Matheus Emerick de Magalhães

A utilização de MACHINE LEARNING na identificação de elementos textuais geográficos

Atena Editora
Ano 2022

Matheus Emerick de Magalhães

A utilização de MACHINE LEARNING na identificação de elementos textuais geográficos

Atena Editora
Ano 2022

CIÊNCIAS EXATAS E DA TERRA E ENGENHARIAS

Editora chefe
Profª Drª Antonella Carvalho de Oliveira
Editora executiva
Natalia Oliveira
Assistente editorial
Flávia Roberta Barão
Bibliotecária
Janaina Ramos
Projeto gráfico
Bruno Oliveira
Camila Alves de Cremo
Luiza Alves Batista
Natália Sandrini de Azevedo
Imagens da capa
iStock
Edição de arte
Luiza Alves Batista

CIÊNCIAS EXATAS E DA TERRA E ENGENHARIAS

A utilização de machine learning na identificação de elementos textuais geográficos

Diagramação: Natália Sandrini de Azevedo
Correção: Yaiddy Paola Martinez
Indexação: Amanda Kelly da Costa Veiga
Revisão: Daysianne Kessy Mendes Isidorio
Autor: Matheus Emerick de Magalhães

Dados Internacionais de Catalogação na Publicação (CIP)

M188 Magalhães, Matheus Emerick de
A utilização de machine learning na identificação de elementos textuais geográficos / Matheus Emerick de Magalhães. – Ponta Grossa - PR: Atena, 2022.

Formato: PDF
Requisitos de sistema: Adobe Acrobat Reader
Modo de acesso: World Wide Web
Inclui bibliografia
ISBN 978-65-258-0731-7
DOI: https://doi.org/10.22533/at.ed.317221011

1. Inteligência artificial - Aplicações educacionais. I. Magalhães, Matheus Emerick de. II. Título.
CDD 004.01

Elaborado por Bibliotecária Janaina Ramos – CRB-8/9166

Atena Editora
Ponta Grossa – Paraná – Brasil
Telefone: +55 (42) 3323-5493
www.atenaeditora.com.br
contato@atenaeditora.com.br

DECLARAÇÃO DO AUTOR

O autor desta obra: 1. Atesta não possuir qualquer interesse comercial que constitua um conflito de interesses em relação ao conteúdo publicado; 2. Declara que participou ativamente da construção dos respectivos manuscritos, preferencialmente na: a) Concepção do estudo, e/ou aquisição de dados, e/ou análise e interpretação de dados; b) Elaboração do artigo ou revisão com vistas a tornar o material intelectualmente relevante; c) Aprovação final do manuscrito para submissão.; 3. Certifica que o texto publicado está completamente isento de dados e/ou resultados fraudulentos; 4. Confirma a citação e a referência correta de todos os dados e de interpretações de dados de outras pesquisas; 5. Reconhece ter informado todas as fontes de financiamento recebidas para a consecução da pesquisa; 6. Autoriza a edição da obra, que incluem os registros de ficha catalográfica, ISBN, DOI e demais indexadores, projeto visual e criação de capa, diagramação de miolo, assim como lançamento e divulgação da mesma conforme critérios da Atena Editora.

DECLARAÇÃO DA EDITORA

A Atena Editora declara, para os devidos fins de direito, que: 1. A presente publicação constitui apenas transferência temporária dos direitos autorais, direito sobre a publicação, inclusive não constitui responsabilidade solidária na criação dos manuscritos publicados, nos termos previstos na Lei sobre direitos autorais (Lei 9610/98), no art. 184 do Código Penal e no art. 927 do Código Civil; 2. Autoriza e incentiva os autores a assinarem contratos com repositórios institucionais, com fins exclusivos de divulgação da obra, desde que com o devido reconhecimento de autoria e edição e sem qualquer finalidade comercial; 3. Todos os e-book são *open access, desta forma* não os comercializa em seu site, sites parceiros, plataformas de *e-commerce,* ou qualquer outro meio virtual ou físico, portanto, está isenta de repasses de direitos autorais aos autores; 4. Todos os membros do conselho editorial são doutores e vinculados a instituições de ensino superior públicas, conforme recomendação da CAPES para obtenção do Qualis livro; 5. Não cede, comercializa ou autoriza a utilização dos nomes e e-mails dos autores, bem como nenhum outro dado dos mesmos, para qualquer finalidade que não o escopo da divulgação desta obra.

SUMÁRIO

RESUMO

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.

PALAVRAS-CHAVES: Identificação de conteúdo geográfico, Extração de Informação e Reconhecimento de entidades mencionadas

ABSTRACT

The Brazil is a vast and dynamic country. Identifying the new elements inaugurated or updated is a task that involves great effort, political and informative. The need for fine-grained geographic recognition needs geographic services that created a demand for fine-grained recognition of a high degree of specificity. As the internet makes available and integrates several sources of information, mainly in news from the most diverse media, about elements inaugurated in our country, state, and street, it becomes necessary to recover the city and structure this information in order to be able to relate it to the context. and reality of local systems where they are sustainable through automated methods and systems. Public bodies also have a need to identify new geographic elements, however, so that useful information must have more accurate geographic elements, for activities such as renaming tasks. For this, a geographic need of the needs is necessary the georeferencing of news, that is, to identify the geographic entities with their spatial location. The present work is an approach to creating grammar rules that allows a fine-grained identification of geographic elements that supports the creation and updating of news-based geographic dictionaries. The results present a usefulness of the approach for the creation of a tool to support the creation of specialties that supports the gazetteers and related activities such as reambulation tasks.

KEYWORDS: Identification of geographic content, Extraction of Information and Recognition of mentioned entities

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.

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.

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.

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.

Embora os dicionários geográficos apresentem informações detalhadas de

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

diversos países, a quantidade e o nível de detalhes sobre informações locais são pobres ou insuficientes para a tomada de decisão ou para resolver alguns problemas pontuais (GELERNTER et al., 2013).

Assim, um dos objetivos dessa 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. Neste 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.

DELIMITAÇÃO E OBJETIVOS

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.

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.

Para a realização dos objetivos propostos, esse trabalho foi dividido em 3 etapas: 1) o pré-processamento de notícias, 2) reconhecimento de entidades mencionadas, e 3) formação e visualização do dicionário geográfico.

REFERENCIAL TEÓRICO

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).

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.

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.

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).

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).

Extração de informação geográfica em notícias

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 (OVERELL; RÜGER, 2007).

As diversas fontes de dados como notícias, sites, blogs, páginas da Wikipedia 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 baseadas: na localização geográfica do editor, em informações geográficas do conteúdo do texto, e na localização dos leitores.

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.

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 classificá-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, 2014).

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.

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 *Wikipedia* (WIKIPEDIA, 2015), *OpenStreetMap* (“OpenStreetMap Brasil”, 2016), *Google Places* (GOOGLE, 2016), que fornecem informações enciclopédicas sobre entidades semiestruturadas 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).

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.

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 1 para a frase de exemplo “João Rua Prefeito”:

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

Tabela 1: Exemplo de categorização ambígua de entidades.

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 2.

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

Tabela 2: Notícia anotada com a notação IO.

Conforme apresentado na Tabela 3, 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 3 apresenta uma notícia anotada utilizando a notação BIO.

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

Tabela 3: Notícia anotada com a notação BIO.

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 1 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.

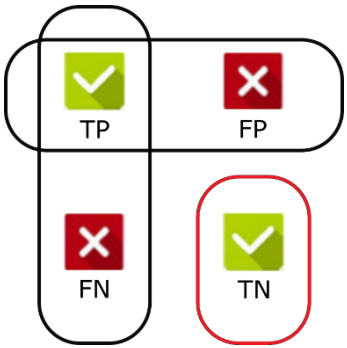

Figura 1: Relevância dos termos.

A Figura 2 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).

