Danielly Sorato

SELEÇÃO E AVALIAÇÃO EXPERIMENTAL DE FERRAMENTAS PARA CLASSIFICAÇÃO MORFOSSINTÁTICA AUTOMÁTICA DE TEXTOS

Monografia submetida ao Programa de Graduação em Ciências da Computação para a obtenção do Grau de Bacharel.
Orientador: Prof. Dr. Renato Fileto
Coorientador: Me. Fábio Bif Goularte

Florianópolis

2016

Ficha de identificação da obra elaborada pelo autor através do Programa de Geração Automática da Biblioteca Universitária da UFSC.

A ficha de identificação é elaborada pelo próprio autor

Maiores informações em:
http://portalbu.ufsc.br/ficha

Danielly Sorato

SELEÇÃO E AVALIAÇÃO EXPERIMENTAL DE FERRAMENTAS PARA CLASSIFICAÇÃO MORFOSSINTÁTICA AUTOMÁTICA DE TEXTOS

Esta Monografia foi julgada aprovada para a obtenção do Título de "Bacharel", e aprovada em sua forma final pelo Programa de Graduação em Ciências da Computação.

Florianópolis, 30 de Outubro 2016.

Prof. Dr.
Mário Antônio Ribeiro Dantas
Universidade Federal de Santa Catarina

Banca Examinadora:

Me. Fábio Bif Goularte
Coorientador

Prof. Dr. Renato Fileto
Orientador

Profa. Dra. Silvia M. Nassar
Universidade Federal de Santa Catarina

Prof. Dr. Elder R. Santos
Universidade Federal de Santa Catarina

RESUMO

Documentos disponíveis na Web (e.g., conteúdos de bibliotecas digitais) e postagens em mídias sociais (e.g., Twitter, Facebook) são fontes abundantes de informações. Nesses textos pode-se encontrar componentes semanticamente ricos denominados palavras relevantes. Essas palavras podem ser, por exemplo, entidades nomeadas (i.e., menções a locais, pessoas, instituições, etc.) ou componentes e expressões que possuem valor sintático e semântico relevantes (e.g., substantivos, verbos, adjetivos, adjuntos). Atualmente, existe uma grande variedade de ferramentas para reconhecimento de palavras relevantes em textos. Tais ferramentas possibilitam extrair, desambiguar e classificar informações valiosas a partir de textos de diversas fontes (literatura, notícias, microblogs, etc). Porém, o desempenho computacional e a qualidade dos resultados produzidos por estas ferramentas costumam ser degradados quando o texto usado é oriundo de mídias sociais. Isso acontece porque o texto de mídias sociais apresenta conteúdo informal, possuindo erros ortográficos e gramaticais, acrônimos, gírias, etc. Este trabalho apresenta uma revisão da literatura sobre técnicas e ferramentas para a extração de palavras relevantes de textos e uma análise experimental de ferramentas para anotação morfossintática automática, com foco em mídias sociais, especialmente microblogs, como o Twitter. As ferramentas de *PoS Tagging* selecionadas são avaliadas em 3 estudos de caso: (i) um *benchmarch* de classificação morfossintática de textos de *tweets* com regras ouro para mensurar precisão e cobertura; (ii) uma amostra do *corpus* histórico do português Tycho Brahe e (iii) um volume considerável de *tweets*. Com isso pretende-se analisar a cobertura e precisão usando o *benchmark* além do desempenho em dados reais de um*corpus* eletrônico e de um microblog. Os resultados mostram evidência experimental de que os resultados das ferramentas de *PoS Tagging* para textos oriundos de mídias sociais são piores do que para textos de linguagem formal.

Palavras-chave: Processamento de linguagem natural, reconhecimento de palavras relevantes, extração de informações, anotação morfossintática, *tweets*.

ABSTRACT

Documents available on the web (eg, digital library content) and posts on social media (eg, Twitter, Facebook) are abundant sources of information. In these texts it is possible to find semantically rich components called relevant words. These words may be, for example, named entities (i.e., references to places, people, institutions, etc.) or components and expressions that have relevant syntactic and semantic value (e.g., nouns, verbs, adjectives, adjuncts). Currently, there is a wide variety of tools for relevant words recognition in texts. Such tools make possible the extraction, disambiguation and classification of valuable information from various sources (literature, news, microblogs, etc). However, the computational performance and quality of results produced by these tools tend to be degraded when the text used is from social media. This happens because the social media text has informal content, spelling and grammatical errors, acronyms, slang, etc. This paper presents a literature review of techniques and tools for the extraction of relevant words of text and an experimental analysis tools for automatic morphosyntactic annotation, focusing on social media, especially microblogs like Twitter. The selected PoS Tagging tools are evaluated in three case studies: (i) a benchmarch for morphosyntactic classification in tweets texts with golden rules for measuring accuracy and coverage; (ii) a sample of the corpus, Tycho Brahe, wich contains historical texts in portuguese (iii) a considerable volume of tweets. This is intended to analyze the coverage and accuracy using the benchmark and the performance on real data from an electronic historic corpus and a microblog. The results have shown experimental evidence that the results of PoS Tagging tools for texts from social media are worse than for formal language texts

Keywords: Natural processing language, relevant words recognition, information extraction, morphosyntactic annotation, tweets.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

SUMÁRIO

1 INTRODUÇÃO

A quantidade de dados disponíveis em meio digital no mundo atual vêm crescendo exponencialmente (HABIB; KEULEN, 2014). A World Wide Web (ALBERT; JEONG; BARABÁSI, 1999) e mídias sociais (e.g. Facebook, Twitter, Instagram) têm contribuído de maneira significativa para este crescimento. Por exemplo, milhões de *tweets* são postados diariamente. Tais dados podem conter componentes relevantes e semanticamente ricos, tais como entidades nomeadas (menções a pessoas, locais, organizações, datas, etc.) (SANG; MEULDER, 2003), além de palavras com classificação gramatical (substantivos, adjetivos, verbos, etc.) que podem carregar semântica relevante (MARTIN; JURAFSKY, 2000).

Tais componentes linguísticos podem auxiliar no entendimento do que se passa com os usuários que fazem as postagens, pois permitem identicar eventos ocorridos ao longo do espaço e do tempo (WANG; STEWART, 2015; ANANTHARAM et al., 2015; XIA et al., 2015), locais frequentemente visitados (FILETO et al., 2015) ou até alimentar com informação precisa métodos de análise de informação (LEV; THIAGARAJAN, 1993; SACENTI et al., 2015), mineração de dados (BERRY; LINOFF, 1997), análise de sentimento (DAS; ACHARJYA; PATRA, 2014) e recomendação (PAZZANI; BILLSUS, 2007), entre outras possibilidades.

Dada a grande quantidade de aplicações para dados textuais disponíveis na Web, surge a necessidade de extração de informações (Information Extraction - IE) de textos. Porém, dados da Web e particularmente postagens em mídias sociais apresentam um novo e desafiador estilo de texto para as tecnologias de processamento de linguagem natural (Natural Language Processing - NLP) (CHOWDHURY, 2003). Diversas aplicações que usam dados de mídias sociais precisam analisar eficientemente a sintaxe e a semântica de grandes quantidades de dados gerados constantemente, cujos padrões linguísticos podem ser dinâmicos, com natureza informal e muitos ruídos, tais como erros ortográficos e gramaticais, presença de acrônimos, gírias, etc. Tais fatores tornam o processo de extração e classificação de informações de mídias sociais particularmente complexo, sendo difícil obter bons resultados por meio de métodos e ferramentas de NLP e alternativas clássicas.

Várias técnicas têm sido propostas para a extração de conhecimento em textos. Essas técnicas visam a extração de palavras relevantes, classificando-as de acordo com taxonomias predefinidas e em alguns casos, desambiguando-as através de URIs (Uniform Resource Identi-

fier - URI) que identificam entidades do mundo real (RIZZO; TRONCY, 2012). Atualmente tecnologias semânticas (HENDLER, 2009) vêm se espalhando em vários domínios de aplicação como um meio confiável e consistente para enfrentar os desafios relacionados com a organização, manipulação, visualização e intercâmbio de dados e conhecimento.

A exploração do conteúdo de textos em mídias sociais, bem como as ferramentas para a extração de informações significativa destes é uma área emergente no cenário de tecnologia e pesquisa, que ainda está em desenvolvimento. Nos últimos anos, uma série de ferramentas surgiram combinando NLP e Dados Ligados (Linked Data - LD) (BIZER; HEATH; BERNERS-LEE, 2009) para reconhecimento e desambiguação de entidades nomeadas, assim como para a realização de anotação morfossintática automática. Há ferramentas implementadas em várias linguagens de programação, com suporte a variados idiomas, que apresentam saídas em diversos formatos e que têm variadas opções de funcionamento (e.g., via serviço, localmente instaladas, programaticamente). Porém falta suporte para as línguas latinas, pois existe uma preponderância de foco na língua inglesa. Por outro lado, o pré-processamento do conteúdo a ser analisado (através de técnicas como normalização (TOMAN; TESAR; JEZEK, 2006) ou fonetização (BIGI; HIRST, 2012)) permite minimizar os erros nas classificações realizadas pelas ferramentas, mesmo em textos com problemas linguísticos como os encontrados em mídias sociais.

A análise de palavras relevantes presentes nos textos de conjuntos de dados tão extensos torna-se inviável sem o uso de tais ferramentas. Além disso a baixa incidência de entidades nomeadas em certos textos da Web, principalmente em mídias sociais, sugere a necessidade de classificação morfossintática para então tentar associar as palavras classificadas com recursos léxicos. Para tanto, é necessário que estas produzam resultados de boa qualidade, com alta precisão e cobertura. Por conseguinte, o foco deste trabalho se volta para a análise do desempenho de ferramentas de classificação morfossintática com dados Web e de mídias sociais, principalmente microblogs, como o Twitter.

1.1 MOTIVAÇÃO E JUSTIFICATIVA

Reconhecer e extrair palavras relevantes é uma tarefa fundamental e o núcleo de NLP. Apesar das pesquisas já realizadas sobre métodos automáticos para reconhecimento de entidades nomeadas, desambiguação e anotação morfossintática automática em textos, muitos

desafios ainda persistem, principalmente quando o texto é oriundo de mídias sociais. Pode-se citar por exemplo a corretude da classificação, resolução de ambiguidade, detecção de sinônimos, correferência e a variabilidade. Vários métodos têm sido utilizados para melhorar a predição das classes corretas, variando de abordagens clássicas, baseadas em dicionário ou em regras e identificação de padrões, até técnicas de aprendizagem de máquina (JABEEN; SHAH; LATIF, 2013).

Como há diferentes métodos usados para tentar contornar os problemas acima, existe uma grande variação nas técnicas utilizadas pelas ferramentas para anotação morfossintática automática dependendo da aplicação. Isso resulta em qualidade dos resultados e desempenhos discrepantes conforme o texto recebido como entrada.

Dada a variação dos resultados das ferramentas em relação aos textos disponibilizados como entrada, é necessário realizar uma avaliação e seleção cuidadosa para descobrir quais as ferramentas mais adequadas para o tipo de texto que se deseja analisar. Isso torna-se ainda mais crucial no contexto de conteúdos textuais provenientes de mídias sociais.

Este estudo parte da literatura sobre extração e classificação de palavras relevantes para focar na análise dos resultados de ferramentas para anotação morfossintática automática de textos em português, principalmente de microblogs. A seleção das ferramentas do estado-da-arte a partir da literatura deverá determinar aquelas que têm obtido melhores resultados em trabalhos similares feitos anteriormente. As ferramentas selecionadas a partir da análise da literatura são então avaliadas nestre trabalho através de experimentos que visam apontar as que são mais aptas para realizar a classificação morfossintática de *tweets*.

1.2 OBJETIVOS

O objetivo geral deste trabalho é fazer uma revisão do estado da arte de métodos e ferramentas para extração de palavras relevantes em texto e avaliar experimentalmente ferramentas selecionadas para a classificação morfossintática automática de textos, com foco em *tweets*.

1.2.1 Objetivos Específicos

Os objetivos específicos desse trabalho são:

1. Realizar uma revisão bibliográfica sobre o estado da arte em classificação morfossintática de textos, com ênfase em microblogs e língua portuguesa.

2. Selecionar métodos e ferramentas identificados na pesquisa bibliográfica.

3. Determinar um conjunto de dados de *tweets* escritos em português, a partir da base de dados disponível no Laboratório de Integração de Sistemas e Aplicações (LISA).

4. Elaborar um padrão ouro para o conjunto de dados de *tweets* criado.

5. Comparar a eficiência e a qualidade dos resultados dos métodos e ferramentas em experimentos com o conjunto de dados de *tweets* e o experimento com o *corpus* eletrônico anotado histórico do português Tycho Brahe. [1]

6. Documentar e caracterizar os resultados obtidos em manuais simplificados das ferramentas analisadas, artigos e monografia.

1.3 METODOLOGIA

Este trabalho de conclusão de curso foi dividido nas seguintes fases: revisão da literatura, análise das abordagens utilizadas pelas ferramentas, seleção dos métodos e ferramentas, preparação do conjunto de dados para teste, confecção do padrão ouro, comparação da performance dos métodos e ferramentas e disseminação dos resultados.

A revisão da literatura buscou técnicas e ferramentas do estado-da-arte para a classificação morfossintática automática, incluindo técnicas voltadas para a língua portuguesa e mídias sociais. Foram usados artigos obtidos nas bibliotecas digitais IEEE Xplore, SpringerLink, ACM Digital Library e ResearchGate assim como trabalhos escritos por colaboradores do LISA. Priorizou-se a inclusão de publicações mais recentes (dos últimos dez anos), com a exceção de trabalhos altamente conceituados e que apresentam soluções sólidas para a resolução do problema de anotação morfossintática. A pesquisa bibliográfica também incluiu alguns trabalhos sobre Reconhecimento de Entidades Nomeadas (Named Entity Recognition - NER) (SANG; MEULDER, 2003) e Desambiguação

[1] Disponível em http://www.tycho.iel.unicamp.br/corpus/

de Entidades Nomeadas (Named Entity Disambiguation - NED) (CUCERZAN, 2007) com o objetivo de incluir uma visão geral do processo de reconhecimento de palavras relevantes, porém o foco do trabalho mantém-se em classificação morfossintática.

O *dataset* utilizado para o experimento com *tweets* é uma amostra que faz parte de uma base de dados coletados no Laboratório de Integração de Sistemas e Aplicações através da ferramenta SeMovGet/Tweet e recursos da LinkedGeoData e da DBPedia através da ferramenta SeMovGet/LOD (KLEIN, 2015). Para o segundo experimento, foi utilizado o *corpus* eletrônico anotado histórico do português Tycho Brahe que possui textos em português escritos por autores nascidos entre 1380 e 1881. Esse *corpus* foi escolhido pelo fato de possuir textos com escrita formal, seguindo rigidamente regras gramaticais e portanto criando um contraste com o conjunto de dados de *tweets* que possui linguagem contemporânea e informal. Nessa etapa serão repetidos os experimentos, porém usando uma amostra do *corpus* selecionado, para então realizar a análise comparativa dos resultados.

A avaliação dos resultados será feita através das métricas de precisão, cobertura e medida-F, cujas fórmulas podem ser averiguadas no apêndice A deste trabalho.

1.4 ORGANIZAÇÃO DO TRABALHO

O restante deste trabalho está dividido nos seguintes capítulos:

1. Fundamentação teórica: detalha conceitos utilizados no decorrer deste estudo, assim como os restringe ao escopo do trabalho;
2. Trabalhos relacionados: apresenta pesquisas recentes, relacionados à proposta apresentada;
3. Análise das ferramentas: descreve as ferramentas levantadas na pesquisa bibliográfica;
4. Experimentos: detalha a execução dos experimentos, bem como os resultados obtidos.
5. Conclusões e trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

Esta seção primeiramente define e classifica os problemas envolvidos no reconhecimento de palavras relevantes em textos e, posteriormente, delineia as principais abordagens da literatura para solucioná-los.

2.1 RECONHECIMENTO DE PALAVRAS RELEVANTES

O Reconhecimento de Palavras Relevantes (Relevant Words Recognition - RWR) consiste em identificar e classificar entidades nomeadas e outros componentes com valor sintático e/ou semântico significativo em textos (DOWNEY; BROADHEAD; ETZIONI, 2007). RWR é considerado em diversas áreas de pesquisa, sendo que cada uma dessas utiliza diferentes abordagens para resolver seus próprios subproblemas. RWR pode ser subdividido em extração de tokens (tokenização), segmentação de sentenças (*chunking*), reconhecimento de entidades nomeadas (NER), desambiguação de entidades nomeadas (NED), *Shallow Parsing* e *POS Tagging*. A Figura 1 ilustra esses subproblemas de RWR, que podem ser vistos como um processo onde somente a tokenização é uma tarefa obrigatória.

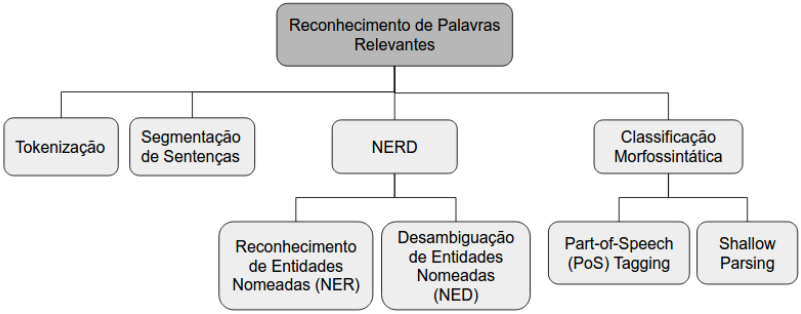

Figura 1 – Subproblemas de RWR.

Técnicas para tratar NER/NED fazem uso de regras sintáticas, dicionários de nomes e/ou *machine learning*. Todavia, tais tarefas só

identificam e classificam um subconjunto de palavras relevantes, as entidades nomeadas. *Shallow Parsing* e *PoS Tagging* (DANIEL; JAMES, 2009), que por sua vez, são tarefas de NLP bastante similares entre si, podem ser usadas para complementar o processo de extração de palavras relevantes, identificando componentes morfossintáticos, tais como verbos (que podem ajudar no entendimento de ações, por exemplo) e adjetivos (que podem ser úteis para detectar polaridade, entre outras possibilidades).

2.1.1 Tokenização

A tarefa de **tokenização** (LABOREIRO et al., 2010) consiste em extrair *tokens* (símbolos separados por espaços e tabulações, tais como palavras e sinais de pontuação) para um conjunto de caracteres. A tokenização pode ser feita com relativa facilidade, por exemplo usando autômatos finitos. Porém, a tokenização tem importância fundamental, pois é uma etapa inicial necessária para todas as outras tarefas (WEBSTER; KIT, 1992). A Figura 2 mostra um exemplo de tokenização, onde cada componente sublinhado da sentença resultante corresponde a um *token*.

