UMA FERRAMENTA DE APOIO NA IDENTIFICAÇÃO DE NOVOS ELEMENTOS GEOGRÁFICOS DE BAIXA GRANULARIDADE EM NOTÍCIAS PARA A ATUALIZAÇÃO DE DICIONÁRIOS GEOGRÁFICOS

Matheus Emerick de Magalhães

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Engenharia de Sistemas e Computação, COPPE, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Mestre em Engenharia de Sistemas e Computação.

Orientador: Geraldo Bonorino Xexéo

Rio de Janeiro
Julho de 2016

UMA FERRAMENTA DE APOIO NA IDENTIFICAÇÃO DE NOVOS ELEMENTOS GEOGRÁFICOS DE BAIXA GRANULARIDADE EM NOTÍCIAS PARA A ATUALIZAÇÃO DE DICIONÁRIOS GEOGRÁFICOS

Matheus Emerick de Magalhães

DISSERTAÇÃO SUBMETIDA AO CORPO DOCENTE DO INSTITUTO ALBERTO LUIZ COIMBRA DE PÓS-GRADUAÇÃO E PESQUISA DE ENGENHARIA (COPPE) DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM CIÊNCIAS EM ENGENHARIA DE SISTEMAS E COMPUTAÇÃO.

Examinada por:

_____
Prof. Geraldo Bonorino Xexéo, D.Sc.

_____
Prof. Jano Moreira de Souza, Ph.D.

_____
Prof. Julia Celia Mercedes Strauch, D.Sc.

RIO DE JANEIRO, RJ - BRASIL
JULHO DE 2016

Magalhães, Matheus Emerick de

Uma ferramenta de apoio na identificação de novos elementos geográficos de baixa granularidade em notícias para a atualização de dicionários geográficos / Matheus Emerick de Magalhães. – Rio de Janeiro: UFRJ/COPPE, 2016.

X, 87 p.: il.; 29,7 cm.

Orientador: Geraldo Bonorino Xexéo

Dissertação (mestrado) – UFRJ/ COPPE/ Programa de Engenharia de Sistemas e Computação, 2016.

Referências Bibliográficas: p. 80-87.

1. Identificação de conteúdo geográfico. 2. Extração de informação. 3. Reconhecimento de entidades mencionadas. I. Xexéo, Geraldo Bonorino. II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia de Sistemas e Computação. III. Título.

Agradecimentos

Primeiramente, agradeço ao único, poderoso e digno Deus, por ter me dado sabedoria, força de vontade e as demais condições necessárias para concluir este trabalho.

Agradeço aos meus pais Osmar e Marilsa por estarem sempre apoiando todas as minhas decisões e ajudando nos momentos difíceis.

Agradeço ao meu orientador Geraldo Bonorino Xexéo, pela dedicação, paciência e boa vontade em me auxiliar durante toda a realização desse trabalho.

Aos professores Jano Moreira de Souza e Julia Celia Mercedes Strauch, por aceitarem fazer parte da banca examinadora.

Aos amigos Joaquim Viana Junior, Carlos Eduardo Barbosa, Daniel Schneider, Rafael Ris-Ala José Jardim, Andreia Oliveira, Egberto Caetano, Rogério Borba, Fellipe Duarte, Lilian Magalhães e a tantos que contribuíram com meu trabalho.

Resumo da Dissertação apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Mestre em Ciências (M.Sc.)

UMA FERRAMENTA DE APOIO NA IDENTIFICAÇÃO DE NOVOS ELEMENTOS GEOGRÁFICOS DE BAIXA GRANULARIDADE EM NOTÍCIAS PARA A ATUALIZAÇÃO DE DICIONÁRIOS GEOGRÁFICOS

Matheus Emerick de Magalhães

Julho/2016

Orientador: Geraldo Bonorino Xexéo

Programa: Engenharia de Sistemas e Computação

O Brasil é um país vasto e dinâmico. Identificar os novos elementos inaugurados ou atualizados é uma tarefa que envolve grande esforço financeiro, político e informacional. A necessidade por informações precisas sob o espaço geográfico que vivemos, criou uma demanda por serviços automatizados de reconhecimento de endereços geográficos de baixa granularidade e alto grau de especificidade. Como a internet disponibiliza e integra diversas fontes de informações, principalmente em notícias dos mais diversos meios, sobre elementos inaugurados em nosso país, estado, cidade e rua torna-se necessário recuperar e estruturar essas informações de forma a poder relacioná-las com o contexto e realidade dos locais em que vivemos através de métodos e sistemas automatizados. Órgãos públicos também possuem a necessidade de identificar os novos elementos geográficos, contudo, para que a informação seja útil deve possuir elementos geográficos mais precisos, para apoiar em atividades como a tarefa de reambulação. Para isso uma das necessidades é possibilitar o georreferenciamento de notícias, ou seja, identificar as entidades geográficas presentes e associá-las com sua correta localização espacial. O presente trabalho propõe uma abordagem para criar regras gramaticais que possibilitem a identificação de elementos geográficos de baixa granularidade que apoie na criação e atualização de dicionários geográficos baseado em notícias. Os resultados apresentam a utilidade da abordagem para a criação de uma ferramenta de apoio à identificação de endereços geográficos que apoie ao enriquecimento de dicionários geográficos e às atividades relacionas as tarefas de reambulação.

Abstract of Dissertation presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Master of Science (M.Sc.)

SUPPORTING TOOL IN IDENTIFYING NEW LOW GEOGRAPHIC ELEMENTS GRANULARITY IN NEWS FOR UPDATING DICTIONARIES GEOGRAPHIC

Matheus Emerick de Magalhães

July/2016

Advisor: Geraldo Bonorino Xexéo

Department: Computer and Systems Engineering

Brazil is a vast, dynamic country. Identifying new updated or opened elements is a task that involves great financial, political and informational effort. The need for accurate information about the geographical space we live in has created a demand for automated services of recognition of geographical addresses of low granularity and high degree of specificity. As the internet provides and integrates various sources of information, especially in the news from a variety of means on opened elements in our country, state, city and street, it becomes necessary to recover and structure this information in order to relate them with the context and the reality of the places in which we live through automated methods and systems. Public bodies also have the need to identify new geographic elements. However, in order for the information to be useful, it needs to have more precise geographic features to support activities such as the task of collecting geographical names. For that, one of the needs is to enable the news georeferencing, i.e., to identify the present geographical entities and associate them with their correct spatial location. This paper proposes an approach to create grammatical rules that allow the identification of geographic features of low granularity which support the creation and update of geographical dictionaries based on news. The results demonstrate the usefulness of the approach in creating a tool to support the identification of spatial addresses that supports the enrichment of geographical dictionaries and the activities related to the task of collecting geographical names.

Sumário

1 Introdução 1
1.1 Contextualização do Problema 1
1.2 Delimitação 4
1.3 Objetivo 5
1.4 Organização do Trabalho 5
2 Referencial Teórico 7
2.1 Extração da Informação 7
2.1.1 Extração de Informação Geográfica em Noticias 8
2.2 Sistema de Informações Geográficas 9
2.2.1 Geoprocessamento 10
2.2.2 Georreferenciamento 11
2.2.3 Identificador de Referências Geográficas 13
2.2.4 A Geocodificação de Elementos Geográficos 13
2.3 Dicionários Geográficos 14
2.3.1 Principais Dicionários Geográficos 16
2.3.2 Trabalhos relacionados 18
2.3.3 Identificadores de Localidade 23
3 Tarefas para o Reconhecimento de Elementos Geográficos em Texto 25
3.1 Reconhecimento de Entidades Mencionadas 25
3.1.1 Conferências de REM 25
3.1.2 As Tarefas de REM 29
3.1.3 Resolução e Anotação de Topônimos 31
3.2 Medidas 32
3.2.1 Abrangência 34
3.2.2 Precisão 35
3.2.3 Medida-F 35
3.3 Aprendizado de Máquina 35
3.3.1 Aprendizado supervisionado 37
3.3.2 Aprendizado não supervisionado 38
3.3.3 Aprendizado Semi-Supervisionado 39
4 GeoNewsBR Dicionário Geográfico 41
4.1 Etapas de Pré-processamento 41
4.2 Coleta de Notícias 42
4.3 Criação de Regras para Logradouro 44
4.3.1 Divisão das Notícias em Sentenças 44
4.3.2 Avaliação das Sentenças 46
4.3.3 Janelar Sentenças 47
4.3.4 Identificar Classes Gramaticais 48
4.3.5 Criação das Regras e Definição dos Limiares 48
4.3.6 Criação das Regras 49
4.3.7 Definição dos Limiares 51
4.3.8 Base de Regras 56
4.4 Identificação de Janelas de Logradouro 57
4.4.1 Filtragem das Regras Válidas 57
4.4.2 Armazenar as Janelas de Logradouro Válidas 58
4.5 Aprendizado de Máquina 58
4.5.1 Stanford NER 59
4.5.2 Anotação das Entidades 60

4.5.3 Criação das Etapas de Treinamento/Validação e Testes 62
5 Formação e Visualização do GeoNewsBR 65
5.1 Formação do Dicionário Geográfico 65
5.2 A Geocodificação do Dicionário Geográfico 67
5.3 Apresentação das Informações Contidas no Dicionário Geográfico 68
5.4 Identificação de Endereços em Notícias da Internet 69
5.5 Visualização das Localidades no Mapa 70
6 Avaliação dos Resultados utilizando o GeoNewsBR 71
6.1 Formação dos Corpora 71
6.2 Primeira Execução dos Experimentos: Baseline 72
6.2.1 Experimento 1 – Influência do Sistema de Apoio 73
6.2.2 Experimento 2 - Notação de Entidades Padrão x BIO 73
6.3 Segunda Execução dos Experimentos: Corpus (C1) 74
6.3.1 Experimento 3 - Notação de Entidades com as Notações IO e BIO 75
6.4 Discussão dos Resultados 75
7 Conclusão 78
7.1 Trabalhos Futuros 79
Referências Bibliográficas 80

Índice de Figuras

Figura 1: Visão Geral de Georreferenciamento (GROVER *et al.*, 2010; TOBIN *et al.*, 2010) 12
Figura 2: Árvore de categorias do segundo HAREM – adaptado de (OLIVEIRA et al., 2008). 29
Figura 3: Relevância dos Termos 34
Figura 4: Fronteiras de Classificação dos Termos - Baseado de *Precision* e *Recall* (KONKOL, 2012). 34
Figura 5: Conjuntos A e B e suas intercessões 35
Figura 6: Estrutura Simplificada da Técnica de Aprendizado Supervisionado 38
Figura 7: Estrutura de Aprendizado Não Supervisionado, adaptado de (SCIKIT-LEARN DEVELOPERS, 2016). 39
Figura 8: Estrutura de Treinamento Semi-Supervisionado 40
Figura 9: Estrutura Inicial do Pré-processamento 42
Figura 10: Estrutura Inicial de Coleta de Notícias 44
Figura 11: Etapas de Criação de Regras de Logradouro 44
Figura 12: Divisão das Notícias em Sentenças 45
Figura 13: Criação das Sentenças Utilizando NTLK 46
Figura 14: Avaliação das Sentenças 46
Figura 15: Janelar Sentenças 47
Figura 16: Identificar Classes Gramaticais 48
Figura 17: Definir Limiares e Criação das Regras 49
Figura 18: Criação das Regras 50
Figura 19 - Processo de Criação de Regras para as Gramas 51
Figura 20: Definição dos Limiares 51
Figura 21: Intervalo de valores para os Limiares 52
Figura 22: Processo Combinação de Limiares 53
Figura 23: Relação de Generalidade por Grupos 54
Figura 24: Fluxo de Identificação de Janelas com indicador de Logradouro Válido 57
Figura 25: Filtragem das Regras Válidas 58
Figura 26: Janelas de Logradouro Válidas 58
Figura 27: Processo de Aprendizado de Máquina 59
Figura 28: Melhor Resultado de Parâmetros 63
Figura 29: Processo de Testes 63
Figura 30: Fluxo de Atividades Formação e Visualização do GeoNewsBR 65
Figura 31: Processo de Formação do Dicionário Geográfico 66
Figura 32: Processo de Geocodificação 68
Figura 33: Tela com as Informações do Dicionário Geográfico 68
Figura 34: Processo de Identificação de Endereços em Notícias da Internet 69
Figura 35: Identificação de Endereços Geográficos da Internet 69
Figura 36: Tela de Visualização no Mapa 70
Figura 37: Estrutura de Criação dos Corpos dos Experimentos 72
Figura 38: Comparativo Experimento 1 73
Figura 39: Comparativo Experimento 2 74
Figura 40: Comparativo Experimento 3 75
Figura 41: Percentual de Ganho Consolidado nos Experimentos 77

Índice de Tabelas

Tabela 1: Exemplo de Registro Simples em um Dicionário Geográfico 15
Tabela 2: Principais Dicionários Geográficos de Livre Acesso - adaptado de (BALLATORE; WILSON; BERTOLOTTO, 2013). 18
Tabela 3: Trabalhos Relacionados 22
Tabela 4: Dicionário de Palavras-Chave para Logradouro 24
Tabela 5: Categorias do segundo HAREM – adaptado de (OLIVEIRA et al., 2008). 28
Tabela 6: Exemplo de Categorização Ambígua de Entidades 31
Tabela 7: Notícia Anotada com a Notação IO 32
Tabela 8: Notícia Anotada com a Notação BIO 32
Tabela 9: Método Dedutivo x Indutivo 36
Tabela 10: Sentenças para Avaliação 46
Tabela 11: Exemplo de Janelamento 47
Tabela 12: Estrutura da Janela Refinada 50
Tabela 13: Exemplo de regras criadas para 1 janela 51
Tabela 14: Intervalo de Valores para os Limiares 52
Tabela 15: Exemplo de Valores de Limiares 53
Tabela 16: Relação de Peso por Grama 54
Tabela 17: Estabelecimento de Regras Válidas 55
Tabela 18: Relação das Melhores Regras 56
Tabela 19: Exemplificação das Base de Regras Geradas 56
Tabela 20: Uso da Máquina de Aprendizado 59
Tabela 21: Representação IO 61
Tabela 22: Representação BIO 61
Tabela 23: Parâmetros Utilizados 62
Tabela 24: Demais Categorias de Entidades 66
Tabela 25: Resultados do Experimento 1 73
Tabela 26: Resultados do Experimento 2 74
Tabela 27: Resultados do Experimento 3 com Anotações IO e BIO 75

1 Introdução

Devido às diversas mudanças que acontecem diariamente na configuração física das edificações e elementos de geográficos no território nacional, aliada ao avanço de políticas de incentivo a expansão, modernização e construção de novas edificações ocorridas nos últimos anos, principalmente nos países em desenvolvimento, deflagrou uma necessidade crescente de informações atualizadas de elementos de geográficos para as atividades de planejamento e tomada de decisão.

As entidades de planejamento urbano enfrentam a dificuldade em obter de forma mais atualizada, detalhada, de fácil acesso e em português, informações consolidadas sobre as evoluções de novas edificações e construções ocorridas no território nacional, para atualizar suas bases de informações e dar continuidade nos processos relativos a cada órgão, principalmente as tarefas relacionadas a tomada de decisões e planejamento geográfico, urbano, social ou comportamental.

Atualmente a maneira mais atual para a identificação de novos elementos físicos no espaço geográfico é feita principalmente através da figura do *reambulador*, uma pessoa ou organização que visita determinada localidade e realiza pesquisas e medições, entre outras técnicas, para identificar informações relativas e caracterizar determinados elemento geográfico e sua respectiva localização.

No Brasil a reambulação é tardia e não atende as constantes atualizações no território nacional, gerando desconhecimento, falta de informação atualizada e útil de grande valor cultural e econômico. De maneira geral, esse trabalho é de difícil execução, economicamente custoso, devido aos recursos pessoais, temporais e tecnológicos necessários para sua realização.

A notícia como fonte de insumo é importante nas atividades e tarefas de identificação de localidades geográficas de baixa granularidade para o apoio na criação e atualização automática de dicionários geográficos mais específicos. Esses dicionários geográficos disponibilizam em uma base de dados informações mais adequadas e precisas, além de reduzir os custos e aceleram a execução no fornecimento de informações com mais especificidade.

1.1 Contextualização do Problema

A reambulação é parte essencial do processo de formação e entendimento do território, auxiliando principalmente nas atividades pertinentes ao mapeamento geográfico em suas subdivisões político-administrativas. Sua atividade consiste na

verificação de informações sobre toponímia[1] das localidades geográficas, além de sanar dúvidas sobre feições geográficas que não puderam ser plenamente definidas a partir de outras fontes de informação, como fotos aéreas e cartas do mapeamento sistemático do Brasil.

O IBGE (2005) define a reambulação como uma pesquisa de campo com o objetivo de elaborar e descrever os elementos geográficos de origem naturais como montanhas, morros, mares, dentro outros. Os Outros elementos não naturais, são elementos artificiais produzidos pelo homem como casas, edifícios, construções, pontes, dentre outros.

A reambulação consiste nas atividades de coletar, confirmar ou destacar informações sobre determinada localidade territorial. Possui o objetivo de criar, atualizar ou incrementar as cartas (mapas) de forma que identifique o lugar e suas características (SANTOS, 2008). Ainda segundo esse autor as atividades ligadas a reambulação "podem contribuir para o fomento de áreas como: o comércio e negócios, como as indicações geográficas; censos demográficos e estatísticas nacionais; direitos de propriedade e cadastro; planejamento urbano e regional; gestão ambiental; comunicação rápida e eficiente nos socorros em desastres naturais, prontidão em situações de emergência e recepção de ações humanitárias, estratégias de segurança e missões de paz; produção de produtos didáticos como mapas e atlas; navegação terrestre, marítima; turismo; diversos aspectos históricos e culturais locais; aspectos lexicográficos e linguísticos; e nas pesquisas acadêmicas."

Segundo o IBGE (2005), para executar o processo de reambulação, é necessário um conjunto de ferramentas auxiliares, que apoiem as atividades desempenhadas destinadas à identificação, localização, denominação e esclarecimentos de elementos geográficos.

O último trabalho de reambulação do IBGE foi realizado no ano de 2012. Segundo entrevista feita entre os analistas do IBGE no Rio de Janeiro. Para o ano de 2016 está previsto o início do ciclo de novas atividades de reambulação, contudo devido às projeções negativas na economia para esse ano, esse processo está em dificuldade.

Essa pesquisa consiste na identificação de elementos geográficos de baixa granularidade, pois esse é parte integrante do processo de criação e atualização de dicionários geográficos que possuem maior nível de detalhamento geográfico.

[1] Toponímia é consiste no estudo da designação dos lugares pelos seus nomes correspondentes.

A escolha da pesquisa na tarefa de identificação de endereço geográfico de baixa granularidade em notícias como objeto desse trabalho deve-se também ao fato de instituições como o IBGE necessitarem de informações mais precisas e assertivas sobre os novos elementos encontrados para o povoamento de suas bases contínuas[2] ou estruturadas. Outro fator de escolha deve-se ao fato de que as notícias são fontes atualizadas de informações com elevado grau de credibilidade.

A identificação de endereços geográficos de baixa granularidade é importante para apoiar a criação e atualização de dicionários geográficos de maneira a retratar a realidade do território nacional devido às diversas e constantes modificações ocorridas no território brasileiro nos últimos anos.

Diversas ferramentas, metodologias e tecnologias são empregadas em auxilio às atividades de reambulação. Contudo, devido a necessidade de obter informações atualizadas, a utilização de dicionários geográficos baseados em notícias aparece como uma ferramenta informacional de apoio para execução das atividades relacionadas a reambulação.

De maneira geral os dicionários geográficos, principalmente no Brasil, não incluem informações geográficas específicas e tão pouco incluem instalações políticas urbanas de menor granularidade de escala como escolas, hospitais, bibliotecas, atrações turísticas e outros pontos de interesse.

Em detrimento de fatores como tempo elevado de elaboração, alto grau de esforço humano e tecnológico e custo elevado na execução das atividades, os dicionários geográficos são difíceis de criar e atualizar com um nível de periodicidade que acompanhe as mudanças no cenário atual. A América do Norte e a Europa possuem uma cobertura e atualização de dados em diversos dicionários geográficos que maximizam a utilização e entendimento do seu território. Contudo no Brasil, as características de algumas instalações geográficas são irregulares, inexistentes ou de atualização tardia, principalmente nas áreas urbanas. Atualmente o uso de tecnologias e ferramentas de apoio tecnologias é extremamente relevante nesse contexto (MACHADO *et al.,* 2011).

Embora os atuais dicionários geográficos apresentem informações detalhadas de diversos países, a quantidade e o nível de detalhes sobre informações locais são pobres

[2] A base cartográfica contínua é um conjunto de dados, estruturados em Banco de Dados Geoespaciais, com a possibilidade de recuperação dos elementos geográficos existentes do país.

ou insuficientes para a tomada de decisão ou para resolver alguns problemas pontuais (GELERNTER et al., 2013).

Nesse trabalho foi proposto um conjunto de etapas de pré-processamento de notícias, para criar um dicionário de regras que delimite as notícias com maior nível de aceitabilidade em possuir elementos geográficos que caracterize a presença de um ou mais logradouros. Isto proporciona a tarefa de aprendizado de máquina, notícias em um formato mais adequado, que permita obter melhores resultados nas métricas utilizadas para identificar os elementos geográficos de baixa granularidade necessários para a criação de dicionário geográfico que apoie nas atividades ligadas as tarefas de reambulação.

1.2 Delimitação

O escopo desse trabalho consiste na identificação de elementos geográficos de baixa granularidade em notícias em português do Brasil, através da formação e utilização de regras gramaticais, para apoiar a criação e atualização de dicionários geográficos em língua portuguesa.

A tarefa de identificar informações geográficas das variadas fontes existentes de notícias é uma tarefa complexa e por isso exige uma área de estudos dedicada somente a ela. Uma notícia pode ter diferentes formatos para apresentar seu conteúdo geográfico, tais como a representação textual explícita de um logradouro, cidade, estado, país, dentre outros, ou com utilização de elementos específicos de delimitação geográfica como coordenadas geográficas ou o CEP, por exemplo.

Neste trabalho, utilizamos apenas as notícias que possuam a presença explícita de indicadores de logradouro em seu conteúdo ou corpo, semelhante ao trabalho de GOUVÊA (2008) e (MACHADO *et al.,* 2011). Este trabalho utilizou como fonte de dados para as etapas de treinamento, validação e teste, um corpus especialmente criado para os experimentos, em vez de utilizar corpus ou corpora pré-existentes como, por exemplo, o corpus de avaliação da HAREM. Com essa restrição espera-se aproveitar as características intrínsecas a esse domínio.

Em meio ao conhecimento prévio de que existe uma grande diversidade entre os fatores que indicam localidade e as novas construções e edificações nas notícias, nesse trabalho as palavras chave de pesquisa são previamente definidas em um dicionário de palavras chave (DPC) para logradouro.

1.3 Objetivo

O objetivo deste trabalho é a realização de estudo e aplicação de metodologias sobre a identificação de endereços geográficos em textos, considerando sua localização geográfica detalhada, baixa granularidade, utilizando notícias presentes em publicações oficiais, sites e diários oficiais, redigidas na língua portuguesa, como insumo para apoiar a criação de um dicionário geográfico que possa ser utilizado nas tarefas relacionadas à atividade de reambulação.

Para a realização dos objetivos propostos, esse trabalho foi dividido em 3 etapas:

1. O pré-processamento de notícias:
    - Divisão das notícias em sentenças;
    - Avaliação das sentenças;
    - Janelamento das sentenças;
    - Identificar classes gramaticais;
    - Criação das regras e definição dos limiares.
2. Reconhecimento de Entidades Mencionadas
    - Utilização da máquina de aprendizado supervisionado.
3. Formação e Visualização do Dicionário Geográfico
    - Formação dos componentes necessários para a criação de um dicionário geográfico;
    - Visualização em uma ferramenta desenvolvida para comprovar, auxiliar e apoiar na identificação de endereços geográficos e disponibilizar o acesso as informações contidas no dicionário geográfico.

1.4 Organização do Trabalho

Este trabalho está organizado em 7 capítulos, organizados da seguinte forma: No capítulo 2 é apresentada a revisão literária realizada durante esse trabalho e os trabalhos relacionados. Essa revisão fornece a base necessária para que o leitor tenha uma visão geral sobre as técnicas utilizadas e o contexto dos problemas propostos.

No capítulo 3 é apresentado os elementos relacionados ao reconhecimento de entidades mencionadas, métricas de avaliação e aprendizado de máquina.

No capítulo 4 é detalhada a tarefa de pré-processamento das notícias, utilizado para atender aos objetos da pesquisa e as principais tarefas que compõem essa atividade de reconhecimento de entidades.

No capítulo 5 é apresentado a ferramenta criada neste trabalho, denominada de GeoNewsBR que permite a comprovação da eficácia das tarefas de pré-processamento e aprendizado de máquina.

No capítulo 6 são detalhados os experimentos realizados na concepção do sistema de Reconhecimento de Entidades Nomeadas e apresenta discursões sobre os resultados deste trabalho.

No capítulo 7 são apresentados os resultados obtidos nos experimentos, as conclusões sobre este trabalho e as sugestões propostas para trabalhos futuros.

2 Referencial Teórico

Esse capítulo apresenta uma revisão dos principais conceitos relacionados à extração de informação e dicionário geográfico. Além disso, descreve conceitos geográficos, contribuições relevantes e trabalhos relacionados que influenciaram no desenvolvimento dessa dissertação, já que para contextualizar o problema alvo do trabalho, é necessário compreender o processo para a identificação do contexto geográfico em textos. O capítulo busca analisar a área de extração de informações, sistema de informações geográficas e dicionários geográficos.

2.1 Extração da Informação

As informações estão frequentemente disponíveis em artigos de jornais, revistas, sites e arquivos de texto. Os sistemas de Extração da Informação (EI) auxiliam na tarefa de coletar dados em textos com relevância e valor para a temática abordada. A EI inicia suas atividades com a coleção de textos coletados, em seguida, transforma esses dados em informação de compreensão (COWIE; LEHNERT, 1996).

Dentre as diversas atribuições, a área de EI está preocupada em identificar conteúdo a partir de textos não estruturados ou totalmente organizados (SILVA; BARROS; PRUDÊNCIO, 2005). Essa tarefa envolve a identificação de entidades, relacionamentos e contextos específicos, com o propósito de refinar a coleção total de textos para representações mais entendíveis de acordo com as necessidades humanas. A EI pode ser utilizada em diferentes domínios de aplicação, sendo que cada um possui particularidades. No caso da extração de entidades nomeadas especificas, palavras chave e de articulação devem ser inseridas com participação humana para produzir resultados mais adequados para cada cenário (ELLOUMI *et al.*, 2013).

Embora a área de EI utilize ferramentas para o reconhecimento de elementos humanos presentes nas mais diversas formas de expressar informações, seu potencial ainda não foi totalmente utilizado, principalmente em fontes de notícias governamentais, que muitas vezes não despertam o interesse de usuários comuns e pesquisadores, devido a fatores políticos, culturais ou simplesmente pela falta de conhecimento da existência dessas fontes.

Devido à grande quantidade de informação e à maneira como utiliza os mais diversos sistemas de informação, um sistema de EI é capaz de transformar os diversos dados intrínsecos em informação com algum valor agregado para determinado cenário (DODDINGTON *et al.*, 2004).

Segundo Sawaragi (2007), a EI se refere à atividade de extração automática de informações em textos, como entidades, relações e atributos. Sua estrutura possui por essência o Processamento de Linguagem Natural (PLN) e abrange os temas como a aprendizagem de máquina, recuperação de informação, banco de dados, web, análise de documentos, permitindo que consultas complexas possam ser executadas em dados não estruturados.

Para alcançar os objetivos necessários na elaboração desse trabalho a EI atua principalmente na subárea de Reconhecimento de Entidades Mencionada (REM), na identificação e extração de entidades a partir de fontes de dados textuais, em particular, elementos textuais das notícias em linguagem natural e na tarefa de transformação desses dados em uma informação mais adequada para gerir informações de conteúdo relevante a determinados contextos (ELLOUMI et al., 2013; WEIKUM et al., 2009).

Os tipos de informações extraídas de textos podem apresentar variação estrutural em detrimento como sua forma de apresentação em suas respectivas fontes. Diferentes tarefas são necessárias para aplicar a EI (ELLOUMI *et al.*, 2013).

2.1.1 Extração de Informação Geográfica em Noticias

Em seu trabalho, Brisaboa (2010) observou que nas últimas décadas, houve um forte crescimento no número de referências geográficas presentes em notícias, páginas da internet, e outros elementos relevantes de informação.

Embora seja comum a presença de informação geográfica em documentos de textos, raramente as referências geográficas são extraídas em sistemas de recuperação de informação. Poucos algoritmos são projetados considerando a natureza espacial das referências geográficas embutidas dentro de textos e em suas aplicações para outros sistemas (BRISABOA *et al.*, 2010).

Ao longo dos anos, as áreas de EI e de sistemas de informação geográfica, foram exaustivamente estudadas sob as mais diversas perspectivas, e mais recentemente, essas áreas foram transformadas em uma única área de aplicação denominada de Recuperação de Informação Geográfica, em inglês, *Geographic Information Retrieval* (GIR). A GIR uniu as vertentes e perspectivas dessas duas áreas em um único âmbito de trabalho, combinando as melhores práticas e perspectivas com o objetivo de fomentar suas pesquisas (BRISABOA et al., 2010).

