Análise dos Algortimos base do Graphia

André Felipe
Karen Teixeira

Visão Geral

O projeto tem como objetivo analisar os resultados de extração de relações e entidades usando os algoritmos de pos-tagger e analise sintática baseados no Graphia, para atribuir semântica a essas relações.

Texto Escolhidos

- Rio de Janeiro
- Rede Globo
- Roberto Marinho

Inspiração: Escolher textos em português que envolvessem uma pessoa, um lugar e uma empresa que possuíssem relações. No nosso trabalho, utilizamos textos retirados da Wikipedia em português, para que pudéssemos extrair automaticamente os textos.

Passo a Passo da extração das árvores de análise sintática

Passo 1:

Recuperação de um dump da Wikipedia no dia 04/11/2014 e usamos o script em python chamado WikiExtractor.py criado por um italiano que pega o dump que vem em XML e transforma num texto limpo, descartando qualquer tipo de informação ou anotação presentes na página da Wikipedia, tais como imagens, referências, tabelas e listas.

Passo a Passo da extração das árvores de análise sintática

Passo 2:

Criação de um script em Java para automatizar a separação de sentenças pertencentes em um texto, onde o retorno da execução é um arquivo texto composto pelas sentenças extraídas do texto selecionado pelo usuário, separadas por linhas.

Passo a Passo da extração das árvores de análise sintática

Passo 3:

Passagem do arquivo retornado pelo algoritmo de separação de sentenças no script parser (algoritmo de análise sintática baseado no Graphia), que usa os pos-taggers mWANN-Tagger e o LX Tagger para “tagueamento” dos termos de cada senteça que serão formatados em forma de árvore e impressos ou no terminal ou em um arquivo texto.

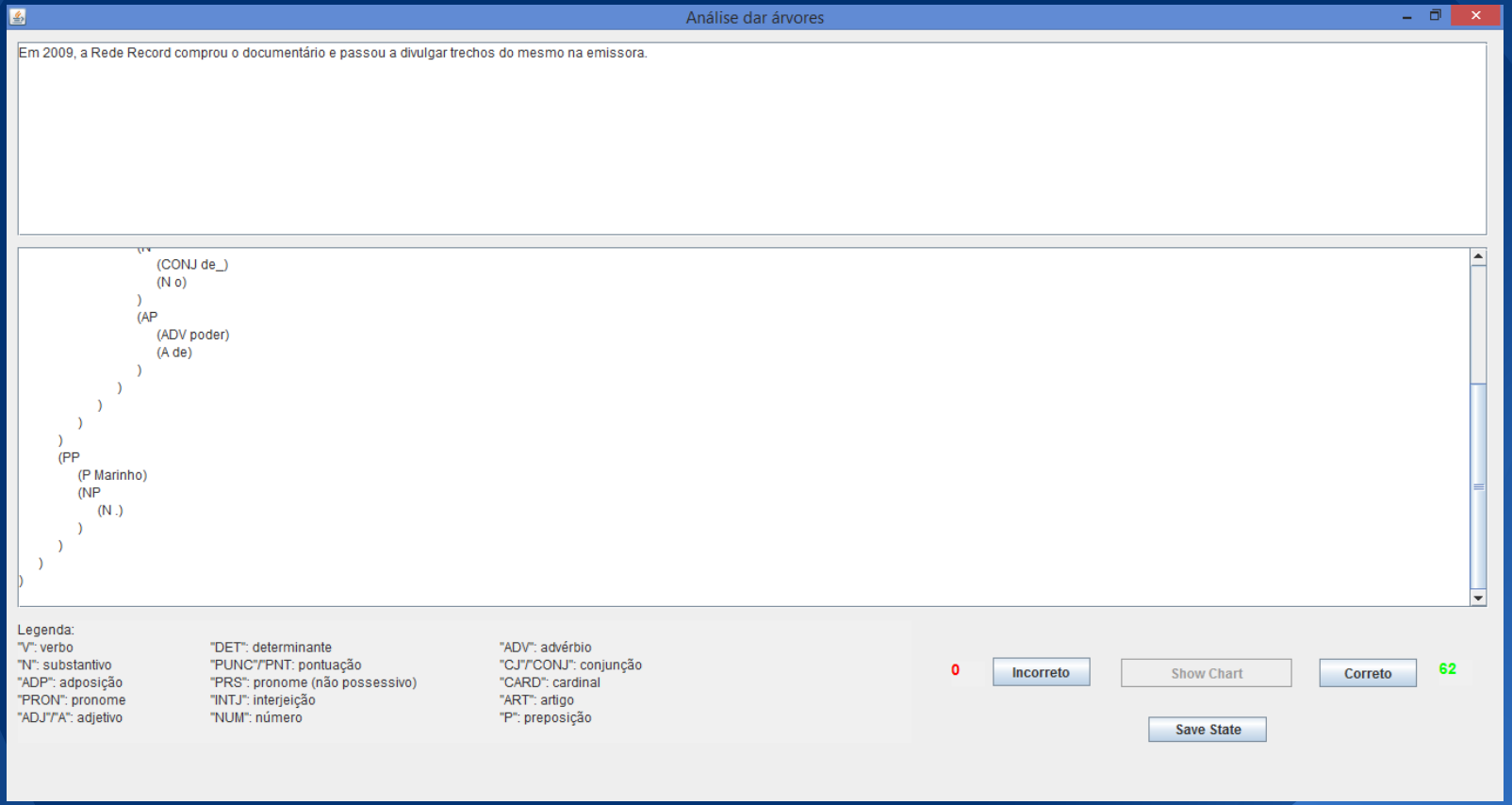
Algoritmo usado na análise das sentenças