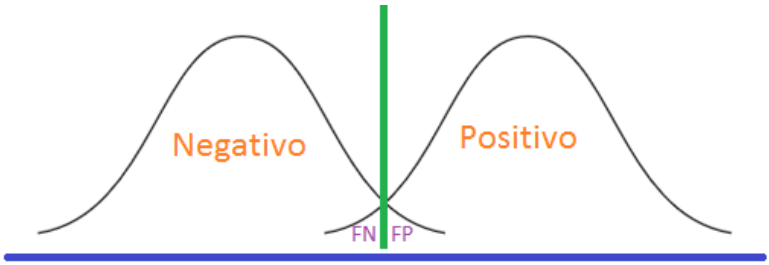

Figura 2: Fronteiras de classificação dos termos - baseado de precision e recall (KONKOL, 2012).

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 3, os valores dos que atendem a medida de abrangência, são expressos por:

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

Precisão

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

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

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 3 representa essa maneira:

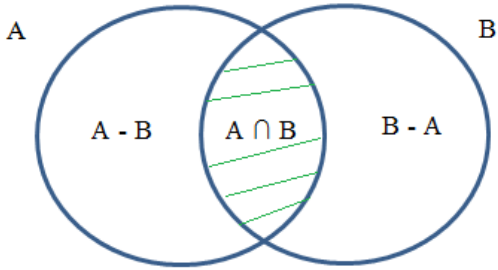

Figura 3: Conjuntos A e B e suas intercessões.

Observe que de acordo com a Figura 3, os valores dos conjuntos que atendem a medida de precisão, são expressos por:

$$Precisão = \frac{A \cap B}{B}$$

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).

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

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[1], 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 4 apresenta um prisma da diferença do método dedutivo para o indutivo.

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

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.

Tabela 4: Método dedutivo x indutivo.

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 este trabalho. 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 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.

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 4 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 4: Estrutura simplificada da técnica de aprendizado supervisionado.

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).

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 5 apresenta as etapas de um processo de aprendizado não supervisionado:

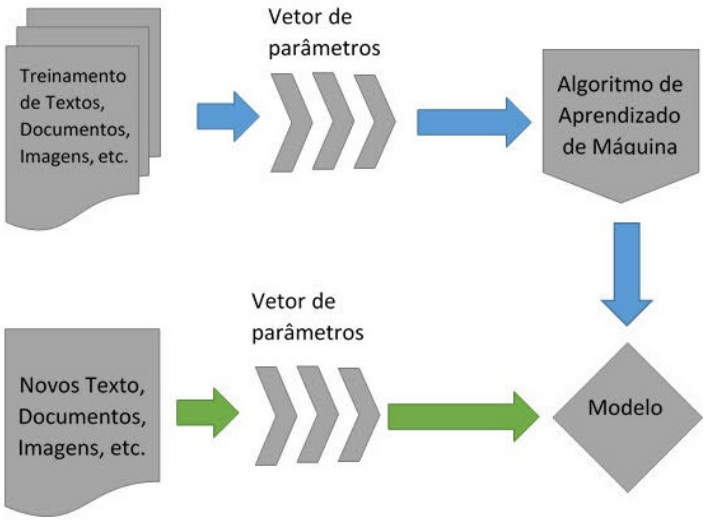

Figura 5: Estrutura de aprendizado não supervisionado, adaptado de (SCIKIT-LEARN DEVELOPERS, 2016).

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).

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

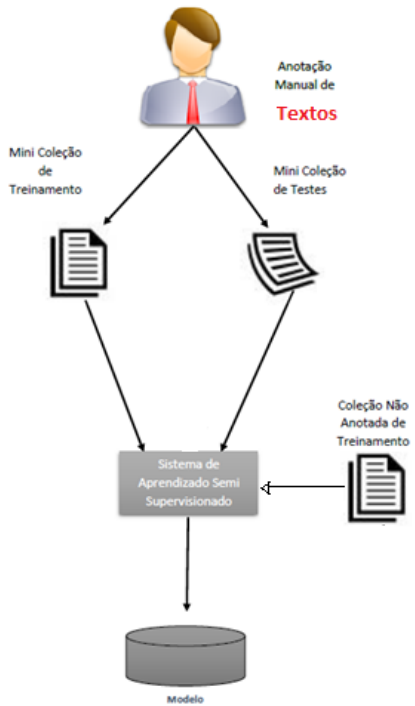

Figura 6: Estrutura de treinamento semi-supervisionado.

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.

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.

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

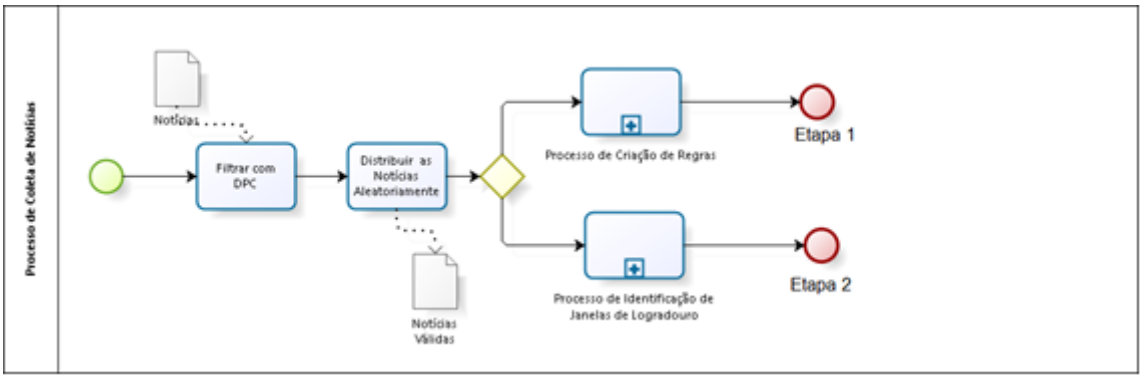

Figura 7: Estrutura inicial de coleta de notícias.

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 8 apresenta as tarefas necessárias para a criação das regras.

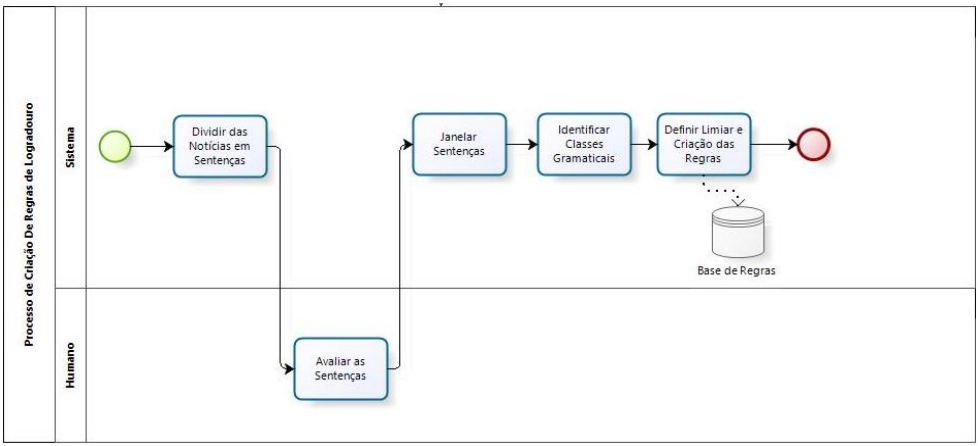

Figura 8: Etapas de criação de regras de logradouro.

O processo de criação de regras de logradouro é composto dos seguintes processos: a) divisão das notícias em sentenças, b) avaliação das sentenças, c) janelamento das sentenças; d) identificação das classes gramaticais - *postag* das janelas, e e) definição dos limiares e criação das regras.

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 9 apresenta essa etapa no fluxo geral de criação de regras.

