

**USO DE GAZETTEERS FOCADOS PARA
GEOPARSING**

BRUNO RABELLO MONTEIRO

**USO DE GAZETTEERS FOCADOS PARA
GEOPARSING**

Tese apresentada ao Programa de Pós-Graduação em Ciência da Computação do Instituto de Ciências Exatas da Universidade Federal de Minas Gerais como requisito parcial para a obtenção do grau de Doutor em Ciência da Computação.

ORIENTADOR: CLODOVEU AUGUSTO DAVIS JR.

Belo Horizonte

Março de 2021

BRUNO RABELLO MONTEIRO

**USAGE OF FOCUSED GAZETTEERS IN
GEOPARSING**

Dissertation presented to the Graduate Program in Computer Science of the Federal University of Minas Gerais in partial fulfillment of the requirements for the degree of Doctor in Computer Science.

ADVISOR: CLODOVEU AUGUSTO DAVIS JR.

Belo Horizonte

March 2021

© 2021, Bruno Rabello Monteiro.
Todos os direitos reservados.

Monteiro, Bruno Rabello.

M775u Usage of focused gazetteers in geoparsing [manuscrito] /
Bruno Rabello Monteiro. – 2021.
xx, 88 f. il.

Orientador: Clodoveu Augusto Davis Junior.

Tese (Doutorado) - Universidade Federal de Minas Gerais,
Instituto de Ciências Exatas, Departamento de Ciência da
Computação.

Referências: f.59-71.

1. Computação – Teses. 2. Sistemas de informação
geográfica – Teses. 3. Geoparsing – Teses. 4. Problema de
resolução de escopo geográfico – Teses. I. Davis Junior,
Clodoveu Augusto. II. Universidade Federal de Minas Gerais,
Instituto de Ciências Exatas, Departamento de Ciência da
Computação. III. Título.

CDU 519.6*74(043)

Ficha catalográfica elaborada pela bibliotecária Belkiz Inez Rezende Costa
CRB 6ª Região nº 1510



UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

FOLHA DE APROVAÇÃO

Usage of Focused Gazetteers in Geoparsing

BRUNO RABELLO MONTEIRO

Tese defendida e aprovada pela banca examinadora constituída pelos Senhores:

PROF. CLODOVEU AUGUSTO DAVIS JUNIOR - Orientador
Departamento de Ciência da Computação - UFMG

PROF. FREDERICO TORRES FONSECA
IST - Pennsylvania State University

PROF. JUGURTA LISBOA FILHO
Departamento de Informática - UFV

PROF. CLÁUDIO DE SOUZA BAPTISTA
Centro de Engenharia Elétrica e Informática - UFCG

PROFA. MIRELLA MOURA MORO
Departamento de Ciência da Computação - UFMG

Belo Horizonte, 26 de Março de 2021.

Acknowledgments

First of all, I would like to thank my beloved wife, Ane, who was the person who most believed in me and supported me most on this journey. Thank you for all the friendship, care, and love. We are life partners.

I also thank my parents, Carlos and Luna. Thanks for all the education I received and for the opportunity that you gave me to make my dreams come true. Without your love, none of this would be possible. I also thank my in-laws, Joana and Wylson, for all the care.

I would like to express my deep thanks to my advisor, Professor Dr. Clodoveu A. Davis Jr., for his guidance, encouragement, and patience. All conversations and advice were crucial to the completion of this work.

I would also like to extend my thanks to all the people who, in some way, contributed: Tiago and Michele for the remote support with the LabCS + X laboratory, and Amanda, Alessandro, Levi, Rodrigo, and Herbert, for giving me physical and mental health to continue my studies.

My thanks also go to the Federal University of Ouro Preto (Ufop), which provided me an opportunity to leave my academic attributions for four years to pursue my doctorate.

Finally, I thank all of those who contributed directly or indirectly to the execution of this work.

Resumo

Geoparsing é a tarefa de recuperação de informação geográfica que lida com o reconhecimento das referências a lugares contidas nos textos. Além do geoparsing, duas outras tarefas são usadas para resolver o Problema de Resolução de Escopo Geográfico (PREG), as tarefas de resolução das referências e determinação das referências. O PREG visa determinar o escopo geográfico de documentos, ou seja, os locais ou regiões relevantes, considerando o conteúdo do documento. Vários trabalhos que tratam do PREG ou de suas tarefas focam principalmente o método de solução em si. Além disso, cada trabalho testa o algoritmo usando diferentes conjuntos de dados e fontes de conhecimento externas, como os gazetteers. Esta tese propõe uma metodologia para avaliar os gazetteers ao invés dos algoritmos. A abordagem varia o tamanho e a cobertura dos gazetteers, delimitando-os geograficamente, enquanto mantém o conjunto de dados e os algoritmos fixos. Gazetteers focados podem aumentar a precisão (com baixa perda de recall) na tarefa de geoparsing em comparação com os gazetteers generalistas. Além disso, os gazetteers focados reduzem consideravelmente o número de candidatos ambíguos para cada topônimo encontrado no geoparsing.

Palavras-chave: Recuperação de Informação Geográfica, Problema de Resolução de Escopo Geográfico, Geoparsing, Gazetteers Focados, Geoparsing, Ambiguidade, Precisão.

Abstract

Geoparsing is the geographic information retrieval task that deals with the recognition of references to places contained in texts. Besides geoparsing, two other tasks are used to solve the Geographic Scope Resolution Problem (GSRP), the reference resolution and the grounding references tasks. The GSRP aims to determine the geographic scope of documents, i.e., the locations or regions relevant, considering the document content. Several works that deal with the GSRP or with its tasks focus mainly on the solution method itself. Also, each work test the algorithm using different datasets and external knowledge sources, such as a gazetteer. This thesis proposes a methodology to evaluate the gazetteers instead of the algorithm. Our approach varies gazetteer size and coverage, delimiting it geographically, while keeping the dataset and algorithms fixed. We show that focused gazetteers can increase precision (with low recall loss) in geoparsing compared to generalist gazetteers. We also show that focused gazetteers considerably reduce the number of ambiguous candidates to each toponym found on geoparsing.

Palavras-chave: Geographic Information Retrieval, Geographic Scope Resolution Problem, Geoparsing, Focused Gazetteers, Geoparsing, Ambiguity, Precision.

List of Figures

1.1	GIR System and related components	1
2.1	The Geographic Scope Resolution Problem	8
2.2	Geoparsing Solutions Classification	12
2.3	Reference resolution approaches	16
2.4	Grounding references classification	18
2.5	GeoNames result for Belo Horizonte	20
2.6	GeoNames result for Belo Horizonte on a map	21
2.7	The Getty Thesaurus of Geographic Names result for Belo Horizonte	22
3.1	Proposed methodology workflow	23
3.2	News example from G1 portal	25
3.3	JSON news example	26
3.4	N-gram frequency over toponyms in GeoNames	32
3.5	Website application	34
4.1	Confusion Table	41
4.2	News item from G1 Centro-Oeste sub-section	42
4.3	Belo Horizonte news evaluation	46
4.4	Ambiguous candidates with GeoNames	50
4.5	Ambiguous candidates with country-focused gazetteer	51
4.6	Ambiguous candidates with state-focused gazetteer	51

List of Tables

3.1	News' total from <i>Centro-Oeste</i> , <i>Nordeste</i> , and <i>Norte</i>	27
3.2	News' total from <i>Sudeste</i> and <i>Sul</i> regions	28
3.3	Equivalence between G1 editorial and IBGE's mesoregions	30
3.4	Geoparsing result for a single news text	32
4.1	Validation dataset	36
4.2	Stats of the validation dataset	36
4.3	Stats considering sub state focused gazetteers	37
4.4	Number of Toponyms to each Focused Gazetteers	38
4.5	Volunteered contributions	39
4.6	Belo Horizonte sub state news evaluation	43
4.7	Sub-state gazetteers evaluations	47
4.8	All 504 news evaluations	48
4.9	Number of Toponyms to each Validation Focused Gazetteers	49
4.10	Number of Toponyms considering all dataset	53
A.1	Centro-Oeste sub-state news evaluation	73
A.2	Grande Minas sub-state news evaluation	76
A.3	Sul de Minas sub-state news evaluation	79
A.4	Triângulo Mineiro sub-state news evaluation	81
A.5	Vales de Minas sub-state news evaluation	83
A.6	Zona da Mata sub-state news evaluation	86

Contents

Acknowledgments	ix
Resumo	xi
Abstract	xiii
List of Figures	xv
List of Tables	xvii
1 Introduction	1
1.1 Motivation	4
1.2 Objectives and Contributions	5
1.3 Organization	6
2 Background	7
2.1 Geographic Scope Resolution Problem	7
2.1.1 GSRP solutions	10
2.2 Geoparsing	11
2.2.1 Geoparsing solutions	12
2.3 Reference Resolution and Grounding References tasks	15
2.4 Gazetteers	19
3 Methodology to Evaluate Focused Gazetteers	23
3.1 Preparation	24
3.2 Use of a Focused Gazetteer	29
3.3 Validation	33
4 Analysis and Discussion	35
4.1 Validation Dataset	35

4.2	Human Contributions	39
4.3	Analysis of the Results	40
4.4	Ambiguity Analysis	48
5	Conclusions and Future Work	55
5.1	Future Work	58
	Bibliography	59
	Appendix A Results for G1 Minas Gerais sub-section news	73

Chapter 1

Introduction

The amount of geographic data available and the demand for it, whether for mobile applications or users on the Web, has never been higher [Elwood et al., 2012; Miller and Goodchild, 2015; Singleton and Arribas-Bel, 2021]. Moreover, a substantial share of the information available on the Web is geographically specific [Aloteibi and Sanderson, 2014; Delboni et al., 2007; Sanderson and Kohler, 2004; Vaid et al., 2005; Vasardani et al., 2013]. Hu and Adams [2020] even emphasize that geographic data harvested from unstructured texts have unique merits as it reflects real-time situations or records essential historical information.

One valuable resource that can support such growth and availability is geoparsing, i.e., the task of recognizing references to places in digital text documents. Automatically performing this task is crucial but complex, and besides that, addressing this problem is one of the pressings tasks for Geographic Information Retrieval (GIR) research [Purves et al., 2018]. Figure 1.1 shows a conceptual schema of a generic GIR system with some related components.

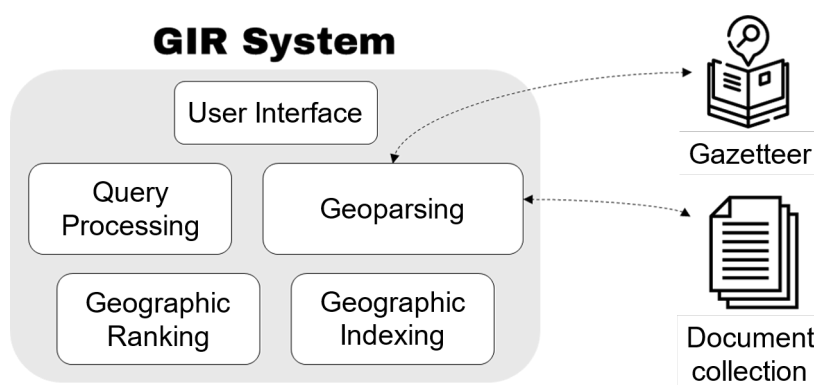


Figure 1.1: GIR System and related components

GIR is an extension of Information Retrieval [Baeza-Yates and Ribeiro-Neto, 2011] to handle geographic locations and metadata [Jones and Purves, 2009], taking it beyond the use of keywords. GIR studies methods and techniques for information retrieval from unstructured or partially structured sources, including relevance ranking, based on queries that specify both theme and geographic scope [Jones and Purves, 2008, 2009].

A GIR system usually has a user interface and functions or modules such as query processing, geographic ranking, geographic indexing, and geoparsing. Moreover, a GIR system may use other elements such as document collection and an external knowledge source. In Figure 1.1, a gazetteer represents the external knowledge source. A gazetteer contains structured information about place names, associating the place names with their geographic location, besides other descriptive information [Goodchild and Hill, 2008; Purves et al., 2018]. GeoNames¹ and The Getty Thesaurus of Geographic Names² are examples of global gazetteers. This thesis uses GeoNames as the primary gazetteer, creating focused gazetteers from it. Focused gazetteers are gazetteers whose contents cover a definite geographic region, for instance, a country, a state, or a city.

Besides geoparsing for references to places in texts, there are efforts to identify these references in other media, such as photos and videos [Luo et al., 2011], including implicit references. With documents accurately and efficiently linked to places mentioned in their content (directly or indirectly), it becomes possible to improve and innovate in areas such as geographic indexing and geographic querying [Yadav et al., 2020]. Moreover, it becomes feasible to obtain relationships based on spatial proximity or containment and detecting localized trends for events and phenomena mentioned in social media.

References to geographic locations in texts can appear in many forms. The most straightforward form is geographic coordinates (latitudes and longitudes values). Additionally, postal addresses (zip code in the United States, or CEP³ in Brazil) and telephone numbers (the country code +55 refers to Brazil, and Brazilian DDD⁴ codes indicate Brazilian states) are references to geographic locations. Historical dates (September 11 refers to New York), demonyms (Japanese refers to Japan and Ottawans point to Ottawa city, in Canada), and typical food (wine and baguette remit to France) are indirect references to places. Even non-spatial words such as *mountain* and *beach* can indicate locations in a given context [Adams and Janowicz, 2012]. Often, humans

¹<http://www.geonames.org/>

²<http://www.getty.edu/research/tools/vocabularies/tgn/>

³Postal Addressing Code is a code created by postal administrations to facilitate the logistical organization and spatial location of an address

⁴ DDD are codes adopted for long distance dialing

recognize these references, but correlation does not come so directly to automated systems.

However, one of the most common form to reference locations in texts are the toponyms. In short, a toponym is a name or a label to a particular place. This thesis considered only toponyms as references to geographic locations. A broader discussion about the concept of toponym can be found in Gritta et al. [2019].

Many queries also include toponyms and other geographic terms [Delboni et al., 2007; Sanderson and Kohler, 2004; Silva et al., 2006]. Hence, there is a demand for mechanisms to search for documents both thematically (for instance, using a set of keywords) and geographically, based on places mentioned or referenced by the text [Zong et al., 2005]. Similar techniques and resources can also apply to stream data, such as Twitter messages [Di Rocco et al., 2020] or RSS feeds, providing the opportunity to index content in near-real-time, based on place references.

These references to places, specially toponyms, can be ambiguous and uncertain since not all of them are straightforward and distinct as geographic coordinates. Two or more places around the globe can share the very same toponym⁵. Also, a toponym can have an identical spelling as common language words⁶ and proper names⁷. The first type of ambiguity occurs when the same toponym references multiple places, the Geo/Geo ambiguity. The latter type is called Geo/Non-Geo ambiguity, which occurs when both a location and a non-location share the same name [Amitay et al., 2004]. Besides those, Clough et al. [2004] suggest reference ambiguity, which occurs when different names reference the same place. For instance, the New York city is also known as *NYC* and *The Big Apple*.

Geoparsing is a task that helps to solve the Geographic Scope Resolution Problem (GSRP). In short, the objective of the GSRP is to assign a set of places referenced by and relevant to the document’s content. To Andogah et al. [2012], “every document has a geographical scope” and, even keyword queries used in search engines can have a geographic scope [Alexopoulos and Ruiz, 2012; Silva et al., 2006] since query words embed the user’s intentions in the search.

Besides the geoparsing, GSRP solutions need two other tasks: reference resolution and grounding references. The geoparsing aims to find geographic references present in the text. The vast majority of works in the literature only address searching for toponyms. The reference resolution task, if necessary, disambiguate these toponyms

⁵Paris, besides being the capital of France, is also the name of sixty other places, according to GeoNames

⁶Park, Hope, and Independence are American cities

⁷Washington and Virginia are American states

giving them a unique distinct location. The grounding reference task uses each disambiguated toponyms to generate a geographic scope (footprint) to the document.

GSRP solutions (or individual task solutions) often need an external knowledge source. Each knowledge source has limitations, such as size, language, structure, or range. This thesis focuses only on the geoparsing task and gazetteers as an external knowledge source.

1.1 Motivation

Algorithms, datasets, and gazetteers have a direct impact on the solutions for the GSRP. Several works that deal with the GSRP or individual tasks focus mainly on the solution method itself. These works include experiments varying the proposed algorithms, the dataset, and the gazetteers, making it difficult to compare different solutions.

Since algorithms, datasets, and gazetteers impact GSRP (and its tasks), one of the first questions to come up is, how can the impact of just one of these factors be estimated? Then, a second question, why not to analyze the gazetteer's influence, instead of proposing a new solution method?

The idea was to show the influence of adopting different gazetteers in the geoparsing task. This idea motivated the proposal of a methodology approach that did not vary all three elements but just one. Keeping the geoparsing algorithm solution and datasets fixed, the question was how to evaluate the gazetteer, changing it in size and scope?

It is reasonable to assume that broader and comprehensive gazetteers might provide more elements to recognize references to place names in the text. However, they might generate more ambiguity in the reference resolution task (i.e., a term will be equal to multiple entries in the gazetteer). Still, a more restricted and small size gazetteer may help to avoid ambiguity. But a smaller gazetteer can miss some toponyms in the geoparsing task (i.e., the candidate string from the text will not match any entry in the resource). Then, how large is the influence, in numbers, of the gazetteer in a solution in the geoparsing task?

Furthermore, the geotagged datasets used in experiments for different works, in general, are not freely available to use [Gritta et al., 2018]. Several of these works create datasets for the experiments, while others use some paid or benchmarking datasets. Thus, a second motivation was to create a freely available dataset to be used in the proposed methodology.

1.2 Objectives and Contributions

The main objective is to show that a focused gazetteer improves the association of places to documents in the geoparsing task. A focused gazetteer means a gazetteer limited by the expected geographic scope of the data. In other words, limiting the gazetteer enhances the results in relating places to documents. For instance, news texts can have a primary geographic scope associated with them, depending on the news portal's section that holds the news. With this geographic information, it is possible to generate focused gazetteers, limiting them based on the news dataset geographic focus.

This thesis discusses the hypothesis that a focused gazetteer can provide increased precision in the geoparsing task while generating less ambiguity in the process. A focused gazetteer can find fewer ambiguous candidates for the reference resolution task compared to a complete gazetteer. This work proposes a methodology that changes the size and the scope of the gazetteer (with focused gazetteers) while keeping algorithms and datasets unchanged. This methodology approach suits verifying the thesis hypothesis. Also, this work proposes a framework for manual validation by humans.

Besides, this research offers the following contributions:

- A survey about the GSRP and its tasks. The survey organizes consistent terminology related to this problem. Also, it contains a classification of the solutions present in the literature for the GSRP and its tasks [Monteiro et al., 2016];
- A methodological approach to evaluate gazetteers used in the geoparsing task. This methodology compares different gazetteers while keeping the method and the dataset fixed;
- A dataset with 529,585 news in Portuguese. All news items are associated with a primary geographic scope;
- A small Portuguese news dataset (504 news) with places annotated by people. This kind of dataset is valuable to compare the quality of the results of geoparsing algorithms or even to compare different solutions to the geoparsing;
- A well-defined approach to evaluate the influence of focused gazetteers in associating places to text, specifically in the geoparsing task;
- A confirmation that focused gazetteers increase the precision in the geoparsing task while generating less ambiguity. Precision, recall, and F1 score support the potential ambiguity reduction found.

