COMPREENDENDO A INTELIGÊNCIA ARTIFICIAL GENERATIVA NA PERSPECTIVA DA LÍNGUA

DUQUE-PEREIRA, Ives
Doutorando no Programa de Pós-Graduação em Cognição e Linguagem na Universidade Estadual do Norte Fluminense (UENF)
Bolsista CAPES
E-mail: ivesduque@gmail.com

MOURA, Sérgio
Professor no Programa de Pós-Graduação em Cognição e Linguagem na Universidade Estadual do Norte Fluminense (UENF)
E-mail: arruda@uenf.br

Resumo – A Inteligência Artificial Generativa (IA Gen) representa um marco na junção da Ciência da Computação e da Linguística, com foco expressivo no Processamento da Linguagem Natural (PLN). O artigo se debruça sobre o funcionamento da tecnologia de geração de texto, explorando suas interações com a língua. A IA Gen, cujo desenvolvimento remonta às décadas de 1940 e 1950, progrediu de sistemas avançados de computação para a especificidade do Aprendizado de Máquina. A IA, ao emular habilidades humanas, pode ser categorizada em abordagens centradas em "pensar" e/ou "agir" como humanos, com o teste de Turing ainda servindo como uma referência de validação de sucesso. A integração da IA com a ciência cognitiva oferece perspectivas sobre os processos mentais humanos, tornando a definição de "inteligência" fundamental. No contexto das ferramentas que operam no paradigma texto-para-texto, os Grandes Modelos de Língua (LLMs), como o GPT-4, LLAMA 2 e BERT, atestam a habilidade da IA Gen de criar e inferir conteúdo a partir de dados de treinamento. Estes modelos sustentam ferramentas como ChatGPT, Meta AI e BARD, baseadas na revolucionária Arquitetura de *Transformers*, introduzida em 2017 pela Google, tendo como um de seus fundamentos a semântica distribucional, que infere significados de palavras com base em seus contextos de uso. Os *Transformers* utilizam um mecanismo de atenção, permitindo-lhes perceber relações entre palavras e, consequentemente, prever sequências de palavras em um processo que envolve codificação e decodificação. Este mecanismo habilita os LLMs a interpretar e relacionar termos, capturando contextos completos e gerando interpretações mais precisas e naturais. No entanto, apesar de suas inovações, os LLMs enfrentam

desafios, como a geração de respostas sem sentido, conhecidas como "alucinações". A engenharia e o design de prompts são vitais para otimizar sua eficácia. Para concluir, a fusão da Inteligência Artificial com a Linguística sinaliza uma redefinição promissora na interação entre humanos e máquinas. Contudo, desafios relacionados à ética, privacidade e precisão dos dados ainda persistem e requerem contínua atenção e pesquisa. Assim, este trabalho visa elucidar os fundamentos da IA Generativa, particularmente no contexto da língua. Através desta exposição, busca-se não apenas educar os leitores sobre a natureza e capacidades da IA Gen, mas também fomentar reflexões sobre o papel da língua no universo das IAs generativas, pavimentando o caminho para futuras pesquisas.

Palavras-chave: Inteligência Artificial Generativa, Língua, Linguística, ChaGPT.

Introdução

A geração de texto por meio de Inteligência Artificial (IA) tem se tornado uma das áreas mais promissoras e desafiadoras da Ciência da Computação e Linguística Computacional. A capacidade de um Modelo de IA de entender e processar a língua e gerar textos de forma coerente e contextual é um testemunho da interseção entre tecnologia e humanidades.

Os Modelos de texto-para-texto, especificamente, referem-se à capacidade de uma IA de receber um texto/comando de entrada (*input*) chamado de "*prompt*" e produzir um texto de saída (*output*) em resposta. Esta técnica, que dá origem ao que chamamos de Inteligência Artificial Generativa (IA Gen), tem aplicações vastas, como robôs de conversação (*chatbots*) e criação de conteúdo textual de todo tipo. Existem outros Modelos como texto-para-imagem, texto-para-3D, texto-para-tarefa e texto-para-vídeo, mas iremos focar no presente trabalho no Modelo de texto-para-texto.

Língua e linguagem são termos frequentemente usados de forma intercambiável, mas, na realidade, possuem significados distintos, especialmente quando analisados sob a perspectiva da Linguística. Segundo Martelotta (2011), a linguagem pode ser entendida, num sentido mais amplo, como capacidade dos seres humanos de se comunicar e expressar pensamentos, emoções, desejos e informações. Nesse sentido, as línguas naturais, como o português, são formas de linguagem utilizadas como instrumento de comunicação, mas não o único. A linguagem não se restringe apenas ao domínio verbal ou escrito, mas abrange uma

variedade de modalidades, incluindo gestos, expressões faciais, movimentos do corpo e até mesmo música e outras formas de arte.

Para os linguistas, a linguagem é uma habilidade estritamente humana de se comunicar, podendo ser pelo uso da língua, ou seja, uma faculdade cognitiva. Nestes termos, língua é um sistema específico de signos e regras usado por uma comunidade para se comunicar. É uma manifestação particular da linguagem e está sujeita a regras gramaticais, léxicas e fonéticas. Cada país ou região pode ter sua própria língua – em algumas vezes mais de uma – e essa língua ter diversos dialetos ou variações regionais. A língua carrega características de especificidade de regras e estrutura, padronização gramatical e lexical com variações por mudanças ao longo do tempo e espaço. (MARTELOTTA, 2011, p. 16).

A distinção entre os termos "língua" e "linguagem" é crucial para a introdução deste trabalho, dada a natureza interdisciplinar da Inteligência Artificial (IA). Em um campo onde diversas áreas convergem, a falta de padronização terminológica pode levar a ambiguidades e mal-entendidos que queremos evitar. A IA Gen que opera através do que comumente é chamado de Processamento de Linguagem Natural (PLN), ilustra bem essa questão.

Enquanto "Processamento de Linguagem Natural" é uma nomenclatura comum no campo da Computação e amplamente aceita em outras áreas, alguns pesquisadores da Linguística preferem "Processamento de Língua Natural". Esta última denominação enfatiza a "língua" como o objeto do processamento, enquanto a "linguagem" é vista mais amplamente como uma capacidade cognitiva, ou seja, processos que estão na base de um instrumento de comunicação.

Similarmente, o termo "Modelo de Linguagem" é amplamente adotado, mas não sem contestações. Há quem defenda "Modelo de Língua" como uma alternativa mais precisa. Devido à natureza do presente trabalho ter como objeto a IA Gen de texto-para-texto e seu campo de inserção na Linguística, optamos pela abordagem de adotar os termos “Processamento de Língua Natural” (PLN) e “Modelo de Língua” (ML). Nesse contexto, só será adotada, no presente trabalho, o termo “Linguagem” no lugar de “Língua” quando se tratar de uma referenciação a um autor que utiliza desta forma.

