UNIVERSIDADE FEDERAL DO CEARÁ

CENTRO DE CIÊNCIAS

DEPARTAMENTO DE COMPUTAÇÃO

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

MESTRADO ACADÊMICO EM CIÊNCIA DA COMPUTAÇÃO

BRENO ALEF DOURADO SÁ

RESOLUÇÃO DE TOPÔNIMOS EM TEXTOS NÃO ESTRUTURADOS BASEADA EM HEURÍSTICAS

FORTALEZA

2022

BRENO ALEF DOURADO SÁ

RESOLUÇÃO DE TOPÔNIMOS EM TEXTOS NÃO ESTRUTURADOS BASEADA EM HEURÍSTICAS

Dissertação apresentada ao Curso de Mestrado Acadêmico em Ciência da Computação do Programa de Pós-Graduação em Ciência da Computação do Centro de Ciências da Universidade Federal do Ceará, como requisito parcial à obtenção do título de mestre em Ciência da Computação. Área de Concentração: Ciência da Computação.

Orientador: Prof. Dr. José Antônio Fernandes de Macêdo.
Coorientadora: Prof.ª Dra. Ticiana Linhares Coelho da Silva.

FORTALEZA

2022

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Sistema de Bibliotecas
Gerada automaticamente pelo módulo Catalog, mediante os dados fornecidos pelo(a) autor(a)

S11r Sá, Breno Alef Dourado.
Resolução de topônimos em textos não estruturados baseada em heurísticas / Breno Alef Dourado Sá. – 2022.
66 f. : il. color.

Dissertação (mestrado) – Universidade Federal do Ceará, Centro de Ciências, Programa de Pós-Graduação em Ciência da Computação, Fortaleza, 2022.
Orientação: Prof. Dr. José Antônio Fernandes de Macêdo.
Coorientação: Profa. Dra. Ticiana Linhares Coelho da Silva.

1. Geocoding. 2. Resolução de topônimos. 3. Desambiguação baseada em heurística. 4. Topônimos adjetivos. 5. Tipos de topônimo. I. Título.

CDD 005

BRENO ALEF DOURADO SÁ

RESOLUÇÃO DE TOPÔNIMOS EM TEXTOS NÃO ESTRUTURADOS BASEADA EM HEURÍSTICAS

Dissertação apresentada ao Curso de Mestrado Acadêmico em Ciência da Computação do Programa de Pós-Graduação em Ciência da Computação do Centro de Ciências da Universidade Federal do Ceará, como requisito parcial à obtenção do título de mestre em Ciência da Computação. Área de Concentração: Ciência da Computação.

Aprovada em: 22/11/2022

BANCA EXAMINADORA

Prof. Dr. José Antônio Fernandes de Macêdo (Orientador)
Universidade Federal do Ceará (UFC)

Prof.ª Dra. Ticiana Linhares Coelho da Silva (Coorientadora)
Universidade Federal do Ceará (UFC)

Prof. Dr. Flávio Rubens de Carvalho Sousa
Universidade Federal do Ceará (UFC)

Dr. Vinícius Cezar Monteiro de Lira
Istituto di Scienza e Tecnologie dell'Informazione (ISTI-CNR)

A Deus.

À minha família.

Aos amigos que me apoiaram nessa caminhada.

Aos mestres.

AGRADECIMENTOS

A Deus, por tudo.

À minha família e amigos, pelo apoio e incentivo nas horas difíceis.

Ao meu orientador, Prof. Dr. José Macêdo, e à minha coorientadora, Prof.ª Dr.ª Ticiana Linhares, pelo acompanhamento e orientação ao longo desta pesquisa, além da colaboração na escrita e publicação do artigo resultante.

Aos membros da banca examinadora, Dr. Vinícius Monteiro e Prof. Dr. Flávio Sousa, pelo tempo dedicado a analisar esta dissertação e pelas valiosas contribuições e sugestões.

Aos professores do Programa de Pós-Graduação em Ciência da Computação da Universidade Federal do Ceará, por sua contribuição para minha formação acadêmica.

Aos demais colegas do *Insight Data Science Lab* e da Universidade Federal do Ceará, em especial ao Me. Anderson Severo de Matos, pelo suporte e repasse de conhecimento.

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), pelo apoio financeiro por meio da bolsa #133938/2019-0.

Ao Dr. Ednardo Moreira Rodrigues, e seu assistente, Alan Batista de Oliveira, pela adequação do *template* utilizado neste trabalho para que o mesmo ficasse de acordo com as normas da biblioteca da Universidade Federal do Ceará.

“Sabiamente comecei com um mapa [...]”

(TOLKIEN, 2006, p. 171)

RESUMO

Cotidianamente, pessoas utilizam nomes de lugares e relações espaciais para dar direções e informar o local de eventos. Menções a locais, também chamadas de topônimos, estão presentes nos mais variados tipos de documentos com conteúdo geográfico, como artigos, blogs, relatórios e relatos criminais. As informações geográficas extraídas desses documentos podem ser utilizada em aplicações de resposta à emergências, monitoramento de epidemias, agrupamento de notícias, planejamento turístico, entre outros. No entanto, devido a ausência de metadados, a extração dessas informações a partir de textos não estruturados não é uma tarefa trivial. Um dos desafios nesse processo é o mapeamento dos topônimos para coordenadas geográficas devido a ambiguidade dos nomes dos locais, que comumente possuem homônimos. O processo de resolução de topônimos para suas coordenadas, obtendo candidatos e os desambiguando, chama-se *geocoding*. O presente trabalho propõe e avalia duas heurísticas para *geocoding*: normalização de topônimos adjetivos e otimização geométrica por tipo de topônimo. Inicialmente, o *baseline* é definido através de experimentações com heurísticas. Em seguida, são criados dois *geocoders* modificando o *baseline* para utilizar cada uma das heurística proposta neste trabalho. Por fim, um terceiro *geocoder* é criado de forma semelhante para utilizar a combinação das duas heurísticas. Os resultados indicam uma melhora de desempenho do *geocoding* utilizando essas heurísticas em comparação com o *baseline*, chegando a superar *geocoders* do estado da arte nas bases de dados avaliadas.

Palavras-chave: *geocoding*; resolução de topônimos; desambiguação baseada em heurística; topônimos adjetivos; tipos de topônimo.

ABSTRACT

Everyday, people use place names and spatial relationships to give directions and inform the location of events. Mentions of places, also called toponyms, are present in the most varied types of documents with geographic content, such as articles, blogs, reports and criminal reports. The geographic information extracted from these documents can be used in emergency response applications, epidemic monitoring, news gathering, tourism planning, among others. However, due to the absence of metadata, extracting this information from unstructured texts is not a trivial task. One of the challenges in this process is the mapping of toponyms to geographic coordinates due to the ambiguity of the names of the places, which commonly have homonyms. The process of solving toponyms to their coordinates, obtaining candidates and disambiguating them, is called geocoding. The present work proposes and evaluates two heuristics for geocoding: normalization of adjectival toponyms and geometric optimization by toponym type. Initially, the baseline is defined through experiments with heuristics. Then, two geocoders are created by modifying the baseline to use each of the heuristics proposed in this work. Finally, a third geocoder is similarly created to use the combination of the two heuristics. The results indicate an improvement in the performance of geocoding using these heuristics compared to the baseline, even surpassing state-of-the-art geocoders in the databases evaluated.

Keywords: geocoding; toponyms resolution; heuristic-based disambiguation; adjectival toponyms; toponym types.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AUC	*Area Under the Curve*/Área Sob a Curva
CLAVIN	*Cartographic Location And Vicinity INdexer*
IDF	*Inverse Document Frequency*/Inverso da Frequência em Documentos
NER	*Named Entity Recognition*/Reconhecimento de Entidade Nomeada
TF	*Term Frequency*/Frequência do Termo

SUMÁRIO

1 INTRODUÇÃO

Cotidianamente, pessoas utilizam nomes de lugares para dar direções, informar localizações de eventos e transmitir informações espaciais baseadas no conhecimento comum desses nomes no geral (VASARDANI *et al.*, 2013). Menções a locais frequentemente se fazem presentes nos mais variados tipos de documentos com conteúdo geográfico, como notícias, blogs, relatórios e até mesmo mensagens em redes sociais. Tais menções podem ser referências diretas a locais, como a cidade de um evento, ou mesmo conceitos não locacionais que sejam relacionados a um lugar, como o embaixador de um país.

Informações geográficas podem ser utilizadas para diversas aplicações. Com elas é possível desenvolver sistemas de notificação de desastres (WU; CUI, 2018), resposta à emergências (SINGH *et al.*, 2019), monitoramento de epidemias (LAMPOS; CRISTIANINI, 2012), prevenção de crimes (VOMFELL *et al.*, 2018), agrupamento de notícias (ABDELKADER *et al.*, 2015), planejamento turístico (COLLADON *et al.*, 2019), entre outros. Entretanto, para utilizar essas informações prontamente, é necessário que elas sejam acessíveis.

O processo de extrair as informações geográficas de interesse embutidas em documentos em formato texto, como é o caso de notícias, não é trivial. A complexidade se dá devido a natureza não estrutural dos textos, isto é, a ausência de metadados nesses conteúdos escritos que impossibilita a indexação e mapeamento dessas informações prontamente para campos em uma base de dados. Dessa forma, é preciso processar os documentos obtendo as informações geográficas embutidas para que possam ser utilizadas.

O processamento automático de informações geográficas de notícias, por exemplo, permite aos pesquisadores extrair informações de eventos e utilizá-las para observar e obter informações de acontecimentos políticos relevantes à medida que estes ocorrem. O SPERG (GUNASEKARAN *et al.*, 2018) é uma dessas iniciativas que faz uso desse tipo de dado, focando-se primariamente em notícias de eventos políticos com o intuito de obter precisamente as localizações de todos os lugares mencionados no texto. Outra iniciativa é o ICEWS (O'BRIEN, 2010), que provê advertências antecipadas de conflitos violentos. Cientistas políticos utilizam essas informações para diversos propósitos de estudo, incluindo o impacto, perfil e local de concentração dos eventos.

Um outro exemplo de aplicação política da extração de informações geográficas são sistemas epidemiológicos de advertência antecipada. Dados epidemiológicos necessitam de tempo para tornarem-se disponíveis devido aos longos testes de laboratório. Por outro lado, dados

de mídias sociais, como Twitter e Facebook, têm sido utilizados para estudos epidemiológicos de diferentes doenças infecciosas, tais como Influenza (ALLEN *et al.*, 2016), Dengue (ALBINATI *et al.*, 2017), COVID-19 (JIANG *et al.*, 2021), etc. Identificar geograficamente esses dados textuais permite às autoridades planejar e agir apropriadamente com intervenções efetivas para controlar doenças infecciosas, reduzindo a mortalidade e morbidade na população.

O processo de mapeamento de referências a locais dentro de um texto, também chamadas de topônimos, para seus respectivos identificadores geográficos é conhecido como *geoparsing* e pode ser dividido nas etapas de reconhecimento, ou *geotagging*, e resolução, ou *geocoding* (GRITTA *et al.*, 2019). Devido à diversidade linguística dos domínios, ambiguidade dos nomes e outros problemas, essa área ainda é considerada um grande desafio (GRITTA *et al.*, 2018).

Nesse contexto, um dos grandes desafios é a obtenção das coordenadas geográficas referentes às localizações identificadas em um documento, pois é possível que um mesmo nome designe mais de um lugar no mundo. Um texto que menciona um local chamado "Springfield", por exemplo, pode estar referenciando mais de 30 lugares ao redor do mundo, como *Springfield em Oregon*, nos Estados Unidos, ou *Springfield em Queensland*, na Austrália. Relacionar corretamente os lugares em um texto às suas coordenadas correspondentes no planeta é fundamental para o uso eficiente desses dados pelas aplicações. Além disso, topônimos podem aparecer também em formas adjetivas, como na frase "O presidente **francês** iniciou o seu novo plano de governo".

Para melhor compreensão do processo e implicações de possíveis erros, pode-se considerar o seguinte exemplo: "O departamento de polícia de **Springfield** em **Oregon** reportou problemas de falta de pessoal". Nesse caso, deseja-se obter as coordenadas de latitude e longitude (lat=44.04624, lon=-123.02203), para "Springfield", e (lat=44.00013, lon=-120.50139), para "Oregon". Entretanto, existem diversos locais ao redor do globo nomeados assim, como pode-se visualizar na Figura 1, onde as localizações corretas encontram-se destacadas com a cor verde. Para esse exemplo, caso a resolução das referências a locais apontasse equivocadamente para lugares na Austrália, tal confusão prejudicaria consideravelmente uma aplicação que fizesse uso dessa informação, uma vez que estão a uma grande distância das coordenadas corretas.

Portanto, devido à alta ambiguidade dos nomes de lugares, é necessário que um sistema de *geocoding* faça mais do que simplesmente checar uma lista de nomes de lugares pelo mundo. Este trabalho objetiva experimentar um conjunto de heurísticas para realização de

Figura 1 – Potenciais coordenadas para as referências do exemplo

Fonte: Elaborado pelo autor
Destacados em verde estão as localizações esperadas para as referências “Springfield” e “Oregon” na frase “O departamento de polícia de Springfield em Oregon reportou problemas de falta de pessoal”. Em vermelho estão os demais possíveis lugares.

geocoding.

1.1 Objetivos

Dada a necessidade das aplicações, conforme mencionado anteriormente, de capturar as informações geográficas em textos, o presente trabalho tem como objetivo geral propor e avaliar novas estratégias de resolução de referências a lugares para coordenadas geográficas.

O escopo geral do projeto pode ser dividido nos seguintes objetivos específicos:

OE1 Criar um dicionário geográfico para obtenção de coordenadas geográficas;

OE2 Avaliar heurísticas estabelecidas de resolução de topônimos e elaboração de um *baseline*;

OE3 Elaborar estratégia de aprimoramento do *baseline* para topônimos em forma adjetiva;

OE4 Elaborar estratégia de aprimoramento do *baseline* baseada na diferenciação de topônimos literais e associativos;

OE5 Comparar resultados das estratégias de aprimoramento com o desempenho do *base-*

line;

OE6 Comparar desempenho das estratégias de aprimoramento com abordagens do estado da arte.

1.2 Contribuições

Neste trabalho, são propostas duas estratégias de aprimoramento para o *geocoding* utilizando heurísticas. A primeira delas trata de topônimos em forma adjetiva (e.g., "A fábrica **chinesa**") e possibilita o *geocoding* de topônimos que, por estarem flexionados em uma forma diferente do nome real do local, não poderiam ser resolvidos corretamente. Já a segunda estratégia trata do *geocoding* de topônimos diferenciando os literais (e.g., "A cidade de **Londres**") dos associativos (e.g., "A rainha da **Inglaterra**"), de forma a reduzir a influência de conceitos relacionados a locais na resolução de referências diretas a locais físicos.

Os seguintes itens resumem as principais contribuições deste trabalho:

- Estratégia de *geocoding* tratando topônimos em forma adjetiva;
- Estratégia de *geocoding* diferenciando topônimos literais e associativos na desambiguação.

1.3 Publicações

Os esforços durante o processo de pesquisa para este trabalho possibilitaram a publicação do seguinte artigo:

- Sá, B. D.; SILVA, T. Coelho da; MACEDO, J. A. Fernandes de. Enhancing geocoding of adjectival toponyms with heuristics. In: Proceedings of The LREC 2022 workshop on Natural Language Processing for Political Sciences. Marseille, France: European Language Resources Association, 2022. p. 37–45.

1.4 Estrutura

Este documento está organizado da seguinte forma.

- O Capítulo 2 apresenta em detalhes os conceitos fundamentais para compreensão do trabalho e define formalmente o problema atacado;
- O Capítulo 3 trata dos trabalhos relacionados encontrados na literatura;
- O Capítulo 4 apresenta os dados utilizados, métricas de avaliação utilizadas e métodos aplicados;

- O Capítulo 5 apresenta os resultados das experimentações realizadas;
- O Capítulo 6 apresenta as conclusões e sugestões para trabalhos futuros.

2 DEFINIÇÃO DO PROBLEMA E CONCEITOS PRELIMINARES

Neste capítulo são apresentados os conceitos fundamentais para a compreensão do restante do trabalho. Além disso, este capítulo apresenta também uma definição formal do problema atacado.

2.1 Topônimo

Um topônimo é uma entidade nomeada que rotula uma localização específica, a qual pode ser qualquer um dos infinitos pontos físicos na Terra identificáveis por coordenadas. Em outras palavras, pode-se dizer que os topônimos são um subconjunto desses pontos, uma vez que nem todos podem ser identificados por nomes.

Apesar da definição geral de topônimos ser útil para diferencia-los de outras entidades nomeadas, a ausência de uma classificação mais detalhada resultou em várias divergências entre trabalhos que tratam desse tipo de entidade. Objetivando resolver esse problema, no trabalho em que buscaram padronizar as definições de topônimos, (GRITTA *et al.*, 2019) estabeleceram uma taxonomia onde essas entidades nomeadas são classificadas de acordo com a posição em que ocorrem dentro de sintagmas.

Em termos gramaticais, sintagmas são unidades sintáticas compostas de um ou mais vocábulos e formam as orações. Dependendo de seu núcleo, podem ser classificados como nominais, verbais, adjetivais, adverbiais ou preposicionais. Para classificação de topônimos, apenas os sintagmas nominais, onde o núcleo é um substantivo, são de interesse, pois são neles que essas referências a lugares aparecem (GRITTA, 2019).

A classificação das entidades nomeadas em questão possui dois níveis: categoria e tipo. Os topônimos são categorizados como Literais ou Associativos dependendo da oração em que o sintagma está contido. Em seguida, os topônimos são classificados em tipos, que dependem de sua categoria e serão explicados em detalhes nas subseções que se seguem.

2.1.1 Topônimos Literais

Os topônimos literais são aqueles que se referem a lugares onde algo acontece ou está fisicamente localizado. Pode-se entender a importância de diferenciar topônimos literais e associativos analisando as frases “Trabalhadores **brasileiros**” e “Trabalhadores **no Brasil**”. Na primeira frase, os trabalhadores podem estar espalhados por diversas localizações geográficas,

enquanto na segunda frase o topônimo indica especificamente o território da República Federativa do Brasil.

Um topônimo é categorizado como literal dependendo do seu contexto, isto é, da oração em que o sintagma que o contém se encontra. Se o contexto indica um literal (e.g., "A reunião foi realizado em **Paris**") ou é ambíguo (e.g., "A cidade de **Londres** votou contra o Brexit"), o topônimo é literal. Já se o contexto for associativo, o topônimo é literal apenas se o núcleo que modifica for literal, isto é, se o substantivo que caracteriza indica que o topônimo referencia o local físico onde algo ocorre ou está localizado (e.g., "O escritório de **São Paulo** contratou 5 funcionários").

