Níveis de processamento linguístico - 07/02/2022

Enumera as camadas linguísticas que um bom processador de língua natural deve percorrer, além de trazer aspectos de divisão no campo da linguísticas e notas de implementação**[i]**

Introdução

Pardo traz questões iniciais sobre a nossa relação com a linguagem, com o mundo pela linguagem e mesmo se a nossa cognição sofisticada depende da linguagem. A linguagem seria algo genético? É uma manifestação do inconsciente (Lacan)? Pensamos em alguma língua, por exemplo, em português? Harari[ii], conforme mostra Pardo, destaca o papel da linguagem em nossa evolução e cunha a Revolução Cognitiva em um período entre 70 a 30 mil anos atrás, tendo como teoria mais aceita mutações genéticas acidentais. Já Daniel Everett[iii], que morou no Brasil, traz o pensamento de Chomsky da linguagem inata, a gramática universal, codificada geneticamente. Se é assim, já nasceríamos com a gramática. Mas Everett rechaça essa ideia, por falta de provas de tal hereditariedade ou gene da linguagem (FOXP 2), existente em outros animais. É aí que entra o dilema da IA, a dificuldade de _nós_ explicarmos _nosso_ cérebro. E a dificuldade de um cérebro artificial[iv]. Isso mostra como essas questões linguísticas influenciam a tecnologia da linguagem.

Níveis de processamento da língua

Para observar os níveis linguísticos podemos nos utilizar da ferramenta do LX-Center[v]. Há o _silabificador_ que quebra sílabas das palavras, segmentando-as; o POS-tag (etiquetador de partes do discurso) anota as classes gramáticas[vi] das sentenças; parser (analisador) de dependência, é um analisador sintático, além das classes gramaticais: sujeito, predicado, etc.; parser de constituintes, que monta a estrutura sintática em forma de árvore, grupos nominais, grupos verbais; reconhecimento de entidades: pessoa, organização, localidade, etc., precisa de contexto e análise semântica[vii]; entre outros. É uma boa ferramenta, mas podemos ver que ainda comete erros. Mas, afinal, pergunta Pardo: de que uma máquina precisa para entender a nossa fala e interagir adequadamente?

Há vários níveis de conhecimento que se dividem em uma cadeia de complexidade e abstração. Abaixo a categorização dos conhecimentos, mostrando o eixo do texto falado separadamente[viii], embora os níveis estejam todos conectados.

[](https://blogger.googleusercontent.com/img/a/AVvXsEg5jbi8BzYLmEmowex8PAcbl67r5OQ57gZ57fbXLOPqbWmdCkOwZMBHrIaZjRy20lB-Idx533kSedBmx-Ho-kOByyqjEg2OYlnAsTaUPV73-9ZW4gcdU-r8h1j_S5P2quvWDhzEUuK7cnpTFsSqawfRFLvggNFS3PqMCWXgEb8kVQOOFJ5T_qqJU0NL=s853)

Fonética e Fonologia. O primeiro é o sistema físico de produzir sons e o segundo está no contexto da língua, fonemas, transcrição fonética[ix], como pronunciar cada som.

Morfologia. São os componentes das palavras, morfemas: raiz ou radical, vogal temática, prefixo, sufixo, etc.

Morfossintaxe. Junta palavras e frases para achar as classes gramaticais, vê o comportamento da palavra dentro da frase, substantivo, preposição, verbo, etc. É a etiquetação morfossintática (POS-tag). Segundo Pardo, chegam a um acerto de 98% em português. Aí estaria o "2 milhões de pessoas morreram", exemplo de Thiago, se milhões é numeral ou substantivo, etc. A dificuldade com os advérbios, que estariam na "caixinha do resto", que não couberam em outras classes gramaticais.

Sintaxe. A sintaxe foca na formação das sentenças, como as palavras se combinam. Ocorre no nível da frase, em que posição as palavras ocorrem. As funções: sujeito, predicado, objetos, etc. A estruturação: sintagma nominal, sintagma verbal. É possível ver graficamente e ver os constituintes, fazer a sua análise de dependência. As árvores chegam a 96, 97% de acerto.

