XXXI Congresso de Iniciação Científica Unicamp 2023

ETIQUETAGEM MORFOLÓGICA NO PORTUGUÊS BRASILEIRO: INVESTIGAÇÃO E DESENVOLVIMENTO DE UM ETIQUETADOR AUTOMÁTICO ATRAVÉS DE DADOS DE FALA DIRIGIDA À CRIANÇA

Palavras-Chave: linguística computacional, etiquetagem automática, aquisição de linguagem.

Autores:
EVELYN ELLEN SOARES PASSOS [IEL UNICAMP]
Prof. Dr. PABLO PICASSO FELICIANO DE FARIA (orientador) [IEL UNICAMP]

INTRODUÇÃO:

Uma tarefa fundamental no processamento de linguagem natural consiste na análise e classificação de palavras em categorias gramaticais. Essa etapa, chamada de etiquetagem, permite com que sejam desenvolvidas ferramentas tecnológicas complexas no que diz respeito ao processamento de informação, como tradutores automáticos, sumarizadores, softwares para reconhecimento de fala, entre outros. No entanto, quando realizado exclusivamente de modo manual o processo de etiquetagem pode se tornar extenso, conforme constatado durante o desenvolvimento dessa pesquisa. Desse modo, dispor de uma ferramenta que anote palavras automaticamente torna-se imprescindível do ponto de vista prático, considerando a limitação de recursos que geralmente estão disponíveis ao pesquisador. Portanto, propõe-se como um dos objetivos principais dessa pesquisa de Iniciação Científica o desenvolvimento de um etiquetador morfológico automatizado para o Português Brasileiro.

METODOLOGIA:

A literatura da área da Linguística Computacional apresenta três abordagens para o desenvolvimento de um etiquetador morfológico automático, entre elas, o modelo baseado em *regras*, o modelo *probabilístico* e o modelo *híbrido*, que consiste na junção das anteriores. O modelo baseado em *regras* é constituído a partir de um conjunto de regras previamente estabelecidas manualmente através de uma análise informacional da palavra e do contexto em que ela se encontra. Um etiquetador desenvolvido com base nesse modelo é treinado exclusivamente com as informações fornecidas sobre as regras definidas anteriormente. O modelo *probabilístico*, por sua vez, atribui etiquetas às palavras de acordo com a estimativa probabilística de frequência de uso em um determinado contexto. Dada essa perspectiva estocástica, tem-se o modelo de *n-gramas*, que atribui probabilidades não somente a uma unidade linguística mas a sequências de *n* palavras, de modo que é possível a verificação da probabilidade de palavras futuras em uma frase e até mesmo de uma sentença completa. O modelo de n-gramas é baseado na hipótese de Markov, que busca calcular a probabilidade de ocorrência de uma dada palavra-alvo através da(s) palavra(s) anterior(es) e/ou posteriores. A depender do tamanho da sequência que contém a palavra-alvo e as palavras de contexto considerado na análise, temos as seguintes classificações: bigramas ($n = 2$), trigramas ($n = 3$), quadrigramas ($n = 4$). Por fim, há o modelo *híbrido*, em que o desenvolvimento de um etiquetador envolve a associação de regras previamente estabelecidas e modelos probabilísticos baseados em n-gramas.

Neste trabalho, a metodologia adotada para o desenvolvimento do etiquetador morfológico automático se baseou em uma abordagem híbrida. Em um primeiro momento, foram extraídos dados de fala dirigida à criança do corpus CHILDES para análise e anotação morfológica manual. No decorrer do processo de anotação, verificou-se a complexidade em que se insere a classificação morfológica, dada a possibilidade de ambiguidade na definição lexical. Dada a extensão dessa tarefa, um corpus de aproximadamente 2 mil palavras foram anotadas manualmente e divididas para uso posterior em treinamento e teste do etiquetador automático. A anotação morfológica fundamentou-se em um sistema de etiquetas baseado no conjunto proposto no projeto Universal Dependencies para o Português Brasileiro. A seguir, os modelos mencionados foram testados individualmente, a fim de estudar as

possibilidades oferecidas por cada um, de modo que ao final da pesquisa será apresentado um etiquetador baseado na abordagem híbrida.