Os topônimos literais podem ser classificados de acordo com os tipos listados a seguir:

- **Literal**: uma referência direta a um local físico, com a semântica e o contexto indicando que o topônimo é literal. Exemplo: "A temperatura no **Ceará** está muito alta";
- **Coerção**: uma entidade polissêmica que normalmente seria classificada como não-topônimo, mas que em um contexto literal têm seu sentido coagido para uma localização física. Exemplo: "O jóquei belga marcou sua primeira **Copa do Mundo de Dubai** em nove tentativas";
- **Misto**: ocorre em um contexto ambíguo com características de topônimo associativo, podendo ter um significado associativo ou literal. Esses casos são um meio-termo entre topônimos literais e associativos, mas são categorizados como literais. Exemplo: "**São Paulo** está gerando muita poluição";
- **Literal Embutido**: um não-topônimo contido dentro de uma entidade maior. É semanticamente, mas não sintaticamente, semelhante a um Modificador Literal. Exemplo: "Olimpíadas de **Tóquio**";
- **Modificador Substantivo**: um topônimo que modifica um núcleo literal de um sintagma, podendo o contexto ser literal ou não, desde que o núcleo o seja. Exemplo: "Ouvimos falar muito bem das belas praias do **Rio de Janeiro**";
- **Modificador Adjetivo**: apresenta o mesmo padrão de um modificador substantivo, mas com o topônimo na forma adjetiva. Muitas vezes pode ser rotulado como uma nacionalidade ou grupo político. Exemplo: "A tundra **russa** apresenta temperaturas baixíssimas".

2.1.2 Topônimos Associativos

Apesar de muitas vezes os topônimos encontrados em um texto serem literais, nem sempre esse é o caso. Em vez disso, topônimos podem frequentemente se referir, substituir ou modificar outros conceitos que não são locacionais mas estão relacionados a lugares. Tais topônimos são chamados de associativos.

De uma forma geral, um topônimo é categorizado como associativo sempre que o contexto for associativo, isto é, se indica que o topônimo está relacionado a um conceito não locacional. Entretanto, caso o topônimo modifique um núcleo, a categoria só é associativa se esse núcleo não for literal. Em outras palavras, para adjetivos e locuções adjetivas o topônimo só pertence a essa categoria se o substantivo que caracteriza não indicar que ele é uma referência a um local físico onde algo ocorre ou está localizado (e.g., "O jogador do **Uruguai** foi titular no Brasileirão").

Os topônimos associativos podem ser classificados de acordo com os tipos listados a seguir:

- **Gentílico**: é derivado de um topônimo e designa os habitantes de uma região. Exemplo: "O autor **brasileiro** Monteiro Lobato escreveu muitos livros";
- **Língua**: indica um idioma e não deve ser interpretado como um modificador. Exemplo: "Como se diz 'olá' em **vietnamita**?";
- **Metonímia**: uma figura de linguagem onde um conceito é substituído por outro relacionado. Exemplo: "**Barcelona** e **Liverpool** se enfrentam hoje";
- **Modificador Substantivo**: topônimos que modificam o núcleo de sintagmas nominais em um contexto associativo. Exemplo: "A chanceler da **Alemanha** visitou a Embaixada da **República Federal da Alemanha** no **Brasil**";
- **Modificador Adjetivo**: é semanticamente idêntico a um modificador substantivo associativo, mas na forma adjetiva. Exemplo: "Estamos todos torcendo para a seleção **brasileira**";
- **Associativo Embutido**: um não-topônimo contido dentro de uma entidade maior. Semanticamente, mas não sintaticamente, semelhante a modificadores associativos. Exemplo: "A assessoria informou ao **Washington** Post que a viagem foi pré-planejada";
- **Homônimo**: uma palavra com a mesma escrita, mas com significado diferente. Exemplo: "**Chelsea** sentou-se próximo a **Paris** na sala".

2.2 Gazetteer

Um *gazetteer* é um dicionário geográfico que contém informações sobre lugares, como nomes, coordenadas geográficas e categorias dos locais (HILL, 2000; GOODCHILD; HILL, 2008). Além disso, esses dicionários podem trazer outras informações, como população, nomes alternativos, etc. Por auxiliarem no mapeamento de nomes de locais para identificadores geográficos, esses dicionários constantemente estão presentes nos métodos de associação dos topônimos às suas coordenadas. Como exemplo de *gazetteers* conhecidos e bastante utilizados, é possível listar: GeoNames[1], OSM[2] e Google Maps[3]. Dentre eles, o mais popular em trabalhos de *geocoding* é o GeoNames.

O GeoNames é uma base de dados geográfica que contém mais de 12 milhões de registros sob uma licença *Creative Commons*, sem custos para download. Dentre as informações disponíveis pra cada registro, pode-se destacar as seguintes:

- **Id**: Identificador do registro no GeoNames;
- **Nome**: Nome completo do local em UTF-8;
- **Nome ASCII**: Nome completo do local em ASCII;
- **Nomes Alternativos**: Nomes alternativos do local, em ASCII, separados por vírgula;
- **Latitude**: Latitude em graus decimais (WGS 84);
- **Longitude**: Longitude em graus decimais (WGS 84);
- ***Feature Class***: Classe do local[4], podendo ser A (país, estado, região, etc.), H (rio, lago, etc.), L (parque, área, etc.), P (cidade, vila, etc.), R (estrada, ferrovia, etc.), S (prédio, fazenda, etc.), T (montanha, colina, etc.), U (locais submarinos) ou V (floresta);
- ***Feature Code***: Código tipificando o local[5] de acordo com seu *feature class*, podendo ser PCLI (entidade política independente), OCN (oceano), ST (rua), PPL (local populado), entre outros;
- **Código do País**: Um código de 2 letras, de acordo com o padrão ISO-3166, para o país do registro (e.g., "BR" para o Brasil);
- **Código de Nível Administrativo 1**: Um código para o primeiro nível administrativo do registro (e.g., 06 para o Ceará);

[1] https://www.GeoNames.org/
[2] https://www.openstreetmap.org/
[3] https://www.google.com.br/maps
[4] http://www.geonames.org/export/codes.html
[5] http://www.geonames.org/export/codes.html

- **Código de Nível Administrativo 2**: Um código para o segundo nível administrativo do registro (e.g., 2304400 para Fortaleza);
- **População**: Número estimado de habitantes do local;
- **Data de Modificação**: Data da última alteração do registro.

2.3 Geoparsing

O *Geoparsing* é um processo especial de extração e resolução de topônimos em texto livre para identificadores geográficos não ambíguos, como coordenadas de latitude e longitude. Um sistema que realiza a tarefa de *geoparsing* é denominado *geoparser* e é compreendido neste trabalho, adaptando a definição de Nizzoli *et al.* (2020), da seguinte forma:

Definição 2.3.1 *Dada uma sequência de tokens* $W =< w_1, w_2, \ldots, w_n >$*, um Geoparser é um modelo* G_p *tal que* $G_p(W) =< p_1, p_2, \ldots, p_m >$*, onde* p_i*,* $0 \leq i \leq m$*, é a tupla (latitude, longitude) correspondente ao topônimo* $t_i =< w_a, \ldots, w_b >$*,* $0 \leq a \leq b \leq n$*, detectado em* $\boldsymbol{W}$*.*

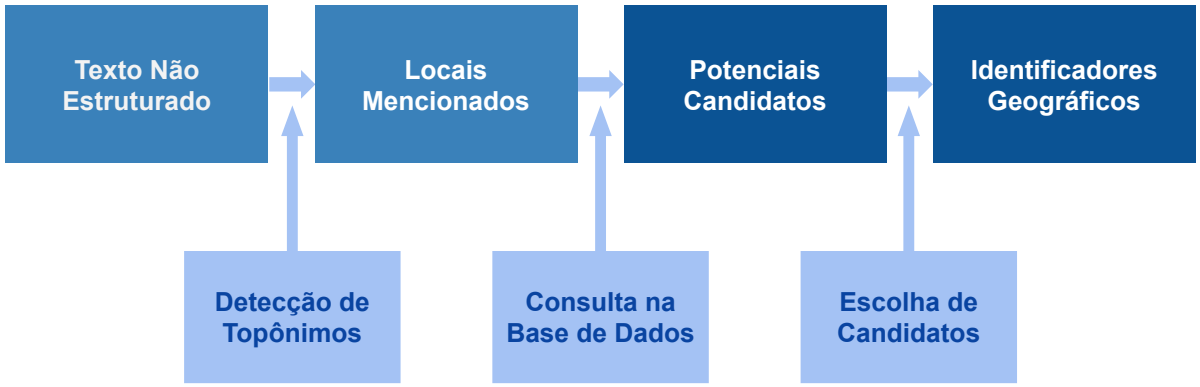
O primeiro passo que compõe o processo de *geoparsing* é chamado de *geotagging* e consiste em reconhecer partes do texto que correspondem à entidade geográficas, como endereços ou menções de lugares. Esse passo também é chamado de reconhecimento ou extração de topônimos. Por tratar de entidades nomeadas presentes em um texto, o *geotagging* é também considerado um caso especial de *Named Entity Recognition*/Reconhecimento de Entidade Nomeada (NER) (GRITTA *et al.*, 2018). Um sistema que realiza essa tarefa é denominado *geotagger*.

O segundo passo do *geoparsing* chama-se *geocoding* e consiste em atribuir às entidades reconhecidas os identificadores geográficos mais prováveis, como por exemplo coordenadas de GPS. Esse passo também é chamado de resolução de topônimos. Um sistema que realiza essa tarefa é chamado de *geocoder*.

Dependendo do escopo da aplicação, o processo de resolução de topônimos pode exigir um passo de desambiguação. Ao lidar com regiões maiores para atribuição das coordenadas, a probabilidade de existirem locais homônimos torna-se maior também. Essa ambiguidade está fortemente presente no *geocoding* de localizações granulares, como prédios e ruas, mas também pode ocorrer com cidades e até países, como é o caso do país Angola e a cidade de Angola no estado americano de Indiana. A desambiguação nesses casos pode ser realizada através de estratégias baseada em mapa, bases de conhecimento ou mesmo por técnicas de aprendizado de máquina Buscaldi e Rosso (2008).

A Figura 2 apresenta uma visão geral de um *geoparser*. Inicialmente o texto não estruturado passa por um processo de detecção de topônimos. Em seguida, por meio de consultas às bases de dados, são obtidos os possíveis identificadores geográficos para os topônimos encontrados no texto. Após uma filtragem e escolha dos candidatos mais adequados, conforme a estratégia de desambiguação, o *geoparser* retorna o conjunto de identificadores geográficos dos topônimos do texto.

Figura 2 – Visão Geral de um *Geoparser* Genérico



Fonte: Elaborado pelo autor

2.4 Definição do Problema

Este trabalho enquadra-se na área de *geoparsing*, porém concentra-se na etapa de *geocoding* do processo. Estabelecidos os conceitos previamente definidos, o problema que pretende-se resolver pode ser formalmente definido da seguinte maneira:

Definição 2.4.1 *Dada uma lista de topônimos* $< t_1, t_2, \ldots, t_n >$ *extraídos de um texto não estruturado e um conjunto* M *de tuplas (latitude, longitude) correspondentes às localizações conhecidas, construir um geocoder* G_c *tal que* $G_c(< t_1, t_2, \ldots, t_n >) =< p_1, p_2, \ldots, p_n >$*, onde* $p_i \in M$*,* $0 \leq i \leq n$*, é a tupla (latitude, longitude) referente a* t_i*.*

3 TRABALHOS RELACIONADOS

Os trabalhos que tratam de *geocoding* possuem metodologias distintas dependendo do objetivo da tarefa. Alguns trabalhos buscam resolver um documento inteiro para um único local, como é o caso de Rahimi *et al.* (2015) que trata da localização de usuários do Twitter. O presente trabalho, porém, assim como os demais listados neste capítulo, tem como objetivo resolver cada uma das localizações mencionadas no texto para as suas respectivas coordenadas.

Os métodos de resolução de topônimo atuais, de uma forma geral, podem ser divididos entre abordagens baseadas em regras ou heurísticas, abordagens estatísticas e abordagens baseadas em aprendizado de máquina. A seguir, são descritos alguns trabalhos que utilizam cada uma dessas estratégias.

3.1 Resolução Baseada em Heurísticas

A utilização de heurísticas é bastante comum na tarefa de *geocoding* desde estudos iniciais. Pode-se citar os trabalhos de (RAUCH *et al.*, 2003) e (AMITAY *et al.*, 2004), que fazem uso da informação de população como critério de desambiguação, e também (CLOUGH, 2005), que usa como critério os *feature classes*, dando prioridade para países sobre cidades.

O trabalho de (LEIDNER, 2008) é uma das primeiras pesquisas abrangentes da área. O *geocoder* proposto pelo autor baseia-se nas heurísticas de um sentido por discurso e minimização geométrica, isto é, assume que apenas um local pode ser referido com um determinado nome no mesmo texto e que as localizações mencionadas estão próximas umas das outras. O autor compara seu *geocoder* com sua réplica do PERSEUS, um sistema apresentado por (SMITH; CRANE, 2001) que também faz uso de uma heurística de minimização das distâncias mútuas entre os topônimos.

O *Edinburgh Geoparser*, apresentado por (GROVER *et al.*, 2010), faz a resolução dos locais mencionados utilizando *feature class*, população e código de país, desambiguando candidatos através de heurísticas de distância mútua, proximidade para a localização de referência fornecida pelo usuário e relações do contexto, que são definidas por regras e detectadas na etapa de *geotagger*. O *geocoder* pode ser utilizado com diversos dicionários geográficos providos pela Universidade de Edinburgh, como uma cópia do GeoNames ou mesmo um *gazetteer* de nomes históricos de lugares no Reino Unido.

Proposto por (KARIMZADEH *et al.*, 2013) como uma API limitada e depois apri-

morado em (KARIMZADEH *et al.*, 2019), o GeoTxt baseia-se em múltiplas heurísticas e utiliza informações de nível administrativo, população, distância de Levenshtein do termo consultado para o nome do candidato e distância mútua entre os topônimos para realizar o *geocoding*. Sua obtenção de candidatos é feita através do Solr [1], uma ferramenta de consulta baseada no Lucene, com incrementos de pontuação para determinados *feature classes* e lugares dentro dos EUA. Já a sua desambiguação é realizada juntando pares de locais, consecutivos no texto ou não, e atribuindo um valor quantitativo para a relação entre as localizações (cidade-estado, cidade-país, etc.). Com o score da consulta, relações e distâncias mínimas, os candidatos com melhor pontuação são escolhidos.

O *Cartographic Location And Vicinity INdexer* (CLAVIN) [2] é um *geocoder* de código aberto que obtém candidatos através do Lucene[3], realizando um incremento de score para alguns campos e valores de forma semelhante ao GeoTxt. Entretanto, diferente do GeoTxt o *CLAVIN* realiza a desambiguação calculando um score para combinações de candidatos baseando-se na comunalidade de países e estados. Em outras palavras, quando há mais de um candidato para um local, dá-se prioridade para o candidato pertencente à região administrativa onde há mais topônimos já desambiguados.

Por fim, (ALDANA-BOBADILLA *et al.*, 2020) propõe uma abordagem que utiliza uma estratégia de desambiguação dinâmica para resolução de topônimos. O método de *geocoding* dos autores utiliza um conjunto de regras simulando o processo que um ser humano utiliza para desambiguar locais, derivando relações entre os topônimos. Em última instância, quando ainda há ambiguidade após a aplicação das regras, o candidato com o nível administrativo mais alto é escolhido.

3.2 Resolução Baseada em Estatística

Diferente das abordagens baseadas em heurísticas, em que a resolução é feita por meio de regras, as abordagens baseadas em estatística buscam resolver o problema através de modelos de distribuição. Essa estratégia é utilizada em vários trabalhos que focam na geolocalização de documentos inteiros, como em (BUTT; HUSSAIN, 2013) e (HULDEN *et al.*, 2015), mas também pode ser aplicada para topônimos individuais.

O TopoCluster, proposto por (DELOZIER *et al.*, 2015) aprimora o trabalho de

[1] https://solr.apache.org/
[2] https://github.com/Novetta/CLAVIN
[3] https://lucene.apache.org/

(BUTT; HUSSAIN, 2013) e faz o *geocoding* através de pseudo-documentos contendo as palavras do contexto do topônimo, utilizando janelas de 15 palavras para cada lado. Sua resolução funciona dividindo o planeta Terra em uma grade com células de 0.5x0.5 degrau e modela a distribuição geográfica das palavras na grade. Com sua análise de *hot spots*, o TopoCluster atribui os topônimos às células com maior sobreposição das distribuições individuais das palavras. No mesmo trabalho o autor apresenta também uma versão do TopoCluster chamada TopoClusterGaz que utiliza um dicionário geográfico híbrido de GeoNames e Natural Earth[4]. Essa versão consulta o *gazetteer* ao fim do processo e atribui ao topônimo as coordenadas do candidato mais próximo da célula predita.

3.3 Resolução Baseada em Aprendizado de Máquina

Estratégias de resolução baseada em aprendizado de máquina usam modelos treinados para predizer coordenadas geográficas para os topônimos. Dentre os métodos atuais, o uso de representações como saco-de-palavras combinado com Máquinas Suporte de Vetores ou Regressão Logística têm apresentado bons resultados (GRITTA *et al.*, 2018).

O CamCoder, proposto por (GRITTA *et al.*, 2018), divide o mundo em uma grade com células de 60 km de lado e faz uso de um vetor chamado MapVec para modelar a distribuição geográfica de menções a lugares. Utilizando GloVe para representação vetorial das palavras, o CamCoder submete à uma rede neural profunda o MapVec, a localização alvo e seu contexto de duzentas palavras em cada direção. Essa rede neural então prediz uma célula da grade correspondente à localização alvo. Com a saída da rede, o CamCoder escolhe o candidato para a localização através de um escore que considera a população do candidato e sua distância para a célula predita.

O trabalho de Chen *et al.* (2019) propõe um método para resolução de topônimos utilizando clusterização com um algoritmo semelhante ao DBSCAN. Para consulta dos candidatos de cada topônimo, utiliza os dicionários geográficos GeoNames, Nominatim[5] e GoogleV3[6]. A estratégia proposta pelos autores resolve topônimos considerando a proximidade de todos os candidatos, de forma que o candidato escolhido na desambiguação do topônimo é aquele que está no *cluster* mais populoso, ou seja, que está próximo da maior quantidade de outros candidatos para as localizações do texto.