Contudo, ressalta-se que a evolução tecnológica da IA Gen têm se encaminhado para o uso de ferramentas multimodais, que trabalham não somente com a língua, mas outros tipos de linguagem como imagens, sons, movimentos corporais, etc. Assim, nesses outros contextos, o uso de “linguagem” no lugar de “língua” pode ser o mais adequado. Ainda se trata de um campo de debates e de necessária delimitação teórica.

No Brasil, onde a diversidade Linguística e cultural é imensa, a capacidade de geração de texto de forma autônoma e contextualizada por ferramentas de IA pode ser uma maneira valiosa para superar barreiras da língua, promover a inclusão e expandir o acesso à informação. Além disso, a geração de texto-para-texto, pela IA Gen, representa um desafio técnico e ético. Técnico, pois exige Modelos cada vez mais sofisticados e treinados em grandes volumes de dados para garantir precisão e fluidez. Ético, pois levanta questões sobre autoria, veracidade e potencial de uso indevido da tecnologia.

Nosso trabalho se debruça sobre um campo emergente de estudos que é o da Inteligência Artificial Generativa na perspectiva da língua. Neste contexto, nosso principal objetivo é compreender os fundamentos e funcionamento da IA Generativa e as interseções com a Língua, que possibilita a utilização de ferramentas voltadas à geração texto-para-texto, como o *ChatGPT*. Através deste trabalho, almejamos não apenas capacitar os leitores a compreenderem o que é uma IA Gen, mas instigar uma reflexão profunda sobre o papel da língua no universo das IAs generativas, abrindo novos campos de pesquisa.

1. Inteligência Artificial e seu fundamento na Linguística

A Inteligência Artificial (IA) representa um dos campos mais inovadores das Ciências e Engenharia. Surgindo logo após a Segunda Guerra Mundial, ainda na década de 1940 a partir do desenvolvimento da lógica, o termo "Inteligência Artificial" foi estabelecido durante a década de 1950. No entanto, definir IA é uma tarefa complexa e em constante evolução. Para Carle (2023), IA é um termo usado constantemente para definir sistemas avançados de computador. Entende-se a IA especificamente

como Aprendizado de Máquina ("*machine learning*"), sistemas de computador com habilidade de aprender depois de conhecer exemplos.

A IA pode ser conceituada como uma especialização dentro da ciência da computação, caracterizada pela elaboração e aprimoramento de sistemas computacionais que emulam habilidades tradicionalmente associadas à inteligência humana. Estes sistemas são projetados para executar tarefas que, em circunstâncias normais, dependeriam do discernimento e raciocínio humanos. Um dos pilares fundamentais da IA é o Aprendizado de Máquina. Esta vertente se concentra na criação de programas e sistemas que, ao serem alimentados com dados, são capazes de treinar Modelos de IA. O diferencial do Aprendizado de Máquina reside na capacidade de permitir que os computadores aprendam e se adaptem a partir dos dados fornecidos, sem a necessidade de instruções explícitas. Assim, a máquina não apenas processa informações, mas também evolui e se aperfeiçoa com base nas experiências e nos dados que assimila. (GOOGLE, 2023)

Quando observamos a IA sob a lente dos processos de pensamento, encontramos duas abordagens principais. A primeira enxerga a IA como uma entidade capaz de "pensar" à semelhança dos seres humanos, quase como máquinas dotadas de mentes. A segunda abordagem percebe a IA de uma forma mais racional, onde Modelos computacionais emulam faculdades mentais, como percepção, raciocínio e ação. Outra abordagem conceitual se concentra no "comportamento". Nesse contexto, a IA é vista tanto como uma entidade que age como o ser humano, realizando tarefas que, quando executadas por humanos, requerem inteligência, quanto como uma entidade que age racionalmente, utilizando "inteligência" computacional para cumprir suas funções programadas. (NORVIG, RUSSELL, 2013, p. 33)

Ao longo da história da IA, diversas estratégias e abordagens foram adotadas por diferentes pesquisadores, cada uma com seus métodos e focos. Enquanto uma abordagem centrada no ser humano pode ser vista como uma ciência empírica, baseada em hipóteses e experimentação, uma perspectiva mais racionalista combina elementos da matemática e engenharia para entender e desenvolver a IA. Nosso interesse recai sobre as abordagens de IA que procuram entender e desenvolver a tecnologia a partir de um "agir" e "pensar" como seres humanos.

Se tratando de “agir” como seres humanos, Alan Turing, em 1950, introduziu uma proposta revolucionária que buscava estabelecer uma definição operacional para a Inteligência Artificial. Esse conceito, conhecido como o "teste de Turing", sugeria que um computador poderia ser considerado "inteligente" se, ao ser interrogado por um humano através de perguntas escritas, suas respostas fossem indistinguíveis daquelas de um ser humano. Ou seja, se o interrogador não conseguisse discernir se as respostas provinham de uma máquina ou de uma pessoa, o computador seria considerado bem-sucedido no teste. (TURING, 1980, p. 434)

Existe uma extensão mais abrangente desse teste, denominada "teste de Turing total". Nesta versão, não se limita apenas à interação textual, mas envolve uma interação física direta entre o interrogador e o computador. Para ser bem-sucedido neste teste ampliado, o computador deve possuir habilidades em seis disciplinas fundamentais. Primeiro, deve ter a capacidade de Processamento de Linguagem Natural, permitindo a comunicação em um idioma comum. Segundo, precisa de Representação de Conhecimento, armazenando e recuperando informações. Terceiro, deve ser capaz de Raciocínio Automatizado, utilizando as informações armazenadas para responder perguntas e formular conclusões. Além disso, o Aprendizado de Máquina é essencial para que o computador se adapte e evolua com base em novos dados. A Visão Computacional, que permite ao computador reconhecer e interpretar objetos, e a Robótica, que lhe confere a habilidade de manipular objetos e se movimentar, são as duas últimas disciplinas necessárias. Estas seis áreas formam a espinha dorsal da Inteligência Artificial. Portanto, mesmo após décadas desde sua proposição, o teste de Turing continua sendo uma referência significativa e relevante na avaliação e compreensão da IA. (NORVIG, RUSSELL, 2013, p. 35)

A Inteligência Artificial busca emular, em muitos aspectos, o "pensar" humano, situando-se na confluência com a ciência cognitiva, um campo interdisciplinar que estuda os processos mentais. No entanto, antes de avançarmos na compreensão da IA, é fundamental abordar a própria definição de "inteligência", um conceito que, até hoje, não alcançou um consenso na comunidade científica. Para alguns psicólogos cognitivos, inteligência é a habilidade inerente ao ser humano de aprender a partir de

experiências vivenciadas. Por outro lado, há aqueles que a definem com base na interação do indivíduo com seu ambiente, considerando as respostas e adaptações a estímulos externos.

