DOI: 10.1590/s0103-4014.2021.35101.005

Inteligência Artificial e os rumos do processamento do português brasileiro

MARCELO FINGER [I]

Introdução

ESTE ARTIGO apresenta um posicionamento frente aos desafios da área do Processamento de Língua Natural (PLN),[1] em particular em Língua Portuguesa. Essa área se encontra na confluência de diversas outras, como Ciência da Computação, Linguística, Lógica, Psicologia, dentre outras, e requer por natureza um tratamento multidisciplinar. Além disso, esse posicionamento está sendo feito no momento de grande explosão de pesquisas e aplicações dessa área do conhecimento, em que o que era antes tratado como ficção científica passa a ser visto como ciência de fato e é encontrado rotineiramente na vida das pessoas, com impactos acadêmicos, sociais e econômicos.

Os objetivos muitas vezes incertos ou mal definidos da área hoje se consolidam, ao menos no ponto de vista desse posicionamento, como a capacidade de capturar os fenômenos linguísticos dentro de um contexto em que diversos fenômenos cognitivos, sociais e até econômicos se inter-relacionam. Esta apresentação se centra nos esforços para a captura desse contexto.

Chegamos a esse ponto na história do processamento de língua natural surfando na confluência de diversas ondas. Por um lado, temos uma onda que começou lá atrás com o início da área de inteligência artificial na década de 1950 e com os estudos computacionais da linguagem humana; os estudos sobre a linguagem humana, é importante dizer, se iniciaram basicamente junto com a própria filosofia há mais de 2.500 anos. Por outro lado, temos a onda que foi formada pelo aumento da disponibilização de dados em formato digital acarretado pela explosão do uso da internet nos últimos 25 anos. Também precisamos apontar a onda gerada pelo barateamento do hardware que fornece a capacidade computacional necessária para o processamento moderno.

A Inteligência Artificial (IA), que abarca a linguística computacional, também chamada de processamento de língua natural, é uma área que sempre foi carregada de grandes expectativas, mas que espelhando o desenvolvimento da maioria das realizações humanas teve o seu início no ambiente inóspito e carente de recursos (computacionais e de dados), encontrou pontos em que quase foi à extinção, mas conseguiu se apoiar em eventos que ocorreram ao seu redor para florescer e se expandir.

Toda essa riqueza de recursos que encontramos nesse momento com o foco de estudos na área deixe de se centrar na falta de recursos e passe a se centrar nos reais objetivos da área. Temos que encarar as questões sobre se a captura dos fenômenos linguísticos em contexto é realmente o objetivo que deveríamos estar buscando. Devemos agora abordar quais são as consequências reais das ferramentas que empregamos nessa tarefa e os impactos positivos e negativos que elas podem ter na vida das pessoas que começam a interagir corriqueiramente com essa tecnologia.

Precisamos também entender a posição que o processamento de Língua Portuguesa ocupa nesse cenário e quais as melhores direções a serem tomadas para o seu desenvolvimento pleno.

Para apresentarmos esse posicionamento, iniciamos com uma breve apresentação do início da pesquisa na área, aquilo que chamei de desenvolvimento em um ambiente inóspito e de baixos recursos na segunda seção. Em seguida, na terceira seção, passamos a descrever inflexão exponencial gerenciada pela área. Com isso nos posicionamos para discutir os desafios a serem enfrentados no futuro próximo e nem tão próximo, quarta seção, para então podermos concluir sobre a nossa visão desse entrocamento de possibilidades em que nos encontramos após uma surpreendente subida e sobre os temores e responsabilidades que contemplamos nessa paisagem em que a pesquisa acadêmica deságua no mundo real.

A consolidação do Processamento de Língua Natural

Neste trabalho oferecemos uma visão particular sobre os avanços, a princípio lentos, e então desnorteantes, da evolução do processamento de língua natural. Vamos apresentar o desenvolvimento deste processamento como uma sucessão de técnicas que tentam conquistar uma importante noção linguística, que apesar de não ser inicialmente óbvia, vai se consolidando ao longo do tempo: a noção de *contexto linguístico*.

O processamento computacional da linguagem pode ter diversas aplicações, desde a tradução automática entre linguagens, passando pela identificação de opiniões favoráveis ou desfavoráveis ao objeto do texto (uma atividade conhecida como análise de sentimento) até a mera identificação de qual elemento do texto é referido por um pronome resolução de anáfora pronominal). No entanto, ao descrevermos a evolução das técnicas de processamento, não vamos nos ater a qualquer tarefa em específico, visto que as técnicas parecem evoluir independentemente da tarefa em questão. Tampouco vamos tratar aqui do processamento da fala, que também tem experimentado importantes avanços, mas vamos nos ater ao processamento da linguagem escrita.

Regras

O grande marco para o desenvolvimento do processamento de língua natural, pelo menos do ponto de vista da ciência da computação, se deu com o enfoque matemático da linguagem proposto pelo trabalho de Chomsky. Esse

trabalho identifica um conjunto de linguagens de fácil trato computacional, chamadas de linguagens livres de contexto (Chomsky, 1965). Essas ideias inovadoras na teoria linguística ocorreram ao mesmo tempo que nascia uma área que foi chamada de Inteligência Artificial (Newell; Simon, 1963), e que teve como uma das primeiras iniciativas a ideia de construir tradutor automático do russo para o inglês (Buchanan, 2005); era então o auge da guerra fria.

O fracasso resultante dessa primeira iniciativa de tradução automática evidenciou o alto grau de complexidade dessa tarefa e motivou o desenvolvimento de uma teoria que ficou conhecida como Teoria da Complexidade Computacional (Papadimitriou, 1994).

As gramáticas livres de contexto serviram de base para o desenvolvimento de linguagens artificiais de programação, impulsionadas pelo sucesso na construção dos primeiros *compiladores*, tradutores automáticos de linguagens de programação para as linguagens de máquina dos computadores daquela época (Hopcroft; Ullman, 1979). A sistematização do desenvolvimento dos compiladores investigou e explorou as propriedades computacionais das linguagens livres de contexto (Aho et al., 1986).

As gramáticas livres de contexto se assemelham muito a um conjunto de regras lógicas, e sua implementação em sistemas de computador se parece com o processo de inferência a partir de regras lógicas. Elas têm um aspecto como o seguinte.

Oração → Sujeito, Predicado

Predicado → Verbo_Intransitivo

Predicado → Verbo_Transitivo_Direto, Objeto

É importante salientar que a formulação gramatical chomskiana *não* se utiliza de termos como Sujeito ou Objeto como categorias básicas, e esses foram utilizados aqui apenas como exemplos de apelo familiar a um grupo mais amplo de leitores. Note-se a dupla possibilidade de leitura de uma regra. Por exemplo, para gerar um predicado, temos que produzir um verbo intransitivo ou um verbo transitivo direto seguido de um objeto; por outro lado, a partir de um verbo transitivo direto seguido de um objeto, podemos constituir um predicado. Uma sequência de regras lógicas aplicadas na geração de uma sentença nos fornece também a sua estrutura sintática. Outra característica importante está no fato, que muitas vezes passa despercebido, que as categorias que compõem a estrutura sintática não são observáveis diretamente da sentença. Dessa forma, dada uma frase, é necessário um especialista humano para classificar cada expressão na sua correspondente categoria sintática, não havendo nenhum outro método de inspeção direta. O estudo de gramática foi mediado por seres humanos até recentemente, quando surgiram os primeiros trabalhos sobre *indução gramatical* (Clark; Lappin, 2010).