Semântica. É o significado, há vários modelos e representações. Significado de palavras, expressões, orações e mesmos textos inteiros. Lexical, composicional e textual. Bola para ser chutada. Bota: calçado. Bater as botas? É para tirar a terra ou morrer? Significado não composicional. Pode-se tentar categorizar cada palavra, conforme o exemplo abaixo.

[![](https://blogger.googleusercontent.com/img/a/AVvXsEiHSblMkV0Evv0uQsDk5dkyXTcU6Qa3ob8dh9QK8ikljw4IJZotTzsMDChv_hZIIVkBjh3wObpg8y0SBr2PYpvnFwHWhNFZ2yRc_PiC3k0DyyrWA6VTT_MYp5aFejccq5LtQrJMgxbzlRNL93gT6Ft9YsMGx7hTbeez0BSYvJvNQbogqJ7UiopAiQ1H=w400-

h130)](https://blogger.googleusercontent.com/img/a/AVvXsEiHSblMkV0Evv0uQsDk5dkyXTcU6Qa3ob8dh9QK8ikljw4IJZotTzsMDChv_hZIIVkBjh3wObpg8y0SBr2PYpvnFwHWhNFZ2yRc_PiC3k0DyyrWA6VTT_MYp5aFejccq5LtQrJMgxbzlRNL93gT6Ft9YsMGx7hTbeez0BSYvJvNQbogqJ7UiopAiQ1H=s754)

Significado por classes, como traços semânticos a partir de taxonomias do concreto e o abstrato, animado e inanimado, etc., são as ontologias. É o que os word embeddings tentam fazer. Há papéis semânticos, como o agente, o tema, o instrumento, etc. “O menino chutou a bola”. Menino é agente, é humano. Bola é tema, artefato. Há relações lexicais: sinônimos (casa/lar), antônimos, hiperonímias, holonomia (todo e parte) e assim por diante. Expressões idiomáticas, metáforas, ironias, etc. Banco de sangue ou banco para sentar: polissêmico, pois há primitiva em comum, sempre é um depósito. Manga de camisa ou manga fruta? É homonímia, pois não tem primitiva em comum, não se sabe o processo que chegou a esses significados, talvez pela falta de conhecimento da origem de cada uso da palavra.

Pragmática e Discurso. O último nível, acima da semântica, está além da sentença, é o nível do todo. Está no nível do relacionamento entre as frases, correferências, intenções, tópicos. Por exemplo, a dificuldade de um assistente virtual de manter o contexto, o histórico da conversa. De uma indecisão de duas frases pode surgir uma dúvida em uma terceira. Há conexões no nível discursivo.

Já a pragmática tem a ver com o contexto de uso, os participantes do diálogo. Força, educação, hierarquia, atitude, etc. Quem está falando? Coisas que estão fora do texto, estilos de escrita e fala, formalidades, protocolos. São coisas da esfera da sociolinguística.

Implementação

Então, tais níveis de conhecimento devem ser formalizados para uso por computadores, e levando em consideração a interação entre eles, ou seja, a simultaneidade. Levar em consideração ambiguidades, variedades, vagueza. São situações que humanos podem tratar, mas que as máquinas precisam ser preparadas. “O coelho foi servido” e “O homem foi servido” são exemplos ilustrativos.

Diante disso, devemos quebrar o problema em partes: a _fase linguística_ , onde o foco é o corpus, _fase de representação_ na qual são formalizadas as regras, criação de embeddings e _fase de implementação_ que é o desenvolvimento, pré-processamento, confecção da interface, ou seja, o sistema em si[x]. No trajeto tenta-se chegar ao máximo possível na fase de

implementação. Sem esquecer que cada área tem que aprender um pouco das outras áreas, seja um informata saber de Saussure ou um linguista de Turing.

Podem haver sistemas com pouco conhecimento linguístico (mais simples) e outros mais profundos, mais semânticos. O conhecimento pode ser representado simbolicamente com árvores, com tabelas, estatísticas. Regras são mais rígidas e padrões mais facilmente aprendidos. O conhecimento também pode ser obtido por linguistas ou de maneira automática e depois revisto.