[4] https://www.naturalearthdata.com/
[5] https://nominatim.openstreetmap.org
[6] https://developers.google.com/maps/documentation/geocoding/overview

(NIZZOLI *et al.*, 2020) adota uma estratégia diferente dos trabalhos anteriores ao utilizarem uma anotação semântica com um grafo de conhecimento para identificar localizações no texto e mapeá-las para coordenadas geográficas. O método proposto envolve percorrer os nós do grafo expandindo a lista de candidatos para os topônimos e, ao fim do processo, escolher um deles utilizando árvores de decisão. A utilização dessa abordagem necessita de um grafo de conhecimento com os locais de interesse da aplicação, que nem sempre existe.

3.4 Análise Comparativa

Quase todos os trabalhos relacionados tratam de textos em inglês. A exceção é o Adaptive Geoparsing, que trata de notícias em espanhol, porém sua estratégia pode ser aplicada em outros idiomas com as devidas adaptações. Pela maior disponibilidade de bases de dados anotadas em inglês, o presente trabalho, assim como os relacionados, trata de textos nesse idioma.

O Quadro 1 apresenta um resumo dos trabalhos relacionados, comparando-os com o que é desenvolvido neste trabalho. Suas colunas indicam detalhes sobre o processo que utilizam para resolver os topônimos, podendo haver ou não a presença de algumas características.

A maioria dos trabalhos faz uso de algum dicionário geográfico, como mostrado na coluna "*Gazetteer*". Dentre os que utilizam, quase todos fazem uso do GeoNames ou alguma modificação dele. O TopoCluster é a única estratégia que não faz uso de um dicionário geográfico, porém seu desempenho é inferior ao da versão que utiliza (GRITTA, 2019). O presente trabalho também utiliza uma versão do GeoNames como dicionário geográfico.

Quanto a utilização de *features classes*, indicada na coluna correspondente, as estratégias ficam divididas, com metade delas aproveitando essa informação de alguma forma. Edinburgh Geoparser, GeoTxt, CLAVIN e Adaptive Geoparsing utilizam esse dado para priorizar candidatos de nível administrativo mais alto no momento da desambiguação. Já o Geo-semantic-parsing utiliza essa informação como entrada para suas árvores de decisão. Este trabalho também faz uso de *feature classes* para priorizar candidatos.

A população dos locais na desambiguação é utilizada por três abordagens, como indicado na coluna correspondente. O GeoTxt e o Edinburgh Geoparser utilizam essa informação para priorizar locais mais populados. Já o CamCoder faz uso desse dado para criar o MapVec e também para calcular as pontuações para desambiguação dos candidatos. Neste trabalho a população também é utilizada para desambiguar candidatos priorizando lugares mais povoados.

Tratando do uso das distâncias mútuas entre os topônimos mapeados para coordenadas geográficas, apresentada na coluna respectiva, apenas o GeoTxt, o Edinburgh Geoparser e o DensityK aproveitam essa informação para realizar o *geocoding*. O primeiro calcula um centroide dos locais resolvidos para priorizar candidatos mais próximos e os outros dois usam essa informação para criar *clusters* e dar prioridades a candidatos em aglomerações de acordo com seus critérios. Neste trabalho a heurística de distância mútua para priorizar locais mais próximos também está presente.

A relação hierárquica entre as localizações referenciadas é utilizada para desambiguação por quatro dos trabalhos relacionados. O Edinburgh Geoparser prioriza candidatos caso haja uma relação, detectada em sua etapa de *geotagging*, indicando que um local está contido em outro. O GeoTxt e o Adaptive Geoparsing verificam se há uma relação entre as localizações durante a desambiguação. Já o CLAVIN 3.0, através de seu score, prioriza candidatos que estejam contidos no mesmo estado ou país. Este trabalho não processa os textos para extrair relações entre os topônimos e não assume que elas existam ao realizar o *geocoding*. Portanto, este trabalho não utiliza a relação hierárquica entre os topônimos como critério de desambiguação.

Em relação ao processamento de topônimos na forma adjetiva, apenas a estratégia de DeLozier *et al.* (2015) lida com essa questão. Todavia, como reportado pelo próprio autor em sua publicação, a abordagem não apresenta bons resultados para esse tipo de entidade nomeada. O presente trabalho apresenta uma estratégia para tratar desse tipo de referência a localizações, traduzindo topônimos em forma adjetival para sua forma substantiva através de um dicionário de adjetivos pátrios oficiais.

Outro diferencial deste trabalho para os relacionados é a utilização dos tipos de topônimos no processo de *geocoding*. Os trabalhos relacionados tratam todos os topônimos igualmente, sem fazer distinção entre eles na hora de desambiguar candidatos. O presente trabalho apresenta uma estratégia de *geocoding* que diferencia topônimos literais e associativos durante a desambiguação, buscando eliminar a influência de topônimos associativos na resolução das referências aos locais físicos dos quais os textos tratam.

Quadro 1 – Comparativo de abordagens para *geocoding*

Geocoder	***Gazetteer***	**Utiliza *Feature Class***	**Utiliza População**	**Utiliza Distância Mútua**	**Utiliza Relação Hierárquica**	**Formas Adjetivas**	**Tipos de Topônimo**
Edinburgh Geoparser	GeoNames Natural Earth	Sim	Sim	Sim	Sim	Não	Não
GeoTxt	GeoNames	Sim	Sim	Sim	Sim	Não	Não
CLAVIN 3.0	GeoNames	Sim	Não	Não	Sim	Não	Não
Adaptive Geoparsing	Próprio	Sim	Não	Não	Sim	Não	Não
TopoCluster	-	Não	Não	Não	Não	Sim	Não
TopoClusterGaz	Híbrido GeoNames + Natural Earth	Não	Não	Não	Não	Sim	Não
CamCoder	GeoNames	Não	Sim	Não	Não	Não	Não
DensityK	Nominatim GoogleV3 GeoNames	Não	Não	Sim	Não	Não	Não
Geo-semantic-parsing	DBpedia	Sim	Não	Não	Não	Não	Não
Este Trabalho	GeoNames	Sim	Sim	Sim	Não	Sim	Sim

Fonte: Elaborado pelo autor.
A coluna **"Formas Adjetivas"** indica se a estratégia trata de topônimos que aparecem dessa maneira no texto. A coluna **"Tipos de Topônimo"** indica se a estratégia diferencia os topônimos por tipo durante o *geocoding*. As colunas preenchidas com **"Sim"** ou **"Não"** indicam se o *geocoder* apresenta a característica.

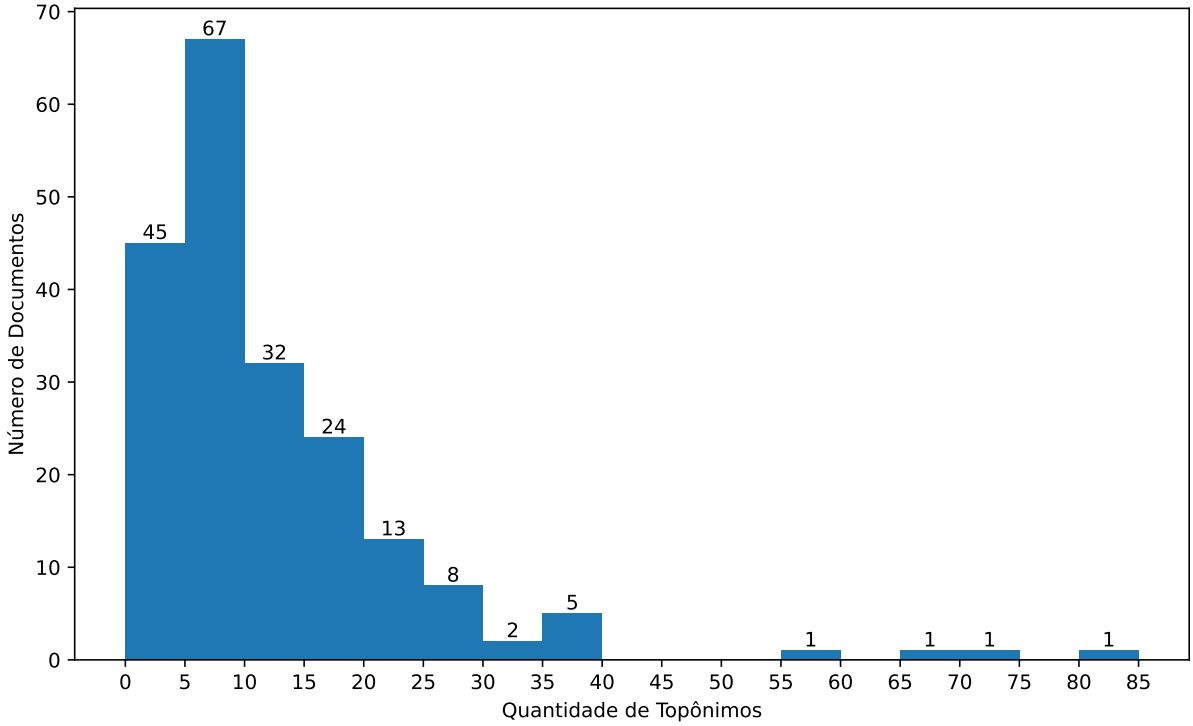
4 DADOS E MÉTODOS

Este capitulo apresenta os dados utilizados neste trabalho, as métricas de avaliação e a descrição dos experimentos. A metodologia descrita neste capítulo foi parcialmente publicada em Sá *et al.* (2022).

4.1 Base de Dados

Este trabalho faz uso do GeoWebNews (GRITTA *et al.*, 2019), uma base de dados para avaliação de *geoparsers*, de forma a padronizar a comparação entre diferentes abordagens de *geoparsing*. Ela é composta por uma coleção de 200 notícias em inglês, contendo o título e o corpo, coletadas de fontes de diversos países ao redor do mundo durante os primeiros oito dias de abril de 2018. A Figura 3 apresenta um histograma com a distribuição dos topônimos por documento, onde o eixo x indica intervalos de contagem de topônimos e o eixo y indica o número de notícias que apresenta aquela quantidade de topônimos.

Figura 3 – Distribuição de topônimos por documento

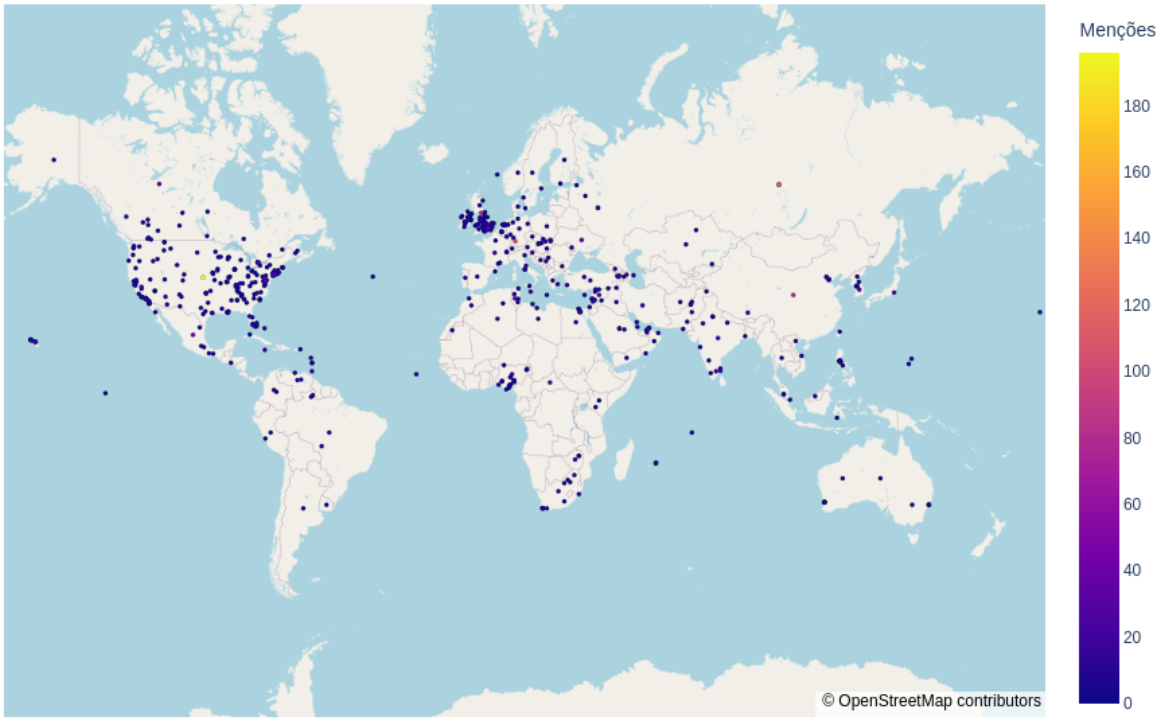


Fonte: Elaborado pelo autor

Cada uma das entradas da base de dados possui anotações dos topônimos e seu tipo, assim como a anotação da latitude e longitude, incluindo o identificador do GeoNames caso exista. A Figura 4 traz um mapa com a distribuição geográfica dos locais mencionados nos textos,

com a cor indicando a quantidade de vezes que são referenciados, para melhor visualização.

Figura 4 – Distribuição dos locais mencionados nos textos do GeoWebNews pelo mundo



Fonte: Elaborado pelo autor

Os topônimos do GeoWebNews são classificados em concordância com a taxonomia de (GRITTA *et al.*, 2019) e, portanto, seguem a mesma utilizada por este trabalho, que foi descrita anteriormente. A Tabela 1 traz as categorias e tipos de cada classe utilizada na base de dados.

Uma vez que o objetivo é atribuir coordenadas geográficas às entidades nomeadas, são consideradas para experimentação apenas aquelas anotadas como topônimos, sejam eles literais ou associativos, resultando em 2401 localizações distribuídas entre as 200 notícias da base de dados. A Figura 5 traz um resumo da distribuição dos tipos de entidade nos textos de notícias da base de dados.

4.2 Processo de *Geoparsing*

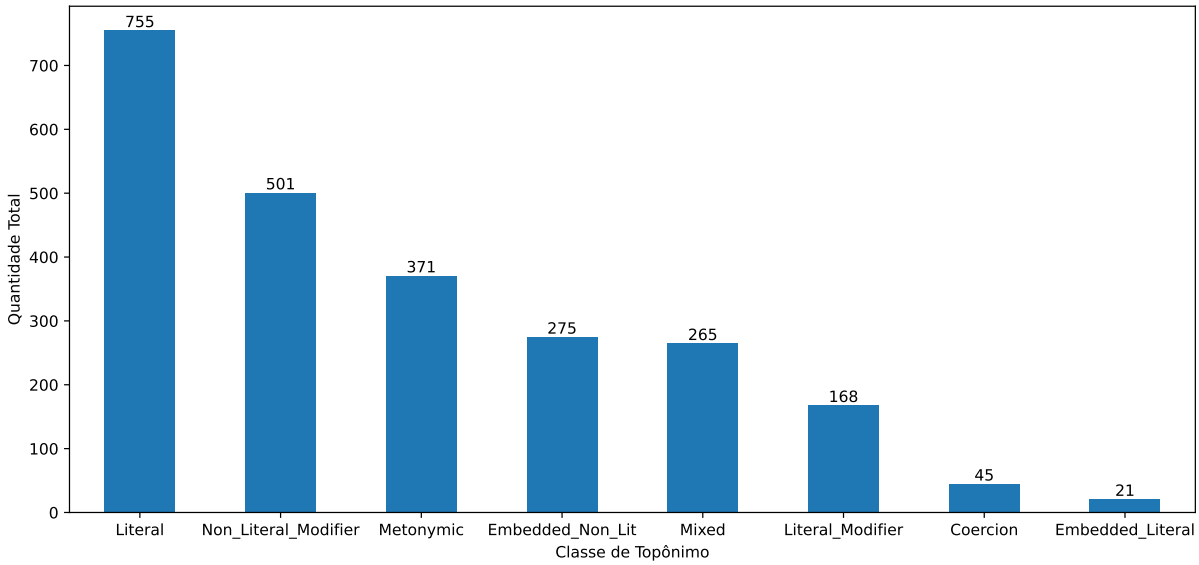
O processo de *geoparsing* dos textos é realizado em duas etapas sequenciais. Inicialmente, é feito o reconhecimento dos topônimos do texto, realizando o *geotagging*, e em seguida os topônimos encontrados são resolvidos para localizações geográficas, realizando o *geocoding*.

Tabela 1 – Categorias e tipos das classes utilizadas pelo GeoWebNews

Classe	Categoria	Tipo
Literal	Literal	Literal
Coercion	Literal	Coerção
Mixed	Literal	Misto
Embedded_Literal	Literal	Literal Embutido
Literal_Modifier	Literal	Modificador Substantivo Modificador Adjetivo
Demonym	Associativo	Gentílico
Language	Associativo	Língua
Metonymic	Associativo	Metonímia
Non_Literal_Modifier	Associativo	Modificador Substantivo Modificador Adjetivo
Embedded_Non_Lit	Associativo	Associativo Embutido
Homonym	Associativo	Homônimo

Fonte: Elaborado pelo autor.

Figura 5 – Contagens de tipos de topônimos no GeoWebNews



Fonte: Elaborado pelo autor

As experimentações tem como foco a etapa de resolução dos topônimos.

4.2.1 Reconhecimento de Topônimos

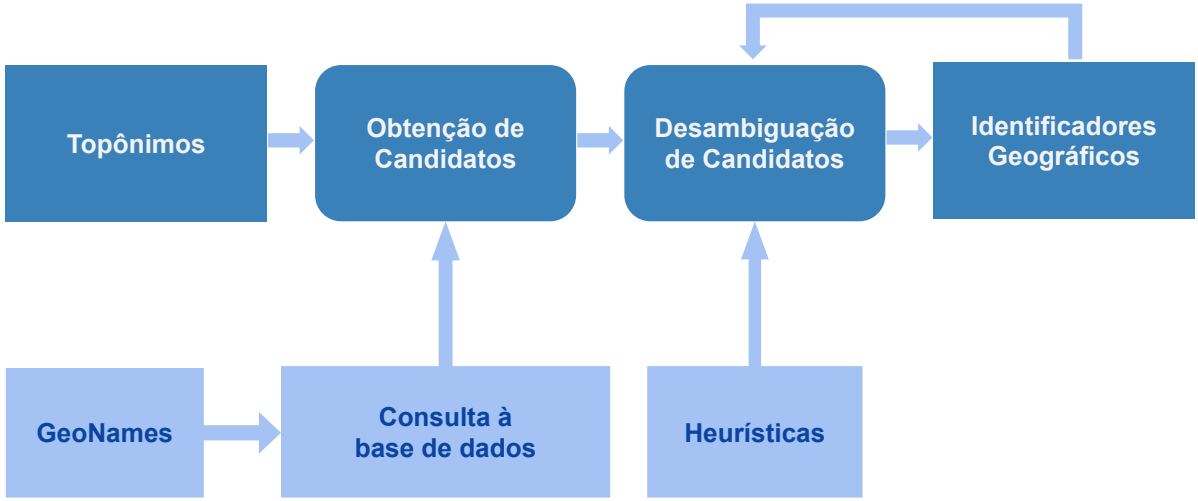
Caso seja utilizada, a etapa de *geotagging* é realizada através da ferramenta *off-the-shelf* spaCy (HONNIBAL *et al.*, 2020), que possui um módulo destinado para NER e por isso pode ser utilizado para encontrar os topônimos dos textos. Já quando a etapa de *geotagging* não é utilizada, o *geocoder* recebe como entrada todos os topônimos anotados, assumindo um *geotagger* fictício com reconhecimento perfeito chamado de Oracle. Uma vez que a etapa de *geotagging* não é o foco deste trabalho, detalhes adicionais, experimentos para essa tarefa e seus

resultados encontram-se no Apêndice A.