Um dos principais objetivos da GIR é utilizar a tarefa de extração de informações para a identificação de referências geográficas contidas em seu conteúdo

textual. De maneira geral, a GIR busca extrair nomes de localidades utilizando textos "livres" e ausentes de ambiguidades semânticas, com o objetivo final de produzir insumos e resultados disponibilizados posteriormente em metadados (OVERELL; RÜGER, 2007).

As diversas fontes de dados como notícias, sites, blogs, páginas da Wikipédia e redes sociais podem conter informações geográficas, que permitem que estas sejam georreferenciadas de forma rápida e com alta disponibilidade ao acesso a essas novas informações (LUO et al., 2011; TEITLER *et al.*, 2008).

As fontes de conteúdo mais importantes para obter um contexto geográfico são os conteúdos baseado em textos, presentes em notícias e páginas na internet (LUO et al., 2010). Os sites de notícias mais populares entre os internautas como o Google News, Yahoo! e Globo possuem apenas um conhecimento simplório da importância e implicação que as informações geográficas exercem sobre as notícias apresentadas (TEITLER *et al.*, 2008).

Ainda segundo Teitler et al. (2008), na tarefa de extração de informações geográficas em notícias existem três grupos de extração de informação que podem fornecer os recursos necessários para desempenhar as atividades ligadas ao georreferenciamento e geocodificação dos textos, são elas:

1. Baseado na localização geográfica do editor;
2. Baseado em informações geográficas do conteúdo do texto;
3. Baseado na localização dos leitores.

2.2 Sistema de Informações Geográficas

Durante a década de 1980, houve um aumento no desenvolvimento e disseminação dos Sistemas de Informações Geográficas (SIG), tanto no setor privado quanto no público motivado pelo interesse de órgãos governamentais. Iniciando uma demanda crescente por dados espaciais digitais, enriquecido por informações de modelos tridimensionais, textos, fotos e multimídia em geral. (SALAMUNI; STELLFELD, 2001).

Devido ao surgimento de novas tecnologias e ferramentas de apoio às atividades de informações geográficas, a academia pode utilizar as novidades tecnológicas com simplicidade e eficiência para facilitar o acesso a essas informações (SCHARL, 2007).

Conquistas como a exploração do espaço foram fundamentais para as pessoas melhorarem a compreensão da importância ao utilizar e gerenciar informações

georreferenciadas (SCHARL, 2007). A evolução tecnológica, nas áreas de aviação, fotografia e computação aliado as novas técnicas de processamento e aos novos sensores, permitiram avanços constantes nas áreas ligadas ao referenciamento e identificação geográficas mapeamento de feições geográficas[3]. (GRIGIO, 2003; LUCHIARI, 2011).

Muitas das atividades de identificação e mapeamento foram desenvolvidas decorrentes da colaboração de usuários com as mais variadas fontes de dados sem, contudo, levar em conta o fato da massa de dados ser proveniente de diferentes pessoas o que a torna propensa a uma variedade de ruídos em relação a qualidade dos dados apresentados. Além disso, a confiabilidade dos dados pode se tornar questionável, pois pode haver fatores políticos, regionais ou pessoais que impactam os resultados esperados (LUO et al., 2010).

A utilização de dados espaciais, definidos como dados que possuem componentes espaciais, torna possível uma melhor compreensão do espaço geográfico, e podem ser. Estes componentes espaciais podem ser representações da superfície terrestre como construções, acidentes geográficos, estruturas hidrográficas, entre outros e estão relacionados com seu posicionamento no espaço geográfico.

Os dados geoespaciais, incluem os mais variados tipos de coleções de dados georreferenciados ou metadados, como textos, imagens de satélite, informações de dispositivos móveis, dados de localização de pessoas, objetos e estruturas no globo terrestre utilizando redes sociais, dados demográficos entre outros (WEBER et al., 1999).

Na tarefa de reconhecimento de elementos geográficos nos textos, podem existir dois grupos de contextos geográficos, os mais óbvios de serem determinados (por exemplo, logradouro, localização, latitude e longitude) e os não tão óbvios, como textos distintos apresentados em arquivos que não possuem uma relação semântica interessante, e esses devem ser tratados diferentemente para que possibilite sua utilização na tarefa (LUO *et al.*, 2010).

2.2.1 Geoprocessamento

O geoprocessamento é um conjunto de técnicas que permitem analisar, manipular e gerir informações geoespaciais (geometria, textos, fotos, sistema de posicionamento

[3] Feições geográficas são os elementos que possuem coordenadas ligadas ao planeta terra. Por exemplo, estados, rios, rodovias, praças, escolas, etc.

global, dentre outros) e informações de áreas adjacentes à geográfica para associar informações de conteúdo relevante em determinados contextos (CARVALHO; LEITE, 2009).

Segundo Silva (2000), o geoprocessamento se refere "ao conjunto de técnicas computacionais, utilizando bases de dados georreferenciadas, para transformar as informações em conteúdo relevante para determinado cenário".

De acordo Zhao et al.,(2012), o geoprocessamento pode ser definido como um conjunto de técnicas para manipular dados espaciais através da utilização de arquiteturas SOA[4], padrões e ferramentas para atender os requisitos de análise, modelos e metodologias específicas para suportar ferramentas de mineração de dados avançados, e atualmente em elementos da internet.

O desenvolvimento do paradigma de geoprocessamento foi influenciado pelos avanços na tecnologia da informação e suas normativas técnicas. O geoprocessamento é composto dos seguintes tipos de processamento (ZHAO; FOERSTER; YUE, 2012):

- Processamento espacial;
- Processamento temático;
- Processamento temporal;
- Processamento de metadados.

A acessibilidade e a disponibilidade dos recursos de geoprocessamento buscam obter melhores resultados na aplicação de dados geoespaciais em vários domínios e contribuir para o incremento do conhecimento geoespacial. Atualmente, diversos esforços estão concentrados no geoprocessamento, contudo a falta de disponibilidade das informações para executar essas tarefas ainda gera atrasos nas evoluções no campo de pesquisa nessa área (LUO *et al.*, 2010).

2.2.2 Georreferenciamento

Na área de EI, o georreferenciamento pode ser representado por um conjunto de métricas, ferramentas e técnicas com o objetivo de identificar nomes de lugares em documentos textuais, utilizando as atividades de extração de informação e posteriormente a atribuição de coordenadas, geralmente latitude e longitude, que identifique essa localidade no globo terrestre (TOBIN *et al.*, 2010).

[4] SOA é arquitetura orientada a serviço, muito utilizado em webservices.

Especialmente no contexto de EI, o georreferenciamento está subdividido em dois componentes principais: (i) geoidentificador, em inglês *geotagger*, responsável principalmente pelo reconhecimento do nome da localidade, e (ii) geosolucionador, em inglês *georesolver*, responsável pela atribuição de coordenadas.

A Figura 1 apresenta a estrutura básica proposta dos componentes envolvidos nas tarefas de georreferenciamento: geoidentificador e geosolucionador (GROVER *et al.*, 2010; TOBIN *et al.*, 2010*).*

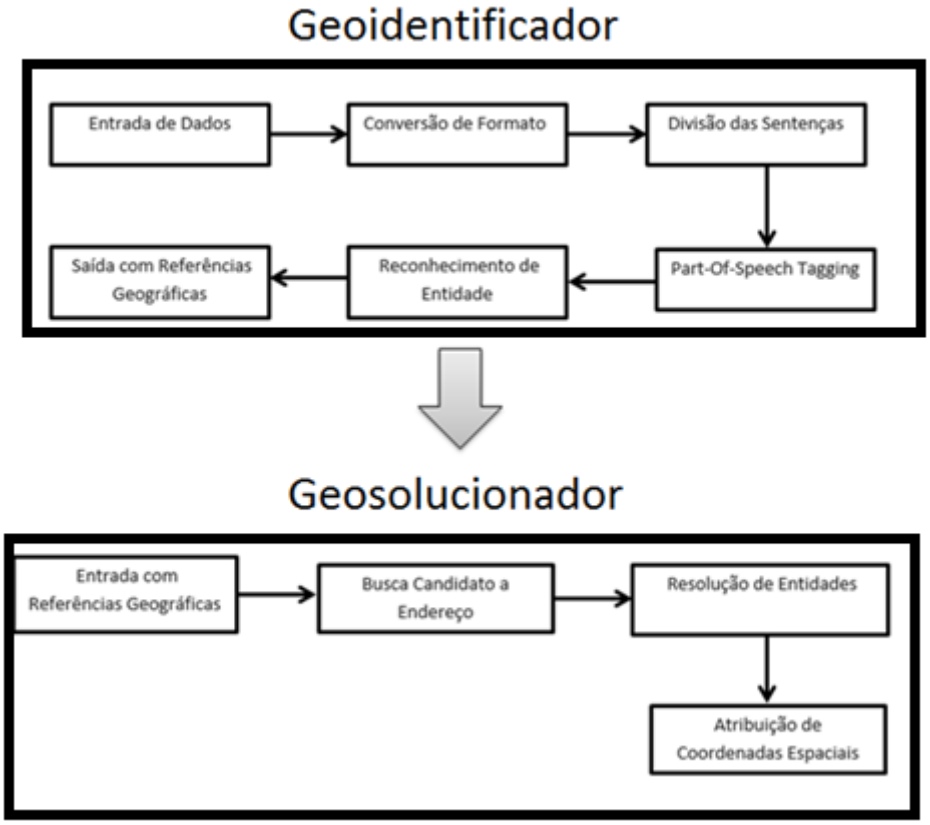

Figura 1: Visão Geral de Georreferenciamento (GROVER *et al.*, 2010; TOBIN *et al.*, 2010)

Nos trabalhos de Clough (2005) e Gouvêa et al. (2008), as etapas de georreferenciamento são organizadas em dois tópicos, identificador de referências geográficas e atribuição ou geocodificação de coordenadas espaciais. Embora os nomes das etapas sejam diferentes do processo apresentado por (TOBIN *et al.*, 2010) (GROVER et al., 2010), as tarefas pertinentes são semelhantes.

Nas seções posteriores são apresentados mais detalhes sobre as etapas de identificador de referências geográficas e geocodificação de coordenadas espaciais

2.2.3 Identificador de Referências Geográficas

A tarefa de identificação de referências geográficas, em inglês *geoparsing*, pode ser considerada como uma aplicação específica da tarefa de REM, que está preocupada com a detecção automática de elementos.

A literatura apresenta diversas nomenclaturas para o termo "identificação de referências geográficas", em inglês *geoparser*, *geotagger* ou *georecognition*, e o termo mais conhecido em português é "reconhecimento de topônimos geográficos" ou simplesmente "reconhecimento de topônimos" (LEIDNER; LIEBERMAN, 2011).

Para Densham and Reid (2003), a identificação de referências geográficas em texto, inicia-se no processo de divisão da notícia em unidades menores como sentenças ou janelas, com o objetivo de tratar as unidades como "candidatas" a possuir elementos que referencie um endereço de uma localidade geográfica.

Diversidades são encontradas na identificação de topônimos geográficos, como a resolução de ambiguidade denominada "geo ambiguidade" que ocorre quando uma determinada expressão dentro de uma sentença se refere a dois topônimos diferentes, por exemplo, a palavra "Flamengo" pode-se referir ao clube de regatas Flamengo, ao bairro do Flamengo ou ao nome da Praia do Flamengo. Outros desafios encontrados são a forma de lidar com erros ortográficos na descrição dos nomes próprios, o uso de acrônimos e o uso coloquial de elementos (LEIDNER; LIEBERMAN, 2011).

Em função da tarefa de reconhecimento de entidades mencionadas estar fortemente relacionada com a identificação de referências geográficas, os algoritmos, heurísticas e métodos probabilísticos associados à identificação de elementos geográficos em textos, apresentam-se como alternativas para a solução do problema de identificação, principalmente em contextos com acrônimos e nomes ambíguos presentes em textos não estruturados (CLOUGH, 2005).

Segundo Rupp et al., (2013), as principais tarefas relacionadas com a identificação de elementos geográficos em textos, consistem na identificação dos nomes de lugares e resolução de ambiguidade, tarefas desempenhadas, principalmente, pelas atividades dos sistemas de REM.

2.2.4 A Geocodificação de Elementos Geográficos

Segundo GOLDBERG et al. (2007), o termo geocodificação, em inglês, *geocoding*, significa atribuir um código geográfico para uma localização, de maneira

que com auxílio de métricas de conversão seja possível identificar a relação de localidade entre um ponto a outro no globo terrestre.

Para LEIDNER; LIEBERMAN (2011), a geocodificação é o processo de conversão de endereços em coordenadas geográficas, que geralmente representam coordenadas geográficas de acordo com sua respectiva latitude e longitude.

Segundo MONCLA et al. (2014), a geocodificação pode ser definida como a associação de um topônimo a uma determinada coordenada, latitude e/ou longitude, considerando os fatores de ambiguidade toponímica.

Em geral, as informações disponíveis em dicionários geográficos no contexto de ruas, cidades, estados e países, possuem pouca ou baixa qualidade de informações em relação a suas coordenadas. Em contextos em que o nível de granularidade da localização é menor, como na identificação de elementos geográficos especificados em logradouros, a qualidade da geocodificação e atribuição do elemento geográfico correto são de grande importância (MONCLA et al., 2014).

2.3 Dicionários Geográficos

Os dicionários geográficos, em inglês, *gazetteers* "(...) são fontes de informação organizada sobre determinados locais e podem contribuir decisivamente com a solução de problemas de recuperação de informações geográficas" (MACHADO *et al.*, 2011). Os dicionários geográficos são componentes que compõem os SIG (GOODCHILD, 2009).

Um dicionário geográfico pode ser considerado um ativo de importante valor para auxiliar na tarefa de identificação e atribuição de conteúdo geográficos como nomes de localidades e lugares, presentes em documentos textuais e associar coordenadas geográficas para esses documentos (SOUZA *et al.*, 2005).

O processo de criação das bases de informações de um dicionário geográfico é uma tarefa que necessita de elevado investimento financeiro e tende a ser uma tarefa de difícil criação e atualização das informações contidas por parte das organizações mantenedoras. Os dicionários geográficos possuem cobertura global, embora que apresentem pouca ou nenhuma informação sobre o Brasil, e em sua maioria não possui o nível de granularidade adequado para o georreferenciamento de baixa granularidade de edificações e construções.

Embora existam grandes e famosos dicionários geográficos, no contexto do Brasil, estes são falhos em relação a cobertura das informações. Informações relevantes

de baixa granularidade geográfica como nomes, escolas, hospitais, bibliotecas e outras edificações e construções (GOUVÊA *et al.*, 2008).

Tradicionalmente, os dicionários geográficos não apresentam em tempo adequado as mudanças geográficas que ocorrem no espaço geográfico, gerando uma distanciação entre a realidade existente no espaço geográfico da apresentada no dicionário geográfico.

A importância dos dicionários geográficos está intimamente ligada com sua utilização nas mais diversas áreas, por exemplo, na desambiguação de nomes, na reambulação de itens geográficos, em sistemas de classificação de posição por relevância ou importância, e na identificação de conteúdo geográficos.

A presença de coordenadas geográficas que representem uma determinada localidade é em geral uma exigência para a criação de um dicionário geográfico. Apesar da existência de diversificadas estruturas, em geral, os dicionário geográficos apresentam basicamente a seguinte estrutura (HILL, 2000):

- Nome do elemento (topônimo);
- Localização em coordenadas (latitude e longitude);
- Tipo.

A Tabela 1 apresenta um trecho de exemplo de um registro simples em um dicionário geográfico:

Tabela 1: Exemplo de Registro Simples em um Dicionário Geográfico

Nome do elemento	**Localização**	**Tipo**
Miguel Couto	Latitude -22.977863; Longitude -43.223855	Hospital

Em geral, o nome do elemento é o atributo que define qual o "tipo" do termo baseado na predefinição criada pelo autor do dicionário, por exemplo, nas categorias definidas pelo dicionário da Alexandria Digital Library (2014) inclui áreas administrativas, recursos hídricos, elementos de relevo e físico. Enquanto no dicionário GeoNames (2015), a representação do tipo da entidade é dividida em 9 classes principais, e utiliza como insumo dados públicos do governo dos Estados Unidos e a artigos extraídos da Wikipédia (2015).

A presença de coordenadas geográficas que representa uma determinada localidade é em geral uma exigência para a criação de um dicionário geográfico. A utilização de coordenadas geográficas em dicionário é necessária para definir a localidade aproximada do elemento. Contudo, devido a problemas como ambiguidade semântica na identificação da localização, localidades com o mesmo nome e

localizações de difícil delimitação, a definição de coordenadas pode não ser precisas o suficiente, sendo impossíveis de serem identificadas automaticamente sem a intervenção humana.

Devido ao alto custo e tempo empenhado no processo de criação e atualização de dicionários geográficos, órgãos públicos e privados utilizam, em sua maioria, recursos não atualizados ou imprecisos para o apoio de suas atividades. O processo de criação e atualização dos dicionários geográficos é financeiramente caro e difícil de ser atualizado e mantido, o que em muitos casos culmina com a sua construção manual (TORAL; MUNOZ, 2006) (POPESCU; GREFENSTETTE; MOËLLIC, 2008).

Existem dois tipos de dicionários geográficos: os *trigger gazetteers* e os dicionários de entidades definidas. Os *trigger gazetteers* são baseados na existência de palavras-chave encontradas no texto para a delimitação e criação de grupos ou categorias. Por outro lado, os dicionários de entidades definidas contêm suas próprias entidades, geralmente identificados gramaticalmente como nomes próprios. Por exemplo, no texto em que aparece a "Brasil", este termo por si próprio pode ser a entidade (TORAL; MUNOZ, 2006).

Em geral, os dicionários geográficos não possuem uma grande quantidade de elementos, sobretudo no Brasil. A falta de elementos é agravada, principalmente no tocante de elementos de menor granularidade, como elementos de edificações e construções públicas das cidades. Algumas dessas dificuldades podem ser superadas pela utilização de serviços de geocodificação disponíveis no mercado.

2.3.1 Principais Dicionários Geográficos

Nassã sessão são apresentados alguns dicionários importantes, dentre eles estão relacionados Geonames, WordNet, Wikipedia, Getty Thesaurus of Geographic Names e Alexandria. A seguir, são descritos com mais quantidade detalhes:

O dicionário geográfico *GeoNames*, combina diversas fontes, utilizando principalmente as informações do Governo dos Estados Unidos da América, e da multidão que colabora *online*. Sua base de informações contém mais de 10 milhões topônimos presentes nas mais diversas categorias e possui poucas informações sobre o Brasil. No tocante de elementos de localização, seus valores estão organizados hierarquicamente de forma que é possível classificá-los em diferentes níveis de granularidade continentes, países, regiões, estado, cidade, etc. (BALLATORE; WILSON; BERTOLOTTO, 2013; GEONAMES, 2006).

O dicionário Alexandria Digital Library (2014). é composto de elementos geográficos de duas agências federais dos EUA (*US National Imagery e Mapping Agency*), além do serviço geológico dos USA (HILL, 2000).

O dicionário geográfico *Getty Thesaurus of Geographic Names* (2015) é um dos dicionários geográficos mais conhecidos. Sua base é composta de milhões de nomes de lugares em diferentes linguagens, representando de forma hierárquica (mundo, continente, país, estado, região, cidade, etc.) e apresenta contexto em diferentes tipos de entidades.

O Wordnet (2015) possui uma área específica para as informações geográficas em seu dicionário, que permite o acesso as informações do conteúdo do seu dicionário geográfico (GIUNCHIGLIA et al., 2010). O conjunto de dados da Wordnet está disponibilizado no formato RDF[5] para utilização de seus recursos utilizando *webservices* e sua ontologia possui uma alta qualidade, no correspondente aos conceitos e relações geográficas, devido ao uso de especialistas que participam do processo de criação e manutenção do dicionário (BALLATORE; WILSON; BERTOLOTTO, 2013).

O Wordnet possui um projeto que disponibiliza informações em língua portuguesa, denominado de WordnetPT (2015), contém cerca de 19000 expressões, agrupadas em detrimento do valor semântico dos termos. É um dicionário utilizando para recuperação e identificação de indicadores geográficos, sendo utilizando para resolução de problemas de ambiguidade. Por exemplo, identificar o termo e determinar qual o Tipo da localidade (se é um país ou estado) como no caso da expressão “Rio de Janeiro” (PAIVA; RADEMAKER; MELO, 2012).

A Wikipédia, com seu aspecto principal de ser composta por dados colaborativos da multidão, tem sido utilizada como dicionário de geográfico para a identificação e reconhecimento de localidades geográficas e na geração de ontologias, com trabalhos em português e no contexto brasileiro. (ODON DE ALENCAR; DAVIS; GONÇALVES, 2010). A Tabela 2 apresenta os dicionários geográficos mais importantes de livre acesso:

[5] RDF ou *Resource Description Framework* é uma linguagem para representar informação na Internet.

Tabela 2: Principais Dicionários Geográficos de Livre Acesso - adaptado de (BALLATORE; WILSON; BERTOLOTTO, 2013).

Nome do Dicionário	Ano de Disponibilização	Estrutura e Descrição	Corpora	Formato disponibilizado de saída
WordNet	1985	Rede semântica, dicionário, enciclopédias; 117 mil palavras e seus sinônimos.	Corpus próprio, inicialmente criado por especialistas.	OWL/RDF
Conceptnet	2000	Ontologia, Rede Semântica; 700.000 sentenças de linguagem natural.	Wikipedia, WordNet e outros.	JSON
Wikipédia	2001	Rede semântica, dicionário, enciclopédias; Mais de 3.9 milhões de artigos.	Multidão de colaboradores.	XML
OpenStreetMap	2004	Mapa de vetor, dicionário geográfico; Mais de 1.2 bilhões de elementos.	Multidão de colaboradores e *datasets* de GIS.	XML
GeoNames	2006	650 classes e mais de 10 milhões de topônimos	Outros dicionários geográficos, Wikipédia e a Multidão de colaboradores.	OWL/RDF
Yago	2006	Ontologia, Rede Semântica; Mais de 10 Milhões de entidades.	Wikipédia, GeoNames e WordNet.	RDF
DBpedia	2007	Ontologia, Rede Semântica; 320 classes.	Wikipedia, WordNet e outros dicionários.	OWL/RDF
Freebase	2007	Ontologia, base de conhecimento derivada da multidão;	Multidão de colaboradores.	TXT
LinkedGeoData	2009	Dicionário geográfico, mais de 380 mil entidades geográficas.	Informações da base do OpenStreetMap.	RDF
Geo WordNet	2010	Rede semântica, enciclopédias, dicionário geográfico; 330 classes, 3.6 milhões de entidades.	WordNet, GeoNames e MultiWordNet.	RDF

2.3.2 Trabalhos relacionados

Na tarefa de criação, atualização e manutenção automática de dicionários geográficos utilizando a extração de informações geográficas presentes em texto, neste caso em notícias, a literatura apresenta diversificadas soluções, utilizando tarefas e métodos como: o aprendizado supervisionado, não supervisionado ou semi-supervisionado, REM no texto, heurísticas de Inteligência Artificial, dentre outras.

Dentre as diversas áreas abordadas nesse trabalho, nessa seção são apresentados os principais trabalhos relacionados à identificação de elementos geográficos

relacionados com a criação e atualização de dicionários geográficos. Esses trabalhos estão sumarizados na Tabela 3 e descritos a seguir.

O trabalho de Odon de Alencar; Davis; Gonlçaves (2010) propõem a identificação de localidades geográficas no Brasil, utilizando como fonte de dados, notícias de jornais, mais especificamente, jornais do estado de Minas Gerais. A ontologia proposta nesse trabalho é iniciada a partir de um conjunto específico de localidades, é realizada a busca de entidades em artigos da Wikipédia, principalmente no título e subtítulo dos artigos. Os termos encontrados são indicados por uma medida que os classifica em relação a sua importância (frequência), os mais importantes são organizados como evidencias de localidades. Para a tarefa de classificação é utilizada a técnica de Naive Bayes (LEWIS, 1998).

O trabalho de Souza *et al.* (2005) apresenta a criação de um dicionário geográfico para atender as necessidades na área de turismo. Utiliza um algoritmo para determinar as relações de proximidades entre as entidades encontradas em páginas da internet, que indicam localização escritas em linguagem natural. O resultado consiste na extração de entidades como cidades, rios, montanhas, e ruas, estruturas urbanas, dentre outras.

O trabalho de Nadeau; Turney; Matwin (2006), utiliza o aprendizado não supervisionado para a tarefa de reconhecimento de entidades, sem a necessidade de treinamento prévio de máquina de aprendizagem ou o uso de um dicionário geográfico como referência. Seu método é baseado em um algoritmo de extração de entidade nomeada proposta por Etzioni *et al.,* (2005), combinado com heurísticas de amostragem proposta por Ciravegna *et al.,* (2003). Contudo, apesar de utilizar o aprendizado não supervisionado, são necessárias algumas atividades de avaliações humanas para obter um maior percentual de precisão e abrangência na tarefa de reconhecimento de entidades das diversas fontes.

O trabalho de Popescu *et al.,* (2008), apresenta o Gezetiki, uma ferramenta automatizada para incrementar informações do dicionário geográfico Geonames, integrando informações da Wikipédia, fotografias aéreas (Panoramio) e buscadores da internet. Utiliza um algoritmo de REM e treinamento não supervisionado proposto por (TORISAWA, 2007), para analisar termos georreferenciados do conteúdo da Wikipédia, integrando as entidades existentes no Panoramio, proporcionando uma interação entre entidades textuais e fotografias das entidades.

O trabalho de Toral; Munoz (2006), apresenta uma tentativa de automatizar a criação e inserção de novas entidades em dicionário geográficos, utilizando a abordagem de REM para reconhecer de localidade (em nível de cidade e estado) combinada a uma heurística proposta pela Wordnet, com o objetivo de identificar se a entidade extraída pertence a uma classe de entidades. Nesse trabalho a utilização de *Part of Speech Tagger* (BRILL, 1992), também conhecido como *Postag* ou *POStag*, não é obrigatória.

O trabalho de Machado *et al.,* (2011) propõe a criação de um dicionário geográfico de novas entidades geográficas, utilizando notícias brasileiras oriundas da internet. Foi desenvolvida uma ontologia, utilizando um algoritmo de reconhecimento de entidades, baseada em análise léxica dos textos e na utilização da técnica de palavras combinadas combinado com heurísticas para a resolução de ambiguidade de topônimos.

O trabalho de Gouvêa *et al.,* (2008) é o trabalho mais próximo dessa dissertação, propondo a criação e atualização de dicionário geográfico baseado na identificação de entidades geográficas encontradas em textos de notícias, sem a necessidade de um corpus de anotação como referência. Para a identificação dos topônimos geográficos, utilizou tarefas de processamento de linguagem natural, como o uso de expressões regulares, com o objetivo final de atribuir as notícias as suas respectivas localidades geográficas. Semelhante ao trabalho desenvolvido nessa dissertação, o autor apresenta a identificação de rua e cidade em notícias, contudo, essa relação é estabelecida utilizando um cálculo de distância entre os termos que indicam localidade, rua e cidade, não utilizando do recurso de criação de regras baseado em classes gramaticais dos termos, conforme é utilizado nesse trabalho.

O trabalho de Overell; Ruger (2007) utiliza a técnica de aprendizado supervisionado para a identificação de nomes que são relacionados com localizações geográficas. O autor justifica em seu trabalho a utilização dessa abordagem devido ao alto grau de precisão no reconhecimento de entidade e da necessidade de atender melhor a criação de dicionários geográficos de corpus de menor quantidade de informações.

O trabalho de Gelernter et al., (2013) utiliza algoritmo de agrupamento Fuzzy, utilizando aprendizado de máquina (SVM), para determinar as entidades existentes em aplicações geográficas da internet, utilizando referencias geográficas de fontes colaborativas, com o objetivo de incrementar dicionário geográfico com novas entidades de menor granularidade de assuntos regionais ou local, visto que esses dicionários possuem sérias restrições. Para identificação dos topônimos é necessário que

os textos possuam marcadores de texto das respectivas aplicações geográficas, que restringe fortemente as fontes analisadas.

O trabalho de Moncla *et al.,* (2014) utiliza as tarefas de georesolução e geocodificação com o apoio das tarefas de aprendizado de máquina híbrido, mas especificamente, aprendizado não supervisionado e supervisionado, para o reconhecimento de entidades. Esse trabalho se diferencia dos demais em relação à identificação elementos geográficos de menor de granularidade.

O trabalho de Gómez (2012) desenvolveu um sistema automático para extrair a localização de notícias a partir de notícias em espanhol, mostrando lugares associados em um mapa. O autor apresenta técnicas de extração de palavras-chaves para identificação de elementos e o aprendizado supervisionado de máquina para o reconhecimento das entidades.

No trabalho desenvolvido por Goldberg *et al.*, (2009), foi proposto um sistema automatizado gerador de dicionário geográfico, onde o método proposto extrai informações de diferentes fontes de dados, principalmente informações disponíveis na internet, utilizando técnicas de extração das informações e aprendizado de máquina semi-supervisionado para reconhecimento de entidades.

