional e ações estratégicas de colaboração internacional."[17] Como se vê, o PLN é um ator estratégico para o avanço tecnológico nacional.

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.

O PLN vem sendo cada vez mais usado por empresas e startups da área, cujo número vem crescendo muito em nosso país - em 2024 são mais de 13 mil. 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

[14]Graphics Processing Unit – unidade de processamento gráfico.
[15]https://www.youtube.com/watch?v=prJrQ8XL-AY&ab_channel=BrasileirasemPLN
[16]TPUs (Unidades de Processamento de Tensor) são aceleradores de treinamento e geração de modelos de machine learning.
[17]https://www.gov.br/mcti/pt-br/acompanhe-o-mcti/noticias/2024/07/cct-aprova-a-proposta-do-plano-brasileiro-de-inteligencia-artificial

área (e a IA de modo geral) nos coloca, bem como sobre nosso papel no esclarecimento à sociedade sobre seus limites.

36.2 Há limites para o PLN?

Os sistemas do estado-da-arte em PLN têm impressionado seus usuários com desempenho até melhores que os humanos em alguns cenários (Capítulo 20). No entanto, já conhecemos algumas de suas falhas e isso nos leva às seguintes questões: como saberemos se já alcançamos a(s) meta(s) do PLN? quais são essas metas? como saberemos quando alcançarmos um limite que impeça que essas metas sejam alcançadas? Essas perguntas necessitam que outras, mais fundamentais, sejam respondidas: O que é a linguagem? O que é significado? como se forma o significado? o que está envolvido numa comunicação? A forma como o computador representa e processa a linguagem permite que ele compreenda a linguagem tal como nós compreendemos?

Se o significado tem sido o maior gargalo do PLN por décadas, a IA generativa, por meio dos grandes modelos de língua, e apesar de grandes desafios a serem transpostos, parece ser o caminho mais promissor, aquele que mais aproxima a representação do significado como uma construção social, e não como uma etiqueta simbólica. A comunidade de PLN deve, portanto, seguir por esse caminho enquanto ele se mostrar promissor.

O PLN tem sido apresentado como uma área comum a duas disciplinas, Computação e Linguística e durante muito tempo olhar para e processar apenas a parte estrutural da língua pareceu possível ou mesmo suficiente. Com o passar do tempo, a evolução das máquinas e das técnicas, isso mudou e essa língua em uso no cenário digital atual só pode ser tratada de forma transdisciplinar. De fato, 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 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 ou duas disciplinas apenas.

O caminho para o sucesso do PLN não é simples, nem cômodo. 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 evidencia que os 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 valores probabilísticos ou parâmetros numéricos das redes neurais), não há um controle de quando e como ele será usado. Todos

esses efeitos colaterais preocupam a sociedade e trazem para a comunidade de PLN desafios e responsabilidades não existentes antes. E o esforço mundial para a regulamentação da criação e do uso de IA vai nos levar a cenários distintos dos atuais.

Convidamos você a nos acompanhar nessa jornada.

Referências

PIRES, R. et al. **Sabiá: Portuguese Large Language Models**. (M. C. Naldi, R. A. C. Bianchi, Eds.)Intelligent Systems. **Anais**...Cham: Springer Nature Switzerland, 2023.

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.