4.2.2 Resolução de Topônimos

Esta é a etapa foco das experimentações realizadas neste trabalho e consiste na atribuição de latitudes e longitudes aos topônimos de um texto. Isso é realizado obtendo-se candidatos para os topônimos e subsequentemente selecionando-se heuristicamente os mais adequados em um processo de desambiguação. A Figura 6 traz um diagrama ilustrando o processo de resolução dos topônimos.

Figura 6 – Processo de Resolução de Topônimos



Fonte: Elaborado pelo autor

A obtenção de possíveis locais para um topônimo é realizada por meio de consultas ao dicionário geográfico, relacionando nomes de locais aos seus respectivos identificadores geográficos. Para estas consultas é utilizada a ferramenta *off-the-shelf* ElasticSearch[1], uma interface para o *Lucene* (DIVYA; GOYAL, 2013), pois, como demostrado por (CLEMENS, 2015), seu ranqueamento baseado em pontuações dinâmicas, que depende da consulta, retorna resultados melhores do que os de ferramentas que utilizam pontuações estáticas.

O índice criado no ElasticSearch é populado com os dados do GeoNames. Para permitir consultas mais variadas, campos textuais do dicionário como Nome e Nomes alternativos são inseridos no índice como *text* e *keyword*, possibilitando que consultas sejam realizadas por fragmentos ou por valores completos. Os campos do tipo *text* são tratados pelo analisador padrão do ElasticSearch[2] tanto no momento de inserção no índice quanto durante a consulta. Já para os

[1] https://www.elastic.co/
[2] https://www.elastic.co/guide/en/elasticsearch/reference/current/analysis-standard-analyzer.html

campos do tipo *keyword*, os valores cadastrados e termos das consultas são apenas transformados para caixa baixa.

As consultas com ElasticSearch neste trabalho utilizam Okapi BM25, uma função de similaridade proposta por Robertson *et al.* (1995) que utiliza *Term Frequency*/Frequência do Termo (TF) e *Inverse Document Frequency*/Inverso da Frequência em Documentos (IDF) e é implementada por padrão no ElasticSearch[3]. O topônimo é consultado em todos os campos de nome no índice para obter a pontuação dos locais do dicionário geográfico[4,5].

A Equação 4.1 traz a fórmula utilizada para calcular o score de um local no índice ao consultar um topônimo (e.g., "United States of America") em um campo de nome (e.g., Nome ASCII). O score é obtido avaliando o valor D do campo (e.g., "United States") para uma consulta $Q =< q_1, q_2, \ldots, q_n >$, onde cada q_i é um termo da consulta. Na fórmula, $f(q_i, D)$ indica o número de vezes que q_i ocorre em D, $\|D\|$ é o tamanho de D em palavras e $avgdl$ é tamanho médio dos valores do campo no índice. Já $IDF(q_i)$ é o peso do termo q_i da consulta, calculado através da fórmula da Equação 4.2, onde N é a quantidade total de valores do campo no índice e $n(q_i)$ é o número de locais cujo campo contém o termo q_i.

$$score(D,Q) = \sum_{i=1}^{n} IDF(q_i) * \frac{f(q_i,D) * (1.2+1)}{f(q_i,D) + 1.2 * \left(1 - 0.75 + 0.75 * \frac{\|D\|}{avgdl}\right)} \tag{4.1}$$

$$IDF(q_i) = ln\left(\frac{N - n(q_i) + 0.5}{n(q_i) + 0.5} + 1\right) \tag{4.2}$$

A escolha dos candidatos mais adequados após a obtenção é realizada com base em heurísticas. Em casos em que há apenas um candidato para o topônimo consultado, ele é o escolhido. Já nos casos em que a consulta retorna mais de uma possível localização para o nome buscado, são aplicadas as heurísticas habilitadas como critério para a desambiguação. Em casos em que nenhum candidato é retornado durante a busca, o topônimo é deixado como não resolvido, pois não é possível atribuir uma coordenada geográfica para ele.

[3] https://www.elastic.co/blog/practical-bm25-part-2-the-bm25-algorithm-and-its-variables

[4] https://www.elastic.co/guide/en/elasticsearch/reference/current/query-dsl-multi-match-query.html#type-best-fields

[5] https://www.elastic.co/guide/en/elasticsearch/reference/current/query-dsl-bool-query.html

4.3 Métricas de Avaliação

Para a etapa de *geocoding* não há um conjunto de métricas padronizado, por isso as métricas escolhidas são aquelas recomendadas por (GRITTA *et al.*, 2019): Erro médio, Acurácia@161km e *Area Under the Curve*/Área Sob a Curva (AUC). As métricas são calculadas para cada documento e, por fim, são obtidas a média e o desvio padrão de cada uma delas.

O erro médio é uma medida da distância do locais preditos para os locais esperados em um documento. Com a informação das coordenadas geográficas reais e as preditas pelo *geocoder*, a distância ortodrômica[6] é calculada entre os dois pontos na superfície da Terra. Com todas as distâncias calculadas, o erro médio é definido como uma simples média de todas elas. Considerando um texto que contém dois topônimos e eles sejam resolvidos para localizações a 5 km e 15 km das coordenadas esperadas, por exemplo, o erro médio é de 10 km.

A acurácia até uma certa distância, denotada como Acurácia@Xkm, onde X é a distância máxima, é uma medida que indica a porcentagem de localizações resolvidas para até X quilômetros das coordenadas esperadas. Em outras palavras, essa métrica aponta a proporção de topônimos do documento resolvidos corretamente, isto é, com um erro menor do que o limiar definido. Essa métrica é utilizada em diversos trabalhos com diferentes valores para a distância limite. O valor utilizado neste trabalho é de 161 km, previamente utilizado em outros trabalhos como (DELOZIER *et al.*, 2015; GRITTA *et al.*, 2019; WANG; HU, 2019). A razão para essa escolha é o possível desalinhamento entre coordenadas anotadas para um lugar na base de dados e as coordenadas da mesma localização no dicionário geográfico utilizado. Se uma base de dados anota as coordenadas de uma cidade em um certo ponto, mas o registro correspondente a ela no *gazetteer* utilizado está a 20km desse ponto, por exemplo, a predição do *geocoding* ainda será considerada correta, pois está dentro da tolerância de 161km.

A terceira e última métrica utilizada para a etapa de *geocoding* é a AUC, uma métrica recente que quantifica o desvio entre as localizações preditas e as esperadas (WANG; HU, 2019). Não deve ser confundida com a AUC da curva ROC e outras medidas homônimas. O valor da área sob a curva de erros de *geocoding* é calculado utilizando a Regra do Trapezoide[7] para integração e depois normalizado para o intervalo $[0, 1]$. A Equação 4.3 traz a fórmula utilizada para o cálculo da métrica para N topônimos. Cada x_i denota a distância entre as coordenadas escolhidas pelo *geocoder* para o i-ésimo topônimo e as coordenadas esperadas. O valor 20039 é

[6] https://geopy.readthedocs.io/en/stable/#geopy.distance.great_circle
[7] https://docs.scipy.org/doc/numpy/reference/generated/numpy.trapz.html

o maior erro de *geocoding* possível, quando a distância entre as coordenadas é aproximadamente metade da circunferência da Terra em quilômetros (20038 Km). Isso ocorre quando a localização predita está diametralmente oposta à esperada na superfície do planeta.

$$AUC = \frac{\sum_{i=1}^{N-1} ln(x_{i-1}+1) + ln(x_i+1)}{2*(N-1)*ln(20039)} \tag{4.3}$$

A AUC é uma medida para lidar com o problema de *outliers*, que quando aparecem tendem a distorcer o erro médio, ajudando a avaliar a maioria dos erros que sem ela seriam suprimidos por esses valores atípicos. Um valor considerado bom para essa métrica é aquele que se aproxima ao máximo de 0, indicando poucos erros. Quanto maior o valor da AUC, pior é considerada a resolução dos topônimos.

Para entender melhor essa última métrica, pode-se considerar dois *geocoders* resolvendo os mesmos três topônimos. O primeiro apresenta erros, em quilômetros, de 10, 100 e 1000, tendo portanto um erro médio de 370 km e uma AUC de aproximadamente 0.468. Já o segundo apresenta erros de 200, 400 e 510, tendo um erro médio também de 370 km, mas uma AUC de aproximadamente 0.594. O melhor *geocoder* nesse caso é o primeiro, que acertou duas localizações com uma tolerância de 161 km, enquanto o segundo não acertou nenhuma, embora apresentem o mesmo erro médio. A superioridade do primeiro *geocoder* é refletida pelo valor da AUC, que é mais próximo de 0.

4.4 Geocoder Heurístico Base

O *geocoder* heurístico base divide a tarefa de *geocoding* em duas etapas onde heurísticas diferentes podem ser utilizadas. O primeiro passo é a obtenção de candidatos consultando o dicionário geográfico. Já o segundo passo consiste na desambiguação dos candidatos obtidos. Ao fim do processo, o *geocoder* escolhe um registro do *gazetteer* para o topônimo da entrada.

O *geocoder* recebe uma lista de topônimos como entrada e tem como saída coordenadas de acordo com os seguintes parâmetros:

- **Obtenção de Candidatos**:
  - **Correspondência Exata**: indica se devem ser considerados apenas candidatos com uma correspondência exata ao termo consultado;
  - **Score Mínimo**: indica o menor score para que um resultado da consulta seja considerado um candidato para desambiguação;

  - **Ordenação**: indica se a ordenação dos candidatos deve ser por score, *feature class* ou população;
- **Desambiguação de Candidatos**:
  - **Distância Mútua**: indica se deve ser utilizada a distância mútua como critério de desambiguação de candidatos no *geocoding*;
  - **Top-K**: indica quantos candidatos do topo do *ranking* após ordenação devem ser considerados para a desambiguação;

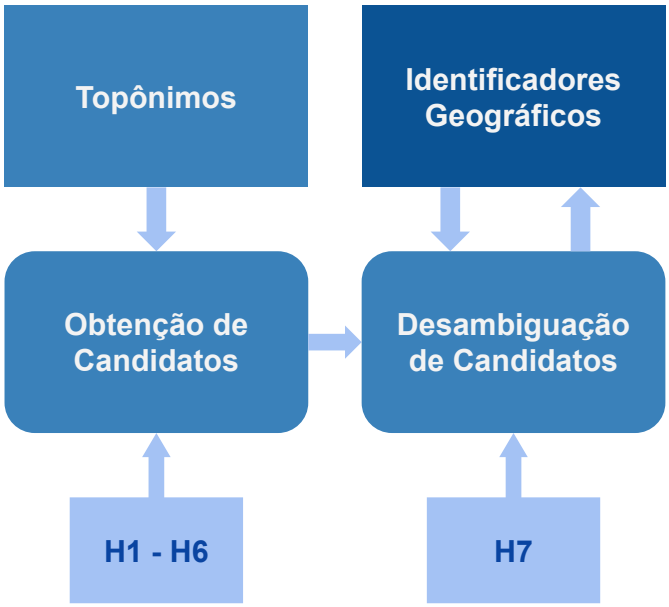
Dessa forma, o processo de *geocoding* é realizado através das seguintes heurísticas:

- **Correspondência Exata (H1)**: considera uma localização como um candidato apenas quando um de seus nomes é exatamente igual ao termo consultado. Isso significa que a entrada para os Estados Unidos é considera ao consultar "United States" ou "USA", que são nomes alternativos no *gazetteer*, mas não "States of America";
- **Correspondência Relaxada (H2)**: considera uma localização como um candidato se há uma correspondência parcial entre um de seus nomes e o texto consultado. Isso significa que consultar "States of America" retornará a entrada para os Estados Unidos como um candidato;
- **Pontuação Mínima de Candidato (H3)**: descarta resultados da consulta com um score inferior ao limiar, considerando como candidatos apenas aqueles com uma pontuação igual ao superior ao valor mínimo fornecido;
- **Ordenar Candidatos por Score (H4)**: ordena candidatos com base na pontuação padrão do ElasticSearch. O score depende do nome do lugar e do texto consultado;
- **Ordenar Candidatos por *Feature Class* (H5)**: ordena candidatos com base em seu *feature class* no GeoNames. Isso significa que o país Angola aparece primeiro na lista de candidatos que a cidade de Angola, Indiana;
- **Ordenar por População (H6)**: ordena candidatos com base em sua população. Nesse caso, a entrada para a República da Coreia aparece antes da correspondente à República Popular Democrática da Coreia na lista de candidatos;
- **Minimização Geométrica (H7)**: minimiza a distância média entre todos os topônimos resolvidos. Isso é feito escolhendo candidatos com base em sua distância para os topônimos resolvidos anteriormente, assumindo que locais mencionados em um texto estão o mais próximo possível uns dos outros.

A Figura 7 traz um diagrama com as heurísticas mencionadas e sua localização no

processo de *geocoding*. As heurísticas de H1 a H6 pertencem ao passo de obtenção de candidatos. Já a heurística H7 está relacionada ao passo de desambiguação dos candidatos após a consulta.

Figura 7 – Heurísticas no processo de resolução de topônimos



Fonte: Elaborado pelo autor

A Tabela 2 traz os valores possíveis de cada um dos parâmetros do *geocoder*. Quando não especificados, os valores utilizados para os parâmetros são aqueles destacados em negrito. Para definição da configuração utilizada como *baseline* são realizados experimentos avaliando o desempenho das combinações de parâmetros, variando-os individualmente e em conjunto. A combinação com melhor desempenho é utilizada para comparação com outros *geocoders*.

Tabela 2 – Resumo dos Parâmetros Avaliados na Escolha do *Geoparser* Base

Parâmetro	**Valores**
Correspondência Exata	Sim, **Não**
Score Mínimo	**0**, 10, 15, 20
Ordenação	**Score**, Feature Class, População
Distância Mútua	Sim, **Não**
Top-K	**5**, 10, 20

Fonte: Elaborado pelo autor.
Valores possíveis para cada um dos parâmetros do *geoparser*.
Em negrito os valores padrão para os parâmetros.

4.4.1 Experimentos de Obtenção de Candidatos

Para a etapa de obtenção de candidatos, os experimentos avaliam os parâmetros de correspondência exata, score mínimo e ordenação dos candidatos. Cada um desses parâmetros é variado individualmente, enquanto os demais permanecem com o valor padrão.

4.4.1.1 Correspondência Exata

Esse experimento avalia o impacto da rigidez da pesquisa de candidatos no desempenho do *geoparser*. No primeiro caso é utilizada uma pesquisa menos rígida, acrescentando à lista de candidatos aqueles que em seu nome possuem algum elemento do texto utilizado para consulta. Já no segundo caso, são acrescentados à lista apenas locais que possuem em seu nome ou em nomes alternativos o texto exato utilizado na consulta.

4.4.1.2 Score Mínimo

Esse experimento avalia a influência de um filtro de score mínimo na consulta de candidatos. Para o primeiro caso não é utilizado nenhum filtro de score mínimo. Já nos dois casos seguintes, são utilizados como limiar os valores 10 e 15, indicando que entradas com um score menor do que esses valores não devem ser retornadas na consulta. É importante ressaltar que este score é calculado pelo ElasticSearch com base no texto consultado.

4.4.1.3 Ordenação

Esse experimento avalia a influência do atributo utilizado para ordenação dos candidatos na etapa de *geocoding*. O experimento verifica três regras para a ordenação: maior score do ElasticSearch, menor valor de *feature class* e maior população.

4.4.2 Desambiguação de Candidatos

Para a etapa de desambiguação dos candidatos, o experimento avalia o impacto da utilização da distância mútua entre os candidatos e os topônimos previamente resolvidos na etapa de *geocoding*, assim como a influência da quantidade máxima de candidatos usados na desambiguação. No caso em que não é utilizada a distância mútua, o primeiro candidato é sempre o escolhido. Já para os casos em que a distância é utilizada, o primeiro topônimo é

resolvido escolhendo o primeiro candidato, porém os topônimos que se seguem são resolvidos escolhendo o candidato cujo o somatório das distâncias para os locais previamente resolvidos é mínimo. São avaliados os desempenhos dessa estratégia de desambiguação usando os Top-5, Top-10 e Top-20 candidatos para cada topônimo.

4.4.3 Combinações de Parâmetros

Ao final dos experimentos avaliando os parâmetros individualmente, esse último experimento compara os resultados das combinações dos parâmetros no intuito de escolher como configuração do *geocoder* base, a combinação de heurísticas com o melhor desempenho, ou seja, a configuração em que a maior parte das métricas avaliadas atinge o melhor resultado. Os valores utilizados nessa experimentação são escolhidos baseado nos resultados dos experimentos anteriores, descartando configurações que apresentam uma piora no resultado. O *geocoder* base definido por meio desse experimento é nomeado daqui em diante como **HG**.

4.5 Estratégias de Aprimoramento

Depois de definido o *baseline*, as estratégias de aprimoramento podem ser avaliadas e comparadas a ele. Nesta seção são descritas as estratégias propostas.

4.5.1 Processamento de Topônimos Adjetivos

A língua inglesa, em que estão escritos os textos tratados neste trabalho, possui palavras que modificam substantivos em termos de país de origem, continente ou mesmo cidade (ROBERTS, 2017). Alguns exemplos incluem: *Australian*, *European*, *New Yorker*, etc. Esses termos são conhecidos em português como adjetivos pátrios.

Ao realizar o processo de *geocoding* esses termos podem ser problemáticos, pois essas formas adjetivas não se encontram listadas entre os nomes oficiais ou nomes alternativos dos locais. Ao consultar um termo como "*Dutch*", por exemplo, a Holanda não é retornada na consulta, o que dificulta o mapeamento do topônimo para a coordenada geográfica correta.