Tabela 3: Trabalhos Relacionados

Trabalho	Corpora	Solução	Pontos Positivos e Observações
(SOUZA *et al.*, 2005)	Páginas da Internet	Expressões de localidade identificadas utilizando PLN	O contexto abordado (turismo) foi relevante e a relação de similaridade apresentada entre as entidades foi um marco importante, além da especificidade do produto final ser para o Brasil. Técnica não eficiente e baixa qualidade e quantidade das fontes analisadas.
(NADEAU; TURNEY; MATWIN, 2006)	Notícias	Aprendizado não supervisionado e técnicas de amostragem	Não necessita de corpus manual de anotação para reconhecimento de entidades. Positivo resultado na criação de grandes dicionários geográficos. A abordagem necessita de algumas atividades humanas não mencionadas no texto.
(POPESCU; GREFENSTETTE; MOËLLIC, 2008)	Wikipédia e Panorama	REM com Aprendizado não supervisionado	Utilização de textos não estruturados ou semiestruturados. Dados restritos apenas de 15 grandes cidades.
(TORAL; MUNOZ, 2006)	Wikipédia e Wordnet	REM na Wikipédia e Heurísticas e NPL no Wordnet.	A tarefa de reconhecimento é executada mais rapidamente. Propõe utilizar heurísticas de distância entre termos. *Postag* é opcional, dificuldades de identificar os elementos corretos no texto.
(MACHADO *et al.*, 2011)	Notícias	REM e heurísticas de resolução de topônimos	Mantêm nomes alternativos, abreviações e acrônimos para as entidades encontradas. As informações contidas no dicionário geográfico são sobre topônimos encontrados no Brasil. Problema na resolução de topônimos e baixa quantidade do dataset de testes.
(GOUVÊA *et al.*, 2008)	Notícias	Processamento de linguagem natural (expressões regulares) e probabilidade.	Construção e criação de dicionário geográfico de indicadores de localidade, por exemplo, relacionados com figuras públicas, relativos a entidades como hospitais, aeroportos, museus, universidades, relacionados com lugares geográficos como estradas, parques, construções, edifícios e assim por diante. Baixa quantidade no dataset e experimento restrito a apenas 9 grandes cidades.
(GOLDBERG; WILSON; KNOBLOCK, 2009)	Fontes diversas	REM com Aprendizado Semi-supervisionado	Um sistema automatizado gerador de dicionário geográfico, que extrai as informações utilizando técnicas de extração das informações e soluções de aprendizado de máquina semi-supervisionado.
(ODON DE ALENCAR; DAVIS; GONÇALVES, 2010)	Jornais do estado de Minas Gerais	Frequência de termos e *Naive Bayes*	Identificação de elementos geográficos em notícias do Brasil, estado de minas gerais, com nível de granularidade até cidade. Contexto limitado aos dados da Wikipédia.
(GÓMEZ *et al.*, 2012)	Notícias	Aprendizado supervisionado e dicionário de palavras	Extração de informações de baixa granularidade de notícias em espanhol e uso de dicionário de palavras e sinônimos e aprendizado supervisionado. Criação de uma ferramenta de visualização dos resultados.
(GELERNTER *et al.*, 2013)	Aplicações Geográficas	Lógica Fuzzy e aprendizado supervisionado	Incremento de dicionários geográficos, com novas entidades regionais, provendo maior nível de detalhamento aos dicionários. É necessário que os textos possuam marcadores geográficos das respectivas aplicações geográficas, restringindo seu uso.
(MONCLA *et al.*, 2014)	Textos de Viagens	Aprendizado híbrido de máquina	A utilização de técnicas de agrupamento para identificar os conjuntos (clusters) com a maior quantidade de topônimos e o objetivo de criar e incrementar dicionários geográficos com topônimos de menor granularidade. Experimentos em francês, italiano e espanhol.

2.3.3 Identificadores de Localidade

Na criação e atualização de dicionários geográficos são necessárias a extração e a identificação de elementos geográficos em textos. Segundo ZHANG; TSAI (2009), o objetivo na detecção de novas sentenças está relacionado à capacidade de remover as inconsistências, encontrar padrões e principalmente em eliminar redundâncias e ambiguidades.

Para a identificação de localidade em textos, este trabalho segue a metodologia baseada em trabalhos como (BRISABOA et al., 2010; LEIDNER; LIEBERMAN, 2011; MIKHEEV; MOENS; GROVER, 1999) e outros descritos na seção de trabalhos relacionados, nos quais estão apresentados as características, técnicas e recomendações para a extração e identificação de elementos que indiquem a presença de localidades geográficas em texto.

Para a realização da tarefa, diversas características podem ser utilizadas, como análise léxica, gramatical ou utilizando palavras previamente conhecidas, que expressem indícios da presença de termos de localidade na sentença ou texto, por exemplo, “rua”, “avenida”, “localizada”, dentre outras (MIKHEEV; MOENS; GROVER, 1999).

Diversos tipos de indicadores ou referencias geográficas podem ser encontrados em notícias, e os indicadores mais encontrados são as expressões geográficas de maior granularidade ou escala, como países (por exemplo, "Brasil", “Estados Unidos” e “Canadá”), regiões do país ("Sul" e "Norte"), cidades (“Rio de Janeiro”, “Juiz de Fora” e “Salvador”), bem como lugares de destaque como ("Copacabana") (LEIDNER; LIEBERMAN, 2011).

Indicadores de menor granularidade ou escala são mais difíceis de serem identificados, por exemplo, podem ser considerados, locais como bairro (“Botafogo”, “São Mateus”, “Cascatinha”), rua ("Barão do Rio Branco"), logradouros e número ("Rua João Henrique Dielle, 36"), cruzamentos de rua ("Rua Oswaldo Cruz e Marquês de Abrantes"), centros de cidades ("centro de São Paulo") e edifícios ("Prédio da Petrobrás") (LEIDNER; LIEBERMAN, 2011).

O trabalho de (BRISABOA *et al.*, 2010) procura palavras ou termos “candidatos” a pertencer ao grupo de localidade utilizando o reconhecimento de entidades e palavras chave.

Nesse trabalho utilizamos para o auxílio na identificação de elementos geográficos de menor granularidade, foi criado um dicionário de palavras-chave (DPC). Esse dicionário foi criado a partir da palavra logradouro e palavras de sinônimos de representatividade, semelhante aos principais dicionários geográficos que agrupam as palavras e palavras correspondentes. A Tabela 4 apresenta um conjunto de termos que compõem o dicionário de palavras de logradouro:

Tabela 4: Dicionário de Palavras-Chave para Logradouro

Palavras de Logradouro
Rua
Avenida
Travessa
Rodovia
Alameda
Beco
Largo

3 Tarefas para o Reconhecimento de Elementos Geográficos em Texto

Esse capítulo apresenta as ferramentas e métodos necessários para o reconhecimento de entidades mencionadas em texto, as métricas utilizadas para a medição e alguns tipos de aprendizado de máquina apropriados ao problema da pesquisa.

3.1 Reconhecimento de Entidades Mencionadas

Dentre essas diferentes técnicas de extração de informação destaca-se o Reconhecimento de Entidades Mencionadas (REM) que consiste em tarefas de detecção e rastreamento de entidades com o objetivo de coletar as entidades presentes no texto e classifica-las de acordo com a estrutura vigente e suas regras (DODDINGTON et al., 2004; ELLOUMI et al., 2013).

O termo Reconhecimento de Entidade Mencionada (REM) ou Reconhecimento de Entidade Nomeada (REN), tradução usada pela comunidade de língua portuguesa para o original em inglês, *Named Entity Recognition* (NER), foi historicamente mais utilizado no mercado a partir da década de 90, principalmente pela evolução de uma grande quantidade de componentes de Extração de Informação para os documentos publicados na internet (RIZZO; TRONCY, 2011).

Com o aumento na ênfase nas técnicas de reconhecimento de linguagem natural, o REM se consolidou como um componente essencial para o campo de Extração de Informação. Atualmente, o REM é muito utilizado para extração de conhecimento específico em fontes de dados textuais e nas tarefas de classificação de tipos de categorias pré-definidas, como pessoa, organização e localização.

3.1.1 Conferências de REM

Com o objetivo de aprimorar o estado da arte e incentivar novas iniciativas de reconhecimento de entidades, foram criadas conferências de avaliação de resultados denominadas: *Message Understanding Conference* (MUC), *Automatic Content Extraction* (ACE) e conferencia em português de reconhecimento de entidades denominada de avaliação conjunta na área do reconhecimento de entidades mencionadas em português (HAREM).

As conferencias tratam de avaliar os resultados obtidos entre os diversos sistemas, utilizando métricas como precisão, abrangência e medida-F.

Essas conferências proporcionaram grande avanço nas tarefas de reconhecimento de entidades ao longo dos anos. Na seção a seguir estão apresentadas as principais conferências e suas características, estudos e resultados relevantes.

3.1.1.1 A Conferência de Entendimento de Mensagens (MUC)

Com a evolução dos sistemas utilizando tarefas de extração da informação, houve a iniciativa de criar as conferências para avaliar e melhorar os desempenhos das tarefas relacionadas dessa atividade.

Nesse contexto sugiram as conferencias de entendimento de mensagens, em inglês, *Message Understanding Conference* (MUC), que inicialmente aconteceu em 1987 MUC-1. As MUCs foram iniciadas e financiadas pelo Governo dos Estados Unidos com o intuito de promover a inserção de outras organizações e de gerar melhoras nos sistemas de EI.

O MUC 2 em 1989 foi destinado a extrair informações em mensagens navais. Nas duas edições seguintes, MUC 3 e MUC4, o trabalho de extração foi concentrado em utilizar as notícias coletadas sobre os incidentes terroristas em países latino-americanos. O MUC5 trouxe uma nova melhoria baseado na taxa erro e evoluções na medida de desempenho (GRISHMAN; SUNDHEIM, 1996).

Apesar das conferências MUC, as tarefas relacionadas ao reconhecimento de entidades mencionadas para o idioma inglês, foram introduzidas apenas na sua sexta versão (GRISHMAN; SUNDHEIM, 1996). Nessa versão da conferência, ainda foram apresentados as entidades dos tipos apresentados a seguir (MIKHEEV; GROVER; MOENS, 1998):

- Examex: é formada por nomes próprios, organizados nas categorias pré-definidas como Pessoa, Organização e Local;
- Timex: é formada pelas representações temporais de data e hora;
- Numex: é formada por expressões monetárias representadas em percentual.

Houve ainda a sétima versão da MUC, última versão da conferencia e ocorreu em 1997-1998, com a participação de 200 artigos e com resultado expressivo (mais de 90%) na métrica denominada “medida F” para as atividades de reconhecimento de entidades mencionadas na língua inglesa. (MIKHEEV; GROVER; MOENS, 1998).

3.1.1.2 Conferências de Extração Automática (ACE)

A partir do ano de 1999 iniciaram às conferências de extração automática de conteúdo, em inglês, *Automatic Content Extraction* (ACE). Essa conferência, tomou maior notoriedade durante os anos de 2001-2002 com a adição das tarefas de identificação e classificação das entidades, tarefa denominada de em *Entity Detection and Tracking* (EDT), além de contemplar categorias não apresentadas anteriormente pela MUC (DODDINGTON et al., 2004).

O passo seguinte da conferência ocorreu nos anos de 2002-2003, no estabelecimento de relação entre pares de entidades, ou relacionamento entre as entidades, tarefa denominada como *Relation Detection and Characterization* (RDC) (DODDINGTON et al., 2004).

Em 2004 foi a vez das tarefas relacionadas a extração e caracterização da presença de eventos, denominado *event extraction*. Outro avanço foram as conferencias de *Automatic Content Extraction* (ACE) com o desenvolvimento da capacidade de extrair o significado de fontes de diversos tipos de fontes de dados, incluindo texto, áudio e imagens (DODDINGTON et al., 2004).

3.1.1.3 Avaliação Conjunta de REM em Português (HAREM)

A HAREM foi a primeira e mais importante conferência de reconhecimento de entidade em língua portuguesa. Em seu trabalho (MOTA; SANTOS, 2008a), define a conferencia como "um conjunto de avaliações de sistemas de reconhecimento de entidades mencionadas em português, com o objetivo de avaliar o sucesso na identificação e classificação automática dos nomes próprios na língua portuguesa".

Nas conferências apresentadas anteriormente (MUC e ACE) o termo utilizado para determinar a área, é o Reconhecimento de Entidades Nomeadas. Contudo, o HAREM utiliza o termo Reconhecimento de Entidade Mencionada.

O HAREM difere de MUC principalmente pela "liberdade" na tarefa de encontrar entidades. Enquanto em MUC as entidades estavam restritas aos tipos: pessoa, localização e organização, no HAREM os trabalhos apresentados estavam interessado em todos os nomes próprios (SANTOS, 2007).

Em 2004 ocorreu a primeira conferência HAREM, e apresentou dois aspectos principais:

- A classificação e identificação de um termo;

- A possibilidade de que um determinado termo possa ser classificado em mais de uma categoria, caso seu contexto seja indeterminado de ser solucionado.

Em 2008, correu a segunda conferência HAREM, apresentando um total de 10 categorias e 42 tipos de elementos identificáveis, conforme apresentadas na Tabela 5

Tabela 5: Categorias do segundo HAREM – adaptado de (OLIVEIRA et al., 2008).

Categoria	**Tipo**
Abstração	Disciplina, Estado, Ideia, Nome, Outro.
Acontecimento	Efemeridade, Evento, Organizado, Outro.
Coisa	Classe, MembroClasse, Objeto, Substancia, Outro.
Local	Físico Humano, Virtual.
Obra	Arte, Plano, Reproduzida, Outro.
Organização	Administração, Empresa, Instituição, Outro.
Pessoa	Cargo, GrupoCargo, GrupoInd, GrupoMembro, Individual, Membro, Povo, Outro.
Tempo	Duração, Frequência, Genérico, TempoCalend, Outro.
Valor	Classificação, Moeda, Quantidade, Outro.
Outro	

As categorias representadas na segunda HAREM são de alta granularidade, ou seja, nível macro dos elementos e os resultados obtidos nas tarefas e métricas sejam em relação a essa escala. A Figura 2 apresenta uma representação explicitando as categorias em menor nível de granularidade. Dentre as representações, ressaltamos a granularidade da categoria Local seguida por suas arestas Humano seguido por seus itens “País”, “Região”, “Divisão”, “Rua” e “Construção”.

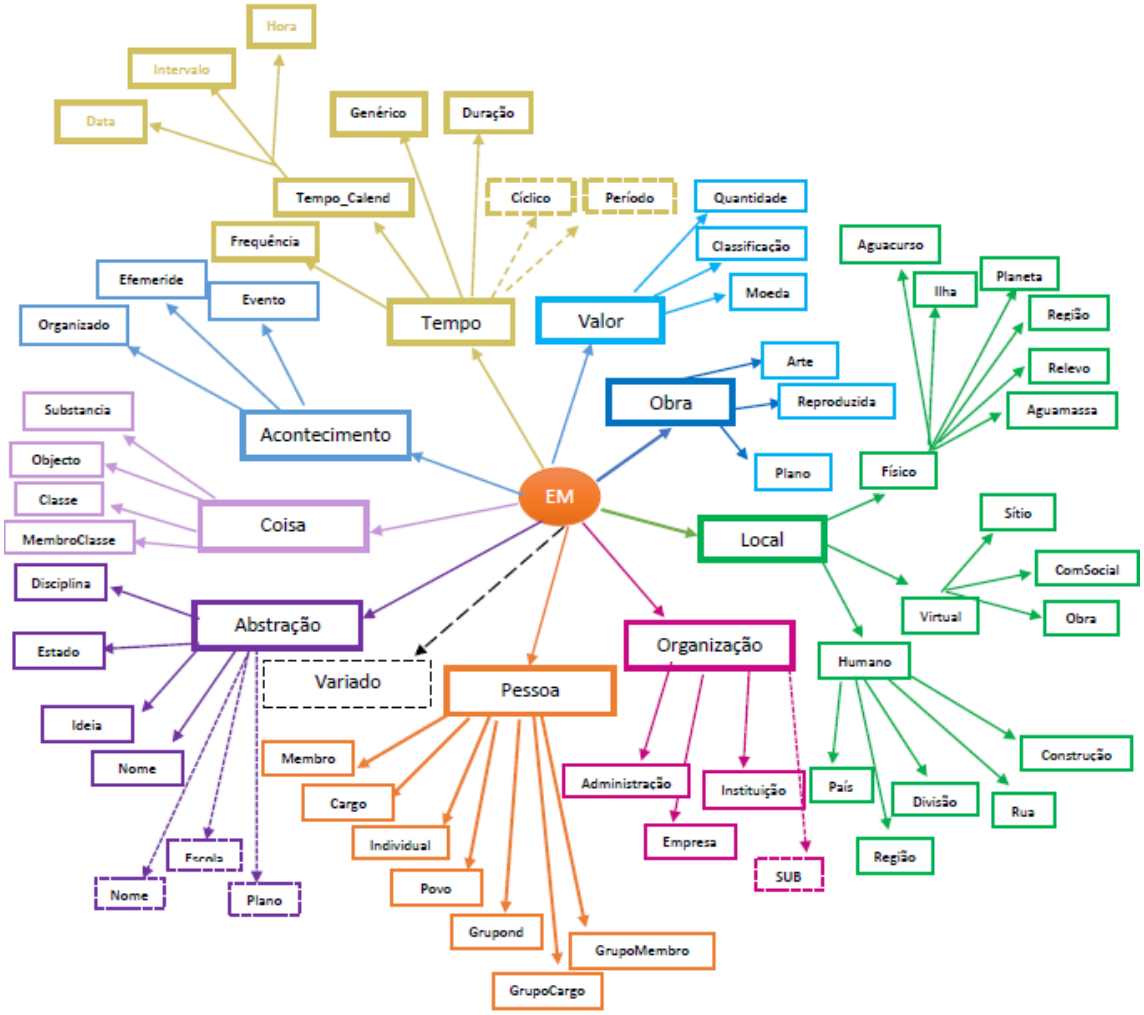

Figura 2: Árvore de categorias do segundo HAREM – adaptado de (OLIVEIRA et al., 2008).

3.1.2 As Tarefas de REM

As tarefas de REM, mas especificamente as relacionadas à classificação de termos e expressões, têm sido utilizadas com frequência principalmente em redes sociais e notícia com objetivo de identificar elementos relevantes denotados em linguagem natural. No entanto, o reconhecimento das entidades nomeadas torna-se um desafio devido a fatores como a grande quantidade de entidades de dados e pelo excessivo número de possibilidades de entidades (RITTER et al., 2011).

Um aspecto importante nas atividades de REM é estabelecer quais são os tipos de entidades necessários alvo de identificação no texto. Algumas categorias de entidades nomeadas são mais fáceis de encontrar do que outras, dependendo das especificidades desejadas (GRISHMAN; SUNDHEIM, 1996).

As características estruturais, semânticas e ortográficas pertencentes aos termos que compõem a sentença, podem ser vistos como indícios da ocorrência de um tipo de entidades. A seguir são apresentadas algumas características para a tarefa de REM em textos (SUNDHEIM, 1996):

• *Características das palavras*: as palavras podem ser utilizadas como um recurso importante para listar entidades. Elas são úteis tanto para compor o dicionário de nomes a partir do *corpus* de treinamento, quanto para capturar certas propriedades das palavras, podem também servir para indicar ou desencadear a ocorrência de uma entidade, por exemplo, o termo “”sr.” acrônimo da palavra “senhor”, em geral indica que a próxima palavra seja o nome da pessoa.

• *Características ortográficas*: propriedades ortográficas das palavras podem ser de grande importância, recursos como a utilização de letras em maiúsculo, a presença de símbolos especiais e caracteres alfanuméricos;

•*Características morfossintáticas*: a morfossintaxe (morfologia + sintaxe) é um recurso associado importante, principalmente a morfologia, no que se refere à classe gramatical de uma palavra (nome, adjetivo, artigo, pronome, quantificador, advérbio, preposição, conjunção, interjeição);

• *Características de pesquisa em dicionários geográficos*: conhecimentos adicionais podem ser adicionados aos sistemas de aprendizagem utilizando uma base de dados de entidades existente. Essa base de dados pode adicionar características e correspondência (*match*) entre a palavra do texto e a encontrada no dicionário da base de dados.

Outro recurso recente adotado para o reconhecimento de entidades mencionadas é o uso de fonte de recursos externos, como a Wikipédia (WIKIPEDIA, 2015), OpenStreetMap (“OpenStreetMap Brasil”, 2016), Google Places (GOOGLE, 2016), que fornecem informações enciclopédicas sobre entidades semi-estruturadas e podem ser usadas para criar automaticamente um banco de dados estruturado ou um dicionário geográfico (RAUCH; BUKATIN; BAKER, 2003; TEITLER et al., 2008).

A literatura apresenta diversos métodos probabilísticos que podem ser utilizados na tarefa de reconhecimento de entidades mencionadas, principalmente utilizando o aprendizado supervisionado de máquina. Dentre os modelos apresentados destacam-se o modelo de entropia máxima (GULL; DANIELL, 1984), *Hidden Markov Models* (EDDY, 1998) e o *Conditional Random Fields* (SUTTON; MCCALLUM, 2006).

3.1.3 Resolução e Anotação de Topônimos

A tarefa de resolução de topônimos consiste em métodos para encontrar determinado topônimo em um conjunto finito de texto, lidando com subjetividades e fatores de ambiguidade de determinação de classificação do termo.

Autores como ZHANG; TSAI (2009), DODDINGTON (2004) e RATINOV; ROTH (2009), apresentam diversas soluções para a categorização e anotação correta das entidades, para evitar que uma entidade possa ser referenciada como outra, ou seja, ambiguidade, evitando que a máquina de aprendizado lide com problemas de dualidade de categorias.

O caso mais comum são os problemas relatados a categorização de ambígua entre as fontes de dados, no qual uma entidade em uma fonte de dados é caracterizada como sendo da Categoria X e em outra fonte de dados é categorizada como na Categoria Y. Conforme apresenta na Tabela 6 para a frase de exemplo "João Rua Prefeito":

Tabela 6: Exemplo de Categorização Ambígua de Entidades

Conjunto de Dados 1		Conjunto de Dados 2	
Palavras	**Categorias pré-definidas**	**Palavras**	**Categorias pré-definidas**
João	Nome	João	Nome
Rua	Nome	Rua	Logradouro
Prefeito	Cargo	Prefeito	Cargo

Nesse caso a palavra "Rua" teria uma classificação ambígua, resultando em problema para a categorização. RATINOV et al., (2009) apresentam que uma maneira de auxiliar na resolução dos problemas de categorização de topônimos é analisar contextos distintos, atribuindo especificidade ao contexto, ou seja, analisar os grupos de entidades em separado.

A anotação (em inglês: *tagging)* de topônimos são necessários para classificar um termo, sendo parte importante para o sucesso da tarefa de reconhecimento de entidade. Em geral, essa é uma tarefa manual e que exige grande esforço e dedicação, sendo tradicionalmente feita por especialistas, como na HAREM, onde sua coleção de treinamento para a entidade LOCAL foi realizada por quatro pessoas.

A anotação das entidades auxilia no processo de não ambiguidade, principalmente se analisarmos em categorias em separadas. O trabalho de DODDINGTON et al. (2004) apresenta a abordagens para a resolução do problema de

categorização ambígua, com a criação e anotação manual de categorias em situações em que as coleções de dados analisados são difíceis de categorizar. As anotações das entidades podem seguir o padrão apresentado na Tabela 7.

Tabela 7: Notícia Anotada com a Notação IO

Termos	Anotação
Na	O
Rua	LOGRADOURO
Praia	LOGRADOURO
Do	LOGRADOURO
Flamengo	LOGRADOURO
402	LOGRADOURO

Conforme apresentado na Tabela 8, podem aparecer entidades que estão presentes em mais de um termo ou palavra, por exemplo, no elemento "Praia do Flamengo 402". Delimitar e conectar termos que possuem correspondência é importante para representar elementos relacionados ou que não podem ser divido, como, o nome de uma rua, cidade, região. Em seu trabalho, Ratinov (2009), apresenta outra anotação de entidades, denomina de notação BIO. Neste trabalho vamos apresentar a anotação BIO, que consiste do significado das letras, *B:begin, I:inside: O: others*. Nessa anotação as palavras além de serem anotadas conforme sua entidade correspondente possui uma anotação adicional que explicita sua relação entre as palavras anteriores e posteriores. A Tabela 8 apresenta uma notícia anotada utilizando a notação BIO.

Tabela 8: Notícia Anotada com a Notação BIO

Termos	Anotação
Na-O	O
Rua-B	LOGRADOURO
Praia-I	LOGRADOURO
Do-I	LOGRADOURO
Flamengo-I	LOGRADOURO
402-I	LOGRADOURO

3.2 Medidas

Com as evoluções entre os trabalhos apresentadas nas conferências MUC e ACE de REM, surgiu a necessidade de medir os diversos resultados encontrados nas atividades entre os projetos participantes dessas conferências. Partindo da premissa, foram criadas diversas medidas ou métricas que permite medir os resultados, dentre elas destacam-se as medidas de precisão, recuperação e medida-F ou também conhecida como *F-score* (COWIE; LEHNERT, 1996; KONKOL, 2012).

Semelhantemente as avaliações para MUC e ACE, a segunda conferência de HAREM apresentou as medidas de avaliação de resultado para as tarefas de reconhecimento de entidades, dentre elas estão: precisão, abrangência e Medida F (OLIVEIRA et al., 2008).

Para obter os resultados das métricas propostas é necessária a classificação dos termos (objetos) é denotada por dois valores, positivos e negativos, e suas relações que resultam em quatro possíveis tipos de classificação (KONKOL, 2012).

- Verdadeiros Positivos, em inglês, *True Positive* (TP), são itens relevantes ao contexto que corretamente são identificados como positivos ou relevantes.
- Verdadeiros Negativos, em inglês, *True Negative* (TN), são itens irrelevantes ao contexto que corretamente são identificadas como falsos ou irrelevantes.
- Falsos positivos, em inglês, *False Positive* (FP), são itens irrelevantes ao contexto que incorretamente são identificados como positivos ou relevantes.
- Falsos negativos, em inglês, *False Negative* (FN), são itens relevantes ao contexto que incorretamente são identificadas como falsos ou irrelevantes.

A Figura 3 ilustra uma representação gráfica da relevância da classificação dos termos, os termos com delimitação na cor preta são denominados relevantes (TP, FP, FN), pois estes são utilizados para mensurar o modelo proposto pela máquina de aprendizado, através do cálculo baseado nas medidas apresentadas nesse trabalho.

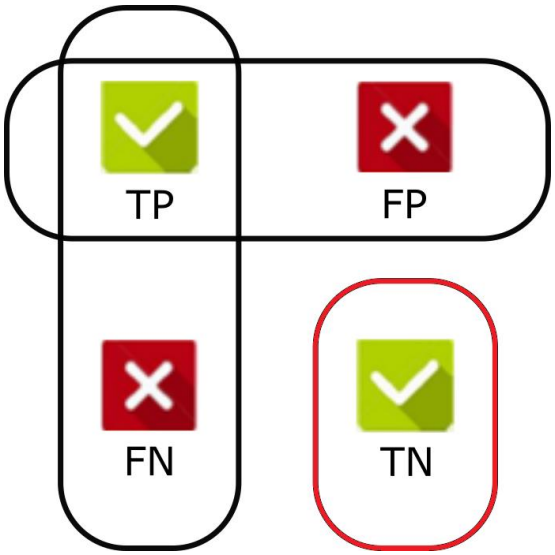

Figura 3: Relevância dos Termos

A Figura 4 apresenta essa classificação, onde as curvas mostram a distribuição de objetos positivos e negativos e a linha tracejada mostra a divisa da decisão do classificador. Nas áreas marcadas como FN e FP são alguns objetos marcados incorretamente (KONKOL, 2012).

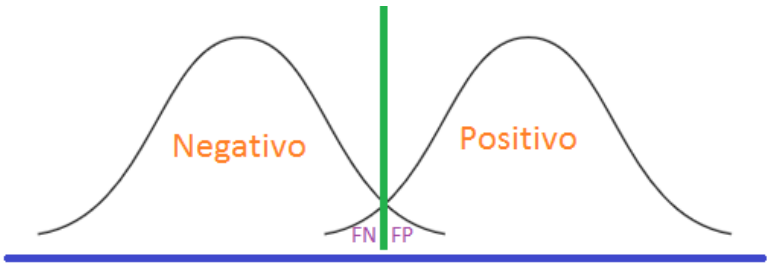

Figura 4: Fronteiras de Classificação dos Termos - Baseado de *Precision* e *Recall* (KONKOL, 2012).