Nesse debate, a contribuição de Jean Piaget destaca-se. Ele elucidou o processo pelo qual os indivíduos assimilam aspectos do ambiente com base em aprendizados anteriores e, simultaneamente, adaptam-se, alterando comportamentos. Esse ciclo contínuo de assimilação e acomodação leva ao desenvolvimento de estruturas cognitivas mais complexas, caracterizando a inteligência como uma entidade dinâmica e em constante evolução. (MEDEIRO, 2018, p. 19)

Dessa forma, a IA pode ser entendida como uma tentativa de replicar, em sistemas programáveis, os mecanismos intrincados que operam no cérebro e na mente humana. A busca é por criar máquinas que não apenas processem informações, mas que também "pensem" e "aprendam" de maneira análoga aos seres humanos. Assim, podemos considerar o “pensar” da IA como sendo uma simulação da inteligência humana caracterizada pela capacidade de resolução de problemas, aprendizado com o ambiente, desenvolvimento do estruturas “cognitivas”, orientação e metas. (MEDEIRO, 2018, p. 19)

Como um campo profundamente interdisciplinar, a IA recebeu contribuições, perspectivas e técnicas de diversas áreas do conhecimento. Segundo Norvig e Russell (2013), a IA é fundamentada em pilares que abrangem filosofia, matemática, economia, neurociência, psicologia, engenharia de computadores, teoria de controle e cibernética e a Linguística.

É justamente na intersecção da IA com a Linguística que nosso interesse se acentua. Norvig e Russell (2013) sublinham a relação entre linguagem e pensamento. Eles abordam Modelos behavioristas de aprendizado linguístico e, em contrapartida, a crítica incisiva de Chomsky, que introduziu sua teoria de estruturas sintáticas. Essas discussões pavimentaram o caminho para a evolução da Linguística moderna, culminando na interação entre a IA e a Linguística. Dessa confluência, nasceu um campo híbrido: a Linguística Computacional, também conhecida como Processamento de Linguagem Natural. Aqui, o foco recai sobre a compreensão

profunda dos temas em seus contextos específicos, transcendendo a mera análise estrutural das frases e buscando uma compreensão mais holística e integrada da linguagem.

Segundo Sandoval (2016), a Linguística Computacional e o Processamento de Linguagem Natural são o mesmo: estudo, compreensão e desenvolvimento de sistemas que emulem a capacidade Linguística inerente aos seres humanos. Inseridas no contexto da Inteligência Artificial, dedicam-se ao estudo e à criação de sistemas computacionais capazes de compreender e gerar línguas naturais. O propósito central é elaborar Modelos computacionais da linguagem com tal riqueza de detalhes que viabilizem o desenvolvimento de softwares aptos a executar diversas tarefas que envolvam a linguagem natural. Essa busca por simular a complexidade da linguagem humana em máquinas é um desafio contínuo e fascinante, demonstrando o potencial e a amplitude da intersecção entre Linguística e tecnologia.

Othero e Menuzzi (2005) conceituam a Linguística Computacional como um campo interdisciplinar que investiga as interseções entre a Linguística e a informática. Esta interação possibilita o desenvolvimento de sistemas capazes de reconhecer e gerar linguagem natural. Segundo esses autores, o Processamento de Linguagem Natural (PLN) é uma especialização dentro da Linguística Computacional, coexistindo com a Linguística de Corpus. Enquanto a Linguística de Corpus se dedica à coleta e análise de vastos conjuntos de dados linguísticos, conhecidos como corpora, representativos da linguagem natural, o PLN foca na aplicação prática desses estudos. Seu objetivo é desenvolver softwares e sistemas computacionais avançados, como os robôs de conversação, popularmente chamados de *chatbots*. Portanto, a distinção e complementaridade entre essas subáreas evidenciam a amplitude e profundidade da Linguística Computacional no cenário contemporâneo.

Desenvolver uma estrutura de língua natural que seja ao mesmo tempo robusta e fluente é um desafio complexo. Isso porque a língua, em sua essência, é multifacetada, englobando aspectos sintáticos, semânticos e discursivos que se manifestam de maneira única em cada uso. O PLN, em sua essência, é uma interseção da Linguística com a ciência da computação. Seu principal objetivo é capacitar máquinas para entender, interpretar e responder à língua humana de

maneira coerente e contextual. O PLN não é apenas uma técnica, mas um vasto campo de estudo que abrange desde a análise sintática de sentenças até a extração de sentidos dos textos.

Dentro do campo da análise temos o processamento de sentenças isoladas e a determinação de sua inserção em uma estrutura discursiva. A análise de sentenças subdivide-se em análise sintáxica e análise semântica. É essencial determinar como cada sentença se encaixa em uma estrutura discursiva mais ampla. Além disso, a análise de uma única sentença pode ser desmembrada em duas vertentes: a análise sintáxica, que se preocupa com a estrutura gramatical, e a análise semântica, voltada para o significado. O refinamento e a expansão contínua dessas estruturas analíticas são fundamentais no processo de engenharia do Processamento de Linguagem Natural (PLN), demonstrando a complexidade e a riqueza desse campo de estudo. (GRISHMAN, 1986, p. 8)

Em meio a essa complexidade, a Linguística Computacional e o PLN têm se mostrado ferramentas indispensáveis para a evolução da interação humano-computador. À medida que a tecnologia avança, a necessidade de sistemas que compreendam e respondam de forma mais humana à língua natural torna-se cada vez mais premente. A capacidade de analisar não apenas a estrutura, mas também o significado e o contexto das sentenças, é o que diferencia os sistemas de PLN de alta qualidade dos demais. A integração entre a análise sintáxica e semântica, conforme destacado por Grishman (1986), é apenas uma das muitas camadas que compõem este campo intrincado. Assim, ao olharmos para o futuro, podemos antecipar um cenário em que a Linguística Computacional desempenhará um papel ainda mais central na construção de uma interação mais fluida e natural entre humanos e máquinas, reafirmando sua relevância e potencial transformador no mundo da tecnologia.

2. Inteligência Artificial Generativa

Segundo Carle (2023), os Modelos generativos, também conhecidos como Inteligência Artificial Generativa vão além da simples previsão, sendo capazes de gerar conteúdos completamente novos, fundamentados nos dados com os quais

foram treinados. Isso só é possível pelo desenvolvimento da Aprendizagem de Máquina.

O Aprendizado de Máquina, uma subárea da IA, refere-se à capacidade de sistemas computacionais de aprender e se adaptar a partir de dados, sem serem explicitamente programados para uma tarefa específica. Em essência, é o processo no qual os computadores desenvolvem habilidades para reconhecer padrões, tomar decisões e realizar predições com base nas informações que lhes são fornecidas. (GÉRON, 2019, p. 4)

Segundo Géron (2019), dentro do Aprendizado de Máquina, existem diferentes abordagens que determinam como os sistemas aprendem, podendo ser classificados entre aqueles que aprendem com a supervisão humana, grau de rapidez e por comparação. Assim, temos o aprendizado supervisionado, não supervisionado, semi-supervisionado, por reforço, online e em lote, baseado em Instância Versus e o baseado no Modelo. Veremos a seguir aqueles que se destacam mais no campo da IA Gen.