Dessa forma, visando melhorar o desempenho quando se trata de topônimos dessa natureza, propõe-se realizar uma normalização de topônimos antes de realizar a consulta por candidatos. Isso significa a utilização de uma nova heurística:

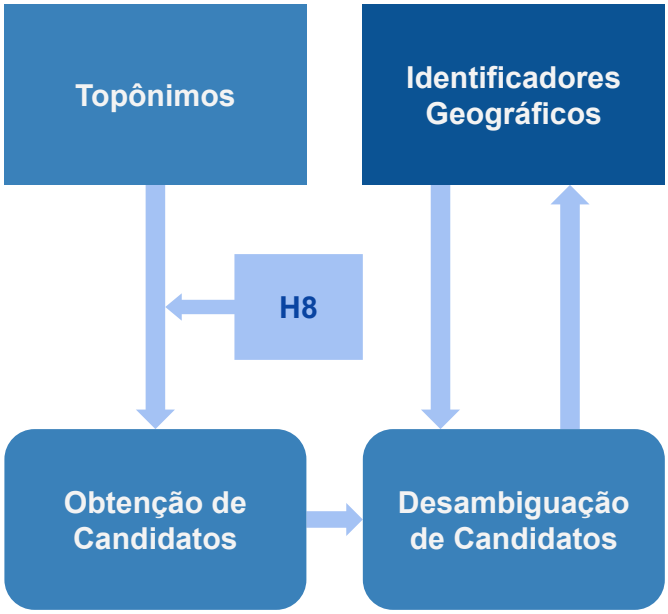
- **Normalização de Topônimos Adjetivos (H8)**: normaliza topônimos adjetivos para a sua

forma substantiva no início do *geocoding*. Nesse caso, em vez de consultar "*Dutch*", o *geocoder* obterá candidatos para "*Kingdom of the Netherlands*".

Para realizar essa normalização, um índice do ElasticSearch é criado para ser utilizado como dicionário. Esse índice então é populado com uma lista de nomes de países, conforme aparecem no GeoNames, e suas formas adjetivas e gentílicas. Por exemplo, a entrada correspondente ao Reino da Dinamarca possui o nome "*Kingdom of Denmark*", a forma adjetiva "*Danish*", que descreve algo originário do país, e o gentílico "*Danes*", que descreve o seu povo. Os adjetivos pátrios utilizados para preencher o índice são extraídos da lista de nacionalidades da Wikipédia[8].

Assim, a estratégia para processar topônimos adjetivos consiste em adicionar um passo antes da obtenção dos candidatos. Os topônimos são consultados no índica do ElasticSearch do dicionário e substituídos por uma versão normalizada. Isso significa que o topônimo "*Danish*" é normalizado para "*Kingdom of Denmark*" antes de consultar o dicionário geográfico por candidatos. O *geocoder* aprimorado com essa estratégia é nomeado daqui em diante como **HG-A**. A Figura 8 ilustra a localização de H8 no processo de resolução de topônimos.

Figura 8 – Heurística de normalização de topônimos adjetivos no processo de *geocoding*



Fonte: Elaborado pelo autor

[8] https://en.wikipedia.org/wiki/List_of_adjectival_and_demonymic_forms_for_countries_and_nations

4.5.2 Otimização Geométrica por Tipo de Topônimo

Uma das heurísticas avaliadas na escolha do *baseline* consiste em escolher o candidato com a menor distância para os locais já resolvidos no texto. Essa heurística baseia-se na ideia de que locais mencionados em um texto concentram-se em uma mesma região geográfica. Entretanto, alguns topônimos podem ser enganosos na hora de descobrir o foco de um texto.

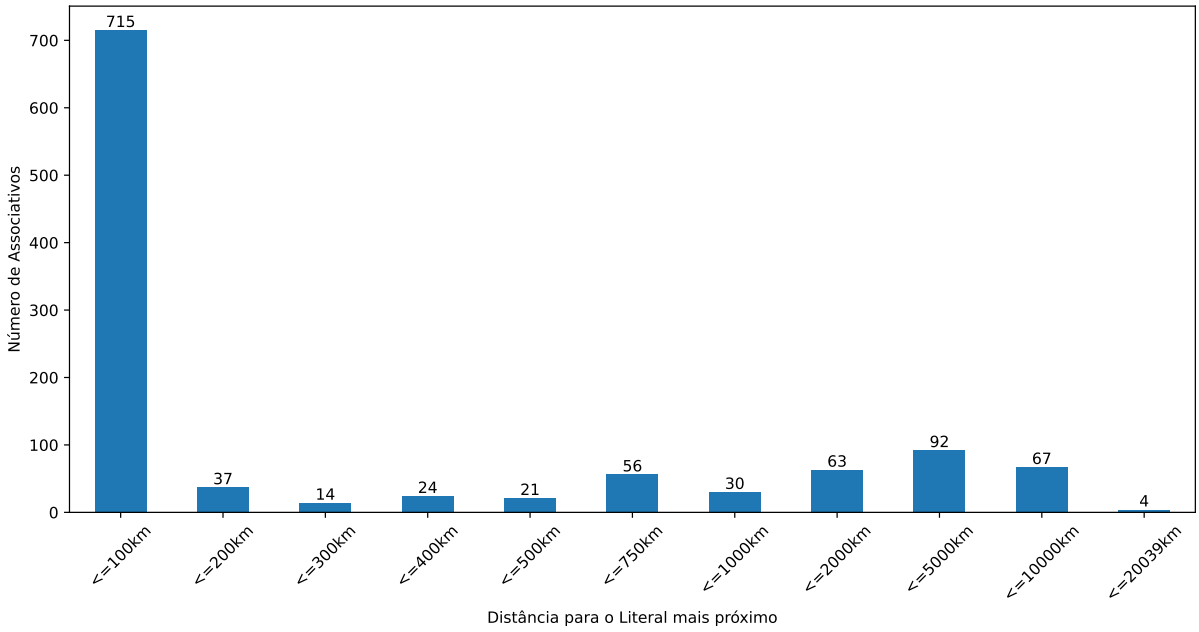
Considerando-se uma notícia informando que o presidente dos Estados Unidos visitou a embaixada americana no México, por exemplo, compreende-se que seu foco é no país latino. Porém, ao utilizar a heurística de distância, o *geocoder* poderia entender erroneamente que o texto fala apenas sobre lugares nos EUA e predizer a cidade de México no Missouri. Tal erro se deveria ao maior foco no topônimo modificador "Estados Unidos" e no topônimo de coerção "americana", não ocorrendo caso o topônimo literal "México" fosse priorizado. Por outro lado, se o *geocoder* resolver "México" e depois tentar resolver "americana" considerando apenas a distância para as coordenadas do país, esse topônimo pode ser resolvido para Cerro La Americana em Chihuahua, México.

A Figura 9 apresenta um histograma com as quantidades de topônimos associativos para várias faixas de distância para o literal mais próximo. Pode-se observar que aproximadamente 63.67% dos associativos estão bem próximos de um literal. Outros 16.21% estão até 1000 km do literal mais próximo. Todavia, os 20.12% restantes dos topônimos associativos estão a uma grande distância dos demais topônimos, prejudicando a utilização de uma heurística de minimização de distâncias como H7.

Baseando-se na hipótese de que alguns tipos de topônimo são mais importantes para a definição da região geográfica em foco, pode-se utilizar uma estratégia que consiste em considerar as distâncias de forma condicional, dependendo do tipo do topônimo sendo resolvido e dos que já foram associados às suas coordenadas. Essa estratégia parte da ideia de que topônimos da categoria Literal estão mais relacionados ao foco do texto do que os da categoria Associativo. Além disso, para evitar que um local seja escolhido apenas por estar próximo quando há outro candidato mais relevante a uma distância maior, pode-se utilizar também a população dos locais. Isso resulta em uma nova heurística:

- **Otimização Geométrica de Topônimos por Tipo (H9)**: escolhe o candidato utilizando um critério diferente dependendo do tipo do topônimo. Para tipos da categoria Associativo, o primeiro candidato é escolhido. Para tipos da categoria Literal, a escolha é feita maximizando a razão entre a população do candidato e a média de suas distâncias para os

Figura 9 – Distribuição de *outliers* associativos, topônimos não literais que se referem a localizações a X km do topônimo literal mais próximo no documento

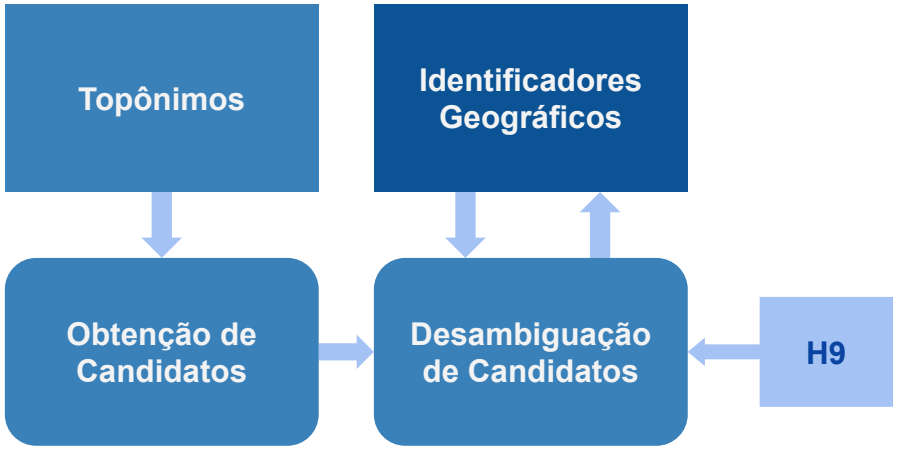


Fonte: Elaborado pelo autor

outros literais já resolvidos.

O *geocoder* base HG é então estendido com essa heurística, alterando o seu processo de desambiguação de topônimos. O *geocoder* resultante dessa alteração de HG é nomeado daqui em diante como **HG-T**. A Figura 10 traz um diagrama ilustrando a localização de H9 no processo de *geocoding*.

Figura 10 – Heurística de otimização geométrica de topônimos por tipo no processo de *geocoding*

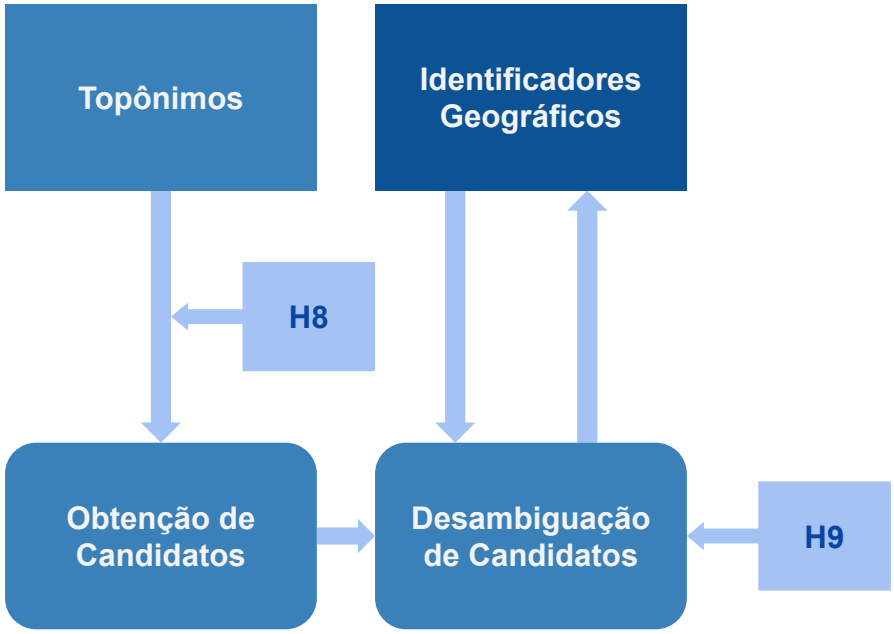


Fonte: Elaborado pelo autor

4.6 Combinação das Estratégias de Aprimoramento

Além de HG-A e HG-T, criados utilizando as heurísticas H8 e H9, respectivamente, propõe-se também a criação de outro geocoder denominado **HG-AT**. Este geocoder é resultante da combinação de ambas as heurísticas, dessa forma encorporando em seu processo de resolução tanto a normalização de topônimos adjetivos quanto a utilização das distâncias dos candidatos para as localizações já resolvidas de acordo com o seu tipo. A Figura 11 traz um diagrama ilustrando HG-AT como a combinação de HG-A e HG-T.

Figura 11 – Combinação das estratégias de aprimoramento no processo de *geocoding*



Fonte: Elaborado pelo autor

4.7 Comparação com Estado da Arte

Finalizadas as experimentações das estratégias de aprimoramento, aquelas com melhor desempenho são comparadas com outras abordagens do estado da arte. Todavia, nem todos os *geoparsers* são capazes de realizar o *geocoding* separadamente. Para realizar a comparação nesses casos, o método utilizado consiste em processar os textos realizando o *geoparsing* completo e depois submeter os topônimos reconhecidos aos *geocoders* propostos neste trabalho. Assim, apenas o *geocoding* é comparado, evitando possíveis diferenças decorrentes de estratégias de *geotagging* diferentes.

Os *geoparsers* utilizados na comparação são: CLAVIN e CamCoder. A comparação

é feita utilizando os arquivos de ground-truth[9] providos pelo EUPEG (WANG; HU, 2019). Os testes são executados no GeoWebNews e no TR-News, uma base de dados proposta por Kamalloo e Rafiei (2018) contendo 118 notícias anotadas por humanos de fontes locais e globais.

A escolha dessas duas bases de dados deve-se a cobertura que apresentam de topônimos adjetivos[10], permitindo uma melhor comparação dos *geocoders* HG-A e HG-AT com o estado da arte. Outras bases de dados, como Geovirus (GRITTA *et al.*, 2018), apresentam anotações incompletas ou nenhuma de topônimos adjetivos. Apesar do LGL (LIEBERMAN *et al.*, 2010) também apresentar esse tipo de topônimo, ele não é utilizado, uma vez que suas localizações são altamente específicas de regiões, tornando-as muito difíceis de serem desambiguadas sem informações adicionais.

Para o caso do CLAVIN, como o *geoparser* não permite a execução do *geocoder* separadamente, o método descrito anteriormente é utilizado. A realização dos testes é feita utilizando a sua versão REST[11].

No caso do CamCoder, o *geoparser* permite a execução do *geocoder* separadamente desde que um arquivo formatado de *ground-truth* seja provido. Assim, os arquivos para o GeoWebNews e TR-News, providos pelo EUPEG, são utilizados para fazer a resolução dos topônimos anotados conforme aparecem nos textos. Os testes são realizados utilizando o código provido no repositório do CamCoder no Github[12]. Adicionalmente, o banco de dados do CamCoder é atualizado antes da comparação com a mesma versão dos registros do GeoNames utilizados para popular o índice do ElasticSearch em que os *geocoders* propostos nesse trabalho operam.

[9] https://github.com/geoai-lab/EUPEG/
[10] Aproximadamente 14.9% dos topônimos do GeoWebNews e 10.7% do TR-News estão em forma adjetiva.
[11] https://hub.docker.com/r/novetta/clavin-rest
[12] https://github.com/milangritta/Geocoding-with-Map-Vector

5 RESULTADOS EXPERIMENTAIS

Neste capítulo, são apresentados os resultados obtidos a partir da execução das heurísticas explicadas no capítulo anterior. Os resultados apresentados neste capítulo foram reportados parcialmente em Sá *et al.* (2022), mas alguns valores divergem, pois, após a publicação, o dicionário geográfico foi atualizado e os experimentos foram refeitos. Vale ressaltar que as métricas descritas anteriormente foram calculadas para cada documento e, por fim, foram obtidas a média e o desvio padrão de cada uma delas.

5.1 Geocoder Heurístico Base

Os experimentos para a escolha do *geocoder* base avaliaram parâmetros em duas etapas do *geocoding*: obtenção de candidatos e desambiguação de candidatos. A seguir, são apresentados os resultados obtidos em cada etapa.

5.1.1 Obtenção de Candidatos

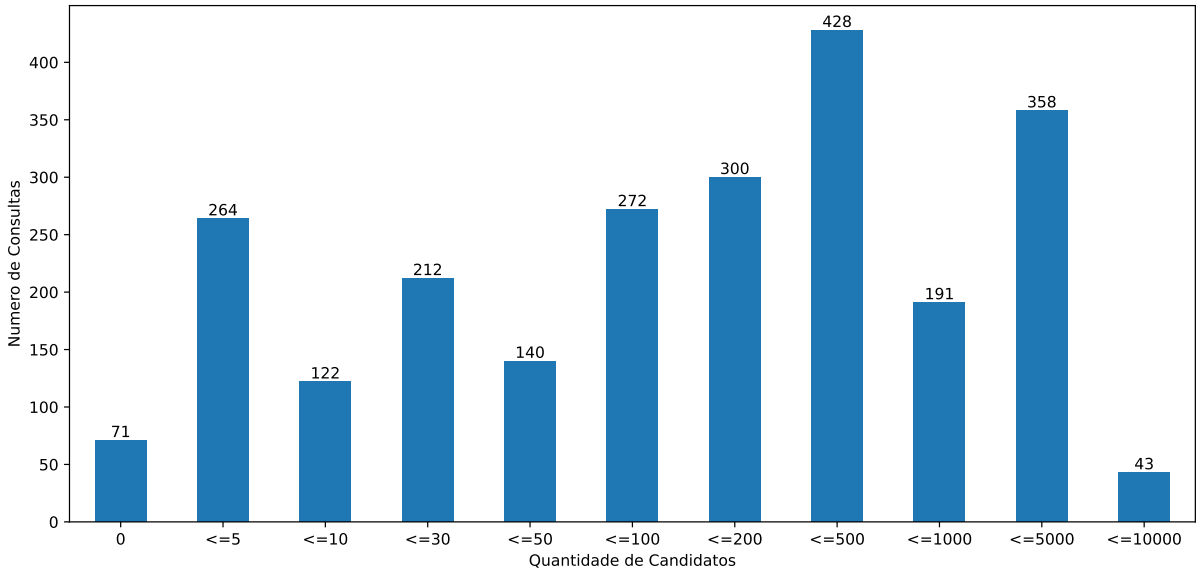
Para a etapa de obtenção de candidatos foram realizados experimentos para avaliar os parâmetros de score mínimo, correspondência exata e ordenação dos candidatos.

5.1.1.1 Correspondência Exata

Esse experimento avaliou o impacto do nível de rigidez da consulta. As Figuras 12 e 13 trazem as quantidades de candidatos retornadas para as consultas utilizando correspondência relaxada e exata, respectivamente. Pode-se observar que consultas utilizando correspondência relaxada trazem uma grande quantidade de candidatos, com 401 consultas retornando mais de 1000 candidatos. Já com a correspondência exata, as consultas retornaram menos candidatos no geral, com apenas 2 casos em que a consulta retornou mais de 1000 candidatos. Por outro lado, utilizar a correspondência exata, apesar de diminuir a ambiguidade, também aumentou o número de casos em que a quantidade de candidatos retornados foi 0.