1.3 Organization

The structure of this document is as follows:

- Chapter 2 sweeps the terminological variations to the GSRP and presents established concepts and terms for the remainder of this document. It also describes the problem and its tasks, specifically, the geoparsing. It contains the main application areas related to and classification on solutions to geoparsing. Also, it presents a solution classification to the other tasks, and to the GSRP, according to the algorithm used.
- Chapter 3 details the methodology employed to evaluate a focused gazetteer. Specifically, the chapter presents (1) the creation of a raw Portuguese news dataset; (2) the use of a naive lookup-based geoparsing method with different focused gazetteers; (3) the process to validate the results found in (2) by contributions made by people.
- Chapter 4 shows the steps taken to analyze the results obtained. Precision and recall metrics are used to assess gazetteers focused on the Geoparsing task. Also, the potential of focused gazetteers about ambiguity is discussed and analyzed.
- Chapter 5 presents the conclusions and final considerations, and highlights future works.

Chapter 2

Background

This chapter presents the necessary concepts for understanding this thesis. Firstly, Section 2.1 overviews the GSRP, indicating all related components. Also, this section discusses GSRP solutions. Section 2.2 presents the geoparsing task with several solutions classified by the approach used.

A brief overview of the reference resolution and grounding references tasks is in Section 2.3. Lastly, Section 2.4 formally describes the gazetteers. A large part of this chapter is also in the survey “A survey on the Geographic Scope of Textual Documents” [Monteiro et al., 2016].

2.1 Geographic Scope Resolution Problem

The Geographic Scope Resolution Problem [Alexopoulos and Ruiz, 2012; Andogah et al., 2012; Alexopoulos et al., 2013] consists in discovering places related to the contents of a textual document, disambiguating them if necessary, and using the resulting set of unique places to build the overall geographic scope. In the literature, the GSRP is known under other names, with equivalent definition, except for terminological differences: *Place Name Assignment Problem* [Zong et al., 2005; Amitay et al., 2004] or *GeoReferencing* [Gouvêa et al., 2008; Zubizarreta et al., 2008].

Although references to places can be of various forms (as mentioned in Chapter 1) and the external knowledge base can be databases or ontologies, this thesis uses only toponyms as references to places and gazetteers as a knowledge source. A toponym is a general name for any location or geographical entity¹ or a name for a topographical feature [Gritta et al., 2018]. Different places around the Earth can share the same

¹According to The United Nations Conference on the Standardization of Geographical Names: <http://unstats.un.org/unsd/geoinfo/UNGEGN/>

name, and they can share the same name with people and other entities [Jones and Purves, 2008; Habib and van Keulen, 2013]. Leidner [2007, 2008] presents an extended discussion on place names, including a historical perspective about them. In short, gazetteers are dictionaries of toponyms [Hill, 2000, 2006]. Section 2.4 expands this concept of gazetteers.

Formally, the geographic scope of documents, or $GS(D)$, is the set of places associated with the content of a document D [Monteiro et al., 2016; Andogah et al., 2012; Buyukkokten et al., 1999; Ding et al., 2000]. The $GS(D)$ considers the places mentioned in the document but does not necessarily need to admit all of them. In its most simple form, the $GS(D)$ corresponds to a set of coordinates of places mentioned in the document. However, the $GS(D)$ might represent these places more broadly, if necessary. For instance, a document D mentioning *Minas Gerais* cities such as *Ouro Preto*, *Mariana*, *Diamantina*, and *Tiradentes* can have a $GS(D) = Minas Gerais$ or $GS(D) = Estrada Real$ ². *Geographic document footprint* [Fu et al., 2005; Silva et al., 2006], *geographic path* [Vargas et al., 2012b], and *geographic focus* [Amitay et al., 2004; Chen et al., 2010; Zubizarreta et al., 2008] are equivalent terms to geographic scope.

Figure 2.1 shows a system that solves the GSRP schematically.

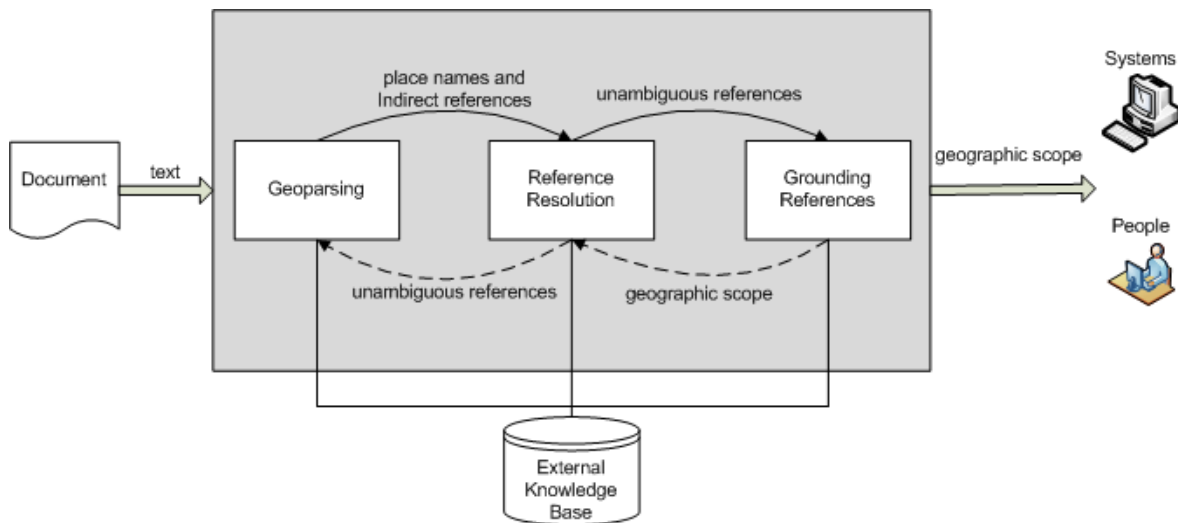


Figure 2.1: The Geographic Scope Resolution Problem

The input is a document or a set of documents, and the output is the geographic scope. Solutions for the GSRP need to execute three tasks over this input. The first one, the geoparsing, identifies the toponyms in the document's content. Then, if necessary, reference resolution algorithms disambiguate these toponyms, indicating a

²The name "Estrada Real" refers to any land route that, at the time of Colonial Brazil, was covered in the process of settlement and economic exploitation

unique place for them. Lastly, the grounding reference creates the geographic scope of the document using these disambiguated references.

In Figure 2.1, the white rectangles indicate the geoparsing, reference resolution, and grounding references tasks. Solid arrows indicate the direct sequence of steps to solve the GSRP. Dashed arrows mean that a later step result can contribute to solving a previous one in further iterations. For instance, a disambiguated toponym can contribute to finding other toponyms. Or the geographic scope of a document can help to disambiguate the toponyms. There are studies that show that the geoparsing and the reference resolution task can be highly dependent³ [Habib and van Keulen, 2011], and Andogah et al. [2012] used a preliminary geographic scope of texts to help with disambiguation. Geoparsing and grounding references are necessary to solve the GSRP. Despite not being mandatory, toponym ambiguity demands executing the reference solution task [Amitay et al., 2004].

The output of a system that solves the GSRP is the geographic scope of the document. Tools and techniques from different areas can benefit from this output, for instance, information retrieval (IR) and web data mining.

IR techniques as geographic indexing, filtering, and ranking can use the geographic scope to become more efficient. Documents with equal $GS(D)$ can be indexed together, and the $GS(D)$ can help find the most suitable physical location for documents in a distributed storage infrastructure [Lieberman et al., 2010; Vaid et al., 2005]. Still, the $GS(D)$ can expand queries using or toponyms topologically-related or places that belong to the same subdivision hierarchy [Andogah et al., 2012; Delboni et al., 2007; Machado et al., 2011; Moura and Davis Jr., 2014]. The geographic scope can contribute to search engines, helping them recognize the toponyms in queries to improve their results [Cardoso, 2011; Delboni et al., 2007; Martins et al., 2007].

Moreover, techniques of geographic filtering and document classification can improve using the geographic scope. The $GS(D)$ can provide criteria for filtering documents according to the users' location [Lieberman and Samet, 2012b; Ribeiro Jr. et al., 2012]. Besides, the $GS(D)$ can serve as criteria for classifying or grouping documents. This classification makes it possible to create spatially-aware services such as map-based news selection applications [Alencar et al., 2010; Alencar and Davis Jr., 2011; Morimoto et al., 2003; Teitler et al., 2008; Silva et al., 2006].

Next, Section 2.1.1 describes solutions that resolve the GSRP as a whole.

³This is called "reinforcement effect".

2.1.1 GSRP solutions

Initial works infer the geographic scope using the physical location of the web server that hosts the document [Buyukkokten et al., 1999] or using the estimated localization of the reader of the text [Wang et al., 2005]. Even though such correspondence may exist in many cases, nothing keeps a Web server from hosting content that is unrelated to its place. Hybrid approaches tried to correct this problem using server location and document contents [Ding et al., 2000; McCurley, 2001]. More recent solutions tend to use the content of the documents. Besides the toponyms, some works solve the GSRP finding other reference types such as urban addresses, telephone, and postal area codes [McCurley, 2001; Borges et al., 2011, 2007].

The geoparsing for the references to places uses from own techniques to third-party tools. Zubizarreta et al. [2008] use a controlled dictionary in geoparsing and heuristics to resolve the ambiguity in the reference resolution step, while Silva et al. [2006] use a NER (Named Entity Recognition) method plus a graph-ranking algorithm for the same two tasks. Zong et al. [2005] use third-party software to perform the geoparsing and a rule-based approach to disambiguate the toponyms. Graphs [Zubizarreta et al., 2008; Silva et al., 2006] or trees [Zong et al., 2005] serve to structure the geographic scope.

Vargas et al. [2012a,b], Zhang et al. [2012], and Andogah et al. [2012] also have a multistage method. Both Vargas et al. [2012a,b] and Zhang et al. [2012] use a third-part tool in the geoparsing task. However, they use a different strategy in the reference resolution task. Vargas et al. [2012a,b] use GIS functions, while Zhang et al. [2012] run a disambiguation procedure, GeoRank, which adapts PageRank [Page et al., 1999] to solve the Geo/Geo ambiguity, and heuristics to solve the Geo/Non-Geo. Andogah et al. [2012] first find the geographic scope, then applies the reference resolution task.

Some works structure the process differently or start with a definite set of assumptions and boundary conditions. The K-Locator system needs the document domains and a comprehensive ontology covering these domains [Alexopoulos et al., 2013]. Borges et al. [2011, 2007] geoparses indirect references such as urban addresses and their components: street names, telephone area codes, urban landmarks, and postal addresses, using an ontology and regular expressions.

Many works that solve the GSRP outputs only a single integrated scope from the set of geoparsed references [Amitay et al., 2004; Campelo and Baptista, 2008; Chen et al., 2010]. The outputted $GS(D)$ uses a hierarchical structure built from relationships obtained using a gazetteer. Amitay et al. [2004] present some heuristics to resolve references based on the principle “one sense per discourse”. This principle

states, if an ambiguous toponym appears more than once in a text, all duplicates should correspond to the same place.

A significant issue is the lack of an established benchmark to compare solutions approaches for GSRP. Each article defines its dataset, algorithms, knowledge source, and comparison methodology. Therefore, there is currently no direct way to establish which approaches are more efficient. Empirically, Anastácio et al. [2009] tried to compare some proposals. They analyze the Web-a-Where system [Amitay et al., 2004], the spatial overlap-based method proposed in the GIPSY project [Woodruff and Plaunt, 1994], the graph-based algorithm (the GREASE project) [Silva et al., 2006], and three simple baseline methods. They concluded that the Web-a-Where system achieved the best results, closely followed by the GraphRank [Silva et al., 2006], and by the baseline based on the most frequently occurring place. Nevertheless, there are challenges to use the comparison methodology used by Anastácio et al. [2009] more broadly. In general, it is necessary to modify the GSRP solutions to fit them into the Anastácio et al. [2009] methodology.

In summary, the firsts GSRP solutions aimed to define the geographic scope based on the infrastructure that hosted the documents and web pages. Actual works use the references present in the text and follow multistage solutions (Figure 2.1) to solve the GSRP. A broader discussion over these solutions is in [Monteiro et al., 2016]. Several solutions tend to concentrate on a single task that compounds the GSRP due to the increasing complexity. The focus on solutions to part of the GSRP allows for more compartmentalized solutions, with sets of techniques directed at each part of the problem. Section 2.2 details the geoparsing task.

2.2 Geoparsing

This thesis mainly focuses on the geoparsing task. Geoparsing is the first required task in solving the GSRP. The objective is to find all references to places in a text document. These references can be direct as a toponym or indirect ones such as people, dates, and other elements linked to geographic locations. Indirect references are called *implicit geographic evidence* [Cardoso et al., 2008] or *location indicators* [Leveling et al., 2006]. They are urban addresses, references to related entities or landmarks, nicknames, or even sets of coordinates. Some works use other names such as *toponym recognition* or *toponym extraction* [Habib and van Keulen, 2013; Jones and Purves, 2008].

The geoparsing task is a specialized version of the Named Entity Recognition problem [Nadeau and Sekine, 2007]. NER intends to identify the entities (people,

places, organizations) mentioned in natural language sentences. So, it is possible to use some NER algorithms to solve the geoparsing, seeking just for locations. In addition to NER techniques, other solutions use lookup methods, matching candidate terms to a gazetteer. Also, some solutions rely on heuristics or rules to find all references to places.

Among the proprietary and open-source tools used to solve geoparsing are C&C tagger⁴, Apache’s OpenNLP⁵, OpenCalais⁶, Yahoo!’s Placemaker⁷, Sheffield’s GATE⁸, Ling-Pipe⁹, Stanford NER¹⁰, and spaCy NER¹¹. Section 2.2.1 describes proposed solutions to the geoparsing task.

2.2.1 Geoparsing solutions

According to [Monteiro et al., 2016], there are three solution approaches to solve the geoparsing task: *lookup-based*, *rule-based*, and *supervised* ones. Also, each solution can be language-dependent or language-independent. Supervised methods are generally language-independent, while most other methods are language-dependent. Some authors consider only two approaches for geoparsing: *machine learning*, and *rule-based* [Habib and van Keulen, 2013], instead of the three defined in [Monteiro et al., 2016; Leidner and Lieberman, 2011]. Figure 2.2 shows the geoparsing solution approaches.

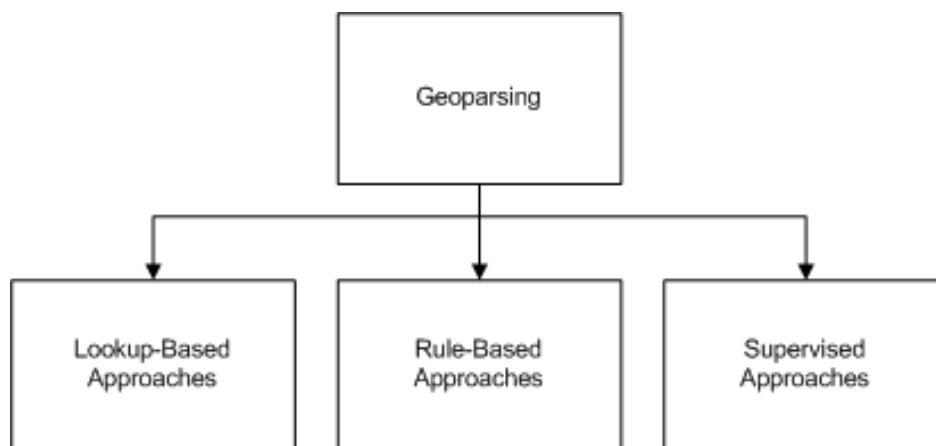


Figure 2.2: Geoparsing Solutions Classification

⁴<http://svn.ask.it.usyd.edu.au/trac/candc/wiki/Taggers>

⁵<https://opennlp.apache.org/>

⁶<http://www.opencalais.com/>

⁷<http://www.programmableweb.com/api/yahoo-placemaker>

⁸<https://gate.ac.uk/>

⁹<http://alias-i.com/lingpipe/>

¹⁰<http://nlp.stanford.edu/software/tagger.shtml>

¹¹<https://spacy.io/>

Solution approaches *lookup-based* analyze each document, matching each candidate word (or set of words) against a gazetteer¹². The quality of the gazetteer has a direct impact on the quality of results. Incomplete gazetteers can miss many references to places, and broader ones might generate more ambiguity. *Lookup-based* methods are language-dependent if the gazetteer is likewise language-dependent. Multilingual gazetteers can offer language independence.

Several works also use heuristics to identify candidates and then compare them to an external knowledge source [Alexopoulos et al., 2013; Amitay et al., 2004; Chen et al., 2010; Clough, 2005; Pouliquen et al., 2006; Purves et al., 2007; Zubizarreta et al., 2008]. There are also more complex proposals using a more varied strategy applied for obtaining candidates. Borges et al. [2011, 2007] use a gazetteer and an ontology plus regular expression to look for toponyms and indirect references. Similarly, Shi and Barker [2011] combine gazetteers and linguistic heuristics¹³ to geoparse for elements as provider location and domain, markup components with coordinates, indirect references (postal codes, phone numbers), and toponyms. Di Rocco et al. [2020] uses tweets with a geographic scope of city granularity to geoparse toponyms at the sub-city level, and Nizzoli et al. [2020] use a reference knowledge graph to geoparsing in free text and extract the geographic coordinates.

Rule-based approaches use heuristics and a set of symbolic rules, such as regular expressions or context-free grammars. Both heuristics and rules are encoded in a domain-specific language, resulting in, in most cases, language-dependent methods. Woodruff and Plaunt [1994] pioneered using a rule-based solution to geoparse texts, looking for place names near lexical constructs that indicate spatial position. Silva et al. [2006] and Lieberman and Samet [2011] use a part of speech (POS) tagger to find proper nouns since toponyms tend to start with capital letters. These works are language-dependent, built for texts in English. Twaroch et al. [2008] also use regular expressions and filters to find toponyms. Still, Pouliquen et al. [2004] present a rule-based approach for geoparsing multilingual texts. First, they discover the text language. Then, they use a set of regular expressions to find country and city names.

Supervised methods use training annotated corpus containing the text associated with the expected set of related places. In general, these training datasets need to be extensive and balanced to achieve good results [Di Rocco et al., 2020]. The annotated dataset train the solution algorithms, using features such as infrequent strings, length, capitalization, and other. When the training is complete, the algorithms run over non-labeled data and compare the same training features. The most common *supervised*

¹²Or using any other external knowledge source, such as ontologies or databases

¹³Prepositions near the toponym and spatial relationship terms

algorithms used in geoparsing are the Support Vector Machine (SVM), the Hidden Markov Model (HMM), and the Conditional Random Field (CRF) [Lieberman and Samet, 2011].

SVM is a machine learning algorithm used in data mining tasks, such as classification and regression analysis. The SVM model maps points in space to represents classified examples. In this way, a hyperplane separates instances of different categories. Then, this same space can map new cases and predict their classes [Hearst, 1998]. SVM can work with two (binary SVM) or more (multi-class SVM) classes. Another example of a machine learning classifier is HMM, which is a sequence classifier. More precisely, HMM is a ubiquitous tool for representing probability distributions over sequences of observations [Fujiwara et al., 2009]. HMM classifies single objects considering their characteristics in the neighborhood. The CRF approach is also a supervised learning technique that builds a statistical model using sequence data. It uses an undirected graphical model that defines a log-linear distribution over sequences, given a particular one.

