INVESTIGANDO DIFERENTES CLASSIFICAÇÕES DE PALAVRAS PARA FINS DE APRENDIZAGEM DISTRIBUCIONAL

Palavras-Chave: Aprendizagem distribucional, Aquisição da linguagem, Modelagem computacional

Autores/as:

LORENA DE LA TORRE [UNICAMP]

Prof. Dr. PABLO DE FARIA (orientador/a) [UNICAMP]

INTRODUÇÃO:

Este estudo se insere no contexto do trabalho já em andamento do Prof. Pablo de Faria (Faria e Ohashi, 2018; Faria, 2019a, 2019b), no qual tem sido elaborada uma modelagem computacional com propósito de simular aspectos da aprendizagem de categorias sintáticas, no processo de aquisição da linguagem. Através desse modelo, é possível experimentar isoladamente uma série de parâmetros, avaliando os seus resultados tendo em vista tanto a teoria linguística quanto resultados de estudos experimentais. É a partir desse aprendiz virtual que este projeto investigou a aprendizagem distribucional de categorias sintáticas das palavras no português brasileiro (PB), tendo como foco a forma como classificações de referência diversas, utilizadas para avaliar a performance do modelo, poderiam interferir nos resultados.

OBJETIVOS

O objetivo geral deste projeto é avaliar a adequação teórico-empírica da classificação de referência adotada no estudo em Faria (2019b); ou seja, tendo em vista as leituras e experimentos realizados ao longo do projeto, buscou-se analisar criticamente a classificação atualmente utilizada nos estudos de Faria. Como objetivos específicos:

- Especificar duas classificações de referência alternativas, explorando aspectos teóricos e empíricos, e refazer as simulações computacionais para comparar os resultados com o estudo original;

- Compreender um pouco mais da área de modelagens cognitivas computacionais;

METODOLOGIA:

A fim de atingir os objetivos especificados, foram conduzidas duas etapas: (1) um estudo bibliográfico e (2) uma série de experimentos exploratórios. A primeira delas, sob orientação, consistiu em uma série de leituras sobre aquisição da linguagem, aprendizagem de categorias de palavras, aprendizagem distribucional e sobre o modelo computacional utilizado no projeto.

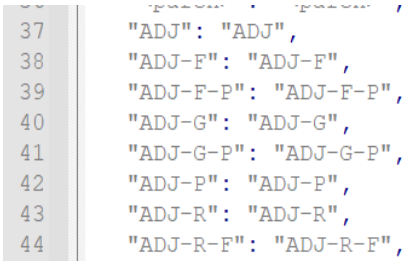

Figura 1: recorte da lista de "tags" no notepad++

Com o estudo bibliográfico como ponto de partida, foram iniciados uma série de experimentos exploratórios. Com o objetivo de avaliar as classificações das palavras, o trabalho teve como material a lista de *tags* (etiquetas morfossintáticas de palavras); essa lista tem como base as classificações sintáticas elaboradas para o corpus *Tycho Brahe* e forma um sistema de equivalências.

No sistema do Tycho Brahe, há 288 *tags* estabelecidas, que servem como base para a classificação das palavras a partir de suas informações distribucionais. Para realizar os estudos sobre as categorias no modelo, elas são restringidas e reagrupadas, o que é possível da seguinte forma: a sigla ao lado esquerdo da coluna corresponde à classificação no corpus Tycho Brahe e a sigla à direita é sua adaptação para as finalidades do modelo, de forma que aquilo que seria categorizado de uma forma (“ADJ-F”) agora será classificado, por exemplo, como um novo agrupamento (“ADJ”).

"ADJ": "**ADJ**", "ADJ-F": "**ADJ**", "ADJ-F-P": "**ADJ**"...

Vale notar que, diferentemente das etiquetas do Tycho Brahe, a classificação de Reddington (1998) tem um pequeno número de etiquetas, contando com 12 *tags*: substantivo, adjetivo, numeral, verbo, artigo, pronome, advérbio, preposição, conjunção, interjeição, contração simples e contração complexa. O intuito desse estudo foi alcançar um “meio termo”, isto é, um sistema mais adequado, que não reduza tanto o número de categorias, mas que também não chegue ao extremo das 288 *tags* do corpus Tycho Brahe.

