

Introdução ao processamento de linguagem natural - 26/01/2022

Trata-se do primeiro texto de uma série que visa desvelar, mais tecnicamente do que filosoficamente, o processamento de linguagem natural, mas não sem buscar por bases linguísticas e refletir sobre pontos críticos**[i]**

A despeito de um uso já bastante disseminado, como, por exemplo, os corretores ortográficos, tradutores automáticos, reconhecedores de fala, assistentes virtuais e sistemas de recomendação, o processamento de linguagem natural (PLN) também avança em tarefas menos cotidianas (ainda), entre elas decodificação de códigos secretos[ii], tradutores universais capazes de ouvir um idioma e fazer a conversão a outro simultaneamente[iii] e, por fim, o famoso projeto GPT[iv], da empresa OpenAI, que é capaz de escrever diversos tipos de gêneros textuais.

Se a maioria das aplicações acima estão em franca evolução, Pardo ressalta alguns desafios, a começar pela exigência científica de que as ferramentas não só sejam competentes, mas tenham que desempenhar[v]. Isso implica, por exemplo, que o PLN deve ter uma decodificação universal, mas que não podemos nos esquecer de reflexões sobre a própria linguagem, entre elas, sua contemporaneidade com a evolução do cérebro/mente[vi], mesmo suas implicações com consciência humana, ou o que nos diferenciaria dos animais[vii], a semântica da língua, contexto linguístico [cultural], estrutura sintática e morfológica, que ela é hierárquica e recursiva, etc. Tudo isso leva ao objetivo de construir a inteligência artificial [de máquina], referente ao processamento de linguagem natural, com sua dependência de técnicas e dados.[viii]****

PLN é transdisciplinar[ix], pois envolve fundamentação linguística, representações linguístico-computacionais, estatística e métodos computacionais, que envolvem técnicas simbólicas, redes neurais, etc., além de representação do conhecimento, que veremos aqui em um texto a parte.

De acordo com Pardo, consta que _2001 - Uma Odisseia no Espaço_ inspirou a ciência e a tecnologia, inclusive com a busca do PLN pela criação de um mecanismo que possibilitasse copiar HAL – o robô da nave, ou seja, com muito conhecimento linguístico: fonética, fonologia, morfologia, vocabulário, semântica, uso das palavras, gramática, composição do discurso (conexão de várias frases) e, por fim, conhecimento de mundo para poder ter um diálogo inteligível com um ser humano, quer dizer, aprendizado do senso comum.

Ora, considerando a língua humana uma língua natural, _código de comunicação_, e a linguagem como algo que envolve os _mecanismos_ físicos e cognitivos utilizados na comunicação, etc., deveríamos falar de um processamento de _língua_ natural, à despeito do termo popularizado processamento de _linguagem_ natural, que é uma sub área da IA. PLN, então, visa habilitar o computador a lidar com a língua, como se fosse uma pessoa.

O PLN, conforme o material, teria surgido na 2ª Guerra Mundial, através das tentativas de ingleses e americanos de descobrir os códigos russos. Embora tenha criado expectativas, logo houve uma frustração que passa a trazer resultados por volta dos anos 2000, não como uma inteligência que nos substituirá, mas que serve para nos apoiar.

Pardo lembra das origens das pesquisas, como o programa Eliza[x] que não tinha conhecimento suficiente de mundo e não era capaz de entender o contexto linguístico, por exemplo, não saber o que é era uma alergia. Isso significa que dá para _conversar_ com a máquina, mas fazê-la _entender_ é algo ainda difícil de conseguirmos implementar.

Já o GPT, citado anteriormente, usa rede neural em camadas com mecanismos de atenção (transformer) com capacidade de identificar o que é mais importante. Ele gera o texto a partir dos corpora que ele compõe, mas, ainda que muito superior à Eliza está muito longe do ser humano[xi]. Por fim, Pardo enumera outra série de ferramentas e aplicações que podem ser analisadas durante as pesquisas de PLN, como WolframAlpha (<<https://www.wolframalpha.com/input/?i=what+time+is+it+now>>). Start para perguntas e respostas (<<http://start.csail.mit.edu/answer.php?query=what+time+is+it+now%3F>>), os famosos Watson (IBM) que usa computação cognitiva e teria superado humanos em um jogo, os assistentes virtuais Siri (Apple), Alexa (Amazon) e o da Google.

Por fim, Pardo cita o LX-Suite[xii] que permita fazer testes de análises linguísticas e um primeiro contato com os termos da área de PLN. Ele faz o parse das sentenças e termos como o VISL (<<https://visl.sdu.dk/visl/pt/>>), que é um dos mais usados em língua portuguesa. Há também o OpenWordnet-PT que pode ser usado para desenvolvimento de programas em NLTK usando Python.

Bom, como pudermos ver nessa primeira aula, há muito para aprender e percorrer no caminho do PLN. Mãos à obra!

* * *

[i] Essa reflexão / artigo é uma síntese da aula introdutória da disciplina espelhada “SCC0633/5908 Processamento de Linguagem Natural”, ministrada pelo Prof. Thiago A. S. Pardo no ICMC-USP-SC e, na presente data (25/01/2022), com material aberto e disponível na internet.

[ii] Conforme slide do professor Pardo _How Revolutionary Tools Cracked a 1700s Code_, <<http://www.nytimes.com/2011/10/25/science/25code.html>>.

[iii] Idem, Google Pixel Buds: [Novos Google Pixel Buds mostram por que a tecnologia é maravilhosa - TecMundo](<https://www.tecmundo.com.br/produto/122679-novos-google-pixel-buds-mostram-tecnologia-maravilhosa.htm>).

[iv] Ibidem: <<https://olhardigital.com.br/2020/08/25/noticias/gpt-3-o-mais-poderoso-sistema-de-inteligencia-artificial-ja-criado/>>.

[v] Sobre desempenho, podemos lembrar da crítica foucaultiana de Byung-Chul Han (<<https://www.reflexoesdofilosofo.blog.br/2018/05/paradigmas-do-seculo-xxii.html>>), mas mais concretamente não podemos nos esquecer das críticas de um Ellul (<<https://www.reflexoesdofilosofo.blog.br/2021/07/em-busca-do-metodo-mais-eficaz.html>>) ou Feenberg.

[vi] Ou se estaria à frente, conforme Leroi-Gourhan: <<https://www.reflexoesdofilosofo.blog.br/2021/09/a-mao-que-liberta-lidera-mas-ate-quando.html>>.

[vii] Talvez uma discussão inócua, mas que passa pela capacidade de comunicação dos animais e a nossa, “mais avançada”, comunicação pela linguagem.

[viii] Pardo cita Big Data e mineração de dados, inclusive as críticas relacionadas e “desperdício de processamento”, isso consequência da dataficação, como sabemos.

[ix] Termo criado por Piaget em 1979, é o conhecimento de uma forma plural (<<https://www.significados.com.br/transdisciplinaridade/>>), ou seja, holística e contextualizada, rompendo a divisão das áreas do conhecimento e visando a compreensão dos fenômenos em sua totalidade.

[x] Ver: <<https://pt.wikipedia.org/wiki/ELIZA>>.

[xi] Pode ser testado aqui: <<https://app.inferkit.com/demo>>.

[xii] <<http://lxcenter.di.fc.ul.pt/>>.