Um dos problemas em se impor um tratamento baseado em regras lógicas a expressões de língua natural está no fato de que todas as línguas humanas apre-

sentam o fenômeno da *ambiguidade*. A ambiguidade se apresenta em diversos níveis da linguagem, seja no contexto sonoro, no contexto lexical (palavras ambíguas), no contexto sintático, semântico, seja até mesmo pragmático. E o fato é que a resolução das ambiguidades necessita explorar o contexto linguístico em que as expressões ambíguas ocorrem, se ele existir, ou até o contexto cultural. As regras gramaticais são capazes de captar fenômenos como a ambiguidade sintática associada, bem como a ambiguidade semântica associada a ela. Porém, as regras, se livres de contexto, não são capazes de resolver essa ambiguidade (Carpenter, 1997).

Os enfoques baseados em regras ainda hoje trazem interesse na pesquisa, pois são capazes de apresentar uma abordagem composicional do tratamento da linguagem e, de acordo com essa visão, a semântica de uma sentença está diretamente associada à sua estrutura sintática (Benthem, 1995; Moortgat, 1997). Isso foi reconhecido como uma propriedade importante da análise de linguagem desde o início da Inteligência Artificial (Lambek, 1958), e já na década de 1980, induziu a um tratamento linguístico acoplado às tecnologias de inteligência artificial que foram produzidas naquela época (Pereira; Shieber, 1987). A abordagem composicional busca obter o significado de expressões linguísticas a partir do significado dos seus componentes e da estrutura sintática utilizada na sua composição.

As regras e métodos simbólicos também são úteis para explicar fenômenos linguísticos, mesmo que a automação das explicações seja inviável na prática. Isso explora um aspecto que permanece bastante interessante em relação às abordagens baseadas em regras lógicas, por sua capacidade de capturar relações causais. Voltaremos a falar em causalidade mais para a frente. O importante é notar que, já na década de 1980, ficou claro que a abordagem estritamente lógica era rígida demais para o desenvolvimento de aplicações que possam lidar com as nuances e a complexidades de fenômenos linguísticos.

Probabilidades

Uma das primeiras propostas de generalização das gramáticas livres de contexto se deu pela atribuição de probabilidades a cada uma das regras gramaticais que poderiam ser aplicadas num determinado ponto (Charniak, 1993). Essa extensão visava resolver ambiguidades sintáticas por meio da escolha de uma dentre as diversas possíveis estrutura da sentença, de forma a priorizar aquela de maior probabilidade. No entanto, esse enfoque ainda assume que as regras gramaticais e suas probabilidades são entidades independentes umas das outras, o que faz que esse formalismo não seja capaz de capturar as interdependências entre as expressões e seu contexto (Manning; Schütze, 1999).

Pobabilidades, no entanto, possuem uma série de propriedades interessantes, por terem a capacidade de expressar um resumo de toda uma configuração. Os modelos probabilísticos divergem da abordagem composicional, considerando que o significado de uma expressão é dado "pela companhia que ela man-

tém'', ou seja, o significado de uma expressão é dado pelos contextos em que ela ocorre (Manning; Schütze, 1999). Nessa visão, o contexto acaba sendo o elemento de atribuição da semântica das expressões linguísticas. Muito se critica essa visão do ponto de vista filosófico, pois ela não fornece os elementos básicos de atribuição de significado e permite uma recorrência infinita no processo de construção de significados. Porém, do ponto de vista computacional, essa visão possui o atrativo de não requerer nenhuma referência externa a não ser as próprias palavras que estão no texto, e boa parte do trabalho realizado em linguística computacional desde os anos 1990 se baseia nessa visão.

Dessa forma, foram surgindo diversos modelos probabilísticos de linguagem, dentre os quais destacamos os modelos baseados em Cadeias de Markov e os modelos baseados em n-gramas (Damerau, 1971). Uma cadeia de Markov é um processo estocástico em que o estado seguinte depende apenas do estado atual, e é independente de todo o histórico anterior dado o estado atual. A restrição de depender apenas de um estado anterior numa sequência discreta pode facilmente ser estendida para um número qualquer predeterminado de estados, conhecido como janela de observação. Dentro dessa janela, um processo markoviano é capaz de detectar as interdependências entre os elementos. Novamente a complexidade computacional cobra um preço, pois o número de relações de probabilidades que devem ser computadas explode exponencialmente como o número de elementos da janela. Assim, os processos markovianos analisam tipicamente janelas muito estreitas, de no máximo cinco elementos. Nenhuma janela de tamanho limitado é capaz de dar conta de diversos fenômenos linguísticos que ocorrem em todas as línguas humanas conhecidas, chamados de *dependências de distância ilimitada*. Por exemplo, nas expressões comparativas que utilizam o par mais/que, essas duas palavras podem ocorrer a uma distância qualquer e ilimitada:

Ela estudou *mais que* eu.

Eu como *mais* doces *que* ela.

Ele passou *mais* horas tocando piano *que* todo o resto da turma junto.

Mais vale um asno que me carregue *que* um cavalo que me derrube.

Nesse último exemplo há ainda outro "que", pronome relativo, que aparece em posição intermediária e não faz parte da comparação. Não é à toa que a ambiguidade associada à palavra "que" é um pesadelo para o processamento do português.

Um modelo simplificado de linguagem muito usado no contexto probabilístico é o de considerar sentenças como "saco de palavras" (*bag of words*), ignorando a ordem em que as palavras ocorrem na sentença, reduzindo a sentença a uma contagem de seus componentes. Por exemplo, a última sentença do exemplo anterior é reduzida aos seguintes pares:

asno:1, carregue:1, cavalo:1, derrube:1, mais:1, me:2, que:3, vale:1, um:2.

Nenhuma estrutura sintática foi mantida, inclusive a conjunção e o pronome relativo que foram contabilizados como se fossem a mesma coisa.

Para aumentar a sensibilidade ao contexto, uma ideia de natureza probabilística que expande esse modelo é o chamado modelo de n-gramas. No caso de n = 1, ele é chamado de modelo de unigramas, cuja aplicação dá origem aos sacos-de-palavras. No caso n = 2, usamos pares de palavras na sequência que ocorrem. Desta forma ficaríamos com os bigramas: mais vale, vale um, um asno etc. Por exemplo, experimentos mostram que a utilização de bigramas para medir as probabilidades na análise de sentimento, ou seja, medir com que frequência uma sequência de duas palavras ocorre numa expressão positiva, negativa, ou neutra, acaba tendo uma eficiência muito boa em domínios limitados.

Ao aumentar o tamanho da sequência, no entanto, tratando de trigramas, tetragramas etc., temos um outro problema de natureza estatística que é a *esparsidade* dos elementos. Ou seja, quanto mais longa for a expressão, mais rara ela será. Por exemplo, a probabilidade de ocorrência do trigrama "asno que me" é bastante baixa, isso pode fazer que diversos n-gramas sejam vistos pouquíssimas veze. Existe uma grande chance de que, em um texto nunca visto antes, haja algum n-grama inédito, o que pode levar a probabilidade total do texto seja tratada como nula; isso é contraditório, dado que o texto efetivamente existe. Diversos métodos de *suavização de probabilidades* foram propostos para lidar com esses casos (Jurafsky; Martin, 2000).