Foi criado um script em java para facilitar a visualização das sentenças.

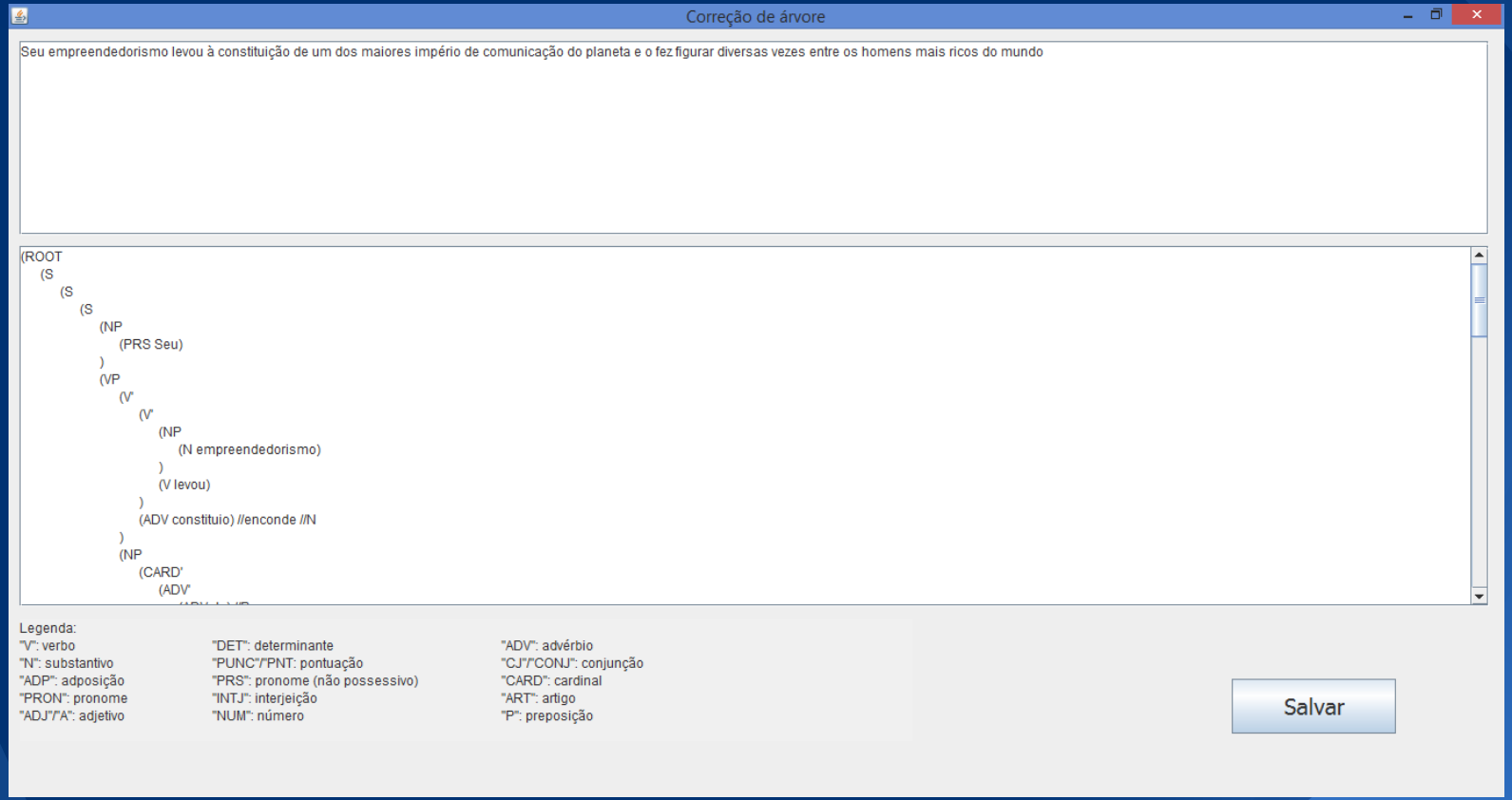
Pelo qual, é possivel visualizar os textos originais e os em forma de árvore resultado dos algorítmo de extrações, consultar uma legenda que indica o significado das tags, e identificar se a sentença foi corretamente parseada ou não e corrigir as sentenças incorretas.