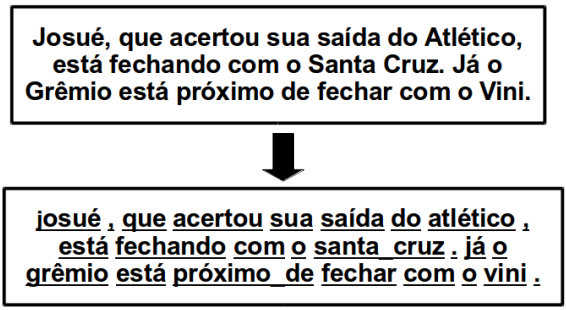

Figura 2 – Processo de tokenização.

Apesar da simplicidade dos métodos de tokenização, são encontrados algumas dificuldades, como a divisão incorreta dos elementos que compõem uma URL, expressões compostas (e.g., São Paulo, próximo de, Cruzeiro do Sul) e palavras que contém símbolos especiais (e.g.,

caixa d'agua). Outro problema encontrado nessa etapa, trata-se do uso de palavras em caixa alta, com letras escritas separadamente (e.g., A M O, D E M A I S). Uma etapa de tokenização bem feita é crucial para o sucesso dos demais processos.

2.1.2 Segmentação de Sentenças

A tarefa de **segmentação de sentenças**, ou *chunking* (SANG; BUCHHOLZ, 2000), consiste em dividir um texto de entrada em sentenças. Gerealmente, alguns sinais de pontuação (e.g., ponto final, exclamação, interrogação) são usados para identificar o final de uma sentença. Os limites de uma sentença são facilmente identificáveis quando há um sinal de pontuação seguido de iniciadores de sentença (i.e., palavra começando com letra maiúscula, travessão), porém são encontrados problemas com siglas, nomes de marcas ou títulos e abreviações. A Figura 3 mostra um exemplo de segmentação de sentenças, onde um texto de entrada foi separado em duas sentenças.

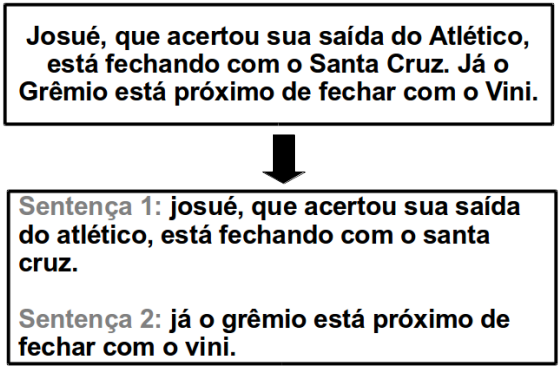

Figura 3 – Processo de segmentação de sentenças (*chunking*).

2.1.3 Entidades Nomeadas

Uma entidade nomeada é uma instância de alguma classe como pessoa, lugar, instituição, número de telefone, valor monetário, etc.

Entidades nomeadas podem ser mencionadas em textos diversos e associados cada qual a uma única descrição de identidade ainda que o nome usado na menção possa se referir a diferentes objetos (e.g., São Paulo estado, cidade, time de futebol, etc). Esta subseção aborda as tarefas de reconhecimento de entidades nomeadas (NER) e desambiguação de entidades nomeadas (NED) em textos.

2.1.3.1 Reconhecimento de Entidades Nomeadas

Named Entity Recognition (NER) (SANG; MEULDER, 2003) é uma tarefa de NLP que consiste em identificar menções a entidades nomeadas em um texto e classificar cada referência com o tipo correspondente (e.g., pessoas, lugares, instituições). NER frequentemente alimenta outras tarefas mais complexas, tais como *Relation Extraction* (DODDINGTON et al., 2004; BOWMAN; DEBRAY; PETERSON, 1993), *Question Answering*(VOORHEES et al., 1999) e NED.

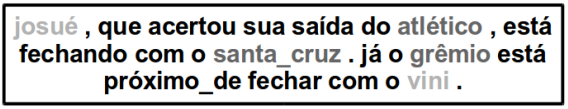

Figura 4 – Reconhecimento de Entidades Nomeadas.

A Figura 4 mostra um exemplo do processo de NER, onde as palavras "josué" e "vini" (em cinza claro) foram reconhecidas e classificadas como pessoas enquanto "atlético", "santa_cruz" e "grêmio" (em cinza escuro) como organizações.

2.1.3.2 Desambiguação de Entidades Nomeadas

Named Entity Disambigation (NED) (AUER et al., 2007) visa ligar cada menção de entidade nomeada detectada em um texto à sua definição em uma base de conhecimento (Knowledge Base - KB). Apesar de o conjunto que forma o domínio de todas as entidades nomeadas que existem ser possivelmente infinito, o conjunto que forma a imagem para essa função de mapeamento, isto é, as definições das entidades nomeadas em uma KB é finito (KLEIN, 2015). Isso pode levar a proble-

mas adicionais no processo de desambiguação, quando a menção não corresponde a entidade alguma na KB.

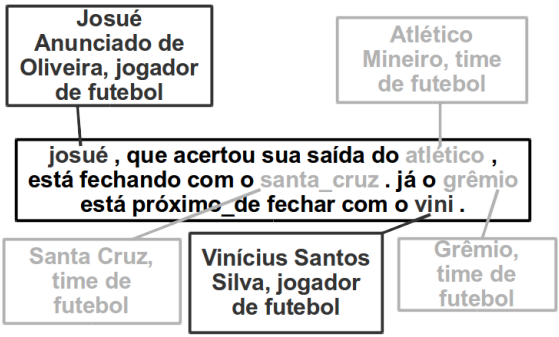

Figura 5 – Desambiguação de Entidades Nomeadas.

Com isso, pode-se inferir que existe o problema de ausência de definições para determinadas entidades nomeadas. É possível também a ocorrência de sinonímia, onde entidades distintas possuem nomes idênticos (e.g., a cidade São Paulo e o time de futebol São Paulo), e a polissemia, onde uma mesma entidade nomeada pode ter diversas menções. As ferramentas para NED utilizam técnicas determinísticas ou de aprendizado de máquina para efetuar a desambiguação com base em informação de contexto e outras que estejam disponíveis (BUNESCU; PASCA, 2006).

A Figura 5 mostra um exemplo de desambiguação para as entidades nomeadas encontradas no texto. Possíveis candidatos encontrados na KB para a desambiguação das entidades nomeadas classificadas como pessoas "josué" e "vini" foram os jogadores de futebol Josué Anunciado de Oliveira e Vinícius Santos Silva, respectivamente. Ao passo que possíveis definições para as entidades nomeadas do tipo organização "atlético", "santa_cruz" e "grêmio" são os times de futebol Atlético Mineiro, Santa Cruz e Grêmio.

2.1.4 Anotação Morfossintática

Esta subseção abordada as tarefas que compreendem a anotação morfossintática de palavras ou expressões.

2.1.4.1 PoS Tagging

A tarefa de ***Part-of-Speech Tagging*** (Part-of-Speech Tagging - PoS Tagging) (VOUTILAINEN, 2003), anota cada palavra em um texto com sua classe morfossintática, com base tanto em definições quanto nos contextos desses componentes no texto. *PoS Tagging* é um problema central em NLP. A identificação precisa dos elementos morfossintáticos de uma sentença é de grande importância, pois ao classificar uma única palavra incorretamente, pode-se gerar erros de processamento subsequentes (TOUTANOVA et al., 2003). Existem diversas técnicas para a implementação de *PoS Taggers*, usando técnicas determinísticas e de aprendizado de máquina, os *PoS Taggers* os mais notáveis são **baseados em regras**, **híbridos** e **estocásticos** (HASAN; UZZAMAN; KHAN, 2007). *Pos Taggers* baseados em regras, como (GREENE; RUBIN, 1971; KLEIN; SIMMONS, 1963; HARRIS, 1962) tentam atribuir *tags* às palavras usando um conjunto de regras escritas à mão. Uma dificuldade deste método é a dependência de especialistas para a criação das regras. Devido à performance superior das técnicas atuais, *PoS Taggers* baseados em regras não são mais utilizados. Os *PoS Taggers* estocásticos, ou probabilísticos, como por exemplo (CUTTING et al., 1992) utilizam um *corpus* de treino para aprender regras baseadas em estatísticas que permitem escolher a *tag* e são baseados em *Hidden Markov Models* (HMM) (RABINER; JUANG, 1986). As abordagens híbridas combinam as duas anteriores, usando por exemplo, a *tag* mais provável com base em um *corpus* de treinamento e depois aplica-se um determinado conjunto de regras para ver se a *tag* deve ser alterada. O uso de aprendizado de máquina em *PoS Taggers* pode explorar dados de treinamento rotulados para se adaptar a novos gêneros ou mesmo línguas, por meio de aprendizagem supervisionada(DERCZYNSKI et al., 2013).

A Figura 6 mostra um exemplo do processo de *PoS Tagging*. Pode-se observar que todos os elementos do texto (incluindo pontuações) receberam uma *tag* que corresponde à sua classe morfossintática.

2.1.4.2 Shallow Parsing

Shallow Parsing é uma alternativa a *parsers* completos (MUNOZ et al., 2000). Em vez de produzir uma análise sintática completa de frases, é executada a análise apenas parcial das estruturas sintáticas

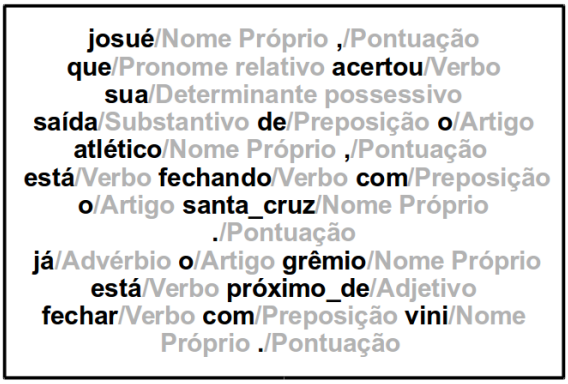

Figura 6 – PoS Tagging.

em um texto (HARRIS, 1957). *Shallow Parsing* identifica os principais constituintes de uma sentença (e.g., frases, orações, grupos de verbos, grupos de substantivos), gerando uma árvore sintática incompleta, mas sem especificar sua estrutura interna e o papel de cada constituinte específico, reduzindo substancialmente a quantidade de memória gasta para realizar a tarefa.

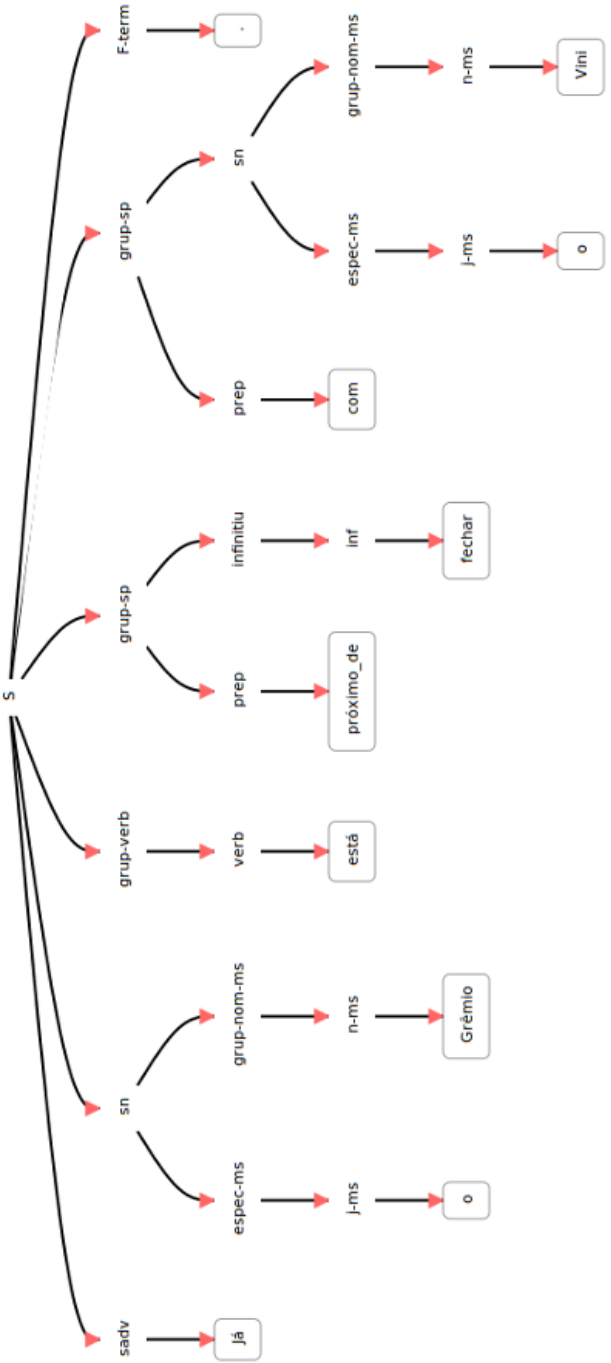

Figura 7 – Shallow Parsing.

A Figura 7 mostra um exemplo da árvore de *parse* incompleta de uma sentença após o processo de *Shallow Parsing*. Pode-se observar que a sentença é dividida em grupos (de verbos, por exemplo), e os nós folha da árvore contêm (grupos de) palavras com as respectivas classes morfossintáticas.

2.2 ABORDAGENS PARA RWR

RWR pode ser realizado com diversas abordagens. Esta seção apresenta algumas das técnicas usadas para realizar a anotação morfossintática, bem como NER/NED.

2.2.1 Técnicas Determinísticas

Dicionário de nomes é uma das técnicas mais simples para NER, onde cada entrada corresponde a um par <nome de superfície, entidade nomeada>. A busca por um nome de superfície em um dicionário seleciona todas as entidades com tal nome (KLEIN, 2015). Contudo, essa abordagem é restrita (só reconhece entidades previstas no dicionário) e não trata ambiguidade (um nome de superfície associado a várias entidades). Assim, geralmente se utilizam dicionários de nomes em conjunto com outras técnicas.

Regras sintáticas, assim como **gramáticas**, se baseiam na estrutura sintática das línguas para identificar padrões em texto. Essas técnicas apresentam grandes dificuldades e taxas de erros altas, assim como performance pior devido à quantidade de informações que precisam ficar armazenadas durante sua execução.

2.2.2 Técnicas Baseadas em Aprendizado de Máquina

Apesar da facilidade de implementação das técnicas determinísticas, métodos que fazem uso de aprendizado de máquina têm gerado melhores resultados atualmente. A maioria das abordagens de aprendizado de máquina nesse âmbito tratam o problema de **classificação de sequências**. Em NLP, o problema da classificação de sequências pode ser usado nas tarefas de *PoS Tagging*, NER e *Shallow Parsing*.

Classificação de sequências é uma técnica robusta que classifica uma sentença por vez. Dada uma sequência de *tokens*, gera-se uma lista de marcadores que indicam a presença de menções. Uma vez

que o classificador de sequências é treinado, utiliza-se o algoritmo de Viterbi (FORNEY, 1973) para calcular uma sucessão de estados que mais provavelmente gera uma coleção de observações, através de programação dinâmica. Problemas de classificação contemporâneos geralmente utilizam classificadores avançados baseados em modelos estatísticos, como por exemplo *Hidden Markov Models* (HMM) (RABINER; JUANG, 1986), *Support Vector Machines* (HEARST et al., 1998), *Conditional Random Fields* (CRF) (LAFFERTY; MCCALLUM; PEREIRA, 2001) e modelos de máxima entropia (RATNAPARKHI et al., 1996). A classificação de sequências pode se beneficiar do uso de *bootstrapping*, também conhecido como auto-treino, quando não existe grande quantidade de dados anotados (KLEIN, 2015).

HMM é uma técnica de modelagem estatística para problemas lineares, tais como séries temporais e análise de sequências (EDDY, 1996). Trata-se de um modelo finito que descreve a probabilidade de distribuição sobre um número infinito de possíveis sequências. O modelo é composto por um número N de estados, que são ligados por transições de distribuição de probabilidade dependentes dos estados anteriores. A máquina de estados emite um símbolo de acordo com a probabilidade de distribuição após uma transição ser feita, que depende do estado atual. Ao chegar no estado final terá sido criada uma sequência de símbolos emitidos pelos estados. O HMM foi usado por muitos anos para resolver problemas de classificação de sequências e inspirou a criação de muitos outros modelos.

Em (BRANTS, 2000), o autor apresenta um ***tagger* estatístico baseado em trigramas** (Trigrams'n'Tags - TnT). O *tagger* usa modelos de Markov de segunda ordem, onde os estados dos modelos representam *tags* e as saídas representam as *palavras*. As probabilidades de transição dependem dos estados, enquanto as probabilidades de saída dependem somente da categoria mais recente. Como trigramas costumam sofrer com o problema de dados esparsos, é usada uma técnica de suavização. Para lidar com palavras desconhecidas, o *tagger* usa o sufixo da palavra para inferir sua classificação morfossintática.

Aprendizagem baseada em memória (Memory-Based Part of Speech Tagger-Generator - MBT) é uma forma de aprendizado supervisionado fundada em raciocínio baseado em similaridade. A anotação de uma palavra em um contexto particular é extrapolada a partir dos casos mais semelhantes presentes em memória (DAELEMANS et al., 1996). Algumas das vantagens que esse método apresenta são o tamanho mínimo necessário relativamente pequeno do *corpus* anotado para treinamento, bons resultados para palavras desconhecidas sem análise

morfológica e a aprendizagem incremental e rápida.

O trabalho de (BRILL, 1995) utiliza **aprendizagem orientada pelo erro e baseada em transformações** (Transformation-Based Error-Driven Learning - TBL) para realizar anotação. A Figura 8 ilustra como o método funciona. Inicialmente o texto de entrada sem anotações passa por um anotador de estados inicial. Esse anotador inicial pode variar em complexidade e uma vez que o texto é processado será então comparado com um *corpus* anotado manualmente que é usado como referência. Uma lista ordenada de transformações é aprendida e pode ser aplicada à saída do anotador inicial para torná-lo melhor e mais semelhante ao *corpus* usado como referência.

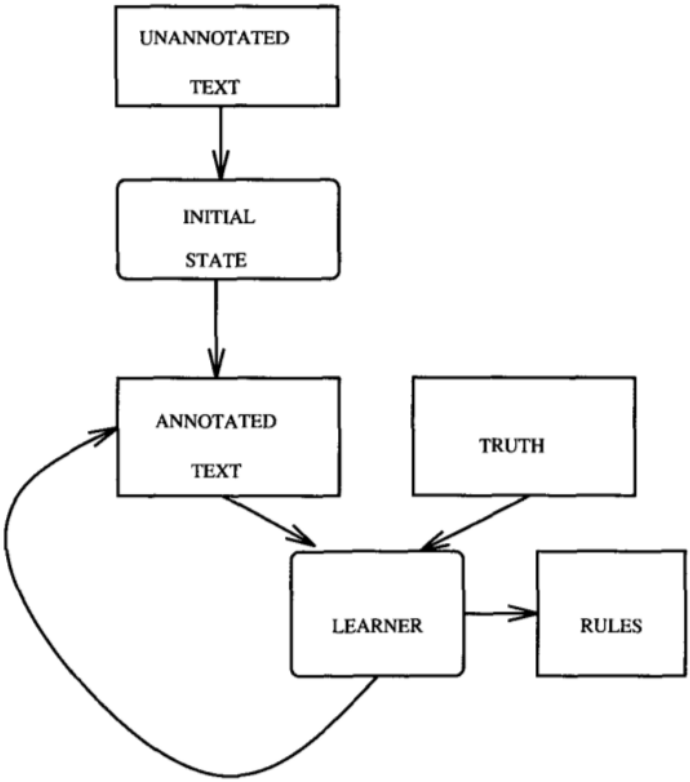

Figura 8 – Método de aprendizado baseado em transformação. Retirado de (BRILL, 1995).