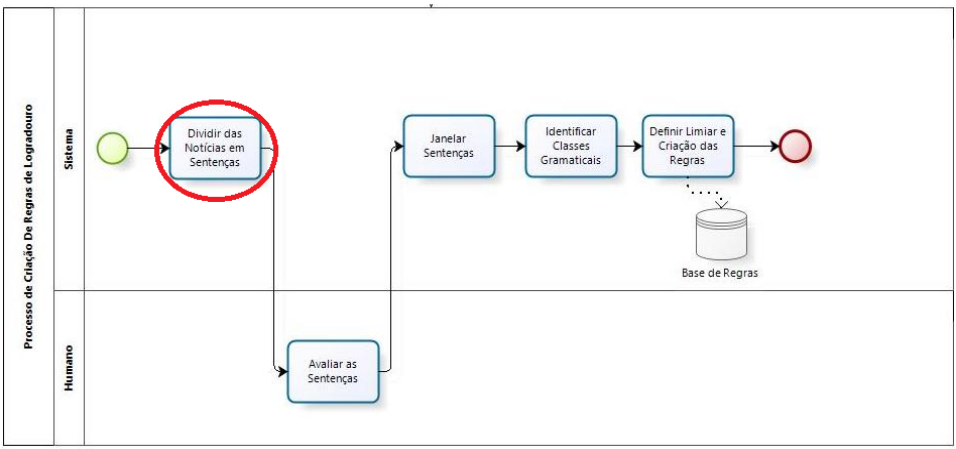

Figura 9: Divisão das notícias em sentenças.

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).

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 10 apresenta essa etapa no fluxo geral de criação de regras.

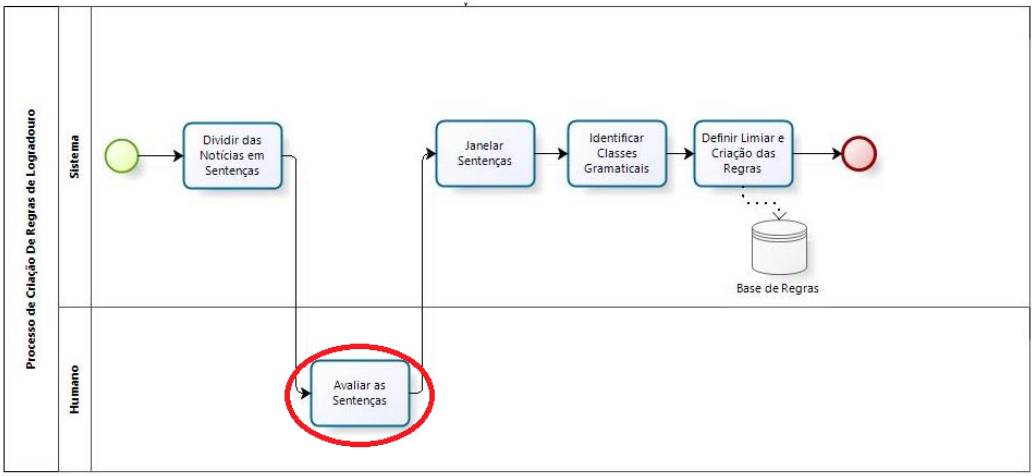

Figura 10: Avaliação das sentenças.

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 5 apresenta exemplo das sentenças a serem avaliadas:

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

Tabela 5: Sentenças para avaliação.

Janelar sentenças

O janelamento das sentenças é a tarefa de dividir a sentença em uma unidade de maior delimitação. A Figura 11 apresenta essa etapa no processo de criação de regras.

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 6.

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.

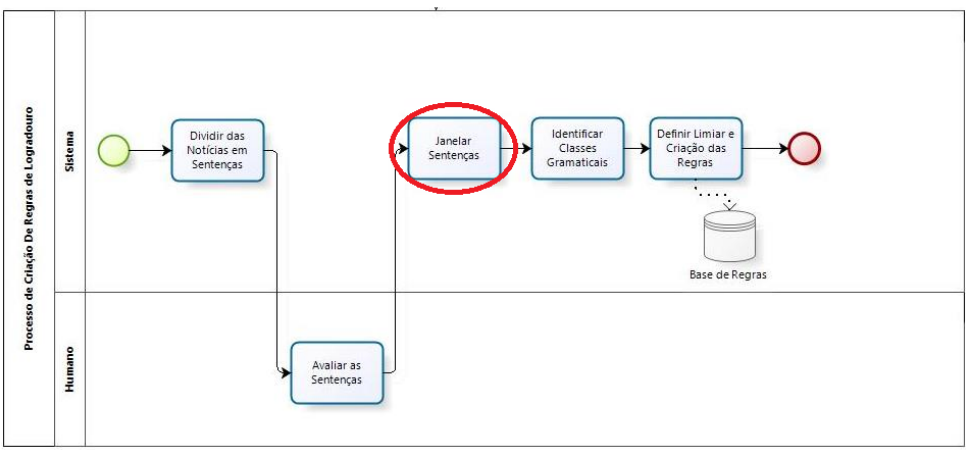

Figura 11: Janelar sentenças.

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.

Tabela 6: Exemplo de janelamento.

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:

$$Janelamento = Sentença\,(Palavra\,DPC, Palavra[1], \ldots Palavra[5])$$

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.

Identificar classes gramaticais

Essa tarefa é responsável por identificar as classes gramaticais dos termos que compõem a janela das notícias. A Figura 12 apresenta essa etapa no fluxo geral de criação de regras.

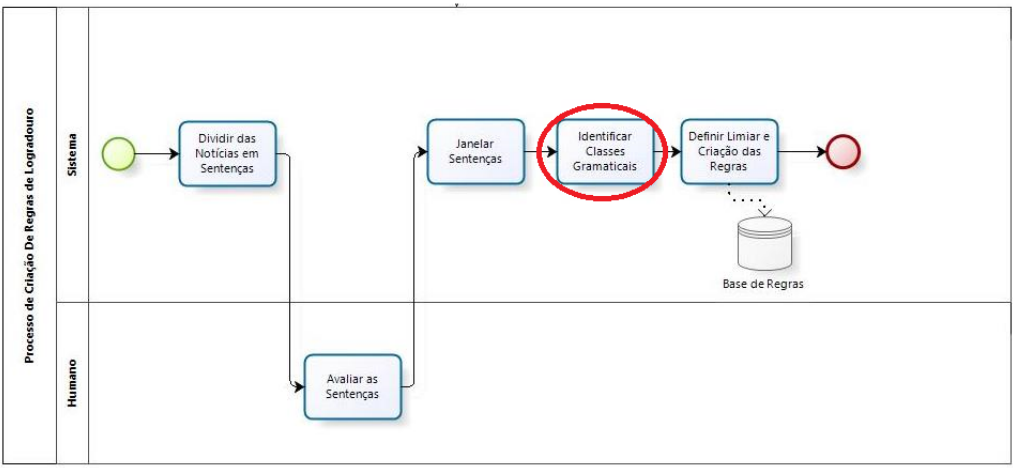

Figura 12: Identificar classes gramaticais.

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.

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 13 apresenta essa atividade no fluxo de criação de regras.

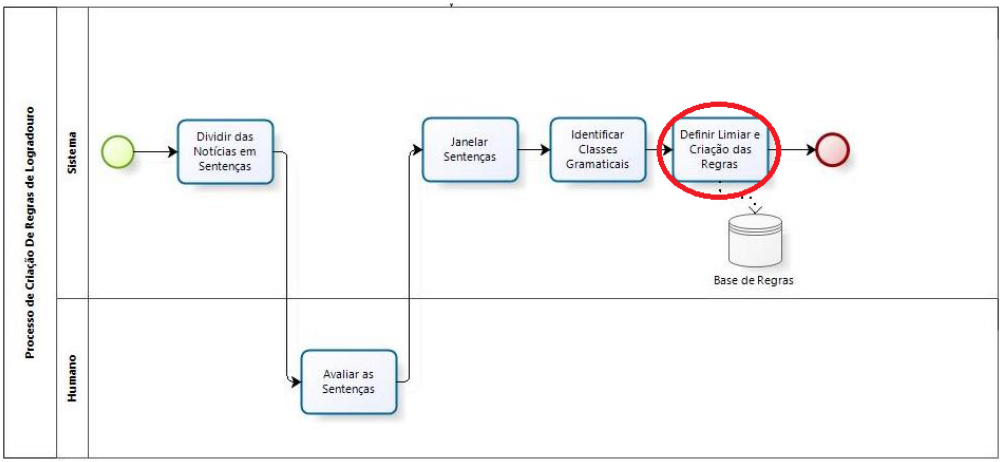

Figura 13: Definir limiares e criação das regras.