O aprendizado supervisionado é uma dessas abordagens, em que o Modelo é treinado com um conjunto de dados previamente rotulado. Segundo Géron (2019), uma tarefa típica do aprendizado supervisionado é a classificação, como um filtro de spam no e-mail. Rotulação se refere ao treinamento de um sistema como, por exemplo, para reconhecer imagens de gatos, fornecendo várias imagens já identificadas. Nesse exemplo, o sistema sabe de antemão quais imagens são de gatos e quais não são. O sistema, então, aprende a distinguir gatos com base nesses dados rotulados.

Por outro lado, no aprendizado não supervisionado, o sistema é exposto a dados não rotulados e deve identificar padrões ou agrupamentos por conta própria. Por exemplo, um conjunto de dados com informações sobre hábitos de compra de clientes. Sem dizer explicitamente ao sistema quais clientes têm comportamentos similares, ele pode identificar grupos de consumidores com preferências semelhantes.

Já o aprendizado semi-supervisionado combina elementos das duas abordagens anteriores. Utiliza-se tanto dados rotulados quanto não rotulados para treinar o Modelo. Por exemplo, ao treinar um sistema para reconhecer a fala, podemos

fornecer algumas gravações transcritas (rotuladas) e outras não transcritas. O sistema aprende com as informações explícitas das gravações rotuladas e generaliza esse aprendizado para as gravações não rotuladas, melhorando sua capacidade de reconhecimento. Géron (2019) exemplifica com o serviço de hospedagem de fotos Google Fotos que, ao carregar fotos da sua família, o aplicativo reconhece um membro automaticamente em algumas fotos (aprendizado não supervisionado) pelas características comuns, mas também necessita te perguntar se trata-se do mesmo membro da família em outras fotos (aprendizado supervisionado).

A evolução e eficácia dos sistemas de Inteligência Artificial que observamos atualmente são amplamente atribuídas à implementação de técnicas avançadas de Aprendizado de Máquina, em particular, o "Aprendizado Profundo" ou "*Deep Learning*". Esta técnica se destaca pelo uso de Redes Neurais Artificiais (RNA), estruturas computacionais inspiradas na complexidade e funcionalidade do cérebro humano. Essas redes são compostas por uma série interconectada de "nós" ou "neurônios" que, assim como os neurônios biológicos, são capazes de processar informações, aprender com elas e, posteriormente, realizar predições ou classificações. (GÉRON, 2019, p. 259)

O Aprendizado Profundo, representa uma evolução no campo do Aprendizado de Máquina. Utilizando Redes Neurais Artificiais com múltiplas camadas, o aprendizado profundo é capaz de processar grandes volumes de dados e identificar padrões complexos, tornando-se uma ferramenta poderosa para tarefas como reconhecimento de imagem, tradução automática e, claro, compreensão de língua natural. O Aprendizado de Máquina e suas diversas abordagens representam a vanguarda da tecnologia, permitindo que sistemas computacionais se aproximem cada vez mais da capacidade cognitiva humana.

Segundo Géron (2019), o que torna o Aprendizado Profundo particularmente poderoso é a sua capacidade de utilizar múltiplas camadas de neurônios artificiais que executam equações matemáticas binárias de entrada e saída de informações. Estas camadas permitem que o sistema aprenda tanto com dados previamente rotulados, que fornecem informações específicas sobre uma tarefa, quanto com dados não rotulados.

Este método semi-supervisionado, é fundamental porque, ao rotular apenas uma fração dos dados e permitir que a RNA processe também os dados não rotulados, potencializa-se a capacidade da Rede de generalizar e adaptar-se a novos exemplos e cenários. Assim, o aprendizado profundo não apenas otimiza a performance dos sistemas de IA, mas também amplia seu escopo de aplicação e adaptabilidade.

A IA Gen representa uma evolução notável no campo da aprendizagem profunda. Essencialmente, ela opera por meio de redes neurais que processam tanto dados rotulados quanto não rotulados, abrangendo abordagens supervisionadas, não supervisionadas e semi-supervisionadas. O cerne da IA Gen é sua capacidade de gerar novos conteúdos, fundamentados em aprendizados de conjuntos de dados anteriores. Esta capacidade não surge aleatoriamente, mas é moldada por sofisticados Modelos estatísticos e equações matemáticas.

Quando interagimos com uma IA Gen, fornecendo-lhe um comando ou "*prompt*", ela não simplesmente responde com informações pré-programadas. Em vez disso, recorre ao seu Modelo estatístico para antecipar e gerar uma resposta (*output*) que se alinhe ao contexto e sequência das palavras fornecidas. Esta previsão é um testemunho da complexidade e profundidade de seu treinamento.

O PLN permitiu o surgimento dos Modelos de Língua que representam uma evolução significativa na capacidade de sistemas computacionais de prever a próxima palavra em uma sequência, garantindo que o resultado faça sentido no contexto. Esses Modelos são aprimorados por meio de treinamento com vastos volumes de texto, utilizando-se de técnicas probabilísticas e cálculos matemáticos para determinar a palavra mais adequada para um determinado contexto. Naturalmente, quanto mais abrangente e diversificado for o conjunto de dados utilizado para treinar o Modelo, mais refinadas e precisas serão suas previsões. (CARLE, 2023)

À medida que a tecnologia avançou, os Modelos de Língua evoluíram para o que agora chamamos de Grandes Modelos de Língua (LLM, do inglês "Large Language Models"). Estes LLMs são caracterizados por terem bilhões ou até trilhões de parâmetros, tornando-os extremamente poderosos em termos de capacidade de processamento e precisão. Eles são capazes de gerar textos coerentes, responder perguntas, traduzir idiomas e até mesmo escrever código de programação.

O LLM é fundamental para a IA Generativa, pois, ao contrário de Modelos mais simples que apenas classificam, sugerem ou reconhecem padrões, os LLMs têm a capacidade de criar, de gerar novos conteúdos baseados em seu treinamento. Eles não apenas reproduzem o que foi visto em seus dados de treinamento, mas também fazem inferências, combinam informações de diferentes contextos e geram respostas originais. Os Grandes Modelos de Língua, ao serem alimentados com combinações extensas de texto, têm a capacidade de transformá-las, gerando conteúdos inéditos e relevantes, refletindo a complexidade e o potencial da interação entre Linguística e tecnologia. (CARLE, 2023)

No cenário contemporâneo da Inteligência Artificial Generativa, destacam-se os Grandes Modelos de Língua que têm revolucionado o campo do Processamento da Língua Natural. Entre os mais importantes, encontramos o GPT-4, que é a força motriz por trás do *chatGPT* desenvolvido pela OpenAI, o PALM-2/Bard da Google, o Claude da Antropic e o LLAMA 2 da Meta. É importante ressaltar que, dentre esses Modelos, o LLAMA 2 se distingue por ser *open source*. Isso significa que seu código é aberto e acessível ao público, permitindo que qualquer indivíduo possa acessar, copiar, modificar e redistribuir o software sem custos. Esta característica fomenta uma cultura colaborativa, na qual uma comunidade de desenvolvedores contribui para aprimoramentos e criação de novas aplicações. A disponibilidade e transparência de um Modelo como o LLAMA 2 não apenas democratiza o acesso à tecnologia, mas também incentiva a inovação e a disseminação do conhecimento no campo da Inteligência Artificial.