3.2.1 Abrangência

Abrangência, em inglês, *Recall* é e de uma medida em que todos os objetos positivos são marcados (selecionados).

$$Abrangência = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} x\ 100\%$$

Conforme a representação dos conjuntos A e B da Figura 5, os valores dos que atendem a medida de abrangência, são expressos por:

$$Abrangência = \frac{\mathrm{A} \cap B}{\mathrm{A}}$$

3.2.2 Precisão

A precisão é uma medida que os objetos marcados como positivos, são realmente denotados como positivos.

$$Precisão = \frac{\text{TP}}{\text{TP} + \text{FP}}\ x\ 100\%$$

A precisão também pode ser entendida utilizando as teorias de conjuntos matemáticos. Considerando que um conjunto A que representa os dados classificados como VP, e o conjunto B que são os dados a ser avaliados. A Figura 5 representa essa maneira:

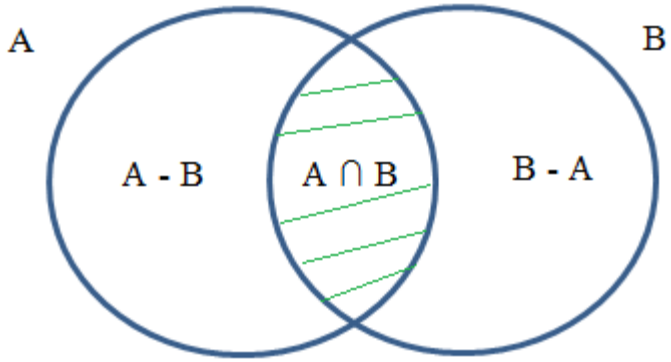

Figura 5: Conjuntos A e B e suas intercessões

Observe que de acordo com a Figura 5, os valores dos conjuntos que atendem a medida de precisão, são expressos por:

$$Precisão = \frac{\text{A} \cap B}{\text{B}}$$

3.2.3 Medida-F

A métrica denominada medida-F, mais especificamente a medida-F1, é responsável em determinar a acurácia, no qual é feita a contagem dos objetos de acordo com a classificação de seus termos em P, N, FP, FN (KONKOL, 2012).

$$Medida\ F = \frac{2\ .\ \text{Precisão}\ .\ \text{Abrangência}}{\text{Precisão} +\ \text{Abrangência}} x\ 100\%$$

3.3 Aprendizado de Máquina

O aprendizado de máquina (em inglês: *machine learning*) pode ser definido como a área computacional, mais especificamente de Inteligência Artificial, cujo objetivo é o desenvolvimento automático de conhecimento computacional baseado em outras formas de aprendizado (MONARD; BARANAUSKAS, 2003).

Diferentemente dos métodos baseados em dedução, o aprendizado de máquina utiliza como premissa o conceito de inferência indutiva[6], com o objetivo de obter determinada conclusão sobre o cenário não específico. A inferência indutiva permite obter conclusões genéricas com objetivo de predição, sobre determinado conjunto de dados a partir de exemplos apresentados (MICHALSKI; CARBONELL; MITCHELL, 2013). A Tabela 9 apresenta um prisma da diferença do método dedutivo para o indutivo.

Tabela 9: Método Dedutivo x Indutivo

Dedutivo	**Indutivo**
Se a sentença que contém a palavra "rua" é uma localização válida, então as novas sentenças que também tiverem a palavra "rua" devem ser um endereço válido.	Se a sentença que contém a palavra "rua" é uma localização válida, então é provavelmente verdadeiro que as novas sentenças que também tiverem a palavra "rua" devem ser um endereço válido, mas a proposição não necessariamente é verdadeira.

O uso de inferência indutiva pode gerar incertezas e questionamentos principalmente com a sua relação com os métodos de composição empírica, visto que ambos se concentram em uma amostragem/observação para predizer o futuro (FRIEDRICH, 2015).

Os algoritmos de aprendizado de máquina podem ser classificados de acordo suas especificidades e características, dentre as mais usuais destacam-se dois grupos (BROWNLEE, 2013):

- O primeiro é um grupo de algoritmos pelo estilo de aprendizagem.
- O segundo é um agrupamento de algoritmos por similaridade dos dados introduzidos.

Devido à grande quantidade de algoritmos de aprendizado de máquina existentes, nesse trabalho estão listados os mais relevantes para essa dissertação. No primeiro grupo, existem diferentes maneiras de modelar uma necessidade ou problema, principalmente relacionadas com a interação, experiência ou o ambiente dos dados introduzidos. Os algoritmos mais relevantes nessa etapa são do tipo de aprendizado supervisionado e não supervisionado.

No segundo grupo de algoritmos relacionados por similaridade, a relação consiste em agrupar os termos baseados na semelhança das informações, dentre os diversos

[6] A **indução** é o raciocínio que, após considerar um número suficiente de casos particulares, conclui uma possível verdade geral.

métodos destacam-se os baseados em árvore de decisão e rede neural (BROWNLEE, 2013).

O aprendizado indutivo pode ser dividido em supervisionado ou não supervisionado. No aprendizado supervisionado é fornecido um conjunto de exemplos de treinamento (indutor), necessário na aplicação do algoritmo. No aprendizado não supervisionado não se utiliza indutor, ou seja, não ocorre um treinamento com o conjunto de treinamento. Os tipos de aprendizado incluindo o aprendizado semi-supervisionado são descritos nas seções decorrentes.

3.3.1 Aprendizado supervisionado

A técnica de aprendizado supervisionado (MØLLER, 1993; MONARD; BARANAUSKAS, 2003) tem sido muito utilizada na resolução da tarefa de reconhecimento de entidades mencionadas.

A técnica de aprendizado supervisionado, conforme apresentado anteriormente, necessita de um objeto de entrada, um indutor, que consiste de um conjunto de dados de aprendizado, denominado de treinamento, com o objetivo de obter um bom classificador utilizando o conjunto de dados de treinamento utilizando na máquina de aprendizado.

O processo de saída da máquina de aprendizado consiste em um classificador de novos conjuntos de dados, denominado de testes, desconhecidos pela máquina, com a finalidade de predizer as possíveis entidades inerentes aos termos pertencentes ao novo conjunto de dados.

Os resultados dos processos de treinamento/validação e testes permitem a geração de uma modelo que pode ser avaliado baseado nas métricas de precisão, abrangência e medida-F, para determinar a qualidade do modelo analisado.

A Figura 6 representa a estrutura do aprendizado supervisionado. Para o início das atividades é necessário a participação humana na tarefa de anotação manual dos termos. Em geral, essas atividades são desempenhadas por especialistas, mas podem também ser feitas de forma colaborativa por pessoas engajadas a participar do processo. Em seguida, o conjunto de textos anotados é dividido em dois conjuntos, treinamento/validação e de teste, delimitando o percentual ou quantidade de registros necessários para compor cada atividade.

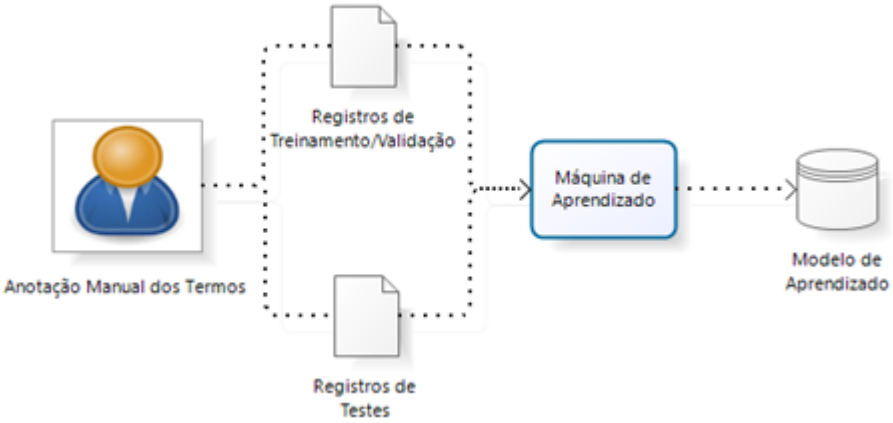

Figura 6: Estrutura Simplificada da Técnica de Aprendizado Supervisionado

3.3.2 Aprendizado não supervisionado

O aprendizado não supervisionado também conhecido como aprendizado por observação e descoberta, consiste na extração de informação sem supervisão humana, e sem a necessidade de um corpus de treinamento para a tarefa de classificação das entidades (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

Segundo ETZIONI et al., (2005), "como os sistemas de extração não supervisionados não requerem intervenção humana, eles podem recursivamente descobrir novas entidade, relações, atributos e instâncias, de forma escalável totalmente automatizado".

Um modelo de aprendizagem baseado em aprendizado não supervisionado tenta adequar os parâmetros existentes no contexto em estudo ao conjunto definido de dados, de modo a melhor resumir regularidades encontradas nos dados. A Figura 7 apresenta as etapas de um processo de aprendizado não supervisionado:

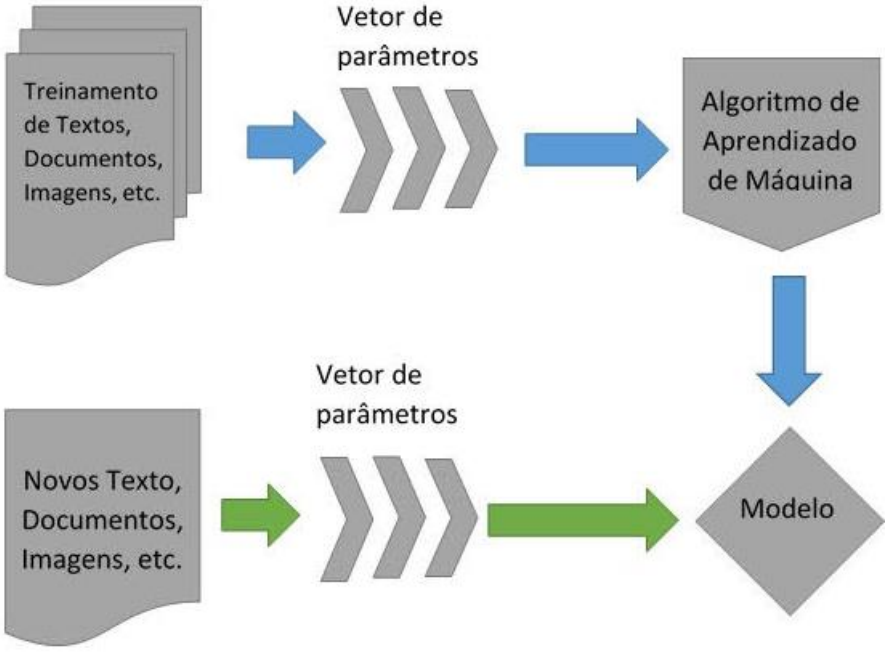

Figura 7: Estrutura de Aprendizado Não Supervisionado, adaptado de (SCIKIT-LEARN DEVELOPERS, 2016).

Dentre os algoritmos mais conhecidos, estão listados principalmente os algoritmos de clusterização como o *K-Means* (HARTIGAN; WONG, 1979) e de associação como o Apriori (CHEUNG et al., 1996).

3.3.3 Aprendizado Semi-Supervisionado

A abordagem de treinamento semi-supervisionado é derivada do aprendizado supervisionado, mas possui um diferencial de necessitar de um conjunto menor de dados para treinamento. As tarefas iniciais de entrada são minimizadas nesse processo, decorrente da minimização da quantidade de elementos necessários para as etapas de treinamento e testes.

Semelhante ao aprendizado supervisionado, o semi-supervisionado inicia as atividades com a anotação manual do conjunto de dados de textos, e consequentemente a divisão em treinamento e testes, com as respectivas separações de percentual ou quantidade de registros.

O processo de sistema de aprendizado semi-supervisionado tem como entrada as coleções ou minicoleções de treinamento e testes anotadas e a coleção de dados não anotada com entidades. A partir dessas entradas, o sistema busca generalizar as inferências existes, com o objetivo de encontrar as entidades da base não anotada tendo como base os padrões e regras existentes no contexto das bases anotadas. A Figura 8

apresenta a estrutura básica de composição da abordagem de treinamento semi-supervisionado.

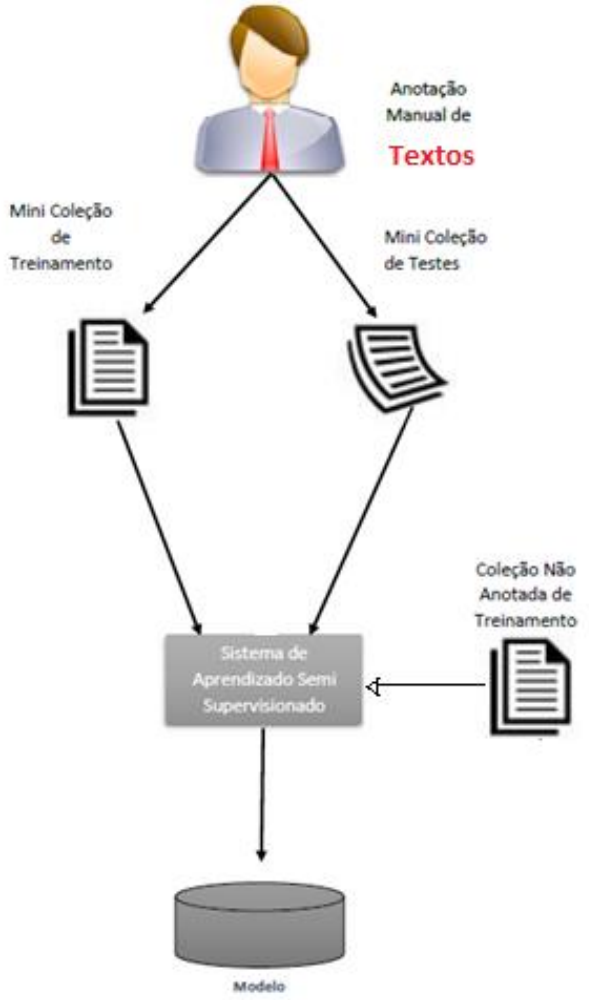

Figura 8: Estrutura de Treinamento Semi-Supervisionado

4 GeoNewsBR Dicionário Geográfico

Em adjacência aos trabalhos apresentados por TOBIN et al.,(2010), GROVER et al., (2010), CLOUCH (2005) e GOUVÊA et al., (2008), esse capítulo propõe a criação de uma estrutura de apoio para a identificação de novos elementos geográficos que permite encontrar endereços de baixa granularidade em notícias para a criação automática de um dicionário geográfico de construções e edificações.

Para a criação da estrutura de apoio, nesse capítulo apresentamos diversas ferramentas, métodos e tecnologias que permitam estabelecer os objetos propostos, através dos passos necessários como a utilização de coletor de notícias, identificação de classes gramaticais, divisão de notícias em sentenças, criação automática de regras, dentre outros. As tarefa e etapas apresentadas nessa seção são necessárias para identificar os elementos geográficos de baixa granularidade e apoiar a criação do dicionário geográfico GEONEWSBR.

Esse capítulo está dividido em três seções: a primeira consiste na criação de regras gramaticais que possuem indicadores de localidade, a segunda está focada na aplicação das notícias sobre as regras geradas, de maneira que apenas as notícias validadas no processo possam ser utilizadas para a identificação dos elementos geográficos, e a terceira consiste no aprendizado de máquina.

4.1 Etapas de Pré-processamento

O pré-processamento foi utilizado nesse trabalho para coletar, filtrar, restringir, filtrar e organizar as notícias, transformando-as em janelas válidas para o processo de aprendizado de máquina possa desempenhar suas tarefas. O processo de pré-processamento é necessário para extrair os conteúdos geográficos de baixa granularidade.

Esse processo está organizado em três etapas, primeiro consiste na coleta de notícias, o segundo na criação de regras e a terceira na identificação de janelas de logradouro. A Figura 9 descreve as três etapas que compõem essa estrutura.

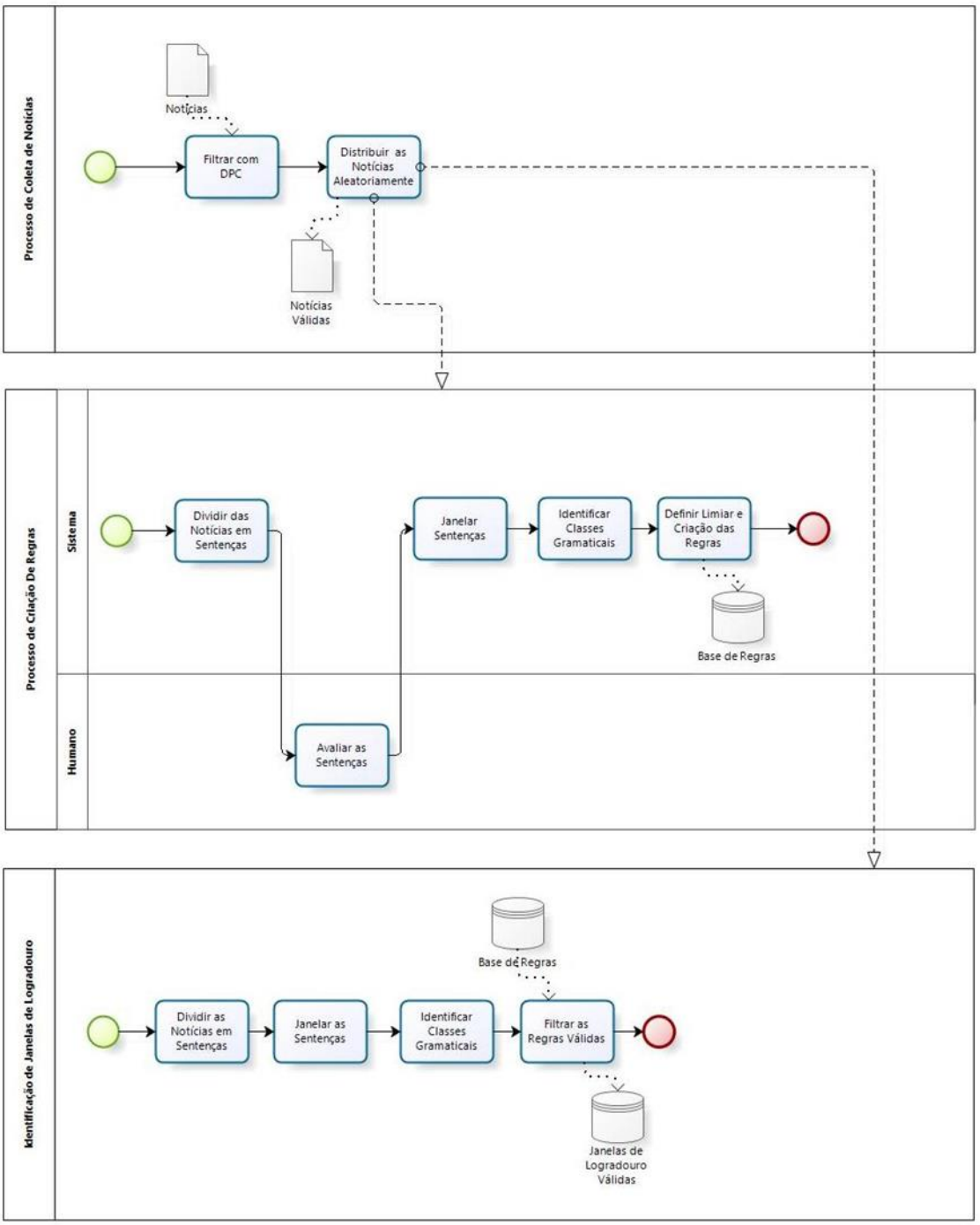

Figura 9: Estrutura Inicial do Pré-processamento

4.2 Coleta de Notícias

O processo de obtenção de notícias é fundamental para a realização de trabalho. Nessa seção apresentamos as etapas que compõem esse processo, proporcionando notícias com elementos válidos para as seções sequentes.

Para a realização das etapas de pré-processamento das notícias criado nesse trabalho, foram utilizados diversos filtros de indicadores de localidade e de novidade no texto. As etapas de coleta de notícias utilizada nesse trabalho são:

1. O processo inicia com a obtenção da fonte de dados. Em geral, notícias, extraídas de *webservice* ou outro sistema de apoio, ou através do sistema web desenvolvido nesse trabalho. Embora esse não seja objeto de estudo desse trabalho, foi criado uma implementação que facilita o processo de obtenção das notícias.
2. O sistema recebe um XML, texto plano ou o *link* que contenha o corpo da notícia, especificando um campo ou delimitador que contenha os textos. As notícias têm seu conteúdo adequado para texto plano, com a remoção de caracteres especiais e marcadores específicos do formato utilizado.
3. Nessa etapa, de filtro simples, são escolhidos os parâmetros que serão utilizados para delimitar as notícias contidas na base de dados:
   a. Termos que indicam novidade, por exemplo, *novo*, *aberto*, *inaugurado*, *inauguração*, entre outros.
   b. Palavras ou expressões que apresentam conteúdo geográfico explícito pertencente ao dicionário de palavras chave (DPC).
4. Aleatoriamente as notícias são separadas em dois grupos, o primeiro grupo é consideravelmente menor que o primeiro, este é utilizado para as tarefas de treinamento/validação e testes. O segundo grupo consiste dos registros restantes no dicionário de dados, e será utilizado a partir do modelo criado no primeiro grupo para a delimitação de seu conteúdo geográfico e nas tarefas de visualização dos dados apresentados no capítulo 5.

A Figura 10 apresenta a etapa de coleta de notícias, que serve de insumo para a atividade de pré-processamento das notícias.

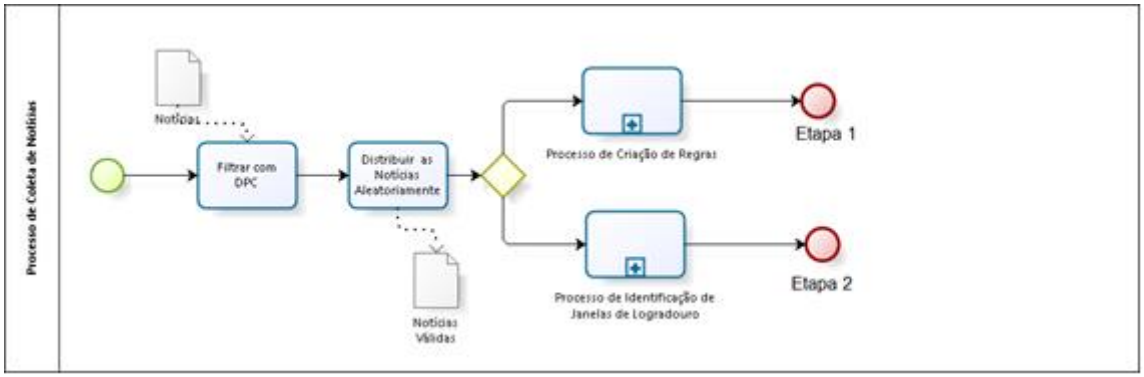

Figura 10: Estrutura Inicial de Coleta de Notícias

4.3 Criação de Regras para Logradouro

Essa seção é responsável pela descrição das etapas necessárias para a criação de regras válidas no processo de identificação de logradouro. A Figura 11 apresenta as tarefas necessárias para a criação das regras.

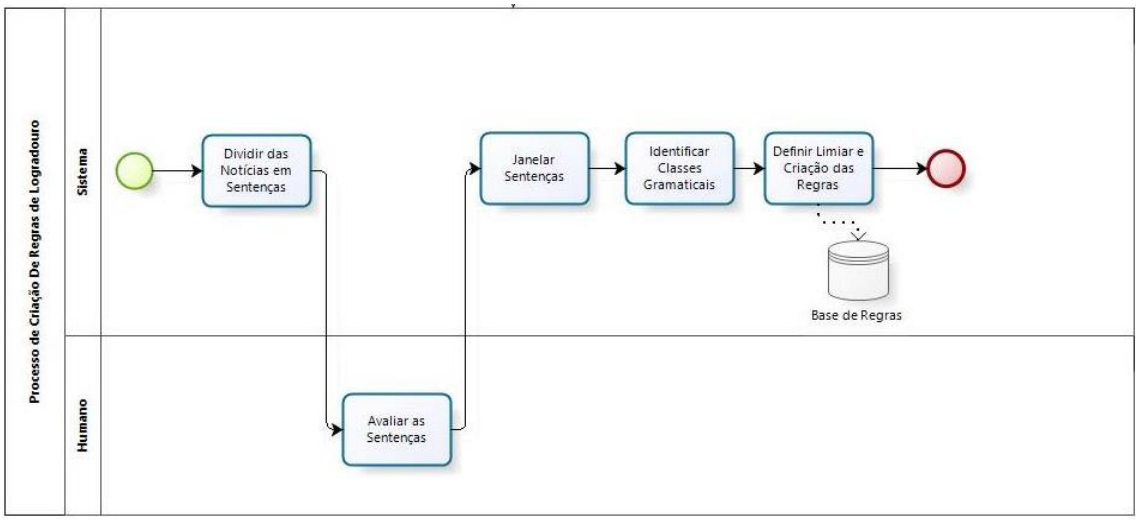

Figura 11: Etapas de Criação de Regras de Logradouro

O processo de criação de regras de logradouro é composto dos seguintes processos:

- Divisão das Notícias em Sentenças;
- Avaliação das Sentenças;
- Janelamento das Sentenças;
- Identificar Classes Gramaticais (*Postag* das Janelas);
- Definição dos Limiares e Criação das Regras.

4.3.1 Divisão das Notícias em Sentenças

A divisão das notícias em sentenças é a tarefa de dividir a notícia em uma unidade menor que possa ser analisada. A Figura 12 apresenta essa etapa no fluxo geral de criação de regras.

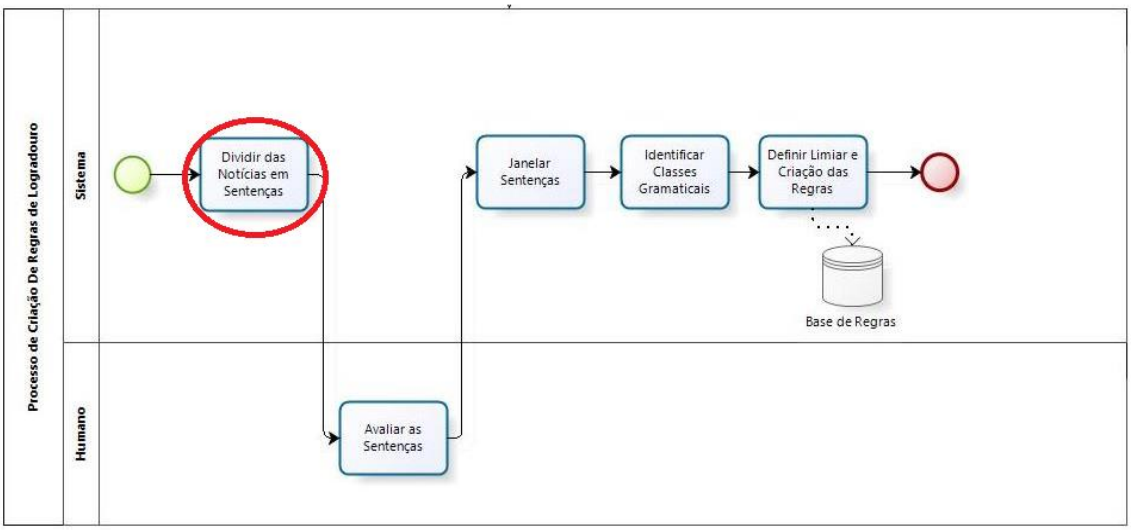

Figura 12: Divisão das Notícias em Sentenças

Essa etapa é responsável subdividir as notícias em sentenças. As notícias que atenderam ao filtro simples implementado na tarefa de coleta de notícia da seção 4.1.

Em geral, as notícias podem conter diversas sentenças, e não necessariamente todas as sentenças são referentes ao contexto da notícia ou possuem conteúdo relevante para esse trabalho.

Para a tarefa de separar as notícias em sentenças, neste trabalho foi utilizado o framework NTLK (BIRD, 2006). A Figura 13, apresenta o código necessário para a transformação das notícias em sentenças.

```
import csv
import nltk

# codigo para fazer a tokenização utilizando o NLTK para fazer o janelamento


def doQuery() :
    import psycopg2
    hostname = 'localhost'
    username = 'postgres'
    password = 'root'
    database = 'ufrj'
    conn = psycopg2.connect( host=hostname, user=username, password=password, dbname=database, port=5433 )
    cur = conn.cursor()

    ### Limpa tabela #######
    cur.execute("truncate table sentencas_noticias_Inaug")
    conn.commit()
    conn.close()

    ################### Gera as sentenças #############
    conn = psycopg2.connect( host=hostname, user=username, password=password, dbname=database, port=5433 )
    cur = conn.cursor()

    cur.execute( "SELECT id, descricao from inauguracao order by id" )
    #idx = 0

    for id, descricao in cur.fetchall() :
       #print ('##', i, '\n')
       text = descricao.replace('\n', ' ')
       for i, sent in enumerate(nltk.sent_tokenize(text)):
           cur.execute('INSERT INTO sentencas_noticias_Inaug (id_noticia, sequencia, linha) VALUES (%s, %s, %s)', (id, i, sent))



    conn.commit()
    conn.close()


doQuery()

print ('Finalizado')
```

Figura 13: Criação das Sentenças Utilizando NTLK

4.3.2 Avaliação das Sentenças

A avaliação das sentenças é a tarefa que envolve a participação humana em avaliar a validade de uma sentença para a extração de logradouro. A Figura 14 apresenta essa etapa no fluxo geral de criação de regras.