Disputas linguísticas

Pardo cita Robert Dale, sobre combinar técnicas simbólicas (racionalistas) e não simbólicas (empiristas). São correntes filosófico-linguísticas. Chomsky impulsionou o racionalismo entre os anos de 60-85, postulando uma linguagem inata devido à complexidade para aquisição da linguagem. É a linha da gramática universal, do gerativismo. Então deveríamos olhar para dentro do cérebro, como pensamos, isto é, extrair regras de inferência por meio da inteligência artificial. Chomsky enfatiza um órgão da linguagem que realiza cálculos combinatórios e permite a recursividade. O empirismo anterior a Chomsky (20-60) se baseava em operações gerais de associação e reconhecimento de padrões, ressaltando a importância dos estímulos sensoriais.

Nesse sentido, vem a necessidade do recorte de textos pela criação de córpus. Daí vem a pergunta inversa: como aprendemos tão pouco se há tantos dados? É um direcionamento para o aprendizado de máquina, que vem com a importância que o empirismo adquire novamente. Chomsky estava mais interessado na competência linguística, no conhecimento do falante, ao passo que os empiristas focam no desempenho linguístico, isto é, o uso. Busca de padrões e convenções, não se guiando tanto pelos princípios categóricos que podem variar. Olham-se hoje os córpus e a busca por erros. Já para Eric Laporte, essas diferenças já não são tão evidentes. Há um ciclo entre intuição e exemplos.

Histórico PLN e teoria

Por outro lado, PLN fundou-se nas décadas de 40 e 50 com o uso de gramáticas e autômatos. Os próximos 20 anos dividiram-se entre simbólicos e estatísticos e a criação dos primeiros córpus on-line. A década de 70 trouxe quatro paradigmas: estocástico, lógico, interpretação textual e discurso. Já a partir de 80 o empirismo voltou com força e destacou-se a medição de dados e avaliação de resultados. O fim da década de 90 é de fortalecimento da área de PLN junto com a web, modelos baseados em dados e exploração comercial. De 2000 para cá a predominância é do aprendizado de máquina (redes neurais), sejam supervisados ou não, aprendizado profundo, grandes conjuntos de dados e

modelos distribucionais (numéricos).

Em PLN há uma classificação básica de _recursos estáticos_ (dicionários, léxicos, córpus com ou sem anotação), _ferramental de processamento_ (tokenizadores, analisadores sintáticos, classificadores de polaridades) e _aplicações_ (ferramentas de usuário final, tradutores, revisores ortográficos, minerador de opinião). Em termos de tendências, disparam tópicos de pesquisas relacionadas ao processamento dos textos de redes sociais, análises de sentimentos, assistentes inteligentes, abordagem multimodal (vídeo, imagem, som, legendas). Os embedding treinados pela indústria (word2vec, BERT). Também tratamento de língua cruzada, escalável e agnóstico ao uso específico. Por outro lado, há usos minoritários como preservação de línguas indígenas. E é por aí que vamos nos achando...

* * *

[i] Esse texto é uma síntese da aula 2 da disciplina de Processamento de Linguagem Natural, ministrada pelo Prof. Thiago A. S. Pardo no ICMS-USP-SC.

[ii] Sapiens: Uma Breve História da Humanidade é um livro de Yuval Harari publicado primeiramente em 2014, embora tenha sido lançado originalmente em Israel em 2011, com o título Uma Breve História do Gênero Humano (Wikipédia).

[iii] Linguagem: a história da maior invenção da humanidade. Daniel L. Everett.

[iv] Sobre o tema ver: <https://www.reflexoesdofilosofo.blog.br/2021/10/ia-na-base-da-antitese-homem-maquina.html>.

[v] Recursos linguísticos e tecnologia para português - Universidade de Lisboa | NLX - Grupo de Linguagem Natural e Fala: https://portulanclarin.net/workbench/lx-syllabifier/.

[vi] Preposição, pontuação, nome próprio, etc.

[vii] Por exemplo, cita Pardo, nas bases do ICMC: jornalística e tweets – há muita diferença.

[viii] Um corpus não é _meramente_ um dataset, pois tem fenômenos linguísticos!!

[ix] IPA
(https://www.internationalphoneticassociation.org/):
is the major as well as the oldest representative organization for
phoneticians

[x] Proposta de Bento Carlos Dias da Silva
(https://bv.fapesp.br/pt/pesquisador/91982/bento-carlos-dias-da-silva/
).