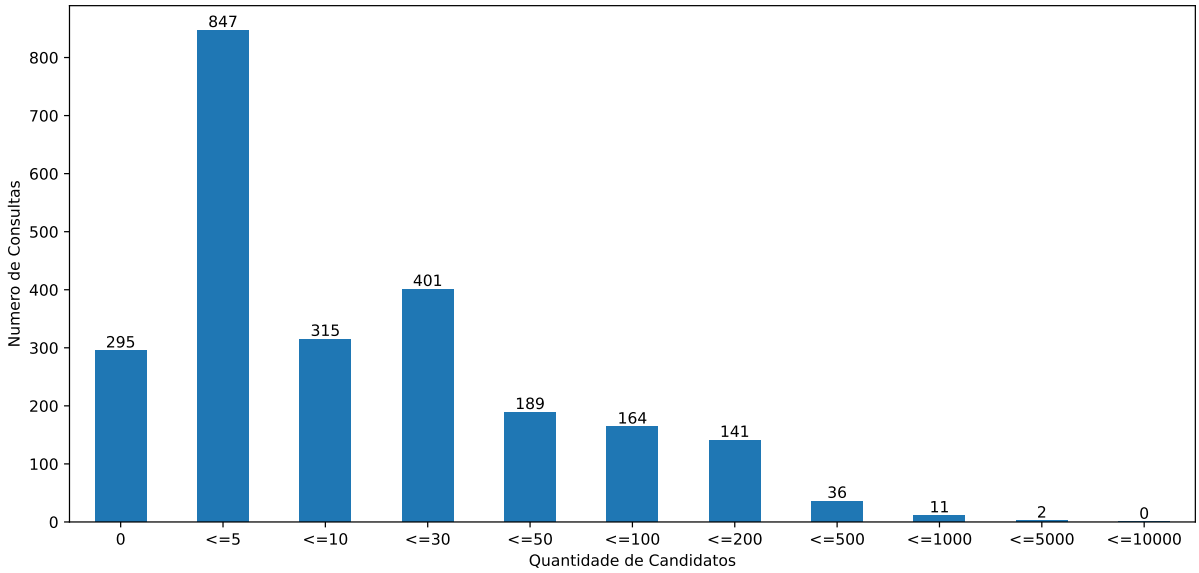
Na Tabela 3 estão os resultados do experimento. Nota-se que o caso em que não foi utilizada a correspondência exata apresentou melhor desempenho na distância média. Entretanto, o caso em que foi utilizada uma consulta mais rígida apresentou um desempenho melhor para Acurácia@161km e AUC. Isto se mantém mesmo ao comparar as métricas considerando somente

Figura 12 – Contagens de candidatos para as consultas com correspondência relaxada



Fonte: Elaborado pelo autor

Figura 13 – Contagens de candidatos para as consultas com correspondência exata



Fonte: Elaborado pelo autor

os topônimos que são resolvidos por ambas as estratégias. É importante ressaltar que os valores de score mínimo, ordenação e distância mútua foram os valores padrão, conforme descrito na Tabela 2.

5.1.1.2 Score Mínimo

Esse experimento avaliou a influência de um filtro de score mínimo na consulta de candidatos. Na Tabela 4 estão os resultados do experimento. Pode-se verificar que o melhor desempenho, em todas as métricas, foi o caso em que utilizou-se um filtro de score mínimo

Tabela 3 – Resultados do experimento com correspondência exata

Correspondência Exata	Distância Média	Acurácia@161km	AUC	# Topônimos
Não	**5212.033266 ±3439.733375**	0.234228 ±0.278889	0.669890 ±0.228203	2330
Sim	5635.218296 ±4066.462333	**0.347782 ±0.307579**	**0.566512 ±0.280165**	2106
Não	**5304.859257 ±3557.676932**	0.239143 ±0.289233	0.661394 ±0.244958	2106
Sim	5635.218296 ±4066.462333	**0.347782 ±0.307579**	**0.566512 ±0.280165**	2106

Fonte: Elaborado pelo autor.

com valor 20. Entretanto, analisando o número de topônimos que foram resolvidos, nota-se que esse desempenho se deu em troca de uma cobertura significativamente menor dos topônimos, retornando coordenadas geográficas para menos de um quinto de todas as localizações presentes nos textos. Apesar da melhora no desempenho, conclui-se que esta não é uma boa estratégia pois a redução na quantidade de topônimos resolvidos é substancial. É importante ressaltar que os valores de correspondência exata, ordenação e distância mútua foram os valores padrão, conforme descrito na Tabela 2.

Tabela 4 – Resultados do experimento com score mínimo

Score Mínimo	Distância Média	Acurácia@161km	AUC	# Topônimos
0	5212.033266 ±3439.733375	0.234228 ±0.278889	0.669890 ±0.228203	2330
10	5242.334061 ±3450.768567	0.234064 ±0.279066	0.670151 ±0.229129	2308
15	5013.586600 ±3604.883161	0.241374 ±0.285875	0.651854 ±0.240591	2001
20	**3138.836767 ±4094.170301**	**0.398328 ±0.434430**	**0.319002 ±0.355399**	413

Fonte: Elaborado pelo autor.

5.1.1.3 Experimento de Ordenação

Esse experimento avaliou a influência do atributo utilizado para ordenação dos candidatos na etapa de geocoding. A Tabela 5 traz os resultados do experimento. Verifica-se que o melhor desempenho, em todas as métricas utilizadas, foi o caso em que os candidatos foram ordenados por população. É importante ressaltar que os valores de correspondência exata, score mínimo e distância mútua foram os valores padrão, conforme descrito na Tabela 2.

Tabela 5 – Resultados do experimento de ordenação de candidatos

Ordenação	Distância Média	Acurácia@161km	AUC	# Topônimos
Score	5212.033266 ±3439.733375	0.234228 ±0.278889	0.669890 ±0.228203	2330
Feature Class	2879.740288 ±2257.998568	0.400149 ±0.313739	0.506514 ±0.232278	2330
População	**1228.269533 ±1583.496541**	**0.718704 ±0.249427**	**0.270962 ±0.187262**	2330

Fonte: Elaborado pelo autor.

5.1.2 Desambiguação de Candidatos

Para essa etapa, o experimento avaliou o impacto da utilização da distância mútua entre os candidatos e os topônimos previamente resolvidos na etapa de geocoding, assim como a influência da quantidade máxima de candidatos usados nessa desambiguação. A Tabela 6 traz os resultados do experimento. Nota-se que o caso em que foram utilizados os 5 primeiros candidatos na desambiguação por distância mínima foi o que apresentou a melhor distância média e área sob a curva. Entretanto, o melhor resultado para Acurácia@161km foi utilizando os Top-10 candidatos.

Tabela 6 – Resultados do experimento de desambiguação por distância mútua

Top-K + Distância Mútua	Distância Média	Acurácia@161km	AUC	# Topônimos
Não	5212.033266 ±3439.733375	0.234228 ±0.278889	0.669890 ±0.228203	2330
Top-5, Sim	**4365.741923 ±3701.190136**	0.256257 ±0.307421	**0.637616 ±0.244143**	2330
Top-10, Sim	4397.783715 ±3820.217714	**0.259079 ±0.319223**	0.638395 ±0.251320	2330
Top-20, Sim	4477.473471 ±3978.830264	0.253136 ±0.325534	0.645094 ±0.253533	2330

Fonte: Elaborado pelo autor.

É importante ressaltar que os valores de correspondência exata, score mínimo e ordenação foram os valores padrão, conforme descrito na Tabela 2. Além disso, outro ponto importante é que a distância mútua sempre é utilizada em conjunto com o Top-K, pois quando a distância mútua não está habilitada o candidato escolhido é sempre o primeiro.

5.1.3 Combinações de Parâmetros

Este último experimento avaliou as combinações dos parâmetros previamente testados visando escolher a melhor combinação como *geoparser* base ou HG. Os valores utilizados nessa experimentação encontram-se na Tabela 7. Esses valores foram escolhidos com base nos resultados dos experimentos anteriores.

Tabela 7 – Valores dos Parâmetros Usados no Experimento de Combinações

Parâmetro	**Valores**
Correspondência Exata	Sim, Não
Score Mínimo	0
Ordenação	Feature Class, População
Top-K + Distância Mútua	Não Top-5, Sim Top-10, Sim

Fonte: Elaborado pelo autor.

Na Tabela 8 estão dispostos os resultados para cada combinação de parâmetros avaliada. Nota-se que o melhor resultado foi aquele em que utilizou-se correspondência exata, ordenação por população e sem desambiguação por distância mútua, ou seja, a combinação das heurísticas H1 e H6. Portanto, essa configuração é escolhida como o *baseline* HG.

A Figura 14 apresenta os resultados de HG separados por classe do GeoWebNews. Cada barra mostra a distribuição de saídas do processo de *geocoding* por classe de topônimo, indicada no eixo y. A cor vermelha indica topônimos para os quais nenhum candidato foi encontrado no *gazetteer*, azul denota lugares resolvidos para as coordenadas esperadas e roxo indica referências a locais resolvidas para coordenadas a mais de 161 km das esperadas. Para topônimos da classe "Literal", referências diretas a localizações físicas (e.g. "*Harvests in* ***Australia***"), o *geocoder* apresenta um alto número de predições corretas. Entretanto, para os da classe "Non_Literal_Modifier", topônimos que modificam um conceito não locacional associado a um lugar (e.g. "***British*** *voters*"), há muitos casos em que nenhum candidato foi encontrado ou as coordenadas atribuídas foram muito distantes das esperadas.

5.2 Estratégias de Aprimoramento

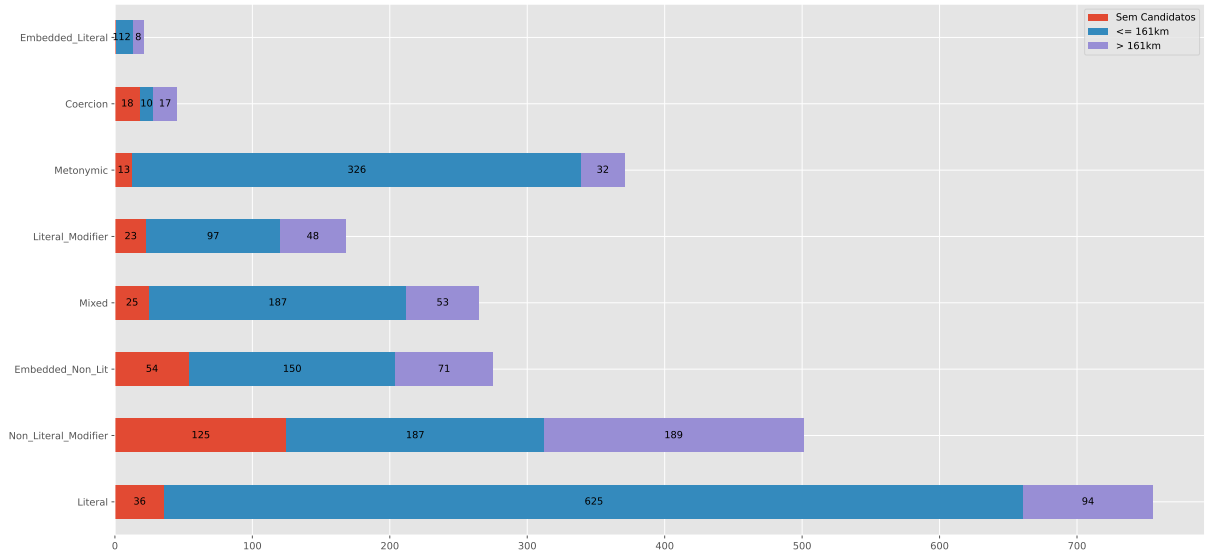
Escolhido o *geocoder* base, deu-se início a aplicação de estratégias para o aprimoramento de seu desempenho. A Tabela 9 apresenta os resultados dos experimentos realizados.

Tabela 8 – Resultados dos experimentos combinando os parâmetros

Correspondência Exata	Ordenação	Top-K + Distância Mútua	Distância Média	Acurácia@161km	AUC	# Topônimos
Sim	População	Não	**1162.944522 ±1612.917557**	**0.771268 ±0.251918**	**0.204715 ±0.177872**	2106
Sim	População	Top-5, Sim	1447.626662 ±1947.277502	0.657638 ±0.319900	0.293033 ±0.244578	2106
Sim	População	Top-10, Sim	1540.183033 ±2122.660416	0.638705 ±0.325980	0.310784 ±0.251467	2106
Sim	Feature Class	Não	2667.204955 ±2794.158863	0.551150 ±0.316706	0.373510 ±0.248511	2106
Sim	Feature Class	Top-5, Sim	2358.257848 ±2959.087513	0.526916 ±0.350471	0.384513 ±0.276446	2106
Sim	Feature Class	Top-10, Sim	2280.393966 ±2938.879049	0.524966 ±0.354939	0.384846 ±0.278361	2106
Não	População	Não	1228.269533 ±1583.496541	0.718704 ±0.249427	0.270962 ±0.187262	2330
Não	População	Top-5, Sim	1378.690984 ±1887.402444	0.607141 ±0.309531	0.367361 ±0.231923	2330
Não	População	Top-10, Sim	1578.729626 ±2088.028937	0.548879 ±0.334898	0.412774 ±0.243208	2330
Não	Feature Class	Não	2879.740288 ±2257.998568	0.400149 ±0.313739	0.506514 ±0.232278	2330
Não	Feature Class	Top-5, Sim	2659.848372 ±2502.937725	0.387915 ±0.333626	0.522184 ±0.247749	2330
Não	Feature Class	Top-10, Sim	2675.901519 ±2779.186909	0.374877 ±0.346944	0.536862 ±0.253505	2330

Fonte: Elaborado pelo autor.

Figura 14 – Resultados do HG no GeoWebNews por tipo de topônimo



Fonte: Elaborado pelo autor

A estratégia de HG-A buscou, através de um dicionário de adjetivos pátrios, melhorar o desempenho utilizando os nomes reais do locais em vez das formas adjetivas extraídas do texto. Verifica-se que a adição desse passo no processo de *geocoding* melhorou consideravelmente o desempenho do *geocoder* quando comparado ao *baseline* HG. Vale ressaltar que HG-A utiliza as

Tabela 9 – Resultados dos experimentos das estratégias de aprimoramento

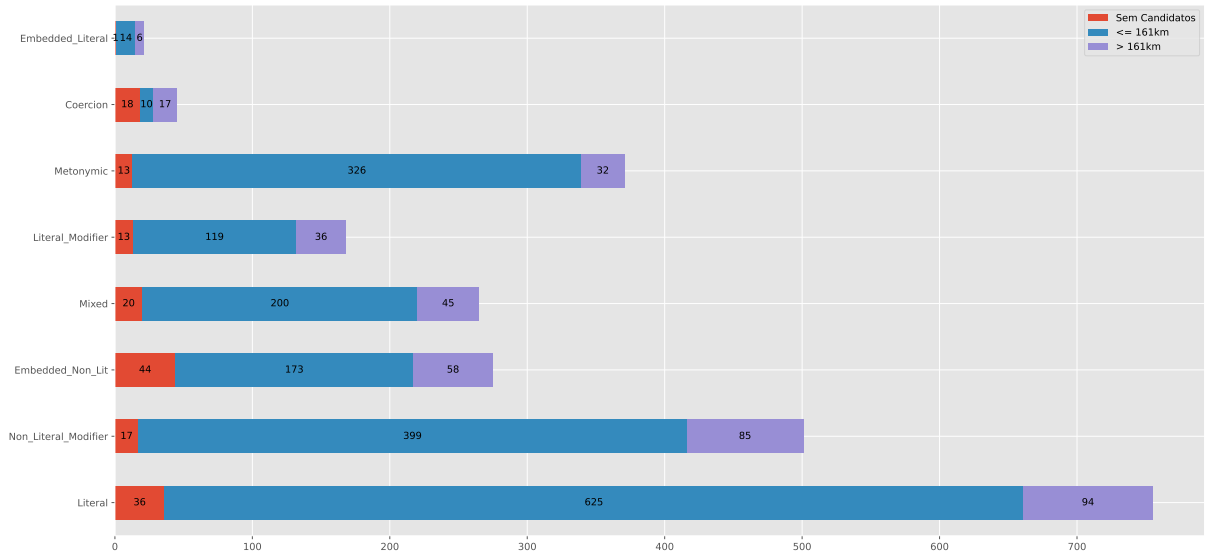
Geoparser	Distância Média	Geocodificados até 161 km	AUC	# Topônimos
HG	1162.944522 ±1612.917557	0.771268 ±0.251918	0.204715 ±0.177872	2106
HG-A	729.976503 ±1340.185650	0.818751 ±0.241200	0.163382 ±0.165491	2239
HG-T	1065.829266 ±1550.477554	0.777777 ±0.249302	0.198074 ±0.174734	2106
HG-AT	**633.379932 ±1231.057199**	**0.825252 ±0.237108**	**0.156947 ±0.160361**	2239

Fonte: Elaborado pelo autor.

heurísticas H1, H6 e H8.

A Figura 15 apresenta os resultados do HG-A separados por classe do GeoWebNews. Comparado ao *baseline*, o *geocoding* apresentou uma melhora de desempenho para diversos tipos, aumentando a quantidade de topônimos resolvidos corretamente. Topônimos associativos, indicados pela classe "Non_Literal_Modifier", são os que apresentam a melhora mais notável. Além disso, nota-se também que houve um aumento na quantidade de locais nos textos para os quais foi possível encontrar pelo menos um candidato.

