UMA ANÁLISE QUALITATIVA DA GRADUALIDADE EM UM MODELO DISTRIBUCIONAL DE AQUISIÇÃO DE CATEGORIAS SINTÁTICAS

Palavras-Chave: aquisição da linguagem; modelagem distribucional; aquisição de categorias sintáticas; análise qualitativa.

Autores:

Maria Paula Cardeliquio Orfanelli [UNICAMP]

Prof. Dr. Pablo Picasso Feliciano de Faria (orientador) [UNICAMP]

INTRODUÇÃO:

Se tornar um usuário competente de qualquer língua é uma tarefa bastante complexa que envolve a aquisição de diversos conhecimentos, sejam fonéticos, fonológicos, semânticos ou sintáticos. Apesar disso, é comum observar crianças de apenas 5 anos que já dominam estruturas complexas de suas respectivas línguas maternas sem terem recebido nenhuma instrução formal para isso. Além disso, as línguas variam entre si e, ainda assim, o processo de aquisição é universal na medida em que independe da língua ou da comunidade em que a criança se encontra inserida.

Um problema que crianças em fase de aquisição enfrentam é o chamado “Problema de Platão”, que diz respeito à lacuna entre os estímulos aos quais ela está exposta e ao conhecimento linguístico adquirido, já que os dados não apresentam de forma explícita todas as informações necessárias para uma aquisição bem sucedida e, mesmo assim, a criança é capaz de generalizar todas as regras necessárias para a língua sendo aprendida (LOPES, 2019).

É possível dividir o problema da aquisição da linguagem em três campos de estudo: "o quê”, “quando” e “como” (PEARL, 2010). “O quê” se refere ao conhecimento linguístico adquirido e requer descrições detalhadas das línguas sendo adquiridas e seu funcionamento. “Quando” se refere ao período em que a criança se torna competente em determinado aspecto da língua sendo adquirida e é um aspecto que tem sido estudado através da observação naturalística e de diversos designs experimentais. Por fim, “como” diz respeito aos mecanismos que permitem que a aquisição da linguagem se dê de forma natural e sem esforço. Apesar dos avanços com imageamento cerebral e novos designs

experimentais, estudar tais mecanismos ainda constitui um grande desafio na área de aquisição, já que, por exemplo, estudos experimentais que limitem a quantidade de input recebida pela criança ou que tentem isolar mecanismos exclusivos da linguagem através de restrições impostas ao processo desenvolvimental da criança são antiéticos.

Logo, a modelagem computacional se tornou um meio bastante produtivo para os estudos de aquisição da linguagem, principalmente em relação a "como" a criança adquire a língua materna. E uma das tarefas envolvidas nesse processo de aquisição é determinar quais seriam as categorias gramaticais, já que tal classificação desempenha um papel importante na organização de sentenças nas línguas e é fundamental para que qualquer pessoa consiga se comunicar efetivamente em sua língua materna. Dentre as várias abordagens para esse problema, uma que vem ganhando força por se apoiar em resultados tanto psicolinguísticos quanto computacionais é a abordagem distribucional, ou seja, aquela que leva em conta os padrões de co-ocorrência para classificar palavras.

O MODELO:

Um algoritmo distribucional para a aquisição de categorias sintáticas foi apresentado em um artigo publicado em 1998 por Redington, Chater e Finch. O modelo apresentado pelos autores mede a distribuição de cada palavra, o que consiste em coletar os contextos em que a palavra ocorre. Um contexto é composto pelas palavras que ocorrem em proximidade a palavra alvo e é armazenado em uma tabela de contingências, com linhas contendo as palavras alvo e colunas com o conjunto de palavras contexto enquanto cada célula grava o número de vezes que a palavra contexto relevante co-ocorreu na posição apropriada com respeito a palavra alvo. Cada fileira da tabela de contingências é o vetor de contexto daquela palavra alvo.

Em seguida é necessário comparar as distribuições de pares de palavras, pois espera-se que palavras com a mesma categoria sintática possuam contextos similares. A medida de similaridade aplicada entre os vetores contexto é coeficiente de correlação de classificação Spearman. Finalmente, deve-se agrupar as palavras com distribuição similares, o que requer um tipo de classificação não hierárquica no espaço de similaridade. Então os agrupamentos formados pelo modelo são comparados com a classificação de referência das palavras alvo para avaliar a performance do algoritmo.

Os autores mostram que as informações distribucionais são de fato úteis para a classificação gramatical no inglês, além de constituírem um mecanismo psicologicamente plausível para a criança. Como a capacidade de adquirir uma língua é

algo universal, em 2019 Faria replicou o modelo do inglês utilizando dados do português brasileiro (PB) para verificar se a informação distribucional continua útil. O modelo se manteve bastante informativo, o que indica a utilidade desse tipo de informação para a aquisição de categorias sintáticas.

Nos modelo citado acima, há a idealização de que o aprendiz recebe e processa todos os dados de uma vez. Por ser assim, é possível selecionar as 1000 palavras mais frequentes para classificação e as 150 palavras mais frequentes como itens relevantes de contexto (pois como vimos no exemplo introdutório, palavras menos frequentes produzem esparsidade na matriz). Afirmar a plausibilidade empírica dessa idealização seria equivalente a sugerir que a criança tem uma experiência instantânea de tudo que ouve durante seus 2 a 3 primeiros anos de vida, podendo então fazer um uso altamente seletivo dos dados. Obviamente, não é este o caso. Sabemos que a criança é exposta aos dados gradualmente e precisa aprender a cada passo, tirando o máximo de informação de toda e qualquer experiência, apesar de limitações de processamento e memória imperfeita.