Chasin et al. [2013] compare an SVM approach, an HMM method (implemented by the LingPipe library), and the Stanford NER to solve geoparsing. A Google Geocoder¹⁴ lookup determines whether each candidate was a toponym. Habib and van Keulen [2013] use HMM and SVM in geoparsing from a set of holiday home descriptions. HMM, trained using manual annotations, is used to extract candidate toponyms from the document. Candidates are then matched against the GeoNames, generating two sets of features (positive and negative candidates) that train an SVM classifier. Then, SVM is used as a reference resolution technique, reinforcing the results from the original extraction.

Works such as Gelernter and Mushegian [2011]; Gelernter et al. [2013]; Hu et al. [2019]; Karimzadeh et al. [2013]; Lieberman et al. [2010] use some NER tools (such as Stanford NER, OpenCalais, or spaCy NER) as part of their workflows to solve geoparsing. In a different approach, Nissim et al. [2004] use the Curran and Clark maximum entropy tagger [Curran and Clark, 2003] to geoparse in historical descriptions of Scotland. The tagger, as a supervised approach, is trained using 10-fold cross-validation on an annotated dataset. The method shows significant improvement in precision, recall, and f1 score over a lookup-based that used a custom-built Scottish gazetteer.

Results from *lookup-based* methods are highly dependent on the external knowledge sources, and each proposal potentially uses a different one, including custom-built

¹⁴<https://developers.google.com/maps/documentation/geocoding/>

ones. *Rule-based* approaches use resources such as regular expressions and linguistic heuristics, custom-built in many cases, with language variations and adaptations tuned to a particular problem. *Supervised* methods require labeled training data, and no standard for comparing the performance of methods has emerged so far. Notice, however, that rule-based and supervised approaches also use external knowledge sources. Thus, their results are also dependent on the quality of these knowledge sources.

In summary, most geoparsing solutions fall into *lookup* and *rule-based* approaches. And these approaches frequently are language-dependent. More recent solutions usually use training to their *supervised* methods. Also, these methods easily support language independence. A detailed classification over geoparsing solutions is in [Monteiro et al., 2016].

The thesis focus is to analyze the performance of the geoparsing using focused gazetteers. In this way, Section 2.3 briefly explains the two other tasks involved in GSRP solutions. And Section 2.4 details the concept of gazetteers.

2.3 Reference Resolution and Grounding

References tasks

Reference resolution is the process of mapping a toponym or a reference to unambiguous identification of the place. This task is mandatory to solve the GSRP whenever the data contain ambiguities, which is a common problem with toponyms. As with geoparsing, other works give this task different names such as *toponym resolution*, *toponym disambiguation*, or *geographical entity resolution* [Habib and van Keulen, 2013; Alexopoulos and Ruiz, 2012; Li et al., 2002].

According to Habib and van Keulen [2013], around 46% of the toponyms found in GeoNames refer to more than one place. For instance, Springfield toponym corresponds to more than one hundred and eighty world geographic locations, including cities in the U.S. and Jamaica, an Australian park, and a New Zealand district. The reference resolution task usually relies on an external geographic resource, such as a gazetteer. Toponyms and other references are used as keywords to search these resources for candidate places. With these candidates and other evidence, the algorithms decide which location is more likely to correspond to the reference. The quality of the results in this step is dependent on the quality and coverage of the external knowledge source. Reference resolution approach solutions classify as *map-based*, *knowledge-based*, and *supervised*. Figure 2.3 shows the proposed classification for the reference resolution task [Monteiro et al., 2016].

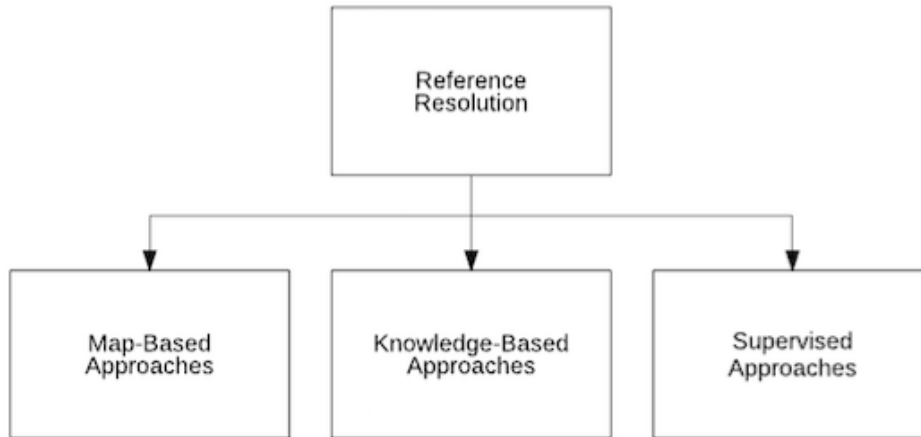


Figure 2.3: Reference resolution approaches

Map-based methods use geometric algorithms or topological functions, such as *disjoint*, *union*, and *intersection*. These solutions are very similar to each other. They usually distinguish themselves only on the geometric algorithm or topological function used. Shi and Barker [2011] disambiguate using resolved toponyms closest to an ambiguous toponym. Smith and Crane [2001] use the centroid concept and distance to this centroid in a map to disambiguate all references. Vargas et al. [2012a,b] use a similar function based on a polygon containing the unambiguous entities, and Zong et al. [2005] use a gazetteer supporting the reference resolution task. Leidner et al. [2003] define a *disambiguating context* as a region in whose confines most unresolved toponyms become unique.

A match against a gazetteer represents the simplest example of the *knowledge-based* approach [Chen et al., 2010; Pouliquen et al., 2004; Rauch et al., 2003; Olligschlaeger and Hauptmann, 1999]. Knowledge-based rely on toponym relations extracted from gazetteers or other external knowledge sources such as Wikipedia Habib and van Keulen [2013]. Works that entrust on heuristics and custom-built rules are also knowledge-based approaches [DeLozier et al., 2015]. Amitay et al. [2004] use several heuristics for each ambiguous toponym (tokens in the vicinity, largest population, and one reference per discourse principle). Some works assign scores combined with some heuristics to disambiguation candidates [Alexopoulos et al., 2013; Li et al., 2006; Silva et al., 2006; Volz et al., 2007]. Other highlights solutions are the GeoRank [Zhang et al., 2012], which uses the same PageRank voting process, the works of Clough [2005], and Purves et al. [2007] that explore the hierarchy of the external knowledge sources to provide a default sense to an ambiguous toponym. Monteiro et al. [2016] made a detailed

classification over reference resolution solutions.

Supervised approaches are those based on standard machine learning techniques. They follow the usual machine learning methodology: build a training set composed by disambiguated toponyms, and then run a machine learning algorithm, such as Naïve Bayes classifiers [Smith and Mann, 2003], Bayesian classifiers [Adelfio and Samet, 2013], Random Forests [Lieberman and Samet, 2012a], clustering methods [Habib and van Keulen, 2012], co-occurrence model [ju2016], or a combination of multiple learning features. Santos et al. [2015]. Santos et al. [2015] and Speriosu and Baldrige [2013] use indirect supervision, while Garbin and Mani [2005] used a gazetteer-based statistical classifier.

Buscaldi and Rosso [2008] compared *map-based* and *knowledge-based* approaches. They conclude that the *knowledge-based* approach was better with a small context, such as a sentence or a paragraph, and the *map-based* obtained the best results when the context was the whole document. The number of works that deal with the reference resolution problem is high. Most solutions require an external knowledge source. As a result, the quality of the knowledge source is crucial to achieving better results. There are some efforts in creating and enriching these external knowledge sources. Machado et al. [2011] proposed an ontological gazetteer that includes geographic elements such as spatial relationships, concepts, and terms related to places. Moura and Davis Jr. [2014] create a gazetteer from linked data sources and found several issues with the data quality in such reference databases.

The grounding references task is the process of mapping all location references to a geographic scope, which may be a set of latitude and longitude coordinates or polygons set representing geographic boundaries. There are granularity problems when a document references multiple locations. For instance, if the document mentions neighboring cities, the geographic scope can be a set of cities or a single region containing these cities. It is the geographic scope form that classifies the grounding reference solutions, not the algorithms and techniques. There is a further classification on *multiple place* geographic scope. This scope can be *structured* (using a data structure) or *non-structured*.

Methods that inform the geographic scope as a *single place* include techniques that consider the most representative place to be the geographic scope. It also can use generalization, grouping the locations found in the document in a single place higher in a geographic or administrative hierarchy [Amitay et al., 2004; Borges et al., 2011; McCurley, 2001; Ding et al., 2000; Buyukkokten et al., 1999; Woodruff and Plaunt, 1994]. Silva et al. [2006] and Wang et al. [2005] use a data structure to calculate the importance of the toponyms identified in the geoparsing task, considering only the

most representative toponym found as the geographic scope. Figure 2.4 presents the organization proposed by Monteiro et al. [2016].

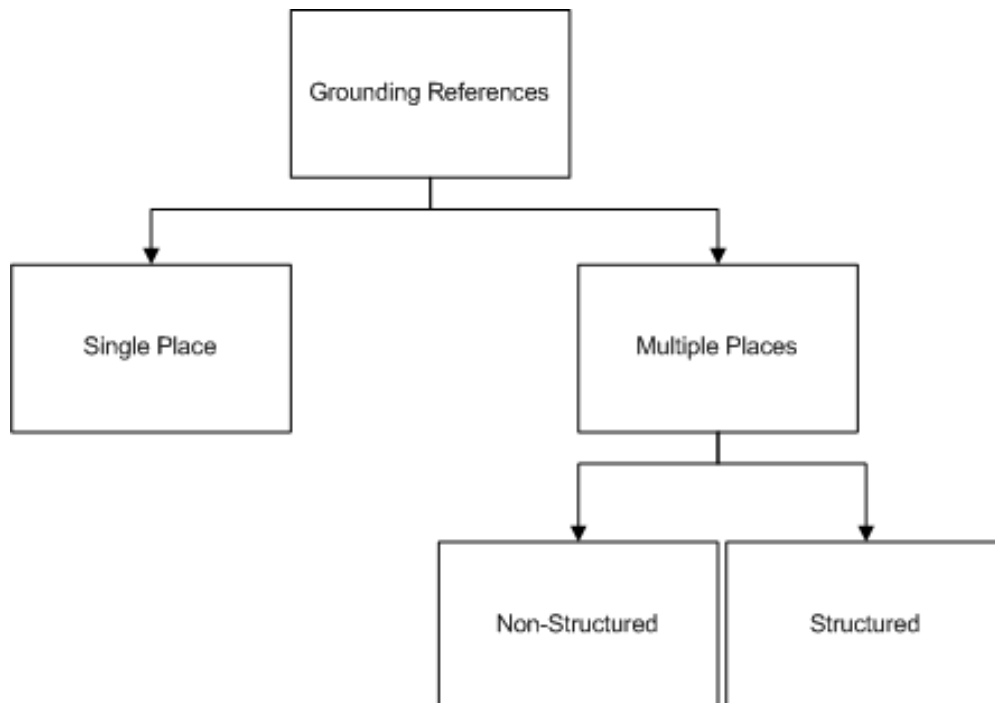


Figure 2.4: Grounding references classification

Several solutions inform the geographic scope considering *multiple places* and define some data structure to organize them, such as a tree [Zhang et al., 2012; Chen et al., 2010; Campelo and Baptista, 2008] or a graph [Zubizarreta et al., 2008]. Tree data structures represent a hierarchy of spatial subdivisions that parent nodes spatially contain their children. In the initial tree data structure levels, it is possible to go from individual references to a single all-encompassing one that serves as a general result. Andogah et al. [2012] propose a particular data structure based on the assumption that places of the same type or under the same administrative jurisdiction or adjacent to each other are more likely to be mentioned in a given discourse unit. Another alternative is a simple list of locations related to the text [Alexopoulos et al., 2013; Vargas et al., 2012b; Zong et al., 2005], in no particular ordering, which corresponds to a *non-structured* geographic scope.

There is a confusing nomenclature regarding the grounding references. While some terms are very similar such as grounding and localization [Amitay et al., 2004], others may mean different tasks in similar fields. For instance, the *geotagging* used by Lieberman and Samet [2011] is the same as the grounding references mentioned. However, usually, the term *geotagging* means a process of creating tags that allow linking

the document (or other types of Web objects, such as photos or videos) to a location or set of locations [Amitay et al., 2004; Teitler et al., 2008]. Another erroneously used term is *geocoding*. The most common meaning is the process of locating points on the surface of the Earth based on alphanumeric address information [Davis Jr. and Fonseca, 2007]. More broadly, *geocoding* means the location of places based on any textual description [Goldberg et al., 2007].

According to Monteiro et al. [2016], grounding reference solutions are equally distributed on the *single place*, *multiple places non-structured*, and *multiple places structured*.

2.4 Gazetteers

This section describes the gazetteers, considering the relevant aspects to the development of this thesis. A gazetteer is a repository of georeferenced toponyms. This kind of knowledge resource usually includes more than just place names. It has further information, such as the type or class of the location, the geographic coordinates¹⁵ (pair of latitude and longitude), and some conceptual or territorial hierarchy [Hill, 2006; Leindner et al., 2003]. In other words, gazetteers are dictionaries of toponyms [Hill, 2000, 2006].

More formally, a gazetteer is a collection of entries. Each one contains, at a minimum, the tuple $(N; T; G)$ where N is a toponym, T is a typing scheme of categories for places, and G represents coordinates indicating a point, line, or areal extent. For instance, a gazetteer may contain the tuple $(Belo Horizonte; city administrative level; (-19^\circ 55', -43^\circ 56'))$ to point to the city of *Belo Horizonte* with a $-19^\circ 55'$ latitude and $-43^\circ 56'$ longitude. Applications often require relationships between these entries [Hill, 2009].

GeoNames, the Columbia Gazetteer of the World Online¹⁶, and The Getty Thesaurus of Geographic Names are three broader generalist worldwide gazetteers. Other gazetteers are more restricted geographically than GeoNames. The Geographic Names Information System¹⁷ (GNIS) contains information about two million physical and cultural geographic features in the United States. And the National Street Gazetteer¹⁸ (NSG) has records of all England and Wales streets.

¹⁵Also called the *footprint* of the toponym

¹⁶<http://www.columbiagazetteer.org/main/Home.html>

¹⁷<https://geonames.usgs.gov/index.html>

¹⁸<https://www.geoplace.co.uk/addresses-streets/street-data-and-services/national-street-gazetteer>

GeoNames is a partly autocratic and partly crowdsourced gazetteer with over 25 million geographic toponyms, including features such as populated places and alternate names [Di Rocco et al., 2020]. GeoNames data are accessed by web services¹⁹ or by database export²⁰. Figure 2.5 shows the top 15 results for *Belo Horizonte* toponym using the GeoNames website.

Name	Country	Feature class	Latitude	Longitude
Belo Horizonte B.H., BH, BHZ, Bel Horizonte, Belo Horizonte, Belo Horizonte, Belo Hte, Belo Horizonte, Bel...	Brazil, Minas Gerais Belo Horizonte	seat of a first-order administrative division population 2,373,224, elevation 888m	S 19° 55' 15"	W 43° 56' 16"
Belo Horizonte BH, Belo Horizonte, Belo Horizonte, Belo Hte, Belo Horizonte, Belo Horizonte, Belo Horizonte, Ur...	Brazil, Minas Gerais Belo Horizonte	second-order administrative division population 2,375,444	S 19° 55' 34"	W 43° 56' 23"
Belo Horizonte Belo Horizonte	Angola, Bie	populated place	S 11° 54' 48"	E 16° 53' 59"
Carlos Drummond de Andrade Airport Aeroporto da Pampulha, PLU, SDBH	Brazil, Minas Gerais Belo Horizonte	airport elevation 789m	S 19° 51' 4"	W 43° 57' 2"
Carlos Prates Airport Aeroporto Carlos Prates, SBPR	Brazil, Minas Gerais Belo Horizonte	airport	S 19° 54' 36"	W 43° 59' 23"
Belo Horizonte Aeroporto Internacional Tancredo Neves, Aeroporto de Confins, CNF, SBCF	Brazil, Minas Gerais Confins	airport elevation 827m	S 19° 38' 1"	W 43° 58' 7"
Barracão Belo Horizonte Barracão Belo Horizonte, Barracão Belo Horizonte, Barracão Belo Horizonte, Barracão Belo Horizonte, Be...	Brazil, Rondônia Guajará-Mirim	populated place	S 11° 53' 0"	W 64° 10' 0"
Belo Horizonte Belo Horizonte, Belo Horizonte, Santo Antonio	Brazil, Amazonas Envira	populated place	S 7° 18' 0"	W 69° 35' 0"
Metropolitana De Belo Horizonte	Brazil, Minas Gerais	region	S 19° 29' 52"	W 44° 1' 40"
Pampulha Pampulha	Brazil, Minas Gerais Belo Horizonte	populated place	S 19° 54' 0"	W 43° 56' 0"
Municipal Park Parque Municipal, Parque Municipal Americo Renné Giannetti, Parque Municipal Americo Renné Giannetti	Brazil, Minas Gerais Belo Horizonte	park	S 19° 55' 24"	W 43° 55' 59"
Sítio Belo Horizonte Sítio Belo Horizonte, Sítio Belo Horizonte	Brazil, Paraná Congonhinhas	populated place	S 23° 33' 37"	W 50° 31' 53"
Sítio Belo Horizonte Sítio Belo Horizonte, Sítio Belo Horizonte	Brazil, Paraná Congonhinhas	populated place	S 23° 37' 38"	W 50° 30' 4"
Belo Horizonte Belo Horizonte, Belo Horizonte	Brazil, Pará Altamira	populated place	S 5° 18' 0"	W 52° 52' 0"
Belo Horizonte	Brazil, Alagoas Campo Alegre	populated place	S 9° 48' 0"	W 36° 20' 0"

Figure 2.5: GeoNames result for Belo Horizonte

Figure 2.6 displays the first *Belo Horizonte* toponym occurrence in a map on the GeoNames website. Besides the map, the GeoNames show features such as the population of 2,373,224 inhabitants, the geographic coordinates (latitude $-19^{\circ}55'15''$ and longitude $-43^{\circ}56'16''$), and the subdivision hierarchy. This *Belo Horizonte* is a first-order administrative division (city) belonging to *Minas Gerais* state, in *Brazil* country.

For comparison, Figure 2.7 shows the search for *Belo Horizonte* toponym, using the Getty Thesaurus of Geographic Names website. While GeoNames find 429 records, the Getty Thesaurus of Geographic Names finds only two.