Após analizar todas as sentenças, é gerado um arquivo com o relatório da análise do texto. Além disso, também é possível gerar um gráfico comparativo entre as sentenças corretas e incorretas.

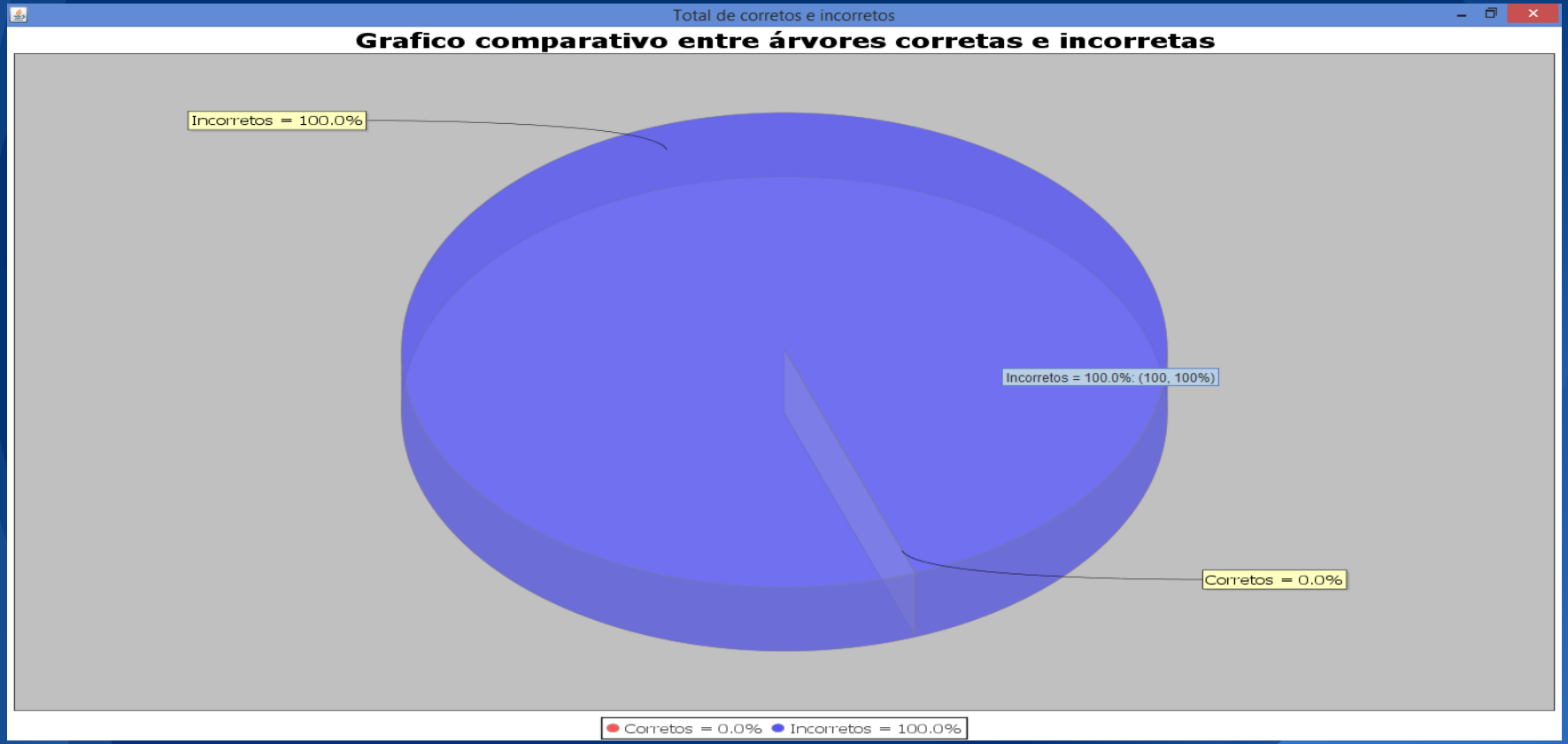
Sistema de análise (Tela de análise)