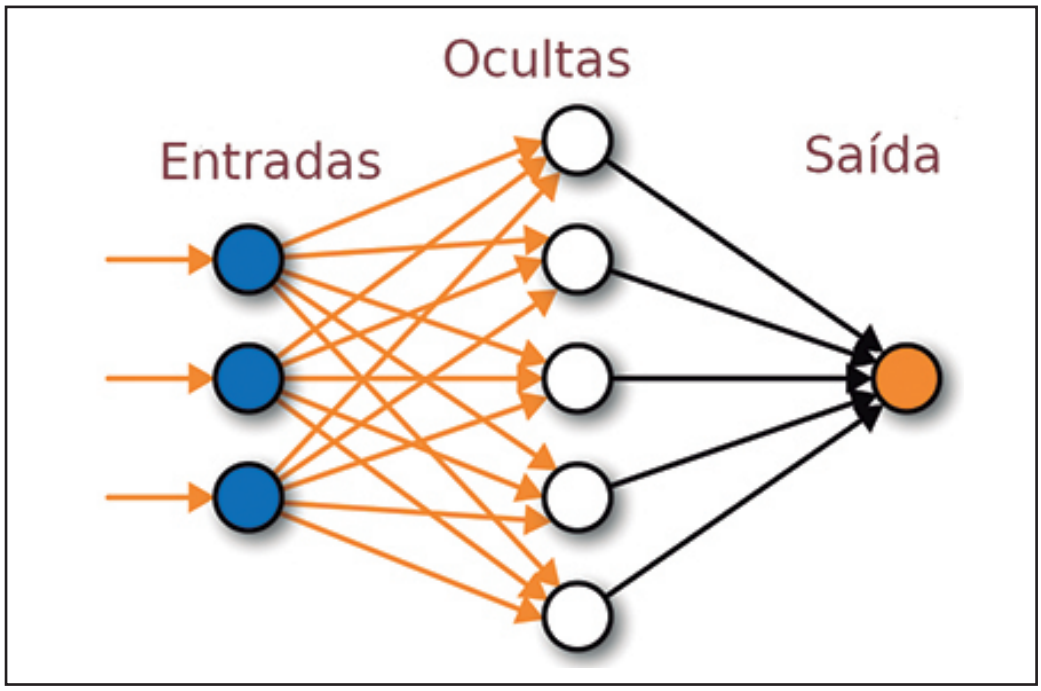
Por sua simplicidade, os modelos de n-gramas permanecem até hoje como uma ferramenta para ser usada em casos de poucos dados ou de necessidade de algum resultado com baixo tempo de desenvolvimento. As ideias de janela fixa e de saco-de-palavras são usadas em domínios específicos e até em modelos neurais (Mikolov et al., 2013). Modelos probabilísticos mais sofisticados também foram empregados utilizando redes bayesianas de arquitetura baseada em moldes (templates), para tratar de tarefas sofisticadas como a classificação de tópicos de um conjunto de textos, porém com problemas de eficiência devido ao alto custo computacional (Blei, 2003).

Modelos probabilísticos igualmente sofisticados tiveram grande aplicação na tradução de textos, com a utilização de corpus paralelos, que impulsionaram as pesquisas em tradução automática (Och; Ney, 2002; Koehn, 2009). Com essas técnicas, houve uma rápida melhora na qualidade das traduções, cujas limitações só foram ultrapassadas com o uso de redes neurais.

Redes neurais

Redes neurais são uma classe de programas que se especializam no reconhecimento de padrões de dados apresentados (dados de treinamento) e que então são utilizados para processar conjuntos de dados desconhecidos (dados de testes). Desde sua concepção, esses sistemas são centrados na captura de elementos contextuais e seu desenvolvimento se deu a partir do aumento da capacidade de processamento e da complexidade dos padrões capazes de serem detectados pelas redes neurais.

As redes neurais possuem duas habilidades básicas, como classificadoras de dados e como preditoras de valores (interpoladoras e extrapoladoras) e ambas as capacidades foram exploradas ao longo do seu desenvolvimento. A história das redes neurais começa com um algoritmo chamado de Percéptron (Rosenblatt, 1958), seguida pelo desenvolvimento de métodos para treinar de forma supervisionada, fornecendo exemplos e respostas, mecanismos esses que sempre foram programas de computador baseados em álgebra linear (Novikoff, 1962; Rosenblatt, 1962; Aizerman et al., 1964). Esses Percéptrons foram combinados e organizados em redes multicamadas. Por uma analogia superficial percebida com circuitos de células do sistema nervoso, ficaram conhecidas como as redes neurais artificiais, e depois apenas como redes neurais, com suas diversas camadas conhecidas como camada de entrada, uma ou mais camadas ocultas e uma camada de saída (Figura 1).



Fonte: Elaboração própria.

Figura 1 – Ilustração de rede neural artificial, com camada de entrada, uma ou mais camadas ocultas e uma camada de saída.

O desenvolvimento das redes neurais experimentou muitos altos e baixos. Depois da atenção inicial despertada pelos percéptrons no final da década de 1950 e dos primeiros algoritmos da década de 1960, seguiu-se uma série de estudos teóricos que provaram limitações na sua expressividade (Minsky; Papert, 1969), o que atrasou o desenvolvimento da área em ao menos uma década, é que às vezes é chamado de "inverno das redes neurais". O interesse nas redes neurais foi a reavivado com a publicação, em 1986, de um algoritmo de treinamento

de redes multicamadas chamado de Algoritmo de Retropropagação (Rumelhart et al., 1986a/b), seguido de experimentos mostrando que essas redes podiam aprender a detectar padrões, e de estudos teóricos mostrando a capacidade e expressividade dessas redes de aprender em qualquer função contínua (Hornik et al., 1989; Maass et al., 1994; Bartlett et al., 1998).

Quem disse que procurar padrões seria fácil? Uma nova sequência de resultados desanimadores levantou uma série de problemas com o emprego das redes neurais multicamadas. Em primeiro lugar há uma necessidade muito grande de recursos computacionais para treiná-las. Em seguida, existem dois problemas de natureza estatística coligados, que ocorrem no treinamento dessas redes. Por um lado, essas redes podem sobreajustar e acabar decorando os dados de entrada, perdendo a capacidade de predição (*overfitting*) (Anderson; Burnham, 2004). Por outro lado, para obter um bom grau de generalidade nas predições, são necessárias quantidades muito grandes de dados que não estavam disponíveis aos pesquisadores antes do advento da internet, o que também levava a perda da capacidade de predição destas redes (*underfitting*) (Harrell, 2001). Mais ainda, redes neurais com muitas camadas, chamadas de redes profundas, apresentam fenômenos de instabilidade, conhecidos como desaparecimento de gradientes e explosão de gradientes (Hochreiter et al., 2001). Por fim, houve o amadurecimento de métodos de aprendizado de máquina, como os métodos estatísticos de tradução e o SVM (Vapnik, 1995). Tudo isso fez que, por volta da transição do milênio, o interesse na área Redes Neurais estivesse bastante esfriado nos centros de pesquisa pelo mundo, um segundo "inverno" das redes neurais.