According to Di Rocco et al. [2020], there is no worldwide optimal gazetteer, and there is a deficiency of intra-urban toponyms in generalist gazetteers [Moura and Davis Jr., 2014]. Because of this, some works create their gazetteer or enriching existing ones. For instance, Moura and Davis Jr. [2014] use two linked data sources (GeoNames and DBPedia²¹) to produce the Linked Ontogazetteer²², an integrated and semantically-enriched gazetteer. Gao et al. [2017] create new gazetteer entries with volunteered

¹⁹<http://www.geonames.org/export/ws-overview.html>

²⁰<http://download.geonames.org/export/dump/>

²¹<https://wiki.dbpedia.org>

²²<http://aquio.io/log/>

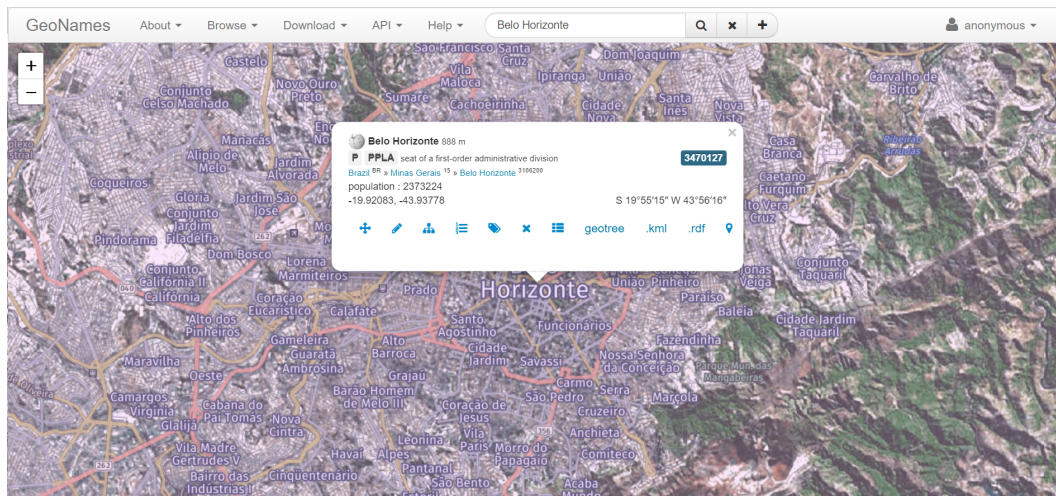


Figure 2.6: GeoNames result for Belo Horizonte on a map


user-tagged photographs, and Amitay et al. [2004] build their gazetteer to associate a geographic scope to web pages. Meanwhile, Amitay et al. [2004] derive a new gazetteer using the GeoNames and the OpenStreetMap²³ for geolocating microblog messages.

Gazetteers are relevant data sources not only for the geoparsing task. As seen in previous sections, even the reference resolution [Souza et al., 2005] and the grounding reference tasks can benefit from gazetteers. Chapter 3 presents the proposed methodology, which compares solutions to the geoparsing task by varying the gazetteer’s size and scope.

²³<https://www.openstreetmap.org>

 **Research**

[Research Home](#) ▶ [Tools](#) ▶ [Getty Thesaurus of Geographic Names](#) ▶ [Search Results](#)

 **Getty Thesaurus of Geographic Names® Online**
Search Results


[New Search](#) [Previous Page](#) [Help](#)

Find Name: **Belo Horizonte** Vernacular Display | **English Display**

Place Type:

Nation: 2 results

[View Selected Records](#) [Select All Records](#) [Clear All](#) [First](#) [Previous](#) [Next](#) [Last](#)
Page: **1**

Click the  icon to view the hierarchy.
Check boxes to view multiple records at once.

1.  **Belo Horizonte** (inhabited place)
(World, Europe, Portugal, Portalegre) [7745314]
2.  **Belo Horizonte** (inhabited place)
(World, South America, Brazil, Minas Gerais) [1020827]

[New Search](#) [First](#) [Previous](#) [Next](#) [Last](#)
Page: **1**

[▲ Back to top](#)

Figure 2.7: The Getty Thesaurus of Geographic Names result for Belo Horizonte

Chapter 3

Methodology to Evaluate Focused Gazetteers

This chapter details the methodology approach to evaluate the use of a focused gazetteer in the geoparsing task. This proposed methodology made it possible to assess the impact of these focused gazetteers to solve the geoparsing. The main idea was to fix the dataset and the geoparsing algorithm solution while used different focused gazetteers over the experiments.

Three main steps constitute the methodology: (1) Preparation; (2) Use of a Focused Gazetteer; and (3) Validation. Figure 3.1 illustrates the workflow of this proposed methodology.

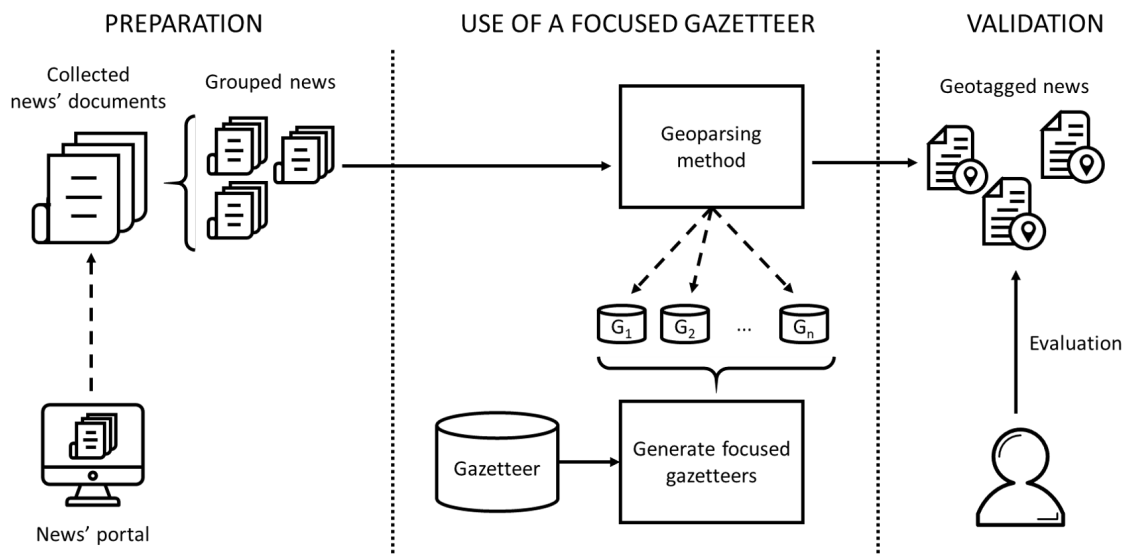


Figure 3.1: Proposed methodology workflow

The leftmost vertical dotted line indicates the end of the Preparation step, i.e., to find or generate the news dataset to use in the process. As said in Gritta et al. [2018], “there exists a challenge that all researchers in this field currently face, and it is the lack of freely available geotagged datasets.”. This step includes the creation of a raw data set consisting of a set of news texts.

The central part of Figure 3.1 shows the second step, the use of a focused gazetteer to solve the geoparsing task. Here, the idea is to use the same dataset created in step (1) and the same algorithm (Section 3.3), with different focused gazetteers (G_1, G_2, \dots, G_n). Each focused gazetteer is a variation of the the reference data provided by GeoNames gazetteer. Finally, the right side shows the final step of the methodology, Validation (3), that uses human evaluations to assess the results from (2) over a sample of the dataset. This final step is necessary since the dataset generated in (1) is not previously labeled.

Section 3.1 describes step (1) and introduces the news dataset generated. Section 3.2 shows the process of using the focused gazetteer, step (2), and preliminary results. Lastly, Section 3.3 details the process to execute the Validation step (3).

3.1 Preparation

There is a lack of freely geographic datasets in the English language, which is also true for Portuguese. Hence, one thesis contribution is to define a news Portuguese dataset, where each news item has a primary geographic scope.

Evaluating a focused gazetteer requires a dataset in which an approximation of the geographic scope is known for each text *a priori*. Based on this, the choice was for a news dataset, because this type of text is, in general, structured and follows grammatical standards and has a lower number of misspellings than social media messages. Also, commonly, news texts are grouped into regional sections, which already provide a preliminary geographical scope. Although there may be news datasets in other languages, such as the Signal Media Group dataset¹ (English news), creating a Portuguese news dataset is important because of the validation by humans in step (3).

The source for the collection of news was the G1 portal². This portal organizes the news texts into several sections, such as *Economia* (Economics), *Educação* (Education), and *Política* (Politics). For this work, the section *Regiões* (Regions) was adequate, since it organizes the news geographically, providing a rough geographical scope for the news texts. This section uses a hierarchy composed, on a first level, by the Regions of Brazil

¹<http://research.signalmedia.co/newsir16/signal-dataset.html>

²<https://g1.globo.com/>

(Midwest, Northeast, North, Southeast, and South), on a second level, by the states and the Federal District of Brazil, and finally, in some cases, a third level, corresponding to some sub-state regions, similar to the mesoregions defined by the Instituto Brasileiro de Geografia e Estatística, IBGE³ (Brazilian Institute of Geography and Statistics). In most states, the G1 portal does not use sub-state regions to organize the news; for instance, in Santa Catarina and Rio Grande do Sul states.



Caminhão tomba na BR-381, em Caeté, na Grande BH

Um caminhão que transportava carvão tombou nesta quarta-feira (7) na BR-381, em Caeté, na Região Metropolitana de Belo Horizonte.

De acordo com a Polícia Rodoviária Federal (PRF), o acidente foi no km 436, no sentido Vitória (ES).

Parte da carga se espalhou na pista, que também foi parcialmente interditada. Houve congestionamento. Ninguém se feriu.

Figure 3.2: News example from G1 portal

This news was published in the *Regiões* editorial organized as *Sudeste*→*Minas Gerais*→*Belo Horizonte e Região*. The first level, *Sudeste*, indicates the region, the

³<https://www.ibge.gov.br/geociencias/cartas-e-mapas/15778-diviso-es-regionais-do-brasil.html>

second level, *Minas Gerais*, indicates the state, and the third level *Belo Horizonte e Região* indicates a sub-state. This news has some primary geographic scopes, from a large grained one such as the *Brazil* itself to a more specific one, the *Belo Horizonte e Região* sub-state, given by this organization.

The crawler generated a JSONL⁴ file for each group of news that share the same primary geographic scope (same region, state, and sub-state in the G1 portal hierarchy). JSONL files are convenient for situations where it is required to iterate over JSON⁵ (JavaScript Object Notation) objects. In a JSONL file, each line represents a JSON object, and in this case, structured with the properties such as URL, source, date, time, title, subtitle, and the news text itself. Figure 3.3 shows the JSON object for the news example in Figure 3.2.

```
{
  "url":      "https://g1.globo.com/mg/minas-gerais/noticia/2018/11/07/caminhao-
tomba-na-br-381.ghtml",
  "source":   "G1.com",
  "date":     "07/11/2018",
  "time":     "07h24",
  "title":    "Caminhão tomba na BR-381, em Caeté, na Grande BH",
  "subtitle": "De acordo com a PRF, o acidente aconteceu no sentido Vitória (ES).",
  "text":     "Um caminhão que transportava carvão tombou nesta quarta-feira (7) na BR-
381, em Caeté, na Região Metropolitana de Belo Horizonte. \n\n De acordo com a Polícia
Rodoviária Federal (PRF), o acidente foi no km 436, no sentido Vitória (ES). \n\n Parte da carga
se espalhou na pista, que também foi parcialmente interditada. Houve congestionamento.
Ninguém se feriu. \n\n Às 11h, a PRF informou que o caminhão havia sido destombado e a
pista, liberada. \n\n"
}
```

Figure 3.3: JSON news example

The news items were collected using a Python⁶ web crawler, using libraries such as Beautiful Soup⁷, requests⁸, and JSON⁹. Also, each one of the RSS feeds¹⁰ from the G1 portal was necessary. The creation of the dataset and the collection of news took place throughout the year 2019, in alternating periods.

⁴<http://jsonlines.org/>

⁵<https://www.json.org/>

⁶<https://www.python.org/>

⁷<https://www.crummy.com/software/BeautifulSoup/bs4/doc/>

⁸<https://requests.readthedocs.io/en/master/>

⁹<https://docs.python.org/3/library/json.html>

¹⁰<http://g1.globo.com/tecnologia/noticia/2012/11/siga-o-g1-por-rss.html>

A data cleaning process was necessary for the collected news, first to remove duplicate news of the same JSONL file, and second because even using the specific RSS feeds for each location, news that cited different cities or geographic regions could be included in two different feeds. This cleaning process used the news’s URL address to eliminate duplicate texts. For instance, for *Goiás* state news, texts would only remain in the dataset if the URL contained the token */go/*; for *Caruaru and Região* sub-state news, inside *Pernambuco* state, the URL would have to have the token *pe/caruaru-region*. This data cleaning procedure eliminated approximately 50% of the news texts collected for the dataset.

Table 3.1 and Table 3.2 show the composition of the dataset. Table 3.1 shows the collected news from regions *Centro-Oeste*, *Nordeste*, and *Norte*. Note that only *Pernambuco* and *Pará* states have sub-state regions defined within the G1 portal.

Table 3.1: News’ total from *Centro-Oeste*, *Nordeste*, and *Norte*

Region	State	Sub-state	News	
Centro-Oeste	Distrito Federal		8,977	
	Goiás		9,637	
	Mato Grosso		14,199	
	Mato Grosso do Sul		9,759	
Nordeste	Alagoas		14,512	
	Bahia		16,482	
	Ceará		10,952	
	Maranhão		5,798	
	Paraíba		3,281	
	Pernambuco		Caruaru e Região	10,870
			Petrolina e Região	5,842
			Recife e Região	11,688
	Piauí		8,267	
	Rio Grande do Norte		9,590	
Sergipe		10,819		
Norte	Acre		12,650	
	Amapá		12,154	
	Amazonas		7,373	
	Pará		Belém e Região	17,554
			Santarém e Região	10,385
	Rondônia		7,669	
	Roraima		9,070	
	Tocantins		16,495	
			244,023	

Meanwhile, Table 3.2 shows the news collected from *Sudeste* and *Sul* regions. The dataset contains 529,585 news texts, organized in 55 JSONL files, occupying a

total of 1.13 GB.

Table 3.2: News' total from *Sudeste* and *Sul* regions

Region	State	Sub-state	News
Sudeste	Espírito Santo		8,409
	Minas Gerais	Belo Horizonte e Região	9,177
		Centro-Oeste	7,779
		Grande Minas	6,272
		Sul de Minas	7,443
		Triângulo Mineiro	10,259
		Vales de Minas Gerais	4,540
		Zona da Mata	10,385
	Rio de Janeiro	Norte Fluminense	6,041
		Região dos Lagos	6,779
		Região Serrana	5,723
		Rio de Janeiro e Região	16,614
		Sul e Costa Verde	10,328
	São Paulo	Bauru e Marília	6,083
		Campinas e Região	10,595
		Itapetininga e Região	3,972
		Mogi das Cruzes e Suzano	12,590
		Piracicaba e Região	8,339
		Presidente Prudente e Região	3,972
		Ribeirão Preto e Franca	8,535
Santos e Região		13,236	
São Carlos e Araraquara		10,611	
São José do Rio Preto e Araçatuba		4,648	
São Paulo e Região		12,475	
Sorocaba e Jundiaí		8,511	
Vale do Paraíba e Região		10,587	
Sul	Paraná	Campos Gerais e Sul	4,108
		Curitiba e Região	11,716
		Norte e Noroeste	6,303
		Oeste e Sudoeste	4,394
	Rio Grande do Sul		16,779
	Santa Catarina		14,668
			285,375

3.2 Use of a Focused Gazetteer

With the news dataset created, the next step is defining focused gazetteers to be used in geoparsing. As mentioned in Chapter 1, a focused gazetteer means a gazetteer covering a previously defined region. For this work, the hierarchy of the section *Regiões* (Regions), provided by the G1 portal, was used to generate the focused gazetteers needed.

The gazetteer chosen to be evaluated was GeoNames, a partially authoritative and partially crowdsourced gazetteer [Di Rocco et al., 2020]. The reasons for this choice were that it supports queries in Portuguese (as GeoNames records many places outside Brazil and Portuguese-speaking countries that have alternative names in Portuguese; e.g., “Londres”, for London) and it is a well-known resource used in several researches, directly or indirectly [Amitay et al., 2004; Martins et al., 2006; Popescu et al., 2008; Teitler et al., 2008; Serdyukov et al., 2009; Luo et al., 2011; Andogah et al., 2012; Brun et al., 2015; Rafiei and Rafiei, 2016; Inkpen et al., 2017; Santos et al., 2018]. In short, each GeoNames entry has a toponym, a coordinate pair (latitude and longitude), and a corresponding place category.

Instead of physically creating the focused gazetteer, this work chose to simulate the focused gazetteers with the use of geographic filters during the execution of the geoparsing method. Thus, each filter is a polygon that represents the expected geographic boundaries of the G1 portal’s sections. All entries whose footprint falls within the region boundaries are then considered as being part of the focused gazetteer.

The process of creating these filters used the QGIS tool¹¹ and geographic data used also by Freitas et al. [2012]. This was a manual process, since G1 sub-state divisions are not the same as the IBGE’s mesoregions. By sampling and using IBGE’s micro-regions¹², it was verified which region each news section contents supposedly covers. For instance, news about *Unaí*, a city that is the most important in the micro-region belonging to the *Noroeste de Minas* macro-region are included in the *Grande Minas* section, but news about *Salinas* and *Montes Claros*, micro-regions of *Norte de Minas* macro-region, are also in *Grande Minas* section. This situation caused many overlaps between IBGE’s mesoregions and G1 sections. Table 3.3 shows the equivalence between such state subdivisions for Minas Gerais state.

¹¹<https://qgis.org/>

¹²<https://www.ibge.gov.br/geociencias/cartas-e-mapas/15778-diviso-es-regionais-do-brasil.html>

Table 3.3: Equivalence between G1 editorial and IBGE’s mesoregions

Minas Gerais State	
G1 Editorial	IBGE’s Mesoregions
Belo Horizonte e Região	Jequitinhonha Metropolitana de Belo Horizonte Zona da Mata
Centro-Oeste	Central Mineira Metropolitana de Belo Horizonte Oeste de Minas Triângulo Mineiro e Alto Paranaíba
Grande Minas	Central Mineira Jequitinhonha Noroeste de Minas Norte de Minas
Sul de Minas	Campos das Vertentes Oeste de Minas Sul e Sudoeste de Minas
Triângulo Mineiro	Noroeste de Minas Triângulo Mineiro e Alto Paraíba
Vales de Minas Gerais	Jequitinhonha Vale do Mucuri Vale do Rio Doce
Zona da Mata	Campos das Vertentes Sul e Sudoeste de Minas Zona da Mata

In total, 67 geographic filters were created: one filter for Brazil; 5 for the major Brazilian regions; 27 for the Brazilian states and the federal district; and 34 for the sub-state regions, considering the G1 portal’s spatial hierarchy. All generated filters are defined in WKT¹³ (*Well-Known Text*) format. The WKT format represents the geographic limits in a string that “provides a means for humans and machines to correctly and unambiguously interpret and utilise a coordinate reference system definition with look-ups or cross references only to define coordinate operation mathematics” [Lott, 2019].

Completing step (2), requires to choose and implement a method to solve the geoparsing task, using the focused gazetteers. A lookup-based method was then built and used, with the help of the freely available GeoNames dump¹⁴ to find the toponym candidates in the dataset created in the previous step (Section 3.1). The method reads

¹³<https://www.ogc.org/standards/wkt-crs>

¹⁴<http://download.geonames.org/export/dump/>

the news text, with the title and the subtitle, and generates all n-grams. Then, these n-grams are matched against the gazetteer to find a corresponding GeoNames entry.

