doi:10.20396/revpibic.v0i0.id

XXVI Congresso de Iniciação Científica Unicamp
17 a 19 de outubro Campinas | Brasil

Investigação de propriedades dos dados de entrada para aquisição da primeira língua através de um estudo sobre aprendizagem distribucional

Giulia O. Ohashi*

Resumo
Em Redington et al. (1998), o potencial da informação distribucional na categorização lexical do inglês como primeira língua é analisado com base em uma série de experimentos computacionais. Tomando-o como base, replicamos parcialmente o estudo para os dados do português brasileiro (PB), contribuindo assim para uma avaliação translinguística. Dois tipos de corpora foram usados: dados de fala dirigida à criança e de diálogos entre adultos. O resultados indicam que a informação distribucional também se mostra útil na aquisição do PB, corroborando os resultados para o inglês obtidos no estudo original.

Palavras-chave:
Aquisição da linguagem, linguística computacional, método distribucional.

Introdução

A aquisição da linguagem é um processo natural do desenvolvimento humano, ou seja, todas as crianças típicas em sua cognição e que são expostas a um ambiente onde exista comunicação linguística conseguirão desenvolver a língua em um mesmo período de tempo e fluentemente. A presente pesquisa visa contribuir no campo da Aquisição da Linguagem ao simular computacionalmente o processo de categorização de palavras, tentando se aproximar analogamente a uma criança adquirindo vocabulário. Mais especificamente, tendo por base o estudo apresentado em Redington *et al*. (1998), investigamos computacionalmente o potencial da informação distribucional como fonte de informação para a classificação de palavras em classes lexicais no português brasileiro (PB).

Resultados e Discussão

Após a obtenção de dois tipos de dados, fala dirigida à criança (dados do Projeto de Aquisição da Linguagem Oral disponível no CEDAE/Unicamp e da base CHILDES) e fala entre adultos (projeto NURC), foram feitas as preparações dos arquivos, que consistiu basicamente em normalização da escrita (p.ex.: de “minina” para “menina”) e retirada de metadados e dos enunciados das crianças. Implementamos o método distribucional de Redington *et al*. (1998) usando a linguagem de programação Python. Esse método constitui-se de três passos principais: (i) medir a distribuição dos contextos de cada palavra-alvo; (ii) comparar a distribuição das palavras entre si; e (iii) agrupar palavras com distribuições semelhantes.

Como ilustração dos experimentos, apresentamos um estudo com os seguintes parâmetros: 150 palavras-alvo, 150 palavras de contexto (ambos os conjuntos sendo as palavras mais frequentes) e janela de contexto incluindo as duas palavras anteriores e as duas palavras posteriores à palavra-alvo. Os agrupamentos foram avaliados quanto à precisão, à completude e à medida $F_{0,5}$. O melhor agrupamento ($F_{0,5}$ = 0,5) foi obtido no ponto de corte 0,68 do dendrograma (Fig. 1), que gerou 13 grupos de similaridade.

Dentre os grupos obtidos, um deles incluiu preposições e artigos, como o “um” e o “em”, outro conteve substantivos e outro, sobretudo, verbos. Ainda obteve-se um grupo com pronomes e advérbios frequentes, como “não” e “já”. Em simulações para quantidades maiores de palavras-alvo, como mil, por exemplo, acreditamos que as grandes classes de substantivos e de verbos influenciem no aumento da completude em detrimento da precisão. No entanto, pode-se observar na Figura 1 que “um” e “uma” possuem distribuição muito similar. Outro ponto a destacar é a presença do verbo “olha” junto com conjunções e alguns demonstrativos, o que, provavelmente, se deve à sua recorrência em começo de frases. Outros experimentos estão sendo conduzidos.

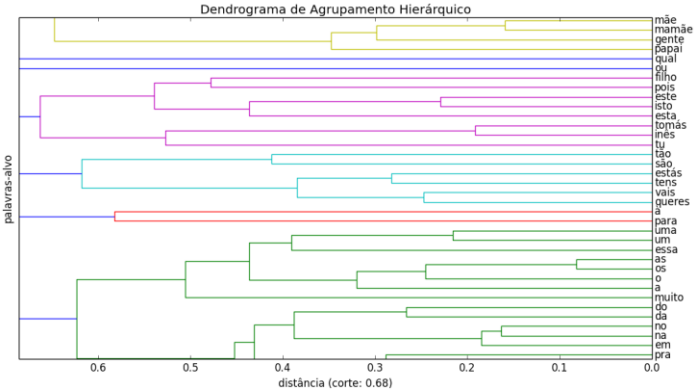

Figura 1. Recorte do dendrograma obtido no experimento.

Conclusões

Este estudo indica que a informação distribucional pode ser bastante informativa para a criança em processo de aquisição da linguagem no PB assim como no inglês. Entretanto, outras fontes de informação são claramente necessárias para que o aprendiz tenha sucesso na tarefa de categorização de palavras.

Agradecimentos

Agradeço ao CNPq/PIBIC/Unicamp que viabilizou a realização desta pesquisa.

Redington, Martin; Chater, Nick; Finch, Steven. Distributional information: A powerful cue for acquiring syntactic categories. Cognitive science, v. 22, n. 4, p. 425-469, 1998.