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 tendem 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.

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 14 apresenta as etapas:

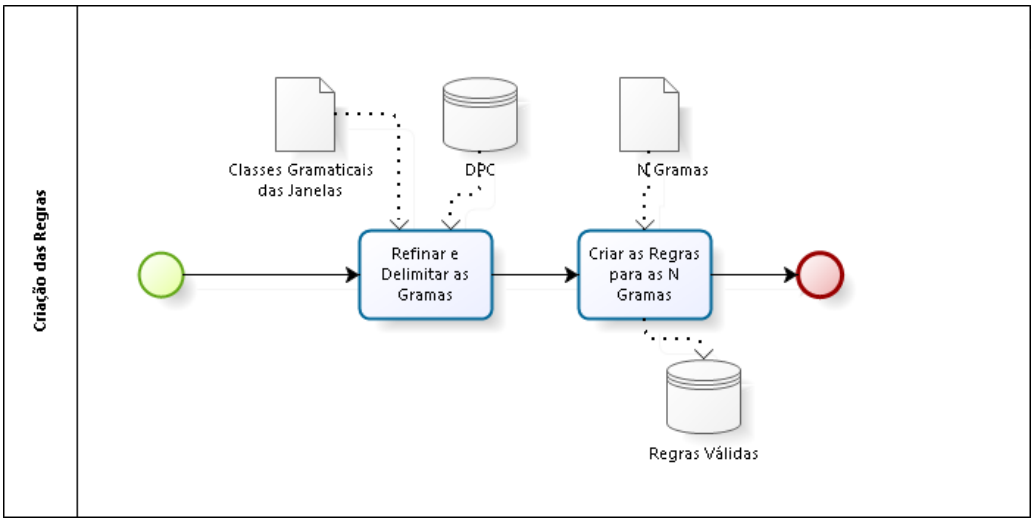

Figura 14: Criação das regras.

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 7 representa a 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

Tabela 7: Estrutura da janela refinada.

Conforme apresentado na Tabela 7, 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.

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 15, apresenta esse processo:

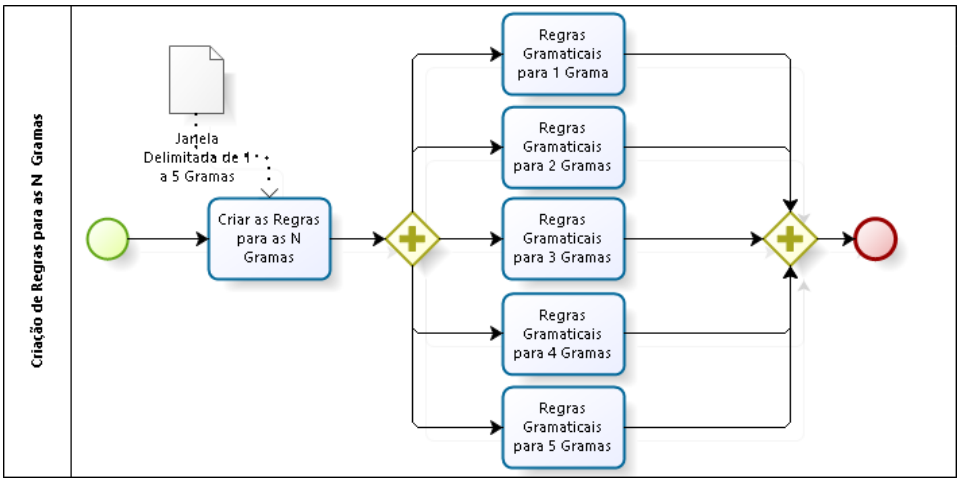

Figura 15 - Processo de criação de regras para as gramas.

Para as janelas que participaram do processo de avaliação de sentenças (4.3.2) são geradas as regras de grama. A Tabela 8 apresenta um exemplo das regras criadas para uma janela de exemplo:

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

Tabela 8: Exemplo de regras criadas para 1 janela.

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 16.

Figura 16: 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 17.

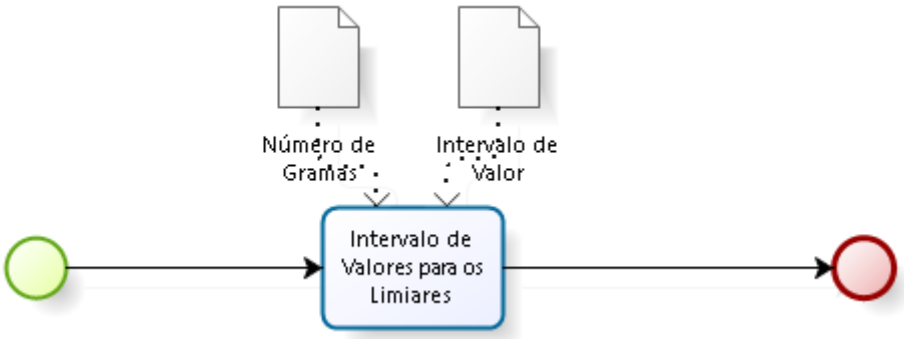

Figura 17: Intervalo de valores para os limiares.

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:

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

A Tabela 9 apresenta o intervalo de valores utilizados para cada grama, valores esses utilizados como “valor atual” na expressão acima.

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

Tabela 9: Intervalo de valores para os limiares.

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 10 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 v**álidas 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.

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

Tabela 10: Exemplo de valores de limiares.

O segundo processo dessa etapa consiste em determinar a melhor combinação de valores dos limiares. A Figura 18 apresenta esse processo.

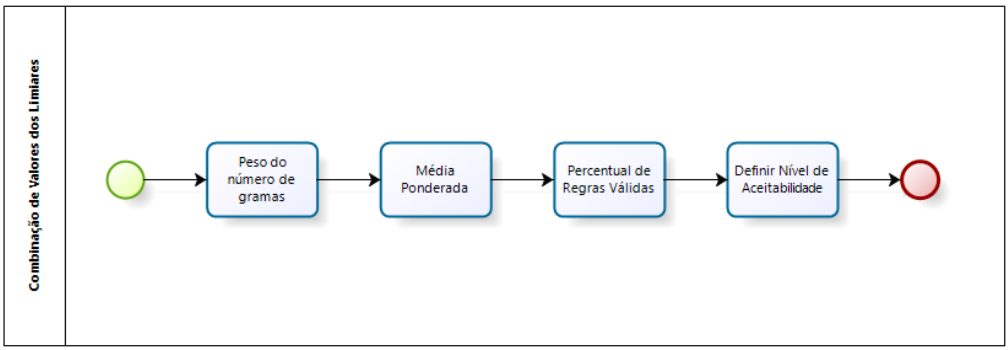

Figura 18: Processo combinação de limiares.

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.

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 11:

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

Tabela 11: Relação de peso por grama.

A Figura 19 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.

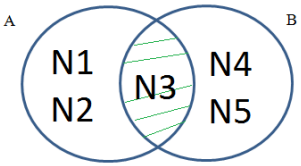

Figura 19: Relação de generalidade por grupos.

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:

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

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 12 apresenta um exemplo de uma regra com número de grama igual a 2, e a regra gramatical: ***preposição + verbo***.

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

Tabela 12: Estabelecimento de regras válidas.

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:

$$\boldsymbol{Percentual\ de\ Regras\ Válidas} = \frac{\boldsymbol{Quantidade\ de\ regras\ válida}}{\boldsymbol{Numero\ de\ quantidade\ da\ regra}}\ \boldsymbol{x\ 100}$$

Nesse caso, temos que:

$$\boldsymbol{Percentual\ de\ Regras\ Válidas} = \frac{\boldsymbol{2}}{\boldsymbol{3}}\ \boldsymbol{x\ 100}$$

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.

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 13 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.

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

Tabela 13: Relação das melhores regras.

Conforme apresentado na Tabela 13 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.