Figura 15 – Resultados do Geocoding com Processamento de Topônimos Adjetivos



Fonte: Elaborado pelo autor

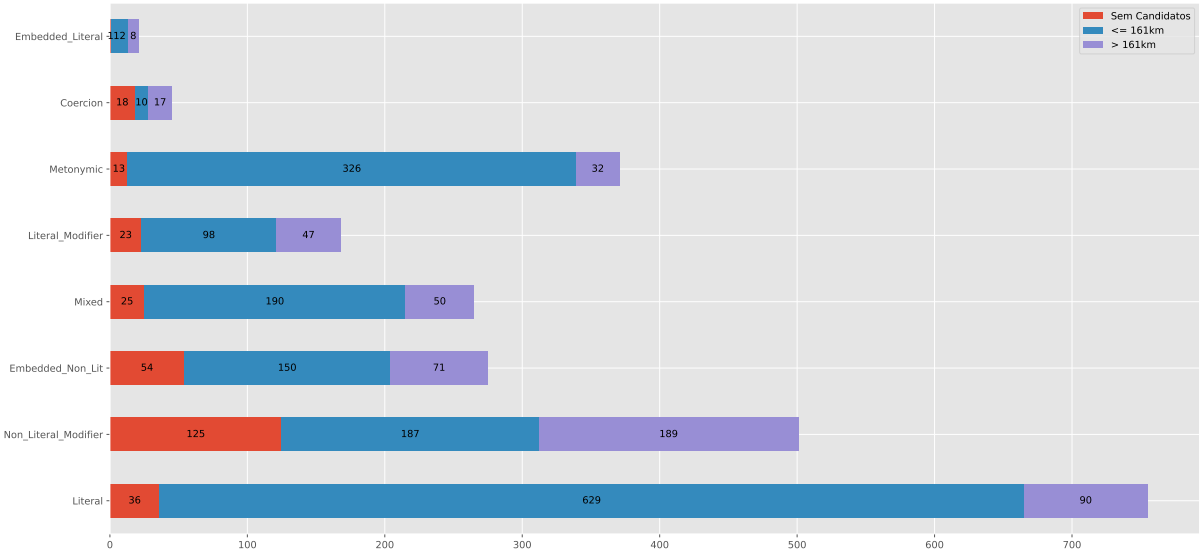
Essa diferença de desempenho se dá pela heurística de processamento de topônimos adjetivos. Por exemplo, considerando a sentença "*They were found in the southern* ***English*** *city of* ***Salisbury***", o *geocoder* HG atribuiria as coordenadas para a cidade de *English, Indiana* em vez das referentes à *Inglaterra* para o topônimo adjetivo "*English*". HG-A consegue lidar com esse topônimo devido ao passo de normalização adicionado, que faz o *geocoder* procurar por

candidatos correspondentes a “*England*”.

A estratégia de HG-T buscou melhorar o desempenho priorizando determinados tipos de topônimo nos textos. As distâncias para topônimos literais já resolvidos foram calculadas para os Top-5 candidatos, pois essa quantidade apresentou os melhores resultados com minimização geométrica nos experimentos para definição de HG, reportados na Tabela 8. Observando a Tabela 9, verifica-se que houve uma leve melhora em relação ao *baseline*. Vale ressaltar que HG-T utiliza as heurísticas H1, H6 e H9.

A Figura 16 traz os resultados de HG-T separados por classe do GeoWebNews. O *geocoding* apresentou uma leve melhora de desempenho para as classes “Literal”, “Literal_Modifier” e “Mixed”, todos pertencentes a categoria Literal, que é o foco da heurística utilizada por HG-T.

Figura 16 – Resultados do Geocoding com Otimização Geométrica por Tipo



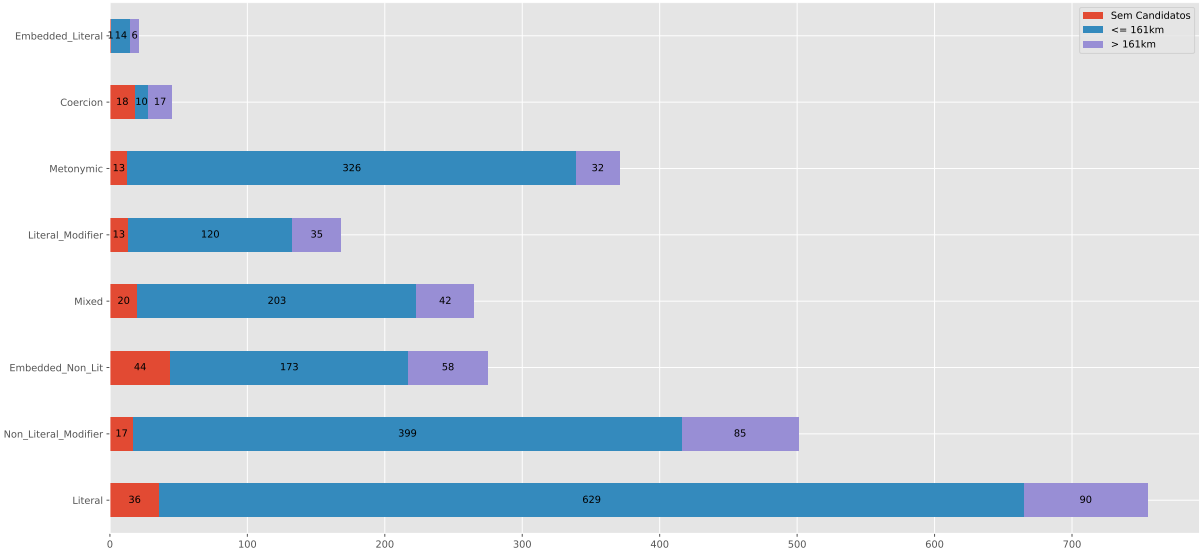
Fonte: Elaborado pelo autor

O *geocoder* HG-AT buscou combinar as estratégias de HG-A e HG-T, assim realizando tanto a normalização de topônimos adjetivos quanto a otimização geométrica por tipo. Assim como HG-T, HG-AT calcula as distâncias para literais resolvidos para os Top-5 candidatos. Observando a Tabela 9, verifica-se que este *geocoder* apresentou o melhor desempenho entre as abordagens apresentadas. Vale ressaltar que HG-T utiliza as heurísticas H1, H6, H8 e H9.

A Figura 17 traz os resultados de HG-AT separados por classe do GeoWebNews. Como este *geocoder* combina as estratégias de HG-A e HG-T, ele apresenta as melhoras de ambos os *geocoders* no processo de resolução dos topônimos das base de dados, superando HG em todas as classes exceto “Coercion” e “Metonymic”, para as quais apresenta a mesma

quantidade de acertos que o *baseline*. Além disso, HG-AT supera também HG-A e HG-T.

Figura 17 – Resultados do Geocoding com Processamento de Topônimos Adjetivos e Otimização Geométrica por Tipo



Fonte: Elaborado pelo autor

5.3 Comparação com Estado da Arte

Após a realização dos experimentos com as estratégias de aprimoramento, deu-se início às comparações com *geoparsers* do estado da arte. Uma vez que podem haver divergências entre os topônimos que são resolvidos por cada *geocoder*, as métricas foram calculadas para todas as localizações resolvidas e também utilizando somente aquelas que são resolvidas por todos os *geocoders* avaliados.

Os resultados dos *geocoders* propostos, quando aplicados aos topônimos do GeoWebNews reconhecidos pelo CLAVIN encontram-se na Tabela 10. Já os resultados para o TR-News podem ser visualizados na Tabela 11. Pode-se observar que as abordagens propostas apresentaram um resultado melhor que o *geocoder* do CLAVIN em ambas as bases de dados. Apesar de para alguns topônimos nenhum candidato ter sido retornado pelos *geocoders* propostos, para vários outros a resolução escolheu candidatos corretamente onde o CLAVIN falhou.

Dentre os *geocoders* avaliados, HG-AT e HG-T obtiveram os melhores resultados. Uma vez que os *geocoders* foram aplicados nos topônimos reconhecidos pelo CLAVIN, que descarta topônimos em forma adjetiva, as diferenças de desempenho se devem à estratégia de resolução adotada. A Figura 18 apresenta os resultados no GeoWebNews, separados por classe, do HG-AT em comparação aos do CLAVIN. Pode-se observar que o HG-AT apresentou um resultado

Tabela 10 – Comparação com CLAVIN no GeoWebNews

Geoparser	Distância Média	Acurácia@161km	AUC	# Topônimos
CLAVIN	790.182865 ±1585.180927	0.810020 ±0.310121	0.126768 ±0.215262	976
HG-A	392.850569 ±1105.984496	0.843183 ±0.291370	0.108625 ±0.173859	959
HG-T	**391.895126 ±1106.020871**	**0.844054 ±0.291196**	**0.108444 ±0.173798**	959
HG-AT	**391.895126 ±1106.020871**	**0.844054 ±0.291196**	**0.108444 ±0.173798**	959
CLAVIN	758.187965 ±1593.402278	0.812387 ±0.319436	0.120451 ±0.212725	959
HG-A	392.850569 ±1105.984496	0.843183 ±0.291370	0.108625 ±0.173859	959
HG-T	**391.895126 ±1106.020871**	**0.844054 ±0.291196**	**0.108444 ±0.173798**	959
HG-AT	**391.895126 ±1106.020871**	**0.844054 ±0.291196**	**0.108444 ±0.173798**	959

Fonte: Elaborado pelo autor.

Tabela 11 – Comparação com CLAVIN no TR-News

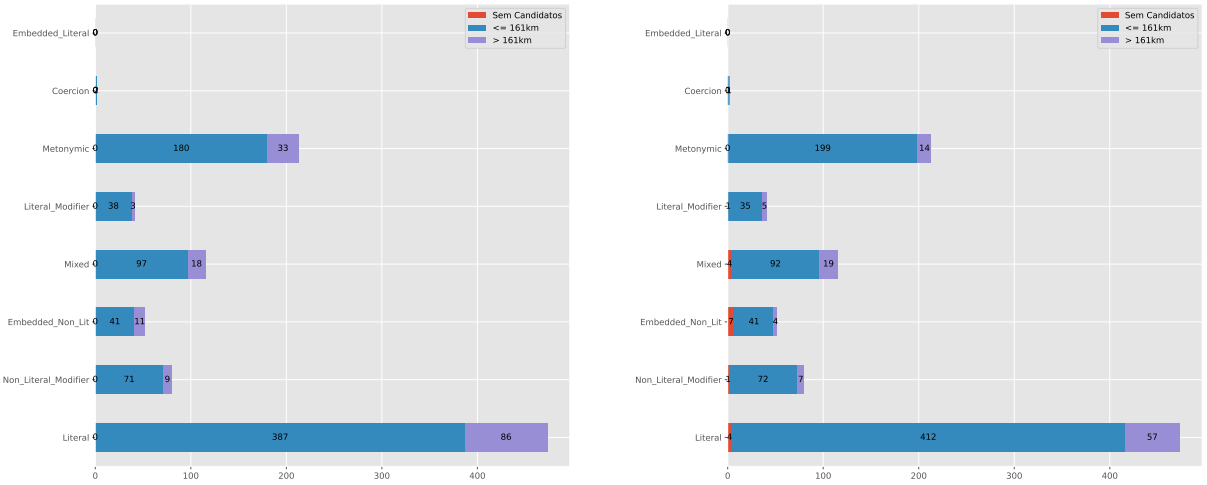
Geoparser	Distância Média	Acurácia@161km	AUC	# Topônimos
CLAVIN	1570.097288 ±2685.965745	0.768716 ±0.327460	0.188900 ±0.258055	666
HG-A	1424.230601 ±2644.265101	0.776993 ±0.334597	0.186386 ±0.259160	649
HG-T	**1409.543667 ±2630.627251**	**0.779558 ±0.333606**	**0.184541 ±0.258646**	649
HG-AT	**1409.543667 ±2630.627251**	**0.779558 ±0.333606**	**0.184541 ±0.258646**	649
CLAVIN	1656.949706 ±2944.170846	0.755328 ±0.347160	0.188443 ±0.264845	649
HG-A	1424.230601 ±2644.265101	0.776993 ±0.334597	0.186386 ±0.259160	649
HG-T	**1409.543667 ±2630.627251**	**0.779558 ±0.333606**	**0.184541 ±0.258646**	649
HG-AT	**1409.543667 ±2630.627251**	**0.779558 ±0.333606**	**0.184541 ±0.258646**	649

Fonte: Elaborado pelo autor.

melhor que o CLAVIN para os topônimos das classes “Metonymic”, “Non_Literal_Modifier” e “Literal”, mas teve um desempenho levemente pior para as classes “Literal_Modifier” e “Mixed”.

Quando comparado ao CLAVIN, no GeoWebNews e no TR-News, o HG-AT demonstrou um resultado melhor para localizações com *feature classes* A e T no GeoNames. Por exemplo, para o texto “*The meeting took place in* ***Brussels, Belgium***”, o CLAVIN traz como resultado para “Brussels” e “Belgium”, respectivamente, as cidades de Brussels e Belgium em Wisconsin, nos Estados Unidos. Isso ocorre devido ao mecanismo do CLAVIN que prioriza

Figura 18 – Comparação com CLAVIN no GeoWebNews por tipo de topônimo



(a) Resultados do CLAVIN

(b) Resultados do HG-AT

Fonte: Elaborado pelo autor
A cor vermelha indica topônimos para os quais nenhum candidato foi encontrado. A cor roxa indica topônimos resolvidos para localizações além da distância máxima de 161 km. A cor azul indica topônimos resolvidos corretamente.

candidatos que compartilham o mesmo código de nível administrativo 1. Já o HG-AT, que se baseia na população e distância do candidato para topônimos já resolvidos, retorna corretamente as coordenadas para a Cidade de Bruxelas e o Reino da Bélgica.

Os resultados da comparação entre o *geocoders* propostos e o CamCoder no GeoWebNews e no TR-News, respectivamente, podem ser visualizados na Tabela 12 na Tabela 13. Pode-se observar que o HG-AT apresentou o melhor resultado para todas as métricas no GeoWebNews. Já no TR-News, o HG-AT obteve os melhores resultados para Acurácia@161km e AUC. É importante ressaltar que nesse caso todos os topônimos anotados foram submetidos aos *geocoders*, de forma a comparar apenas o desempenho da resolução de topônimos.

A Figura 19 traz os resultados do *geocoding* no GeoWebNews para o CamCoder e o HG-AT. Verifica-se que para todas as classes do GeoWebNews, que seguem a taxonomia utilizada neste trabalho, o HG-AT apresentou um resultado melhor que o CamCoder. Além disso, o HG-AT apresenta resultados melhores que o CamCoder para localizações com *feature classes* A, T e H no GeoNames.

A principal vantagem do HG-AT está na resolução dos topônimos adjetivos, que o CamCoder não trata antes de tentar resolver. Por exemplo, para o texto "*The chancellor of a* ***Spanish*** *university [...]*", o CamCoder traz como resultado a cidade de Spanish em Ontario, no Canadá. O CLAVIN ignora o topônimo, pois não o considera uma referência a uma localização. Já o HG-AT, que trata esse tipo de topônimo, retorna corretamente as coordenadas para o Reino

Tabela 12 – Comparação com CamCoder no GeoWebNews

Geoparser	Distância Média	Acurácia@161km	AUC	# Topônimos
CamCoder	1037.378719 ±1514.639870	0.752997 ±0.267948	0.206935 ±0.188238	2102
HG-A	729.976503 ±1340.185650	0.818751 ±0.241200	0.163382 ±0.165491	2239
HG-T	1065.829266 ±1550.477554	0.777777 ±0.249302	0.198074 ±0.174734	2106
HG-AT	**633.379932 ±1231.057199**	**0.825252 ±0.237108**	**0.156947 ±0.160361**	2239
CamCoder	1037.378719 ±1514.639870	0.752997 ±0.267948	0.206935 ±0.188238	2102
HG-A	751.253401 ±1358.503972	0.813993 ±0.242896	0.167742 ±0.166071	2102
HG-T	1060.668945 ±1543.034846	0.778190 ±0.249087	0.197818 ±0.174502	2102
HG-AT	**654.168792 ±1252.058028**	**0.820503 ±0.239015**	**0.161272 ±0.161185**	2102

Fonte: Elaborado pelo autor.

Tabela 13 – Comparação com CamCoder no TR-News

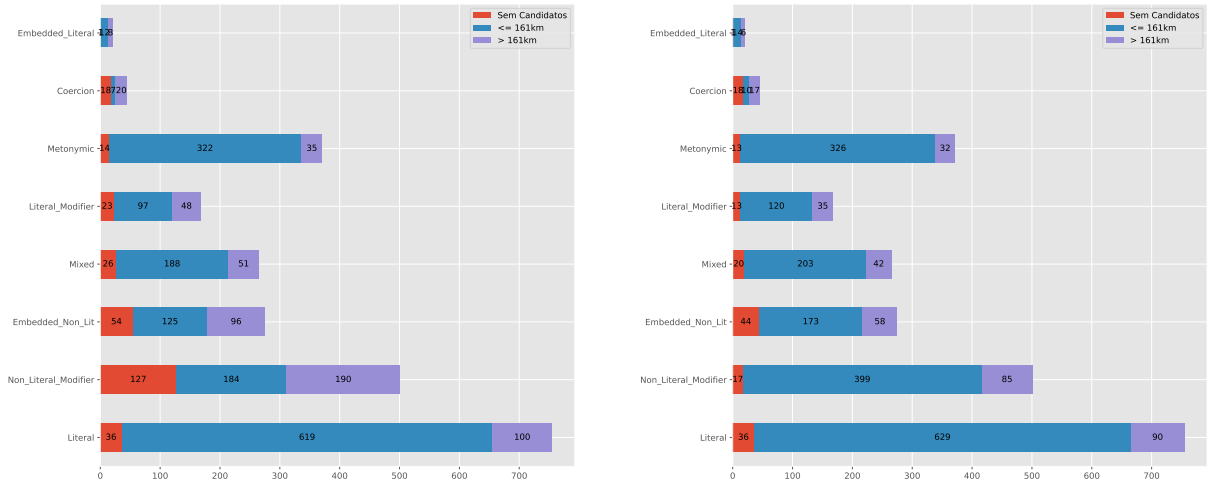
Geoparser	Distância Média	Acurácia@161km	AUC	# Topônimos
CamCoder	**1140.507819 ±1576.841230**	0.793848 ±0.240413	0.200976 ±0.198811	1165
HG-A	1250.090958 ±1597.009169	0.803426 ±0.237952	0.192425 ±0.198230	1245
HG-T	1513.268402 ±1681.162054	0.758852 ±0.252310	0.230948 ±0.206495	1200
HG-AT	1162.391805 ±1526.622529	**0.815258 ±0.235708**	**0.182969 ±0.194414**	1245
CamCoder	**1140.507819 ±1576.841230**	0.793848 ±0.240413	0.200976 ±0.198811	1165
HG-A	1264.822263 ±1605.746453	0.800616 ±0.239104	0.195440 ±0.198440	1165
HG-T	1511.327056 ±1688.305957	0.759861 ±0.251789	0.230999 ±0.206058	1165
HG-AT	1177.129504 ±1536.615613	**0.812449 ±0.237012**	**0.186023 ±0.194835**	1165

Fonte: Elaborado pelo autor.

de Espanha.

Por outro lado, conforme esperado de um *geocoder* com resolução fortemente influenciada pelo valor de população, localizações como prédios, aeroportos, parques, vilas e seções de lugares populados ainda são um problema. Isso é um problema especialmente no TR-News, devido às ambiguidades como no caso de Heathrow, o aeroporto em Londres, na Inglaterra, e Heathrow, a comunidade suburbana na Flórida, nos Estados Unidos. Quando aplicado ao GeoWebNews, esse problema é amenizado através da heurística H9, que considera a

Figura 19 – Comparação com CamCoder no GeoWebNews por tipo de topônimo



(a) Resultados do CamCoder

(b) Resultados do HG-AT

Fonte: Elaborado pelo autor

A cor vermelha indica topônimos para os quais nenhum candidato foi encontrado. A cor roxa indica topônimos resolvidos para localizações além da distância máxima de 161 km. A cor azul indica topônimos resolvidos corretamente.

distância para os topônimos literais na resolução. Entretanto, isto não é possível com o TR-News, uma vez que este não possui anotações dos tipos de topônimo, impossibilitando a diferenciação entre os topônimos no cálculo das distâncias.