Algorithm 1 shows the lookup-based method’s pseudocode. The method was coded in Python and used libraries such as Shapely¹⁵, to manipulate the geometries and simulate the focused gazetteers; Pandas¹⁶ to format the results in spreadsheets; and NLTK¹⁷ to generate news n-grams. The option to use a lookup-based method to solve geoparsing was motivated by its simplicity and by having its results directly impacted by the quality of the gazetteers.

Algorithm 1 Lookup-Based Geoparsing

- 1: **Input:** A *jsonl* file with *news* JSON objects of the same location
 - 2: **Output:** A *table*, with one line for each *news*, with the toponyms candidates in columns

 - 3: **for each** *news* **do**
 - 4: Generate all capitalized *news* n-grams of size 1 to 5
 - 5: Query each n-gram against the GeoNames data
 - 6: Filter the matched n-gram using the country, region, state, and, if it exists, the sub-state polygons
 - 7: **end for**
 - 8: Save a *table* with all matched toponym candidates. Each column represents the results using a specific geographic filter, and each line corresponds to a news text
-

This method generates all capitalized n-grams (size one to five), considering the title, subtitle and news text, which are matched against GeoNames entries. Geographic filters are then used to simulate a focused gazetteer. In summary, each n-gram that has a correspondent in GeoNames, has its coordinates checked in each of the filters (country, region, state, and, if it exists, the sub-state polygon). The output of this method creates a spreadsheet for each news geoparsed. Each spreadsheet contains toponyms found, considering each filter, with its respective candidates.

The maximum n-gram size, five, was determined using a frequency distribution analysis. Figure 3.4 shows the frequency distribution of GeoNames toponyms, based on the number of words that compose them (n-grams). Since toponyms with five words or less correspond to 98.29% of the total GeoNames entries, the method limited the n-grams generated to five tokens (Algorithm 1, line 4). Also, as news are structured texts which follow grammatical standards, a heuristic that considers only capitalized n-grams to be potential toponym candidates was used.

¹⁵<https://pypi.org/project/Shapely/>

¹⁶<https://pandas.pydata.org/>

¹⁷<https://www.nltk.org/>

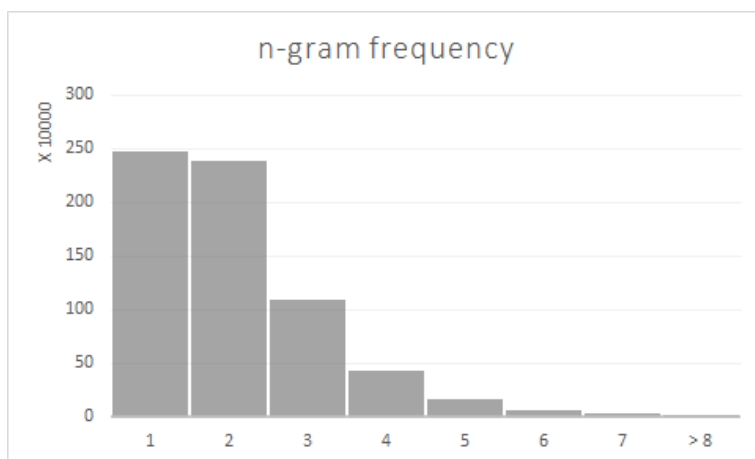


Figure 3.4: N-gram frequency over toponyms in GeoNames

Geographic filtering verifies whether the geographic coordinates of a matched n-gram, obtained from the gazetteer, are within the region’s limits, as defined by a WKT polygon (Algorithm 1, line 6). The method’s output organizes the toponym candidates of a single news text in a table (Algorithm 1, line 8).

Table 3.4 shows the Algorithm 1 table result to the news item illustrated in Figure 3.2.

Table 3.4: Geoparsing result for a single news text

Toponym Candidates	Global	Country	Region	State	Sub-state
Caeté	9	9	4	3	2
Grande	50	4	4	-	-
Vitória	30	27	6	-	-
Belo Horizonte	38	32	6	3	3
Polícia Rodoviária Federal	1	1	-	-	-
Parte	2	-	-	-	-

The results contain six toponym candidates. The columns *Global*, *Country*, *Region*, *State*, and *Sub-State* respectively indicate the geographic filters or focused gazetteers *GeoNames*, *Brazil*, *Sudeste*, *Minas Gerais*, and *Belo Horizonte e Região*. The numbers mean the ambiguous candidates to each toponym candidate. For instance, *Belo Horizonte* matches 38 GeoNames entries and only three candidates with the sub-state filter (*Belo Horizonte e Região* focused gazetteer). The - character indicates no toponym candidates found to that specific focused gazetteer. As a curiosity, the candidate toponym *Parte* has two matches in GeoNames, one being a populated place in the Italian commune of Belmonte Calabro and the other one a populated place in the Spain municipality of Monforte de Lemos. Clearly, this is a false toponym candi-

date to *Belo Horizonte* news. The table result also contains the coordinates (longitude and latitude) of each ambiguous toponym candidate. This information is not shown in Figure 3.2, for visualization reasons.

Algorithm 1 was executed on the entire dataset generated in step (1), resulting in a table for each news, and with each table containing the toponym candidates, considering each focused gazetteer. Chapter 4 details and discusses the results.

3.3 Validation

The last step of the methodology, Validation (3), consists of using humans to verify the use of the focused gazetteer. In other words, a poll with humans serves as a Geo/Non-Geo solver, since people are asked to confirm whether the results obtained in step 2 (Section 3.2) are valid toponyms or not. As mentioned earlier, this step is necessary because the database built in (1) was not previously labeled.

The main idea was to build a Web application to collect contributions from volunteers. Then, submit to them a small part of the dataset (Section 3.1), with the results obtained in step (2) indicated. This Web application allows people to evaluate as many news as they want, indicating which of the toponym candidates found in step (2) they recognize as actual toponyms. The application uses a voting scheme, over evaluations of the same news text, to compute the final result.

For each news item in this validation dataset, people declare whether candidates are actual toponyms or not. People can also declare that they are unsure about a certain candidate, or report a toponym that has not been identified by the geoparsing method (Algorithm 1) in Section 3.2.

This website used the Flask¹⁸ framework, a micro-framework that provides a model for web development. Also, it used PostgreSQL¹⁹ to store the contributions. The website is hosted in <https://pesquisaacademica.herokuapp.com/>.

The initial page contains information about the research project, and a brief explanation on how to participate. Figure 3.5 shows the contribution page. Left side shows the news with toponyms candidates highlighted (in grey), the right side box allows people vote for each candidate. Finally, in the lower section it is possible to type in any toponym that was not highlighted.

There was no limitation on the number of contributions that the same person could make. To select the news to be evaluated, the application chose one text randomly from among those that received three contributions or less. This was devised to achieve

¹⁸<https://flask.palletsprojects.com/en/1.1.x/>

¹⁹<https://www.postgresql.org/>

PESQUISA ACADÊMICA
HOME

Prefeitura de Santos Dumont¹ anuncia cancelamento do carnaval

Executivo alegou dificuldades financeiras. Cidade² não contou com festividade também em 2015.

A cidade de Santos Dumont¹ não contará com programação oficial para o carnaval de rua neste ano. O cancelamento da festividade foi confirmado através de decreto publicado pela Prefeitura no dia 19 de janeiro. A decisão foi justificada por dificuldades financeiras encontradas pelo município.

Segundo⁴ o decreto nº 3.090 de 2018, foi considerado o "alto custo para realização deste evento e que existem outras prioridades no município nas áreas de saúde, educação, infraestrutura urbana e rural, o que torna inviável a realização deste evento".

O G1 entrou em contato com Prefeitura para apurar o valor previsto de investimento para o carnaval 2018 e aguarda retorno. Em 2017, a festa custou aproximadamente R\$ 230 mil aos cofres públicos, valor bem mais alto de que o investimento feito em 2016, que foi de cerca de R\$ 50 mil.

Em 2015, a festa também foi cancelada por causa de crise financeira.

Quais destaques são nomes de lugares?

Santos Dumont¹

Cidade²

Santos Dumont³

Segundo⁴

Se encontrou na notícia nomes de lugares que não foram destacados, informe-os abaixo

Nome do Lugar:

FINALIZAR
E CONTINUAR ↻

FINALIZAR
E SAIR ➔

Figure 3.5: Website application

a minimum of three contributions for each news item, and to avoid that some person, in the same participation session, could evaluate the same text more than once.

With this proposed methodology, it was possible to fix Algorithm 1 and the dataset created on step (1) while varying the focused gazetteers. The Validation step (3) allowed to quantify the correctness of Algorithm 1 with each focused gazetteer. Chapter 4 presents the validation dataset, and discusses the results obtained.

Chapter 4

Analysis and Discussion

This chapter details and discusses the results of the Validation step (Chapter 3, Section 3.3). First, Section 4.1 details the data used to validate and verify the use of the focused gazetteers. Section 4.2 explains the process of collecting contributions of people on the voting process. Last, Section 4.3 presents the final analysis and discusses the results obtained.

4.1 Validation Dataset

The validation by humans used a small part of the dataset generated (Chapter 3, Section 3.1), with its respective Algorithm 1 results obtained in step 2. Initially, 560 news items were randomly selected, 80 news for each section, contemplating all sub-sections inside the *Sudeste*→*Minas Gerais* State editorial. Minas Gerais State comprises 853 municipalities and has rich toponymy set on other geographic elements, such as rivers, mountains, lakes, and other features. Thereby, Minas Gerais and its subdivisions were selected for validation since many volunteers live in the state’s capital, Belo Horizonte, and in nearby cities. The idea was to avoid losses in toponym identification by volunteers due to a lack of knowledge. This situation can occur if the volunteers are not familiar with the region mentioned in the news.

The selection of the news items occurred randomly since the content itself was not relevant to the research. At the end of the Validation step, 56 news items were excluded for not reaching the minimum of three evaluations. Table 4.1 shows the remaining 504 news selected and evaluated, indicating each sub-section.

Table 4.1: Validation dataset

G1 Editorials	News
Belo Horizonte	50
Centro-Oeste	80
Grande Minas	77
Sul de Minas	70
Triângulo Mineiro	79
Vales de Minas	78
Zona da Mata	70

Considering only results given by Algorithm 1, described in Chapter 3, these 504 news items had a total of 6126 toponyms candidates and an average of 12.18 toponyms candidates per news item when considering the gazetteer GeoNames. Beyond that, each one of the candidates has on average 25.43 possibles places for disambiguation.

With focused gazetteers, the average number of candidates for toponym per news decreased, as well as the number of ambiguous candidates for the same toponym. Table 4.2 shows the total of toponyms candidates (TpC), the average of toponym candidates per news (TpC/News), and the average of ambiguous candidates per toponym candidate (AmbC/TpC) for each sub-state editorial. Each of the Red_(i) columns show the percentage reduction of the previous metric considering the immediately preceding gazetteer.

Table 4.2: Stats of the validation dataset

Gazetteer	TpC	Red₁	TpC/News	Red₂	AmbC/TpC	Red₃
GeoNames	6,126	-	12.18	-	25.43	-
Country-focused	3,348	45.35%	6.66	45.32%	7.89	68.97%
Region-focused	2,059	38.50%	4.09	38.59%	3.83	51.46%
State-focused	1,578	23.36%	3.14	23.23%	2.41	37.08%
Sub-state focused	1,014	35.74%	2.02	35.67%	2.00	17.01%

In Table 4.2, the Country-focused line indicates the GeoNames gazetteer, limited by *Brazil* boundaries. Likewise, region and state-focused gazetteers are limited, respectively, by *Sudeste* and *Minas Gerais* boundaries. The last line of Table 4.2 indicates a sum of all values for each of the seven sub-state regions and corresponding focused gazetteers. The toponyms total is reduced by 45.35% with a country-focused gazetteer (6126 to 3348 toponyms) and reaches a 74.24% reduction with state-focused gazetteers compared to GeoNames (6126 to 1578 toponyms).

Moreover, there have been significant reductions in the average number of toponyms candidates per news item, 45.32% (with country-focused gazetteer) up to 74.22% when using the state-focused gazetteer. The average reduction in TpC/News values when using each of the focused gazetteers, when compared to its predecessor, is 35.70%. The average of ambiguous per toponym also reduced. The reduction was 68.97% with the country-focused gazetteer, and it reaches 90.52% when using the state-focused gazetteer. This decrease in AmbC/TpC average means less ambiguity for the toponyms present in the news. Each toponym stops having 25.43 candidates on average, to 2.41 with the state-gazetteer.

Table 4.3 exhibits the same results as Table 4.2, however considering each sub-state focused gazetteers. The TpC values are lower due to the number of news items in each G1 sub-section and the corresponding focused gazetteer. For example, the *Grande Minas* sub-section has 77 news items, and these news items, using the *Grande Minas* sub-state gazetteer, accounted for 92 toponyms (a 1.19 TpC/News and 2.37 AmbC/TpC). For comparison purposes, the same news from the *Grande Minas* sub-section, with GeoNames, accounted for 981 place names, 12.77 TpC/News, and 18.33 AmbC/TpC. The total number of toponyms decreases 90.62%, the average of toponym candidates per news shrinks 90.66%, and the AmbC/TpC reduced 87.07%.

Table 4.3: Stats considering sub state focused gazetteers

Focused Gazetteer	TpC	TpC/News	AmbC/TpC
Belo Horizonte	126	2.57	2.07
Centro-Oeste	171	2.14	1.94
Grande Minas	92	1.19	2.37
Sul de Minas	236	3.37	1.94
Triângulo Mineiro	145	1.84	2.02
Vales de Minas	83	1.06	1.87
Zona da Mata	161	2.30	1.91

For comparison purposes, Table 4.4 shows the size of each gazetteer considering the number of toponyms present. The Reduction column shows the percentage reduction in the number of toponyms considering the focused gazetteer immediately before. In the case of the sub-state-focused gazetteers, the percentage is concerning the state-focused one. It is possible to observe a high reduction in toponyms number when using the focused gazetteer compared to GeoNames. With the country-focused gazetteer alone, the number of toponyms decreased 99.00%.

These reductions are an expected result. Focused gazetteers obviously contain less data than a geographically broader gazetteer, such as GeoNames. In this way, both the number of toponym candidates and the average of ambiguous candidates per toponym tend to be smaller. The reduction in the size of gazetteers is greater than the reduction in the number of candidates per toponym, considering the country and region-focused gazetteers. For instance, the Brazil-focused gazetteer is approximately 1% the size of GeoNames (99.00% reduction), while AmbC/TpC reduced by 68.97% (Table 4.2). However, the AmbC/TpC reduction can reach 90.52% with the state-focused gazetteer and 92.65% with the sub-state-focused one (*Vales de Minas*-focused) when compared to GeoNames. These values are relevant because they indicate that for more geographically restrict gazetteers (sub-state-focused ones), the decrease in the size and number of candidates by toponyms are proportional.

Table 4.4: Number of Toponyms to each Focused Gazetteers

Gazetteer	Toponyms_(total)	Reduction
GeoNames	12,023,361	-
Country-focused	119,711	99.00%
Region-focused	31,421	73.75%
State-focused	11,298	64.04%
Sub-state-focused		
Belo Horizonte	4,361	59.53%
Centro-Oeste	2,006	81.38%
Grande Minas	3,022	71.95%
Sul de Minas	2,099	80.52%
Triângulo Mineiro	1,471	86.35%
Vales de Minas	2,303	78.63%
Zona da Mata	3,015	72.02%

Next section describes the counting and verification of volunteered contributions. Human assessments were necessary to confirm the successes and errors of Algorithm 1 under the effect of various focused gazetteers.

4.2 Human Contributions

The Validation step occurred between August 14, 2020, and October 14, 2020, using a Web application¹, and obtained almost 1800 evaluations. Table 4.5 shows that most news texts received a minimum of three assessments. Only 12% of the news items had five evaluations or more.

The decision on the minimum number of three evaluations for each news item was intentional. With three evaluations per news item, it is possible to untie contributions with different opinions about a toponym candidate, considering volunteered responses as votes. Besides, requesting a small minimum number of evaluations for each news item made it possible to include more news texts in the validation process.

Table 4.5: Volunteered contributions

Evaluations	Quantity
3	268
4	169
5	48
6	10
7	2

Of the 1794 evaluations made by people, 539 had indicated one or more toponyms that were not found by Algorithm 1. These 539 evaluations correspond to 281 distinct news texts. Only reported toponyms with a direct match in the text, with the exact spelling, were considered. The low percentage of distinct news items with additional toponyms supplied by volunteers (15.6%) implies that either people did not want to inform these toponyms, or that the geoparsing method recognized most toponyms. People were not required to report any missing toponym in the news, but they could do so if they wanted to. Therefore, it was difficult to determine why there was a low rate of news items with additional toponyms provided by people, therefore we considered that Algorithm 1 recognized most toponyms.

A manual investigation of these 539 evaluations with toponyms indicated by people found that more than 90% of the toponyms informed represent intracity localities such as street names, parks, and neighborhoods. Other commonly supplied toponyms were fuzzy regions, such as *região sul de Belo Horizonte* (south region of Belo Horizonte) or *leste de Minas Gerais* (Minas Gerais east). The indication of intracity toponyms

¹<https://pesquisaacademica.herokuapp.com/>

was not surprising, as there is a lack of these types of toponyms in the Brazilian part of GeoNames [Borges et al., 2007].

This manual investigation also noted the lack of standardization on how people recognize and inform place names. Different evaluations indicated the equivalent location in different ways. For instance, both *Avenida Juscelino Kubitschek* and *Juscelino Kubitschek* were used to mention the same avenue for a news text. This problem is outside the scope of this work, and it is covered in recent research [Gritta et al., 2019]. Each evaluation includes an unique code, the news id, the evaluation date, a tuple indicating whether, for each toponym candidate, if the user agrees that the candidate is a real toponym. Also, the evaluation contains a list of volunteer indicated toponyms but not identified by Algorithm 1.

To decide whether a toponym candidate from a news item is an actual toponym, an account of votes was made over the evaluations. This counting considered the number of 'Yes', 'No', and 'I don't know' votes that a candidate received. Only toponym candidates with more than 50% 'Yes' votes, with at least three evaluations, were considered to be correct. The toponyms reported by the volunteers were not submitted to any verification. Toponyms indicated in a single volunteer response were assumed to be correct. Once again, notice that reporting unidentified toponyms was not mandatory. So, evaluations with toponyms supplied by volunteers were considered to be more valuable than evaluations that did not indicate the same toponyms for the same news.

In this step, volunteers evaluations served as a Geo/Non-Geo classification of the results. Next section shows and details the analysis of results of this classification.

4.3 Analysis of the Results

The analysis used a confusion table and computed precision, recall, and F1 score metrics over the results of step 3 (Chapter 3). The main objective was to verify if a focused gazetteer can increase precision in the geoparsing task while generating less ambiguity. Figure 4.1 show the confusion table.

The horizontal axis expresses the evaluations made by humans, and the vertical axis represents the results given by the geoparsing method. All toponym candidates found by Algorithm 1 and evaluated as real by people are True-Positives (TP), and False-Positives (FP) are all toponym candidates declared as false by people. False-Negatives (FN) are real toponyms not detected by the geoparsing method and indicated by people. True-Negatives (TN) would correspond to non-toponyms not detected by Algorithm 1 and not reported by people.

