Universidade Federal do Rio Grande do Norte

Rafael Gomes Dantas Gurgel Siqueira

Relatório PLN
Desenvolvimento de um Part-of-Speech Tagger

PLN

2023

Rafael Gomes Dantas Gurgel Siqueira

RESUMO

Testar e validar um algoritmo de Part-of-Speech Tagger (PoS Tagger) sobre a base de dados do Penn Treebank tendo como finalidade aprofundar o entendimento sobre o funcionamento de um PoS Tagger e verificação da acurácia do modelo, foi realizado um estudo com a utilização de monômios para identificar a Tag de cada palavra, com treinamento sobre as sessões 0-18 e validação sobre as sessões 22-24 e obtenção de uma precisão de 87.76%.

Palavras-Chave: Part-of-Speech Tagger, PoS Tagger, Processamento de linguagem natural.

SUMÁRIO

1 INTRODUÇÃO

O Processamento de Linguagem Natural (PLN) é uma área de pesquisa fundamental que tem aplicações em uma ampla variedade de campos, desde assistentes virtuais até tradução automática e análise de texto. Neste contexto, o Part-of-Speech Tagger desempenha um papel crucial, pois permite identificar a classe gramatical de cada palavra em um texto. Este relatório detalha o estudo e a implementação de um algoritmo de Part-of-Speech Tagger.

1.1 Problema de Pesquisa

Como podemos desenvolver e avaliar um algoritmo de Part-of-Speech Tagger usando o Penn Treebank como conjunto de dados de treinamento e validação?

1.2 Objetivos

1.2.1 Objetivo Geral:

O objetivo principal deste projeto é desenvolver um algoritmo eficaz de Part-of-Speech Tagger e avaliá-lo usando a base de dados do Penn Treebank.

1.2.2 Objetivos Específicos:

- Realizar o treinamento do algoritmo com as sessões 0-18 do Penn Treebank.
- Validar o algoritmo com as sessões 22-24 do Penn Treebank.
- Medir a precisão do modelo desenvolvido.

1.3 Justificativa

O estudo e desenvolvimento de algoritmos de Part-of-Speech Tagger são essenciais para a compreensão e aplicação de técnicas de Processamento de Linguagem Natural em diversas aplicações práticas. A precisão do modelo PoS Tagger é um fator crítico para a qualidade de saída em muitas tarefas de PLN.

1.4 Metodologia

Neste projeto, utilizamos o Penn Treebank como nosso conjunto de dados de treinamento e validação. Implementamos um algoritmo em java de Part-of-Speech Tagger que se baseia em monômios para identificar as tags de palavras. O treinamento do modelo foi realizado nas sessões 0-18 do Penn Treebank, e a validação ocorreu nas sessões 22-24. A precisão do modelo foi medida para avaliar seu desempenho.

2 DESENVOLVIMENTO

Nesta seção, descrevemos detalhadamente o processo de desenvolvimento do algoritmo de Part-of-Speech Tagger em java, destacando as etapas-chaves e o fluxo de trabalho do programa.

Coleta de dados do Penn Treebank

Os dados do Penn Treebank foram obtidos a partir do split 0-18 Training, 19-21 Development e 22-24 Testing, os dados são formatados de maneira que cada linha forme uma sentença, e cada palavra da sentença seja seguida pelo caractere “_” (underline) e a sua TAG correspondente, e cada conjunto de palavra + underline + tag separado por espaço.

Pré-processamento dos dados

Os dados foram organizados no código de maneira a otimizar a consulta de comparação entre as tags, por isso foi feito um hashmap entre cada TAG do formato String para um inteiro que corresponde ao índice da tag, e no próprio código cada sentença consiste em uma lista de palavras (String) relacionada a uma lista de inteiros (índice da TAG correspondente).

Desenvolvimento do algoritmo

O algoritmo pode ser alimentado tanto de um arquivo de texto contendo o Penn Treebank formatado como informado anteriormente, ou um arquivo próprio do algoritmo que consiste nos dados já pré-processados, consistindo na primeira linha uma lista de todas as TAGs conhecidas e cada linha posterior uma palavra seguida por uma sequência de números informando a frequência de cada TAG para aquela palavra.

Palavras com sua TAG mais frequente aparecendo menos que 5 vezes ou palavras desconhecidas são classificadas como “UNK”, pois não temos dados suficientes para deduzir uma classe gramatical para ela.

Treinamento e Validação

O treinamento consiste na leitura do arquivo do Penn Treebank (para o exemplo informado neste relatório foram selecionadas as sessão 0-18) e na contagem das respectivas TAGs relacionadas a aquela palavra.

Já a validação consiste em ler alguma sessão do Penn Treebank (para o exemplo informado neste relatório foram selecionada as sessões 22-24) e ao encontrar uma palavra, perguntar ao algoritmo qual seria a TAG que ele daria para ela e comparar com a TAG informada no arquivo do Penn Treebank, caso as TAGs sejam iguais é contabilizado um acerto e caso contrário um erro, ao final é calculado a taxa de acerto geral do algoritmo.

Resultados

Após o treinamento e validação dos dados os resultados obtidos pelo algoritmo foram de 87.76% de precisão ao tentar informar qual TAG dada palavra deveria ter, este resultado se deve principalmente pela escolha de uma grande sessão de treinamento visto que este algoritmo depende de uma larga base de dados e também da escolha de como tratar palavras “desconhecidas”, foi analisado que a alteração do valor limite (no exemplo informado no relatório o valor limite de frequência mínima foi de 5) influencia na taxa de precisão, abaixando demais o valor pode levar a palavras sendo classificadas de maneira errada, o que é pior do que classifica-las como desconhecidas, e subindo muito o valor a taxa de palavras desconhecidas sobre drasticamente o que afeta a precisão já que não conhecer uma palavra consta como erro.

3 CONCLUSÃO

Concluímos que o algoritmo de Part-of-Speech Tagger desenvolvido apresentou uma precisão de 87.76% ao ser validado com as sessões 22-24 do Penn Treebank. Este resultado é promissor e representa um passo importante no aprimoramento das técnicas de PLN. Futuras pesquisas podem focar na melhoria do modelo e na sua aplicação em tarefas específicas.

REFERÊNCIAS

https://drive.google.com/drive/folders/19_F8mmI65lWnL6BmKvtzMX2Z_tcNlXxb - Dados do Penn Treebank.
https://en.wikipedia.org/wiki/Part-of-speech_tagging - Informações sobre PoS Tagger.