Dois fenômenos levaram a um acalorado retorno no interesse nesta área. Um deles foi a disseminação da internet e a consequente explosão na quantidade de dados que se tornaram disponíveis para fins de busca e de pesquisa. O outro fenômeno foi a disseminação dos jogos de computadores que veio com a popularização dos computadores pessoais, causando uma demanda por um tipo de hardware especial para melhorar o desempenho visual dos jogos. Esse hardware se chama comumente de placa de vídeo, ou GPU, e consiste de unidades contendo inúmeras células paralelas e independentes de processamento, que são capazes de realizar apenas operações muito básicas como soma e produto de números. Ocorre que estas operações são justamente aquelas realizadas durante o treinamento das redes neurais, e quando disparadas de forma muito rápida e em processamento paralelo nas inúmeras unidades de uma GPU, ampliam a quantidade de dados que podem ser tratados, levando a uma ampliação das áreas em que as redes neurais podem ser usadas.

Esses avanços fizeram com que as redes neurais pudessem ser aplicadas na identificação de padrões em dados estruturados e em imagens, casos em que todos os dados são apresentados simultaneamente à rede. Mas no caso em que os dados são apresentados sequencialmente, como em textos, sinais de voz, vídeo, música ou qualquer sequência temporal de dados, uma rede neural multicama-

das não é capaz de captar padrões não triviais. Quando muito, as redes multicamadas condensam pequenas sequências e não obtêm expressividade maior que os métodos de sacos de palavras. Então, além da musculatura das GPU e da testosterona abundante nos dados, a área precisou recorrer à inteligência (natural).

Avanço acelerado

A partir do ano 2012, houve uma inflexão na *performance* dos sistemas de processamento de língua natural, a qual passamos a apresentar aqui. Houve um grande avanço nas novas arquiteturas propostas para redes neurais que fizeram que a qualidade do processamento de língua natural tivesse uma aceleração nunca vista, e esse tipo de modelo passasse a dominar a maioria das aplicações em língua natural. Essas novas propostas só se tornaram viáveis por serem acompanhadas pela enxurrada de dados que passaram a estar disponíveis, bem como uma nova classe de equipamentos de hardware que permitiu o seu processamento.

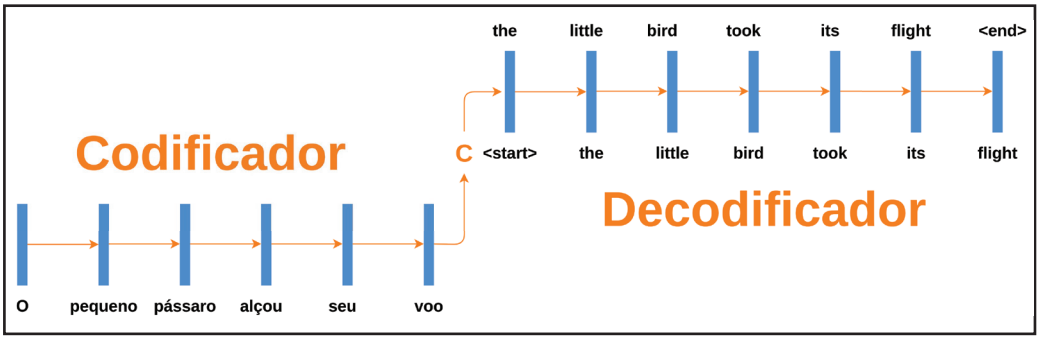
A atividade que mais impulsionou esse avanço foi justamente a tradução automática, tarefa que havia iniciado pesquisas de processamento de língua natural no auge da guerra fria. Até 2012, os métodos mais bem-sucedidos utilizados nessa tarefa tratavam de alinhar textos em duas línguas e computar as probabilidades de um trecho ser traduzido em outro. Com o aumento da disponibilidade de dados e da capacidade de processamento, arquiteturas mais profundas puderam ser exploradas, com diversas soluções para os problemas que já eram conhecidos, e que agora poderiam ser tratados com a abundância de recursos que passou a ficar disponível.

A principal arquitetura que permitiu avanços na tradução ficou conhecida pelo nome de sequência-para-sequência, ou arquitetura codificadora-decodificadora (Sutskever et al., 2014). Esse tipo de rede neural era capaz de computar um valor que possuía a expressividade de resumir (codificar) uma sequência de entrada, em um idioma, e então utilizar esta codificação para gerar uma nova sequência de saída em um outro idioma (Figura 2).

O modelo sequência-para-sequência finalmente superou a limitação imposta pelo processamento baseado em "saco-de-palavras" para permitir o tratamento de textos na ordem em que foram produzidos. No entanto, é importante notar que a expressividade dessa arquitetura decai com o tamanho da expressão, e a codificação consegue apenas captar ou resumir a informação de no máximo uma única sentença.

Esse tipo de arquitetura codificadora-decodificadora teve seus resultados amplificados por uma nova maneira de representar palavras em redes neurais. Com a disponibilidade de grandes repositórios de textos, chamados de córpus de textos, contando com quantidades da ordem de bilhões de palavras em contexto, pode-se representar as palavras em sequências numéricas, tecnicamente chamadas de vetores multidimensionais, que tentam capturar numericamente os contextos mais frequentes em que uma palavra pode acontecer. Essa ideia de

representar palavras por vetores faz parte do repertório do processamento de língua natural desde os primórdios, mas sua representação em modelos neurais profundos ficou conhecida como *word2vec* (uma condensação da expressão "palavra para vetor" em inglês) (Mikolov et al., 2013). Anteriormente a esse tipo de representação, as palavras eram representadas pelo seu índice num dicionário, num tipo de representação unária conhecida como *1-hot*, ou com cada dimensão do vetor associada explicitamente a uma dimensão linguística estática e não contextual. Além de captar o contexto em que uma palavra pode ocorrer, essa nova representação também representa uma compactação na representação de palavras que antes necessitavam de sequências do tamanho do vocabulário, da ordem de dezenas de milhares, e agora passaram a ser representadas por pontos num espaço de dimensão muito menor, da ordem de dezenas.



Fonte: Elaboração própria.

Figura 2 – Ilustração de rede neural artificial na arquitetura sequencia-para-sequência para a tradução entre línguas diferentes.

Essa nova representação foi aplicada em diversas tarefas de processamento de língua natural, como a análise de sentimento e a etiquetagem de trechos relevantes de um texto de acordo com a aplicação. Mas foi novamente da tradução automática que surgiu uma nova melhoria nas arquiteturas neurais, que amplificou ainda mais a sensibilidade ao contexto, muito além das janelas limitadas usadas no *word2vec*.

A arquitetura codificadora-decodificadora trata o texto apenas como uma sequência, ignorando a estrutura sintática e semântica subjacente à comunicação de informações onipresente nas línguas humanas. Em particular, idiomas distintos ordenam os elementos em uma sentença de forma distinta. Por exemplo, em algumas línguas o verbo precede o objeto, enquanto em outras o verbo é o último elemento da sentença. Igualmente, em algumas línguas os adjetivos precedem o substantivo que qualificam, enquanto em outras o nome tende a anteceder seus qualificadores.

O conceito de *atenção neural* busca correlacionar expressões em uma língua, ou melhor dizendo, a codificação de expressões em uma língua, com a codificação de expressões em outra língua. A atenção neural é inserida como uma

camada entre a codificação e a decodificação (Bahdanau et al., 2016; Luong et al., 2015). Dessa forma, a atenção neural correlaciona entradas e saídas, o que permite capturar elementos de ordenação entre expressões relacionadas.