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 14 apresenta uma exemplificação da base de regras para cada grama:

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

Tabela 14: Exemplificação da base de regras geradas.

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.

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 20 apresenta os fluxos que compõem essa etapa.

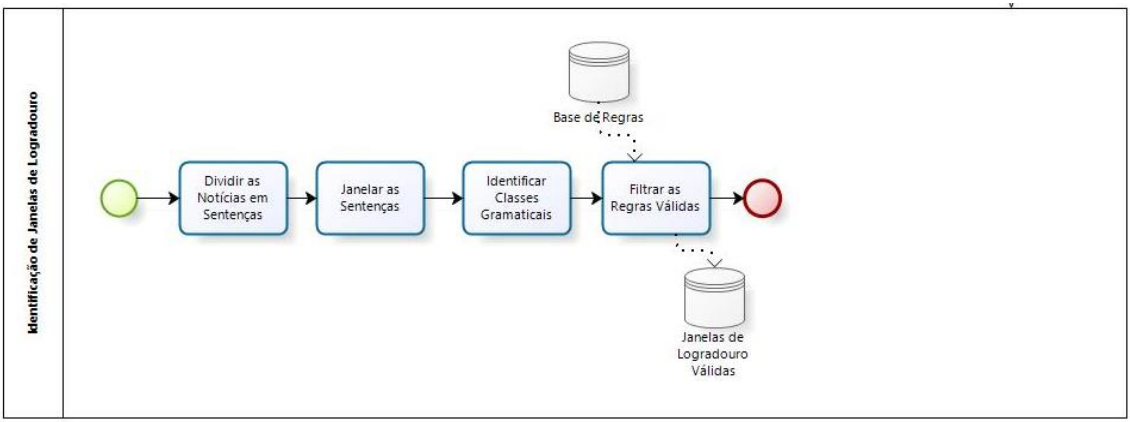

Figura 20: Fluxo de identificação de janelas com indicador de logradouro válido.

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 21 apresenta essa etapa no fluxo de identificação de janelas de logradouro.

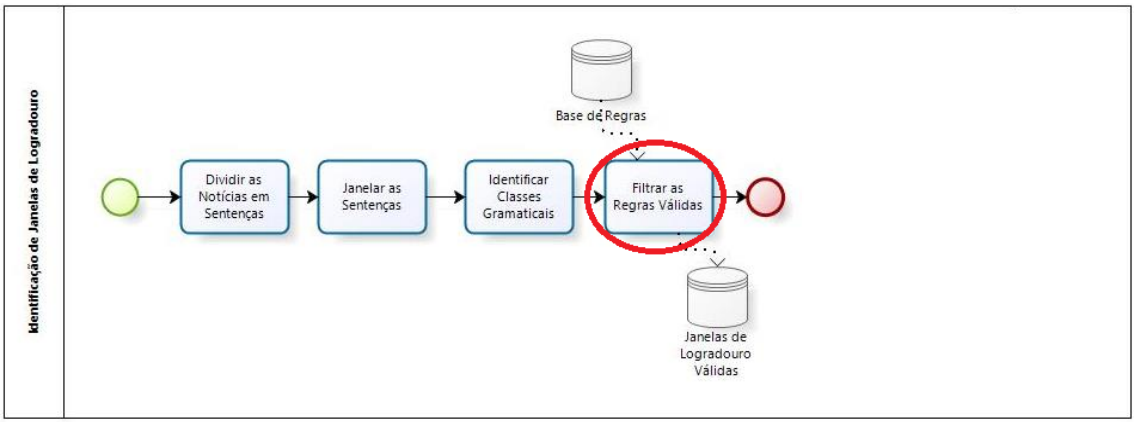

Figura 21: Filtragem das regras válidas.

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.

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 22 apresenta essa etapa no fluxo de identificação de janelas de logradouro.

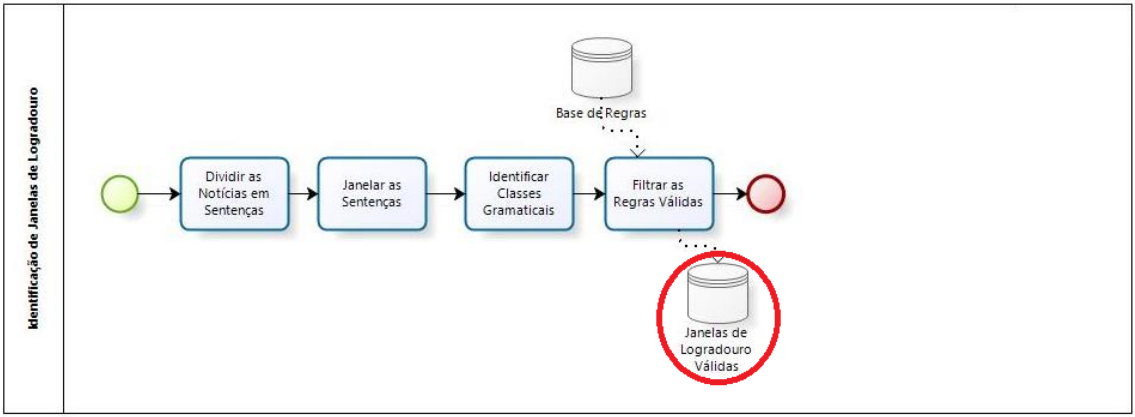

Figura 22: Janelas de logradouro válidas.

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.

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 23 apresenta o fluxo geral do processo de aprendizado de máquina.

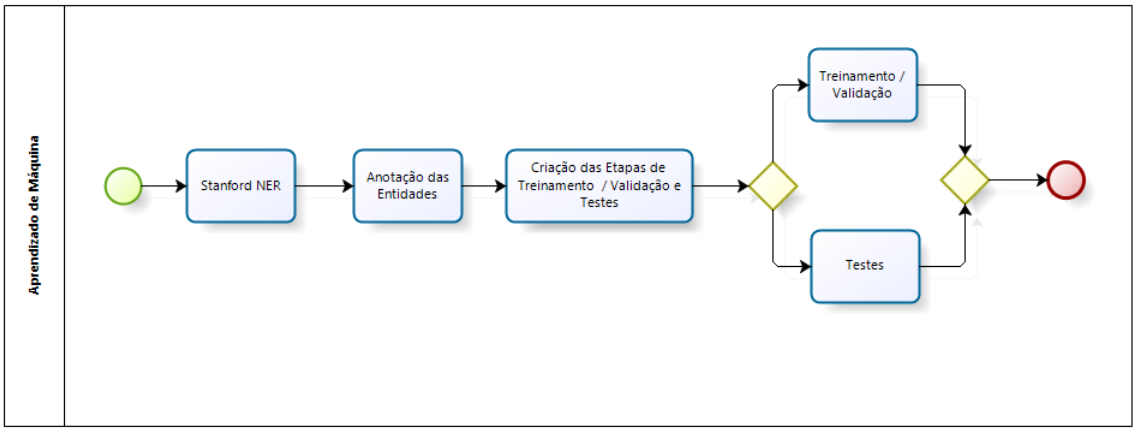

Figura 23: Processo de aprendizado de máquina.

A Tabela 15 apresenta um exemplo da necessidade de utilização da máquina de aprendizado para a identificação de entidades.

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

Tabela 15: Uso da máquina de aprendizado.

A Tabela 15 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.

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, 2008).

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.

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:

$$IO = Termo + espaço + Tipo\ da\ Entidade$$

A Tabela 16 apresenta um exemplo utilizando a notação IO necessário para a criação dos arquivos de anotação de entidades utilizado pelo Stanford NER.

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

Tabela 16: Representação IO.

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.

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

A Tabela 17 apresenta exemplo de anotação BIO necessário para a criação dos arquivos de anotação de entidade utilizado pelo Stanford NER.

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

Tabela 17: Representação BIO.

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*.

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 18, de maneira a identificar quais os parâmetros que possuem os melhores resultados.

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

Tabela 18: Parâmetros utilizados.

A Figura 24 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 24: Melhor resultado de parâmetros.