Sistema de análise (Tela de correção)



Resultados


Total de corretos e incorretos
Grafico comparativo entre árvores corretas e incorretas
Incorretos = 100.0%
Incorretos = 100.0%: (100, 100%)
Corretos = 0.0%
Corretos = 0.0%
Incorretos = 100.0%

Resultados

Exemplo de saída do sistema de análise:

roberto marinho

Número de corretos:
0
0.0%

Número de incorretos:
69
100.0%

Senteças corretas:

Senteças incorretas:
Roberto Marinho
Roberto Pisani Marinho (Rio de Janeiro, — Rio de Janeiro,) foi um jornalista e empresário brasileiro
Proprietário das Organizações Globo de 1925 a 2003, foi um dos homens mais poderosos e influentes do país no século XX

Resultados

A análise dos textos trouxeram um resultado um pouco inesperado.

100% das senteças estavam com algum tipo de incorreção. Salvo apenas por algumas sentenças formadas por títulos que possuem apenas 1 ou 2 palavras, entretanto, essas sentenças são desconsideradas para a anállise.

Problemas encontrados nos resultados

Algumas palavras estavam incompletas devido a presença de algum tipo de acentuação (provavelmente não houve um tratamento de codificação de palavras no algoritmo), e isso fazia com que houvesse uma classificação incorreta.

Além desse problema de encode, a palavra Marinho na entidade nomeada Roberto Marinho foi identificado como adjetivo que é uma das classificações posíveis para a palavra, apesar de estar errado no contexto.

Houveram também, algumas palavras que foram caracterizadas totalmente erradas.

Dificuldades encontradas

- A separação de sentenças foi errada em casos que pontuação não indicassem final de frase, porém foi indiferente nessa fase da análise.

- Alguns bugs foram encontrados no parser, pois ele retornava um erro quando executava com o texto do Rio de Janeiro.

- Dificuldade para relembrar todas as regras gramaticais para realizar a análise.

Trabalhos Futuros

Como trabalho futuro, possuímos a intenção de remodelar o algoritmo do Graphia para que funcione corretamente para textos em português, pois os resultados dessa análise foram negativos e a proposta desse trabalho era avaliar a necessidade de fazer essa remodelagem.

Graphia

From Text to Knowledge Graph