O conceito de atenção neural promoveu uma nova onda de avanços na área, que ficou em emblematizada no *slogan* "atenção é tudo o que se necessita" (Vaswani et al., 2017). Uma nova arquitetura chamada de *Transformer* foi proposta para explorar a autocorrelação, ou autoatenção, de uma longa expressão linguística consigo mesma. A arquitetura Transformer utilizada na tradução entre expressões realiza a autoatenção da entrada com entrada em várias camadas, da saída com a saída, também em várias camadas, além da tradicional correlação entrada com saída.

Em seguida, novas formas de treinar o modelo Transformer levaram ao desmembramento da arquitetura com um aumento significativo da acurácia nas aplicações de processamento de língua natural. O desmembramento do codificador gerou a arquitetura Bidirectional Encoder Representations from Transformers (Bert) (Devlin et al., 2019) e o desmembramento do decodificador gerou arquitetura geradora de textos GPT (Brown et al., 2020).

O sucesso do modelo Bert foi tão grande que gerou uma área de pesquisa chamada pelos seus praticantes de Bertologia, abarcando a série de modelos gerados a partir de modificações do modelo Bert original. Esse modelo é composto de duas fases. A primeira fase é chamada de modelo pré-treinado de linguagem, e consiste em um aprendizado autossupervisionado (uma forma de aprendizado não supervisionado) ao qual são submetidas quantidades enormes de texto; por exemplo, no caso do modelo pré-treinado Bert para o inglês foi treinado com 3,3 bilhões de instâncias de palavras e possui 340 milhões de parâmetros que devem ser aprendidos. O processo de pré-treinamento possui algumas tarefas, dentre as quais omitir algumas palavras de um texto e tentar prever essas palavras e verificar a correta ordenação de sentenças completas conforme o texto original. Os recursos computacionais necessários para o pré-treinamento são consideráveis. A segunda fase se chama processo de refinamento e é uma bem-sucedida aplicação da técnica de aprendizado por transferência. O refinamento parte do modelo pré-treinado com alguma possível extensão e adapta todos os parâmetros para uma tarefa específica que pode ser tanto de classificação quanto de marcação ou etiquetagem de textos. Outros modelos derivados do Bert foram propostos modificando ou simplificando o pré-treinamento, que geraram modelos de nomes como RoBERTa (Liu et al., 2019), XL-NET (Yang et al., 2020), DisltilBERT (Sanh et al., 2020) e CamemBERT (Martin et al., 2020). O modelo pré-treinado para o português brasileiro recebeu o nome de BERTimbau (Souza et al, 2020) e foi obtido a partir de um corpus de 2,7 bilhões de instâncias de palavras BrWaC (Wagner Filho et al., 2018).

O modelo de GPT já está na sua terceira versão, e alcançou a surpreendente marca de mais de um bilhão de parâmetros a serem treinados. Trata-se de um

modelo de geração de texto que causa muita apreensão pois dada uma semente de texto ele é capaz de gerar um parágrafo coerente, o que levanta as sérias preocupações com o emprego ético de uma ferramenta como essa.[2]

Desafios do momento

Viemos falando sobre os desenvolvimentos inéditos da área de processamento de língua natural e só agora, no final da apresentação, é que mencionamos a língua portuguesa. Não obstante os vertiginosos progressos, há muitos desafios a serem explorados tanto no processamento em geral quanto particularmente no processamento da língua portuguesa. Nem todos esses desafios cumprem uma agenda positiva, mas é preciso mencioná-los, o que faremos a seguir.

Sintaxe, semântica ou pragmática?

O grande impulso no processamento de língua natural, quer seja em inglês, em chinês ou em português, se deu utilizando uma classe de algoritmos chamados de redes neurais. Quando mencionamos regras sintáticas ou modelos probabilísticos, sabíamos que domínio da linguística estávamos encarando.

As regras sintáticas claramente descreviam a sintaxe da língua. Por outro lado, a ideia de que a semântica de uma palavra se dá pela companhia que ela mantém tenta capturar a natureza semântica de acordo com um determinado ponto de vista.

Então, cabe a pergunta: a classe de algoritmos que contém as redes neurais trata de qual dimensão linguística? Certamente, nelas não há nenhuma representação explícita da sintaxe da língua, como é o caso das gramáticas formais. Por outro lado, a representação numérica vetorial das palavras usadas pelas redes neurais como codificações intermediárias pode sugerir que essas redes tratam de uma espécie de representação de significado que poderia de ser chamada de *semântica vetorial*. Esse ponto de vista sofre o seguinte questionamento. Ao tratarmos um vetor como a representação de uma palavra em contexto, não sabemos o que esse vetor significa, ou seja, temos uma representação de significado para a qual não conhecemos o significado. Em suma, a semântica vetorial parece ser uma representação sem semântica.

Por outro lado, o seu treinamento é realizado com textos obtidos do uso da linguagem, e portanto deveria pertencer ao domínio linguístico da pragmática, não correspondendo a uma representação do seu significado, mas uma aproximação do seu uso em situações práticas. Igualmente à crítica da semântica vetorial, esta visão de uma representação pragmática inescrutável parece derrotar a própria concepção do conceito de representação.

Por falta de uma classificação melhor, os sistemas de redes neurais acabaram sendo enquadrados como *semânticas distribucionais*. Mais recentemente, o fato de que sistemas baseados em redes neurais podem ser enganados ou desvirtuados por meio de exemplos maldosamente arquitetadas (Bender; Koller, 2020) levou a uma crítica que diz que esses sistemas não representam nem a sintaxe, nem a semântica e nem a pragmática (ibidem). Em particular, a crítica emana

do fato de que não é possível atribuir significado meramente a partir da forma, ou seja, o fato de que as redes neurais estão confinadas a consumir apenas texto faz que esses mecanismos estejam impossibilitados de representação semântica de uma língua natural, uma vez que todas as línguas naturais humanas fazem referência ao mundo exterior, extralinguístico. Essa posição baseia-se no fato de que a aquisição de qualquer linguagem humana se dá pela interação afetiva com outros seres humanos que indicam ao aprendiz os referentes das expressões linguísticas. Assim, a atribuição de semântica para os elementos da linguagem é um processo sociolinguístico, que não pode ser capturado por qualquer algoritmo, rede neural ou outro, que esteja limitado a exemplos do campo linguístico apenas. Os métodos algorítmicos reconhecedores de padrões seriam limitados a capturar os estereótipos existentes na linguagem, que são uma fração superficial do fenômeno linguístico.

Correlação versus *causalidade*

O comportamento de entidades inteligentes envolve a captura e o emprego e relações causais entre conceitos. No caso das ferramentas modernas de aprendizado automático, não está claro nem se elas são capazes de formular conceitos úteis para o raciocínio causal. Além disso, seu funcionamento centrado no aprendizado de padrões mostra que o foco está na obtenção de correlações probabilísticas entre os dados de treinamento. Em particular, a atenção neural é uma forma sofisticada de capturar correlações.

Há uma diferença importante entre correlação e causalidade. O segundo implica o primeiro, mas não o contrário. Dessa forma, é ilusório pensar que uma ferramenta que aprende correlações possui imediatamente um comportamento inteligente. Isso pode ser verificado nas diversas falhas observadas nas redes neurais.