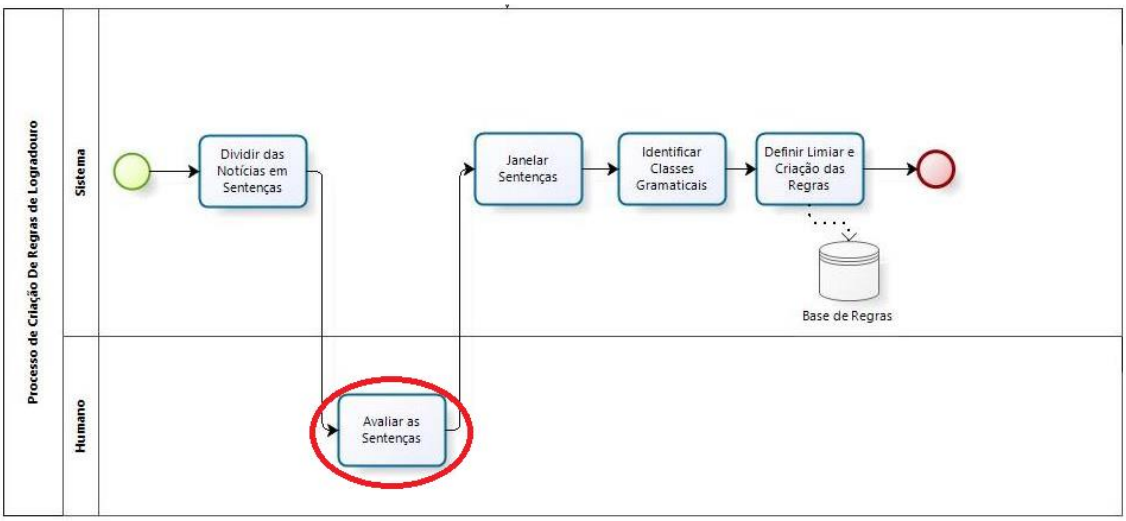

Figura 14: Avaliação das Sentenças

Para o processo de criação de regras é necessário a avaliação das sentenças, a partir da análise humana em determinar a validade ou invalidade das sentenças em relação à presença de logradouro válido.

Uma sentença é denominada válida, quando possui em sua estrutura um endereço válido e completo, conforme apresentado a seguir:

$$Palavra\ do\ (DPC) + nome\ do\ logradouro +\ acréscido\ ou\ não\ do\ número$$

A Tabela 10 apresenta exemplo das sentenças a serem avaliadas:

Tabela 10: Sentenças para Avaliação

Sentença	Válidas ou Inválidas
A inauguração da UPA foi realizada na rua Anália Pereira, 13.	Válida
A população não foi às ruas protestar contra a miséria.	Inválida

4.3.3 Janelar Sentenças

O janelamento das sentenças é a tarefa de dividir a sentença em uma unidade de maior delimitação. A Figura 15 apresenta essa etapa no processo de criação de regras.

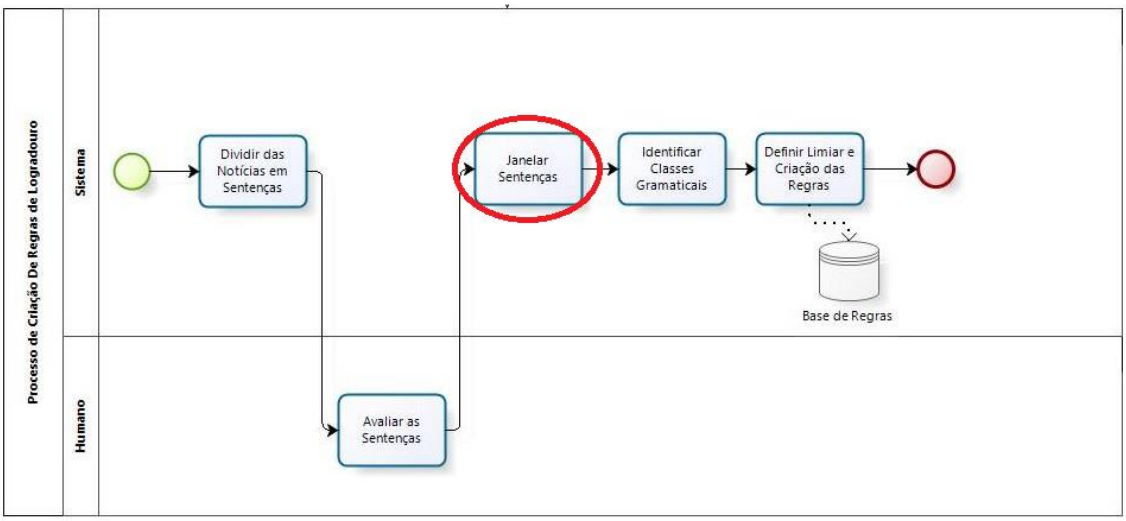

Figura 15: Janelar Sentenças

A separação de sentenças em janelas é uma tarefa muito importante, pois as sentenças podem conter mais de um elemento de logradouro dentro de um mesmo registro, conforme apresentado na Tabela 11.

Tabela 11: Exemplo de Janelamento

Sentença	Janelas
A Praça José de Alencar fica em um quadrilátero formado pelas ruas 24 de Maio e Liberato Barroso e é um dos lugares mais famosos da cidade.	rua 24 de Maio General Sampaio Guilherme Rocha e rua Liberato Barroso e é um.
A Secretaria de Obras investiu R 3 6 milhões nas intervenções das ruas Haddock Lobo e Voluntários da Pátria.	rua Haddock Lobo e rua Voluntários da Pátria.

Após a separação das sentenças em janelas, o sistema filtra as notícias que contenham elementos do DPC. As janelas selecionadas são denominadas como “janelas candidatas” de logradouro.

As sentenças que não atendem o dicionário de palavras são descartadas desse processo. As janelas que atendem ao DPC são utilizadas para o janelamento. O janelamento cria uma nova sentença a partir da sentença original, com a seguinte estrutura:

$$\boldsymbol{Janelamento} = \boldsymbol{Sentença}\,(\boldsymbol{Palavra\ DPC}, \boldsymbol{Palavra}[\mathbf{1}], \ldots \boldsymbol{Palavra}[\mathbf{5}])$$

Contudo a expressão que define o janelamento pode apresentar elementos não relevantes para o georreferenciamento da entidade. Para realizar o melhor refinamento da janela é necessário o uso de uma máquina de aprendizado para refinar as janelas baseada no aprendizado.

4.3.4 Identificar Classes Gramaticais

Essa tarefa é responsável por identificar as classes gramaticais dos termos que compõem a janela das notícias. A Figura 16 apresenta essa etapa no fluxo geral de criação de regras.

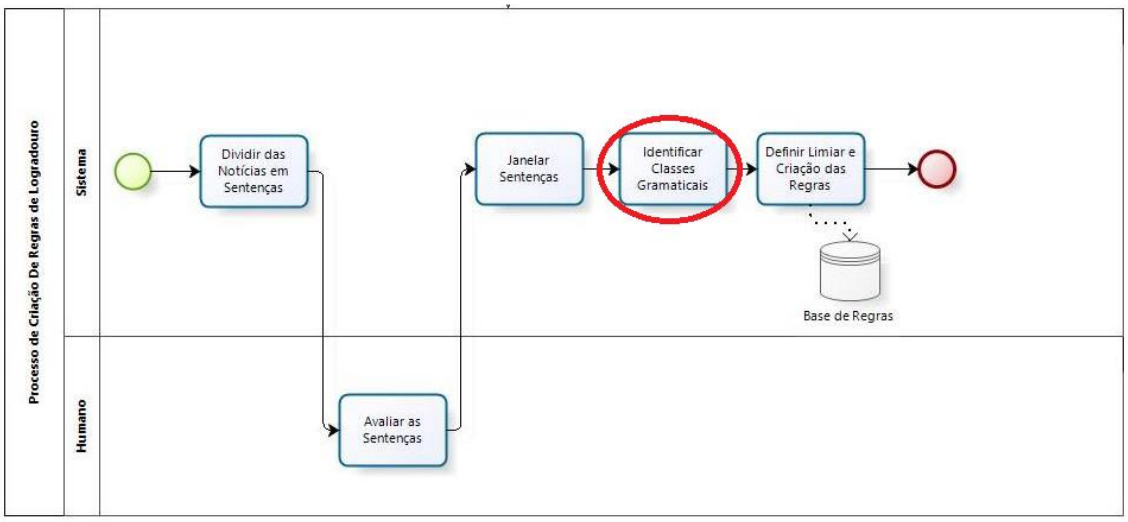

Figura 16: Identificar Classes Gramaticais

O *Postag* ou *PosTagger*, foi utilizado para diversas tarefas de processamento de texto, inclusive para a atividade de reconhecimento de endereços (MARQUES; LOPES, 2001). O resultado dessa etapa são janelas que contém as classes gramaticais dos termos correspondentes, para a formação das regras gramaticais.

4.3.5 Criação das Regras e Definição dos Limiares

A tarefa de criação das regras e definição do limiar é responsável por gerar todas as regras válidas e delimitar o melhor limiar. A Figura 17 apresenta essa atividade no fluxo de criação de regras.

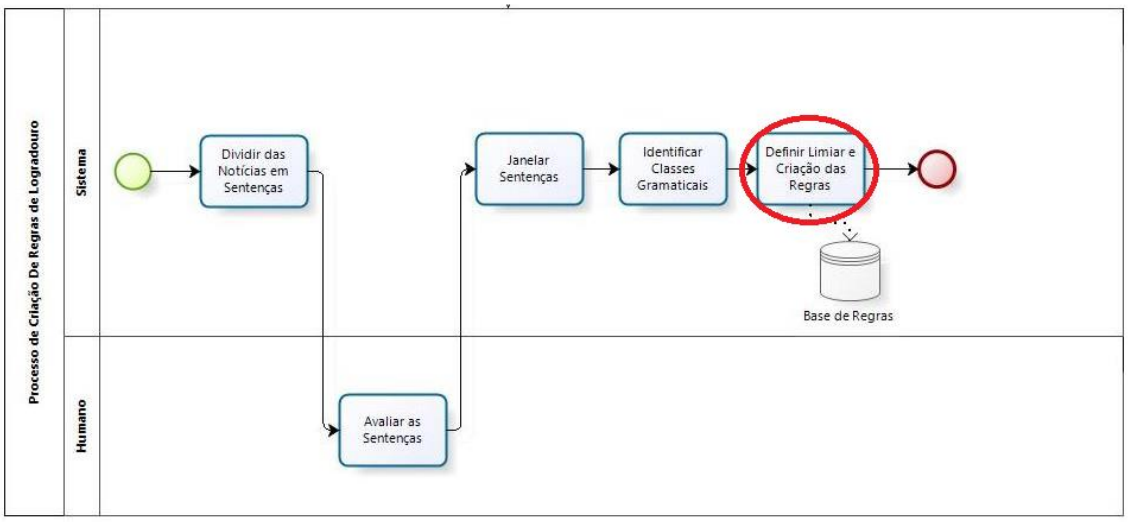

Figura 17: Definir Limiares e Criação das Regras

O estudo sobre o estabelecimento e definição dos valores adequados para determinar os melhores limiares, tem como objetivo estabelecer quais as melhores configurações de valor, que apresente o número adequado de regras válidas e com elevado grau de generalidade.

Os valores de limiares com um percentual de acerto da regra muito grande tende a ter uma grande quantidade de regras específicas, diminuindo a probabilidade de serem aplicadas a bases futuras.

Quando aumentamos a quantidade possível de gramas, consequentemente as regras se tornam mais específicas, ou seja, menor grau de generalidade, não atendendo ao princípio estabelecido para a criação de uma base de regras, o princípio de uma base que possa ser utilizada como referência na identificação de padrões de logradouro.

O processo de definição dos limiares e criação de regras está divido em duas etapas, a primeira etapa é a criação das regras e a segunda etapa é de definição dos limiares. Neste caso foi utilizada uma quantidade amostral de 4800 janelas. As seções a seguir presentam mais detalhes e implementações sobre os fluxos que compreendem as duas etapas.

4.3.6 Criação das Regras

O processo de criação das regras apresenta as etapas de refinamento das janelas e delimitação das gramas, e a criação das regras válidas. A Figura 18 apresenta as etapas:

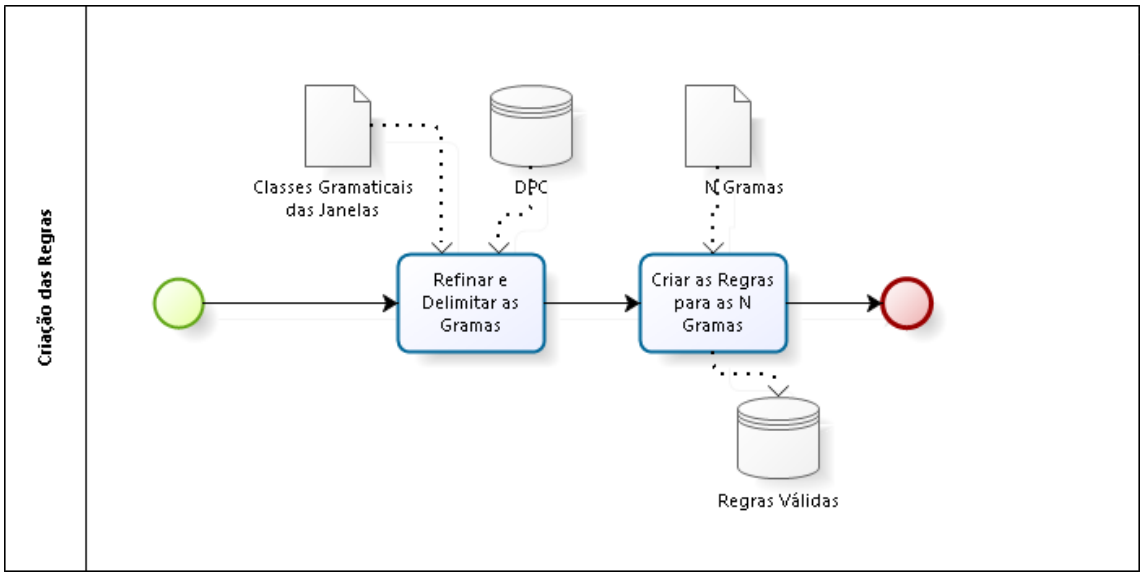

Figura 18: Criação das Regras

4.3.6.1 Refinamento das Janelas e Delimitação das Gramas

A primeira etapa da criação das regras consiste em obter o refinamento das janelas e a delimitação em gramas. As gramas correspondem aos termos, palavras ou nesse caso as classes gramaticais que correspondem a cada um dos termos de uma janela.

Nessa etapa para cada palavra pertencente ao DPC, são extraídos (N) gramas sequentes até um total de 5 gramas. A Tabela 12 representa a estrutura da janela refinada.

Tabela 12: Estrutura da Janela Refinada

Termo do DPC	**N1**	**N2**	**N3**	**N4**	**N5**	**N6**	**N (x)**
Palavra do DPC	Classe gramatical	Classe gramatical	Classe gramatical	Classe gramatical	Classe gramatical	Classe gramatical	Classe gramatical

Conforme apresentado na Tabela 12, as colunas de cor cinza são as gramas que não pertencem ao intervalo de N1 a N5 e serão descartados do processo de refinamento da Janela. O resultado dessa etapa são janelas delimitadas de acordo com o número de gramas.

4.3.6.2 Criação das Regras Válidas

A segunda etapa da criação de regras é o processo de criação e armazenamento das regras válidas. Esse processo é responsável por gerar as possibilidades para cada janela no intervalor de tamanhos de 1 à 5. A Figura 19, apresenta esse processo:

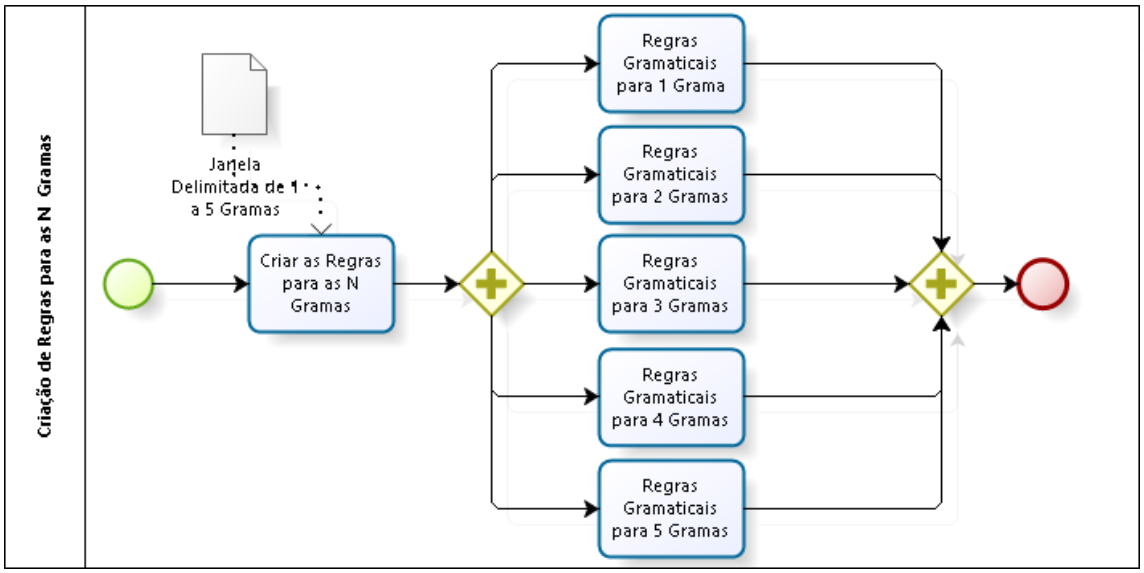

Figura 19 - Processo de Criação de Regras para as Gramas

Para as janelas que participaram do processo de avaliação de sentenças (4.3.2) são geradas as regras de grama. A Tabela 13 apresenta um exemplo das regras criadas para uma janela de exemplo:

Tabela 13: Exemplo de regras criadas para 1 janela

Gramas	Palavra DPC	Número de Grama (1)	Número de Grama (2)	Número de Grama (3)	Número de Grama (4)	Número de Grama (5)
Janela X	Avenida	Substantivo	Substantivo + verbo ou substantivo ou outra classe gramatical	Substantivo + 2 classes gramaticais	Substantivo + 3 classes gramaticais	Substantivo + 4 classes gramaticais

4.3.7 Definição dos Limiares

A definição dos limiares é responsável por duas tarefas, definir o intervalo de valores para os limiares restringir as regras válidas, e a determinação de parâmetros de escolha dos melhores limiares na etapa de combinação de valores dos limiares, conforme apresentado na Figura 20.

Figura 20: Definição dos Limiares

O primeiro processo dessa etapa consiste em determinar um intervalo de valores que os limiares podem assumir para cada número de grama, conforme apresentado na Figura 21.

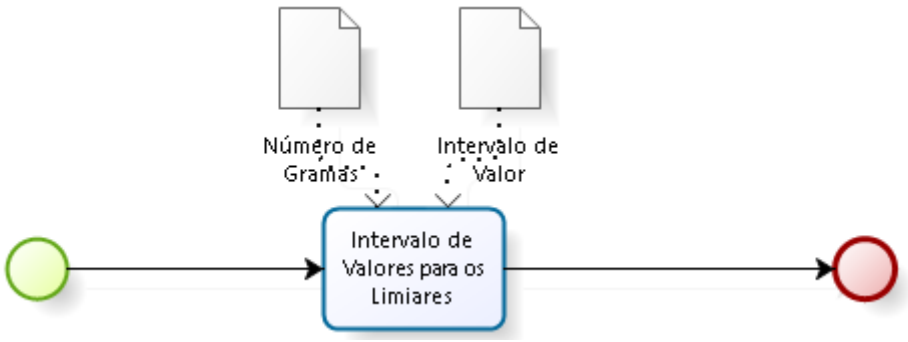

Figura 21: Intervalo de valores para os Limiares

Para essa tarefa foi estabelecido um total de 32.500, combinações possíveis de valores que os valores do limiar entre as gramas N1, N2, N3, N4 e N5. Os valores são testados com o valor atual da grama decrescendo de dois em dois, representada com a expressão:

$$\boldsymbol{Número\ da\ Grama\ =\ Valor\ Atual - 2}$$

A Tabela 14 apresenta o intervalo de valores utilizados para cada grama, valores esses utilizados como "valor atual" na expressão acima.

Tabela 14: Intervalo de Valores para os Limiares

Grama	Valor Inicial	Valor Final
N1	100	82
N2	80	62
N3	60	52
N4	50	42
N5	40	16

O estabelecimento dos valores para os limiares é necessário para filtrar as regras, de modo que uma regra é denominada como "boa" ou "válida" se o seu percentual de regra válida para a grama correspondente for maior que limiar determinado, por exemplo:

Percentual de regras válidas igual a 66%, para regra com número de grama igual a dois, e a regra gramatical: ***preposição + verbo***.

Neste caso os valores de limiar para o número de grama igual a dois, devem ser menores ou iguais a 66% para a regra ser considerada válida. Caso uma regra não seja considerada válida, é necessário continuar analisando as gramas maiores, neste caso, tamanho de grama igual a 3, 4 e 5. A necessidade de verificar as gramas maiores tem

como objetivo analisar a existência de regras de menor generalidade que atendem aos valores de limiar.

A Tabela 15 apresenta um exemplo de valores de limiares. Aplicando esses limiares valores sob a regra de tamanho de grama igual a 1 para a classe gramatical ***preposição***, com percentual de regras válidas de 66%, essa regra é denominada como inválida, pois não atende ao valor do limiar de uma grama, neste caso maior ou igual a 90%. Contudo, analisando número grama igual a 2 para a classe gramatical ***preposição*** + ***verbo***, com percentual de regras válidas de 82%, essa regra é denominada como válida, pois atende ao valor do limiar de duas gramas.

Tabela 15: Exemplo de Valores de Limiares

Gramas	Limiares
1	90%
2	80%
3	70%
4	50%
5	40%

O segundo processo dessa etapa consiste em determinar a melhor combinação de valores dos limiares. A Figura 22 apresenta esse processo.

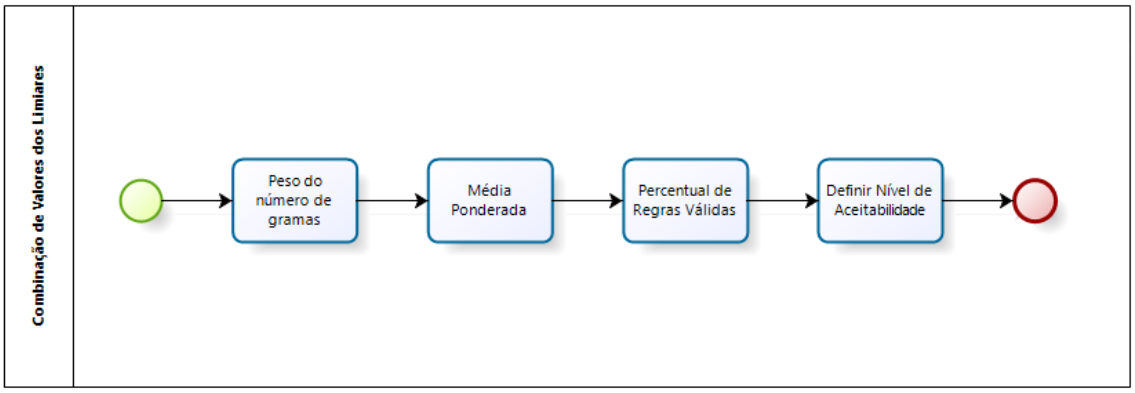

Figura 22: Processo Combinação de Limiares

Estabelecer a melhor combinação de limiares consiste em determinar a melhor relação de valores para a obtenção da maior de quantidade de regras com maior percentual de generalidade.

A maior quantidade de regras e o maior percentual de generalidade são importantes para construção da base de regras de maneira heterogênea, com o intuito de atender aos mais diversos contextos. Para a realização dessas premissas é necessário estabelecer alguns parâmetros:

- Peso do número de gramas;
- Média Ponderada;
- Percentual de Regras Válidas;
- Nível de Aceitabilidade.

4.3.7.1 Peso do Número de Gramas

As regras com maior valor de generalidade e maior percentual de acerto são ditas como melhores regras. Para analisar quais as melhores regras, nesse trabalho foi criada uma estrutura que atribui peso para as regras geradas, conforme o grau de generalidade, representado na Tabela 16:

Tabela 16: Relação de Peso por Grama

Número de Gramas da Regra	1	2	3	4	5
Peso em relação a generalização	5	4	3	2	1

A Figura 23 apresenta o grupo A e B. O grupo A é composto de N1, N2 e N3, este grupo é denominado de grupo com regras com maior generalidade ou mais geral. Por outro lado, o grupo B é composto de N4, N5 e N3, este grupo é denominado de grupo com regras de menor generalidade ou menos geral.

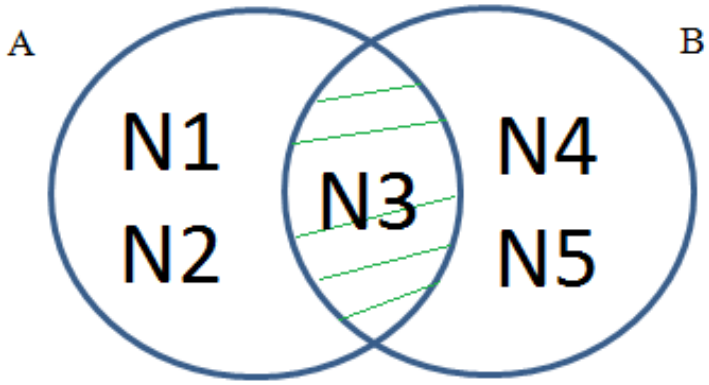

Figura 23: Relação de Generalidade por Grupos

4.3.7.2 Média Ponderada

A média ponderada consiste do somatório da quantidade de regras de cada grama multiplicado por seu peso correspondente, divido pelo número máximo de gramas, cinco nesse caso. A expressão a seguir apresenta o cálculo:

$$Média\ Ponderada = \sum_{i=1}^{n=5} \frac{(qtde\ de\ Regras\ de\ N(i)\ *\ Peso\ de\ N(i)}{n}$$

4.3.7.3 Percentual de Regras Válidas

Após obter as regras possíveis é necessário contabilizar as janelas denominadas válidas, baseado na etapa de avaliação das sentenças.

Na etapa de avaliação das sentenças com a participação humana, o contexto avaliado são as sentenças, contudo as sentenças contêm janelas, e são essas janelas que são avaliadas como validas ou invalidas, ou seja, se a sentença é válida a janela correspondente também é denominada de válida.

Para estabelecer o percentual de regras válidas, precisamos saber a quantidade total de janelas e a quantidade de janelas válidas para cada número de gramas, para cada regra existente. A Tabela 17 apresenta um exemplo de uma regra com número de grama igual a 2, e a regra gramatical: ***preposição + verbo***.

Tabela 17: Estabelecimento de Regras Válidas

Número de gramas	Identificador da Sentença	Regra	Janela da Regra é Válida ou inválida
2	123	Proposição + Verbo	Válida
2	124	Proposição + Verbo	Válida
2	188	Proposição + Verbo	Inválida

No exemplo acima, a regra gramatical ***preposição + verbo*** apareceu três vezes na etapa de avaliação de sentenças. A regra foi denotada como válida por duas vezes, temos então que o percentual de regras válidas da regra é de:

$$\boldsymbol{Percentual\ de\ Regras\ V\acute{a}lidas\ =\ \frac{Quantidade\ de\ regras\ v\acute{a}lida}{Numero\ de\ quantidade\ da\ regra}\ x\ 100}$$

Nesse caso, temos que:

$$\boldsymbol{Percentual\ de\ Regras\ V\acute{a}lidas\ =\ \frac{2}{3}\ x\ 100}$$

Percentual de regras válidas igual a 66%, para a regra com número de grama igual a dois. O mesmo cálculo apresentado anteriormente deve ser aplicado para todas as combinações de gramas e em todos os números de gramas de 1 a 5. O valor do número 5 (cinco) como a quantidade máxima de gramas para a formação de regras e limiares, foi estabelecido devido a necessidade do trabalho em obter regras com maior generalidade, pois conforme o número de gramas aumenta, o grau de generalidade diminui.

4.3.7.4 Nível de Aceitabilidade

O nível de aceitabilidade é a medida que define os melhores limiares. Seu resultado é importante, pois é ele que apresenta quais os limiares que possuem melhor percentual de regras válidas com maior quantidade de regras com maior grau de generalidade. A expressão a seguir apresenta o cálculo:

$$Nível\ de\ Aceitabilidade = \left(\frac{Percentual\ de\ Regras\ Válidas}{Média}\right)$$

Após inserir as possibilidades do intervalo de valores dos limiares, na Tabela 18 apresentamos um estudo com os diferentes valores de limiares gerados e os melhores resultados apresentados sob a perspectiva do melhor nível de aceitabilidade e menor quantidade de regras válidas.