O **modelo de máxima entropia** (RATNAPARKHI et al., 1996) pressupõe que todas as probabilidades são igualmente relevantes e independentes umas das outras. O *tagger* usa elementos de contexto para predizer as anotações. Essa é uma técnica extremamente flexível para a modelagem linguística, que combina vantagens de outras técnicas, uma vez que usa uma representação rica em recursos, como TBL e gera uma distribuição de probabilidade da *tag* para cada palavra, como as técnicas de HMM e árvores de decisão.

Conditional Random Fields (CRF) (LAFFERTY; MCCALLUM; PEREIRA, 2001) usa o princípio de que uma *tag* depende da palavra atual, da imediatamente anterior e da próxima. CRFs podem ser usados para tarefas de NLP como *PoS Tagging*, *Shallow Parsing* e NER (SHA; PEREIRA, 2003).

Ensemble learning (ARBIB, 2003) combina resultados de várias ferramentas. A maneira mais simples de fazer tal combinação é usando algum esquema de votação. Dessa forma, teoricamente os resultados finais são potencialmente melhores, porém obtidos com maiores custos (e.g., tempo de processamento, uso de memória).

3 TRABALHOS RELACIONADOS

Este capítulo discute trabalhos voltados para a proposição e análise de desempenho de métodos e ferramentas de anotação morfossintática automática, visando comparar este trabalho ao estado da arte desta área de conhecimento.

O trabalho de (MEGYESI et al., 2001) compara quatro algoritmos do estado da arte de aprendizado de máquina para realizar *PoS Tagging* em textos da língua sueca. Os algoritmos inclusos no estudo são: HMM (RABINER; JUANG, 1986), máxima entropia (RATNAPARKHI et al., 1996), aprendizado baseado em memória (DAELEMANS et al., 1996) e aprendizado baseado em transformações (BRILL, 1995). Os algoritmos são analisados levando em consideração principalmente o *tagset* utilizado e o tamanho do conjunto de treinamento.

No trabalho de (SANTOS; MILIDIÚ; RENTERÍA, 2008) os autores propõem uma nova abordagem baseada em aprendizado de máquina para *PoS Tagging* em português. A técnica usa transformações guiadas por entropia (Entropy Guided Transformation Learning - ETL) e combina as vantagens de árvore de decisão e TBL. A idéia da abordagem ETL é utilizar indução de árvores de decisão para obter modelos e depois usar o algoritmo TBL para gerar regras de transformação, que são mais eficientes do que árvores de decisão. A técnica é comparada com HMM, TBL, árvores de decisão e Cadeias de Markov de comprimento Variável (Variable Length Markov Chain - VLMC) (BÜHLMANN; WYNER et al., 1999) usando o *corpus* Tycho Brahe[1] e o *corpus* Mac-Morpho (ALUÍSIO et al., 2003). O ETL alcançou resultados superiores em relação às outras abordagens com ambos os *corpus*.

O trabalho de (GIMPEL et al., 2011) faz experimentos com *tweets* em inglês. Durante o estudo foi desenvolvido um *tagset* próprio, anotados manualmente 1.827 *tweets* e então realizados os experimentos, alcançando uma acurácia de até 90%. Como o *tagset* desenvolvido é específico para *tweets*, são trazidas algumas características importantes, como por exemplo a identificação de *hashtags*(#), menções a usuários (forma @usuário), URLs e emoticons. A comparação é feita com o Stanford POS, que usa o modelo de máxima entropia.

O estudo de (RITTER et al., 2011) também foi sob o foco de *tweets* em inglês, em algumas tarefas de NLP como *PoS Tagging*, NER e *Shallow Parsing*. As análises mais importantes envolvem as ferramentas T-NER e T-POS, comparando-a com a Stanford NER e Stanford POS.

[1] Disponível em http://www.tycho.iel.unicamp.br/corpus/

Autor e Ano de Publicação	Foco do Trabalho	Experimentos com Microblogs?	Língua Foco	Abordagens Comparadas
Megyesi, Beáata, 2001	Comparação de algotimos de aprendizado	Não	Sueco	HMM, Máxima Entropia, MBT, TLB
Branco, Antonio e Silva, João, 2004	PoS Tagging para língua portuguesa	Não	Português	TBL, TnT, Máxima Entropia e HMM
Dos Santos, Cícero Nogueira, et al., 2008	Nova abordagem de PoS Tagging	Não	Português	ETL, TBL, TnT, árvores de decisão, HMM, Máxima Entropia e VLMC
Gimpel, Kevin, et al., 2011	Tweets	Sim	Inglês	CRF e Máxima Entropia
Ritter, Alan, et al., 2011	Tweets	Sim	Inglês	CRF e Máxima Entropia

Tabela 1 – Trabalhos Relacionados

A ferramenta T-POS utiliza CRF e obteve acurácia 8% melhor do que a Stanford POS nos experimentos, assim como a ferramenta T-NER que usa CRF, dicionário de nomes e *bootstrapping* obteve resultados notavelmente melhores do que a Stanford NER.

Por fim, em (BRANCO; SILVA, 2004), são mostradas soluções em língua portuguesa específicas para os passos do processo de *PoS Tagging* e também são comparadas as acurácias de quatro *taggers*(TBL (BRILL, 1995), TnT (BRANTS, 2000), QTag(TUFIS; MASON, 1998) e MXPOST(RATNAPARKHI et al., 1996)) usando um *corpus* em português. Note que dentre os 5 trabalhos listados, 2 focam em língua portuguesa e 2 em *tweets*. O primeiro trabalho foi incluido em nossa resenha por se tratar de um comparativo de técnicas de aprendizado de máquina aplicadas à classificação morfossintática. A Tabela 1 mostra um quadro comparativo dos trabalhos relacionados.

Além da defasagem para análises comparativas de métodos para anotação morfossintática automática em português, existe a falta também no âmbito de textos oriundos de mídias sociais. Especialmente, não foram encontradas pesquisas que analisam o desempenho de métodos e ferramentas para *tweets* em português, o que motivou a proposta desse trabalho.

4 ANÁLISE DAS FERRAMENTAS

Esta seção descreve as principais ferramentas para RWR avaliadas no âmbito desta pesquisa e compara suas características. Os critérios de seleção das ferramentas a partir da literatura técnico-científica pesquisada foram: proeminência, grau de recomendação, qualidade da documentação, suporte à Língua Portuguesa ou foco específico no processamento de *tweets*. Foram incluídas neste trabalho algumas ferramentas para NER/NED, para fornecer uma visão geral de RWR. As ferramentas avaliadas foram divididas em ferramentas para *PoS Tagging*, ferramentas para NER/NED e suítes, que são ferramentas que implementam diversas tarefas de NLP.

4.1 FERRAMENTAS PARA *POS TAGGING*

TreeTagger (SCHMID, 2013) é um *PoS Tagger* probabilístico desenvolvido por Helmut Schmid, no Instituto de Linguística Computacional da Universidade de Stuttgart. Ela usa árvores de decisão para contornar o problema de dados esparsos e tem sido aplicado com sucesso em vários idiomas. A árvore de decisão é construida recursivamente a partir de um conjunto de treinamento de trigramas, usando uma versão modificada do algoritmo ID3 (QUINLAN, 1986).

LX-Tagger (BRANCO; SILVA, 2006) é uma ferramenta que faz parte da LX-Suite, desenvolvida pelo *Natural Language and Speech Group* da Universidade de Lisboa, com foco no processamento da Língua Portuguesa, e composta pelos módulos de segmentação de sentença, *PoS-tagger*, tokenizador, *Nominal featurizer* e *Nominal lemmatizer*. O segmentador de sentenças é um autômato de estados finitos, no qual as transições são ativadas por sequências de caracteres especificadas na entrada e os símbolos emitidos correspondem aos limites da sentença e do parágrafo. O *PoS-tagger* utilizado é o (BRANTS, 2000), treinado com 90% de um corpus de 280.000 *tokens*. A tokenização é realizada identificando *tokens* a partir dos espaços em branco. O *Nominal featurizer* é responsável por atribuir *tags* para inflexão (gênero e número) e grau (diminutivo, superlativo e comparativo) para palavras de categorias morfossintáticas nominais. O *Nominal lemmatizer* é responsável pela tarefa de atribuir adjetivos e nomes comuns a uma forma normalizada (masculino singular). A ferramenta pode ser baixada livremente

[1], porém não tem código aberto.

4.2 FERRAMENTAS PARA NER/NED

FOX (*Federated knOwledge eXtraction framework*) (SPECK; NGOMO, 2014) é um *framework* de código aberto que implementa serviços Web RESTful para NER e NED. FOX usa AGDISTIS, Weka e uma rede Perceptron multicamadas (RUCK et al., 1990). AGDISTIS [2] (USBECK et al., 2014) é um *framework* de código aberto para desambiguação de palavras relevantes que consegue relacionar menções a entidades descritas em bases de conhecimento na forma de dados ligados e fazer a desambiguação usando a DBpedia (AUER et al., 2007). Weka (HALL et al., 2009) é um *framework* para mineração de dados. Os algoritmos nele disponíveis baseiam-se principalmente em aprendizado de máquina e podem ser aplicados diretamente a um conjunto de dados usando sua interface ou a partir de código Java. Weka contém ferramentas para pré-processamento de dados, classificação, regressão, *clustering*, regras de associação e visualização. FOX usa *ensemble learning* para combinar os resultados de algumas das melhores ferramentas atuais para NER e NED. Os autores afirmam que a ferramenta consegue atingir um percentual de medida-F de aproximadamente 95.23%. O *workflow* realizado pela ferramenta consiste em quatro passos: No primeiro passo, de pré-processamento, o usuário fornece como entrada uma URL, texto com tags HTML ou texto simples. Se a entrada for uma URL, FOX envia uma requisição à URL fornecida para receber os dados de entrada. Para todos os formatos de entrada, FOX remove as tags HTML e detecta frases e tokens. No segundo passo é realizado o NER, via combinação dos resultados de quatro outras ferramentas: Stanford NER, Illinois Named Entity Tagger, Ottawa Baseline Information Extraction e Apache OpenNLP Name Finder. Na terceira etapa é feita a ligação de entidades usando AGDISTIS. Na última etapa é realizada a serialização dos resultados. O usuário pode escolher dentre os seguintes formatos de saída: JSON-LD, N-Triplas, RDF/JSON, RDF/XML, Turtle, TriG e N-Quádruplas. É possível usar a ferramenta por meio de serviço Web[3] , bem como uma API (a qual possui bindings para Java e Python). FOX é de código aberto, o que possibilita fazer *fork* do seu *baseline* de software, disponível no Github[4].

[1] Disponível em http://lxcenter.di.fc.ul.pt/tools/pt/conteudo/LXTagger.html
[2] Disponível em https://github.com/AKSW/AGDISTIS
[3] Disponível em http://139.18.2.164:4444/demo/index.html#!/demo
[4] Código fonte disponível em https://github.com/AKSW/FOX

Stanford NER (*Stanford Named Entity Recognizer*) (FINKEL; GRENAGER; MANNING, 2005) é um *framework* de código aberto para NER. A ferramenta baseia-se em classificação de sequências e utiliza dez características locais para treinar um CRF. O framework foi implementado em Java pelo grupo de NLP de Stanford, e rotula nomes (de pessoas, empresas, genes e proteínas, entre outros) no texto. Pode-se usar o código da Stanford NER para construir modelos de sequência para reconhecimento de palavras relevantes ou qualquer outra tarefa. Também é possível usá-lo via interface gráfica, linha de comando, através de API *Web Service* ou cópia instalada localmente de maneira *standalone*. A distribuição atual permite o acesso a todos os recursos do pipeline do Stanford CoreNLP, acessados através da classe NERClassifierCombiner. Experimentos com o StanfordNER resultaram em 85,51% na medida-F ao utilizar o *dataset* da CoNLL[5] e 92,29% ao utilizar o *dataset* da CMU [6].

DBpedia Spotlight (MENDES et al., 2011) é uma ferramenta para reconhecimento e desambiguação de entidades nomeadas. A ferramenta usa uma abordagem de recuperação de informação, onde candidatos são gerados pela busca em um dicionário de nomes previamente computado no *DBPedia Lexicalizations dataset*[7]. Para realizar a desambiguação, é feita uma sobreposição entre o contexto textual da menção e de cada candidato (KLEIN, 2015). O DBpedia Spotlight permite muitas formas de uso e ajustes de diversas configurações internas para atender as necessidades do usuário.

T-NER (RITTER et al., 2011) é uma ferramenta capaz de alcançar medida-F de 59% ao classificar entidades nomeadas em três classes (pessoa, local e organização), enquanto o Stanford NER alcançou apenas 29% para tal tarefa usando o mesmo conjunto de dados. O *workflow* do T-NER consiste nas fases de segmentação e de classificação. Na fase de segmentação, cada sentença do texto é rotulada segundo a notação BIO (*Beginning, the Inside and Outside of text segments*) (RATINOV; ROTH, 2009). Um CRF é responsável pelo aprendizado e inferência a partir de características como a ortografia, o contexto textual e um dicionário de nomes e classes extraído do Freebase (BOLLACKER; COOK; TUFTS, 2007). Na fase de classificação, a ferramenta LabeledLDA é utilizada para modelar tópicos de acordo com 10 classes de entidades extraídas da Freebase (RAMAGE et al., 2009). O LabeledLDA é treinado com um dataset de 60 milhões de *tweets*.

[5]Disponível em http://cnts.uia.ac.be/conll2003/ner
[6]Disponível em http://nlp.shef.ac.uk/dot.kom/resources.html
[7]Disponível em http://wiki.dbpedia.org/lexicalizations

Illinois Named Entity Tagger (RATINOV; ROTH, 2009) é um *tagger* que marca texto simples com delimitações e classes de entidades nomeadas encontradas no texto de entrada. Atualmente as entidades detectadas por tal ferramenta podem se encaixar nas categorias pessoa, organização, localização e variados, ou ainda dezoito outros tipos, se a classificação for realizada com base no *corpus* OntoNotes. Ela usa *gazetteers* extraídos da Wikipedia, modelos de classes palavras derivadas de texto não rotulado e recursos não-locais expressivos.

NERD (RIZZO; TRONCY, 2012) é um *framework* que propõe unificar as saídas de dez diferentes extratores NLP publicamente disponíveis na Web. A abordagem baseia-se na ontologia NERD[8], que provê uma interface para anotação de elementos e uma API REST usada para unificar as saídas. NERD é uma aplicação web que funciona sobre várias ferramentas de NLP. Sua arquitetura segue os princípios REpresentational State Transfer (REST) e fornece um acesso HTML web para os usuários e uma API para que computadores possam fazer intercâmbio de conteúdo em formato JSON ou XML. Ambas as interfaces são suportadas pela *engine* NERD REST. A *engine* NERD REST e as ferramentas de NLP usadas pelo NERD fazem o reconhecimento e desambiguação de entidades nomeadas apontando URIs de bases de conhecimento para objetos do mundo real, como eles poderiam ser definidos na Web.

4.3 SUÍTES DE FERRAMENTAS

FreeLing (PADRÓ; STANILOVSKY, 2012) é uma biblioteca C++ (que também possui APIs para Java e Python, entre outras linguagens) que realiza diversas tarefas de NLP, como identificação da língua, tokenização, segmentação de sentenças, análise morfossintática, NER, NED, *PoS tagging*, *Shallow parsing* e resolução de correferência. Inclui dois módulos diferentes capazes de realizar *PoS tagging*. O primeiro é a classe tagger_hmm, baseada no método de (BRANTS, 2000), ou seja, usa o método de aprendizado baseado em transformações. O segundo módulo é o relax_tagger, um sistema híbrido capaz de integrar conhecimentos sobre estatística e códigos feitos a mão, segundo (PADRÓ, 1998). A aplicação pode instanciar qualquer um dos dois módulos.

GATE Twitter PoS Tagger (DERCZYNSKI et al., 2013) é um *tagger* desenvolvido especificamente para a anotação morfossintática de *tweets*. Para reduzir o impacto da dispersão de dados a ferra-

[8] Disponível em http://nerd.eurecom.fr

menta usa *vote-constrained bootstrapping* (uma variação de *bootstrapping*. Também são utilizados métodos para o tratamento de erros característicos de *tweets* e gírias. A ferramenta usa a abordagem de CRF e é destinada a *tweets* na língua inglesa e contém uma instância do Stanford Tagger[9]. A ferramenta tem várias opções de utilização (*plugin* do GATE, *Standalone*, etc) e também está inclusa em um pipeline para processamento de linguagem natural de código aberto customizado para textos de microblogs chamado TwitIE (BONTCHEVA et al., 2013).

OpenNLP[10] é um conjunto de ferramentas baseadas em aprendizagem de máquina visando o processamento de linguagem natural. Possui suporte para tarefas como tokenização, segmentação de sentença, *PoS tagging*, extração de entidades nomeadas, *parsing* e resolução de correferência. O *PoS-tagger* usa o modelo de máxima entropia marcando os *tokens* com o seu tipo de palavra correspondente com base no próprio *token* bem como no seu contexto. A ferramenta usa um modelo de probabilidade para prever a *tag* correta para o *token* a partir do conjunto de *tags*. No site do OpenNLP, existem modelos pré-treinados para dinamarquês, alemão, inglês, holandês, sueco e português. Porém, a ferramenta também permite o treinamento de modelos em outras línguas. Isso pode ser feito por meio de coleções de textos anotados com *tags*.

4.4 RESUMO COMPARATIVO DAS FERRAMENTAS

A Tabela 2 mostra uma comparação de ferramentas para NER/-NED, ao passo que a Tabela 3 mostra uma comparação de ferramentas para *PoS Tagging*. Note que tais ferramentas usam variadas abordagens, costumam oferecer diversas formas de uso, entradas em formato de texto ou HTML e algumas ferramentas NER/NED oferecem anotações semânticas com bases de dados proeminentes e saídas fornecidas em formatos conhecidos na área de Web semântica, além de texto. Todavia, se observa a carência de soluções para a língua portuguesa, principalmente para ferramentas NER/NED.