Embora PLN e o LLM sejam termos frequentemente usados de forma intercambiável, é essencial entender suas diferenças e semelhanças para compreender a amplitude e profundidade da tecnologia Linguística contemporânea. O PLN é uma disciplina que se concentra na interação entre computadores e a língua humana. Seu objetivo é permitir que as máquinas entendam, interpretem e respondam à língua natural de maneira que se assemelhe à compreensão humana.

Por outro lado, os LLM são uma abordagem específica dentro do PLN. Eles são Modelos de aprendizado profundo treinados em vastos volumes de texto, como os conhecidos GPT, PALM e LLMA. O que os tornam notáveis é sua habilidade de

capturar as nuances e o contexto da língua, gerando respostas que, em muitos casos, são quase indistinguíveis das produzidas por seres humanos. O que realmente diferencia os LLM de outras abordagens no PLN é sua capacidade de processar e gerar língua natural com uma precisão sem precedentes. Eles não apenas reconhecem padrões linguísticos, mas também conseguem prever e gerar sequências de palavras de forma coerente e contextualizada. (CHANG et al., 2023, p. 23)

No entanto, é fundamental reconhecer que, apesar de suas distinções, PLN e LLM estão profundamente interligados. Os LLM podem ser vistos como a realização máxima do que o PLN aspira alcançar. Ambos compartilham a missão de aprimorar a comunicação entre humanos e máquinas. Enquanto o PLN fornece as bases teóricas e metodológicas, os LLM elevam esses princípios a um novo patamar, processando informações em uma escala que seria inimaginável para o ser humano.

3. Inteligência Artificial Generativa de texto-para-texto

No vasto universo da Inteligência Artificial, a introdução dos "*Transformers*" em 2017, pela empresa Google, marcou um divisor de águas no campo do PLN fazendo surgir um dos primeiros LLM que foi o BERT. Estes mecanismos, apesar de seu nome sugestivo, não têm relação com robôs que se transformam, mas sim com uma abordagem inovadora de interpretar e gerar linguagem. Segundo Vaswani *et. al.* (2017), os *Transformers* utilizam o chamado mecanismo de atenção, inspirado na faculdade cognitiva humanada da atenção, o Modelo aprende a ver a relação entre as palavras, selecionando as mais importantes para a compreensão geral das sentenças. A partir disso, o Modelo de Língua se utiliza da probabilidade para prever uma sequência de palavras.

Os *Transformers* são estruturados em duas partes fundamentais: codificação e decodificação. Quando uma informação é fornecida como entrada, ela passa por um processo de codificação, transformando-a em uma representação intermediária. Posteriormente, essa representação é decodificada, resultando na saída desejada. Esta abordagem de dois estágios permite que os *Transformers* compreendam e produzam linguagem com uma precisão e contexto sem precedentes, solidificando

sua posição como uma das ferramentas mais poderosas no arsenal da IA Gen. (GOOGLE, 2023)

Ao processar essa frase "Os leões, conhecidos como os reis da selva, são predadores formidáveis e têm uma relação especial com as hienas, que muitas vezes competem pelo mesmo alimento.", um Modelo baseado em *Transformers*, como o GPT, usaria o mecanismo de atenção para identificar a relação entre "leões" e "reis da selva", bem como entre "leões" e "hienas". Ele entenderia que a palavra "predadores" é crucial para compreender a natureza dos leões e que "competem" indica uma relação de rivalidade entre leões e hienas. Se, por exemplo, quiséssemos prever a próxima palavra após "leões são", o Modelo poderia sugerir palavras como "animais", "predadores" ou "grandes", baseando-se na probabilidade e no contexto fornecido pela frase. Esta capacidade de entender e prever com precisão é o que torna os *Transformers* tão importantes no campo do Processamento da Linguagem Natural. (OPENAI, 2023)

Nesta fase, a frase é analisada e transformada em uma representação intermediária, uma espécie de código que encapsula o significado e o contexto da frase original. Esta representação não é algo legível para humanos, mas é extremamente valiosa para a máquina, pois contém todas as nuances e particularidades da frase de entrada. Ou seja, antes de ser processada por um *Transformer,* essa frase seria dividida em *token*s, que são palavras – ou suas partes - que compõe uma sentença. Em um sistema simples, os *token*s poderiam ser ["Os", "leões", ",", "conhecidos", "como", "os", "reis", "da", "selva", ",", "são", "predadores", "formidáveis", "e", "têm", "uma", "relação", "especial", "com", "as", "hienas", ",", "que", "muitas", "vezes", "competem", "pelo", "mesmo", "alimento", "."]. Cada *token* é então convertido em uma representação numérica, geralmente chamada de "*embedding*", que captura seu significado e relação com outros *token*s.

Após a fase de codificação, o processo avança para a decodificação. Com base na representação intermediária criada, o sistema inicia a tarefa de interpretar o significado e formular uma resposta apropriada. Durante a codificação, cada fragmento de informação, ou *token*, é avaliado levando em consideração tanto seu significado intrínseco quanto o contexto proporcionado pelos *token*s adjacentes. Por

exemplo, ao analisar um *token* específico, o sistema pondera sua relevância em relação aos *token*s que vêm antes e depois dele. Na sequência, a fase de decodificação se vale dessas representações ricas em contexto para construir uma resposta coerente e contextualizada como saída (*output*).

Os *Transformers* não se limitam a considerar os *token*s adjacentes, mas sim toda a sequência, permitindo uma compreensão mais profunda do contexto global. Isso é crucial para idiomas onde a ordem das palavras pode alterar significativamente o significado, ou onde um *token* no início da frase pode influenciar um *token* no final. Há uma capacidade de considerar o contexto global da frase, em vez de se limitar a analisar cada palavra isoladamente.

Isso significa que eles podem capturar nuances e ambiguidades, resultando em traduções e interpretações mais precisas e naturais. Uma das razões para sua eficácia é a maneira como lidam com a língua em sua forma mais granular, através de *token*s. Esse método de *tokenização* permite que os *Transformers* processem informações com uma granularidade muito mais detalhada, levando em consideração as sutilezas da língua que muitas vezes são perdidas em abordagens tradicionais.