		People's evaluation	
		True Toponym	False Toponym
Algorithm 1 results	True Toponym	TP	FP
	False Toponym	FN	TN

Figure 4.1: Confusion Table

Precision ($\frac{TP}{TP+FP}$) aims to indicate the proportion of correct answers among the results of Algorithm 1. Recall ($\frac{TP}{TP+FN}$) shows the ratio of correct answers over the real positives. Meanwhile, the F1 score ($2 * \frac{precision * recall}{precision + recall}$) gives a value correlating precision and recall.

Each news had its TP, FP, FN, precision, recall, and F1 score values calculated considering all gazetteers (country, region, state, and sub-state). This work did not need TN values since they are not necessary to calculate precision, recall and F1 score.

To exemplify this process, see Figure 4.2 that shows a news item² from sub-section *Centro-Oeste*.

From the complete GeoNames contents, the geoparsing method identified the following candidates for toponyms, with their respective ambiguous places: *Empresa* (4 candidates), *Itaúna* (11 candidates), *Usiminas* (1 candidate), *Bolsa* (7 candidates), *Senai* (4 candidates), *Rua* (26 candidates), *Antunes* (7 candidates), *Bairro* (8 candidates), *Nogueira* (45 candidates) and, *Machado* (32 candidates). For comparison, using the sub-state focused gazetteer, the method identified only the toponyms *Itaúna* (2 candidates), *Antunes* (1 candidate), and *Machado* (1 candidate).

Looking now at the assessment by volunteers, this news item received four evaluations. They were unanimous in deciding that each of the toponym candidates was real. The only candidate identified as toponym was *Itaúna*, in all three times that this name appears in the news. Additional toponyms reported by the volunteers were *Rua Lília Antunes* (or *Lília Antunes*) and *Bairro Nogueira Machado* (or *Nogueira Machado*). Corroborating the human evaluation, in a closer inspection of the news, it is possible to observe that only *Itaúna*, *Rua Lília Antunes*, and *Bairro Nogueira Machado* are actual toponyms.

Considering the GeoNames, this news had 3 True-Positives, 10 False-Positives,

²<https://g1.globo.com/mg/centro-oeste/concursos-e-emprego/noticia/2018/09/12/empresa-de-siderurgia-abre-inscricoes-para-cursos-de-capitacao-para-itauna-e-regiao.ghtml>



MENU G1 CENTRO-OESTE TV INTEGRACAO

Empresa de siderurgia abre inscrições para cursos de capacitação para Itaúna e região

Programa da Usiminas oferece aulas sobre manutenção industrial e manutenção de equipamentos a diesel para jovens de 18 a 23 anos. Bolsa auxílio é de R\$ 954.

Por G1 Centro-Oeste de Minas
12/09/2018 18h59 · Atualizado há 2 anos

A Usiminas está com inscrições abertas para o segundo semestre do programa de formação de aprendizes da empresa de siderurgia em Itaúna e região. Jovens de 18 a 23 anos podem concorrer à participação em um dos dois cursos de capacitação oferecidos e os selecionados recebem uma bolsa auxílio de R\$ 954.

O "Programa Aprendiz da Mineração" está com 20 vagas abertas para o curso de manutenção industrial e outras 20 para o curso de manutenção de equipamentos a diesel.

As inscrições seguem até o dia 21 de setembro no Senai de Itaúna, na Rua Lília Antunes, 99, no Bairro Nogueira Machado.

Para participar da seleção, os candidatos devem ter concluído ou estar cursando o ensino médio. As aulas devem começar em outubro.

Além da bolsa auxílio, os selecionados recebem vale transporte, alimentação e seguro de vida.

Figure 4.2: News item from G1 Centro-Oeste sub-section

and 3 False-Negatives. Also, it had a 23% precision, 50% recall, and 32% F1 Score. The focus of this work is not on the quality of the geoparsing method on its own. Instead, the focus is on the gain in precision that focused gazetteers can bring to the geoparsing task. Although Algorithm 1 has a low precision value (23%), there is an increase with focused gazetteers.

With the country-focused gazetteer, TP and FN remain the same, three. Just the FP reduced to five. The values of precision, recall, and F1 score were, respectively, 38%, 50%, and 43%. Here, the precision value already had a gain of more than 65%. These values were identical when using region- and state-focused gazetteers. As there were no False-Positives eliminated, precision has not increased using these two more restricted focused gazetteers. However, using the sub-state focused gazetteer (for *Centro-Oeste*), FP dropped to two, and precision reached 60%, and F1 score 55%, while recall keeps

the 50%. For this news, precision almost doubled, from 32% to 60%.

As expected, when a focused gazetteer eliminates False-Positive candidates for not having it, the precision will rise. However, the same does not occur with recall. Recall values for a single news item, considering every focused gazetteer, are constant (50% in the previous example) because the number of toponyms actually mentioned in each news item does not change. The evaluation step processes the use of focused gazetteers after people’s participation. However, recall gains relevance considering the set of news. Thus, all news items evaluated with no missing toponym reported by people ($FN = 0$) have a 100% recall.

Table 4.6 shows precision and F1 score metrics for each news belonging to the *Belo Horizonte e Região* sub-section. Abbreviations *PR*, *RC*, and *F1* mean, respectively, Precision, Recall, and F1 Score. The *N* column shows the news id, and the next five columns show the values of precision and F1 score using the gazetteers GeoNames, country-focused (*Brazil*), region-focused (*Sudeste*), state-focused (*Minas Gerais*), and sub-state-focused (*Belo Horizonte e Região*). As recall values are constant, only the column Geonames presents these values. Appendix A contains raw tables with results for the other *Minas Gerais* news sub-sections.

Table 4.6: Belo Horizonte sub state news evaluation

N	Geonames			Brazil		Southeast		MG		BH	
	PR	RC	F1	PR	F1	PR	F1	PR	F1	PR	F1
50	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
52	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
72	10%	100%	18%	17%	29%	25%	40%	33%	50%	100%	100%
53	11%	50%	18%	100%	67%	100%	67%	100%	67%	100%	67%
23	8%	67%	14%	11%	19%	22%	33%	25%	36%	67%	67%
24	12%	56%	20%	38%	45%	50%	53%	71%	63%	100%	72%
75	4%	100%	8%	6%	11%	8%	15%	14%	25%	33%	50%
30	10%	80%	18%	40%	53%	50%	62%	80%	80%	80%	80%
67	14%	44%	21%	29%	35%	44%	44%	50%	47%	100%	61%
49	13%	83%	22%	19%	31%	33%	47%	38%	52%	80%	80%
2	17%	50%	25%	50%	50%	100%	67%	100%	67%	100%	67%
65	9%	21%	13%	15%	18%	43%	28%	50%	30%	50%	23%

Table 4.6 continued from the previous page

Geonames				Brazil		Southeast		MG		BH	
N	PR	RC	F1	PR	F1	PR	F1	PR	F1	PR	F1
1	38%	50%	43%	50%	50%	75%	60%	75%	60%	67%	50%
28	12%	46%	19%	19%	27%	28%	34%	27%	31%	22%	22%
40	21%	90%	34%	30%	45%	60%	72%	50%	63%	45%	58%
42	42%	100%	59%	93%	96%	92%	96%	90%	95%	83%	91%
43	22%	89%	35%	42%	57%	50%	62%	71%	77%	50%	57%
48	40%	80%	53%	50%	62%	67%	73%	80%	80%	75%	75%
38	44%	88%	59%	67%	73%	60%	67%	60%	67%	100%	80%
29	7%	67%	13%	14%	23%	25%	36%	14%	22%	25%	33%
78	17%	56%	26%	34%	42%	67%	56%	54%	39%	53%	30%
80	27%	75%	40%	38%	50%	75%	75%	67%	67%	100%	80%
25	48%	100%	65%	65%	79%	80%	89%	67%	80%	100%	100%

In total 30 news items (news 4, 6, 8, 17, 19-22, 27, 34, 39, 41, 47, 54-64, 66, 68-71 and 79) left the analysis because they did not reach the minimum of three evaluations.

The first two news texts (news 50 and 52), in Table 4.6, had 0% in all metrics: precision, recall, and F1 score. A manual verification found that these news items did not contain text, only a set of photos under a title. Also, none of them had toponym candidates in their titles founded by Algorithm 1. Gray lines present news that increase the precision constantly consider each focused gazetteer. Many news items achieve 100% precision, and other news items reached their higher precision rate without the need for a more specific focused gazetteer, for instance, the news 53 (line 6). The lowest precision gain was 67% (news 46, last gray line). The precision gain of news items in the first 20 gray lines is above 300%.

On average, the precision gain was 52%. Likewise, the F1 score also increases, from 19% to 22% in the worst case and from 18% to 100% in the best one. On average, the F1 score increases by 17%.

The yellow lines represent news that had a drop in precision with some more restricted-focused gazetteer. The explanation for these cases is that the most restrictive focused gazetteer excludes some True-Positive toponyms (already confirmed by people),

reducing the precision. For instance, if a *Belo Horizonte* news mentions a country or a city that does not belong to these limits, this candidate is no longer recognized by the geoparsing method. Future researches can deal with situations like this.

Focused gazetteers could include some predetermined list of universally well-known toponyms to minimize this problem. Such a list could contain continents and country names or follow the approach used by [Teitler et al., 2008; Lieberman and Samet, 2012b]. They compile a list of city names (all unambiguous), considering a measure of importance to select the homonymous places. A country-focused gazetteer will, of course, contain toponyms of its regions and states. As stated before, the recall value, when looking at the news individually, does not change.

The news items in red lines present a variable precision. For instance, the last one (news 25) rises precision until 80% with the region-focused gazetteer, then the precision value drops to 67% and rises again to 100%. The explanation for these cases is quite direct. While the other news (outside the red area) tends to drop only the False-Positives, keeping True-Positives constant, these news items, considering each focused-gazetteer, decrease both the True-Positives and False-Positives at different rates. Again with news 25, using Geonames, there are $TP = 11$ and $FP = 12$. Using the region-focused gazetteer, $TP = 8$ and $FP = 2$ (a drop in both values). With the state-focused one, $TP = 4$, and FP remains two. Finally, with *Belo Horizonte*-focused gazetteer, $TP = 3$ and $FP = 0$, reaching a 100% precision. All of this explanation is valid for news from the other editorials in Appendix A.

Also, it is possible to analyze these metrics regarding all news from the same G1 sub-section. Figure 4.3 displays a line graph with Precision, Recall, and F1 Score to *Belo Horizonte e Região* G1 sub-section editorial news.

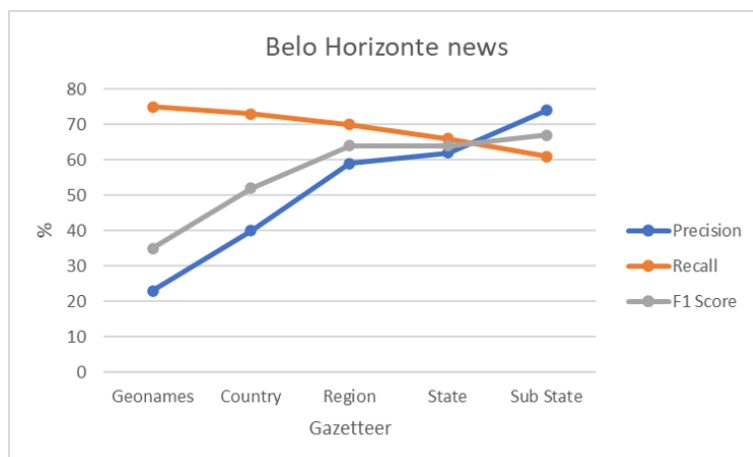


Figure 4.3: Belo Horizonte news evaluation

The blue line represents the precision, increasing from 23% to 74%. The F1 Score, the grey line, goes from 35% to 67%. The recall, represented by the orange line, drops from 75% to 61%. Although there is a slight drop in the recall rate, precision increases significantly with the *Belo Horizonte e Região* focused gazetteer, causing the F1 score to increase. Again, new approaches using lists of well-known toponyms can prevent the decrease in recall values.

Table 4.7 presents precision, recall, and F1 score values to all sub-state focused gazetteers. Each line corresponds to one of the *Minas Gerais* G1 sub-sections: BH (*Belo Horizonte e Região*), CO (*Centro-Oeste*), GM (*Grande Minas*), SM (*Sul de Minas*), TM (*Triângulo Mineiro*), and ZM (*Zona da Mata*). All focused gazetteers improve precision when compared to broader gazetteers. F1 Score values also improve, mainly when the full Geonames results are compared with sub-state focused gazetteers. There is a small drop in recall values but, in the face of the greater improvements in precision, the F1 increases.

Table 4.7: Sub-state gazetteers evaluations

	Geonames			Brasil			Sudeste			Minas Gerais			Sub-state		
	PR	RC	F1	PR	RC	F1	PR	RC	F1	PR	RC	F1	PR	RC	F1
BH	23%	75%	35%	40%	73%	52%	59%	70%	64%	62%	66%	64%	74%	61%	67%
CO	28%	75%	40%	50%	74%	60%	69%	74%	72%	78%	74%	76%	81%	73%	77%
GM	11%	82%	20%	22%	82%	35%	36%	81%	50%	43%	79%	55%	82%	77%	79%
SM	35%	85%	49%	54%	84%	66%	65%	83%	73%	71%	81%	76%	77%	79%	78%
TM	27%	78%	40%	46%	77%	57%	61%	76%	68%	69%	75%	72%	86%	75%	80%
VM	14%	82%	23%	26%	81%	39%	41%	79%	54%	46%	78%	58%	78%	74%	76%
ZM	22%	82%	35%	34%	81%	48%	47%	79%	59%	59%	78%	67%	70%	76%	73%

For instance, news from *Grande Minas* G1 sub-section editorial has a precision reaching 82% with a proper sub-state focused gazetteer. An improvement of more than 600% in precision while recall drops only 6%. The smaller precision gain was 120%, with the *Sul de Minas* G1 sub-section news, and with only a 7.06% drop in the recall. On the other hand, the highest fall in the recall values was 18.68%, with the *Belo Horizonte e Região* sub-section news, but with a precision gain of 221.73%. This increase in precision corroborates the hypothesis raised in this thesis that focused gazetteers can improve the precision in recognizing toponyms present in texts.

Lastly, Table 4.8 presents a summary of the results considering all 504 news of the validation dataset.

Table 4.8: All 504 news evaluations

Geonames			Brasil			Sudeste			Minas Gerais			Sub-state		
PR	RC	F1	PR	RC	F1	PR	RC	F1	PR	RC	F1	PR	RC	F1
23%	80%	36%	40%	79%	53%	55%	78%	64%	63%	76%	69%	78%	74%	76%

These results show that all precision, recall, and F1 score values present a behavior that is similar to previous results with individual sub-section news (Table 4.7). The experiments reveal that whether considering each sub-section individually or considering the set of 504 news items, the results are equivalent. Precision rose from 23% to 74%, which is in line with the precision values shown in Table 4.7. On average, the precision increase was 239% with 504 news items and 243% considering the average increase in individual sub-sections. Also, the F1 rises significantly, by 111%. The recall value continues to present a small drop, of 8%. Again, we observed a considerable increase in the precision of toponym recognition with a low decrease in the recall value.

All precision increases corroborate the hypothesis raised in this thesis, that focused gazetteers can improve the precision in recognizing toponyms present in texts. Besides that, the gains in precision increasingly occur with less restricted focused gazetteers (country-wide), an average increase of 74%.

In addition to improving the precision in the Geoparsing task, focused gazetteers present less ambiguous candidates. Next section discusses the results observing that aspect, considering the toponym ambiguity problem and focused gazetteers.

4.4 Ambiguity Analysis

Beyond increasing the precision in toponym recognition in the text, using focused gazetteers limits the occurrence of ambiguous toponyms. As said in Section 4.1, the focused gazetteers reduce the number of ambiguous candidates per toponym and, this is an expected result. However, this reduction is only good if the correct place remains within the candidate places identified by the gazetteer.

First, remember that focused gazetteers are directly related to the preliminary geographic scope of the data. In other words, it is necessary to have prior information about the geographic region to which the text probably refers in order to be able to select and use a focused gazetteer. In this way, the primary geographic focus serves as a geographic boundary for the focused gazetteer.

Considering the focused gazetteers, Table 4.2 and Table 4.3, in Section 4.1, show a comparison of the average number of ambiguous candidates per toponym in the news. Furthermore, Table 4.4, also in Section 4.1, indicates the size of these gazetteers

compared to the complete Geonames. Table 4.9 presents these data again, but relating them to the precision results from the previous section, which were calculated over the sample of the texts that were submitted to human evaluation.

Table 4.9: Number of Toponyms to each Validation Focused Gazetteers

Gazetteer	Toponyms_(total)	Precision	Recall	AmbC/TpC
Geonames	12,023,361	23%	80%	25.43
Country-focused	119,711	40%	79%	7.89
Region-focused	31,421	55%	78%	3.83
State-focused	11,298	63%	76%	2.41
Sub-state-focused				
Belo Horizonte	4,361	74%	61%	2.07
Centro-Oeste	2,006	81%	73%	1.94
Grande Minas	3,022	82%	77%	2.37
Sul de Minas	2,099	77%	79%	1.94
Triângulo Mineiro	1,471	86%	75%	2.02
Vales de Minas	2,303	78%	74%	1.87
Zona da Mata	3,015	70%	76%	1.91

The Toponyms_(total) column indicates the number of toponyms present in each gazetteer. Precision and Recall columns show, respectively, the percentage of the real toponyms found and the percentage of toponyms actually cited found with the Geoparsing task. In the Validation step (Section 3.3), volunteers validated the precision and recall values. The last column, AmbC/TpC indicates the average of ambiguous candidates per toponym.

Precision gains have already been discussed and analyzed in Section 4.3. In Table 4.9, they are present to confirm the relevance of the reduction in the average number of candidates by toponyms. The drop in the AmbC/TpC value is also relevant. This average falls from 25.43 per toponym, with Geonames, to 2.41 with state-focused gazetteers, a 90.52% reduction. Using the *Vales de Minas*-focused gazetteer to calculate the percentage decrease, it reaches 92.65%.

Still, with the *Vales de Minas*-focused gazetteer, the AmbC/TpC is 1.87, meaning that for each toponym present in the news text, there are, on average, just under two candidates left for the disambiguation process to select. Furthermore, given that the *Vales de Minas*-focused gazetteer contains only toponyms belonging to that geographic region, it can be said that the candidates eliminated correspond to false ones. The 78% precision and 74% recall, to the Geoparsing method, corroborates this observation.

In summary, the *Vales de Minas*-focused gazetteer has less than 0.02% of the toponyms in Geonames, but achieves precision and recall over 70%, with a 92.65% reduction in the average number of toponym candidates per news text. Other results shown in Table 4.9 are consistent with this explanation.

Figure 4.4 exhibits all candidates for toponyms found with the validation dataset using GeoNames. For each candidate found in one of the 504 news items, we plotted all corresponding points using their geographic coordinates. Each point represents a disambiguation location for one of the toponym candidates.