A foram levantadas ferramentas atuais e proeminentes. Foi observada que grande parte das abordagens envolve aprendizado de máquina, ou modelos para lidar com o problema de classificação de sequências (e.g., CRF, *Taggers* baseados em trigramas), o que indica que as técnicas

[9]Disponível em http://nlp.stanford.edu/software/tagger.shtml
[10]Disponível em http://opennlp.apache.org

determinísticas estão caindo em desuso por conta de qualidades inferiores dos resultados. Apesar do intuito de encontrar na literatura ferramentas que tenham suporte específicos para microblogs, como o Twitter, poucas ferramentas foram encontradas com essa funcionalidade, limitando-se a GATE Twitter PoS Tagger e T-NER. O mesmo ocorre para o suporte para a língua portuguesa, tornando evidente que a linguagem foco nessa área de pesquisa ainda é o inglês.

Ferramenta	Modo/ restrições de uso	Abordagem	Formato de Entrada	Formato de Saída	Suporte de línguas	Bases de Conhecimento
FOX	API/ open source	Ensemble Learning	txt, HTML, URI	RDF, JSON, TriG, N-tripla, N-Quad, Turtle	EN, FR, NL, DE	DBPedia
Stanford NER	Standalone, API, linha de comando/ open source	CRF	txt, HTML	XML, txt	EN, ZH, DE	Freebase
DBPedia Spotlight	Standalone, API, linha de comando/ open source	Gazeteers, dicionário de nomes	txt, HTML	XML, JSON, HTML, RDF, NIF	PT, EN ES, FR, IT, RU, DE, NL, HU, TR	DBPedia, Freebase, Schema.org
Illinois Named Entity Tagger	Standalone/ uso livre	Gazeteers, HMM	txt	txt	EN	DBPedia
NERD	API/open source	Ontologia NERD	txt, HTML, URI	JSON	EN, FR, DE, IT, PT, ES RU, SV	DBPedia
T-NER	linha de comando/ open source	Dicionário de nomes, CRF, bootstrapping	txt, HTML, URI	txt	EN	Freebase

Tabela 2 – Comparação de ferramentas para NER/NED.

Ferramenta	Modo/ restrições de uso	Abordagem	Formato de Entrada	Formato de Saída	Suporte de línguas
LX-Tagger	linha de comando/ uso livre	Trigramas	txt, HTML	txt	PT
OpenNLP	API, linha de comando/ open source	Máxima Entropia	txt	txt	PT, DA, DE, EN, ES, NL, SV
Freeling	API, linha de comando, código/ open source	Trigramas	txt	txt	EN, PT, ES, FR, DE, IT, RU, DA, CA, CR, CY, GA, SL, AS, NW
TreeTagger	linha de comando, GUI/ uso livre	Árvores de decisão	txt	tag	DE, EN, FR, IT, NL, ES, BG, RU, PT, GL, ZH, SW, SK, SL, LA, ET, PL, PT, EN
GATE Twitter PoS Tagger	Plugin, Standalone, linha de comando/open source	CRF, vote-constrained bootstraping	txt, HTML	txt	EN

Tabela 3 – Comparação de ferramentas para *PoS Tagging*.

5 EXPERIMENTOS

Este capítulo descreve os experimentos de classificação morfossintática realizados para a avaliação de ferramentas selecionadas sobre uma coleção de *tweets* e o *corpus* Tycho Brahe.

5.1 SELEÇÃO DAS FERRAMENTAS

As ferramentas selecionadas precisam ser capazes de processar eficientemente uma grande quantidade de texto, principalmente, em Língua Portuguesa, anotar morfossintaticamente as palavras e expressões, ter código fonte aberto ou permitir a utilização em pesquisa, permitir aplicação em mídias sociais, além da possibilidade de integração a outras aplicações. Usando esses critérios foram selecionadas para os experimentos: LX-Tagger, TreeTagger, Gate Twitter PoS Tagger, FreeLing e OpenNLP. Todas elas realizam *PoS Tagging*, ou seja, atribuem uma classe gramatical, na forma de *tag*, a cada palavra do texto de entrada.

5.2 CONJUNTOS DE DADOS

Primeiramente, é necessário esclarecer que, no contexto desse trabalho, considera-se *corpus* como sendo um conjunto de dados na forma de textos com anotações linguísticas previamente realizadas por um grupo de especialistas (por exemplo, anotação morfológica, sintática, lematização, entidades nomeadas, padrão-ouro). Caso contrário, podem-se considerar que o conjunto de dados é um *dataset* - uma amostra representativa de um fenômeno linguístico específico em um contexto restrito. Nas subseções seguintes são mostrados alguns pontos importantes de ambos.

Priorizou-se a análise dos resultados de *PoS Tagging* por este ser um componente fundamental na análise linguística, contribuindo para a recuperação de informação e a desambiguação de palavras, fazendo a análise sintática automática de frases em termos de suas funções gramaticais. Ademais, *PoS Tagging* não se restringe a análise de um grupo específico de palavras, tais como entidades nomeadas, pois anota todos *tokens*. Isso o torna uma tarefa com forte potencial de aplicação em pesquisas sobre RWR e enriquecimento semântico de dados desenvolvidas no LISA.

5.2.1 *Dataset* de *tweets*

Para realização dos experimentos com *tweets*, foram selecionados 100 mil *tweets* escritos em português e postados entre 30/11/2015 e 15/12/2015 no território brasileiro. Inicialmente pretendia-se fazer experimentos com todo o mês de dezembro, porém isso resultou em cerca de 67 milhões de *tweets*, ocupando um arquivo texto de 7,3 GB. Em virtude do volume de informação, foi tomada a decisão de selecionar aleatoriamente a amostra de 100 mil *tweets* dentro do período mencionado. A Tabela 4 apresenta estatísticas dos dados para os 100 mil *tweets* selecionados para experimentos:

Descrição	**Quantificação**
Tweets	100000
Tokens	1122561
Caracteres por palavra:	
Média	4,49
Desvio Padrão	3,18
Tamanho do arquivo (MB)	∼6

Tabela 4 – Estatística do dataset

A Figura 9 traz alguns exemplos de *tweets* do conjunto selecionado.

- **A folga do cara https://t.co/lxvzZfkmKK;**
- **@JackJackJohnson: São Paulo!!!!! O show foi incrível, mal posso esperar para voltar para um maior ainda;**
- **Paçoca é muito massa, é doce e salgada ao mesmo tempo;**
- **Dou um doce pra quem conseguir me provar que eu estou errada, e dessa vez não, não estou.**

Figura 9 – Exemplos de *tweets*.

Por exemplo, a frase "Cheguei ontem e apaguei", após passar pelo pelo processamento pode ser representada como *"Cheguei/V ontem/ADV e/CJ apaguei/V"*, com a *tag* V indicando verbo, ADV - advérbio e CJ - conjunção. Também filtrou-se de maneira simples o conteúdo dos *tweets*, com o intuito de retirar os *emoticons* e caracteres especiais que pudessem causar o mau funcionamento das ferramentas.

Dada a dificuldade de criação de um padrão ouro manualmente anotado para todo o *dataset*, uma amostra de 383 tweets do dataset foi selecionada aleatoriamente por sorteio, ao nível de confiança de 95%, para analisar a qualidade dos resultados. Essa amostra foi manualmente anotada conforme as seguintes categorias: adjetivo, advérbio, conjunção, determinante (adjetivos), interjeição, nome (substantivos), pontuação, preposição, pronome, sigla/símbolo/número/outros (chamada "SSN") e verbo. Na categoria SSN foram colocados também elementos como gírias, URIs e *hashtags*. Foi então realizada a normalização das saídas, substituindo as *tags* dos *outputs* das ferramentas por *tags* correspondentes às categorias mencionadas (os motivos dessa normalização serão apresentados na seção de Resultados e Discussão). A amostra de *tweets* apresenta 4.336 *tokens*, os quais verifica-se a distribuição das categorias morfossintáticas conforme a análise das ferramentas no Apêndice B, Tabela 8.

5.2.2 *Corpus* Tycho Brahe

O *corpus* Histórico do Português Tycho Brahe é um *corpus* eletrônico anotado, composto de textos em português escritos por autores nascidos entre 1380 e 1881. Atualmente, 73 textos (3.048.971 palavras) estão disponíveis para pesquisa livre. Destes, 40 possuem anotação morfológica, com um total de 1.728.284 palavras.

Tal *corpus* foi selecionado visando causar contraste em relação ao *dataset* de *tweets*, uma vez que esse *corpus* possui textos formais que seguem a gramática portuguesa rigidamente, ao passo que os *tweets* são altamente informais e contemporâneos.

Para o experimento com o *corpus* Tycho Brahe, primeiramente foi realizada uma análise para determinar a quantidade de sentenças que deveria compor uma amostra estratificada proporcional, com nível de confiança de 95%. Através disso, foi determinado que a amostra do *corpus* deveria conter 368 sentenças, presentes em 37 textos do *corpus*. Os 37 textos completos possuem 1.612.800 palavras, ao passo que a amostra possui 5.330 palavras. A seleção das sentenças dessa amostra

foi realizada por sorteio.

No site do *corpus* eletrônico também são disponibilizadas as versões dos textos morfossintaticamente anotados. Dessa forma, as sentenças morfossintaticamente anotadas correspondentes às selecionadas, foram usadas para compor um padrão ouro para esse experimento. A amostra do *corpus* apresenta 5.736 *tokens*, classificados em substantivos, verbos, adjetivos entre outras categorias morfossintáticas (ver a distribuição das categorias conforme a análise das ferramentas no Apêndice B, Tabela 9).

5.2.3 Padrão Ouro

A construção de um padrão ouro para *PoS Tagging* se dá pela anotação manual de um *dataset*, atribuindo a cada palavra a sua classificação morfossintática correta. A utilização do padrão ouro é imprescindível para o cálculo da precisão, cobertura e por conseguinte da medida-F, que depende das métricas anteriores.

A Tabela 5 mostra a quantidade de *tokens* manualmente anotados na amostra de 383 *tweets*, conforme a distribuição das categorias morfossintáticas. A frequência das classes gramaticais obedece a literatura, pois as classes com mais *tokens* são as de substantivos e verbos.

5.3 RESULTADOS E DISCUSSÃO

Nesta seção são apresentados e discutidos os resultados dos experimentos. Os resultados foram divididos segundo os seguintes aspectos: variabilidade das categorias morfossintáticas, análise da qualidade dos resultados e por fim, análise do tempo de execução das ferramentas.

5.3.1 Variabilidade de Categorias Morfossintáticas

As ferramentas receberam como entrada os textos dos *tweets* e geraram como saída um arquivo texto com cada palavra relevante neles encontrada, anotada com a sua categoria morfossintática. A anotação, denominada *tagset*, variou conforme a especificidade da classificação morfossintática da ferramenta. Por exemplo, a TreeTagger usa as seguintes notações para adjetivo: AO - adjetivo ordinal, AOS - adjetivo ordinal superlativo, AQ - adjetivo qualificativo, AQA - adjetivo qualificativo aumentativo, AQC - adjetivo qualificativo diminutivo, AQS -

Categoria	Quantidade
ADJ	169
ADV	323
CON	114
DET	234
INT	30
NOM	1127
PON	306
PRE	394
PRO	450
SSN	478
VER	787
Total	**4412**

Tabela 5 – *Tokens* do padrão ouro pelas ferramentas para cada categoria morfossintática.
ADJ=Adjetivo, ADV=Advérbio, CON=Conjunção, DET=Determinante,
INT=Interjeição, NOM=Nome, PON=Pontuação, PRE=Preposição, PRO = Pronome, SSN = Sigla/Símbolo/Número/Outros, VER = Verbo.

adjetivo qualificativo superlativo; enquanto a Gate Twitter PoS Tagger usa JJ - adjetivo, JJR - adjetivo comparativo, JJS - adjetivo superativo. Em razão de tais diferenças, subcategorias foram agrupadas em grandes categorias.

Os resultados obtidos estão descritos na tabela e figuras a seguir. A Tabela 6 mostra a quantidade de *tokens* encontrados pelas ferramentas, conforme a distribuição das categorias morfossintáticas. Note que mesmo o número total de *tokens* detectado foi diferente para cada ferramenta. Isso ocorre em função dos diferentes algoritmos adotados nas ferramentas para realizar os processos de tokenização, *chunking* e *PoS Tagging*. A FreeLing encontrou mais *tokens* que as demais porque possui funcionalidades para lidar com texto coloquial, e também por implementar vários módulos que identificam uma grande variedade de *tokens*, conseguindo enriquecer os *tweets* com novas informações linguísticas. Assim como a FreeLing, a TreeTagger faz lematização, gerando anotações mais detalhadas.

Categoria	**Ferramentas**				
	LX-Tagger	**TreeTagger**	**Gate Twitter PoS Tagger**	**Freling**	**OpenNLP**
ADJ	39174	60348	11671	52210	42667
ADV	34520	62672	1376	68227	49509
CON	37561	41654	2457	40475	38229
DET	104520	110358	34581	110062	37870
INT	2069	4167	54847	3739	0
NOM	136539	444741	211192	346059	147713
PON	63108	123546	0	210697	43044
PRE	44188	100019	23832	103938	51170
PRO	35420	77534	13178	109111	87629
SSN	9787	11003	85101	37557	10985
VER	139389	190160	42726	211282	105179
Total	**646275**	**1226202**	**480961**	**1293357**	**613995**

Tabela 6 – Número de *tokens* em *tweets* pelas ferramentas para cada categoria morfossintática.

Legenda ADJ=Adjetivo, ADV=Advérbio, CON=Conjunção, DET=Determinante,INT=Interjeição, NOM=Nome, PON=Pontuação, PRE=Preposição, PRO = Pronome, VER = Verbo, SSN = Sigla/Símbolo/Número/Outros.

A Figura 10 mostra as diferenças entre as médias de ocorrências das categorias morfossintáticas nos resultados das ferramentas. O gráfico do tipo *boxplot* apresentado na Figura 10 permite analisar a variabilidade dos resultados (diferença entre o terceiro e o primeiro quartil), a

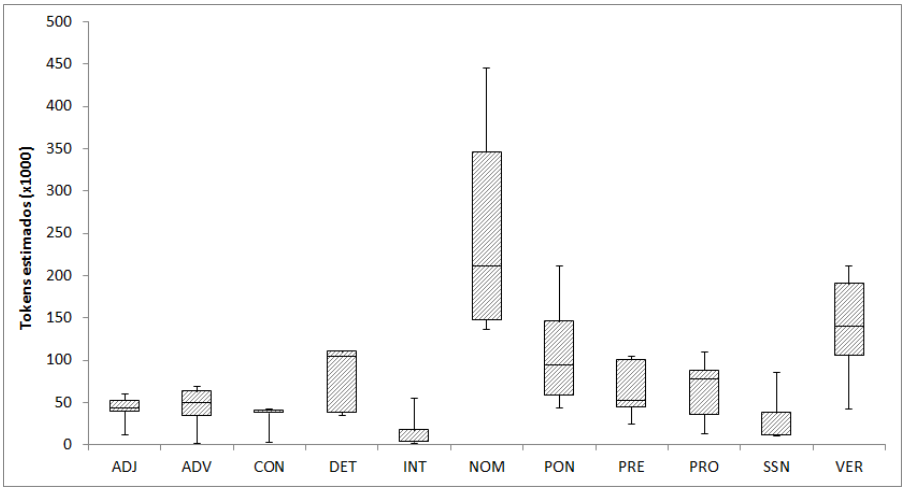

Figura 10 – Variabilidade das categorias morfossintáticas.

mediana (linha central) e os máximos e mínimos (calda inferior e superior). A ordem decrescente de frequência média das categorias (NOM, VER, PON, PRO, PRE, ADJ, SSN, DET, ADV, INT e COM) condiz com a literatura, onde se demonstra que substantivo e verbo são as categorias mais frequentes e importantes para análise linguística. A mediana das categorias em muitos casos está deslocada do centro, e a cauda muito longa ou curta, denotando falta de simetria. Isso está relacionado à grande diferença dos resultados das ferramentas para cada categoria. Observa-se que para as categorias NOM, PON, PRO, SSN e VER a variabilidade é alta. Incluir a lematização das palavras juntamente com a anotação morfossintática ajuda no enriquecimento semântico, mas deve-se ter o cuidado de verificar o contexto da aplicação. Se a quantidade de dados for grande, a inclusão de mais componentes linguísticos no processo de *PoS Tagging* pode acarretar em maior tempo de processamento.

Outra forma de verificar a variabilidade dos resultados é usada na Figura 11, que ilustra o coeficiente de variabilidade (desvio-padrão/ média) das categorias. Quanto menor tal coeficiente, mais homogêneo é o conjunto de dados. Ao observar-se a variabilidade do total de *tokens*, pode-se afirmar que as ferramentas apresentam um desempenho de processamento semelhante. Porém, quando o foco volta-se para as categorias, todos os coeficientes foram maiores que 25%, valor esse

considerado limite para afirmar estatisticamente que não existe homogeneidade nas ferramentas quanto a classificação.

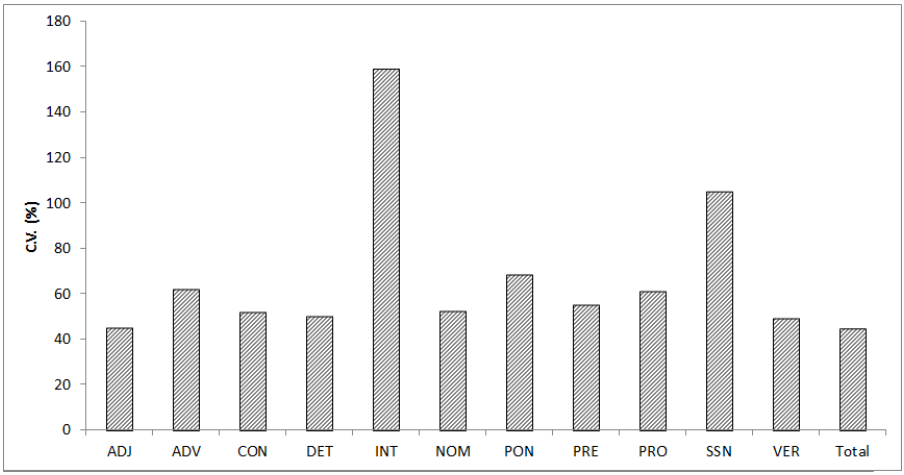

Figura 11 – Coeficiente de variabilidade (CV).

5.3.2 Análise da Qualidade dos Resultados

A avaliação de precisão, cobertura e medida-F foi realizada usando a ferramenta ROUGE (LIN, 2004) com intervalo de confiança de 95%. Para fazer essa análise, os arquivos de saída das ferramentas já normalizados foram alterados da forma "palavra_tag", deixando somente as *tags*. Isso foi necessário pois a ferramenta realiza a avaliação por unigramas. Logo, retirando as palavras, não corre-se o risco de que a ferramenta interprete incorretamente o que é um unigrama, fornecendo maior confiabilidade ao teste. A Figura 15 mostra como fica o arquivo após essa alteração, usando um *tweet* da amostra como exemplo.