Tabela 18: Relação das Melhores Regras

Regra	**Melhores Resultados**								
	N1	**N2**	**N3**	**N4**	**N5**	**Média**	**Percentual de Regras Válidas**	**Quantidade de Regras Válidas**	**Nível de Aceitabilidade**
Regra 1	82	68	54	42	16	69	91%	102	1,3188
Regra 2	84	62	54	42	16	89,80	93%	135	1,0356
Regra 3	90	72	54	42	16	99	94%	151	0,9495
Regra 4	92	62	54	42	16	106,80	94%	160	0,8801

Conforme apresentado na Tabela 18 o melhor resultado dos limiares é apresentado pela regra 1. Essa regra é dita como a melhor pois o valor do nível de aceitabilidade é o maior e a quantidade de regras válidas é a menor, permitindo maior generalidade e com alto percentual de regras válidas.

4.3.8 Base de Regras

A base de regras é responsável por armazenar as regras válidas que passaram pelas restrições de limiares. A base contém informações das regras de acordo com o número de gramas no intervalor de 1 a 5 para cada regra. A Tabela 19 apresenta uma exemplificação da base de regras para cada grama:

Tabela 19: Exemplificação das Base de Regras Geradas

Quantidade de Gramas				
N1	**N2**	**N3**	**N4**	**N5**
Proposição	Proposição + outra classe	Proposição + 2 outras classes	Proposição + 3 outras classes	Proposição + 4 outras classes
Artigo	Artigo + outra classe gramatical	Artigo + 2 outras classes gramaticais	Artigo + 3 outras classes gramaticais	Artigo + 4 outras classes gramaticais

A base de regras é utilizada na próxima seção 4.4 para filtrar as janelas das notícias que possuem em sua estrutura gramatical as regras pertencentes nessa base de regras.

4.4 Identificação de Janelas de Logradouro

Semelhante ao processo de criação das regras para logradouro proposto da seção 4.3, esta etapa utiliza as seguintes fases:

- Divisão das notícias em sentenças;
- Janelar as sentenças;
- Identificar as classes gramaticais.

As etapas que diferenciam as seções 4.3 e 4.4 são de filtragem das regras validas e armazenamento de janelas de logradouro válidas. A Figura 24 apresenta os fluxos que compõem essa etapa.

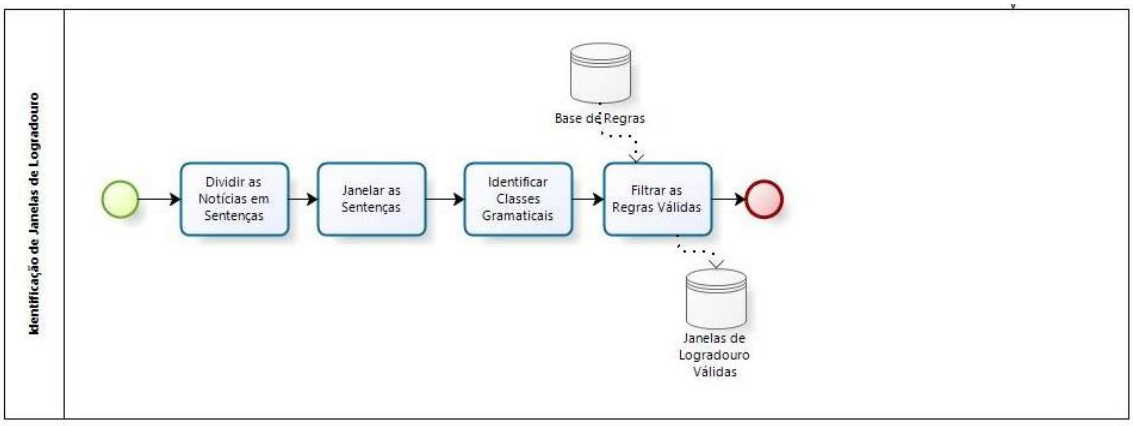

Figura 24: Fluxo de Identificação de Janelas com indicador de Logradouro Válido

4.4.1 Filtragem das Regras Válidas

A filtragem das regras válidas é a tarefa de permitir que apenas as notícias (janelas) válidas continuem no processo. A Figura 25 apresenta essa etapa no fluxo de identificação de janelas de logradouro.

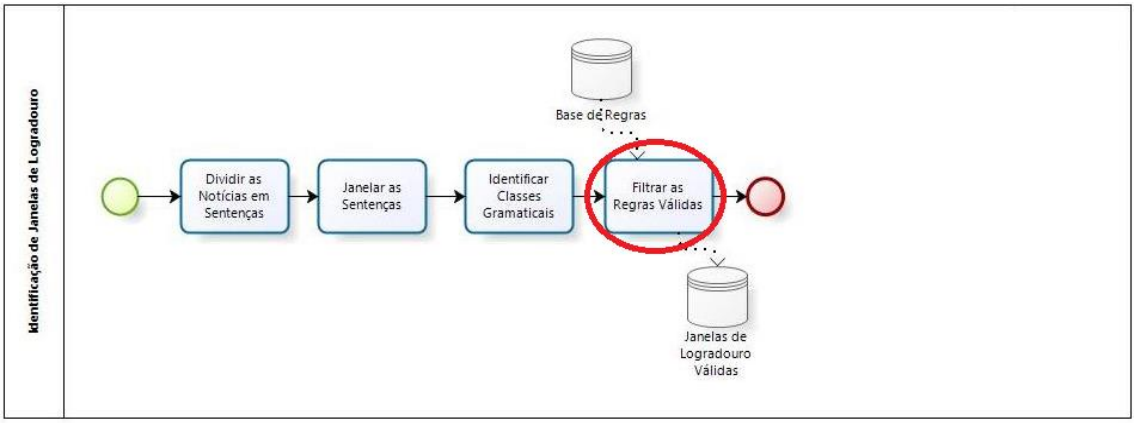

Figura 25: Filtragem das Regras Válidas

O processo de filtragem de regras válidas é responsável por selecionar apenas os registros que atendem as regras criadas de acordo com os limiares. Permitindo que os registros sejam selecionados de acordo com seu percentual de assertividade e generalidade. Os registros que atendem as premissas são armazenados na base de janelas de logradouro.

4.4.2 Armazenar as Janelas de Logradouro Válidas

A base de janelas de logradouro válidas é responsável em armazena os resultados válidos que serão utilizados na tarefa de aprendizado de máquina. A Figura 26 apresenta essa etapa no fluxo de identificação de janelas de logradouro.

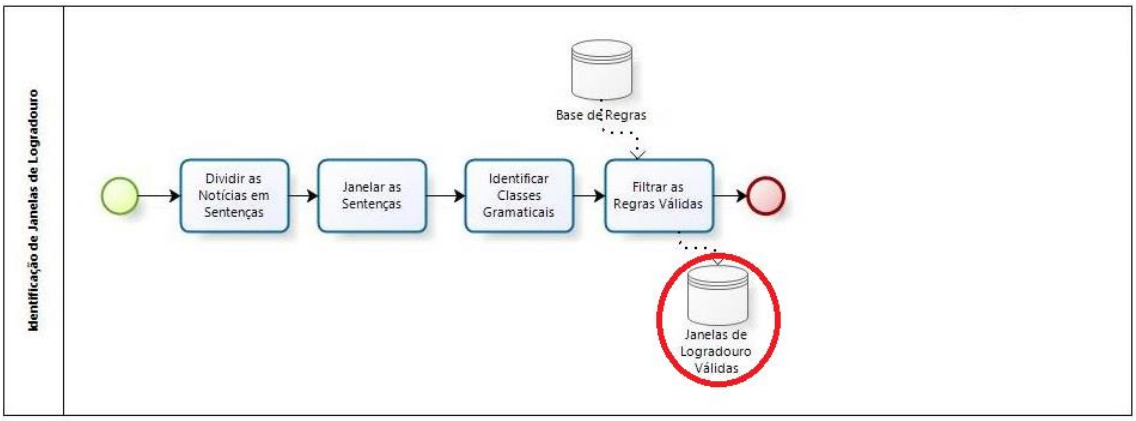

Figura 26: Janelas de Logradouro Válidas

Nessa etapa as janelas que passaram pelos filtros são armazenadas. Contudo a janela não tem seu conteúdo delimitado suficientemente para o reconhecimento de entidade, conforme apresentado na seção 4.3.3. Para a melhor delimitação da janela da notícia, é necessária a participação de uma máquina de aprendizado.

4.5 Aprendizado de Máquina

A tarefa de aprendizado de máquina consiste no refinamento ou delimitação dos registros armazenados na base de janelas de logradouro válidas.

Neste trabalho, o aprendizado de máquina é utilizado para extrair apenas os elementos que representam o grupo de entidades que representam logradouro. A máquina de aprendizado foi utilizada devido a necessidade de obter o refinamento das janelas para que estas tenham apenas entidades que permitem obter logradouros passíveis de serem geocodificados. A Figura 27 apresenta o fluxo geral do processo de aprendizado de máquina.

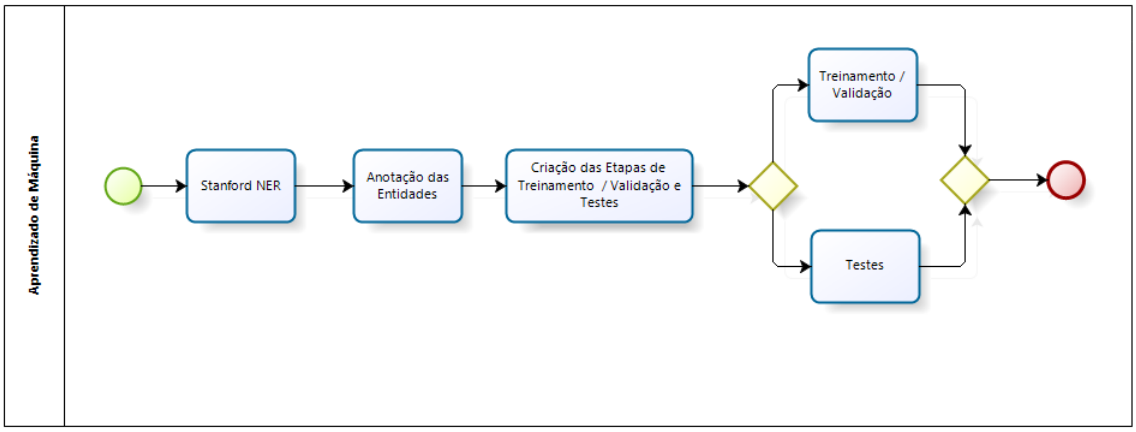

Figura 27: Processo de Aprendizado de Máquina

A Tabela 20 apresenta um exemplo da necessidade de utilização da máquina de aprendizado para a identificação de entidades.

Tabela 20: Uso da Máquina de Aprendizado

Janela	Antes da Máquina de Aprendizado	Depois da Máquina de Aprendizado
	Rua Jose da Silva foi inaugurado	Rua Jose da Silva

A Tabela 20 apresenta um exemplo da importância de utilizar a máquina de aprendizado, visto que apenas com a utilização de regras gramaticais não é possível obter uma delimitação dinâmica que atenda todas as janelas, pois os tamanhos dos logradouros são variados, e neste caso cabe a máquina de aprendizado definir e efetuar a delimitação necessária, removendo os termos que não colaboram nesse processo. As seções a seguir apresentam os elementos de aprendizado de máquina utilizados nesse trabalho.

4.5.1 Stanford NER

Para a tarefa de aprendizado supervisionado de máquina, nesse trabalho utilizamos o Stanford NER (2016). A escolha pelo Stanford NER deve-se ao fato da facilidade de adaptações que a ferramenta dispõem e a implementação nativa do algoritmo de "campos aleatório condicional", em inglês, *Conditional Randon Fields* (CRF) (LAFFERTY; MCCALLUM; PEREIRA, 2001) .

O CRF é um método de modelagem estatística utilizado no reconhecimento de padrões e no aprendizado de máquina, especialmente no aprendizado supervisionado. Uma vez que em geral, os classificadores analisam apenas o termo para determinar o rotulo da entidade, o CRF verifica os termos que estão ao redor para prever o rotulo adequado ao termo. Em resumo podemos assumir que o CRC consiste em uma maneira

de melhorar o nível de aceitabilidade na tarefa de rotular os termos (SUTTON; MCCALLUM, 2006).

A importância de aplicar o CRF para a tarefa de reconhecimento de entidade em textos da língua portuguesa deve-se ao fato de que essa possibilita a extração automática de entidades a partir de um conjunto de dados com uma capacidade de resposta mais rápida do que outras técnicas já utilizadas, como a implantação de heurísticas específicas (MOTA; SANTOS, 2008b).

O método CRF representa o estado da arte para as tarefas desempenhadas, além de prover treinamento discriminativo e com maior grau de sucesso nas tarefas de reconhecimento de entidades mencionadas (FINKEL; GRENAGER; MANNING, 2005).

Na língua portuguesa, a aplicação do CRF para as tarefas de reconhecimento de entidades mencionadas, apresenta bons resultados, principalmente o utilizando aplicado ao processo de aprendizado supervisionado de máquina, pois o algoritmo apresenta uma capacidade de resposta mais rápida e adequada se comparado a outros algoritmos. (AMARAL; VIEIRA, 2014). Exemplos da importância do CRF em português podem ser encontrados nos trabalhos de AMARAL; VIEIRA (2014), DA SILVA; DE MEDEIROS CASELI (2015) e BATISTA et al., (2010) que utilizou CRF para identificar nomes entidades geográficas em textos.

4.5.2 Anotação das Entidades

Para as etapas de treinamento/validação e testes, requeridos no aprendizado de máquina é necessário a anotação manual das entidades existentes na notícia, possibilitando o aprendizado geral da ferramenta de aprendizado supervisionado, ou seja, a tarefa de anotação manual das notícias é importante para o sucesso da tarefa de reconhecimento de entidades mencionadas, pois através dessa atividade a máquina aprende padrões em linguagem natural necessário para a classificação dos novos elementos contidos nos textos.

Na conferência HAREM o processo de anotação manual das notícias, foi realizado com a participação de um grupo de 10 especialistas, que tinham a responsabilidade de anotar as entidades em seus respectivos grupos correspondentes (SANTOS et al., 2007).

A ferramenta Stanford NER apresenta duas formas de anotar as notícias, denominadas IO (padrão Stanford) e BIO. O método de anotação por IO é a mais

popular para a tarefa, principalmente pela significativa rapidez em marcar as entidades nos textos.

Na anotação IO as entidades são denotadas com as seguintes expressões:

$$\boldsymbol{IO = Termo\ +\ espaço\ + Tipo\ da\ Entidade}$$

A Tabela 21 apresenta um exemplo utilizando a notação IO necessário para a criação dos arquivos de anotação de entidades utilizado pelo Stanford NER.

Tabela 21: Representação IO

Termo	Entidade
Foi	O
Inaugurado	O
Hospital	TIPO
Na	O
Rua	LOGRADOURO
Jose	LOGRADOURO
Silva	LOGRADOURO

Os termos que não possui valor significativo para o trabalho são denotados com a letra "O", significando "outros".

Na anotação BIO as entidades são denotadas com o acréscimo de um termo BIO. O termo BIO identifica um termo aglutinado referente a posição do termo no conjunto de elementos. A letra "B" é atribuída ao primeiro termo que comprem a aglutinação, a letra "I" é atribuída aos demais termos que compõem a aglutinação, enquanto a letra "O", semelhante a notação IO, identifica os termos irrelevantes.

$$\boldsymbol{BIO = Termo\ + Hífen + Termo(BIO) +\ espaço\ + Tipo\ da\ Entidade}$$

A Tabela 22 apresenta exemplo de anotação BIO necessário para a criação dos arquivos de anotação de entidade utilizado pelo Stanford NER.

Tabela 22: Representação BIO

Termo	Termo BIO	Entidade
Foi	-O	O
Inaugurado	-O	O
Hospital	-B	TIPO
Na	-O	O
Rua	-B	LOGRADOURO
Jose	-I	LOGRADOURO
Silva	-I	LOGRADOURO

4.5.3 Criação das Etapas de Treinamento/Validação e Testes

Para o aprendizado de máquina é necessário um conjunto de procedimentos que permita que a máquina de aprendizado crie um modelo de aprendizado supervisionado com um alto valor nas medidas-F, precisão e abrangência.

Uma quantidade amostral de registros oriundos da base de janelas de logradouro válidas serve como insumo nesse processo, divididos em conjuntos ou *folds*, e são separados em dois fluxos denominados: Treinamento/Validação e Testes, para a realização do *K-fold Validation*.

4.5.3.1 Treinamento/Validação

O processo de treinamento e validação consiste em treinar a máquina para o reconhecimento de entidades mencionadas e a identificação dos parâmetros que proporcionem as melhores métricas.

Nessa etapa os *folds* correspondentes a treinamento/validação são submetidos ao Stanford NER com diversos parâmetros utilizados pelo algoritmo CRF. A cada mudança do valor dos parâmetros, o algoritmo calcula a média das medidas de precisão, abrangência e medida-F para delimitar qual o melhor parâmetro para a máquina de aprendizado.

Neste trabalho foi testado um total de 1300 variações, com diversos parâmetros e valores "intervalos de valores testados", escolhidos baseado nas características das particularidades do CRF, apresentados na Tabela 23, de maneira a identificar quais os parâmetros que possuem os melhores resultados.

Tabela 23: Parâmetros Utilizados

Parâmetros	Descrição	Intervalo de Valores Testados
maxNGramLen	Define a quantidade máxima de n-gramas que será utilizada para determinar a entropia.	1 a 10
useClassFeature	Insere uma prévia sobre as classes que equivale a quantas vezes o recurso apareceu nos dados de treinamento.	Verdadeiro ou Falso
useNGrams	Análise baseado nas características das letras dos termos, ou seja, *substrings* do termo.	Verdadeiro ou Falso
usePrev	Analisa a entropia utilizando o termo anterior, semelhante ao uso da notação BIO.	Verdadeiro ou Falso
useNext	Analisa a entropia utilizando o termo posterior, semelhante ao uso da notação BIO.	Verdadeiro ou Falso
useSequences	Utilizar o recurso de combinação entre os grupos de entidade para analisar o termo.	Verdadeiro ou Falso

A Figura 28 apresenta o arquivo de parâmetros que apresentou o melhor resultado.

```
#location of the training file
trainFile = C:/base_file.tok
#location where you would like to save (serialize to) your
#classifier; adding .gz at the end automatically gzips the file,
#making it faster and smaller
serializeTo = C:/Desktop/ETLValidacaoKfolds/CROSSVALIDATION/processamento/BaseCompleta.ser.gz
#structure of your training file; this tells the classifier
#that the word is in column 0 and the correct answer is in
#column 1
map = word=0,answer=1
#these are the features we'd like to train with
#some are discussed below, the rest can be
#understood by looking at NERFeatureFactory
useClassFeature=true
useWord=true
useNGrams=true
#no ngrams will be included that do not contain either the
#beginning or end of the word
noMidNGrams=true
useDisjunctive=true
maxNGramLeng=3
usePrev=true
useNext=true
useSequences=true
usePrevSequences=true
maxLeft=1
#the next 4 deal with word shape features
useTypeSeqs=true
useTypeSeqs2=true
useTypeySequences=true
wordShape=chris2useLC
```

Figura 28: Melhor Resultado de Parâmetros

4.5.3.2 Testes

O processo de testes consiste em avaliar a capacidade dá máquina de aprendizado na tarefa de classificação das entidades mencionadas. A Figura 29 apresenta esse processo.

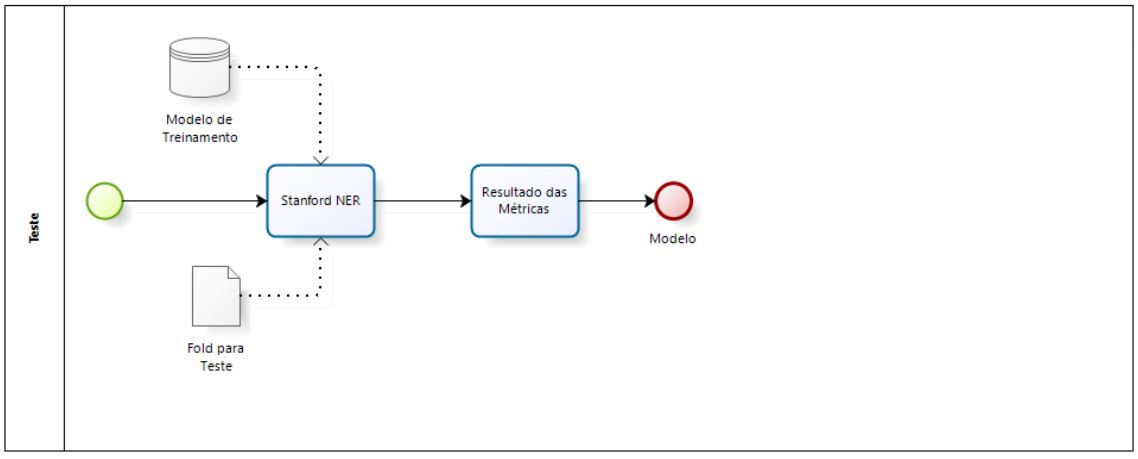

Figura 29: Processo de Testes

Nessa etapa os *folds* do processo de treinamento/validação em conjunto com o fold de teste são submetidos ao processo de *k-fold Validation* utilizado no Stanford NER, esse processo permite que a máquina avalie seu aprendizado médio para a classificação de elementos baseado nas métricas de precisão, abrangência e medida-f.

Essa etapa permita a comparação dos valores médios obtidos no processo, além de disponibilizar o modelo de aprendizado para o sistema web (capítulo 5). Esse sistema Web utiliza o modelo criado para classificar os novos elementos analisados.

5 Formação e Visualização do GeoNewsBR

Esse capítulo apresenta a formação e visualização do GeoNewsBR, uma aplicação web desenvolvida para verificar a validade das etapas propostas de pré-processamento e o reconhecimento de entidades. Nessa seção também foi desenvolvido a funcionalidade de georeferrenciamento dos elementos de baixa granularidade, além da formação do dicionário geográfico.

As funcionalidades disponibilizadas pelo GeoNewsBR incluem: (1) formação do dicionário geográfico; (2) geocodificação dos logradouros presente no dicionário geográfico; (3) apresentação das informações contidas no dicionário geográfico; (4) a identificação de logradouro em notícias da internet; (5) visualização dos logradouros em um mapa. A Figura 30 apresenta as etapas da plataforma GeoNewsBR.

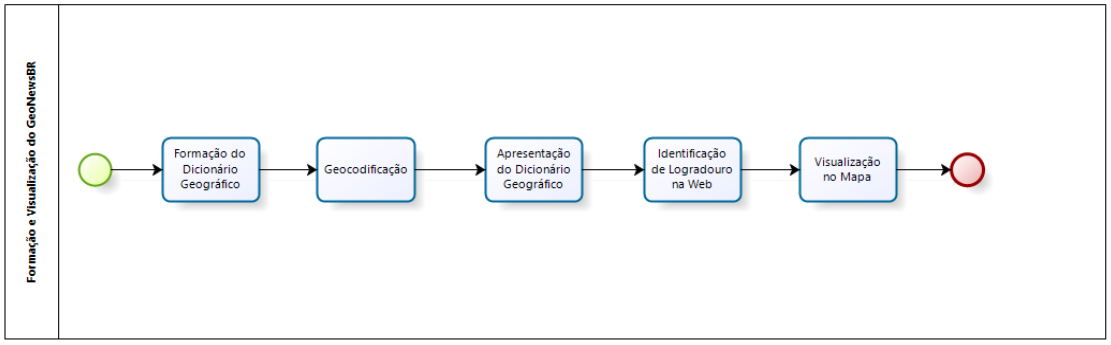

Figura 30: Fluxo de Atividades Formação e Visualização do GeoNewsBR

5.1 Formação do Dicionário Geográfico

O processo de formação do dicionário geográfico está organizado em duas atividades, relacionar os elementos de menor granularidade com os demais grupos de elementos e a criação do dicionário geográfico. A Figura 31 apresenta uma visão dessas atividades:

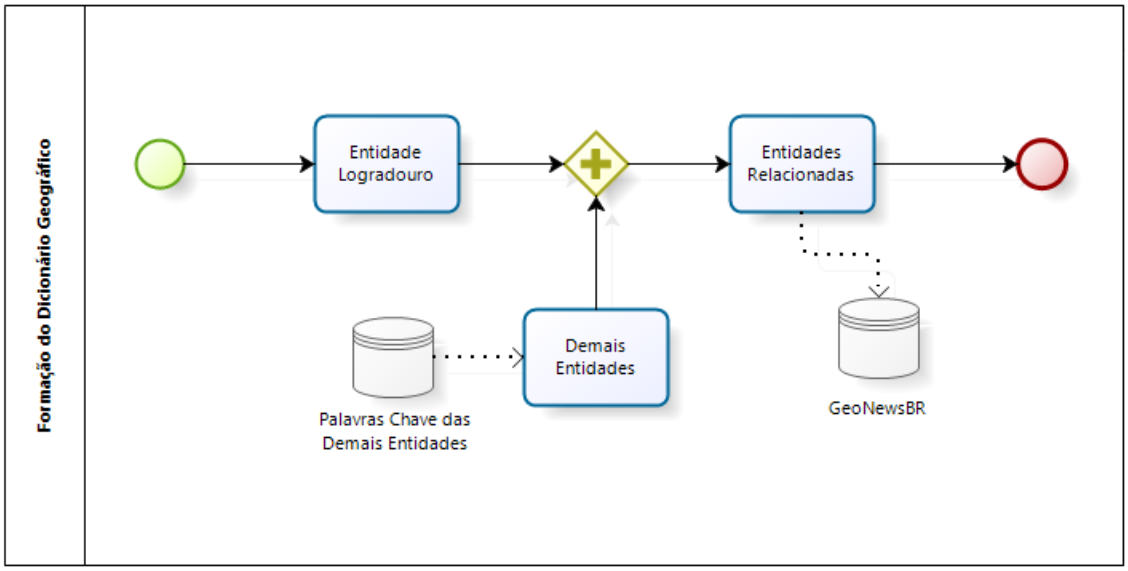

Figura 31: Processo de Formação do Dicionário Geográfico

O processo de relacionar as demais entidades existentes nas notícias que são necessárias para apoiar as atividades de criação do dicionário geográfico. Essas atividades correspondem em relacionar um endereço geográfico de baixa granularidade identificado com um determinado grupo de entidades, neste caso Tipo e Local.

Apesar do trabalho não levar em consideração a identificação de elementos que não pertencentes ao grupo logradouro para cumprir a tarefa de criação do dicionário geográfico e apoiar as atividades ligadas à reambulação, é necessário relacionar os elementos na notícia que identifiquem localidade e tipo na notícia. A Tabela 24 detalha os tipos de categorias e as palavras chave que identificam as categorias nas notícias:

Tabela 24: Demais Categorias de Entidades

Categoria	Descrição	Palavras Chave
Tipo	Descreve qual é a construção ou edificação.	Hospital, UPA, Unidade de pronto atendimento, Unidade básica de saúde, Clinica, Unidade básica de saúde, Maternidade, Pronto socorro, Policlínica, Escola, Universidade, Biblioteca, Creche, Pré-escola, Museu, Rodoviária, Aeroporto, Estrada, Rodovia, BRT, Quadra esportiva, Ginásio, Praça, UPP, Posto policial e Delegacia.
Local	Identifica a localização de maior granularidade em construção ou edificação.	Todas as cidades do Brasil.

A atribuição dos demais elementos é realizada com uma busca simples da ocorrência de palavras chave no conteúdo das notícias. Uma notícia que contenha uma ou mais palavras chave, por exemplo: “foi inaugurado novo hospital na Rua Oswaldo Cruz 13”. Neste caso a palavra hospital é pertencente ao conjunto de palavras chave, atribuindo o elemento Tipo igual a “hospital” para a notícia.

A criação do dicionário geográfico consiste em armazenar os elementos identificados pela máquina de aprendizado em um banco de dados denominado dicionário geográfico.

Para a formação do dicionário geográfico são utilizados os registros da segunda etapa de coleta de notícia, proposta no capítulo 4. Contudo esses registros não completam a formação do dicionário geográfico, necessitando ainda o georeferrenciamento dos elementos geográficos.

5.2 A Geocodificação do Dicionário Geográfico

A etapa de geocodificação dos elementos de baixa granularidade contidos no dicionário geográfico é importante para atribui as coordenadas de latitude e longitude no dicionário geográfico.

Para realizar a geocodificação, utilizamos a *API* do Google de georreferenciamento, acessada via um *webservice* desenvolvido para o GeoNewsBR. A aplicação pesquisa a latitude e longitude dos endereços existentes no dicionário geográfico e atribui as coordenadas correspondentes baseadas na *API*. A Figura 32 apresenta esse processo.