Os *Transformers*, com sua capacidade avançada de capturar contextos e nuances Linguísticas, representam um marco na evolução do PLN. No entanto, para compreender plenamente a profundidade e a complexidade da língua humana, é essencial abordar outro conceito fundamental: a semântica distribucional. Esta teoria, que se baseia na premissa de que o significado de uma palavra pode ser inferido pelo contexto em que aparece, oferece uma perspectiva única sobre como as palavras adquirem significado e como esse significado é representado em Modelos computacionais.

A semântica distribucional é uma abordagem empírica para modelar o significado das palavras, baseada na premissa de que o significado de uma palavra pode ser inferido a partir dos contextos em que ela ocorre. Esta abordagem tem suas raízes na Hipótese Distribucional, que postula que palavras que aparecem em contextos semelhantes tendem a ter significados semelhantes. Assim, o significado de uma palavra está atrelado as palavras que se seguem e antecedem. Originando-se de tradições estruturalistas, a semântica distribucional moderna emprega técnicas

computacionais avançadas e grandes conjuntos de dados para criar Modelos de espaço vetorial. Estes Modelos são capazes de capturar relações semânticas complexas, variações contextuais e nuances de significado que são desafiadoras para abordagens Linguísticas mais tradicionais (BOLEDA, 2020, p. 3)

Pavlick (2022) ressalta que, no cerne do Aprendizado Profundo, encontra-se a aderência à hipótese distribucional do significado. Compreender que o significado semântico de uma palavra pode ser deduzido a partir das estatísticas de co-ocorrência com outras palavras no texto, oferece melhoramentos para o funcionamento das RNA.

As palavras que aparecem em contextos semelhantes tendem a ter significados semelhantes. Por exemplo, as palavras "gato" e "felino" frequentemente coexistem em textos que discutem animais domésticos, e, portanto, podem ser consideradas semanticamente próximas. Da mesma forma, "livro" e "página" são frequentemente encontrados juntos em contextos literários, indicando uma relação semântica entre eles. (OPENAI, 2023)

A IA Generativa texto-para-texto, por sua vez, refere-se a sistemas que são capazes de criar conteúdo textual novo e original com base em dados de treinamento prévios. A aplicação da semântica distribucional na IA Gen pode ser vista em Grandes Modelos de Língua avançados que geram texto. Estes Modelos, ao serem treinados em grandes volumes de texto, aprendem as relações semânticas entre palavras com base em seus contextos de co-ocorrência.

Ao alimentar um Modelo generativo com a frase "O gato está no" (*input*), é provável que ele complete a frase com "telhado" ou "sofá" (*output*), pois aprendeu, através da semântica distribucional, que essas são conclusões plausíveis com base nos contextos em que "gato" frequentemente aparece. Da mesma forma, se fornecermos "Ela abriu o", o Modelo pode sugerir "livro" ou "jornal" como possíveis complementos. Além disso, a semântica distribucional também permite que a IA Gen crie conteúdo que seja contextualmente relevante e coerente. Por exemplo, ao gerar uma história sobre um ambiente de selva, a IA Gen pode usar palavras como "tigre", "floresta" e "rio", pois reconhece a proximidade semântica desses termos com base em seu treinamento. (OPENAI, 2023)

Ao longo dos anos, a semântica distribucional tem ganhado destaque e reconhecimento no âmbito da Linguística Computacional. No entanto, sua influência e integração na Linguística teórica, que busca formular teorias e princípios universais sobre a essência da linguagem, ainda são limitadas. Boleda (2020) sugere que há um vasto potencial para a sinergia entre a semântica distribucional e a Linguística teórica. Ao incorporar *insights* empíricos e modelagem baseada em dados da semântica distribucional, a Linguística teórica pode se beneficiar de uma perspectiva mais rica e detalhada sobre o significado, enquanto a semântica distribucional pode se beneficiar das profundas análises e teorias estabelecidas pela Linguística teórica.

Boleda (2020) argumenta que essa intersecção limitada entre a semântica distribucional e a Linguística teórica pode ser atribuída à abordagem empírica e orientada a dados da primeira. Em contraste, a Linguística teórica tem sua fundamentação nas intuições, julgamentos e análises meticulosas de falantes nativos. A semântica distribucional, por outro lado, é profundamente enraizada em dados e modelagem quantitativa. Essa divergência nas metodologias pode criar obstáculos para uma fusão harmoniosa desses dois campos.

A semântica distribucional, apesar de sua notável aplicabilidade e êxito em diversas tarefas de Processamento de Língua Natural, enfrenta críticas e desafios inerentes à sua natureza. Mesmo com a capacidade da semântica distribucional de proporcionar percepções profundas sobre a essência do significado e de gerar representações detalhadas de palavras e sentenças, sua assimilação ao corpo consolidado da teoria Linguística tem sido um processo intrincado. Um dos desafios prementes é determinar o grau adequado de abstração para as representações lexicais e abordar a complexidade das palavras polissêmicas, que possuem múltiplos significados e sua composição.

A polissemia refere-se ao fenômeno linguístico em que uma única palavra tem múltiplos significados relacionados. Por exemplo, a palavra "banco" pode se referir a uma instituição financeira ou a um objeto no qual as pessoas se sentam. A capacidade de uma palavra ter vários significados é uma característica comum em muitas línguas e apresenta desafios interessantes para a modelagem semântica. No contexto da semântica distribucional, a polissemia pode ser explorada observando-se diferentes

contextos em que uma palavra ocorre e tentando capturar esses múltiplos sentidos em representações vetoriais. (OPENAI, 2023)

A composição, por outro lado, lida com a combinação de significados de palavras individuais para formar o significado de estruturas maiores, como frases ou sentenças. Por exemplo, o significado da frase "cão feroz" é composto pelos significados das palavras "cão" e "feroz". A questão central aqui é como combinar esses significados individuais de maneira a capturar corretamente o significado global. (OPENAI, 2023)

De acordo com Boleda (2020), a semântica distribucional tem demonstrado eficiência na representação de significados de palavras isoladas. No entanto, quando se trata da composição semântica, particularmente em situações polissêmicas, os desafios se intensificam. A complexidade surge, por exemplo, ao tentar integrar de forma adequada os significados de palavras que possuem diversas acepções em um determinado contexto. Apesar desses obstáculos, a comunidade científica tem proposto uma série de técnicas e abordagens visando superá-los. Estudos voltados para a composição semântica têm investigado a maneira como distintos elementos de uma sentença, como modificadores e núcleos de compostos, interagem para formar um significado coeso. Adicionalmente, pesquisas empíricas têm se dedicado a compreender a essência da composicionalidade e o papel desempenhado por diferentes palavras e estruturas nesse processo.