A Figura 13 mostra os resultados obtidos de precisão, cobertura e medida-F para o *dataset* de *tweets* e a amostra do *corpus* Tycho Brahe. Como pode ser observado, para o *dataset* de *tweets* a ferramenta OpenNLP obteve o melhor resultado, enquanto as ferramentas LX-Tagger, Freeling e TreeTagger obtiveram resultados semelhantes, ao passo que a ferramenta GATE Twitter PoS Tagger obteve o pior resultado. Apesar de ser uma ferramenta dedicada especialmente para o processamento de *tweets*, devido à falta de suporte para a língua por-

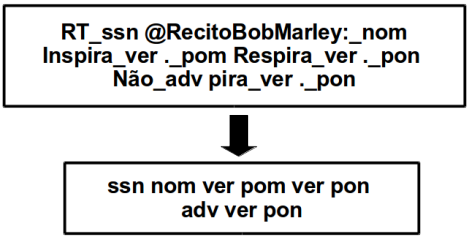

Figura 12 – Exemplo de adequação do arquivo.

Ferramenta	Dataset de Tweets			Amostra do Corpus		
	Precisão	Cobertura	Medida-F	Precisão	Cobertura	Medida-F
LX-Tagger	0.84585	0.78782	0.81580	0.83043	0.71809	0.77019
Freeling	0.80573	0.87251	0.83779	**0.92938**	**0.95848**	**0.94371**
TreeTagger	0.76705	0.82020	0.79274	0.91095	0.95912	0.93441
OpenNLP	**0.92029**	**0.85756**	**0.88782**	0.95385	0.82482	0.88465
GATE Twitter PoS Tagger	0.52059	0.46082	0.48888	0.44195	0.38037	0.40885

Figura 13 – Resultados de precisão, cobertura e medida-F para as amostras.

Figura 14 – Gráfico comparativo dos valores para medida-F.

tuguesa os resultados da ferramenta GATE Twitter PoS Tagger não conseguiu alcançar as outras.

Durante a análise dos resultados, um ponto importante notado foi a disparidade da tokenização e dos *tagsets* usados pelas ferramentas. Por exemplo, a TreeTagger e Freeling consideram a pontuação (e.g., ponto final, vígula, interrogação) como um *token* único, ao passo que OpenNLP, LX Tagger e Gate Twitter PoS Tagger consideram a pontuação junto com a palavra imediatamente anterior um *token*, na maioria do arquivo de saída. Há ferramentas que consideraram uma URI ou menção a usuário (e.g., @fulano_de_tal) como *token*, enquanto outras separam essas estruturas em vários *tokens* (e.g., @, fulano, _de, _tal). Os *tagsets* usados pelas ferramentas não possuem nenhuma padronização (com exceção da TreeTagger e Freeling, que são similares), o que torna o processo de comparação mais difícil, portanto foi adotada a estratégia de agrupamentos em classes maiores, conforme mencionado anteriormente.

Como esperado, houve uma melhora nos resultados obtidos com a amostra do *corpus* Tycho Brahe, em comparação com o *dataset* de *tweets*, como pode ser observado. Apesar da melhora dos resultados, levanta-se a dúvida de qual é o impacto estatístico dos problemas inerentes dos textos de microblogs nos resultados em termos de precisão e cobertura, salvo a necessidade de existirem categorias específicas para termos próprios do Twitter (e.g., *hashtags*, menções a usuário, URI, *retweet*).

Mesmo que os resultados apresentados para o *corpus* Tycho Brahe tenham sido melhores do que os obtidos com o *dataset* de *tweets*, ainda não foram alcançados resultados próximos aos apresentados na literatura (que em média beiram a 99% de precisão e cobertura para outros *corpus* existentes). Apesar das ferramentas utilizadas terem apresentado alguns bons resultados neste estudo, há evidências de que o problema de *PoS Tagging* ainda possui espaço para melhoras, principalmente no âmbito de mídias sociais. Além disso, em uma breve análise verificou-se palavras anotadas com categoria incorreta. Uma solução para os problemas constatados nos resultados das ferramentas seria considerar tarefas que melhoram a qualidade dos dados na etapa de pré-processamento, como a normalização de abreviações e acrônimos principalmente (e.g., sr^{a} seria traduzido para senhora), ou ainda, subdividir o conjunto de dados em um *corpus* de teste e treinamento e repetir os experimentos. A normalização desta linguagem também pode ser crucial para facilitar o processamento de texto e melhorar o desempenho de ferramentas de NLP aplicadas a mídias sociais. Contudo,

dependendo do propósito da aplicação, a normalização torna-se difícil, por exemplo, uma URI pode ser considerada como um *token* ou mesmo como vários (quando utiliza-se palavras com sentido próprio na composição).

Para o caso de microblogs, como o Twitter, vale ainda mais o uso de fonetização para tentar corrigir certos erros gramaticais (e.g., ksa, eh). Também, o uso de recursos extras para a identificação de gírias (e.g., sdds é uma gíria que significa saudades) poderia melhorar a qualidade dos resultados, porém seria necessário analisar o impacto do tempo adicional de processamento que pode ser acarretado pela adição desses recursos de pré-processamento, uma vez que métodos de NLP são computacionalmente pesados.

5.3.3 Análise do Tempo de Execução

Durante os experimentos, também foi medido o tempo de execução das ferramentas. A Tabela 7 mostra a média aritmética de cinco execuções de cada ferramenta para a mesma amostra, onde "Amostra 1" corresponde ao *dataset* com 100 mil *tweets*, "Amostra 2" à amostra desse *dataset*, com 383 *tweets* e por fim, "Amostra 3" à amostra do *corpus* Tycho Brahe com 368 sentenças e 5.330 palavras. Os experimentos, bem como a medição dos tempos foi realizada usando um notebook com processador Intel i5 com sistema operacional Ubuntu 14.04 e todas as ferramentas tiveram suas rotinas de anotação invocadas via terminal. Pode-se observar, que os tempos de execução mantiveram-se na ordem de poucos minutos ou segundos, exceto pelo tempo obtido pela GATE Twitter PoS Tagger para a "Amostra 1", que foi discrepantemente maior que todos os outros. Com isso, pode concluir-se que caso o conjunto de dados possuísse milhões de sentenças, a performance das ferramentas poderia ser um grande gargalo em certos casos, sugerindo a necessidade de paralelização para casos onde o volume de dados é muito grande.

Ferramenta	Amostra 1	Amostra 2	Amostra 3
LX-Tagger	172.54s	2.05s	2.22s
Freeling	138.95s	3.24s	3.29s
TreeTagger	10.64s	0.58s	0.58s
OpenNLP	26.53s	0.69s	0.73s
GATE Twitter PoS Tagger	14340,12s	56.42s	157.43s

Tabela 7 – Média dos tempos de execução das ferramentas.

6 CONCLUSÃO E TRABALHOS FUTUROS

Atualmente, apesar da maior disponibilidade de ferramentas para processamento de texto, muitos problemas ainda persistem: baixa interoperabilidade (principalmente entre ferramentas comerciais); limitações no compartilhamento de recursos (em função da dependência de plataformas específicas para execução); falta de padronização; disponibilização apenas de versões online para teste e por vezes, restrições de capacidade de processamento ao lidar com grandes volumes de dados. Outro ponto crítico é a escassez de recursos apropriados para processar textos naturais de mídias sociais, uma vez que a única ferramenta específica para essa finalidade encontrada foi a GATE Twitter PoS Tagger. Também faltam recursos para o processamento de textos em certos idiomas, como é o caso da Língua Portuguesa.

O estudo do estado da arte mostra que propostas recentes tendem ao uso de técnicas baseadas em aprendizado de máquina, pois estas começam a gerar melhores resultados e proveem a possibilidade do treinamento de um modelo específico.

Este trabalho apresentou uma revisão bibliográfica sobre as tarefas de Reconhecimento de Palavras Relevantes, com foco em anotação morfossintática. Foram apresentadas técnicas e ferramentas consistentes com o estado-da-arte de PoS Tagging. Os experimentos com as ferramentas selecionadas demonstraram que há grande variabilidade na qualidade dos resultados quando o texto é oriundo de mídias sociais. A análise comparativa aqui apresentada de ferramentas para o RWR em textos apresenta as seguintes contribuições:

1. Cobertura de uma variedade de subproblemas, métodos e ferramentas de áreas de pesquisa tradicionalmente separadas;

2. Foco na análise de postagens em mídias sociais usando a língua portuguesa;

3. Resultados experimentais mais abrangentes, analisando as disparidades apresentadas por ferramentas distintas;

4. Evidência experimental de que os resultados de PoS Tagging obtidos com textos oriundos de mídias sociais são piores;

5. Análise do desempenho das ferramentas através da medição do tempo de execução da tarefa de Pos Tagging;

6. Construção de um padrão ouro para *tweets* em português.

Em pesquisas futuras pretende-se:

1. Realizar um teste estatístico para compreender até que ponto a fonte de dados impacta na qualidade dos resultados das ferramentas;
2. Pesquisar métodos de processamento que possam melhorar a qualidade dos resultados;
3. Adicionar recursos para a identificação de gírias e acrônimos nos textos de microblogs;
4. Ligar palavras relevantes a recursos de bases léxicas (e.g., WordNet, VerbNet);
5. Explorar o uso de anotação morfossintática aliada a ontologias, vocabulários controlados e *entity linking* para extração de informações.

REFERÊNCIAS

ALBERT, R.; JEONG, H.; BARABÁSI, A.-L. Internet: Diameter of the world-wide web. **Nature**, Nature Publishing Group, v. 401, n. 6749, p. 130–131, 1999.

ALUÍSIO, S. et al. An account of the challenge of tagging a reference corpus for brazilian portuguese. In: SPRINGER. **International Workshop on Computational Processing of the Portuguese Language**. [S.l.], 2003. p. 110–117.

ANANTHARAM, P. et al. Extracting city traffic events from social streams. **ACM Transactions on Intelligent Systems and Technology (TIST)**, ACM, v. 6, n. 4, p. 43, 2015.

ARBIB, M. A. **The handbook of brain theory and neural networks**. [S.l.]: MIT press, 2003.

AUER, S. et al. Dbpedia: A nucleus for a web of open data. In: **The semantic web**. [S.l.]: Springer, 2007. p. 722–735.

BERRY, M. J.; LINOFF, G. **Data mining techniques: for marketing, sales, and customer support**. [S.l.]: John Wiley & Sons, Inc., 1997.

BIGI, B.; HIRST, D. Speech phonetization alignment and syllabification (sppas): a tool for the automatic analysis of speech prosody. In: **Speech Prosody**. [S.l.: s.n.], 2012. p. 1–4.

BIZER, C.; HEATH, T.; BERNERS-LEE, T. Linked data-the story so far. **Semantic Services, Interoperability and Web Applications: Emerging Concepts**, p. 205–227, 2009.

BOLLACKER, K.; COOK, R.; TUFTS, P. Freebase: A shared database of structured general human knowledge. In: **AAAI**. [S.l.: s.n.], 2007. v. 7, p. 1962–1963.

BONTCHEVA, K. et al. Twitie: An open-source information extraction pipeline for microblog text. In: **RANLP**. [S.l.: s.n.], 2013. p. 83–90.

BOWMAN, M.; DEBRAY, S. K.; PETERSON, L. L. Reasoning about naming systems. **ACM Transactions on Programming**

Languages and Systems (TOPLAS), ACM, v. 15, n. 5, p. 795–825, 1993.

BRANCO, A.; SILVA, J. Evaluating solutions for the rapid development of state-of-the-art pos taggers for portuguese. In: **LREC**. [S.l.: s.n.], 2004.

BRANCO, A.; SILVA, J. R. A suite of shallow processing tools for portuguese: Lx-suite. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. **Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics: Posters & Demonstrations**. [S.l.], 2006. p. 179–182.

BRANTS, T. Tnt: a statistical part-of-speech tagger. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. **Proceedings of the sixth conference on Applied natural language processing**. [S.l.], 2000. p. 224–231.

BRILL, E. Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. **Computational linguistics**, MIT Press, v. 21, n. 4, p. 543–565, 1995.

BÜHLMANN, P.; WYNER, A. J. et al. Variable length markov chains. **The Annals of Statistics**, Institute of Mathematical Statistics, v. 27, n. 2, p. 480–513, 1999.

BUNESCU, R. C.; PASCA, M. Using encyclopedic knowledge for named entity disambiguation. In: **EACL**. [S.l.: s.n.], 2006. v. 6, p. 9–16.

CHOWDHURY, G. G. Natural language processing. **Annual review of information science and technology**, Wiley Online Library, v. 37, n. 1, p. 51–89, 2003.

CUCERZAN, S. Large-scale named entity disambiguation based on wikipedia data. In: **EMNLP-CoNLL**. [S.l.: s.n.], 2007. v. 7, p. 708–716.

CUTTING, D. et al. A practical part-of-speech tagger. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. **Proceedings of the third conference on Applied natural language processing**. [S.l.], 1992. p. 133–140.

DAELEMANS, W. et al. Mbt: A memory-based part of speech tagger-generator. **arXiv preprint cmp-lg/9607012**, 1996.

DANIEL, J.; JAMES, H. Speech and language processing: An introduction to natural language processing. **Computational Linguistics and Speech Recognition, 2nd Ed., Prentice Hall**, 2009.

DAS, T.; ACHARJYA, D.; PATRA, M. Opinion mining about a product by analyzing public tweets in twitter. In: IEEE. **Computer Communication and Informatics (ICCCI), 2014 International Conference on**. [S.l.], 2014. p. 1–4.

DERCZYNSKI, L. et al. Twitter part-of-speech tagging for all: Overcoming sparse and noisy data. In: **RANLP**. [S.l.: s.n.], 2013. p. 198–206.

DODDINGTON, G. R. et al. The automatic content extraction (ace) program-tasks, data, and evaluation. In: **LREC**. [S.l.: s.n.], 2004. v. 2, p. 1.

DOWNEY, D.; BROADHEAD, M.; ETZIONI, O. Locating complex named entities in web text. In: **IJCAI**. [S.l.: s.n.], 2007. v. 7, p. 2733–2739.

EDDY, S. R. Hidden markov models. **Current opinion in structural biology**, Elsevier, v. 6, n. 3, p. 361–365, 1996.

FILETO, R. et al. The baquara 2 knowledge-based framework for semantic enrichment and analysis of movement data. **Data & Knowledge Engineering**, Elsevier, v. 98, p. 104–122, 2015.

FINKEL, J. R.; GRENAGER, T.; MANNING, C. Incorporating non-local information into information extraction systems by gibbs sampling. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. **Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics**. [S.l.], 2005. p. 363–370.

FORNEY, G. D. The viterbi algorithm. **Proceedings of the IEEE**, IEEE, v. 61, n. 3, p. 268–278, 1973.

GIMPEL, K. et al. Part-of-speech tagging for twitter: Annotation, features, and experiments. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. **Proceedings of the 49th**

Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2. [S.l.], 2011. p. 42–47.

GREENE, B. B.; RUBIN, G. M. **Automatic grammatical tagging of English**. [S.l.]: Department of Linguistics, Brown University, 1971.

HABIB, M. B.; KEULEN, M. van. Information extraction for social media. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. [S.l.], 2014.

HALL, M. et al. The weka data mining software: an update. **ACM SIGKDD explorations newsletter**, ACM, v. 11, n. 1, p. 10–18, 2009.

HARRIS, Z. String analysis of language structure. **Mouton and Co., The Hague**, 1962.

HARRIS, Z. S. Co-occurrence and transformation in linguistic structure. **Language**, JSTOR, v. 33, n. 3, p. 283–340, 1957.

HASAN, F. M.; UZZAMAN, N.; KHAN, M. Comparison of different pos tagging techniques (n-gram, hmm and brillâ€™s tagger) for bangla. In: **Advances and Innovations in Systems, Computing Sciences and Software Engineering**. [S.l.]: Springer, 2007. p. 121–126.

HEARST, M. A. et al. Support vector machines. **IEEE Intelligent Systems and their Applications**, IEEE, v. 13, n. 4, p. 18–28, 1998.

HENDLER, J. Web 3.0 emerging. **Computer**, IEEE, v. 42, n. 1, p. 111–113, 2009.

JABEEN, S.; SHAH, S.; LATIF, A. Named entity recognition and normalization in tweets towards text summarization. In: IEEE. **Digital Information Management (ICDIM), 2013 Eighth International Conference on**. [S.l.], 2013. p. 223–227.

KLEIN, D. Estudo de técnicas e ferramentas aplicáveis a mídias sociais para reconhecimento e desambiguação de entidades nomeadas. Universidade Federal de Santa Catarina, p. 9, 2015.

KLEIN, S.; SIMMONS, R. F. A computational approach to grammatical coding of english words. **Journal of the ACM (JACM)**, ACM, v. 10, n. 3, p. 334–347, 1963.

LABOREIRO, G. et al. Tokenizing micro-blogging messages using a text classification approach. In: ACM. **Proceedings of the fourth workshop on Analytics for noisy unstructured text data**. [S.l.], 2010. p. 81–88.

LAFFERTY, J.; MCCALLUM, A.; PEREIRA, F. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: **Proceedings of the eighteenth international conference on machine learning, ICML**. [S.l.: s.n.], 2001. v. 1, p. 282–289.

LEV, B.; THIAGARAJAN, S. R. Fundamental information analysis. **Journal of Accounting research**, JSTOR, p. 190–215, 1993.

LIN, C.-Y. Rouge: A package for automatic evaluation of summaries. In: BARCELONA, SPAIN. **Text summarization branches out: Proceedings of the ACL-04 workshop**. [S.l.], 2004. v. 8.

MARTIN, J. H.; JURAFSKY, D. Speech and language processing. **International Edition**, v. 710, 2000.

MEGYESI, B. et al. Comparing data-driven learning algorithms for pos tagging of swedish. Citeseer, 2001.

MENDES, P. N. et al. Dbpedia spotlight: shedding light on the web of documents. In: ACM. **Proceedings of the 7th international conference on semantic systems**. [S.l.], 2011. p. 1–8.

MUNOZ, M. et al. A learning approach to shallow parsing. **arXiv preprint cs/0008022**, 2000.

PADRÓ, L. A hybrid environment for syntax-semantic tagging. **arXiv preprint cmp-lg/9802002**, 1998.

PADRÓ, L.; STANILOVSKY, E. Freeling 3.0: Towards wider multilinguality. In: **LREC2012**. [S.l.: s.n.], 2012.

PAZZANI, M. J.; BILLSUS, D. Content-based recommendation systems. In: **The adaptive web**. [S.l.]: Springer, 2007. p. 325–341.

QUINLAN, J. R. Induction of decision trees. **Machine learning**, Springer, v. 1, n. 1, p. 81–106, 1986.

RABINER, L.; JUANG, B. An introduction to hidden markov models. **ieee assp magazine**, IEEE, v. 3, n. 1, p. 4–16, 1986.

RAMAGE, D. et al. Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. **Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1**. [S.l.], 2009. p. 248–256.

RATINOV, L.; ROTH, D. Design challenges and misconceptions in named entity recognition. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. **Proceedings of the Thirteenth Conference on Computational Natural Language Learning**. [S.l.], 2009. p. 147–155.

RATNAPARKHI, A. et al. A maximum entropy model for part-of-speech tagging. In: PHILADELPHIA, USA. **Proceedings of the conference on empirical methods in natural language processing**. [S.l.], 1996. v. 1, p. 133–142.