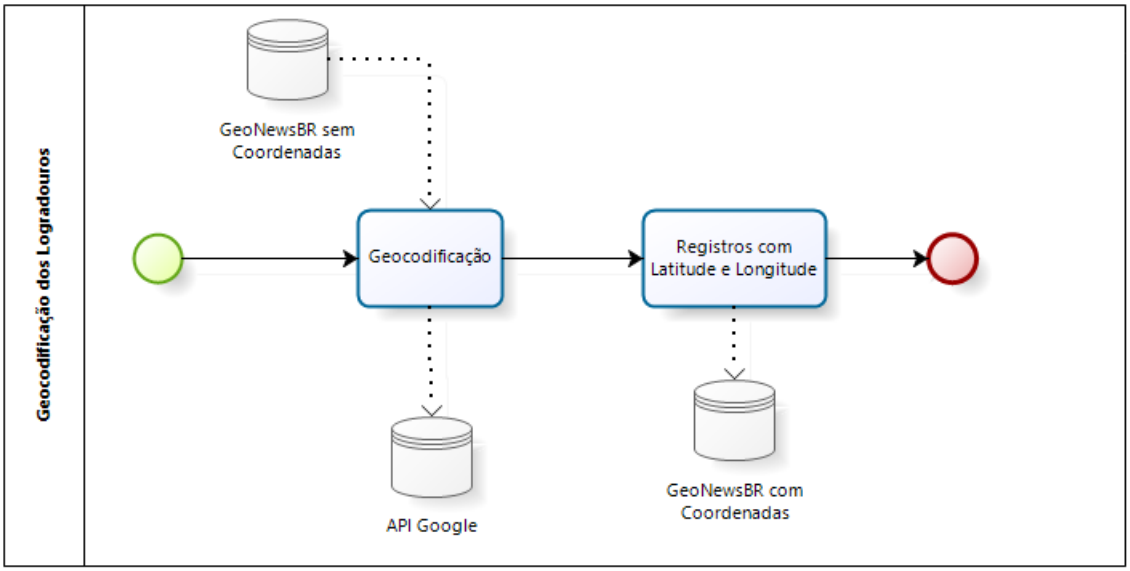

Figura 32: Processo de Geocodificação

5.3 Apresentação das Informações Contidas no Dicionário Geográfico

A etapa de apresentação das informações contidas no dicionário geográfico é responsável por apresentar o dicionário completo, contendo a categoria da entidade, o tipo, logradouro e as coordenadas (latitude e longitude).

Para a apresentação das informações contidas no dicionário geográfico é necessário que o sistema web (GeoNewsBR), faça a leitura dos registros do dicionário geográfico atualizado com as coordenadas obtidas no processo de georreferenciamento.

O desenvolvimento dessa etapa permite que usuários, pesquisadores e órgãos governamentais possam pesquisar as informações contidas nessa base. A Figura 33 apresenta a tela de apresentação das informações contidas no dicionário geográfico.

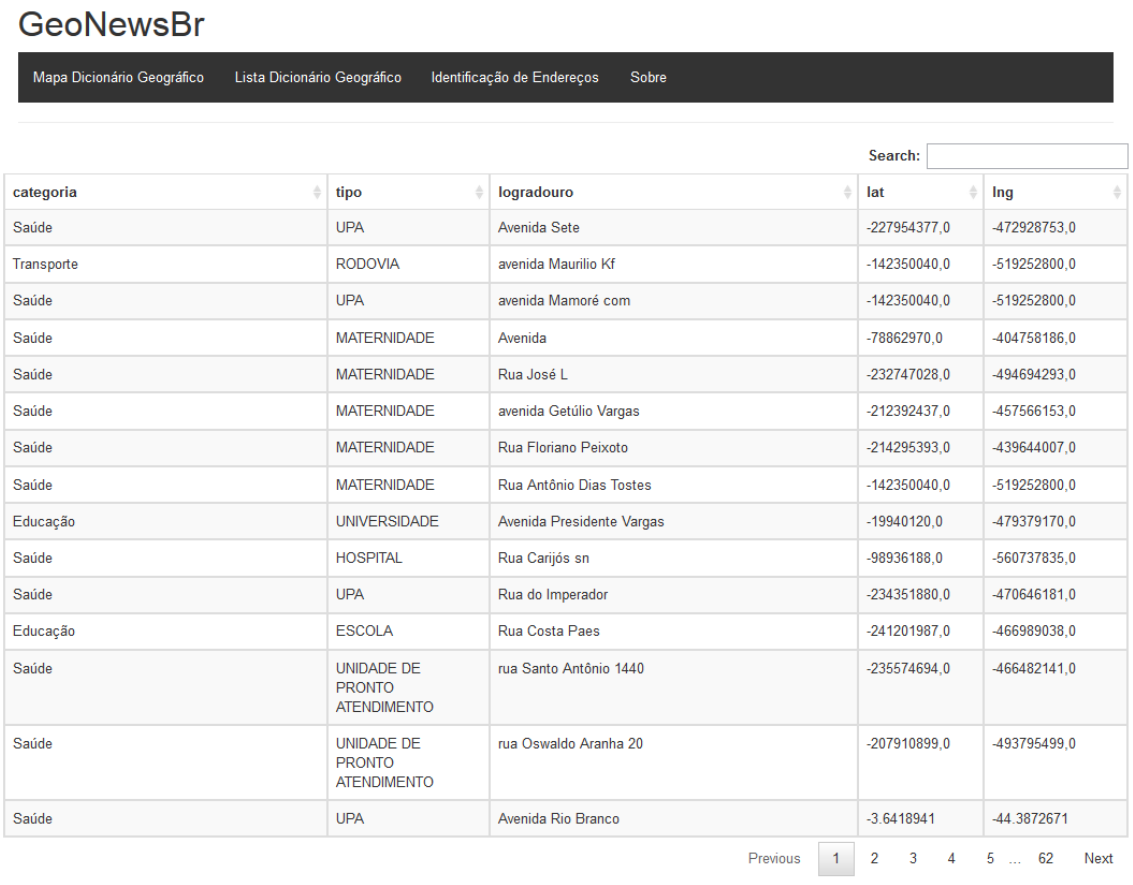

Figura 33: Tela com as Informações do Dicionário Geográfico

5.4 Identificação de Endereços em Notícias da Internet

A etapa de identificação de endereços em notícias da internet, pode ser utilizado para apoiar validar e visualização das etapas propostas. A Figura 34 apresenta as etapas desse processo.

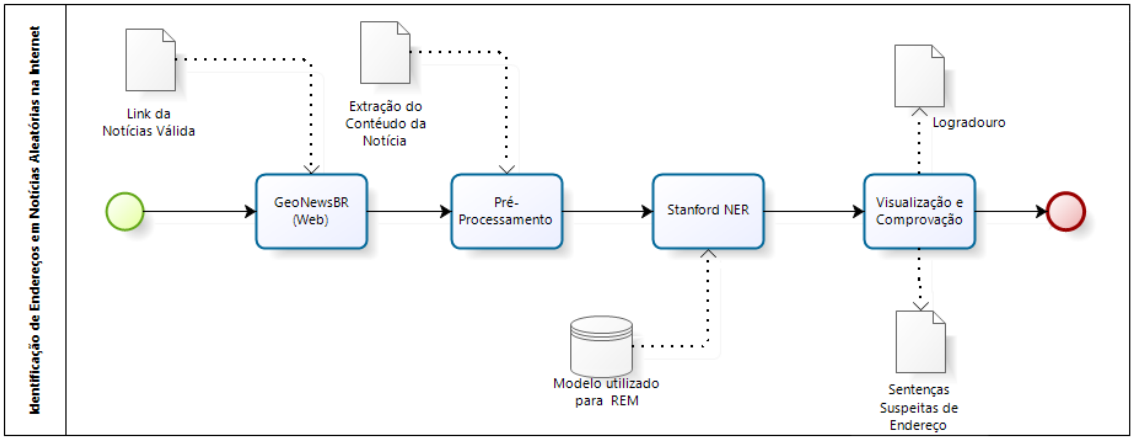

Figura 34: Processo de Identificação de Endereços em Notícias da Internet

Na etapa de identificação de endereços em notícias na internet, há a possibilitando de validar através de uma ferramenta visual as etapas propostas no trabalho. Para realizar a validação e visualização é necessário que um link de notícia válida, contendo endereço de baixa granularidade no corpo do parágrafo, seja coletado manualmente na internet e inserida na plataforma web do GeoNewsBR.

O sistema executa as atividades de pré-processamento e aprendizado de máquina com os modelos e definições apresentados no capítulo 4. Após a execução das tarefas, o resultado é apresentado na tela com as informações de logradouro e as sentenças suspeitas de conter elementos geográficos de baixa granularidade. A Figura 35 apresenta essa funcionalidade.

Figura 35: Identificação de Endereços Geográficos da Internet

5.5 Visualização das Localidades no Mapa

A etapa de visualização das localidades no mapa é a funcionalidade que permite analisar os registros presente na base do dicionário geográfico, através de um mapa interativo, utilizando a API de mapas do Google.

Para a visualização em mapa, primeiramente foi utilizado o processo de identificação de endereços em notícias da internet, proposto na seção 5.4. As coordenadas resultantes servem de parâmetro para a API de mapas do Google, apresentando no mapa a localidade geográfica, conforme apresentado na Figura 36.

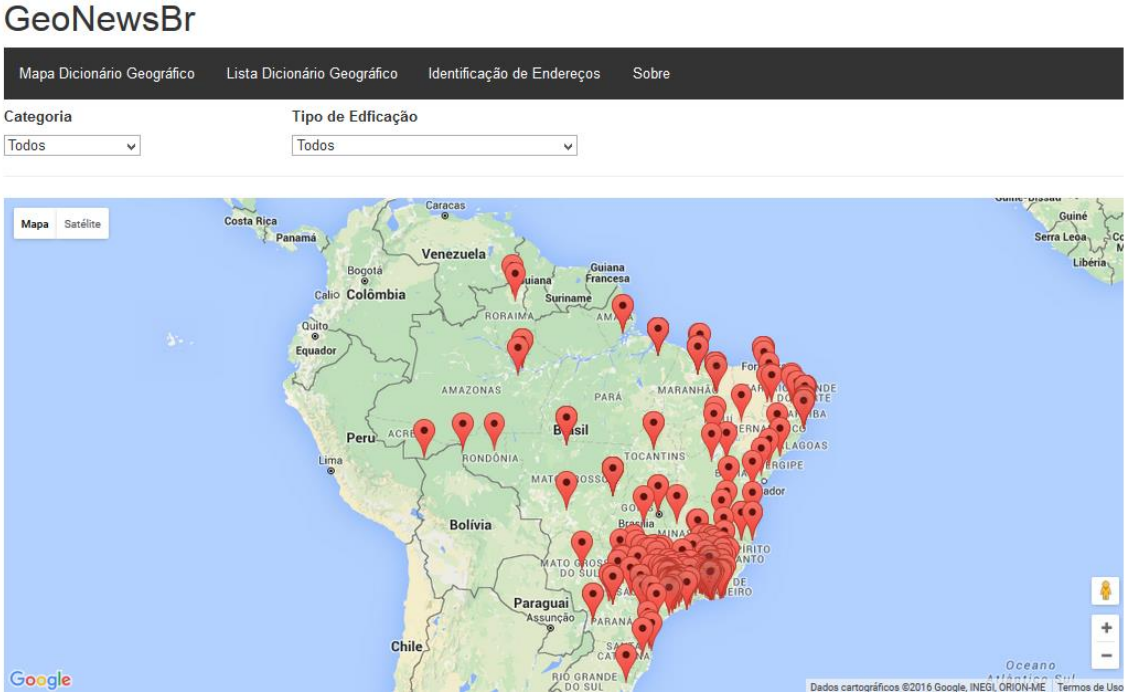

Figura 36: Tela de Visualização no Mapa

6 Avaliação dos Resultados utilizando o GeoNewsBR

Neste capitulo, o objetivo é verificar a necessidade do uso das etapas de pré-processamento, comparar diferentes anotações de entidades e analisar as métricas obtidas na máquina de aprendizado, nas tarefas de identificação do conteúdo geográfico de baixa granularidade.

Este capítulo detalha os principais experimentos realizados durante este trabalho, os métodos utilizados, bem como os resultados obtidos, que são descritos de forma a embasar as conclusões finais desse trabalho.

6.1 Formação dos Corpora

Para comprovar a validade da ferramenta de apoio na identificação de entidades de baixa granularidade, mais especificamente logradouro, foram criados duas *corpora*[7] de notícias.

Os corpora criados possuem um total de 250 notícias aleatoriamente escolhidas através do processo de coleta de notícias, seguindo as regras e filtros apresentados no capítulo 4, seção 4.2. Conforme relatado na seção (1.2) de delimitação do trabalho, visto a particularidade específica dos *corpora*, semelhante aos trabalhos de GOUVÊA (2008) e MACHADO et al., (2011), nesse trabalho não foi utilizado o corpus baseline da HAREM ou de outra conferência. Os *corpora* foram exclusivamente criados com objetivo de avaliar o trabalho proposto e são descritos como: Corpus Baseline e Corpus 1 (C1).

Corpus Baseline: Esse corpus foi criado com o propósito de ser utilizado como fator de comparação. Para essa finalidade, possuem 250 notícias aleatórias resultantes do processo de coleta de notícias. Suas notícias não foram pré-processadas, limpas ou modificadas, preservando a autenticidade e originalidade do texto, conforme disponibilizado em sua fonte de origem.

Corpus C1: Esse corpus foi criado com o propósito de ser utilizado como fator de validação e verificação das etapas e processos apresentados nesse trabalho. Para essa finalidade, possuem as mesmas 250 notícias aleatórias resultantes do processo de coleta de notícias. Suas notícias foram pré-processadas utilizando as etapas do capítulo 4, e as janelas foram delimitas e refinadas utilizando as tarefas e modelo aprendizado de máquina, também proposto nesse trabalho.

---

[7] Plural de corpus.

Para a realização dos experimentos, foi necessário que o corpus Baseline apresentasse alguns padrões que permitam o uso das métricas propostas nesse trabalho, como a anotação das entidades e a utilização do *k-fold Validation* para treinar e testar a máquina de aprendizado em condições semelhantes às utilizadas no pré-processamento. A Figura 37 apresenta o fluxo geral das etapas utilizadas para a realização dos experimentos e analise dos resultados:

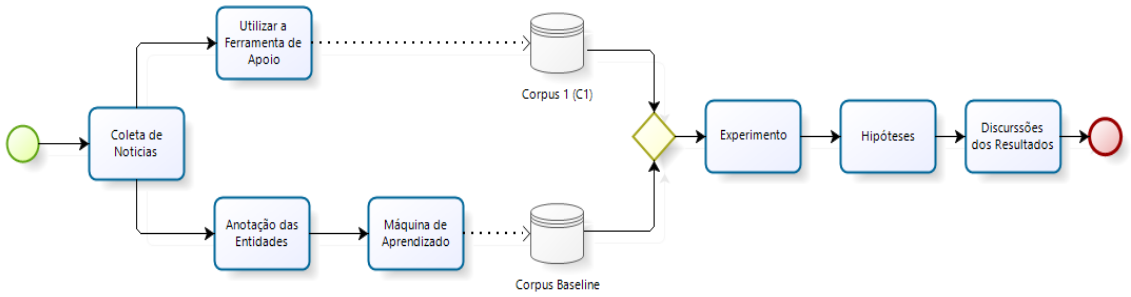

Figura 37: Estrutura de Criação dos Corpos dos Experimentos

6.2 Primeira Execução dos Experimentos: Baseline

A primeira rodada de experimento está interessada no estudo da influência dos elementos de pré-processamento propostos nesse trabalho, através da comparação entre os corpora baseline e C1. Nesse experimento esperamos também, realizar a análise da influência de diferentes padrões de anotação de entidades nomeadas.

Para realização da análise de influência dos diferentes tipos de anotação de entidade, o corpus baseline foi subdividido em baseline com padrão IO e baseline com para BIO.

A análise da influência dos tipos de anotação é importante principalmente devido ao tempo elevo gasto na anotação de entidades, sobretudo quando utilizado a notação BIO, além da necessidade de um maior conhecimento entre as relações dos termos antecessores e sucessores da entidade anotada.

Para a realização e esclarecimentos sobre os experimentos esse trabalho levantamos as seguintes hipóteses:

- Hipótese 1: Utilizando os corpora de baseline e C1, verificar a inferência da metodologia proposta nesse trabalho em comparação com a sua não utilização e analisar sua influência nos resultados baseado nas métricas geradas para o processo de reconhecimento de entidades do grupo logradouro.

- Hipótese 2: Utilizando o corpus de baseline, verificar a inferência do uso do padrão de anotação de entidades BIO em comparação com o padrão de anotação IO, na influência dos resultados baseado nas métricas geradas para o processo de reconhecimento de entidades do grupo logradouro.

A seguir são apresentados os experimentos que permeiam as hipóteses 1 e 2, apresentando os resultados baseados nas métricas e os gráficos necessários para realizar comparações visualmente.

6.2.1 Experimento 1 – Influência do Sistema de Apoio

Conforme apresentado na primeira hipótese, nesse experimento a categoria logradouro foi avaliada, utilizando os corpos baseline e C1 para determinar a inferência do uso da metodologia na tarefa de reconhecimento de entidades geográficas de baixa granularidade. A Tabela 25 apresenta os resultados médios gerados em cada corpus.

Tabela 25: Resultados do Experimento 1

Notação	Categoria	Corpus	Precisão	Abrangência	Medida-F	
					Média	σ
IO	Logradouro	Baseline	0,6453	0,4630	0,5227	0,0698
IO	Logradouro	C1	0,7519	0,6954	0,7215	0,0371

A Figura 38 apresenta um comparativo entre os itens do experimento.

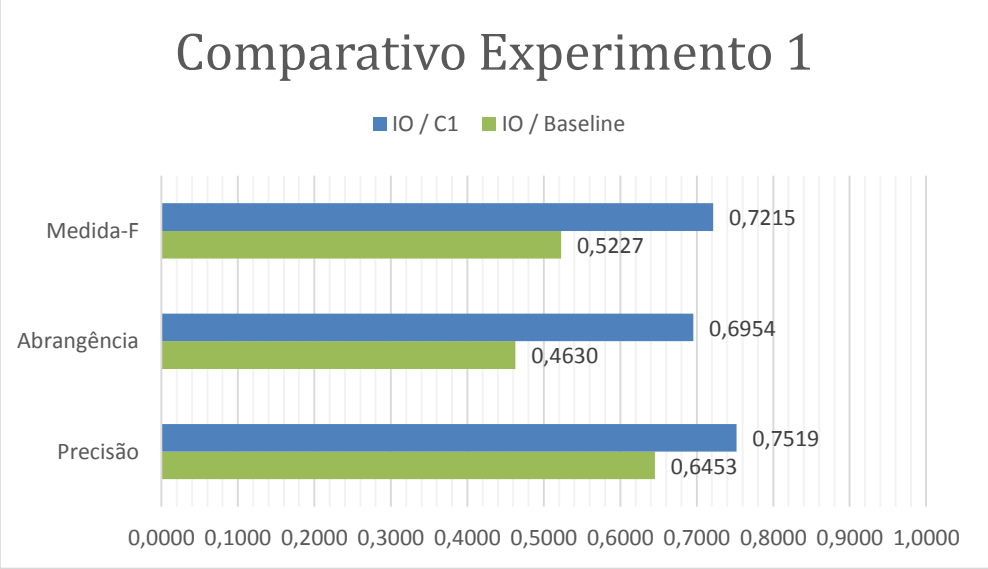

Figura 38: Comparativo Experimento 1

6.2.2 Experimento 2 - Notação de Entidades Padrão x BIO

Conforme apresentado na segunda hipótese, nesse experimento comparamos o grupo de entidade logradouro no corpus baseline utilizando as notações IO e BIO. A Tabela 26 apresenta os resultados gerados.

Tabela 26: Resultados do Experimento 2

Notação	Categoria	Corpus	Precisão	Abrangência	Medida-F	
					Média	σ
IO	Logradouro	Baseline	0,6217	0,4424	0,4973	0,0703
BIO	Logradouro	Baseline	0,6501	0,4350	0,5069	0,0637

A Figura 39 apresenta um comparativo entre os itens do experimento.

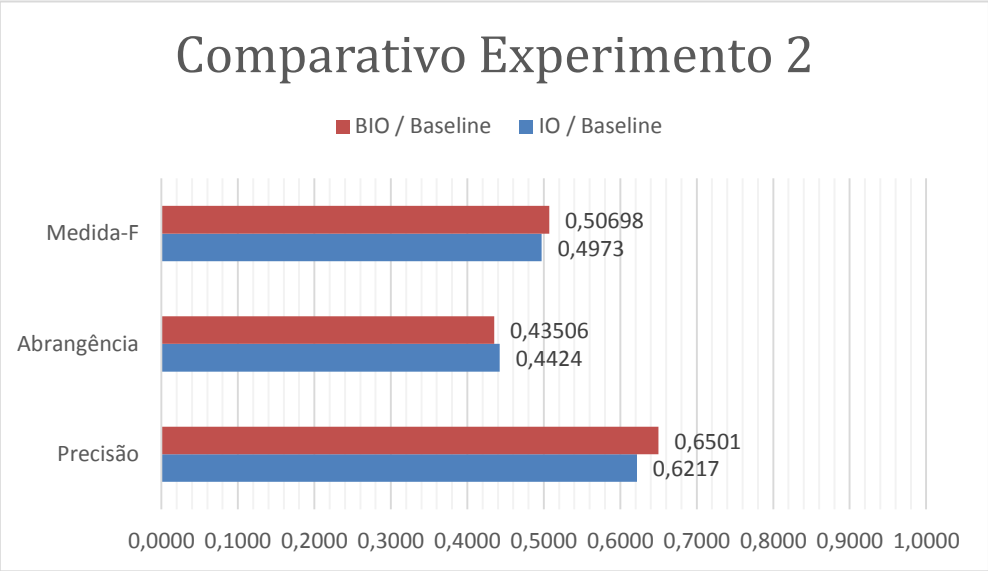

Figura 39: Comparativo Experimento 2

6.3 Segunda Execução dos Experimentos: Corpus (C1)

Após a realização da primeira rodada de experimentos, estabelecemos a relação da utilização do corpus baseline com o padrão de anotação IO comparado ao corpus de C1 com o padrão de anotação IO.

Semelhante ao segundo experimento realizado na seção 6.2, para esse experimento estamos interessados na análise de influência dos diferentes tipos de anotação de entidade, contudo, utilizando o corpus C1. Para a realização dos testes o corpus C1 foi subdividido em baseline com padrão IO e baseline com para BIO.

Para a realização e esclarecimentos sobre os experimentos esse trabalho levantamos a seguinte hipótese:

Hipótese 3: Utilizando o corpus C1, verificar a inferência do uso do padrão de anotação de entidades BIO em comparação com o padrão de anotação IO, na influência dos resultados baseado nas métricas geradas para o processo de reconhecimento de entidades do grupo logradouro.

6.3.1 Experimento 3 - Notação de Entidades com as Notações IO e BIO

Conforme apresentado na terceira hipótese, nesse experimento comparamos os grupos de entidades de localidade e logradouro utilizando a notação IO e BIO. A Tabela 27 apresenta os resultados gerados para cada tipo de notação/categoria.

Tabela 27: Resultados do Experimento 3 com Anotações IO e BIO

Notação	Categoria	Corpus	Precisão	Abrangência	Medida-F	
					Média	σ
IO	Logradouro	C1	0,7480	0,6918	0,7179	0,0364
BIO	Logradouro	C1	0,7519	0,6954	0,7215	0,0371

A Figura 40 apresenta um comparativo entre os itens do experimento:

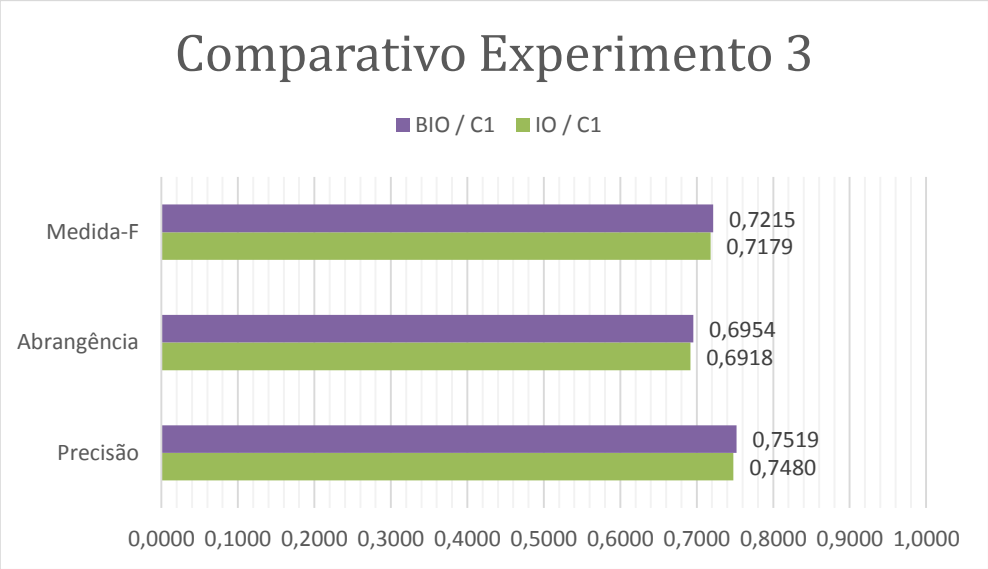

Figura 40: Comparativo Experimento 3

6.4 Discussão dos Resultados

A partir da realização dos experimentos e da análise dos resultados é possível identificar a viabilidade e os desafios relacionados à identificação de endereços geográficos de baixa granularidade, com o objetivo de apoiar a criação de dicionários geográficos e as tarefas ligadas as atividades de reambulação.

Levando em consideração o **Experimento 1**, realizado entre os corpus Baseline e C1 com anotação padrão IO, pode-se verificar a partir dos resultados médios da métrica Medida-F, os seguintes resultados: Comparando os resultados utilizando o sistema de pré-processamento proposto neste trabalho (corpus C1) com os resultados do corpus baseline temos que com a utilização da estrutura proposta, possibilitaram a identificação de endereços geográficos de baixa localidade com maior percentual de

precisão e abrangência, possibilitando ganhos de 16% e 50% respectivamente, possibilitando um incremento na medida-f de 38,03%.

Com relação ao **Experimento 2**, realizado utilizando o corpus baseline e as variações de notação propostas no trabalho, IO e BIO, pode-se analisar a influência da notação para a identificação de topônimos geográficos. Neste experimento houve um aumento de 4,73% e 1,65% respectivamente nas medidas de precisão e abrangência, possibilitando um incremento na medida-f de 1,93%.

No **Experimento 3**, utilizando o corpus C1 e as variações de notação propostas no trabalho, IO e BIO, pode-se analisar a influência da notação para a identificação de topônimos geográficos no sistema de pré-processamento apresentado nesse trabalho. Neste experimento houve um aumento de 0,52% e 0,52% respectivamente nas medidas de precisão e abrangência, possibilitando um incremento na medida-f de 0,50%.

Analisando os resultados do **Experimento 1**, um achado deste estudo é a melhoria significativa em abrangência, recuperação e medida-f, quando comparados ao corpus baseline. Os resultados ilustram o potencial e utilidade das etapas propostas nesse trabalho no auxilio as tarefas de identificação de elementos geográficos de baixa granularidade, mais especificamente em logradouros, em notícias textuais extraídas na Internet.

Analisando os resultados do **Experimento 2**, um achado deste estudo é a pequena diferença de desempenho relativa aos padrões de anotações de entidades, IO e BIO. Os resultados não obtiveram diferença significativa, utilizando o corpus baseline que justificam a utilização do padrão BIO, pois esse possui maior complexidade, tempo de resposta e maior dificuldade em anotar manualmente as entidades.

Analisando os resultados do **Experimento 3**, que comparou os dois padrões de diferentes de anotações de entidades, IO e BIO, utilizando o corpus C1, obtendo resultados ainda mais próximos, quando comparados ao **Experimento 2**. Uma hipótese para esta menor diferença acontecer devido ao janelamento das notícias, fazendo com que a máquina de aprendizado deixe de estabelecer relações entre os termos anteriores e posteriores ao elemento.

A Figura 41 apresenta um gráfico que sintetiza os resultados obtidos nesses experimentos, em relação a medida-F em comparação do corpus C1 e o baseline.