Ao analisar os desdobramentos práticos das abordagens de Aprendizado Profundo, percebemos certas particularidades que merecem destaque. Conforme apontado por Pavlick (2020), Modelos que são treinados com bases de dados volumosas e diversificadas tendem a apresentar resultados mais consistentes e robustos. Uma observação relevante é que sentenças que contêm palavras de conteúdo semelhante tendem a potencializar a precisão desses Modelos. Essas constatações nos conduzem a reflexões sobre a adequação das técnicas atuais, especialmente em cenários onde os dados são escassos ou quando há uma grande variação no conteúdo lexical. A despeito dos desafios inerentes, é inegável que existem vastas oportunidades para aprimoramentos e inovações no campo, principalmente em ferramentas do tipo *chatbot*.

Um dos produtos mais emblemáticos da IA Gen é o *chatGPT*, desenvolvido pela OpenAI, uma companhia de pesquisa e desenvolvimento de IA. O *chatGPT* é um robô de conversação (*chatbot*) lançado em novembro de 2022 alcançando, em dois meses, uma base de 100 milhões de usuários. Esse Modelo, que se baseia na arquitetura de *Transformer* Generativo Pré-Treinado (GPT - Generative Pre-trained Transformer), representa um marco na maneira como as máquinas entendem e geram a língua natural.

O *ChatGPT*, como um Grande Modelo de Língua, vai além de simplesmente recuperar informações, pois ele "pensa", "reflete" e "cria" respostas. Isso é evidente quando fornecemos *prompt* como entrada e não é feita uma busca por resposta pré-programada, mas gera uma com base em seu treinamento. Uma das características notáveis deste LLM é sua capacidade de adaptar-se e aprender com novos dados. Enquanto o *ChatGPT* é pré-treinado em grandes volumes de texto, ele também tem a capacidade de ajustar-se com base nas interações que tem, tornando-o mais adaptável e preciso ao longo do tempo.

A arquitetura *Transformer*, que é a espinha dorsal do *ChatGPT*, utiliza mecanismos de atenção para ponderar diferentes partes de uma entrada, permitindo que o Modelo considere o contexto global ao gerar uma resposta. Essa capacidade de criar conteúdo coerente e contextualizado, seja completando uma frase ou elaborando um texto, é uma demonstração do poder da IA Gen. A capacidade dos *Transformers* de dar atenção a diferentes partes de um texto é o que permite ao *ChatGPT* entender e gerar respostas que são relevantes e contextualizadas.

O *ChatGPT*, treinado em vastos volumes de texto, internaliza nuances Linguísticas, desde a sintaxe até a semântica. Ele não apenas reconhece padrões gramaticais, mas também entende contextos, metáforas e ambiguidades. Palavras polissêmicas, que têm múltiplos significados, são interpretadas pelo *ChatGPT* com base no contexto em que são apresentadas. Assim, ele pode discernir se "banco" se refere a uma instituição financeira ou a um objeto utilizado para se sentar, dependendo da frase. Esse entendimento contextual é um testemunho da profundidade de seu treinamento linguístico, com a utilização da semântica distributiva.

O *chatGPT* introduz uma inovação significativa no campo da Inteligência Artificial: a capacidade de incorporar *feedback* humano diretamente em seu processo de aprendizado. Este modelo opera por meio de um sistema de aprendizagem por reforço, onde os usuários desempenham um papel crucial ao indicar as respostas mais adequadas e pertinentes às suas consultas. Esse mecanismo de *feedback* contínuo não apenas otimiza a performance do *chatGPT*, mas também aprimora sua capacidade de interagir de forma mais natural e coerente com os usuários. Assim, o *chatGPT* não é apenas um *chatbot* convencional, mas representa um avanço na simulação da língua humana, estabelecendo-se como uma ferramenta que se aproxima cada vez mais da complexidade e fluidez da comunicação natural.

No universo dos LLMs, um fenômeno que tem despertado atenção e cautela é o que a comunidade técnica denomina de "alucinação". Estas alucinações referem-se à geração de palavras ou frases pelos LLMs que são desprovidas de sentido, gramaticalmente inconsistentes ou até mesmo incorretas. Diversos fatores podem contribuir para tais inconsistências, incluindo o treinamento do modelo com dados insuficientes, de baixa qualidade ou com ruídos. Além disso, a falta de contexto adequado ou restrições no momento de fornecer um *prompt* também podem induzir o modelo a alucinar. O impacto dessas alucinações não é trivial pois podem resultar em saídas de informações imprecisas e enganosas. Portanto, um dos principais desafios na área atualmente é desenvolver estratégias e técnicas que minimizem a ocorrência dessas alucinações, garantindo respostas mais precisas e confiáveis por parte dos LLMs.

Nesse cenário, surgem diversos campos de estudo, sendo um destes o *design*/engenharia de *prompt*. O *design* de *prompt* e a engenharia de *prompt* representam duas facetas complementares no universo dos Modelos de Língua. O *design* de *prompt* refere-se à formulação cuidadosa de instruções destinadas a guiar o Modelo em uma tarefa específica, considerando o contexto e o objetivo desejado. Por exemplo, ao solicitar a tradução de uma sentença do inglês para o francês, o *prompt* inicial em inglês deve claramente indicar a necessidade de uma resposta em francês. Por outro lado, a engenharia de *prompt* vai além da simples formulação. Trata-se de uma prática meticulosa de desenvolvimento, teste e otimização de

prompts para garantir respostas precisas e relevantes. Isso pode envolver a incorporação de exemplos, palavras-chave ou outros elementos que direcionem o LLM a produzir saídas alinhadas às expectativas. Enquanto o *design* de *prompt* estabelece a base para a interação com o modelo, a engenharia de *prompt* aprofunda-se nas nuances, buscando a excelência na comunicação e na obtenção de resultados. (GOOGLE, 2023)

Assim, para maximizar a eficácia de um LLM, é imperativo não apenas compreender sua arquitetura interna, mas também dominar a arte de formular *prompt*s eficientes. Esta abordagem enfatiza a interdependência entre a qualidade da entrada fornecida e a saída recebida, reforçando a necessidade de uma interação consciente e informada com essas ferramentas avançadas. Enquanto o *design* de *prompt* configura a espinha dorsal da interação com os Modelos, a engenharia de *prompt* é a refinada arte de aprimorar essa interação, especialmente em aplicações que demandam alta precisão e qualidade nas respostas geradas. Ambos são fundamentais para explorar efetivamente o potencial dos LLMs em diversas aplicações.

4. Considerações Finais

A evolução da Inteligência Artificial (IA) nos últimos anos tem sido notável, e sua intersecção com a Linguística Computacional tem aberto portas para inovações sem precedentes. A introdução de técnicas avançadas de Aprendizado de Máquina, em especial o "aprendizado profundo", tem permitido que máquinas se aproximem cada vez mais da capacidade cognitiva humana de comunicação e criação. No cerne dessa revolução, encontram-se as Redes Neurais Artificiais, estruturas inspiradas na complexidade do cérebro humano, que processam, aprendem e fazem predições com uma eficácia surpreendente.