RICCIA, G. D.; KRUSE, R.; LENZ, H.-J. **Computational intelligence in data mining**. [S.l.]: Springer, 2014.

RITTER, A. et al. Named entity recognition in tweets: an experimental study. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. **Proceedings of the Conference on Empirical Methods in Natural Language Processing**. [S.l.], 2011. p. 1524–1534.

RIZZO, G.; TRONCY, R. Nerd: a framework for unifying named entity recognition and disambiguation extraction tools. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. **Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics**. [S.l.], 2012. p. 73–76.

RUCK, D. W. et al. The multilayer perceptron as an approximation to a bayes optimal discriminant function. **IEEE Transactions on Neural Networks**, IEEE, v. 1, n. 4, p. 296–298, 1990.

SACENTI, J. A. et al. Automatically tailoring semantics-enabled dimensions for movement data warehouses. In: SPRINGER. **International Conference on Big Data Analytics and Knowledge Discovery**. [S.l.], 2015. p. 205–216.

SANG, E. F. T. K.; BUCHHOLZ, S. Introduction to the conll-2000 shared task: Chunking. In: ASSOCIATION FOR

COMPUTATIONAL LINGUISTICS. **Proceedings of the 2nd workshop on Learning language in logic and the 4th conference on Computational natural language learning-Volume 7**. [S.l.], 2000. p. 127–132.

SANG, E. F. T. K.; MEULDER, F. D. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. **Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4**. [S.l.], 2003. p. 142–147.

SANTOS, C. N. D.; MILIDIÚ, R. L.; RENTERÍA, R. P. Portuguese part-of-speech tagging using entropy guided transformation learning. In: SPRINGER. **International Conference on Computational Processing of the Portuguese Language**. [S.l.], 2008. p. 143–152.

SCHMID, H. Probabilistic part-ofispeech tagging using decision trees. In: ROUTLEDGE. **New methods in language processing**. [S.l.], 2013. p. 154.

SHA, F.; PEREIRA, F. Shallow parsing with conditional random fields. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. **Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1**. [S.l.], 2003. p. 134–141.

SPECK, R.; NGOMO, A.-C. N. Named entity recognition using fox. In: CEUR-WS. ORG. **Proceedings of the 2014 International Conference on Posters & Demonstrations Track-Volume 1272**. [S.l.], 2014. p. 85–88.

TOMAN, M.; TESAR, R.; JEZEK, K. Influence of word normalization on text classification. **Proceedings of InSciT**, v. 4, p. 354–358, 2006.

TOUTANOVA, K. et al. Feature-rich part-of-speech tagging with a cyclic dependency network. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. **Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1**. [S.l.], 2003. p. 173–180.

TUFIS, D.; MASON, O. Tagging romanian texts: a case study for qtag, a language independent probabilistic tagger. In: **Proceedings of the First International Conference on Language Resources and Evaluation (LREC)**. [S.l.: s.n.], 1998. v. 1, p. 589–596.

USBECK, R. et al. Agdistis-graph-based disambiguation of named entities using linked data. In: SPRINGER. **International Semantic Web Conference**. [S.l.], 2014. p. 457–471.

VOORHEES, E. M. et al. The trec-8 question answering track report. In: **Trec**. [S.l.: s.n.], 1999. v. 99, p. 77–82.

VOUTILAINEN, A. Part-of-speech tagging. **The Oxford handbook of computational linguistics**, Oxford University Press, p. 219–232, 2003.

WANG, W.; STEWART, K. Spatiotemporal and semantic information extraction from web news reports about natural hazards. **Computers, Environment and Urban Systems**, Elsevier, v. 50, p. 30–40, 2015.

WEBSTER, J. J.; KIT, C. Tokenization as the initial phase in nlp. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. **Proceedings of the 14th conference on Computational linguistics-Volume 4**. [S.l.], 1992. p. 1106–1110.

XIA, C. et al. What is new in our city? a framework for event extraction using social media posts. In: SPRINGER. **Pacific-Asia Conference on Knowledge Discovery and Data Mining**. [S.l.], 2015. p. 16–32.

APÊNDICE A – Fórmulas de Precisão, Cobertura e Medida-F

Nesse apêndice são mostradas as fórmulas clássicas da literatura para precisão, cobertura e medida-F, retiradas do livro (RICCIA; KRUSE; LENZ, 2014), aonde P, C, TR, TO significam respectivamente precisão, cobertura, *tag* real e *tag* obtida. A *tag* real é a correta para uma dada palavra (e.g., a palavra "andar" deveria receber a *tag* "verbo") e a *tag* obtida é a que foi atribuída pela ferramenta no processo de *PoS Tagging*.

$$\boldsymbol{P = \frac{TR \cap TO}{TO}} \tag{A.1}$$

$$\boldsymbol{C = \frac{TR \cap TO}{RT}} \tag{A.2}$$

$$\boldsymbol{medidaF = \frac{2 * (P * C)}{(P + C)}} \tag{A.3}$$

APÊNDICE B – Tabelas de *tokens* por Distribuição das Categorias Morfossintáticas das Amostras

A Tabela 8 mostra a quantidade de *tokens* encontrados pelas ferramentas nessa amostra de 383 *tweets*, conforme a distribuição das categorias morfossintáticas.

Categoria	**Ferramentas**				
	LX-Tagger	**TreeTagger**	**Gate Twitter PoS Tagger**	**Freeling**	**OpenNLP**
ADJ	222	204	54	194	235
ADV	176	233	6	258	298
CON	206	159	5	153	163
DET	502	451	166	474	237
INT	9	18	681	17	1
NOM	1269	1747	2116	1285	1239
PON	345	462	0	773	262
PRE	335	421	1	387	318
PRO	216	280	1	309	422
SSN	6	45	711	163	55
VER	827	702	167	779	882
Total	**4113**	**4722**	**3908**	**4792**	**4112**

Tabela 8 – Estimativa de *tokens* da amostra do *dataset* pelas ferramentas para cada categoria morfossintática.

[a] ADJ=Adjetivo, ADV=Advérbio, CON=Conjunção, DET=Determinante, INT=Interjeição, NOM=Nome, PON=Pontuação, PRE=Preposição, PRO = Pronome, SSN = Sigla/Símbolo/Número/Outros, VER = Verbo.

A Tabela 9 mostra a quantidade de *tokens* encontrados pelas ferramentas nessa amostra de 368 sentenças, conforme a distribuição das categorias morfossintáticas.

Categoria	Ferramentas				
	LX-Tagger	**TreeTagger**	**Gate Twitter PoS Tagger**	**Freeling**	**OpenNLP**
ADJ	330	274	98	273	307
ADV	236	348	0	361	401
CON	413	403	0	383	366
DET	915	908	297	911	443
INT	20	4	654	4	1
NOM	1385	1432	3596	1292	1310
PON	215	855	0	859	139
PRE	736	782	255	740	577
PRO	222	500	88	536	703
SSN	5	36	17	34	37
VER	854	949	294	965	1047
Total	**5331**	**6491**	**5299**	**6358**	**5331**

Tabela 9 – Estimativa de *tokens* da amostra do *corpus* pelas ferramentas para cada categoria morfossintática.

[a] ADJ=Adjetivo, ADV=Advérbio, CON=Conjunção, DET=Determinante, INT=Interjeição, NOM=Nome, PON=Pontuação, PRE=Preposição, PRO = Pronome, SSN = Sigla/Símbolo/Número/Outros, VER = Verbo.

APÊNDICE C – Artigo

Seleção e Avaliação Experimental de Ferramentas para Anotação Morfossintática Automática

Danielly Sorato[1], Renato Fileto[1], Fábio B. Goulart[1]

[1]Departamento de Informática e Estatística
Universidade Federal de Santa Catarina (UFSC) – Florianópolis, SC – Brasil

danielly.sorato@grad.ufsc.br, r.fileto@ufsc.br,
fabio.bif@posgrad.ufsc.br

Abstract. *Documents available on the web and posts on social media are abundant sources of information. In these texts is possible to find semantically rich components called relevant words. Currently, there is a wide variety of tools for relevant words recognition in texts. However, the performance and quality of results produced by these tools tend to be degraded when the text used is from social media. This happens because the social media text has informal content, spelling and grammatical errors, acronyms, slang, etc. This paper presents a literature review of techniques and tools for the extraction of relevant words of text and an experimental analysis tools for automatic morphosyntactic annotation, focusing on social media, especially microblogs like Twitter.*

Resumo. *Documentos disponíveis na Web e postagens em mídias sociais são fontes abundantes de informações. Nesses textos pode-se encontrar componentes semanticamente ricos denominados palavras relevantes. Atualmente, existe uma grande variedade de ferramentas para reconhecimento de palavras relevantes em textos. Porém, o desempenho e a qualidade dos resultados produzidos por estas ferramentas costumam ser degradados quando o texto usado é oriundo de mídias sociais. Isso acontece porque o texto de mídias sociais apresenta conteúdo informal, possuindo erros ortográficos e gramaticais, acrônimos, gírias, etc. Este trabalho apresenta uma revisão da literatura sobre técnicas e ferramentas para a extração de palavras relevantes de textos e uma análise experimental de ferramentas para anotação morfossintática automática, com foco em mídias sociais, especialmente microblogs, como o Twitter.*

1. Introdução

A quantidade de dados disponíveis em meio digital no mundo atual vêm crescendo exponencialmente (HABIB; KEULEN, 2014). A World Wide Web e mídias sociais (e.g. Facebook, Twitter, Instagram) têm contribuído de maneira significativa para este crescimento. Por exemplo, milhões de *tweets* são postados diariamente. Tais dados podem conter componentes relevantes e semanticamente ricos, tais como entidades nomeadas (menções a pessoas, locais, organizações, datas, etc.) (SANG; MEULDER , 2003), além de palavras com classificação gramatical (substantivos, adjetivos, verbos, etc.) que podem carregar semântica relevante (MARTIN; JURAFSKY , 2000).

Tais componentes linguísticos podem auxiliar no entendimento do que se passa com os usuários que fazem as postagens, pois permitem identicar eventos ocorridos ao longo do espaço e do tempo (WANG; STEWART, 2015; ANANTHARAM et al., 2015; XIA et al., 2015), locais frequentemente visitados (FILETO et al., 2015) ou até alimentar com informação precisa métodos de análise de informação (LEV; THIAGARAJAN, 1993; SACENTI et al., 2015), mineração de dados (BERRY; LINOFF, 1997), análise de sentimento (DAS; ACHARJYA; PATRA, 2014) e recomendação (PAZZANI; BILLSUS , 2007), entre outras possibilidades.

Dada a grande quantidade de aplicações para dados textuais disponíveis na Web, surge a necessidade de extração de informações (*Information Extraction* - IE) de textos. Porém, dados da Web e particularmente postagens em mídias sociais apresentam um novo e desafiador estilo de texto para as tecnologias de processamento de linguagem natural (*Natural Language Processing* - NLP) (CHOWDHURY, 2003). Diversas aplicações que usam dados de mídias sociais precisam analisar eficientemente a sintaxe e a semântica de grandes quantidades de dados gerados constantemente, cujos padrões linguísticos podem ser dinâmicos, com natureza informal e muitos ruídos, tais como erros ortográficos e gramaticais, presença de acrônimos, gírias, etc. Tais fatores tornam o processo de extração e classificação de informações de mídias sociais particularmente complexo, sendo difícil obter bons resultados por meio de métodos e ferramentas de NLP e alternativas clássicas.

A exploração do conteúdo de textos em mídias sociais, bem como as ferramentas para a extração de informações significativa destes é uma área emergente no cenário de tecnologia e pesquisa, que ainda está em desenvolvimento. Nos últimos anos, uma série de ferramentas surgiram combinando NLP e Dados Ligados (*Linked Data* - LD) (BIZER; HEATH; BERNERS-LEE, 2009) para reconhecimento e desambiguação de entidades nomeadas, assim como para a realização de anotação morfossintática automática.

A análise de palavras relevantes presentes nos textos de conjuntos de dados tão extensos torna-se inviável sem o uso de tais ferramentas. Além disso a baixa incidência de entidades nomeadas em certos textos da Web, principalmente em mídias sociais, sugere a necessidade de classificação morfossintática para então tentar associar as palavras classificadas com recursos léxicos. Para tanto, é necessário que estas produzam resultados de boa qualidade, com alta precisão e cobertura. Por conseguinte, o foco deste trabalho se volta para a análise do desempenho de ferramentas de classificação morfossintática com dados Web e de mídias sociais, principalmente microblogs, como o Twitter.

2. Fundamentos

Esta seção primeiramente define e classifica problemas relacionados ao reconhecimento de palavras relevantes em textos e, posteriormente, delineia as principais abordagens da literatura.

2.1. Definição do Problema

O Reconhecimento de Palavras Relevantes (*Relevant Words Recognition* - RWR) consiste em identificar e classificar entidades nomeadas e outros componentes com valor sintático e/ou semântico significativo em textos (DOWNEY, et al., 2007). RWR é considerado em diversas áreas de pesquisa, sendo que cada uma dessas se utiliza de diferentes abordagens para resolver seus próprios subproblemas. RWR pode ser subdividido em extração de *tokens*, segmentação de sentenças (*Chunking*) Reconhecimento de Entidades Nomeadas (*Named Entity Recognition* – NER), Desambiguação de Entidades Nomeadas (*Named Entity Desambiguation* – NED), *Shallow Parsing* e *PoS Tagging* (*Part-of-Speech Tagging*). A Figura 1 ilustra esses subproblemas de RWR, que podem ser vistos como um processo onde somente a tokenização é uma tarefa obrigatória, mas cujos resultados são potencialmente muito melhores com todas essas tarefas realizadas de maneira ordenada e, por vezes, cooperativa. Ténicas para tratar NER/NED fazem uso de regras sintáticas, dicionários de nomes e/ou *machine learning*. Todavia, tais tarefas só identificam e classificam um subconjunto de palavras relevantes, as entidades nomeadas. *Shallow Parsing* e *PoS Tagging* (DANIEL; JAMES, 2009), que por sua vez, são tarefas de NLP bastante similares entre si, podem ser usadas para complementar o processo de extração de palavras relevantes, identificando componentes morfossintáticos, tais como verbos (que podem ajudar no entendimento de ações, por exemplo) e adjetivos (que podem ser úteis para detectar polaridade, entre outras possibilidades).

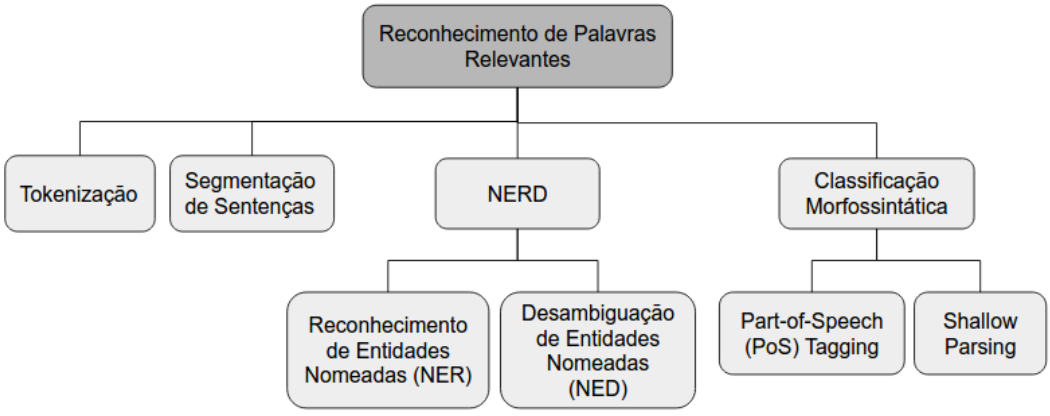

Figura 1: Subproblemas de RWR

2.2. Abordagens

RWR pode ser realizado com diversas abordagens. Esta seção apresenta algumas das técnicas usadas para realizar a anotação morfossintática, bem como NER/NED.

2.2.1. Técnicas Determinísticas

Dicionário de nomes é uma das técnicas mais simples para NER, onde cada entrada corresponde a um par <*nome de superfície, entidade nomeada*>. A busca por um nome de superfície em um dicionário seleciona todas as entidades com tal nome (KLEIN , 2015). Contudo, essa abordagem é restrita e não trata ambiguidade. Geralmente se utilizam dicionários de nomes em conjunto com outras técnicas.

Regras sintáticas, assim como gramáticas, se baseiam na estrutura sintática das línguas para identificar padrões em texto. Essas técnicas apresentam grandes dificuldades e taxas de erros altas, assim como performance pior devido à quantidade de informações que precisam ficar armazenadas durante sua execução.

2.2.2. Técnicas Baseadas em Aprendizado de Máquina

Apesar da facilidade de implementação das técnicas determinísticas, métodos que fazem uso de aprendizado de máquina têm gerado melhores resultados atualmente. A maioria das abordagens de aprendizado de máquina nesse âmbito tratam o problema de classificação de sequências. Em NLP, o problema da classificação de sequências pode ser usado nas tarefas de PoS Tagging, NER e Shallow Parsing.

Classificação de sequências é uma técnica robusta que classifica uma sentença por vez. Dada uma sequência de *tokens*, gera-se uma lista de marcadores que indicam a presença de menções. Uma vez que o classificador de sequências é treinado, utiliza-se o algoritmo de Viterbi (FORNEY , 1973) para calcular uma sucessão de estados que mais provavelmente gera uma coleção de observações, através de programação dinâmica.

Hidden Markov Models (HMM) é uma técnica de modelagem estatística para problemas lineares, tais como séries temporais e análise de sequências (EDDY, 1996). Trata-se de um modelo finito que descreve a probabilidade de distribuição sobre um número infinito de possíveis sequências. O modelo é composto por um número N de estados, que são ligados por transições de distribuição de probabilidade dependentes dos estados anteriores. A máquina de estados emite um símbolo de acordo com a probabilidade de distribuição após uma transição ser feita, que depende do estado atual. Ao chegar no estado final terá sido criada uma sequência de símbolos emitidos pelos estados. O HMM foi usado por muitos anos para resolver problemas de classificação de sequências e inspirou a criação de muitos outros modelos.

Em (BRANTS, 2000), o autor apresenta um *tagger* estatístico baseado em trigramas (*Trigrams'n'Tags* – TnT). O *tagger* usa modelos de Markov de segunda ordem, onde os estados dos modelos representam *tags* e as saídas representam as palavras. As probabilidades de transição dependem dos estados, enquanto as probabilidades de saída dependem somente da categoria mais recente.

Como trigramas costumam sofrer com o problema de dados esparsos, é usada uma técnica de suavização. Para lidar com palavras desconhecidas, o *tagger* usa o sufixo da palavra para inferir sua classificação morfossintática.

Aprendizagem baseada em memória (Memory-Based Part of Speech Tagger-Generator – MBT) é uma forma de aprendizado supervisionado fundada em raciocínio baseado em similaridade. A anotação de uma palavra em um contexto particular é extrapolada a partir dos casos mais semelhantes presentes em memória (DAELEMANS

et al., 1996). Algumas das vantagens que esse método apresenta são o tamanho mínimo necessário relativamente pequeno do *corpus* anotado para treinamento, bons resultados para palavras desconhecidas sem análise morfológica e a aprendizagem incremental e rápida.