Esses tropeços são talvez mais facilmente percebidos quanto as redes neurais são usadas no processamento de imagens. Sistemas treinados para classificar retratos de pessoas nas categorias masculino ou feminino acabam detectando correlações espúrias devido a vieses nos dados. Por exemplo, acabam inferindo que um fundo parecido com uma cozinha é indicador de um retrato feminino, mesmo que a figura no centro da foto seja de uma pessoa calva e com barba.

Esses fenômenos também se repetem no processamento de língua natural. O programa *word2vec*, que computa uma codificação para as palavras baseadas em sua ocorrência em milhares de textos apresentava a princípio comportamentos de aparente inteligência. Em um teste intrínseco ele era capaz de encontrar associações entre pares de palavras, do tipo homem está para rei assim como mulher está para X, encontrando X = rainha; Itália está para Roma assim como o Japão está para X, encontrando X = Tóquio. No entanto, esse tipo de associação logo se mostrou problemático como exemplos do tipo: homem está para cirurgião assim como mulher está para X, encontrando X = enfermeira, em vez de encontrar X = cirurgiã. O que pode aparentar ser um sexismo implícito nas

redes neurais nada mais é que um artefato probabilístico, capturando um viés de caráter cultural, pois nos textos fornecidos para treinamento a probabilidade de enfermeira aparecer no contexto em que a palavra cirurgião também ocorre é muito maior do que a probabilidade de encontrar a palavra cirurgiã. Esse viés advém da identificação de um padrão global do conjunto de textos, mesmo que não haja nenhum texto explicitamente sexista no conjunto utilizado para treinamento. Esse comportamento revela um aprendizado de vieses escondidos, e acaba reproduzindo preconceitos latentes, demonstrando outro problema das redes neurais, que é o de repetir comportamentos do passado sem nenhum filtro crítico.

O mesmo fenômeno já foi observado utilizando outros programas além do word2vec, com a inserção de vieses em programas como Bert, GPT3, dentre outros (Bender; Koller, 2020). E na atual tecnologia, não há nada que possa ser feito a não ser inspecionar os dados de treinamento em busca de vieses conhecidos ou descobertos *a posteriori*, a fim de evitar o viés no próximo retreinamento, sem nenhuma garantia de que outros vieses escondidos não estejam manifestos nos dados.

Dessa forma, a fronteira de pesquisa em novos algoritmos de inteligência artificial em geral, e de processamento de língua natural em particular, deve buscar sanar a deficiência da incapacidade de aprender relações causais. Esse é um problema verdadeiramente difícil, pois até o estabelecimento de relações causais de forma não automática é bastante dispendioso e difícil nas áreas experimentais, tais como na medicina, psicologia e economia.

No momento temos investigações que buscam conectar modelos neurais com outras formas causais de raciocínio, tais como as redes bayesianas e o raciocínio lógico baseado em regras (Scarselli et al., 2009; Dwivedi et al., 2020). Essas pesquisas ainda precisam amadurecer para que possamos discutir os seus resultados, mas indicam uma linha a ser perseguida.

Problemas éticos

Os problemas éticos do processamento de texto baseado em métodos de captura de padrões saltam aos olhos. Qualquer discussão que envolva a geração de artefatos estatísticos baseados em vieses implícitos nos dados tem o potencial de deflagrar inúmeros problemas éticos.

Por exemplo, mencionamos acima que métodos baseados em redes neurais são capazes de fazer emergir posturas sexistas implícitas num conjunto muito grande de dados, mesmo que não possamos apontar um único texto de matiz sexista no conjunto de dados usados para treinamento. Nesse caso, temos um padrão sexista, que sorrateiramente apresenta cirurgiões do sexo masculino ao lado de enfermeiras do sexo feminino, sem nunca apresentar uma cirurgiã do sexo feminino sendo instrumentada por um enfermeiro do sexo masculino. Dessa forma, os textos apenas descrevem cenas corriqueiras para o leitor, o qual não se dá conta do desequilíbrio na frequência em que há alguém do sexo masculino numa posição dominante auxiliado por uma pessoa do sexo feminino numa po-

sição coadjuvante. No entanto, esse padrão corriqueiro é capturado pelas redes neurais em sua análise de conjuntos extensos de dados e, dependendo da aplicação, pode ter o efeito de reproduzir e ampliar esse desequilíbrio nas frequências das observações.

Dessa forma, os programas de processamento de língua natural podem se tornar ferramentas de reforço e propagação de comportamentos eticamente questionáveis.

O que descrevemos acima não é uma característica única de algum tipo de aplicativo, nem de alguma tecnologia específica. Pelo contrário, ele é indissociável de toda atividade que visa a captura de informações a partir de contextos.

Sendo assim, é uma obrigação do treinamento de novos profissionais e especialistas na área de processamento de língua natural fazer com que os programadores, projetistas e demais envolvidos no processo da criação de aplicativos de processamento de linguagem estejam cientes dos problemas que este tipo de processamento pode causar. Já se conhece o suficiente sobre os efeitos colaterais desta área para que o processo de formação de recursos humanos seja necessariamente acompanhado instruções sobre o posicionamento ético dos envolvidos, da mesma forma como outras áreas que afetam a saúde e o bem-estar humano também já tratam durante a formação de profissionais do posicionamento ético frente a questões regularmente enfrentadas pela área.

Desafios para o processamento do Português

Quase tudo que foi mencionado acima, tanto na descrição do estado atual do processamento de língua natural, quanto na descrição dos desafios enfrentados pela área até o momento, omitiu quase que por completo o idioma específico em que esse tipo de processamento ocorre.

Na realidade, em se tratando de problemas dependentes de grandes quantidades de dados, há um enorme desbalanço na quantidade de ferramentas e na qualidade dessas, dependendo da quantidade de recursos linguísticos disponíveis para cada uma das línguas. Nesse cenário, línguas como o inglês e o chinês despontam como dominantes e outras, devido ao número de falantes, vêm logo a seguir, como espanhol e francês. Também há trabalhos que combinam recursos multilínguas, como é o caso da arquitetura Bert, para a qual a um pré-treinamento utilizando 104 línguas em contextos autógrafos extraídos da Wikipédia (Pires et al., 2019).

Embora o português seja uma dentre estas 104 línguas, em muitos casos ele ainda é tratado, devidamente, como uma linguagem de baixos recursos linguísticos, especialmente para o desenvolvimento de ferramentas baseadas em Big Data, com diversos trabalhos aceitos para conferências sobre línguas com baixos recursos (Salvatore et al., 2019).

Esse quadro não é tão crônico quanto o de outras línguas, sendo que para o português já foi organizado um corpus geral na linha de utilizar a web como fonte. Trata-se do corpus BrWaC (The Brazilian Portuguese Web as Corpus)

(Wagner Filho et al., 2018), o que já permite uma série de pesquisas linguísticas e computacionais com suas 2,7 bilhões de instâncias de palavras em português brasileiro. Inclusive, baseado nesse corpus foi treinado uma versão para o português brasileiro da arquitetura pré-treinada BERT, chamado de BERTimbau (Souza et al., 2020).