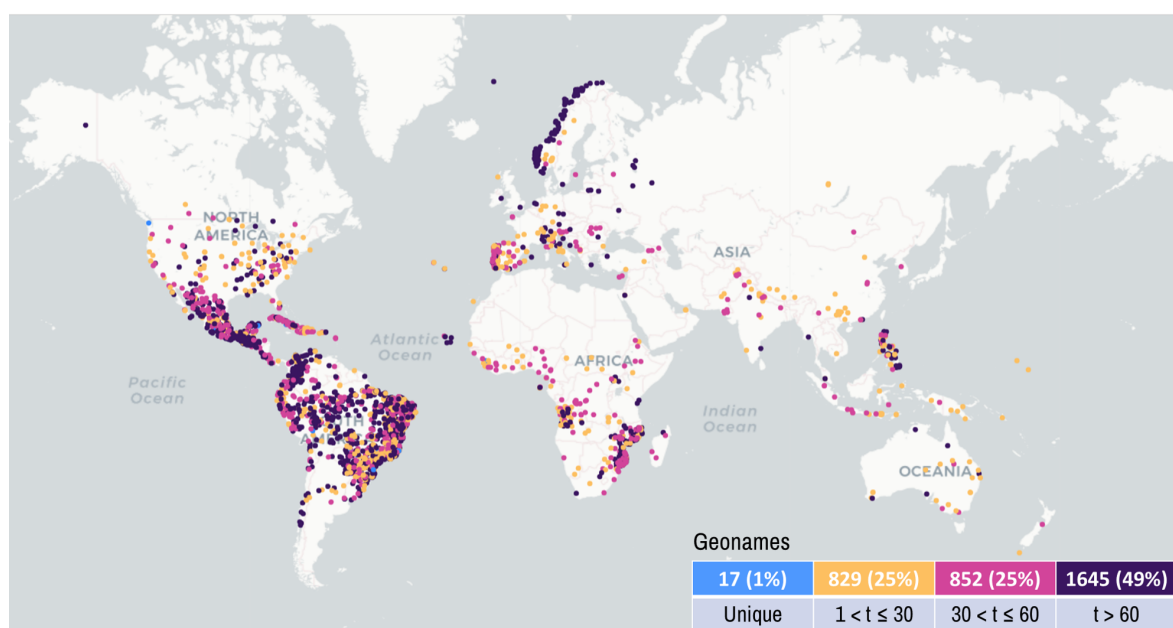


Figure 4.4: Ambiguous candidates with GeoNames

Blue dots indicate unambiguous candidates and represent only 1% of the total. Candidates for toponyms with more than 60 possible locations correspond to 49% of the total. The percentage of toponyms with more than 30 candidate places is about 74%. It is possible to notice that even though news are from *Minas Gerais*, there are points spread worldwide, indicating a high potential to contribute to ambiguity.

Figure 4.5 display the same map with toponym candidates but considering the Brazil-focused gazetteer. The first result is obvious and comes from the definition of the focused gazetteers themselves. The country-focused gazetteer eliminates all candidates outside Brazil's borders.

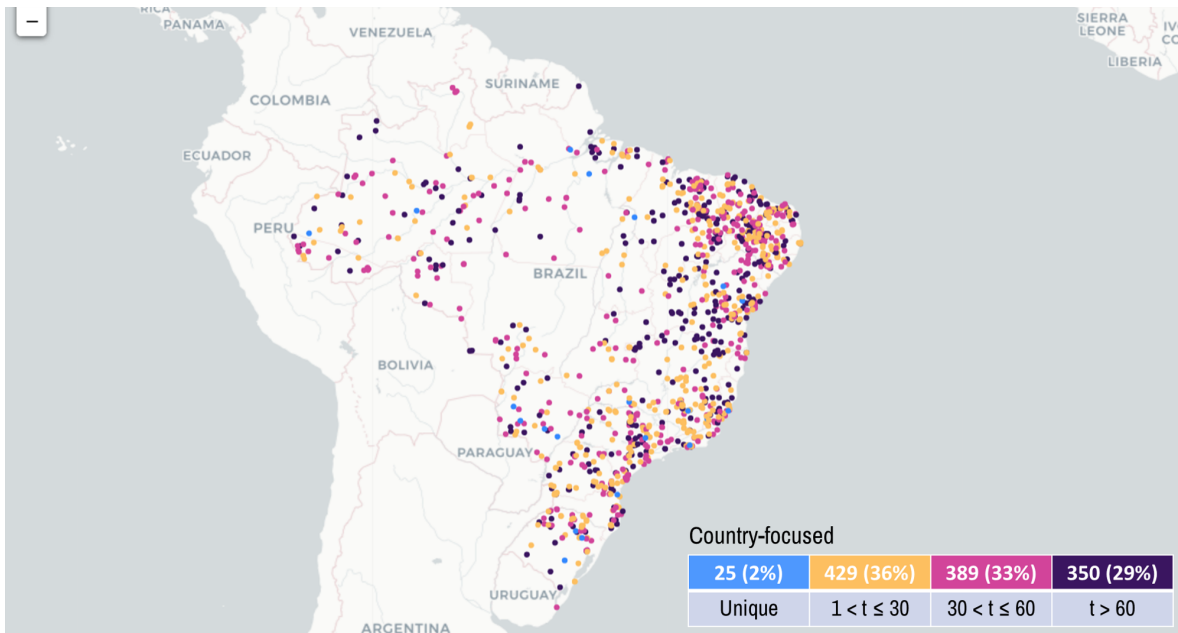


Figure 4.5: Ambiguous candidates with country-focused gazetteer

Again, each point represents a candidate for the disambiguation of a toponym. In addition to eliminating candidates outside Brazil's territorial limits, the country-focused gazetteer decreases the number of ambiguous candidates. The number of blue points rose to 2% of the total, while the number of toponyms with more than 30 candidate places dropped to 62%. In turn, Figure 4.6 presents ambiguous candidates on the map regarding the state-focused gazetteer.

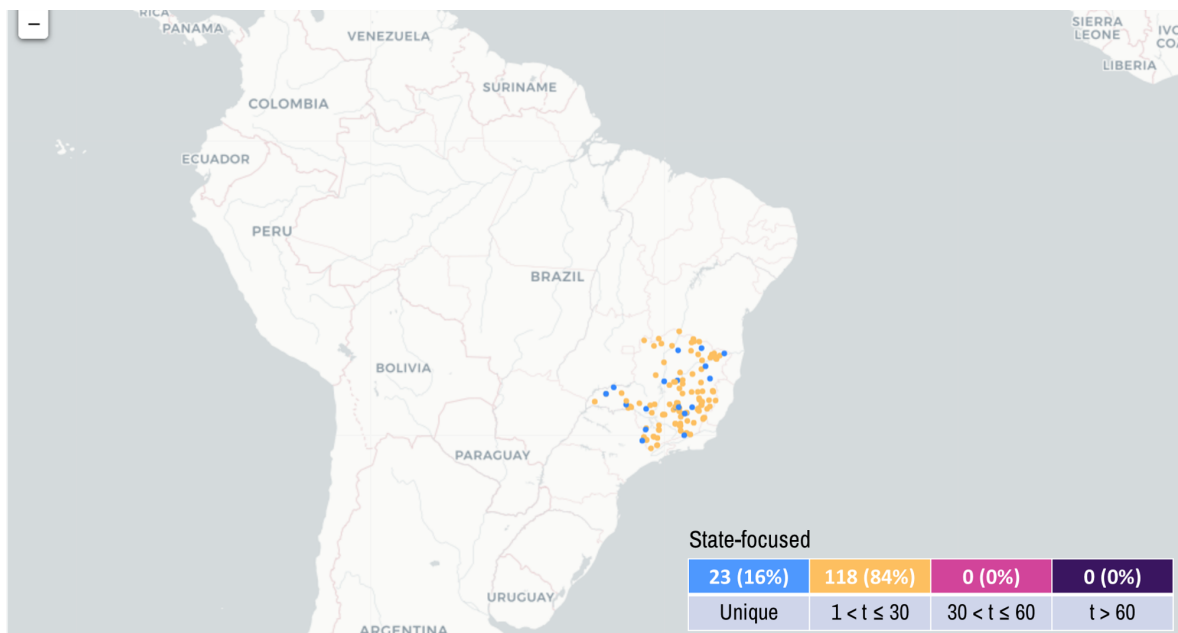


Figure 4.6: Ambiguous candidates with state-focused gazetteer

With each more restricted focused-gazetteer, the maps display more light-colored dots (blue and orange) and fewer dark-colored ones (purple and pink). For this state focused-gazetteer, there are no ambiguous toponyms with more than 30 candidate places. Also, the percentage of unambiguous dots (blue ones) reached 16% of the total. This visual result is relevant considering the precision and recall values obtained with the country-focused gazetteer, shown in Table 4.9.

Table 4.10 exhibits the ambiguity analysis to all datasets created. The table has the sizes of all focused gazetteers ($\text{Toponyms}_{(\text{total})}$) together with the number of news of each G1 section ($\text{News}_{(\text{total})}$) and the average values of ambiguous candidates per toponym (AmbC/TpC). The precision and recall values obtained with the validation dataset corroborate the results shown.

The AmbC/TpC column has values consistent with , indicating a reduction in the number of ambiguous candidates by toponym. The first two lines from Table 4.10 are different from the respective lines in Table 4.9 because the last one considers all 529,585 news, while the first is an analysis made on the 504 news items from the validation dataset. The average number of ambiguous candidates by toponym is 25.43 with the validation dataset and 20.95 considering the complete dataset.

These averages remain the same considering both the country-focused gazetteer and the region-focused-gazetteer. The AmbC/TpC is 7.98 to the validation dataset and 8.81 to the complete one with Brazil-focused. The values of AmbC/TpC remain close with the region-focused gazetteer, 3.83 to validation and 3.77 to the whole dataset.

The reduction in AmbC/TpC values is also similar. The fall from 20.95 using GeoNames to 8.81 using Brazil-focused represents a 58% reduction. With the validation dataset, there was a 69% reduction in the AmbC/TpC values. With regions-focused gazetteers, the average drop is between 77% (with Nordeste-focused) and 89% (with Centro-Oeste-focused). Considering the states focused gazetteers, the AmbC/TpC reduction is over 87%, with São Paulo-focused, reaching 94% with Distrito-Federal-focused gazetteer.

These reductions in the number of ambiguous candidates per toponym are relevant because they indicate the possibility of using simpler and more efficient disambiguation solutions after the geoparsing stage. Chapter 5 discusses the conclusions of this thesis and future work involving focused gazetteers.

Table 4.10: Number of Toponyms considering all dataset

Gazetteer	Toponyms_(total)	News_(total)	AmbC/TpC
Geonames	12,023,361	529,585	20.95
Brasil-focused	119,711	529,585	8.81
Regions-focused			
Norte	17,022	93,350	4.59
Nordeste	33,351	108,110	4.82
Centro-Oeste	10,684	42,572	2.24
Sudeste	31,946	227,407	3.77
Sul	26,708	57,968	2.46
States-focused			
Acre	1,044	12,650	2.67
Amapá	509	12,154	2.36
Amazonas	6,095	7,373	2.62
Pará	4,203	27,939	2.00
Rondonia	1,243	7,669	1.85
Roraima	1,123	9,070	1.89
Tocantins	2,805	16,495	1.79
Alagoas	1,096	14,521	1.59
Bahia	7,620	16,482	2.55
Ceará	6,286	10,952	1.99
Maranhão	4,156	5,798	2.65
Paraíba	2,334	3,281	2.04
Pernambuco	3,336	28,400	2.15
Piauí	1,123	8,267	2.04
Rio Grande do Norte	2,535	9,590	2.17
Sergipe	476	10,819	1.72
Distrito Federal	269	8,977	1.28
Goiás	476	9,637	2.34
Mato Grosso	3,782	9,759	1.87
Mato Grosso do Sul	2,524	14,199	2.04
Espírito Santo	1,457	8,409	1.76
Minas Gerais	11,298	55,529	2.41
Rio de Janeiro	5,705	45,475	1.99
São Paulo	13,486	117,994	2.81
Paraná	17,703	26,521	1.87
Rio Grande do Sul	4,008	16,779	2.13
Santa Catarina	4,497	14,668	1.85

Chapter 5

Conclusions and Future Work

This thesis aimed to show that a focused gazetteer increases the precision in the geoparsing task, generating less ambiguity in the process. Experimental results confirm that, using a broad indication as to a region to which a text is related as a preliminary step, a gazetteer whose contents focus in that region obtains better results, using the same geoparsing algorithm.

We conducted a comprehensive and exhaustive review of the literature about the geographic scope resolution problem. This literature review originated a survey that identifies the tasks that make up the GSRP and tries to group the different nomenclatures existing in the literature [Monteiro et al., 2016]. Besides that, this survey contains classifications of the algorithms and solution methods of both the GSRP and its tasks.

Such a survey enabled to check the difficulty in comparing different works about the GSRP or its tasks. Most of these works change the methods, the dataset, and the external knowledge base. Besides that, most of them focused only on the efficiency and effectiveness of the proposed algorithms. Considering this, we have proposed a methodology for evaluating the external knowledge base. In particular, the gazetteers. Until February 2021, there have been more than thirty citations to this work¹.

This proposed methodology evaluated gazetteers in the geoparsing task. It changed the size and scope of gazetteers while maintained algorithms and datasets constants. Each different gazetteer, named a focused gazetteer, contains only toponyms belonging to a geographical boundary, such as a country, region, or state. This characteristic of keeping the methods and dataset fixed and varying the gazetteers is the main innovation of the proposed methodology.

¹33 citations according to Google Scholar, and 23 according to metrics of ScienceDirect

One of the first difficulties was obtaining a dataset already labeled for use with the geoparsing task. As stated earlier, there is a lack of freely available geographic datasets in general [Gritta et al., 2018]. With that, we created a dataset composed of more than 500,000 news texts in Portuguese. Each news item has a preliminary geographical scope, indicated by the G1 portal editorial that stored the news. All the news items went through a geoparsing method (Algorithm 1) using Geonames. The algorithm worked as a Geo/Non-Geo classifier for each text. However, this classification was not definitive and still needed a manual check to confirm the results.

The Validation step depended on the voluntary participation of people. It was only possible to check a small part of the dataset, about 500 news items. If we had access to a previously labeled dataset, it would be possible to skip the Validation step. Also, it would be possible to analyze the focused gazetteer with the Reference Resolution task.

In the experimental evaluation, significant increases in precision were observed, rising from 23% (with Geonames) to 78% with a sub-state-focused gazetteer. Furthermore, this increase was consistent, not only comparing the full Geonames contents to the sub-state gazetteer, but also with the other focused gazetteers. The lowest precision gain occurs with the gazetteer that comprises toponyms for the entire country of Brazil, an increase of 74%.

Meanwhile, the recall only drops from 80% to 74%. The smaller size of the focused gazetteers explains this recall loss. For comparison purposes only, each sub-state gazetteer has less than 1% of the number of Geonames toponyms. For instance, *Belo Horizonte* news can mention places such as *Minas Gerais* or *Brazil*, and the sub-state-focused gazetteer would miss these two toponyms. So, a possible solution to avoid this recall loss is to keep a well-known toponym list. This list can contain, for instance, the country and the state that covers the sub-state region. This list of toponyms can prevent recall loss, but experiments to prove it still needs to be done.

In addition to increased precision, focused gazetteers also reduce the number of ambiguous candidates per toponym. The geographic delimitation offered by the focused gazetteers eliminates false candidates to the reference resolution step. In manual and punctual checks of some news texts, we noted that the candidates eliminated are really false ones. In this way, the focused gazetteers can contribute to disambiguation solutions, providing less ambiguous candidates by toponym.

The experimental results confirm the hypothesis of this work, and show that it is not necessary to use a broad generalist gazetteer to perform the geoparsing task efficiently. Focused gazetteers are smaller and require less effort for construction and maintenance. Furthermore, focused gazetteers can be built from broader ones using

simple selection techniques, or using gazetteer organization features such as hierarchical territory subdivisions. Also, as shown with the experiments described in this thesis, they produce a more efficient result in the geoparsing task, and facilitate the next step in the determination of the geographic scope of the document, geolocating, due to the significantly reduced toponym ambiguity.

It is worth mentioning that the use of focused gazetteers requires information on a rough geographical scope of the text. This primary scope is necessary to delimit the gazetteer geographically. In this sense, news previously organized by location corresponds to the type of text suitable for use. Nevertheless, the conclusion that a focused gazetteer suggests, as future research, the design of new geoparsing techniques that incrementally expand the scope of the gazetteer used looking for gains in recall, or that try to obtain a focused region from the comparative verification of ambiguous candidates, generated by the broadest gazetteer contents.

The lack of a previously labeled dataset and the dependence on volunteers to confirm the results were the main difficulties and limitations encountered. Extensions and continuations of this work are in Section 5.1.

In summary, the main contributions of this work were:

- A survey about the GSRP and its tasks (Geoparsing, Reference Resolution, and Grounding References), each one is defined considering its needs and their relation with the GSRP. Also, it contains a classification of the solutions present in the literature for the GSRP and its tasks [Monteiro et al., 2016];
- A dataset containing 529,585 news texts in Portuguese. All news items is associated to a primary geographic scope, provided by the identification of the subsection in which they appear in a news portal. Also, the dataset contains the results of Algorithm 1, using Geonames and other focused gazetteers, applied to each news item. This result still needs to be checked since it has false positives;
- A subset of the news dataset (500 news) with places annotated by people. This kind of dataset is useful to compare geoparsing methods or even to compare different solutions to the geoparsing step;
- A methodological approach to evaluate the external knowledge base used in GSRP or its tasks. Our methodology compares different gazetteers while keeping the method and the dataset fixed;
- A verification that focused-gazetteers increase the precision in the geoparsing task with a small loss of recall.

5.1 Future Work

This section describes open questions raised during the research for this thesis, proposed here as future work.

Even though we verified, in a manual check, that there was a reduction in the number of candidates for disambiguation when using focused gazetteers, our methodology does not contemplate this feature. So, future works can deal with reference resolution using focused gazetteers. Intuitively, smaller gazetteers should present fewer candidates for each toponym, and the idea would be to assess the size of this reduction compared to the size of the focused gazetteer. It is necessary to have a previously labeled dataset with all toponyms and their real location or use people to serve as Geo/Geo classifiers.

Another future work that we envision is the use of focused gazetteers with other geoparsing algorithms or even with other GSRP methods. As we describe in Chapter 2, there are types of geoparsing algorithms beyond the lookup-based used in this thesis, such as rule-based and supervised-based. These two types also use external knowledge sources, such as a gazetteer. The idea is to replicate this thesis's methodology, changing the method if the goal is to re-evaluate focused gazetteers. Still, news works can evaluate previous solutions (to Geoparsing, Reference Resolution, or even GSRP) with focused gazetteers to check for increased efficiency.

New solutions to GSRP and its tasks also can be researched with the focused gazetteers. The idea is to consider the preliminary geographic scope of the data to delimit the gazetteer. First, if the data does not have a primary geographic focus, it is necessary to infer one. Artificial intelligence techniques and NLP (Natural Language Processing) methods can help in this task, using metadata such as URLs, toponyms in titles, or web page source code. This process would correspond to the automation of the methodology used in this thesis. With dynamically generated focused gazetteers, there will be less effort to prepare the dataset to be geoparsed or disambiguated.