O trabalho de (BRILL, 1995) utiliza aprendizagem orientada pelo erro e baseada em transformações (Transformation-Based Error-Driven Learning – TBL) para realizar anotação. Inicialmente o texto de entrada sem anotações passa por um anotador de estados inicial. Esse anotador inicial pode variar em complexidade e uma vez que o texto é processado será então comparado com um *corpus* anotado manualmente que é usado como referência. Uma lista ordenada de transformações é aprendida e pode ser aplicada à saída do anotador inicial para torná-lo melhor e mais semelhante ao *corpus* usado como referência.

O modelo de máxima entropia (RATNAPARKHI et al., 1996) pressupõe que todas as probabilidades são igualmente relevantes e independentes umas das outras. O *tagger* usa elementos de contexto para predizer as anotações. Essa é uma técnica extremamente flexível para a modelagem linguística, que combina vantagens de outras técnicas, uma vez que usa uma representação rica em recursos, como TBL e gera uma distribuição de probabilidade da *tag* para cada palavra, como as técnicas de HMM e árvores de decisão.

Conditional Random Fields (CRF) (LAFFERTY; MCCALLUM; PEREIRA, 2001) usa o princípio de que uma *tag* depende da palavra atual, da imediatamente anterior e da próxima. CRFs podem ser usados para tarefas de NLP como *PoS Tagging*, *Shallow Parsing* e NER (SHA; PEREIRA, 2003).

Ensemble learning (ARBIB , 2003) combina resultados de várias ferramentas. A maneira mais simples de fazer tal combinação é usando algum esquema de votação. Dessa forma, teoricamente os resultados finais são potencialmente melhores, porém obtidos com maiores custos (e.g., tempo de processamento, uso de memória).

3. Análise das Ferramentas

Esta seção descreve as principais ferramentas para RWR avaliadas no âmbito desta pesquisa e compara suas características. Os critérios de seleção das ferramentas a partir da literatura técnico-científica pesquisada foram: proeminência, grau de recomendação, qualidade da documentação, suporte à Língua Portuguesa ou foco específico no processamento de mídias sociais. Foram incluídas neste trabalho algumas ferramentas para NER/NED, para fornecer uma visão geral de RWR. As ferramentas avaliadas foram divididas em ferramentas para *PoS Tagging*, ferramentas para NER/NED e suítes, que são ferramentas que implementam diversas tarefas de NLP.

3.1. Ferramentas para PoS Tagging

TreeTagger (SCHMID, 2013) é um *PoS Tagger* probabilístico desenvolvido por Helmut Schmid, no Instituto de Linguística Computacional da Universidade de Stuttgart. Ela usa árvores de decisão para contornar o problema de dados esparsos e tem sido aplicado com sucesso em vários idiomas. A árvore de decisão é construida recursivamente a partir de um conjunto de treinamento de trigramas, usando uma versão modificada do algoritmo ID3 (QUINLAN, 1986).

LX-Tagger (BRANCO; SILVA, 2006) é uma ferramenta que faz parte da LX-Suite, desenvolvida pelo Natural *Language and Speech Group* da Universidade de Lisboa, com foco no processamento da Língua Portuguesa, e composta pelos módulos de segmentação de sentença, *PoS-tagger*, tokenizador, *Nominal featurizer* e *Nominal lemmatizer*. O *PoS tagger* utilizado é o (BRANTS, 2000), treinado com 90% de um corpus de 280.000 *tokens.*

3.2. Ferramentas para NER/NED

FOX (*Federated knOwledge eXtraction framework*) (SPECK; NGOMO, 2014) é um *framework* de código aberto que implementa serviços Web RESTful para NER e NED. FOX usa AGDISTIS, Weka e uma rede Perceptron multicamadas. FOX usa *ensemble learning* para combinar os resultados de algumas das melhores ferramentas atuais para NER e NED. Os autores afirmam que a ferramenta consegue atingir um percentual de medida-F de aproximadamente 95.23%. O *workflow* realizado pela ferramenta consiste em quatro passos: No primeiro passo, de pré-processamento, o usuário fornece como entrada uma URL, texto com tags HTML ou texto simples. Se a entrada for uma URL, FOX envia uma requisição à URL fornecida para receber os dados de entrada. Para todos os formatos de entrada, FOX remove as tags HTML e detecta frases e tokens. No segundo passo é realizado o NER, via combinação dos resultados de quatro outras ferramentas: Stanford NER, Illinois Named Entity Tagger, Ottawa Baseline Information Extraction e Apache OpenNLP Name Finder. Na terceira etapa é feita a ligação de entidades usando AGDISTIS. Na última etapa é realizada a serialização dos resultados.

Stanford NER (*Stanford Named Entity Recognizer*) (FINKEL; GRENAGER; MANNING, 2005) é um *framework* de código aberto para NER. A ferramenta baseia-se em classificação de sequências e utiliza dez características locais para treinar um CRF. O framework foi implementado em Java pelo grupo de NLP de Stanford, e rotula nomes (de pessoas, empresas, genes e proteínas, entre outros) no texto. A distribuição atual permite o acesso a todos os recursos do pipeline do Stanford CoreNLP, acessados através da classe NERClassifierCombiner. Experimentos com o StanfordNER resultaram em 85,51% na medida-F ao utilizar o dataset da CoNLL[1] e 92,29% ao utilizar o dataset da CMU[2].

DBpedia Spotlight (MENDES et al., 2001) é uma ferramenta para reconhecimento e desambiguação de entidades nomeadas. A ferramenta usa uma abordagem de recuperação de informação, onde candidatos são gerados pela busca em um dicionário de nomes previamente computado no *DBPedia Lexicalizations dataset*[3]. Para realizar a desambiguação, é feita uma sobreposição entre o contexto textual da menção e de cada candidato (KLEIN , 2015). O DBpedia Spotlight permite muitas formas de uso e ajustes de diversas configurações internas para atender as necessidades do usuário.

Illinois Named Entity Tagger (RATINOV; ROTH , 2009) é um *tagger* que marca texto simples com delimitações e classes de entidades nomeadas encontradas no

---

1 Disponível em http://cnts.uia.ac.be/conll2003/ner

2 Disponível em http://nlp.shef.ac.uk/dot.kom/resources.html

3 Disponível em http://wiki.dbpedia.org/lexicalizations

texto de entrada. Atualmente as entidades detectadas por tal ferramenta podem se encaixar nas categorias pessoa, organização, localização e variados, ou ainda dezoito outros tipos, se a classificação for realizada com base no *corpus* OntoNotes. Ela usa *gazetteers* extraídos da Wikipedia, modelos de classes palavras derivadas de texto não rotulado e recursos não-locais expressivos.

NERD (RIZZO; TRONCY, 2012) é um *framework* que propõe unificar as saídas de dez diferentes extratores NLP publicamente disponíveis na Web. A abordagem baseia-se na ontologia NERD[4], que provê uma interface para anotação de elementos e uma API REST usada para unificar as saídas. NERD é uma aplicação web que funciona sobre várias ferramentas de NLP. Sua arquitetura segue os princípios REST e fornece um acesso HTML web para os usuários e uma API para que computadores possam fazer intercâmbio de conteúdo em formato JSON ou XML. Ambas as interfaces são suportadas pela *engine* NERD REST. A *engine* NERD REST e as ferramentas de NLP usadas pelo NERD fazem o reconhecimento e desambiguação de entidades nomeadas apontando URIs de bases de conhecimento para objetos do mundo real, como eles poderiam ser definidos na Web.

3.3. Suites de Ferramentas

FreeLing (PADRÓ; STANILOVSKY, 2012) é uma biblioteca C++ que realiza diversas tarefas de NLP, como identificação da língua, tokenização, segmentação de sentenças, análise morfossintática, NER, NED, *PoS tagging*, *Shallow parsing* e resolução de correferência. Inclui dois módulos diferentes capazes de realizar *PoS tagging*. O primeiro é a classe tagger_hmm, baseada no método de (BRANTS, 2000), ou seja, usa o método de aprendizado baseado em transformações. O segundo módulo é o relax_tagger, um sistema híbrido capaz de integrar conhecimentos sobre estatística e códigos feitos a mão, segundo (PADRÓ, 1998). A aplicação pode instanciar qualquer um dos dois módulos.

GATE Twitter PoS Tagger (DERCZYNSKI et al., 2013) é um *tagger* desenvolvido especificamente para a anotação morfossintática de *tweets*. Para reduzir o impacto da dispersão de dados a ferramenta usa *vote-constrained bootstrapping*. Também são utilizados métodos para o tratamento de erros característicos de *tweets* e gírias. A ferramenta usa a abordagem de CRF e é destinada a *tweets* na língua inglesa e contém uma instância do Stanford Tagger[5].

OpenNLP[6] é um conjunto de ferramentas baseadas em aprendizagem de máquina visando o processamento de linguagem natural. Possui suporte para tarefas como tokenização, segmentação de sentença, *PoS tagging*, extração de entidades nomeadas, *parsing* e resolução de correferência. O *PoS-tagger* usa o modelo de máxima entropia marcando os *tokens* com o seu tipo de palavra correspondente com base no próprio *token* bem como no seu contexto. A ferramenta usa um modelo de probabilidade para prever a *tag* correta para o *token* a partir do conjunto de *tags*. No site do OpenNLP, existem modelos pré-treinados para dinamarquês, alemão, inglês, holandês, sueco e

---

4 Disponível em http://nerd.eurecom.fr

5 Disponível em http://nlp.stanford.edu/software/tagger.shtml

6 Disponível em http://opennlp.apache.org

português. Porém, a ferramenta também permite o treinamento de modelos em outras línguas. Isso pode ser feito por meio de coleções de textos anotados com *tags*.

4. Experimentos

Esta seção descreve os experimentos de classificação morfossintática realizados para a avaliação de ferramentas selecionadas sobre uma coleção de *tweets* e o *corpus* Tycho Brahe.

4.1. Seleção das Ferramentas

As ferramentas selecionadas precisam ser capazes de processar eficientemente uma grande quantidade de texto, principalmente, em Língua Portuguesa, anotar morfossintaticamente as palavras e expressões, ter código fonte aberto ou permitir a utilização em pesquisa, permitir aplicação em mídias sociais, além da possibilidade de integração a outras aplicações. Usando esses critérios foram selecionadas para os experimentos: LX-Tagger, TreeTagger, Gate Twitter PoS Tagger, FreeLing e OpenNLP. Todas elas realizam *PoS Tagging*, ou seja, atribuem uma classe gramatical, na forma de *tag*, a cada palavra do texto de entrada.

4.2. Conjuntos de Dados

Para realização dos experimentos com *tweets*, foram selecionados 100 mil *tweets* escritos em português e postados entre 30/11/2015 e 15/12/2015 no território brasileiro. Inicialmente pretendia-se fazer experimentos com todo o mês de dezembro, porém isso resultou em cerca de 67 milhões de *tweets*, ocupando um arquivo texto de 7,3 GB. Em virtude do volume de informação, foi tomada a decisão de selecionar aleatoriamente a amostra de 100 mil *tweets* dentro do período mencionado. Por exemplo, a frase "Cheguei ontem e apaguei", após passar pelo pelo processamento pode ser representada como *Cheguei*/V *ontem*/ADV *e*/CJ *apaguei*/V, com a *tag* V indicando verbo, ADV advérbio e CJ conjunção. Também filtrou-se de maneira simples o conteúdo dos *tweets*, com o intuito de retirar os *emoticons* e caracteres especiais que pudessem causar o mau funcionamento das ferramentas.

Dada a dificuldade de criação de um padrão ouro manualmente anotado para todo o *dataset*, uma amostra de 383 tweets do dataset foi selecionada aleatoriamente por sorteio, ao nível de confiança de 95%, para analisar a qualidade dos resultados. Essa amostra foi manualmente anotada conforme as seguintes categorias: adjetivo, advérbio, conjunção, determinante (adjetivos), interjeição, nome (substantivos), pontuação, preposição, pronome, sigla/símbolo/número/outros (chamada ``SSN'') e verbo. Na categoria SSN foram colocados também elementos como gírias, URIs e *hashtags*. Foi então realizada a normalização das saídas, substituindo as *tags* dos *outputs* das ferramentas por *tags* correspondentes às categorias mencionadas (os motivos dessa normalização serão apresentados na seção de Resultados e Discussão).

O *corpus* Histórico do Português Tycho Brahe é um *corpus* eletrônico anotado, composto de textos em português escritos por autores nascidos entre 1380 e 1881. Atualmente, 73 textos (3.048.971 palavras) estão disponíveis para pesquisa livre. Destes, 40 possuem anotação morfológica, com um total de 1.728.284 palavras. Tal *corpus* foi selecionado visando causar contraste em relação ao *dataset* de *tweets*, uma

vez que esse *corpus* possui textos formais que seguem a gramática portuguesa rigidamente, ao passo que os *tweets* são altamente informais e contemporâneos.

Para o experimento com o *corpus* Tycho Brahe, primeiramente foi realizada uma análise para determinar a quantidade de sentenças que deveria compor uma amostra estratificada proporcional, com nível de confiança de 95%. Através disso, foi determinado que a amostra do *corpus* deveria conter 368 sentenças, presentes em 37 textos do *corpus*. Os 37 textos completos possuem 1.612.800 palavras, ao passo que a amostra possui 5.330 palavras. A seleção das sentenças foi realizada por sorteio.

No site do *corpus* eletrônico também são disponibilizadas as versões dos textos morfossintaticamente anotados. Dessa forma, as sentenças morfossintaticamente anotadas correspondentes às selecionadas, foram usadas para compor um padrão ouro para esse experimento. A amostra do *corpus* apresenta 5.736 *tokens*, classificados em substantivos, verbos, adjetivos entre outras categorias morfossintáticas.

4.3. Padrão Ouro

A construção de um padrão ouro para *PoS Tagging* se dá pela anotação manual de um *dataset*, atribuindo a cada palavra a sua classificação morfossintática correta. A utilização do padrão ouro é imprescindível para o cálculo da precisão, cobertura e por conseguinte da medida-F, que depende das métricas anteriores. A Tabela 1 mostra a quantidade de *tokens* manualmente anotados na amostra de 383 *tweets*, conforme a distribuição das categorias morfossintáticas. A frequência das classes gramaticais obedece a literatura, pois as classes com mais *tokens* são as de substantivos e verbos.

5. Resultados e Discussão

Nesta seção são apresentados e discutidos os resultados dos experimentos. Os resultados foram divididos segundo os seguintes aspectos: variabilidade das categorias morfossintáticas, análise da qualidade dos resultados e por fim, análise do tempo de execução das ferramentas.

5.1. Variabilidade das Categorias Morfossintáticas

As ferramentas receberam como entrada os textos dos *tweets* e geraram como saída um arquivo texto com cada palavra relevante neles encontrada, anotada com a sua categoria morfossintática. A anotação, denominada *tagset*, variou conforme a especificidade da classificação morfossintática da ferramenta. Por exemplo, a TreeTagger usa as seguintes notações para adjetivo: AO - adjetivo ordinal, AOS - adjetivo ordinal superlativo, AQ - adjetivo qualificativo, AQA - adjetivo qualificativo aumentativo, AQC - adjetivo qualificativo diminutivo, AQS - adjetivo qualificativo superlativo; enquanto a Gate Twitter PoS Tagger usa JJ - adjetivo, JJR - adjetivo comparativo, JJS - adjetivo superativo. Em razão de tais diferenças, subcategorias foram agrupadas em grandes categorias.

Os resultados obtidos estão descritos na tabela e figuras a seguir. A Tabela 2 mostra a quantidade de *tokens* encontrados pelas ferramentas, conforme a distribuição das categorias morfossintáticas para a amostra de 100 mil *tweets*. Note que mesmo o número total de *tokens* detectado foi diferente para cada ferramenta. Isso ocorre em

função dos diferentes algoritmos adotados nas ferramentas para realizar os processos de tokenização, *chunking* e *PoS Tagging*. A FreeLing encontrou mais *tokens* que as demais porque possui funcionalidades para lidar com texto coloquial, e também por implementar vários módulos que identificam uma grande variedade de *tokens*, conseguindo enriquecer os *tweets* com novas informações linguísticas. Assim como a FreeLing, a TreeTagger faz lematização, gerando anotações mais detalhadas.

Categoria	Quantidade
ADJ	169
ADV	323
CON	114
DET	234
INT	30
NOM	1127
PON	306
PRE	394
PRO	450
SSN	478
VER	787
Total	**4412**

Tabela 1: *Tokens* do padrão-ouro para cada categoria morfossintática. ADJ=Adjetivo, ADV=Advérbio, CON=Conjunção, DET=Determinante, INT=Interjeição, NOM=Substantivo, PON=Pontuação, PRE=Preposição, PRO = Pronome, SSN = Sigla/Símbolo/Número/Outros, VER =Verbo

A Figura 2 mostra as diferenças entre as médias de ocorrências das categorias morfossintáticas nos resultados das ferramentas. O gráfico do tipo *boxplot* apresentado na Figura 2 permite analisar a variabilidade dos resultados (diferença entre o terceiro e o primeiro quartil), a mediana (linha central) e os máximos e mínimos (calda inferior e superior). A ordem decrescente de frequência média das categorias (NOM, VER, PON, PRO, PRE, ADJ, SSN, DET, ADV, INT e COM) condiz com a literatura, onde se demonstra que substantivo e verbo são as categorias mais frequentes e importantes para análise linguística. A mediana das categorias em muitos casos está deslocada do centro, e a cauda muito longa ou curta, denotando falta de simetria. Isso está relacionado à grande diferença dos resultados das ferramentas para cada categoria. Observa-se que para as categorias NOM, PON, PRO, SSN e VER a variabilidade é alta. Incluir a lematização das palavras juntamente com a anotação morfossintática ajuda no enriquecimento semântico, mas deve-se ter o cuidado de verificar o contexto da aplicação. Se a quantidade de dados for grande, a inclusão de mais componentes

linguísticos no processo de *PoS Tagging* pode acarretar em maior tempo de processamento. Outra forma de verificar a variabilidade dos resultados é usada na Figura 3, que ilustra o coeficiente de variabilidade (desvio-padrão/ média) das categorias. Quanto menor tal coeficiente, mais homogêneo é o conjunto de dados. Ao observar-se que todos os coeficientes foram maiores que 25%, valor esse considerado limite para afirmar estatisticamente que não existe homogeneidade nas ferramentas quanto a classificação.