Testes

O processo de testes consiste em avaliar a capacidade dá máquina de aprendizado na tarefa de classificação das entidades mencionadas. A Figura 25 apresenta esse processo.

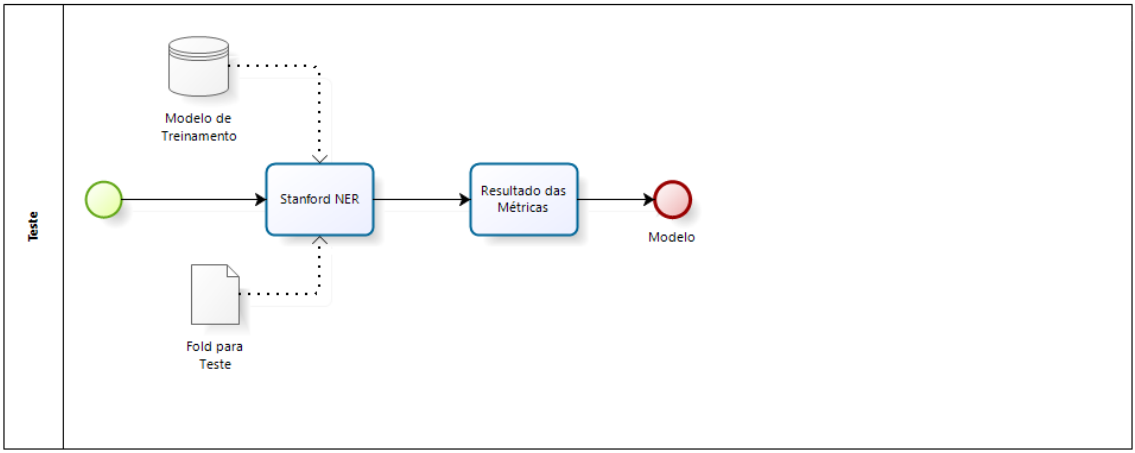

Figura 25: Processo de testes.

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.

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 se**ção** 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 26 apresenta as etapas da plataforma GeoNewsBR.

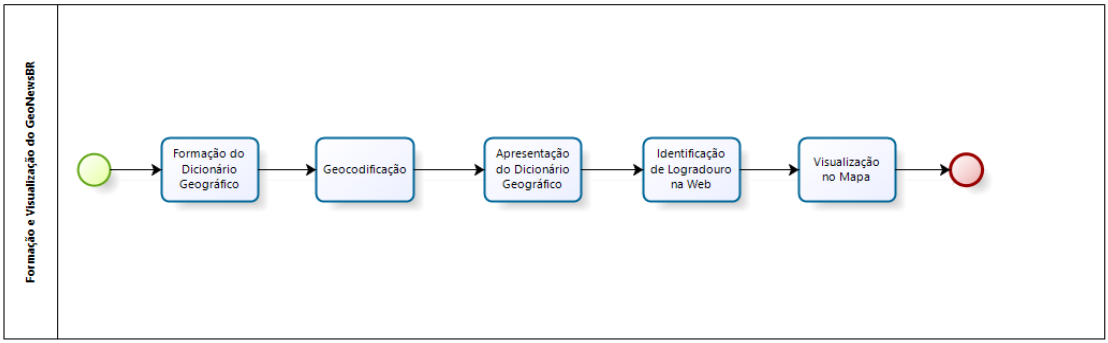

Figura 26: Fluxo de atividades formação e visualização do GeoNewsBR.

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 27 apresenta uma visão dessas atividades:

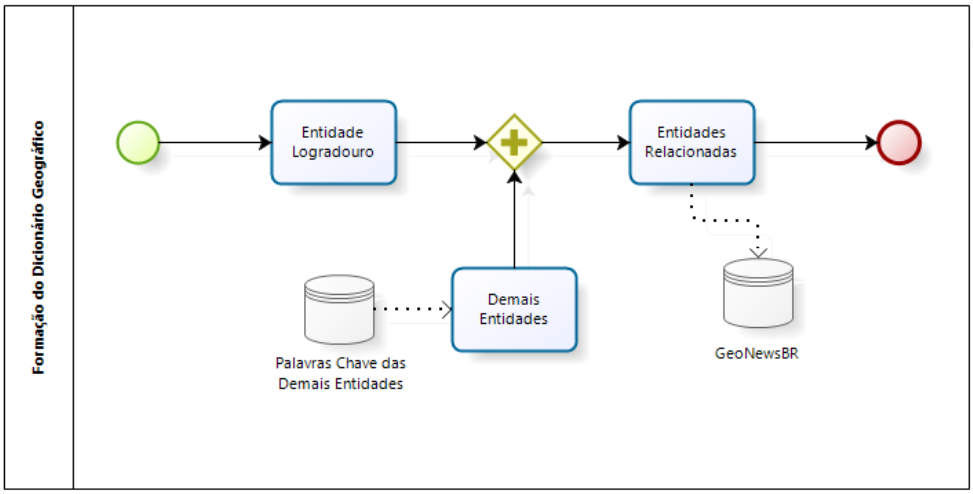

Figura 27: Processo de formação do dicionário geográfico.

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 19 detalha os tipos de categorias e as palavras chave que identificam as categorias nas notícias:

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.

Tabela 19: Demais categorias de entidades.

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 georreferenciamento dos elementos geográficos.

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 28 apresenta esse processo.

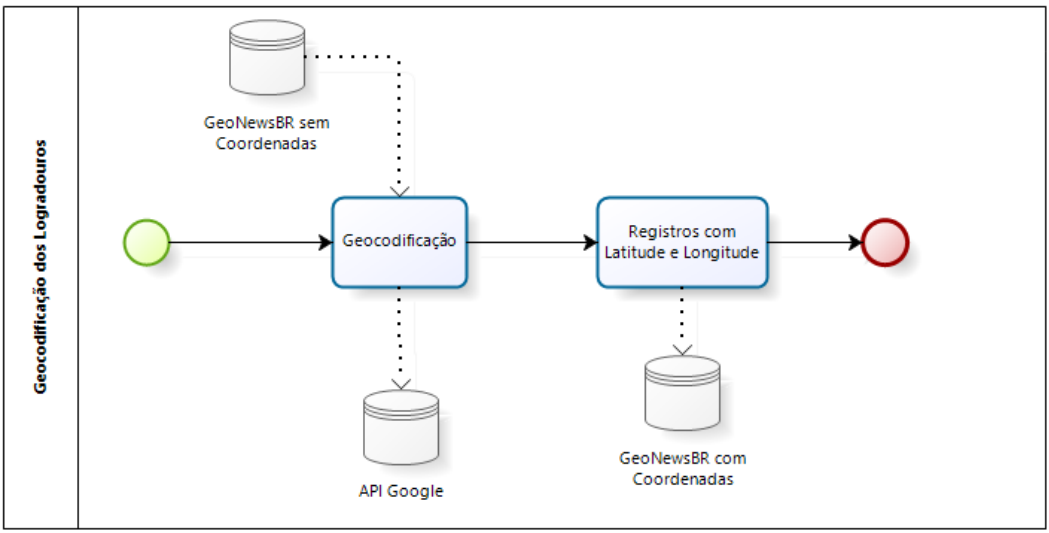

Figura 28: Processo de geocodificação.

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 29 apresenta a tela de apresentação das informações contidas no dicionário geográfico.

Figura 29: Tela com as informações do dicionário geográfico.

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 30 apresenta as etapas desse processo.