Não obstante essa realização recente, há um número muito grande de variantes pertencentes à área da Bertologia que estão surgindo em ritmo acelerado, e para que possamos atingir um nível de desenvolvimento no processamento de língua natural em português que nos remova definitivamente do rol das línguas com baixa quantidade de recursos linguísticos, é necessário um esforço dedicado à produção de recursos para o processamento do português. Tampouco um único corpus de textos em português brasileiro baseado na web é capaz de suprir todas as necessidades da pesquisa em linguística em geral, e linguística computacional em particular. Essas tarefas de geração de recursos e produção de um pipeline de processamento foram assumidas pelo núcleo de processamento de língua natural em português do recém-inaugurado Center for Artificial Intelligence (C4AI), sediado na Universidade de São Paulo e no momento financiado por um projeto IBM Fapesp de centro de engenharia.

Não se trata apenas de continuarmos a produzir corpus gerais bem como corpus etiquetados para serviço de base para o processamento do português, mas também de produzir uma linha de geração, um pipeline, para o treinamento de novas arquiteturas e novos aplicativos que permitam estender ao português os desenvolvimentos já obtidos em outros idiomas e, quem sabe, promover alguns novos desenvolvimentos originários em nossa língua.

Conclusão

Após descrevermos os imensos avanços experienciados pela área de processamento de língua natural em prazos recentíssimos, bem como enumerarmos alguns dos enormes desafios enfrentados pela área no momento, é justo perguntar se todas essas realizações nos trazem algum conhecimento a mais sobre o fenômeno humano da linguagem.

Pode-se argumentar que todas essas realizações tecnológicas, não obstante a geração de produtos para o mercado e de facilitar uma série de serviços na era da informática, parecem não ter trazido nenhuma informação substancial sobre o processo humano de reproduzir e se comunicar por meio da linguagem. Os produtos do processamento de língua natural certamente ajudaram a alterar a forma de trabalho humano, alteração essa que foi imposta pela pandemia e que, ao menos em parte, já transformou de forma perene a maneira que estávamos acostumados a trabalhar. Seguindo essa linha de raciocínio, o processamento de língua natural teria se dissociado do estudo de linguagem.

Um segundo ponto de vista, bastante mencionado nos corredores e salas de café, é que a tecnologia acabará por matar o estudo tradicional da linguagem, relegando o estudo da linguística a um passado pré-tecnológico.

Nossa visão, em particular, considera que ambas estas visões são exageradas. A linguística é absolutamente fundamental para a área de processamento de língua, uma vez que essa tarefa computacional não explica a língua, não ajuda a prever nem a explicar as evoluções naturais das línguas. O processamento se concentra em o que fazer e como fazer, mas não tem muito a dizer sobre o porquê das coisas. Com o passar do tempo, todo o "oba-oba" se dissipará e o que restar poderá então ser apreciado sobre a ótica da racionalidade, e o que sobrar desse processo de filtragem será incorporado como parte da ciência.

Agradecimentos – O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior Brasil (Capes) – Código de Financiamento 001. Este trabalho foi executado no Centro de Inteligência Artificial (C4AI-USP) com apoio da Fundação de Apoio à Pesquisa do Estado de São Paulo (Processo Fapesp 2019/07665-4) e da IBM Corporation. O autor recebeu apoio parcial dos projetos Fapesp 2020/06443-5 (Spira) e do CNPq 303609/2018-4 (PQ).

Notas

1 Estamos dando preferência ao termo "língua natural" como uma tradução melhor da expressão *natural language*.

2 Testes mostram que aplicação popular de IA ainda tem uma compreensão pobre da realidade. Disponível em: <https://www.technologyreview.com/2020/08/22/1007539/gpt3-openai-language-generator-artificial-intelligence-ai-opinion, acessado em 22.2.21.

Referências

AHO, A. V. et al. *Compilers:* principles, techniques, and tools. New York: Addison-Wesley Longman Publishing Co., Inc. 1986.

AIZERMAN, M. A. et al. Theoretical Foundations of the Potential Function Method in Pattern Recognition Learning. *Automation and Remote Control,* n.25, p.821-37, 1964.

ANDERSON, D.; BURNHAM, K. *Model Selection and Multi-Model Inference*. 2.ed. New York: Springer, 2004.

BAHDANAU, D. et al. Neural Machine Translation by Jointly Learning to Align and Translate. 2016. Disponível em: <http://arxiv.org/abs/1409.0473>.

BARTLETT, P. L. et al. Almost Linear Vc-Dimension Bounds for Piecewise Polynomial Networks. *Neural Comput.,* v.10, n.8, p.2159-73, 1998. Disponível em: <http://dblp.uni-trier.de/db/journals/neco/neco10.html#BartlettMM98>.

BENDER, E. M.; KOLLER, A. Climbing Towards NLU: On Meaning, Form, and Understanding in the Age of Data. In: PROCEEDINGS OF THE 58TH ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, p.5185-98. 2020. Online: Association for Computational Linguistics. Disponível em: <https://doi.org/10.18653/v1/2020.acl-main.463>.

BENTHEM, J. van. *Language in Action*. s. l.: MIT Press, 1995.

BLEI, D. M. Latent Dirichlet Allocation. *J. Mach. Learn. Res.*, v.3, p.993-1022, 2003. Disponível em: <https://doi.org/http://dx.doi.org/10.1162/jmlr.2003.3.4-5.993>.

BROWN, T. B. et al. Language Models Are Few-Shot Learners. 2020. Disponível em: <http://arxiv.org/abs/2005.14165>.

BUCHANAN, B. G. A (Very) Brief History of Artificial Intelligence. *AI Magazine*, v.26, n.4, p.53, 2005.

CARPENTER, B. *Type-Logical Semantics*. Cambridge: The MIT Press, 1997.

CHARNIAK, E. *Statistical Language Learning*. Cambridge: The MIT Press, 1993.

CHOMSKY, N. *Aspects of the Theory of Syntax*. Cambridge: The MIT Press, 1965. Disponível em: <http://www.amazon.com/Aspects-Theory-Syntax-Noam-Chomsky/dp/0262530074>.

CLARK, A.; LAPPIN, S. Unsupervised Learning and Grammar Induction. *The Handbook of Computational Linguistics and Natural Language Processing*, n.57, 2010.

DAMERAU, F. J. *Markov Models and Linguistic Theory*. De Gruyter Mouton, 1971. Disponível em: <https://doi.org/doi:10.1515/9783110908589>.

DEVLIN, J. et al. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. 2019. Disponível em: <http://arxiv.org/abs/1810.04805>.

DWIVEDI, V. P. et al. Benchmarking Graph Neural Networks. 2020. Disponível em: <http://arxiv.org/abs/2003.00982>.

HARRELL, F. E. *Regression Modeling Strategies*: With Applications to Linear Models, Logistic Regression, and Survival Analysis. Springer, 2001.

HOCHREITER, S. et al. Gradient Flow in Recurrent Nets: The Difficulty of Learning Long-Term Dependencies. In KREMER, S. C.; KOLEN, J. F. (Ed.) *A Field Guide to Dynamical Recurrent Neural Networks*. s.l.: IEEE Press, 2001.

HOPCROFT, J. E.; ULLMAN, J. D. *Introduction to Automata Theory, Languages, and Computation*. s. l.: Addison-Wesley Publishing Company, 1979.

HORNIK, K. et al. Multilayer Feedforward Networks Are Universal Approximators. *Neural Networks*, v.2, n.5, p.359-66, 1989. Disponível em: <https://doi.org/http://dx.doi.org/10.1016/0893-6080(89)90020-8>.