Categoria	**Ferramentas**				
	LX-Tagger	**TreeTagger**	**Gate Twitter PoS Tagger**	**Freeling**	**OpenNLP**
ADJ	39174	60348	11671	52210	42667
ADV	34520	62672	1376	68227	49509
CON	37561	41654	2457	40475	38229
DET	104520	110358	34581	110062	37870
INT	2069	4167	54847	3739	0
NOM	136539	444741	211192	346059	147713
PON	63108	123546	0	210697	43044
PRE	44188	100019	23832	103938	51170
PRO	35420	77534	13178	109111	87629
SSN	9787	11003	85101	37557	10985
VER	139389	190160	42726	211282	105179
Total	**646275**	**1226202**	**480961**	**1293357**	**613995**

Tabela 2: Estimativa de *tokens pelas ferramentas* para cada categoria morfossintática. ADJ=Adjetivo, ADV=Advérbio, CON=Conjunção, DET=Determinante, INT=Interjeição, NOM=Substantivo, PON=Pontuação, PRE=Preposição,PRO = Pronome, SSN = Sigla/Símbolo/Número/Outros, VER =Verbo

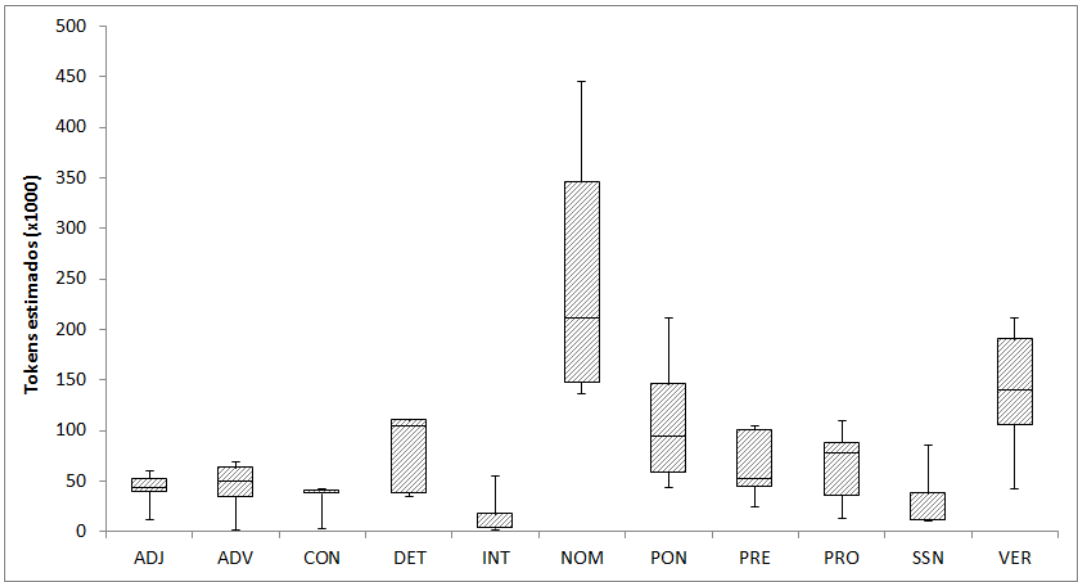

Figura 2: Variabilidade das categorias morfossintáticas.

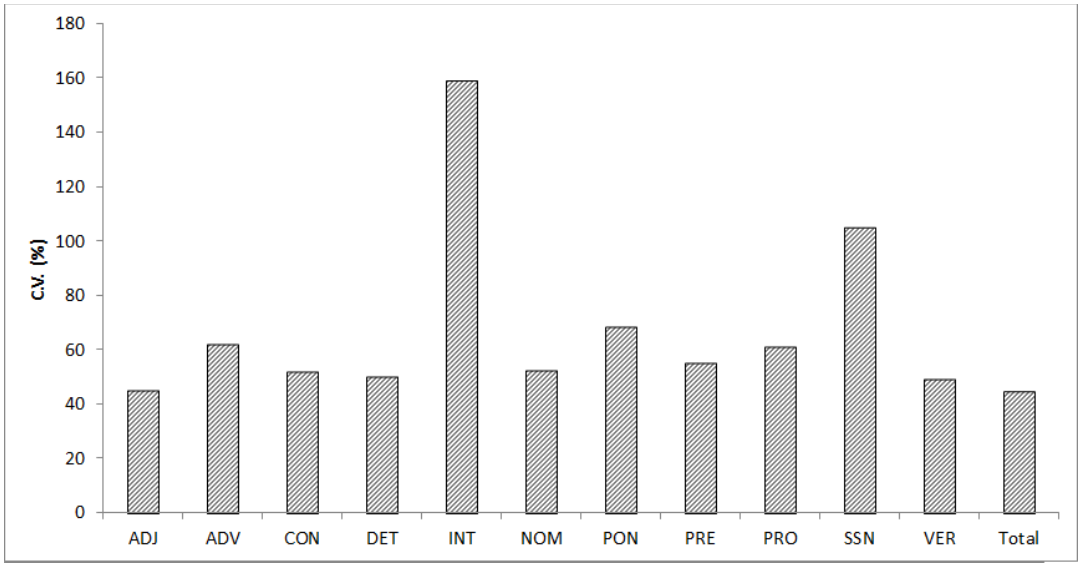

Figura 3: Coeficiente de variabilidade (CV).

5.2. Análise da Qualidade dos Resultados

A avaliação de precisão, cobertura e medida-F foi realizada usando a ferramenta ROUGE (LIN, 2004) com intervalo de confiança de 95%. Para fazer essa análise, os arquivos de saída das ferramentas já normalizados foram alterados da forma "palavra_tag", deixando somente as *tags*. Isso foi necessário pois a ferramenta realiza a avaliação por unigramas. Logo, retirando as palavras, não corre-se o risco de que a ferramenta interprete incorretamente o que é um unigrama, fornecendo maior confiabilidade ao teste. A Tabela 3 mostra os resultados obtidos de precisão, cobertura e medida-F para o *dataset* de *tweets* e a amostra do *corpus* Tycho Brahe. Como pode ser observado, para o *dataset* de *tweets* a ferramenta OpenNLP obteve o melhor resultado, enquanto as ferramentas LX-Tagger, Freeling e TreeTagger obtiveram resultados

semelhantes, ao passo que a ferramenta GATE Twitter PoS Tagger obteve o pior resultado. Apesar de ser uma ferramenta dedicada especialmente para o processamento de *tweets*, devido à falta de suporte para a língua portuguesa os resultados da ferramenta GATE Twitter PoS Tagger não conseguiu alcançar as outras.

Durante a análise dos resultados, um ponto importante notado foi a disparidade da tokenização e dos *tagsets* usados pelas ferramentas. Por exemplo, a TreeTagger e Freeling consideram a pontuação (e.g., ponto final, vígula, interrogação) como um token único, ao passo que OpenNLP, LX Tagger e Gate Twitter PoS Tagger consideram a pontuação junto com a palavra imediatamente anterior um *token*, na maioria do arquivo de saída. Há ferramentas que consideraram uma URI ou menção a usuário como *token*, enquanto outras separam essas estruturas em vários *tokens*. Os *tagsets* usados pelas ferramentas não possuem nenhuma padronização (com exceção da TreeTagger e Freeling, que são similares), o que torna o processo de comparação mais difícil, portanto foi adotada a estratégia de agrupamentos em classes maiores, conforme mencionado anteriormente.

Como esperado, houve uma melhora nos resultados obtidos com a amostra do *corpus* Tycho Brahe, porém, ainda não foram alcançados resultados próximos aos apresentados na literatura (que em média beiram a 99% de precisão e cobertura para outros *corpus* existentes). Apesar das ferramentas utilizadas terem apresentado alguns bons resultados neste estudo, há evidências de que o problema de *PoS Tagging* ainda possui espaço para melhoras, principalmente no âmbito de mídias sociais. Além disso, em uma breve análise verificou-se palavras anotadas com categoria incorreta. Uma solução para os problemas constatados nos resultados das ferramentas seria considerar tarefas que melhoram a qualidade dos dados na etapa de pré-processamento, como a normalização de abreviações e acrônimos principalmente, ou ainda, subdividir o conjunto de dados em um conjunto de teste e treinamento e repetir os experimentos. A normalização desta linguagem também pode ser crucial para facilitar o processamento de texto e melhorar o desempenho de ferramentas de NLP aplicadas a mídias sociais. Contudo, dependendo do propósito da aplicação, a normalização torna-se difícil, por exemplo, uma URI pode ser considerada como um *token* ou mesmo como vários (quando utiliza-se palavras com sentido próprio na composição).

Para o caso de microblogs, como o Twitter, vale ainda mais o uso de fonetização para tentar corrigir certos erros gramaticais. Também, o uso de recursos extras para a identificação de gírias poderia melhorar a qualidade dos resultados, porém seria necessário analisar o impacto do tempo adicional de processamento que pode ser acarretado pela adição desses recursos de pré-processamento, uma vez que métodos de NLP são computacionalmente pesados.

5.3. Análise do Tempo de Execução das Ferramentas

Durante os experimentos, também foi medido o tempo de execução das ferramentas. A Tabela 4 mostra a média aritmética de cinco execuções de cada ferramenta para a mesma amostra, onde "Amostra 1" corresponde ao *dataset* com 100 mil *tweets*, "Amostra 2" à amostra desse *dataset*, com 383 *tweets* e por fim, "Amostra 3" à amostra do *corpus* Tycho Brahe com 368 sentenças e 5.330 palavras. Os experimentos, bem como a medição dos tempos foi realizada usando um notebook com processador Intel i5 com sistema operacional Ubuntu 14.04 e todas as ferramentas tiveram suas rotinas de

anotação invocadas via terminal. Pode-se observar, que os tempos de execução mantiveram-se na ordem de poucos minutos ou segundos, exceto pelo tempo obtido pela GATE Twitter PoS Tagger para a "Amostra 1", que foi discrepantemente maior que todos os outros. Com isso, pode concluir-se que caso o conjunto de dados possuísse milhões de sentenças, a performance das ferramentas poderia ser um grande gargalo em certos casos, sugerindo a necessidade de paralelização para casos onde o volume de dados é muito grande.

Ferramenta	***Dataset* de *Tweets***			**Amostra do *Corpus***		
	Precisão	**Cobertura**	**Medida-F**	**Precisão**	**Cobertura**	**Medida-F**
LX-Tagger	0.84585	0.78782	0.81580	0.83043	0.71809	0.77019
Freeling	0.80573	0.87251	0.83779	0.92938	0.95848	0.94371
TreeTagger	0.76705	0.82020	0.79274	0.91095	0.95912	0.93441
OpenNLP	0.92029	0.85756	0.85756	0.95385	0.82482	0.88465
Gate Twitter PoS Tagger	0.52059	0.46082	0.48888	0.44195	0.38037	0.40885

Tabela 3: Resultados de precisão, cobertura e medida-F para as amostras.

Ferramenta	**Amostra 1**	**Amostra 2**	**Amostra 3**
LX-Tagger	172.54s	2.05s	2.22s
Freeling	138.95s	3.24s	3.29s
TreeTagger	10.64s	0.58s	0.58s
OpenNLP	26.53s	0.69s	0.73s
Gate Twitter PoS Tagger	14340,12s	56.42s	157.43s

Tabela 4: Média dos tempos de execução das ferramentas.

6. Conclusão e Trabalhos Futuros

Atualmente, apesar da maior disponibilidade de ferramentas para processamento de texto, muitos problemas ainda persistem: baixa interoperabilidade (principalmente entre ferramentas comerciais); limitações no compartilhamento de recursos (em função da dependência de plataformas específicas para execução); falta de padronização; disponibilização apenas de versões online para teste e por vezes, restrições de capacidade de processamento ao lidar com grandes volumes de dados. Outro ponto crítico é a escassez de recursos apropriados para processar textos naturais de mídias sociais, uma vez que a única ferramenta específica para essa finalidade encontrada foi a

GATE Twitter PoS Tagger. Também faltam recursos para o processamento de textos em certos idiomas, como é o caso da Língua Portuguesa.

O estudo do estado da arte mostra que propostas recentes tendem ao uso de técnicas baseadas em aprendizado de máquina, pois estas começam a gerar melhores resultados e proveem a possibilidade do treinamento de um modelo específico.

Este trabalho apresentou uma revisão bibliográfica sobre as tarefas de Reconhecimento de Palavras Relevantes, com foco em anotação morfossintática. Foram apresentadas técnicas e ferramentas consistentes com o estado-da-arte de *PoS Tagging*. Os experimentos com as ferramentas selecionadas demonstraram que há grande variabilidade na qualidade dos resultados quando o texto é oriundo de mídias sociais. A análise comparativa aqui apresentada de ferramentas para o RWR em textos distingue-se de trabalhos afins pelas seguintes contribuições: (i) cobertura de uma variedade de subproblemas, métodos e ferramentas de áreas de pesquisa tradicionalmente separadas (ii) foco na análise de postagens em mídias sociais usando a língua portuguesa; (iii) resultados experimentais mais abrangentes, analisando as disparidades apresentadas por ferramentas distintas; (iv) evidência experimental de que os resultados de *PoS Tagging* obtidos com textos oriundos de mídias sociais são piores; (v) análise do desempenho das ferramentas através da medição do tempo de execução da tarefa de *Pos Tagging*; e (vi) construção de um padrão ouro para *tweets* em português.

7. Referências

HABIB, M. B.; KEULEN, M. van. Information extraction for social media. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. [S.l.], 2014.

SANG, E. F. T. K.; MEULDER, F. D. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4. [S.l.], 2003. p. 142–147.

MARTIN, J. H.; JURAFSKY, D. Speech and language processing. International Edition, v. 710, 2000.

WANG, W.; STEWART, K. Spatiotemporal and semantic information extraction from web news reports about natural hazards. Computers, Environment and Urban Systems, Elsevier, v. 50, p. 30–40, 2015.

ANANTHARAM, P. et al. Extracting city traffic events from social streams. ACM Transactions on Intelligent Systems and Technology (TIST), ACM, v. 6, n. 4, p. 43, 2015.

XIA, C. et al. What is new in our city? a framework for event extraction using social media posts. In: SPRINGER. Pacific-Asia Conference on Knowledge Discovery and Data Mining. [S.l.], 2015. p. 16–32.

FILETO, R. et al. The baquara 2 knowledge-based framework for semantic enrichment and analysis of movement data. Data & Knowledge Engineering, Elsevier, v. 98, p. 104–122, 2015.

LEV, B.; THIAGARAJAN, S. R. Fundamental information analysis. Journal of Accounting research, JSTOR, p. 190–215, 1993.

SACENTI, J. A. et al. Automatically tailoring semantics-enabled dimensions for movement data warehouses. In: SPRINGER. International Conference on Big Data Analytics and Knowledge Discovery. [S.l.], 2015. p. 205–216.

BERRY, M. J.; LINOFF, G. Data mining techniques: for marketing, sales, and customer support. [S.l.]: John Wiley & Sons, Inc., 1997.

DAS, T.; ACHARJYA, D.; PATRA, M. Opinion mining about a product by analyzing public tweets in twitter. In: IEEE. Computer Communication and Informatics (ICCCI), 2014 International Conference on. [S.l.], 2014. p. 1–4.

PAZZANI, M. J.; BILLSUS, D. Content-based recommendation systems. In: The adaptive web. [S.l.]: Springer, 2007. p. 325–341.

CHOWDHURY, G. G. Natural language processing. Annual review of information science and technology, Wiley Online Library, v. 37, n. 1, p. 51–89, 2003.

BIZER, C.; HEATH, T.; BERNERS-LEE, T. Linked data-the story so far. Semantic Services, Interoperability and Web Applications: Emerging Concepts, p. 205–227, 2009.

DOWNEY, D.; BROADHEAD, M.; ETZIONI, O. Locating complex named entities in web text. In: IJCAI. [S.l.: s.n.], v. 7, p. 2733–2739, 2007.

DANIEL, J.; JAMES, H. Speech and language processing: An introduction to natural language processing. Computational Linguistics and Speech Recognition, 2nd Ed., Prentice Hall, 2009.

KLEIN, D. Estudo de técnicas e ferramentas aplicáveis a mídias sociai para reconhecimento e desambiguação de entidades nomeadas. Universidade Federal de Santa Catarina, p. 9, 2015.

FORNEY, G. D. The viterbi algorithm. Proceedings of the IEEE, IEEE, v. 61, n. 3, p. 268–278, 1973.

EDDY, S. R. Hidden markov models. Current opinion in structural biology, Elsevier, v. 6, n. 3, p. 361–365, 1996.

BRANTS, T. Tnt: a statistical part-of-speech tagger. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. Proceedings of the sixth conference on Applied natural language processing. [S.l.], 2000. p. 224–231.

DAELEMANS, W. et al. Mbt: A memory-based part of speech tagger-generator. arXiv preprint cmp-lg/9607012, 1996.

BRILL, E. Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. Computational linguistics, MIT Press, v. 21, n. 4, p. 543–565, 1995.

RATNAPARKHI, A. et al. A maximum entropy model for part-of-speech tagging. In: PHILADELPHIA, USA. Proceedings of the conference on empirical methods in natural language processing. [S.l.], 1996. v. 1, p. 133–142.

LAFFERTY, J.; MCCALLUM, A.; PEREIRA, F. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proceedings of

the eighteenth international conference on machine learning, ICML. [S.l.: s.n.], 2001. v. 1, p. 282–289.

SHA, F.; PEREIRA, F. Shallow parsing with conditional random fields. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1. [S.l.], 2003. p. 134–141.

ARBIB, M. A. The handbook of brain theory and neural networks. [S.l.]: MIT press, 2003.

SCHMID, H. Probabilistic part-ofispeech tagging using decision trees. In: ROUTLEDGE. New methods in language processing. [S.l.], 2013. p. 154.

QUINLAN, J. R. Induction of decision trees. Machine learning, Springer, v. 1, n. 1, p. 81–106, 1986.

BRANCO, A.; SILVA, J. R. A suite of shallow processing tools for portuguese: Lx-suite. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics: Posters & Demonstrations. [S.l.], 2006. p. 179–182.

SPECK, R.; NGOMO, A.-C. N. Named entity recognition using fox. In: CEUR-WS. ORG. Proceedings of the 2014 International Conference on Posters & Demonstrations Track-Volume 1272. [S.l.], 2014. p. 85–88.

FINKEL, J. R.; GRENAGER, T.; MANNING, C. Incorporating non-local information into information extraction systems by gibbs sampling. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. [S.l.], 2005. p. 363–370.

MENDES, P. N. et al. Dbpedia spotlight: shedding light on the web of documents. In: ACM. Proceedings of the 7th international conference on semantic systems. [S.l.], 2011. p. 1–8.

RATINOV, L.; ROTH, D. Design challenges and misconceptions in named entity recognition. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. Proceedings of the Thirteenth Conference on Computational Natural Language Learning. [S.l.], 2009. p. 147–155.

RIZZO, G.; TRONCY, R. Nerd: a framework for unifying named entity recognition and disambiguation extraction tools. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics. [S.l.], 2012. p. 73–76.

PADRÓ, L. A hybrid environment for syntax-semantic tagging. arXiv preprint cmp-lg/9802002, 1998.

DERCZYNSKI, L. et al. Twitter part-of-speech tagging for all: Overcoming sparse and noisy data. In: RANLP. [S.l.: s.n.], 2013. p. 198–206.

LIN, C.-Y. Rouge: A package for automatic evaluation of summaries. In: BARCELONA, SPAIN. Text summarization branches out: Proceedings of the ACL-04 workshop. [S.l.], 2004. v. 8.