Quando os experimentos são realizados, com as *tags* adaptadas, o modelo fornece ao pesquisador três resultados numéricos significativos (p: precisão; c: cobertura, f: medida-F), além de um dendrograma, cuja disposição permite uma análise qualitativa dos elementos agrupados. Dessa forma, foram criadas várias adaptações alternativas do sistema de tags com o propósito de observar como certas características poderiam interferir na performance do modelo — e, tendo em vista os resultados alcançados nos primeiros testes exploratórios, novas adaptações, com resultados mais consistentes, foram elaboradas.

RESULTADOS E DISCUSSÃO:

Os experimentos foram realizados tendo em vista a performance da lista de *tags* até então utilizada nas pesquisas de Faria, que tomam como base os estudos de Reddington (1998) — cujas classificações foram pensadas considerando um corpus em inglês —, e a das categorizações do Tycho Brahe, obtidas neste corpus histórico de língua portuguesa.

Dessa forma, cada uma dessas classificações estabelecidas previamente ao projeto — a de Reddington et al. (1998) e a própria classificação do corpus Tycho Brahe — serve como parâmetro para observar os resultados obtidos, cada uma se alocando em polos opostos no que toca a granularidade das *tags*.

- Classificação de Reddington: [best-F] 0.64 (p: 0.71 , c: 0.29);
- Classificação pelo Tycho Brahe: [best-F] 0.51 (p: 0.52 , c: 0.45)

Resultado da classificação alternativa 1

- **Classificação 1:** Verbos no infinitivo e gerúndio

Através da análise qualitativa dos dendrogramas formados pelos experimentos com as listas de referência, nota-se a presença de dois agrupamentos bem claros: verbos no gerúndio e no infinitivo. Com isto em mente, foi elaborada uma classificação que simplificou os verbos da lista do Tycho Brahe em três categorias: **VB** (verbo), **VB-INF** (VB. no infinitivo) e **VB-G** (VB. no gerúndio).

- **Classificação 1. 2:** Verbos no infinitivo e gerúndio e exclusão de todas as marcações de gênero.

  Esta lista foi o resultado de uma mescla da **Classificação 1** com uma outra (classificação com marcações de gênero suprimidas em todas as classes gramaticais), cujos resultados obtidos não foram significativos por si.

Resultado da classificação alternativa 2

- **Classificação 2:** Categoria de palavras funcionais.

De acordo com Faria (2020, p. 108), a presença de “palavras funcionais” — isto é, aquelas que carregam mais significado intralinguístico (dentro da própria estrutura da língua) do que extralinguístico — é importante para a performance do aprendiz. Isso significa que essas palavras geram um contexto mais informativo para a aquisição das outras categorias através da análise distribucional.

Por essa razão, foi criada uma lista em que todas as categorias funcionais estabelecidas pelo Tycho Brahe foram agrupadas em uma só, de modo a não as excluir, mas a simplificá-las.

- **Classificação 2.2:** Junção das Classificações 1 e 2.

A tentativa, intuitivamente promissora, de unir esta lista àquela dos verbos simplificados acrescidos com o infinitivo e o gerúndio não apresentou resultados surpreendentes. Apesar de permitir uma elevação no valor da precisão, houve uma leve queda na cobertura, como podemos observar na tabela na sessão a seguir.

Nome do experimento	Alteração no sistema de etiquetas (com base no Tycho Brahe)	Resultados numéricos
Classif. 1	Agrupamento dos verbos em: VB, VB-INF e VB-G	[best-F] 0.55 (p: 0.57 , c: 0.42)
Classif. 1.2	Exclusão das marcações de gênero + agrupamento: VB, VB-INF e VB-G	[best-F] 0.56 (p: 0.58 , c: 0.42)
Classif. 2	Agrupamento das categorias de palavras funcionais (FUNC)	[best-F] 0.53 (p: 0.54 , c: 0.40)
Classif. 2.2	Agrupamento FUNC + agrupamento: VB, VB-INF e VB-G	[best-F] 0.56 (p: 0.59 , c: 0.39)

Figura 2:Comparação dos resultados numéricos

Discussão:

Para compreender os resultados, é preciso definir o que cada um dos critérios avaliativos. De forma simplificada, a *precisão* diz respeito à porcentagem de pares de palavras que foram agrupados corretamente no modelo. A *cobertura* mede a proporção de pares agrupados na classificação de referência que também foram agrupados pelo aprendiz — isto é, indica quanto daquilo que deveria ter sido agrupado o modelo conseguiu juntar. Ou seja, se todos os itens da mesma natureza A estiveram agrupados, a cobertura será alta, mesmo que esse grupo misture palavras A e B; por outro lado, a precisão será baixa, sendo possível aumentá-la quando os pares gerados forem da mesma categoria que na classificação de referência.