JURAFSKY, D.; MARTIN, J. H. *Speech and Language Processing*: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. New York: Prentice Hall PTR, 2000.

KOEHN, P. *Statistical Machine Translation*. s. l.: Cambridge University Press, 2009. Disponível em: <https://doi.org/10.1017/CBO9780511815829>.

LAMBEK, J. The Mathematics of Sentence Structure. *American Mathematical Monthly*, v.65, p.154-69, 1958.

LIU, Y. et al. RoBERTa: A Robustly Optimized Bert Pretraining Approach, 2019. Disponível em: <http://arxiv.org/abs/1907.11692>.

LUONG, M.-T. et al. Effective Approaches to Attention-Based Neural Machine Translation. 2015. Disponível em: <http://arxiv.org/abs/1508.04025>.

MAASS, W. et al. A Comparison of the Computational Power of Sigmoid and Boolean Threshold Circuits. In: ROYCHOWDHURY, V. et al. (Ed.) *Theoretical Advances in*

Neural Computation and Learning. Boston, MA: Springer US, 1994. p.127-50. Disponível em: <https://doi.org/10.1007/978-1-4615-2696-4_4.>

MANNING, C. D.; SCHÜTZE, H. *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press, 1999.

MARTIN, L. et al. CamemBERT: A Tasty French Language Model. In: PROCEEDINGS OF THE 58TH ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 2020. p.7203-19. Dispnível em: <https://doi.org/10.18653/v1/2020.acl-main.645>.

MIKOLOV, T. et al. Efficient Estimation of Word Representations in Vector Space. 2013. Disponível em: <http://arxiv.org/abs/1301.3781>.

MIKOLOV, T. et al. Distributed Representations of Words and Phrases and Their Compositionality. In: BURGES, C. J. C et al (Ed.) *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2016. v.26. Disponível em: <https://proceedings.neurips.cc/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf>.

MINSKY, M.; PAPERT, S. *Perceptrons*. Cambridge, MA: The MIT Press, 1969.

MOORTGAT, M. Categorial Type Logics. In: BENTHEM, A. van; MEULEN, A. T, (Ed.) *Handbook of Logic and Language*. Elsevier North-Holland: The MIT Press. 1997.

NEWELL, A.; SIMON, H. A. GPS, a Program That Simulates Human Thought. In: FEIGENBAUM, E. A.; FELDMAN, J. (Ed.) *Computers and Thought*. s. l.: McGraw-Hill, 1963. p.279-93.

NOVIKOFF, A. B. On Convergence Proofs on Perceptrons. In: PROCEEDINGS OF THE SYMPOSIUM ON THE MATHEMATICAL THEORY OF AUTOMATA, 12, p.615-22. Polytechnic Institute of Brooklyn, New York, 1962.

OCH, F. J.; NEY, H. 2002. Discriminative Training and Maximum Entropy Models for Statistical Machine Translation. In: PROCEEDINGS OF THE 40TH ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, p.295-302. 2002.

PAPADIMITRIOU, H. *Computational Complexity*. s.l.: Addison-Wesley, 1994.

PEREIRA, F. C. N.; SHIEBER, S. M. *Prolog and Natural-Language Analysis*. s.l.: Center for the Study of Language; Information, 1987.

PIRES, T. et al. How Multilingual Is Multilingual Bert?. 2019. Disponível em: <http://arxiv.org/abs/1906.01502>.

ROSENBLATT, F. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, v.65, n.6, p.386-408, 1958. Disponível em: <https://doi.org/10.1037/h0042519>.

_____. *Principles of Neurodynamics*. New York: Spartan, 1962.

RUMELHART, D. E. et al. Learning Internal Representations by Error Propagation. In: RUMELHART, D. E.; MCCLELLAND, J. L. *Parallel Distributed Processing*: Explorations in the Microstructure of Cognition. Cambridge, MA: The MIT Press, 1986a. v.1: Foundations, p.318-62.

_____. Learning Representations by Back-Propagating Errors. *Nature*, v.323, n.6088, p.533-36, 1986b. Disponível em: <http://dx.doi.org/10.1038/323533a0>.

SALVATORE, F. et al. A Logical-Based Corpus for Cross-Lingual Evaluation. In: *Deep Learning for Low-Resource Nlp Workshop at Emnlp 2019*. 2019.

SANH, V. DistilBERT, a Distilled Version of Bert: Smaller, Faster, Cheaper and Lighter. 2020. Disponível em: <http://arxiv.org/abs/1910.01108>.

SCARSELLI, F. et al. The Graph Neural Network Model. *IEEE Transactions on Neural Networks*, v.20, n.1, p.61-80, 2009. Disponível em: <https://doi.org/10.1109/TNN.2008.2005605>.

SOUZA, F. et al. BERTimbau: Pretrained Bert Models for Brazilian Portuguese. In: CERRI, R.; PRATI, R. C. (Ed.) *Intelligent Systems*. Cham: Springer International Publishing, 2020. p.403-17.

SUTSKEVER, I. et al. Sequence to Sequence Learning with Neural Networks. 2014. Disponível em: <http://arxiv.org/abs/1409.3215>.

VAPNIK, V. N. *The Nature of Statistical Learning Theory*. New York: Springer-Verlag Inc., 1995.

VASWANI, A. et al. Attention Is All You Need. 2017. Disponível em: <http://arxiv.org/abs/1706.03762>.

WAGNER FILHO, J. A. et al. The BrWaC Corpus: A New Open Resource for Brazilian Portuguese. In: PROCEEDINGS OF THE ELEVENTH INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION (LREC 2018). Miyazaki, Japan: European Language Resources Association (ELRA). 2018. Disponível em: <https://www.aclweb.org/anthology/L18-1686>.

YANG, Z. et al. XLNet: Generalized Autoregressive Pretraining for Language Understanding. 2020. Disponível em: <http://arxiv.org/abs/1906.08237>.

RESUMO – Neste artigo apresentamos um posicionamento sobre a área de processamento de língua natural em português, seus desenvolvimentos desde o princípio até a explosão de aplicações modernas baseadas em aprendizado de máquina. Exploramos os desafios que a área necessita enfrentar no momento, tanto de natureza técnica quanto de natureza ética e moral, e concluímos com a inabalável associação do processamento de língua natural com os estudos linguísticos.

PALAVRAS-CHAVE: Processamento de língua natural, Redes neurais, Contexto linguístico, Português brasileiro.

ABSTRACT – This is a position paper on the current state of the field of natural language processing (NLP) in Portuguese, its developments from the beginning, and the explosion of recent applications based on machine learning. We explore the challenges that the field is currently facing, of both technical and ethical and moral nature, and conclude with the unwavering association between natural language processing and linguistic studies.

KEYWORDS: Natural language processing, Neural networks, Linguistic context, Brazilian Portuguese.

Marcelo Finger é professor titular do Departamento de Ciência da Computação do Instituto de Matemática e Estatística da Universidade de São Paulo, e pesquisador principal

do USP-Fapesp-IBM Center for Artificial Intelligence (C4AI), onde coordena a área de Processamento de Língua Natural em Português. @ – mfinger@ime.usp.br / https://orcid.org/0000-0002-1391-1175.

Recebido em 2.3.2021 e aceito em 9.3.2021.

[I] Universidade de São Paulo, Instituto de Matemática e Estatística, São Paulo, Brasil.