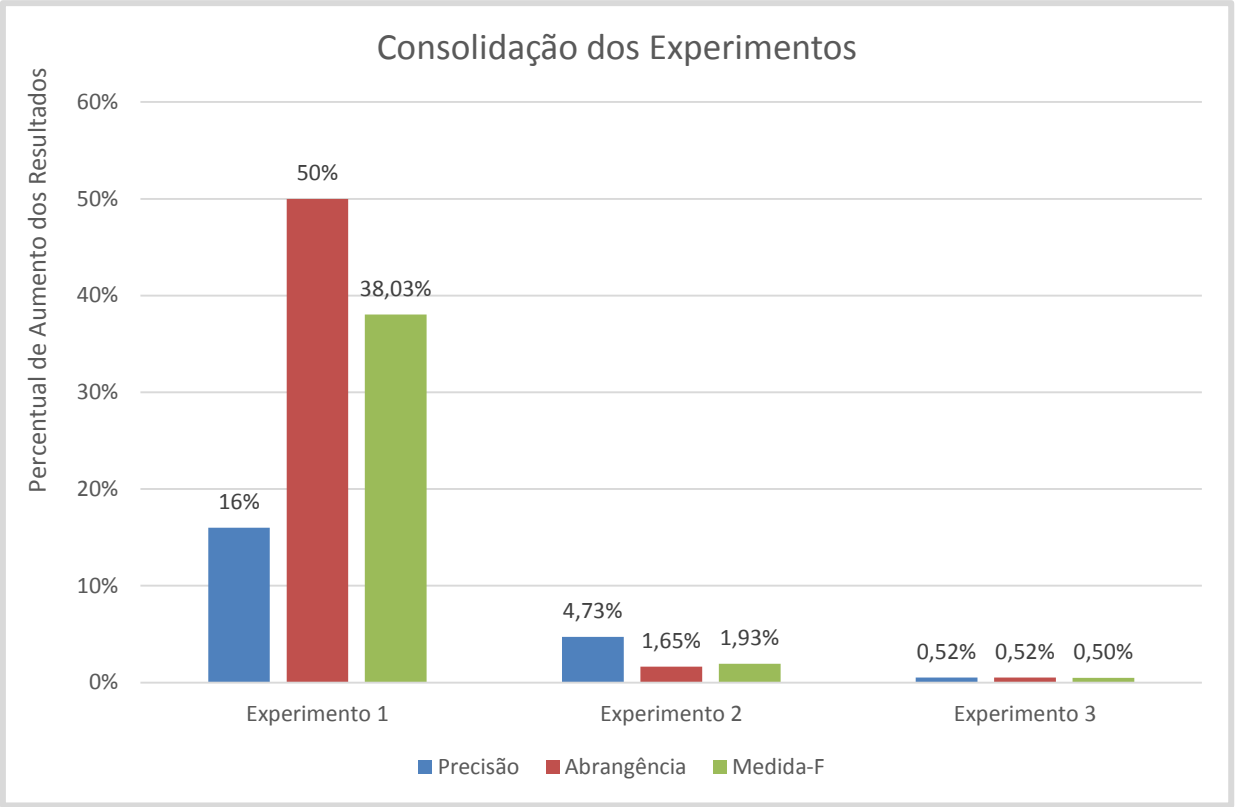

Figura 41: Percentual de Ganho Consolidado nos Experimentos

Conforme apresentado na Figura 41, os resultados comparativos das experiências fornecem evidências que ao utilizar os processos de pré-processamento para delimitar o escopo do conteúdo das notícias aplicando aos processos descritos nesse trabalho há uma redução na complexidade das características intrínsecas ao problema, facilitando o processo de aprendizado supervisionado de máquina, possibilitando maior refinamento na informação gerida pela máquina, conforme apresentado nos resultados obtidos.

Os resultados comparativos do trabalho fornecem também evidencias que ao utilizar as diferentes anotações de entidades, IO e BIO, essas não inferiram em alterações significativas no processo de reconhecimento de entidades nomeadas para logradouro.

7 Conclusão

Esse trabalho demonstrou a importância de identificar elementos geográficos mais específicos tendo como alvo as notícias do Brasil. Aproveitando a ocorrência de elementos em notícias que indiquem a localidade específica e a necessidade de refinar o conteúdo para obter melhores resultados na tarefa de reconhecimento de entidades especifica para logradouro.

A partir dos experimentos realizados alegamos que a identificação de elementos geográficos utilizando as etapas propostas nesse trabalho engendrou melhor resultados e maior poder de generalidade, visto que foram criadas regras que abrangem uma quantidade de padrões consideráveis de endereços, permitindo que os métodos possam ser aplicados em outros contextos, experimentos e cenários.

Os experimentos realizados possibilitam o georreferenciamento das notícias, com alto grau de confiabilidade, pois diferente de outros trabalhos relacionados apresentados, os resultados do processo geográfico serão mais precisos em relação ao posicionamento no globo terrestre em detrimento do fator de granularidade abordado nesse trabalho.

Os resultados fornecem também evidencias que ao utilizar as diferentes anotações de entidades, IO e BIO, essas não inferiram em alterações significativas no processo de reconhecimento de entidades nomeadas para logradouro, permitindo que para essa tarefa a outras pesquisas podem reduzir o esforço na escolha do padrão de anotação de entidades.

Conforme apresentado nessa pesquisa, a abordagem proposta pode ser utilizada para apoiar a criação e atualização automática de dicionários geográficos mais específicos, podendo abranger o país como um todo, possibilitando manter informações geográficas especificas mais detalhadas, com esforço reduzido.

O maior desafio em relação à identificação de elementos geográficos de baixa granularidade - do ponto vista do REM – é obter resultados consistentes e expressivos, ou seja, garantir que as entidades corretas sejam identificadas de maneira que permita o georreferenciamento. A utilidade da abordagem proposta neste trabalho para identificação de Indicadores de logradouro em notícias pôde então ser constatada, podendo ser cada vez mais aperfeiçoada tendo como base as análises e experimentos desenvolvidos e os trabalhos futuros sugeridos.

7.1 Trabalhos Futuros

Com relação à abordagem proposta, sustentamos que ela cumpriu o objetivo proposto de identificação de elementos geográficos de baixa granularidade em notícias em português do Brasil, através da formação e utilização de regras gramaticais em língua portuguesa. Os resultados apresentados ilustraram melhorias nas métricas propostas que permitem obter maior grau de identificação e classificação na tarefa especifica de identificação de endereços geográficos de baixa granularidade. Além de demonstrar os resultados descritos com duas diferentes anotações, possibilitando o estudo das inferências das anotações em outros grupos de entidades.

Para a continuidade desse trabalho, existem oportunidades de atualização e criação de ferramentas, métodos e processos que podem ser desenvolvidos. A seguir estão listadas algumas propostas de trabalho futuro:

- A criação de um sistema Web, que permita maior integração e facilidade no acesso às informações do GEONEWSBR, além de prover o acesso através de um WebService para outras aplicações utilizarem o sistema para identificar logradouros conforme a necessidade de cada sistema.
- O desenvolvimento de uma plataforma colaborativa que permita aos voluntários contribuir identificando notícias e adicionando novos itens na base de dados geográficos, através do envio de bases de dados particulares ou conteúdo de textos extraídos da Internet. Além de permitir que os voluntários avaliem as informações contidas nos dicionários de dados geográficos.
- A criação de uma estrutura de aprendizado de máquina que permita identificar automaticamente os demais grupos de elementos necessários para a criação de um dicionário geográfico.
- Criar uma estrutura multidimensional nos dados que permita a tomada de decisões de forma analítica e possibilite a integração com outros sistemas de *Business Intelligence* compartilhando indicadores, tabelas de dimensões e de fatos.

Referências Bibliográficas

Alexandria Digital Research Library. Disponível em: <http://alexandria.ucsb.edu/>. Acesso em: 20 ago. 2015.

AMARAL, D. O. F. DO; VIEIRA, R. NERP-CRF: uma ferramenta para o reconhecimento de entidades nomeadas por meio de Conditional Random Fields. **Linguamática**, v. 6, n. 1, p. 41–49, 2014.

BALLATORE, A.; WILSON, D. C.; BERTOLOTTO, M. A Survey of Volunteered Open Geo-Knowledge Bases in the Semantic Web. In: PASI, G.; BORDOGNA, G.; JAIN, L. C. (Eds.). . **Quality Issues in the Management of Web Information**. Intelligent Systems Reference Library. [s.l.] Springer Berlin Heidelberg, 2013. p. 93–120.

BATISTA, D. S. et al. **Geographic signatures for semantic retrieval**. Proceedings of the 6th Workshop on Geographic Information Retrieval. **Anais**...ACM, 2010Disponível em: <http://dl.acm.org/citation.cfm?id=1722104>. Acesso em: 7 dez. 2015

BIRD, S. **NLTK: The Natural Language Toolkit**. Proceedings of the COLING/ACL on Interactive Presentation Sessions. **Anais**...: COLING-ACL '06.Stroudsburg, PA, USA: Association for Computational Linguistics, 2006Disponível em: <http://dx.doi.org/10.3115/1225403.1225421>. Acesso em: 12 jun. 2016

BRILL, E. **A Simple Rule-based Part of Speech Tagger**. Proceedings of the Workshop on Speech and Natural Language. **Anais**...: HLT '91.Stroudsburg, PA, USA: Association for Computational Linguistics, 1992Disponível em: <http://dx.doi.org/10.3115/1075527.1075553>. Acesso em: 16 jun. 2016

BRISABOA, N. R. et al. Exploiting geographic references of documents in a geographical information retrieval system using an ontology-based index. **GeoInformatica**, v. 14, n. 3, p. 307–331, 30 jan. 2010.

BROWNLEE, J. **A Tour of Machine Learning Algorithms**. Disponível em: <http://machinelearningmastery.com/a-tour-of-machine-learning-algorithms/>. Acesso em: 11 nov. 2015.

CARVALHO, G. A.; LEITE, D. V. B. Geoprocessamento na gestão urbana municipal– a experiência dos municípios mineiros Sabará e Nova Lima. **SIMPÓSIO BRASILEIRO DE SENSORIAMENTO REMOTO**, v. 14, p. 3643–3650, 2009.

CHEUNG, D. W. et al. **A fast distributed algorithm for mining association rules**. , Fourth International Conference on Parallel and Distributed Information Systems, 1996. **Anais**... In: , FOURTH INTERNATIONAL CONFERENCE ON PARALLEL AND DISTRIBUTED INFORMATION SYSTEMS, 1996. Dezembro 1996

CIRAVEGNA, F. et al. **Integrating Information to Bootstrap Information Extraction from Web Sites**. In: IJCAI'03 Workshop on Intelligent Information Integration. **Anais**...2003

CLG - UNIVERSIDADE DE LISBOA. **Wordnet.PT**. Disponível em: <http://www.clul.ul.pt/clg/wordnetpt/index.html>. Acesso em: 16 set. 2015.

CLOUGH, P. **Extracting Metadata for Spatially-aware Information Retrieval on the Internet**. Proceedings of the 2005 Workshop on Geographic Information Retrieval. **Anais**...: GIR '05.New York, NY, USA: ACM, 2005Disponível em: <http://doi.acm.org/10.1145/1096985.1096992>. Acesso em: 31 ago. 2015

COWIE, J.; LEHNERT, W. Information Extraction. **Commun. ACM**, v. 39, n. 1, p. 80–91, jan. 1996.

DA SILVA, L. H.; DE MEDEIROS CASELI, H. Reconhecimento de entidades nomeadas em textos em português do Brasil no domınio do e-commerce. 2015.

DENSHAM, I.; REID, J. **A Geo-coding Service Encompassing a Geo-parsing Tool and Integrated Digital Gazetteer Service**. Proceedings of the HLT-NAACL 2003 Workshop on Analysis of Geographic References - Volume 1. **Anais**...: HLT-NAACL-GEOREF '03.Stroudsburg, PA, USA: Association for Computational Linguistics, 2003Disponível em: <http://dx.doi.org/10.3115/1119394.1119406>. Acesso em: 26 ago. 2015

DODDINGTON, G. R. et al. **The Automatic Content Extraction (ACE) Program-Tasks, Data, and Evaluation.** LREC. **Anais**...2004Disponível em: <https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/lrec2004-ace-program.pdf>. Acesso em: 2 dez. 2014

EDDY, S. R. Profile hidden Markov models. **Bioinformatics**, v. 14, n. 9, p. 755–763, 1998.

ELLOUMI, S. et al. General learning approach for event extraction: Case of management change event. **Journal of Information Science**, v. 39, n. 2, p. 211–224, 1 abr. 2013.

ETZIONI, O. et al. Unsupervised named-entity extraction from the Web: An experimental study. **Artificial Intelligence**, v. 165, n. 1, p. 91–134, jun. 2005.

FINKEL, J. R.; GRENAGER, T.; MANNING, C. **Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling**. Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. **Anais**...: ACL '05.Stroudsburg, PA, USA: Association for Computational Linguistics, 2005Disponível em: <http://dx.doi.org/10.3115/1219840.1219885>. Acesso em: 26 nov. 2015

FRIEDRICH, H. L. **Newton da Costa e o problema da indução**. Trabalho de Conclusão de Curso - Graduação - Bacharelado. Disponível em: <http://bdm.unb.br/handle/10483/10956>. Acesso em: 8 jun. 2016.

GELERNTER, J. et al. **Automatic Gazetteer Enrichment with User-geocoded Data**. Proceedings of the Second ACM SIGSPATIAL International Workshop on Crowdsourced and Volunteered Geographic Information. **Anais**...: GEOCROWD '13.New York, NY, USA: ACM, 2013Disponível em: <http://doi.acm.org/10.1145/2534732.2534736>. Acesso em: 30 set. 2015

GEONAMES. **Localidades e Municípios Brasileiros**. Disponível em: <https://geonames.wordpress.com/2006/12/17/localidades-e-municipios-brasileiros/>. Acesso em: 26 ago. 2015.

GEONAMES. **The GeoNames geographical database**. Disponível em: <http://www.geonames.org/>. Acesso em: 20 ago. 2015.

GIUNCHIGLIA, F. et al. GeoWordNet: A Resource for Geo-spatial Applications. In: AROYO, L. et al. (Eds.). . **The Semantic Web: Research and Applications**. Lecture Notes in Computer Science. [s.l.] Springer Berlin Heidelberg, 2010. p. 121–136.

GOLDBERG, D. W.; WILSON, J. P.; KNOBLOCK, C. A. From text to geographic coordinates: the current state of geocoding. **URISA journal**, v. 19, n. 1, p. 33–46, 2007.

GOLDBERG, D. W.; WILSON, J. P.; KNOBLOCK, C. A. Extracting Geographic Features from the Internet to Automatically Build Detailed Regional Gazetteers. **Int. J. Geogr. Inf. Sci.**, v. 23, n. 1, p. 93–128, jan. 2009.

GÓMEZ, C. G. et al. Automatic Extraction of Geographic Locations on Articles of Digital Newspapers. In: RODRÍGUEZ, J. M. C. et al. (Eds.). . **Trends in Practical Applications of Agents and Multiagent Systems**. Advances in Intelligent and Soft Computing. [s.l.] Springer Berlin Heidelberg, 2012. p. 101–108.

GOODCHILD, M. F. Geographic Information System. In: LIU, L.; ÖZSU, M. T. (Eds.). . **Encyclopedia of Database Systems**. [s.l.] Springer US, 2009. p. 1231–1236.

GOOGLE. **Google Places API**. Disponível em: <https://developers.google.com/places/?hl=pt-br>. Acesso em: 11 jan. 2016.

GOUVÊA, C. et al. **Discovering Location Indicators of Toponyms from News to Improve Gazetteer-Based Geo-Referencing**. In Simpósio Brasileiro de Geoinformática-GEOINFO. **Anais**...2008

GRIGIO, A. M. Aplicação de sensoriamento remoto e sistema de informação geográfica na determinação da vulnerabilidade natural e ambiental do Município de Guamaré (RN): simulação de risco às atividades da indústria petrolífera. 2003.

GRISHMAN, R.; SUNDHEIM, B. **Message Understanding Conference-6: A Brief History.** COLING. **Anais**...1996Disponível em: <http://www.alta.asn.au/events/altss_w2003_proc/altss/courses/molla/C96-1079.pdf>. Acesso em: 11 dez. 2014

GROVER, C. et al. Use of the Edinburgh geoparser for georeferencing digitized historical collections. **Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences**, v. 368, n. 1925, p. 3875–3889, 28 ago. 2010.

GULL, S. F.; DANIELL, G. J. The Maximum Entropy Method1. 1984.

HARTIGAN, J. A.; WONG, M. A. Algorithm AS 136: A K-Means Clustering Algorithm. **Journal of the Royal Statistical Society. Series C (Applied Statistics)**, v. 28, n. 1, p. 100–108, 1979.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. Unsupervised Learning. In: **The Elements of Statistical Learning**. Springer Series in Statistics. [s.l.] Springer New York, 2009. p. 485–585.

HILL, L. L. Core Elements of Digital Gazetteers: Placenames, Categories, and Footprints. In: BORBINHA, J.; BAKER, T. (Eds.). . **Research and Advanced Technology for Digital Libraries**. Lecture Notes in Computer Science. [s.l.] Springer Berlin Heidelberg, 2000. p. 280–290.

IBGE, I. B. DE G. E E. **Noções Básicas de Cartografia**. Disponível em: <http://www.ibge.gov.br/home/geociencias/cartografia/manual_nocoes/processo_cartografico.html>. Acesso em: 27 maio. 2016.

KONKOL, M. Named Entity Recognition. 2012.

LAFFERTY, J.; MCCALLUM, A.; PEREIRA, F. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. **Departmental Papers (CIS)**, 28 jun. 2001.

LEIDNER, J. L.; LIEBERMAN, M. D. Detecting Geographical References in the Form of Place Names and Associated Spatial Natural Language. **SIGSPATIAL Special**, v. 3, n. 2, p. 5–11, jul. 2011.

LEWIS, D. D. Naive (Bayes) at forty: The independence assumption in information retrieval. In: NÉDELLEC, C.; ROUVEIROL, C. (Eds.). . **Machine Learning: ECML-98**. Lecture Notes in Computer Science. [s.l.] Springer Berlin Heidelberg, 1998. p. 4–15.

LUCHIARI, A. Identificação da cobertura vegetal em áreas urbanas por meio de produtos de sensoriamento remoto e de um Sistema de Informação Geográfica. **Revista do Departamento de Geografia**, v. 14, n. 0, p. 47–58, 5 maio 2011.

LUO, J. et al. Geotagging in multimedia and computer vision—a survey. **Multimedia Tools and Applications**, v. 51, n. 1, p. 187–211, 19 out. 2010.

LUO, J. et al. Geotagging in Multimedia and Computer Vision–a Survey. **Multimedia Tools Appl.**, v. 51, n. 1, p. 187–211, jan. 2011.

MACHADO, I. M. R. et al. An ontological gazetteer and its application for place name disambiguation in text. **Journal of the Brazilian Computer Society**, v. 17, n. 4, p. 267–279, 14 out. 2011.

MARQUES, N. C.; LOPES, G. P. Tagging with Small Training Corpora. In: HOFFMANN, F. et al. (Eds.). . **Advances in Intelligent Data Analysis**. Lecture Notes in Computer Science. [s.l.] Springer Berlin Heidelberg, 2001. p. 63–72.

MICHALSKI, R. S.; CARBONELL, J. G.; MITCHELL, T. M. **Machine Learning: An Artificial Intelligence Approach**. [s.l.] Springer Science & Business Media, 2013.

MIKHEEV, A.; GROVER, C.; MOENS, M. **Description of the LTG system used for MUC-7**. Proceedings of 7th Message Understanding Conference (MUC-7). **Anais**...Fairfax, VA, 1998Disponível em:

<http://staff.um.edu.mt/mros1/csa4050/ie/pdf/mikheev98description.pdf>. Acesso em: 25 nov. 2015

MIKHEEV, A.; MOENS, M.; GROVER, C. **Named Entity Recognition Without Gazetteers**. Proceedings of the Ninth Conference on European Chapter of the Association for Computational Linguistics. **Anais**...: EACL '99.Stroudsburg, PA, USA: Association for Computational Linguistics, 1999Disponível em: <http://dx.doi.org/10.3115/977035.977037>. Acesso em: 25 ago. 2015

MØLLER, M. F. A scaled conjugate gradient algorithm for fast supervised learning. **Neural Networks**, v. 6, n. 4, p. 525–533, 1993.

MONARD, M. C.; BARANAUSKAS, J. A. Conceitos sobre aprendizado de máquina. **Sistemas Inteligentes-Fundamentos e Aplicações**, v. 1, p. 1, 2003.

MONCLA, L. et al. **Geocoding for Texts with Fine-grain Toponyms: An Experiment on a Geoparsed Hiking Descriptions Corpus**. Proceedings of the 22Nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. **Anais**...: SIGSPATIAL '14.New York, NY, USA: ACM, 2014Disponível em: <http://doi.acm.org/10.1145/2666310.2666386>. Acesso em: 15 out. 2015

MOTA, C.; SANTOS, D. **Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM**. [s.l.] Linguateca, 2008a.

MOTA, C.; SANTOS, D. **Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM - Capitulo 3 - R3M**. [s.l.] Linguateca, 2008b.

NADEAU, D.; TURNEY, P. D.; MATWIN, S. Unsupervised Named-Entity Recognition: Generating Gazetteers and Resolving Ambiguity. In: LAMONTAGNE, L.; MARCHAND, M. (Eds.). . **Advances in Artificial Intelligence**. Lecture Notes in Computer Science. [s.l.] Springer Berlin Heidelberg, 2006. p. 266–277.

ODON DE ALENCAR, R.; DAVIS, C. A., Jr.; GONÇALVES, M. A. **Geographical Classification of Documents Using Evidence from Wikipedia**. Proceedings of the 6th Workshop on Geographic Information Retrieval. **Anais**...: GIR '10.New York, NY, USA: ACM, 2010Disponível em: <http://doi.acm.org/10.1145/1722080.1722096>. Acesso em: 9 jun. 2016

OLIVEIRA, H. G. et al. Avaliação à medida no Segundo HAREM. **Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM. Linguateca**, p. 97–129, 2008.

OpenStreetMap Brasil. Disponível em: <http://www.openstreetmap.com.br/>. Acesso em: 11 jan. 2016.

OVERELL, S. E.; RÜGER, S. **Geographic Co-occurrence As a Tool for Gir.** Proceedings of the 4th ACM Workshop on Geographical Information Retrieval. **Anais**...: GIR '07.New York, NY, USA: ACM, 2007Disponível em: <http://doi.acm.org/10.1145/1316948.1316968>. Acesso em: 21 ago. 2015

PAIVA, V. DE; RADEMAKER, A.; MELO, G. DE. OpenWordNet-PT: an Open Brazilian Wordnet for reasoning. 1 dez. 2012.

POPESCU, A.; GREFENSTETTE, G.; MOËLLIC, P. A. **Gazetiki: Automatic Creation of a Geographical Gazetteer**. Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries. **Anais**...: JCDL '08.New York, NY, USA: ACM, 2008Disponível em: <http://doi.acm.org/10.1145/1378889.1378906>. Acesso em: 19 ago. 2015

PRINCETON UNIVERSITY. **What is WordNet?** Disponível em: <https://wordnet.princeton.edu/>. Acesso em: 16 set. 2015.

RATINOV, L.; ROTH, D. **Design challenges and misconceptions in named entity recognition**. Proceedings of the Thirteenth Conference on Computational Natural Language Learning. **Anais**...Association for Computational Linguistics, 2009Disponível em: <http://dl.acm.org/citation.cfm?id=1596399>. Acesso em: 2 dez. 2014

RAUCH, E.; BUKATIN, M.; BAKER, K. **A Confidence-based Framework for Disambiguating Geographic Terms**. Proceedings of the HLT-NAACL 2003 Workshop on Analysis of Geographic References - Volume 1. **Anais**...: HLT-NAACL-GEOREF '03.Stroudsburg, PA, USA: Association for Computational Linguistics, 2003Disponível em: <http://dx.doi.org/10.3115/1119394.1119402>. Acesso em: 9 jan. 2015

RITTER, A. et al. **Named Entity Recognition in Tweets: An Experimental Study**. Proceedings of the Conference on Empirical Methods in Natural Language Processing. **Anais**...: EMNLP '11.Stroudsburg, PA, USA: Association for Computational Linguistics, 2011Disponível em: <http://dl.acm.org/citation.cfm?id=2145432.2145595>. Acesso em: 10 dez. 2014

RIZZO, G.; TRONCY, R. **NERD: Evaluating Named Entity Recognition Tools in the Web of Data**. Titolo volume non avvalorato. **Anais**... In: (ISWC'11) WORKSHOP ON WEB SCALE KNOWLEDGE EXTRACTION (WEKEX'11). Bonn, Germany: 2011Disponível em: <http://porto.polito.it/2440793/>. Acesso em: 2 dez. 2014

RUPP, C. J. et al. **Customising geoparsing and georeferencing for historical texts**. 2013 IEEE International Conference on Big Data. **Anais**... In: 2013 IEEE INTERNATIONAL CONFERENCE ON BIG DATA. Outubro 2013

SALAMUNI, E.; STELLFELD, M. C. Banco de dados geológicos geo-referenciados da Bacia Sedimentar de Curitiba (PR) como base de sistema de informação geográfica (SIG). **B. Paranaense Geoci**, v. 49, p. 21–32, 2001.

SANTOS, C. J. B. D. **A Padronização dos Nomes geográficos num estudo de caso dos municípios fluminenses**. [s.l.] Tese de Doutorado–Universidade Federal do Rio de Janeiro, IGEO. Rio de Janeiro, 2008.

SANTOS, D. O modelo semântico usado no Primeiro HAREM. **Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área**, p. 43–57, 2007.

SANTOS, D. et al. Breve introduçao ao HAREM. **HAREM, a primeira avaliaçao conjunta de sistemas de reconhecimento de entidades mencionadas para português: documentaçao e actas do encontro, Linguateca**, 2007.

SARAWAGI, S. Information Extraction. **Foundations and Trends in Databases**, v. 1, n. 3, p. 261–377, 2007.

SCHARL, P. A. Towards the Geospatial Web: Media Platforms for Managing Geotagged Knowledge Repositories. In: SCHARL, P. A.; TOCHTERMANN, P. K. (Eds.). . **The Geospatial Web**. Advanced Information and Knowledge Processing. [s.l.] Springer London, 2007. p. 3–14.

SCIKIT-LEARN DEVELOPERS. **Machine Learning 101: General Concepts**. Disponível em: <http://www.astroml.org/sklearn_tutorial/general_concepts.html>. Acesso em: 8 jun. 2016.

SILVA, E. F.; BARROS, F. A.; PRUDÊNCIO, R. B. **Uma abordagem de aprendizagem híbrida para extração de informação em textos semi-estruturados**. Anais do XXV Congresso da Sociedade Brasileira de Computação. **Anais**...2005Disponível em: <https://www.cin.ufpe.br/~rbcp/papers/ENIA05.pdf>. Acesso em: 6 jul. 2016

SILVA, J. X. DA. Geomorfologia, Análise ambiental e Geoprocessamento. **Revista Brasileira de Geomorfologia**, v. 1, n. 1, 2000.

SOUZA, L. A. et al. **The Role of Gazetteers in Geographic Knowledge Discovery on the Web**. Proceedings of the Third Latin American Web Congress. **Anais**...: LA-WEB '05.Washington, DC, USA: IEEE Computer Society, 2005Disponível em: <http://dx.doi.org/10.1109/LAWEB.2005.38>. Acesso em: 1 out. 2015

STANFORD NLP GROUP. **Stanford Named Entity Recognizer (NER)**. Disponível em: <http://nlp.stanford.edu/software/CRF-NER.shtml>. Acesso em: 25 jun. 2016.

SUNDHEIM, B. M. **Overview of Results of the MUC-6 Evaluation**. Proceedings of a Workshop on Held at Vienna, Virginia: May 6-8, 1996. **Anais**...: TIPSTER '96.Stroudsburg, PA, USA: Association for Computational Linguistics, 1996Disponível em: <http://dx.doi.org/10.3115/1119018.1119073>. Acesso em: 15 dez. 2014

SUTTON, C.; MCCALLUM, A. An introduction to conditional random fields for relational learning. **Introduction to statistical relational learning**, p. 93–128, 2006.

TEITLER, B. E. et al. **NewsStand: A New View on News**. Proceedings of the 16th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. **Anais**...: GIS '08.New York, NY, USA: ACM, 2008Disponível em: <http://doi.acm.org/10.1145/1463434.1463458>. Acesso em: 8 jan. 2015

THE J. PAUL GETTY TRUST. **Getty Thesaurus of Geographic Names (Getty Research Institute)**. Disponível em: <http://www.getty.edu/research/tools/vocabularies/tgn/?find=&place=escola&nation=&prev_page=1&english=Y&popup=P>. Acesso em: 21 ago. 2015.

TOBIN, R. et al. **Evaluation of Georeferencing**. Proceedings of the 6th Workshop on Geographic Information Retrieval. **Anais**...: GIR '10.New York, NY, USA: ACM, 2010Disponível em: <http://doi.acm.org/10.1145/1722080.1722089>. Acesso em: 25 ago. 2015

TORAL, A.; MUNOZ, R. **A proposal to automatically build and maintain gazetteers for Named Entity Recognition by using Wikipedia**. 2006Disponível em: <http://www.researchgate.net/profile/Antonio_Toral/publication/251868774_A_proposal_to_automatically_build_and_maintain_gazetteers_for_Named_Entity_Recognition_by_using_Wikipedia/links/02e7e53b556dd199e6000000.pdf#page=64>. Acesso em: 15 out. 2015

TORISAWA, K. **Exploiting Wikipedia as External Knowledge for Named Entity Recognition**. [s.l: s.n.].

WEBER, E. et al. Qualidade de dados geoespaciais. **Universidade Federal do Rio Grande do Sul**, 1999.

WEIKUM, G. et al. Database and Information-retrieval Methods for Knowledge Discovery. **Commun. ACM**, v. 52, n. 4, p. 56–64, Abril 2009.

WIKIPEDIA. **Wikipedia**. Disponível em: <https://pt.wikipedia.org/>. Acesso em: 11 jan. 2016.

ZHANG, Y.; TSAI, F. S. **Combining Named Entities and Tags for Novel Sentence Detection**. Proceedings of the WSDM '09 Workshop on Exploiting Semantic Annotations in Information Retrieval. **Anais**...: ESAIR '09.New York, NY, USA: ACM, 2009Disponível em: <http://doi.acm.org/10.1145/1506250.1506256>. Acesso em: 7 jun. 2016

ZHAO, P.; FOERSTER, T.; YUE, P. The Geoprocessing Web. **Computers & Geosciences**, Towards a Geoprocessing Web. v. 47, p. 3–12, Outubro 2012.