O etiquetador proposto está sendo desenvolvido através da plataforma Natural Language Toolkit (NLTK) - uma biblioteca utilizada no desenvolvimento de programas que lidam com dados de linguagem humana -, na linguagem de programação Python. No livro Natural Language Processing with Python, Bird, Klein e Loper (2009, p.198) apontam as classes oferecidas pela NLTK para a etiquetagem automática de textos, são elas: *nltk.DefaultTagger*, *nltk.RegexpTagger*, *nltk.UnigramTagger*, *nltk.BigramTagger* e *nltk.NgramTagger*. Cada um desses módulos realizam a etiquetagem automática através de padrões distintos, por exemplo: o *nltk.DefaultTagger* atribui a mesma etiqueta a cada *token* (palavra), o *nltk.RegexpTagger* atribui etiquetas de acordo com a correspondência de padrões, o *nltk.UnigramTagger* estabelece etiquetas de acordo com seus usos mais comuns e o *nltk.BigramTagger* leva também em consideração o contexto do token anterior para a atribuição de etiquetas. Esses módulos estão sendo explorados durante o desenvolvimento do etiquetador com o objetivo de alcançar o modelo híbrido proposto.

RESULTADOS E DISCUSSÃO:

O presente projeto de pesquisa encontra-se em andamento e, portanto, não apresenta resultados. No entanto, o desenvolvimento da metodologia proposta exibiu pontos de interesse e discussão.

CONCLUSÕES:

O presente projeto de pesquisa encontra-se em andamento e, portanto, não apresenta conclusões.

BIBLIOGRAFIA

BIRD, Steven; KLEIN, Ewan; LOPER, Edward. **Natural Language Processing with Python**; O'Reilly Media, 2009.

HARRIS, Zellig S. Distributional structure. **Word,** v. 10, n. 2-3, p. 146-162, 1954.

JURAFSKY, Daniel; MARTIN, James H. Speech and Language Processing: An introduction to speech recognition, computational linguistics and natural language processing. **Upper Saddle River, NJ: Prentice Hall**, 2008.

PETROV, Slav; DAS, Dipanjan; MCDONALD, Ryan. A universal part-of-speech tagset. **arXiv preprint arXiv:1104.2086**, 2011.

ATWELL, E. S. Development of tag sets for part-of-speech tagging. 2008.

RADEMAKER, Alexandre et al. Universal dependencies for Portuguese. In: **Proceedings of the fourth international conference on dependency linguistics (Depling 2017)**. 2017. p. 197-206.

SAUSSURE, F. de. **Curso de Lingüística Geral**. Tradução Antônio Chelini, José Paulo Paes, Isidoro Blikstein. 28.ed. São Paulo: Cultrix, 2012.

DURAN, Magali Sanches. Manual de Anotação de PoS tags: Orientações para anotação de etiquetas morfossintáticas em Língua Portuguesa, seguindo as diretrizes da abordagem Universal Dependencies (UD). **0103-2569**, 2021.

VIEIRA, Renata; LIMA, Vera Lúcia Strube. Lingüística computacional: princípios e aplicações. In: **Anais do XXI Congresso da SBC. I Jornada de Atualização em Inteligência Artificial**. sn, 2001.

DE ÁVILA OTHERO, Gabriel. **Lingüística Computacional: uma breve introdução**. **Letras de hoje**, v. 41, n. 2, 2006.

DE OLIVEIRA, Lúcia Pacheco. Linguística de Corpus: teoria, interfaces e aplicações. **Matraga-Revista do Programa de Pós-Graduação em Letras da UERJ**, v. 16, n. 24, 2009.

CARVALHO, Cid Ivan da Costa; VASCONCELOS, Davis Macedo; ARARIPE, Leonel Figueiredo de Alencar. **Superando o estado da arte na etiquetagem morfossintática por meio de regras de pós-etiquetagem**. 2012.

AIRES, Rachel Virgínia Xavier. **Implementação, adaptação, combinação e avaliação de etiquetadores para o português do Brasil**. 2000. Tese de Doutorado. Universidade de São Paulo.

MENEZES, Carlos Eduardo Dantas de; NETO, João José. Um método híbrido para a construção de etiquetadores morfológicos, aplicado à língua portuguesa, baseado em autômatos adaptativos. In: **Proceedings of the 2nd Conferencia Iberoamericana en Sistemas, Cibernética e Informática (CISCI'2002), Orlando, USA**. 2002.

FERREIRA, Marcelo; LOPES, Marcos. **Para Conhecer: Linguística Computacional**. Editora Contexto, São Paulo, 2019.