OBJETIVOS E METODOLOGIA:

Neste contexto, Faria (2019) implementa um modelo distribucional de aquisição das categorias sintáticas para o Português Brasileiro. E este trabalho tem por objetivo estudar qualitativamente o modelo, aprofundando o entendimento de variáveis que possam influenciar a sua gradualidade para estabelecer bases que, futuramente, possam tornar o modelo mais incremental, ou seja, capaz de atualizar seus conhecimentos conforme os dados são apresentados.

Para atingir esse objetivo, foram analisados os arquivos de saída do modelo e buscou-se identificar quais categorias compõem os agrupamentos formados, quais os efeitos de diferentes quantidades de corpus, palavras de contexto e palavras-alvo utilizadas. Além disso, foi feita uma reflexão sobre os limites da informação distribucional.

CONCLUSÕES:

Apesar da falta de parâmetros claros para um estudo qualitativo na literatura, este trabalho buscou explorar algumas formas de realizar este estudo: por meio de gráficos, tabelas comparativas, foco nas categorias que de fato formam os agrupamentos e observando se de fato as classificações de referência e as realizadas pelo modelo fazem sentido. É complicado depreender conclusões definitivas e encontrar um sentido global a partir das análises específicas realizadas, pois as

medidas qualitativas não oferecem generalizações e sim especificidades, até pelo seu caráter exploratório.

Ainda assim, é possível observar que os resultados obtidos reforçam o Problema de Platão: não é possível chegar a um aprendizado semelhante ao das crianças em fase de aquisição levando em conta apenas as propriedades superficiais e lineares encontradas no corpus. É necessário definir quais restrições psicológicas estão em jogo em cada estágio da aquisição e quais inferências são plausíveis de serem feitas. Por outro lado, outra questão observada que contribui para a teorização da aquisição é a de que, utilizando apenas a informação distribucional, o modelo foi capaz de identificar subdivisões das categorias de referência que são importantes para o uso pleno da língua. Isso seria mais um indicativo do valor, embora não da suficiência, da informação distribucional para a aquisição da linguagem.

Além disso, é possível refletir sobre diversos aspectos do problema utilizando esses estudos. Ao observar a variação das palavras contexto, palavras-alvo e quantidade do corpus utilizada, podemos observar que utilizando apenas a informação distribucional o modelo não apresenta uma performance que melhora ou piora de maneira unidirecional, porém, o aprendizado da criança não ocorre de maneira estritamente linear também. Seria importante averiguar, em estudos futuros, como outras fontes de informação interagem para o aprendizado gradual e como seria possível modelar os diferentes “estágios” de conhecimento pelos quais a criança passa.

Outro aspecto demonstrado é que, a informação distribucional parece bastante útil para encontrar diferenças na língua e discriminar até mesmo entre subcategorias de verbo, mas quando se trata de agrupar elementos semelhantes (por outros critérios gramaticais), são necessários outros tipos de informação, como, por exemplo, a morfológica, pelo menos da forma modelada aqui.

A utilidade da informação distribucional parece estar bastante ligada à frequência das palavras utilizadas como contexto, já que utilizar categorias individuais de palavras para classificar as outras, sem levar em conta sua frequência, não resultou em um bom desempenho do modelo. Ainda em relação às categorias, os verbos e substantivos costumam ser os mais bem classificados já que ocorrem em diversos contextos. Além disso, as categorias lexicais são agrupadas juntas com maior frequência do que as categorias funcionais.

Outro ponto que fica evidente nos estudos qualitativos são os problemas causados pela falta de anotações detalhadas no corpus (p.e., reformulações) ou de escolhas de classificação de referência, já que algumas palavras podem fazer parte de mais de uma categoria sendo, assim, fundamental que a categoria escolhida corresponda à mais frequente no corpus utilizado. Esses pontos precisam de constante reflexão e testes para determinar quais escolhas produzem um desempenho melhor.

E, por fim, as grandes categorias de referência não parecem refletir precisamente a realidade distribucional da língua ou o que a criança aprende de fato, visto que ela opera muito bem desde cedo com aspectos distribucionais. Além disso, a limitação de uma única categoria por palavra é certamente uma fraqueza. Embora o modelo faça alguns agrupamentos que não estão de acordo com as classificações de referência, se o objetivo na aquisição das categorias sintáticas é saber os ordenamentos permitidos na língua sendo aprendida, muitos dos agrupamentos do modelo fazem mais sentido do que as classificações de referência. Seria interessante conduzir estudos que levem em conta outros critérios para definir as categorias de referência.

BIBLIOGRAFIA

FARIA, Pablo. Learning parts-of-speech through distributional analysis: Further results from Brazilian Portuguese. **Diacrítica**, [s. l.], v. 33, n. 2, p. 229-251, 18 dez. 2019. DOI doi.org/10.21814/diacritica.415.

LOPES, Ruth Elisabeth Vasconcellos. Os cinco problemas para a teoria linguística: O problema de Platão. In: OTHERO, Gabriel de Ávila; KENEDY, Eduardo (org.). **Chomsky: a reinvenção da Linguística.** Editora Contexto, 2019. cap. 7, p. 143-158. ISBN 8552001373.

PEARL, L. (2010). Using computational modeling in language acquisition research. **Experimental methods in language acquisition research**, v. 27, p. 163.

REDINGTON, M., CHATER, N., e FINCH, S. (1998) Distributional information: A powerful cue for acquiring syntactic categories. **Cognitive science**, v. 22, n. 4, p. 425-469.