A semântica distribucional, com sua abordagem empírica de inferir o significado das palavras a partir dos contextos em que aparecem, tem sido uma ferramenta essencial nesse avanço. Ela fornece uma base sólida para a compreensão e geração de língua natural, permitindo que as máquinas reconheçam nuances e ambiguidades Linguísticas. Esta abordagem tem se mostrado eficaz não apenas na representação

de significados de palavras isoladas, mas também na composição semântica de estruturas maiores, como frases e sentenças.

A IA Generativa, com sua capacidade de criar conteúdo novo e original, representa um marco na evolução do Processamento da Língua Natural. A combinação da semântica distribucional com a IA Generativa tem permitido que máquinas gerem conteúdo que soa natural e é contextualmente apropriado. Esta sinergia promete revolucionar a maneira como as máquinas entendem e geram a língua, aproximando-as ainda mais da capacidade humana de comunicação.

No entanto, é crucial reconhecer que, apesar dos avanços significativos, ainda existem desafios a serem superados. A complexidade da língua humana, com suas nuances, ambiguidades e contextos culturais variados, exige uma abordagem contínua e adaptativa. A interação entre a semântica distribucional, a IA Generativa e outras técnicas emergentes de Aprendizado de Máquina deve ser explorada ainda mais profundamente para alcançar uma compreensão mais holística da língua na comunicação e criação de novos textos.

O *chatGPT* combina profundo entendimento linguístico com poderosas capacidades generativas inspiradas na cognição humana e habilidades comunicativas pelo uso da língua. Ele não apenas compreende a língua em sua rica complexidade, mas também a utiliza para criar respostas que são, muitas vezes, indistinguíveis das de um ser humano. Esta fusão de Linguística e IA Generativa permite que o Modelo navegue por conversas complexas, responda a perguntas variadas e até mesmo crie conteúdo original, desde histórias até explicações técnicas.

No entanto, essa capacidade impressionante de Modelos como *chatGPT* não está isenta de desafios e preocupações. Um dos principais problemas reside nos dados de treinamento. Muitas vezes, os modelos são treinados com vastos conjuntos de dados que podem conter informações protegidas por direitos autorais ou dados sensíveis, levantando questões éticas sobre a propriedade e privacidade da informação. Além disso, a natureza abrangente desses dados pode inadvertidamente incorporar preconceitos e estereótipos, refletindo as imperfeições da sociedade em suas respostas.

Outra preocupação significativa é o potencial de mau uso desses modelos avançados. Com sua habilidade de gerar texto convincente, o *chatGPT* pode ser explorado por indivíduos mal-intencionados para produzir notícias falsas, disseminar desinformação ou perpetrar golpes online, complicando ainda mais o cenário da informação na era digital.

Além disso, as "alucinações" – respostas geradas que podem ser sem sentido, gramaticalmente incorretas ou factualmente erradas – são uma preocupação constante. Embora os modelos sejam treinados com vastos volumes de dados, eles ainda podem produzir saídas inesperadas ou imprecisas, especialmente quando confrontados com *prompts* ambíguos ou fora de seu domínio de treinamento.

À medida que continuamos a desenvolver e aprimorar Modelos como o *ChatGPT*, estamos dando passos em direção a um mundo onde a tecnologia entende, responde e colabora conosco de maneiras cada vez mais naturais e intuitivas. Estamos em um momento excitante na interseção da Linguística e da IA. As possibilidades são vastas, e os avanços recentes apenas arranham a superfície do que é possível. À medida que continuamos a explorar e inovar, o futuro da Linguística computacional promete ser repleto de descobertas e transformações que redefinirão a relação entre humanos e máquinas.

Referências

BIRD, Steven. KLEIN, Ewan. LOPER, Edward. **Natural Language Processing**. O'Reilly Média, 2009.

BOLEDA, Gemma. **Distributional Semantics and Linguistic Theory**. Annual Review of Linguistics, v. 6, p. 213-234, 2020. Disponível em: https://doi.org/10.1146/annurev-linguistics-011619-030303 . Acesso em: 23 de set. 2023.

CARLE, Eben. Ask a Techspert: **What is generative AI?.** Google The Keyword. Disponível em:< https://blog.google/inside-google/googlers/ask-a-techspert/what-is-generative-ai/>. Acesso em: 14 de set. de 2023.

CHANG, Yupeng et al. **A survey on evaluation of large language models**. arXiv preprint arXiv:2307.03109, 2023. Disponível em:< https://arxiv.org/pdf/2307.03109.pdf>. Acesso em: 23 de set. de 2023.

GÉRON, Aurélien. **Mãos à Obra: Aprendizado de Máquina com Scikit-Learn & TensorFlow. Conceitos, Ferramentas e Técnicas para a construção de Sistemas Inteligentes.** Rio de Janeiro: Alta Books Editora, 2019.

GOOGLE. Google Cloud Skills Boost. I**ntroduction to Generative AI Learning Path.** Disponível em:< https://www.cloudskillsboost.google/course_templates/536>. Acesso em: 14 de set. de 2023.

GRISHMAN, Ralph. **Computational linguistics: an introduction**. Cambridge University Press, 1986.

MARTELOTTA, Mário (org.). **Manual de Linguística. Linguística-Conceitos básicos**. 2ed. São Paulo: Contexto, 2011.

MEDEIROS, Luciano. **Inteligência Artificial Aplicada: uma abordagem introdutória** [livro eletrônico]. Curitiva: InterSaberes, 2018.

NORVIG, Peter; RUSSELL, Stuart. **Inteligência artificial.** Rio de Janeiro: Grupo GEN, 2013.

OPENAI. ***ChatGPT*.** [S.l.], [2023]. Disponível em: < https://chat.openai.com/>. Acesso em: 24 set. 2023].

OTHERO, Gabriel; MENUZZI, Sérgio**. Linguística computacional: teoria e prática.** São Paulo: Parábola, 2005.

PAVLICK, Ellie. **Semantic Structure in Deep Learning**. Annual Review of Linguistics, Providence, Rhode Island, v. 8, p. 447-471, 2022. Disponível em:< https://www.annualreviews.org/doi/abs/10.1146/annurev-linguistics-031120-122924>. Acesso em: 23 de set. 2023.

SANDOVAL, Antonio Moreno. **Lingüística computacional**. In: Enciclopedia de Lingüística Hispánica. Routledge, 2016.

TURING, Alan M. **Computing Machinery and Intelligence**. Creative Computing, v. 6, n. 1, p. 44-53, 1980.

Vaswani et. al. **Attention is all you need.** In *Proceedings of the 31st International Conference on Neural Information Processing Systems*. 6000–6010, 2017. Disponível em:< https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf>. Acesso em: 23 de set. 2023.