6 CONCLUSÕES E TRABALHOS FUTUROS

Este trabalho teve como objetivo propor e avaliar novas estratégias heurísticas para resolução de topônimos extraídos de textos não estruturados. Inicialmente, nos Capítulos 2 e 3, foram apresentados os conceitos fundamentais para compreensão desta dissertação e os trabalhos relacionados. No Capítulo 2 foi também apresentada uma definição formal do problema tratado por este trabalho. Nos Capítulos 4 e 5 foram apresentados a metodologia utilizada e os resultados obtidos.

O processo de *geocoding* neste trabalho foi dividido nas subtarefas de obtenção de candidatos e subsequente desambiguação. A obtenção de candidatos foi realizada consultando um dicionário geográfico. Para tal, foi utilizada a ferramenta *off-the-shelf* ElasticSearch para consultar um índice populado com dados do GeoNames. Já a resolução de topônimos foi realizada com base em heurística.

Para a definição do *baseline*, inicialmente foram avaliadas as heurísticas de correspondência exata (H1), correspondência relaxada (H2), pontuação mínima de candidato (H3), ordenação por score (H4), ordenação por *feature class* (H5), ordenação por população (H6) e minimização geométrica (H7). Essas heurísticas foram combinadas e avaliadas de forma a eleger um *baseline* HG, que utiliza H1 e H6, para avaliação de estratégias de aprimoramento.

A primeira heurística proposta para aprimoramento do *baseline* foi a normalização de topônimos adjetivos (H8). Essa estratégia consistiu em utilizar um dicionário de formas adjetivas de países para mapear topônimos adjetivos para sua forma substantiva antes de realizar a obtenção de candidatos. Em outras palavras, ao tentar resolver um topônimo como "*Finnish*", que está em forma adjetiva, o *geocoder* busca candidatos para "*Republic of Finland*", a forma substantiva. O *geocoder* aprimorado que utiliza essa heurística foi denominado de HG-A.

A segunda heurística proposta foi a otimização geométrica de topônimos por tipo (H9). Essa estratégia baseou-se na ideia de tratar de forma diferente topônimos literais e associativos no momento da desambiguação. Para os associativos, a heurística não alterou o processo de resolução, mas para os literais os candidatos passaram a ser escolhidos minimizando a razão entre a população do candidato e sua distância média para os outros topônimos literais já resolvidos. O *geocoder* aprimorado que utiliza essa heurística foi denominado de HG-T.

Por fim, foi criado um terceiro *geocoder* aprimorado nomeado de HG-AT. Este geocoder combinou as estratégias de HG-A e HG-T, realizando tanto a normalização dos topônimos adjetivos quanto a otimização geométrica de topônimos por tipo. Dessa forma,

HG-AT utilizou tanto as heurísticas H1 e H6 de HG, quanto as heurísticas H8 e H9 que foram adicionadas aos *geocoders* HG-A e HG-T, respectivamente.

Ao serem avaliados, os três *geocoders* aprimorados com as heurísticas propostas superaram o *baseline* HG na base de dados GeoWebNews. Dentre eles, o *geocoder* com melhor desempenho foi HG-AT. Este resultado positivo se manteve ao comparar os *geocoders* propostos com os *geocoders* do estado da arte CLAVIN e CamCoder nas bases de dados GeoWebNews e TR-News. Por outro lado, o HG-AT teve dificuldade em resolver localizações como prédios, parques, vilas e seções de lugares populados. Essa dificuldade é amenizada pela heurística de otimização geométrica de topônimos por tipo, mas ainda é um problema.

O trabalho realizado ao longo desta pesquisa resultou na publicação de um artigo, Sá *et al.* (2022), mas ainda há muito a ser explorado nesta área. Para trabalhos futuros, mais experimentos podem ser realizados utilizando outras bases de dados, que não sejam compostas de notícias ou que tratem de localizações mais granulares, para verificar diferenças de desempenho e a aplicabilidade das abordagens heurísticas em outros contextos. A normalização de topônimos adjetivos pode ser melhorada adicionando mais formas adjetivas relacionadas a outras regiões administrativas como províncias e cidades. Já a otimização geométrica de topônimos por tipo pode ser melhorada avaliando-se os tipos de topônimo ou *feature classes* que influenciam mais o foco geográfico do texto. Além disso, diferentes estratégias podem ser utilizadas para desambiguar candidatos baseando-se no tipo de topônimo além da distância geográfica. Por fim, o processamento de topônimos embutidos também pode melhorar o *geocoding*.

REFERÊNCIAS

ABDELKADER, A.; HAND, E.; SAMET, H. Brands in newsstand: Spatio-temporal browsing of business news. *In*: **Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems**. [*S. l.*: *s. n.*], 2015. p. 1–4.

ALBINATI, J.; JR, W. M.; PAPPA, G. L.; TEIXEIRA, M.; MARQUES-TOLEDO, C. Enhancement of epidemiological models for dengue fever based on twitter data. *In*: **Proceedings of the 2017 International Conference on Digital Health**. [*S. l.*: *s. n.*], 2017. p. 109–118.

ALDANA-BOBADILLA, E.; MOLINA-VILLEGAS, A.; LOPEZ-AREVALO, I.; REYES-PALACIOS, S.; MUÑIZ-SANCHEZ, V.; ARREOLA-TRAPALA, J. Adaptive geoparsing method for toponym recognition and resolution in unstructured text. **Remote Sensing**, Multidisciplinary Digital Publishing Institute, v. 12, n. 18, p. 3041, 2020.

ALLEN, C.; TSOU, M.-H.; ASLAM, A.; NAGEL, A.; GAWRON, J.-M. Applying gis and machine learning methods to twitter data for multiscale surveillance of influenza. **PloS one**, Public Library of Science San Francisco, CA USA, v. 11, n. 7, p. e0157734, 2016.

AMITAY, E.; HAR'EL, N.; SIVAN, R.; SOFFER, A. Web-a-where: geotagging web content. *In*: **Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval**. [*S. l.*: *s. n.*], 2004. p. 273–280.

BUSCALDI, D.; ROSSO, P. A conceptual density-based approach for the disambiguation of toponyms. **International Journal of Geographical Information Science**, Taylor & Francis, v. 22, n. 3, p. 301–313, 2008.

BUTT, M.; HUSSAIN, S. Proceedings of the 51st annual meeting of the association for computational linguistics: System demonstrations. *In*: **Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations**. [*S. l.*: *s. n.*], 2013.

CHEN, H.; VASARDANI, M.; WINTER, S. Clustering-based disambiguation of fine-grained place names from descriptions. **GeoInformatica**, Springer, v. 23, n. 3, p. 449–472, 2019.

CLEMENS, K. Geocoding with openstreetmap data. **GEOProcessing 2015**, p. 10, 2015.

CLOUGH, P. Extracting metadata for spatially-aware information retrieval on the internet. *In*: **Proceedings of the 2005 workshop on Geographic information retrieval**. [*S. l.*: *s. n.*], 2005. p. 25–30.

COLLADON, A. F.; GUARDABASCIO, B.; INNARELLA, R. Using social network and semantic analysis to analyze online travel forums and forecast tourism demand. **Decision Support Systems**, Elsevier, v. 123, p. 113075, 2019.

DELOZIER, G.; BALDRIDGE, J.; LONDON, L. Gazetteer-independent toponym resolution using geographic word profiles. *In*: **Twenty-Ninth AAAI Conference on Artificial Intelligence**. [*S. l.*: *s. n.*], 2015.

DIVYA, M. S.; GOYAL, S. K. Elasticsearch: An advanced and quick search technique to handle voluminous data. **Compusoft**, COMPUSOFT, An International Journal of Advanced Computer Technology, v. 2, n. 6, p. 171, 2013.

GOODCHILD, M. F.; HILL, L. L. Introduction to digital gazetteer research. **International Journal of Geographical Information Science**, Taylor & Francis, v. 22, n. 10, p. 1039–1044, 2008.

GRITTA, M. **Where are you talking about? advances and challenges of geographic analysis of text with application to disease monitoring**. Tese (Doutorado) – University of Cambridge, 2019.

GRITTA, M.; PILEHVAR, M.; COLLIER, N. Which melbourne? augmenting geocoding with maps. 2018.

GRITTA, M.; PILEHVAR, M. T.; COLLIER, N. A pragmatic guide to geoparsing evaluation. **Language Resources and Evaluation**, Springer, p. 1–30, 2019.

GRITTA, M.; PILEHVAR, M. T.; LIMSOPATHAM, N.; COLLIER, N. What's missing in geographical parsing? **Language Resources and Evaluation**, Springer, v. 52, n. 2, p. 603–623, 2018.

GROVER, C.; TOBIN, R.; BYRNE, K.; WOOLLARD, M.; REID, J.; DUNN, S.; BALL, J. Use of the edinburgh geoparser for georeferencing digitized historical collections. **Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences**, The Royal Society Publishing, v. 368, n. 1925, p. 3875–3889, 2010.

GUNASEKARAN, A. K.; IMANI, M. B.; KHAN, L.; GRANT, C.; BRANDT, P. T.; HOLMES, J. S. Sperg: Scalable political event report geoparsing in big data. *In*: IEEE. **2018 IEEE International Conference on Intelligence and Security Informatics (ISI)**. [*S. l.*], 2018. p. 187–192.

HILL, L. L. Core elements of digital gazetteers: placenames, categories, and footprints. *In*: SPRINGER. **International Conference on Theory and Practice of Digital Libraries**. [*S. l.*], 2000. p. 280–290.

HONNIBAL, M.; MONTANI, I.; LANDEGHEM, S. V.; BOYD, A. **spaCy: Industrial-strength Natural Language Processing in Python**. Zenodo, 2020. Disponível em: https://doi.org/10.5281/zenodo.1212303.

HULDEN, M.; SILFVERBERG, M.; FRANCOM, J. Kernel density estimation for text-based geolocation. *In*: **Proceedings of the AAAI Conference on Artificial Intelligence**. [*S. l.*: *s. n.*], 2015. v. 29, n. 1.

JIANG, Y.; HUANG, X.; LI, Z. Spatiotemporal patterns of human mobility and its association with land use types during covid-19 in new york city. **ISPRS International Journal of Geo-Information**, Multidisciplinary Digital Publishing Institute, v. 10, n. 5, p. 344, 2021.

KAMALLOO, E.; RAFIEI, D. A coherent unsupervised model for toponym resolution. *In*: **Proceedings of the 2018 World Wide Web Conference**. [*S. l.*: *s. n.*], 2018. p. 1287–1296.

KARIMZADEH, M.; HUANG, W.; BANERJEE, S.; WALLGRÜN, J. O.; HARDISTY, F.; PEZANOWSKI, S.; MITRA, P.; MACEACHREN, A. M. Geotxt: a web api to leverage place references in text. *In*: **Proceedings of the 7th workshop on geographic information retrieval**. [*S. l.*: *s. n.*], 2013. p. 72–73.

KARIMZADEH, M.; PEZANOWSKI, S.; MACEACHREN, A. M.; WALLGRÜN, J. O. Geotxt: A scalable geoparsing system for unstructured text geolocation. **Transactions in GIS**, Wiley Online Library, v. 23, n. 1, p. 118–136, 2019.

LAMPOS, V.; CRISTIANINI, N. Nowcasting events from the social web with statistical learning. **ACM Transactions on Intelligent Systems and Technology (TIST)**, ACM New York, NY, USA, v. 3, n. 4, p. 1–22, 2012.

LEIDNER, J. L. **Toponym resolution in text: Annotation, evaluation and applications of spatial grounding of place names**. [*S. l.*]: Universal-Publishers, 2008.

LIEBERMAN, M. D.; SAMET, H.; SANKARANARAYANAN, J. Geotagging with local lexicons to build indexes for textually-specified spatial data. *In*: IEEE. **2010 IEEE 26th international conference on data engineering (ICDE 2010)**. [*S. l.*], 2010. p. 201–212.

NIZZOLI, L.; AVVENUTI, M.; TESCONI, M.; CRESCI, S. Geo-semantic-parsing: Ai-powered geoparsing by traversing semantic knowledge graphs. **Decision Support Systems**, Elsevier, v. 136, p. 113346, 2020.

O'BRIEN, S. P. Crisis early warning and decision support: Contemporary approaches and thoughts on future research. **International studies review**, Blackwell Publishing Ltd Oxford, UK, v. 12, n. 1, p. 87–104, 2010.

RAHIMI, A.; VU, D.; COHN, T.; BALDWIN, T. Exploiting text and network context for geolocation of social media users. **arXiv preprint arXiv:1506.04803**, 2015.

RAUCH, E.; BUKATIN, M.; BAKER, K. A confidence-based framework for disambiguating geographic terms. *In*: **Proceedings of the HLT-NAACL 2003 workshop on Analysis of geographic references**. [*S. l.*: *s. n.*], 2003. p. 50–54.

ROBERTS, M. The semantics of demonyms in english: Germans, queenslanders, and londoners. **The semantics of nouns**, p. 205–220, 2017.

ROBERTSON, S. E.; WALKER, S.; JONES, S.; HANCOCK-BEAULIEU, M. M.; GATFORD, M. *et al.* Okapi at trec-3. **Nist Special Publication Sp**, NATIONAL INSTIUTE OF STANDARDS & TECHNOLOGY, v. 109, p. 109, 1995.

SÁ, B. D.; SILVA, T. C. da; MACÊDO, J. A. F. de. Enhancing geocoding of adjectival toponyms with heuristics. *In*: **Proceedings of the LREC 2022 workshop on Natural Language Processing for Political Sciences**. [*S. l.*: *s. n.*], 2022. p. 37–45.

SINGH, J. P.; DWIVEDI, Y. K.; RANA, N. P.; KUMAR, A.; KAPOOR, K. K. Event classification and location prediction from tweets during disasters. **Annals of Operations Research**, Springer, v. 283, n. 1, p. 737–757, 2019.

SMITH, D. A.; CRANE, G. Disambiguating geographic names in a historical digital library. *In*: SPRINGER. **International Conference on Theory and Practice of Digital Libraries**. [*S. l.*], 2001. p. 127–136.

TOLKIEN, J. R. R. [Carta 144] Para Naomi Mitchison. *In*: CARPENTER, H.; TOLKIEN, C. (Ed.). **As cartas de J. R. R. Tolkien**. Tradução de Gabriel Oliva Brum. Curitiba: Arte Letra, 2006. p. 168–175.

VASARDANI, M.; WINTER, S.; RICHTER, K.-F. Locating place names from place descriptions. **International Journal of Geographical Information Science**, Taylor & Francis, v. 27, n. 12, p. 2509–2532, 2013.

VOMFELL, L.; HÄRDLE, W. K.; LESSMANN, S. Improving crime count forecasts using twitter and taxi data. **Decision Support Systems**, Elsevier, v. 113, p. 73–85, 2018.

WANG, J.; HU, Y. Enhancing spatial and textual analysis with eupeg: An extensible and unified platform for evaluating geoparsers. **Transactions in GIS**, Wiley Online Library, v. 23, n. 6, p. 1393–1419, 2019.

WU, D.; CUI, Y. Disaster early warning and damage assessment analysis using social media data and geo-location information. **Decision support systems**, Elsevier, v. 111, p. 48–59, 2018.

APÊNDICE A – GEOTAGGING

O processo de reconhecimento de topônimos é realizado utilizando o módulo de NER da ferramenta *off-the-shelf* spaCy. O modelo de NER utilizado, especificamente, é o en_core_web_lg, que é provido pela ferramenta e reconhece entidades nomeadas em textos na língua inglesa. Como o *geotagging* trata especificamente de localizações, apenas topônimos são considerados, todas as outras entidades nomeadas são descartadas.

Uma vez que o *geotagging* é uma sub-tarefa de Reconhecimento de Entidade Nomeada, a avaliação de desempenho é feita utilizando as métricas já estabelecidas, isto é, precisão, *recall* e F1-score. As Equações A.1, A.2 e A.3 descrevem o cálculo dessas métricas em ordem, onde S indica os topônimos reconhecidos no texto e T os topônimos esperados.

$$P(S,T) = \frac{|R \cap T|}{|T|} \tag{A.1}$$

$$R(S,T) = \frac{|S \cap T|}{|S|} \tag{A.2}$$

$$F1(S,T) = 2 * \frac{P(S,T) * R(S,T)}{P(S,T) + R(S,T)} \tag{A.3}$$

O impacto da fase de reconhecimento de topônimos no resultado final de um *geoparser* pode ser avaliado verificando-se a influência dos topônimos identificados e não identificados no desempenho final. Em outras palavras, avalia-se o quanto as entidades nomeadas de localização não encontradas durante o *geotagging* afetam a resolução de topônimos para coordenadas geográficas.

Nos experimentos realizados para a etapa de *geotagging*, foram avaliados o spaCy, utilizando o modelo NER en_core_web_lg provido pela ferramenta, e uma ferramenta de NER fictícia com predição perfeita chamada aqui de Oracle, de forma semelhante ao que foi feito por Gritta *et al.* (2019). Para o caso do NER Oracle, os topônimos anotados são submetidos diretamente ao *geocoder* a partir das anotações do conjunto de dados, ou seja, assume-se um processo com 100% de acerto, que encontra exatamente todos os topônimos esperados.

A Tabela 1 traz os resultados do spaCy na etapa de *geotagging* e a Tabela 2 os resultados da etapa de *geocoding*. Pode-se verificar que a distância média, a acurácia@161km e

a área sob a curva foram melhores no caso em que foi utilizado o spaCy. Analisando o número de topônimos resolvidos, porém, verifica-se que essa diferença de desempenho deve-se a falha na identificação de alguns topônimos durante o *geotagging* utilizando spaCy.

Tabela 1 – Desempenho do spaCy no *Geotagging*

Componente NER	**Precisão do Geotagger**	**Recall do Geotagger**	**F1-score do Geotagger**
spaCy	0.503695 ±0.232689	0.767677 ±0.217731	0.589930 ±0.182503

Fonte: Elaborado pelo autor.

Tabela 2 – Resultados finais do *geoparser* com spaCy e com NER perfeito

Componente NER	**Distância Média**	**Acurácia@161km**	**AUC**	**# Topônimos**
Oracle	5212.033266 ±3439.733375	0.234228 ±0.278889	0.669890 ±0.228203	2330
spaCy	**5002.326274 ±3659.636990**	**0.276150 ±0.310968**	**0.627342 ±0.261686**	1629

Fonte: Elaborado pelo autor.