O valor **F** foi uma alternativa, encontrada por Faria (2019b), à medida de *informatividade* utilizada por Reddington, a qual não alcançara uma implementação satisfatória. Esta medida incorpora os outros dois critérios: “The F-score measure is used to integrate these two according to the following general formula: $F\beta = (1 + \beta 2) * \frac{\text{precision} * \text{recall}}{(\beta 2 * \text{precision}) + \text{recall}}$.” (FARIA, 2019b, p. 235), sendo *precision* a precisão e *recall* a cobertura.

Frente aos resultados numéricos, ao contrário do que era esperado, a lista baseada nos estudos de Reddington alcançou a melhor performance e o melhor F. Porém, é possível que ela não seja plenamente adequada aos estudos. O critério de cobertura é muito baixo (c: 0.29) em relação a todas as outras listas de *tags (0.45; o,42; 0,40; 0.39)*, o que significa que, assumindo essa classificação de referência, o modelo não se mostra capaz de juntar as palavras conforme essas classes, talvez em função das particularidades morfossintáticas do PB, realizando agrupamentos muito pouco abrangentes. Isso pode ser observado desde um primeiro momento, tendo como base de comparação a classificação da plataforma Tycho Brahe, cuja cobertura foi bem mais alta (c: 0.45).

A análise qualitativa dos dendrogramas (estruturada como uma árvore, cujas ramificações mostram as relações entre as palavras a partir de seus contextos) obtidos permite observar a

avaliação da performance das etiquetas é dificultada, em parte, pela forma como o modelo realiza sua linha de corte — para obter os agrupamentos, é traçada uma linha que atravessa verticalmente as suas ramificações, de modo que os clusters mantidos à direta são correlacionados às etiquetas.

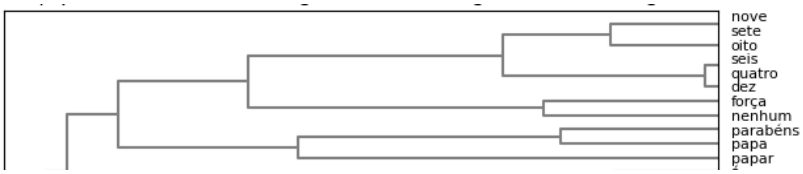

Figura 3: Recorte do dendrograma obtido com a classificação alternativa 1

Ao traçar uma linha reta que atravessa todas os clusters em um mesmo nível, são desconsideradas subdivisões formadas abaixo do corte pelo próprio aprendiz, como podemos observar na Figura 3, com o sub agrupamento preciso de numerais que se funde a outros não relacionados em níveis mais altos da hierarquia. Assim, talvez o que o modelo mais precise seja um método mais dinâmico de avaliar a performance, que leve em conta diferentes níveis do dendrograma.

CONCLUSÃO:

Os experimentos realizados por este estudo, porém, não foram capazes de encontrar uma lista claramente ideal de categorias, como gostaríamos. Os resultados obtidos com classificações alternativas à do estudo original de Reddington et al. (1998), em suas melhores performances, alcançaram valores muito próximos aos da classificação própria do Tycho Brahe, mesmo com uma diminuição substancial do número de *tags*. A nossa hipótese é que a principal limitação do para cálculo das performances das classificações seja, talvez, a forma como os agrupamentos são obtidos, como comentando anteriormente. De todo modo, há inúmeras outras classificações alternativas que poderiam ser avaliadas, para além das apresentadas aqui.

BIBLIOGRAFIA

FARIA, P. (2020). Compreendendo a modelagem computacional de aquisição da linguagem. In: Veredas — Revista de Estudos Linguísticos, v.24, n.1: 94-112.

FARIA, P. (2019a). The role of utterance boundaries and word frequencies in the categorization of words in Brazilian Portuguese through distributional analysis. In: Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics (NAACL'19), 152–159.

FARIA, P. (2019b). Aprendizagem de categorias de palavras por análise distribucional resultados adicionais para Português Brasileiro. Diacrítica, 33(2), 229-251.

FARIA, P. e OHASHI, G. O. (2018). A aprendizagem distribucional no português brasileiro: um estudo computacional. Revista LinguíStica, 14(3): 128–156.

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.