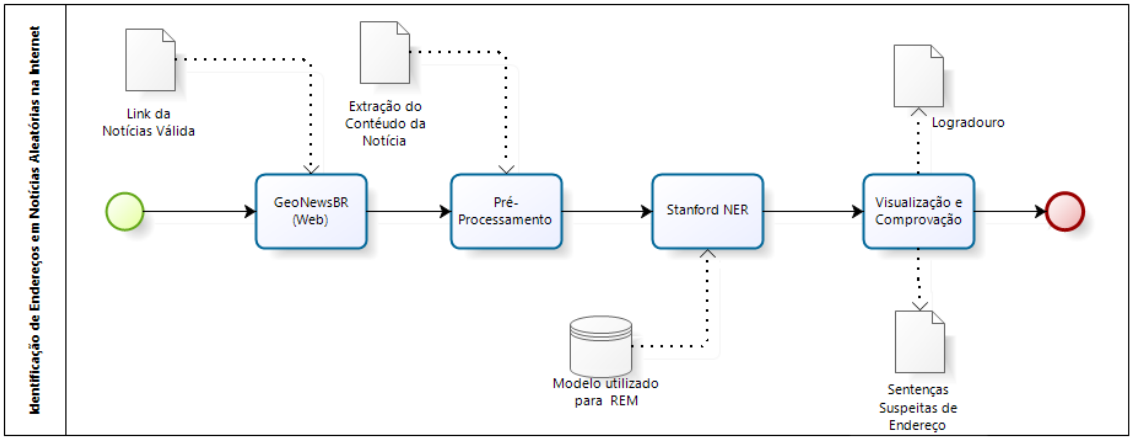

Figura 30: Processo de identificação de endereços em notícias da internet.

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 31 apresenta essa funcionalidade.

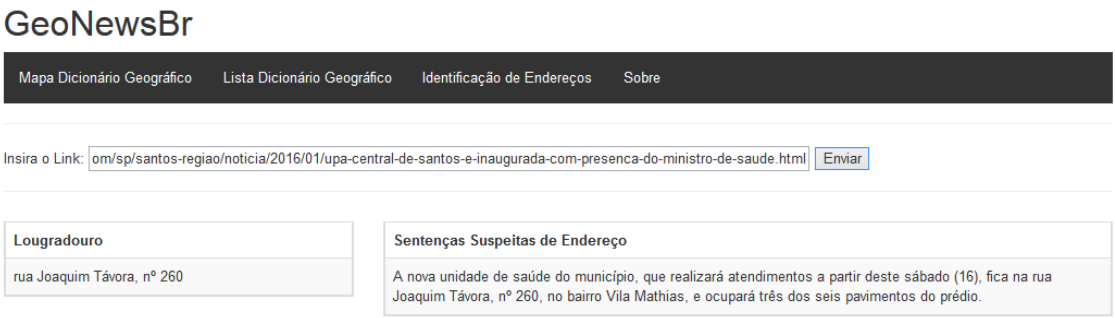

Figura 31: Identificação de endereços geográficos da internet.

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 32.

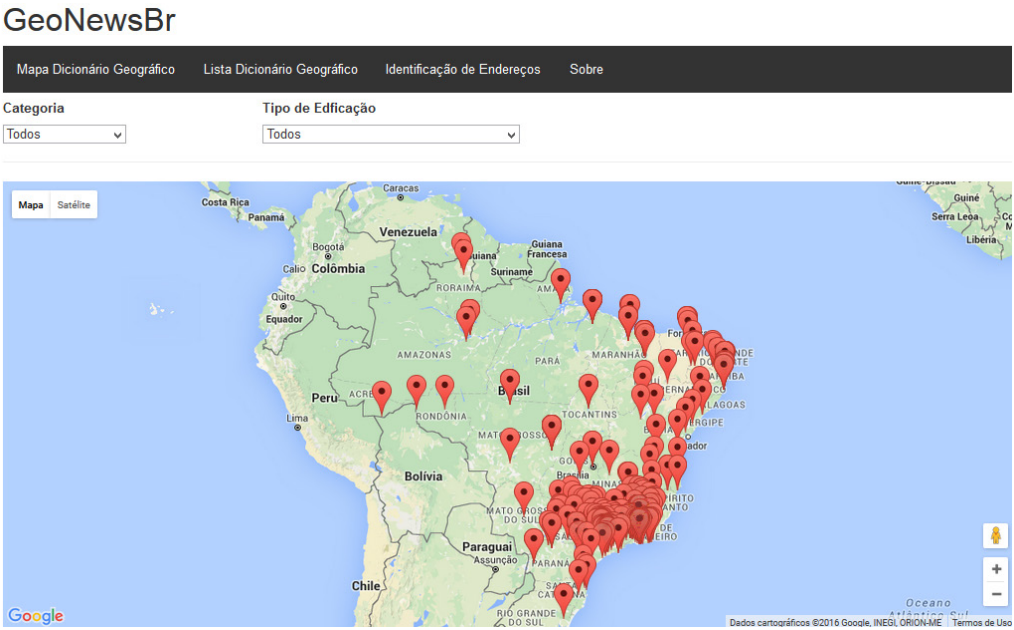

Figura 32: Tela de visualização no mapa.

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.

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*[1] 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.

Para a realização dos experimentos, foi necessário que o corpus Baseline apresentasse alguns padrões que permitam o uso das métricas propostas nesse trabalho,

1. Plural de corpus.

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 33 apresenta o fluxo geral das etapas utilizadas para a realização dos experimentos e analise dos resultados:

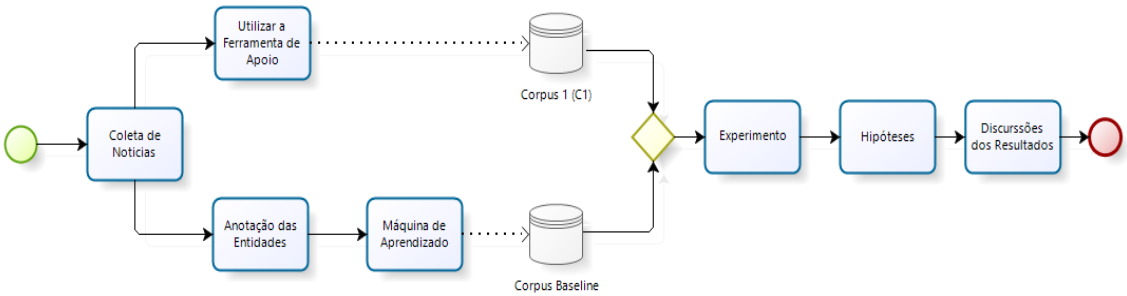

Figura 33: Estrutura de criação dos corpos dos experimentos.

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 pa-

drão 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.

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 20 apresenta os resultados médios gerados em cada corpus.

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

Tabela 20: Resultados do experimento 1.

A Figura 34 apresenta um comparativo entre os itens do experimento.

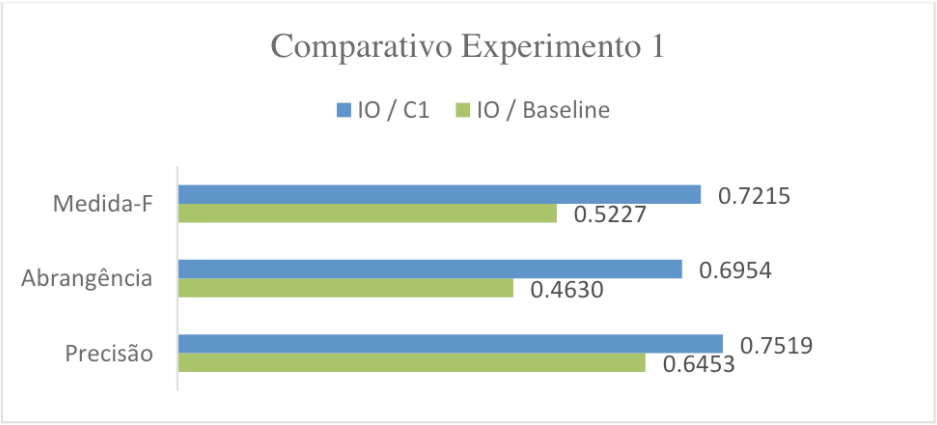

Figura 34: Comparativo experimento 1.

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 21 apresenta os resultados gerados.

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

Tabela 21: Resultados do experimento 2.

A Figura 35 apresenta um comparativo entre os itens do experimento.

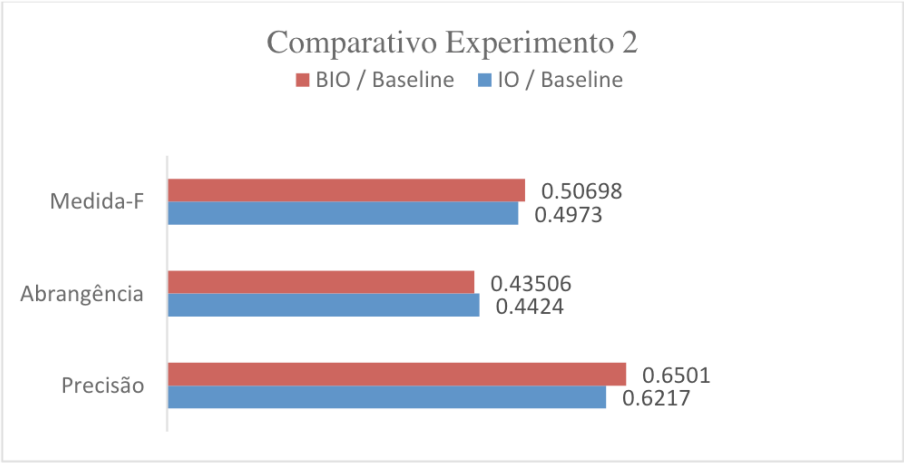

Figura 35: Comparativo experimento 2.

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.

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 22 apresenta os resultados gerados para cada tipo de notação/categoria.

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

Tabela 22: Resultados do experimento 3 com anotações IO e BIO.

A Figura 36 apresenta um comparativo entre os itens do experimento:

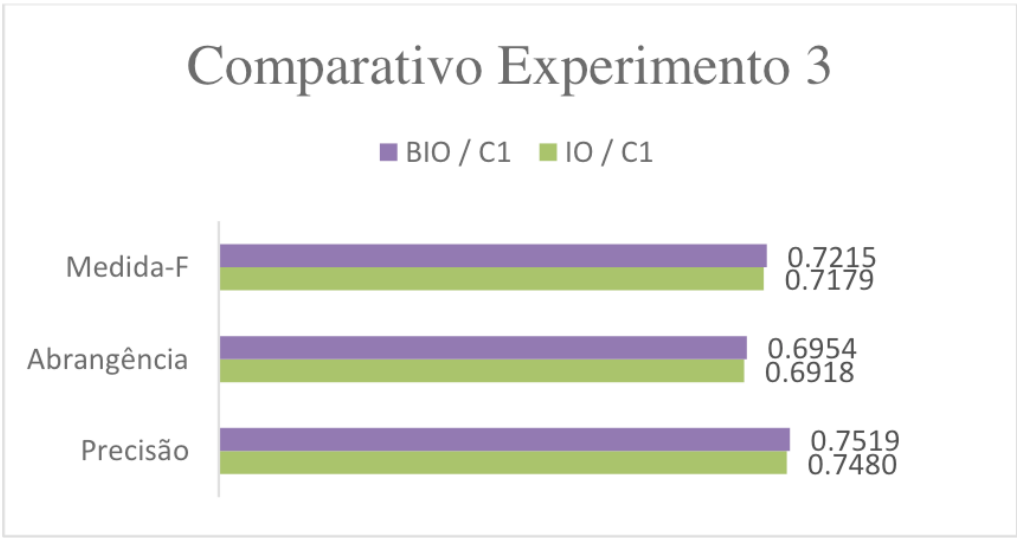

Figura 36: Comparativo experimento 3.

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

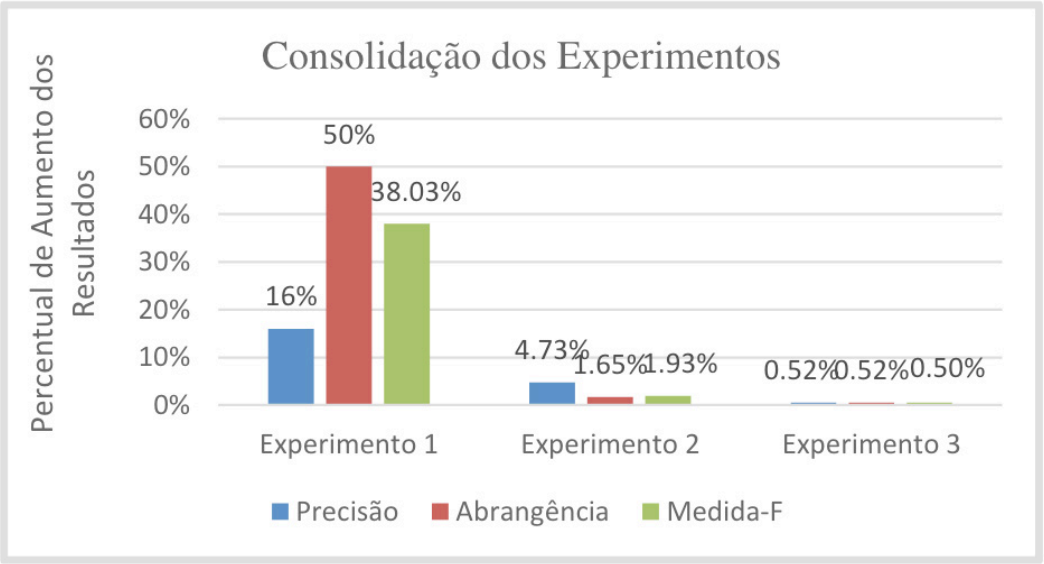

Figura 37: Percentual de ganho consolidado nos experimentos.

Conforme apresentado na Figura 37, 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.

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 melhores 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, tendo em vista que os resultados do processo geográfico serão mais precisos em relação ao posicionamento no globo terrestre em detrimento do fator de granularidade abordado nesse trabalho.

Foi também evidenciado que ao utilizar as diferentes anotações de entidades, IO e BIO, não havia interferência em alterações significativas no processo de reconhecimento de entidades nomeadas para logradouro, permitindo reduzir o esforço na escolha do padrão de anotação de entidades para essa tarefa a outras pesquisas.

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 em trabalhos futuros.

REFERÊNCIAS

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.

BATISTA, D. S. et al. **Geographic signatures for semantic retrieval**. Proceedings of the 6th Workshop on Geographic Information Retrieval. **Anais**...ACM, 2010. Disponível 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.

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, 2010.

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.

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.

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.

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;

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

GRISHMAN, R.; SUNDHEIM, B. **Message Understanding Conference-6: A Brief History.** COLING. **Anais**.1996. Disponível 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.

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.

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. Tese para obtenção do título de doutora em Ciência e Engenharia de Computação da University of West Bohemia. 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.

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.

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.

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

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

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.

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.

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**...2005.

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.

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.

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.

SOBRE O AUTOR

MATHEUS EMERICK DE MAGALHÃES – Mestre e Doutorando em Engenharia da Computação pela Universidade Federal do Rio de Janeiro e Pós-Graduado em Business Intelligence e Gerenciamento de Projetos. Há mais de 10 anos trabalha na área de Business Intelligence e Banco de Dados. Tem experiência em sistemas de apoio a tomada de decisão, data warehouse, business intelligence, tratamento de dados não-estruturados, BigData e gerenciamento de projetos. Atualmente é Capitão de Corveta da Marinha do Brasil, sendo condecorado com a premiação de primeiro colocado em todo o Brasil no curso de formação de oficiais de notório saber.

REVISOR:

DAYSIANNE KESSY MENDES ISIDORIO – Mestre em Ciência dos Materiais pelo Instituto Militar de Engenharia, pós-graduada em Processos de Soldagem e Metalurgia pelo grupo Prominas, graduada em engenharia de materiais pela Universidade Federal do Cariri.

A utilização de

MACHINE LEARNING

na identificação de elementos textuais geográficos

www.atenaeditora.com.br
contato@atenaeditora.com.br
@atenaeditora
www.facebook.com/atenaeditora.com.br

Atena Editora
Ano 2022

A utilização de

MACHINE LEARNING

na identificação de elementos textuais geográficos

www.atenaeditora.com.br
contato@atenaeditora.com.br
@atenaeditora
www.facebook.com/atenaeditora.com.br

Atena Editora
Ano 2022