To generate focused gazetteers dynamically is necessary to think in new structures to store the gazetteers. The intention is to include in the gazetteer the inherent hierarchy between places. For example, countries have many states, and several cities belong to a state. Graphs and hierarchical databases appear as the first suggestions. Another relevant feature to put in gazetteers are the spatial relationships. These relationships allowed us to consult, for example, toponyms considering their geographical neighbors or their position in the hierarchy.

Bibliography

- Adams, B. and Janowicz, K. (2012). On the Geo-Indicativeness of Non-Georeferenced text. In *Proceedings of the International AAAI Conference on Web and Social Media*.
- Adelfio, M. D. and Samet, H. (2013). GeoWhiz: Toponym resolution using common categories. In *Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, SIGSPATIAL'13, pages 532–535, New York, NY, USA. ACM.
- Alencar, R. O. and Davis Jr., C. A. (2011). *Advancing Geoinformation Science for a Changing World*, volume 1, chapter Geotagging Aided by Topic Detection with Wikipedia, pages 461–477. Springer Berlin Heidelberg, Berlin, Heidelberg.
- Alencar, R. O., Davis Jr., C. A., and Gonçalves, M. A. (2010). Geographical classification of documents using evidence from Wikipedia. In *Proceedings of the 6th Workshop on Geographic Information Retrieval*, GIR '10, pages 12:1–12:8, New York, NY, USA. ACM.
- Alexopoulos, P. and Ruiz, C. (2012). Optimizing geographical entity and scope resolution in texts using non-geographical semantic information. In *Proceedings of the 6th International Conference on Advances in Semantic Processing*, SEMAPRO 2012, pages 65–70, Barcelona, Spain. Think Mind. Accessed: 2016-05-30.
- Alexopoulos, P., Ruiz, C., Villazon-Terrazas, B., and Gómez-Pérez, J. M. (2013). KLocator: An ontology-based framework for scenario-driven geographical scope resolution. *International Journal On Advances in Intelligent Systems*, 6(3 and 4):177–187. Accessed: 2016-05-30.
- Aloteibi, S. and Sanderson, M. (2014). Analyzing geographic query reformulation: An exploratory study. *Journal of the Association for Information Science and Technology*, 65(1):13–24.

- Amitay, E., Har'El, N., Sivan, R., and Soffer, A. (2004). Web-a-Where: Geotagging web content. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '04, pages 273–280, New York, NY, USA. ACM.
- Anastácio, I., Martins, B., and Calado, P. (2009). A comparison of different approaches for assigning geographic scopes to documents. In *Proceedings of the 1st InForum-Simpósio de Informática*, INForum '09, pages 285–296, Lisbon, Portugal. Accessed: 2016-05-31.
- Andogah, G., Bouma, G., and Nerbonne, J. (2012). Every document has a geographical scope. *Data & Knowledge Engineering*, 81-82:1–20.
- Baeza-Yates, R. A. and Ribeiro-Neto, B. (2011). *Modern Information Retrieval*. Addison-Wesley Professional, Boston, MA, USA, 2nd edition.
- Borges, K. A. V., Davis Jr., C. A., Laender, A. H. F., and Medeiros, C. B. (2011). Ontology-driven discovery of geospatial evidence in web pages. *GeoInformatica*, 15(4):609–631.
- Borges, K. A. V., Laender, A. H. F., Medeiros, C. B., and Davis Jr., C. A. (2007). Discovering geographic locations in web pages using urban addresses. In *Proceedings of the 4th ACM Workshop on Geographical Information Retrieval*, GIR '07, pages 31–36, New York, NY, USA. ACM.
- Brun, G., Dominguès, C., and Van Damme, M.-D. (2015). TEXTOMAP: Determining geographical window for texts. In *Proceedings of the 9th Workshop on Geographic Information Retrieval*, pages 1–2.
- Buscaldi, D. and Rosso, P. (2008). Map-based vs. knowledge-based toponym disambiguation. In *Proceedings of the 2nd International Workshop on Geographic Information Retrieval*, GIR '08, pages 19–22, New York, NY, USA. ACM.
- Buyukkokten, O., Cho, J., Garcia-Molina, H., Gravano, L., and Shivakumar, N. (1999). Exploiting geographical location information of web pages. In *ACM SIGMOD Workshop on The Web and Databases (WebDB'99)*, pages 91–96. Accessed: 2016-05-30.
- Campelo, C. E. C. and Baptista, C. S. (2008). Geographic scope modeling for web documents. In *Proceedings of the 2nd International Workshop on Geographic Information Retrieval*, GIR '08, pages 11–18, New York, NY, USA. ACM.

- Cardoso, N. (2011). Evaluating geographic information retrieval. *SIGSPATIAL Special*, 3(2):46–53.
- Cardoso, N., Silva, M. J., and Santos, D. (2008). Handling implicit geographic evidence for geographic IR. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management, CIKM '08*, pages 1383–1384, New York, NY, USA. ACM.
- Chasin, R., Woodward, D., Witmer, J., and Kalita, J. (2013). Extracting and displaying temporal and geospatial entities from articles on historical events. *The Computer Journal*, 57(3):403–426.
- Chen, M., Lin, X., Zhang, Y., Wang, X., and Yu, H. (2010). Assigning geographical focus to documents. In *2010 18th International Conference on Geoinformatic*, pages 1–6. IEEE.
- Clough, P. (2005). Extracting metadata for spatially-aware information retrieval on the Internet. In *Proceedings of the 2005 Workshop on Geographic Information Retrieval, GIR '05*, pages 25–30, New York, NY, USA. ACM.
- Clough, P., Sanderson, M., and Joho, H. (2004). Extraction of semantic annotations from textual web pages. Technical Report D15 6201, University of Sheffield. SPIRIT Project (EU IST-2001-35047). Accessed: 2016-05-30.
- Curran, J. R. and Clark, S. (2003). Language independent NER using a maximum entropy tagger. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4, CONLL '03*, pages 164–167, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Davis Jr., C. A. and Fonseca, F. T. (2007). Assessing the certainty of locations produced by an address geocoding system. *GeoInformatica*, 11(1):103–129.
- Delboni, T. M., Borges, K. A. V., Laender, A. H. F., and Davis Jr., C. A. (2007). Semantic expansion of geographic web queries based on natural language positioning expressions. *Transactions in GIS*, 11(3):377–397.
- DeLozier, G., Baldrige, J., and London, L. (2015). Gazetteer-independent toponym resolution using geographic word profiles. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence, AAAI'15*, pages 2382–2388. AAAI Press. Accessed: 2016-05-30.
- Di Rocco, L., Dassereto, F., Bertolotto, M., Buscaldi, D., Catania, B., and Guerrini, G. (2020). Sherlock: a knowledge-driven algorithm for geolocating microblog messages

- at sub-city level. *International Journal of Geographical Information Science*, pages 1–32.
- Ding, J., Gravano, L., and Shivakumar, N. (2000). Computing geographical scopes of web resources. In *Proceedings of the 26th International Conference on Very Large Data Bases, VLDB '00*, pages 545–556, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc. Accessed: 2016-05-30.
- Elwood, S., Goodchild, M. F., and Sui, D. Z. (2012). Researching volunteered geographic information: Spatial data, geographic research, and new social practice. *Annals of the Association of American Geographers*, 102(3):571–590.
- Freitas, A. L., Davis Jr, C. A., and Filgueiras, T. M. (2012). GeoSQL: um ambiente online para aprendizado de SQL com extensoes espaciais. *XIII Simpósio Brasileiro de Geoinformática (GeoInfo 2012)*, 2012:146–151.
- Fu, G., Jones, C. B., and Abdelmoty, A. I. (2005). Building a geographical ontology for intelligent spatial search on the web. In Hamza, M. H., editor, *Proceedings of IASTED International Conference on Databases and Applications, DBA 2005*, pages 167–172, Anaheim, CA, USA. IASTED/ACTA Press. Accessed: 2016-05-31.
- Fujiwara, Y., Sakurai, Y., and Kitsuregawa, M. (2009). Fast likelihood search for Hidden Markov Models. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 3(4):18:1–18:37.
- Gao, S., Li, L., Li, W., Janowicz, K., and Zhang, Y. (2017). Constructing gazetteers from volunteered Big Geo-Data based on Hadoop. *Computers, Environment and Urban Systems*, 61:172–186.
- Garbin, E. and Mani, I. (2005). Disambiguating toponyms in news. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing, HLT '05*, pages 363–370, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Gelernter, J., Ganesh, G., Krishnakumar, H., and Zhang, W. (2013). Automatic gazetteer enrichment with user-geocoded data. In *Proceedings of the Second ACM SIGSPATIAL International Workshop on Crowdsourced and Volunteered Geographic Information*, pages 87–94.
- Gelernter, J. and Mushegian, N. (2011). Geo-parsing messages from microtext. *Transactions in GIS*, 15(6):753–773.

- Goldberg, D. W., Wilson, J. P., and Knoblock, C. A. (2007). From text to geographic coordinates: The current state of geocoding. *URISA Journal (Journal of the Urban and Regional Information Association)*, 19(1):33–47. Accessed: 2016-05-30.
- Goodchild, M. F. and Hill, L. L. (2008). Introduction to digital gazetteer research. *International Journal of Geographical Information Science*, 22(10):1039–1044.
- Gouvêa, C., Loh, S., Garcia, L. F. F., Fonseca, E. B., and Wendt, I. (2008). Discovering location indicators of toponyms from news to improve gazetteer-based georeferencing. In Carvalho, M. T. M., Casanova, M. A., Gattass, M., and Vinhas, L., editors, *Proceedings of the X Brazilian Symposium on GeoInformatics, GeoInfo 2008*, pages 51–62, Porto Alegre, RS, Brazil. SBC. Accessed: 2016-05-31.
- Gritta, M., Pilehvar, M. T., and Collier, N. (2019). A pragmatic guide to geoparsing evaluation. *Language Resources and Evaluation*, pages 1–30.
- Gritta, M., Pilehvar, M. T., Limsopatham, N., and Collier, N. (2018). What’s missing in geographical parsing? *Language Resources and Evaluation*, 52(2):603–623.
- Habib, M. B. and van Keulen, M. (2011). Named entity extraction and disambiguation: The reinforcement effect. In *Proceedings of the 5th International Workshop on Management of Uncertain Data, MUD 2011, Seattle, USA*, volume WP11-02 of *CTIT Workshop Proceedings Series*, pages 9–16, Enschede. Centre for Telematics and Information Technology University of Twente. Accessed: 2016-05-31.
- Habib, M. B. and van Keulen, M. (2012). *Web Engineering: 12th International Conference, ICWE 2012, Berlin, Germany, July 23-27, 2012. Proceedings*, chapter Improving Toponym Extraction and Disambiguation Using Feedback Loop, pages 439–443. Springer Berlin Heidelberg, Berlin, Heidelberg.
- Habib, M. B. and van Keulen, M. (2013). A hybrid approach for robust multilingual toponym extraction and disambiguation. In Kaopotek, M. A., Koronacki, J., Marciniak, M., Mykowiecka, A., and Wierzchon, S. T., editors, *Language Processing and Intelligent Information Systems*, volume 7912 of *Lecture Notes in Computer Science*, pages 1–15. Springer Berlin Heidelberg.
- Hearst, M. A. (1998). Support Vector Machines. *IEEE Intelligent Systems*, 13(4):18–28.
- Hill, L. L. (2000). *Research and Advanced Technology for Digital Libraries: 4th European Conference, ECDL 2000 Lisbon, Portugal, September 18-20, 2000 Proceedings*,

- chapter Core Elements of Digital Gazetteers: Placenames, Categories, and Footprints, pages 280–290. Springer Berlin Heidelberg, Berlin, Heidelberg.
- Hill, L. L. (2006). *Georeferencing: The geographic associations of information*. The MIT Press.
- Hill, L. L. (2009). *Gazetteers*, pages 1217–1218. Springer US, Boston, MA.
- Hu, Y. and Adams, B. (2020). *Harvesting Big Geospatial Data from Natural Language Texts*, chapter 1. Springer.
- Hu, Y., Mao, H., and McKenzie, G. (2019). A natural language processing and geospatial clustering framework for harvesting local place names from geotagged housing advertisements. *International Journal of Geographical Information Science*, 33(4):714–738.
- Inkpen, D., Liu, J., Farzindar, A., Kazemi, F., and Ghazi, D. (2017). Location detection and disambiguation from Twitter messages. *Journal of Intelligent Information Systems*, 49(2):237–253.
- Jones, C. B. and Purves, R. S. (2008). Geographical Information Retrieval. *International Journal of Geographical Information Science*, 22(3):219–228.
- Jones, C. B. and Purves, R. S. (2009). *Encyclopedia of Database Systems*, chapter Geographic Information Retrieval, pages 1227–1231. Springer US, Boston, MA.
- Karimzadeh, M., Huang, W., Banerjee, S., Wallgrün, J. O., Hardisty, F., Pezanowski, S., Mitra, P., and MacEachren, A. M. (2013). GeoTxt: a web API to leverage place references in text. In *Proceedings of the 7th workshop on geographic information retrieval*, pages 72–73.
- Leidner, J. L. (2007). *Toponym Resolution in Text: Annotation, Evaluation and Applications of Spatial Grounding of Place Names*. PhD thesis, University of Edinburgh.
- Leidner, J. L. (2008). *Toponym Resolution in Text: Annotation, Evaluation and Applications of Spatial Grounding of Place Names*. Universal-Publishers.
- Leidner, J. L. and Lieberman, M. D. (2011). Detecting geographical references in the form of place names and associated spatial natural language. *SIGSPATIAL Special*, 3(2):5–11.

- Leidner, J. L., Sinclair, G., and Webber, B. (2003). Grounding spatial named entities for information extraction and question answering. In *Proceedings of the HLT-NAACL 2003 Workshop on Analysis of Geographic References - Volume 1*, HLT-NAACL-GEOREF '03, pages 31–38, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Leveling, J., Hartrumpf, S., and Veiel, D. (2006). Using semantic networks for geographic information retrieval. In Peters, C., Gey, F. C., Gonzalo, J., Müller, H., Jones, G., Kluck, M., Magnini, B., and de Rijke, M., editors, *Accessing Multilingual Information Repositories*, volume 4022 of *Lecture Notes in Computer Science*, pages 977–986. Springer Berlin Heidelberg.
- Li, H., Srihari, R. K., Niu, C., and Li, W. (2002). Location normalization for information extraction. In *Proceedings of the 19th International Conference on Computational Linguistics - Volume 1*, COLING '02, pages 1–7, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Li, Y., Moffat, A., Stokes, N., and Cavedon, L. (2006). Exploring probabilistic toponym resolution for geographical information retrieval. In Purves, R. and Jones, C., editors, *Proceedings of the 3rd ACM Workshop On Geographic Information Retrieval*, SIGIR 2006, Seattle, WA, USA. Department of Geography, University of Zurich. Accessed: 2016-05-30.
- Lieberman, M. D. and Samet, H. (2011). Multifaceted toponym recognition for streaming news. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, pages 843–852, New York, NY, USA. ACM.
- Lieberman, M. D. and Samet, H. (2012a). Adaptive context features for toponym resolution in streaming news. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12, pages 731–740, New York, NY, USA. ACM.
- Lieberman, M. D. and Samet, H. (2012b). Supporting rapid processing and interactive map-based exploration of streaming news. In *Proceedings of the 20th International Conference on Advances in Geographic Information Systems*, SIGSPATIAL '12, pages 179–188, New York, NY, USA. ACM.
- Lieberman, M. D., Samet, H., and Sankaranarayanan, J. (2010). Geotagging with local lexicons to build indexes for textually-specified spatial data. In *Proceedings of*

- the *IEEE 26th International Conference on Data Engineering (ICDE 2010)*, pages 201–212. IEEE.
- Lott, R. (2019). *Geographic Information - Well-Known Text Representation of Coordinate Reference Systems*. Open Geospatial Consortium, Wayland, MA, USA, 2 edition.
- Luo, J., Joshi, D., Yu, J., and Gallagher, A. (2011). Geotagging in multimedia and computer vision - a survey. *Multimedia Tools and Applications*, 51(1):187–211.
- Machado, I. M. R., Alencar, R. O., Campos Jr., R. O., and Davis Jr., C. A. (2011). An ontological gazetteer and its application for place name disambiguation in text. *Journal of the Brazilian Computer Society*, 17(4):267–279.
- Martins, B., Cardoso, N., Chaves, M. S., Andrade, L., and Silva, M. J. (2007). The university of Lisbon at GeoCLEF 2006. In Peters, C., Clough, P., Gey, F. C., Karlgren, J., Magnini, B., Oard, D. W., de Rijke, M., and Stempfhuber, M., editors, *Evaluation of Multilingual and Multi-modal Information Retrieval*, volume 4730 of *Lecture Notes in Computer Science*, pages 986–994. Springer Berlin Heidelberg.
- Martins, B., Silva, M. J., Freitas, S., and Afonso, A. P. (2006). Handling locations in search engine queries. *GIR*, 6:1–6.
- McCurley, K. S. (2001). Geospatial mapping and navigation of the web. In *Proceedings of the 10th International Conference on World Wide Web, WWW '01*, pages 221–229, New York, NY, USA. ACM.
- Miller, H. J. and Goodchild, M. F. (2015). Data-driven geography. *GeoJournal*, 80(4):449–461.
- Monteiro, B. R., Davis Jr., C. A., and Fonseca, F. (2016). A survey on the geographic scope of textual documents. *Computers & Geosciences*, 96:23–34.
- Morimoto, Y., Aono, M., Houle, M. E., and McCurley, K. S. (2003). Extracting spatial knowledge from the web. In *Symposium on Applications and the Internet*, pages 326–333.
- Moura, T. H. V. M. and Davis Jr., C. A. (2014). Integration of linked data sources for gazetteer expansion. In *Proceedings of the 8th ACM SIGSPATIAL Workshop on Geographic Information Retrieval, GIR '14*, pages 5:1–5:8, New York, NY, USA. ACM.

- Nadeau, D. and Sekine, S. (2007). A survey of named entity recognition and classification. *Linguisticae Investigationes*, 30:3–26. Accessed: 2016-05-30.
- Nissim, M., Matheson, C., and Reid, J. (2004). Recognising geographical entities in scottish historical documents. In Purves, R. and Jones, C., editors, *Proceedings of the Workshop On Geographic Information Retrieval, SIGIR 2004, Sheffield, England, July, 25, 2004*, Sheffield, England. Department of Geography, University of Zurich. Accessed: 2016-05-30.
- Nizzoli, L., Avvenuti, M., Tesconi, M., and Cresci, S. (2020). Geo-semantic-parsing: AI-powered geoparsing by traversing semantic knowledge graphs. *Decision Support Systems*, 136:113346.
- Olligschlaeger, A. M. and Hauptmann, A. G. (1999). Multimodal information systems and GIS: The informedia digital video library. In *1999 ESRI User Conference*. Environmental Systems Research Institute (ESRI) Inc. Accessed: 2016-05-30.
- Page, L., Brin, S., Motwani, R., and Winograd, T. (1999). The PageRank citation ranking: Bringing order to the web. Technical Report, Stanford University. Accessed: 2016-05-30.
- Popescu, A., Grefenstette, G., and Moëllic, P. A. (2008). Gazetiki: automatic creation of a geographical gazetteer. In *Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries*, pages 85–93.
- Pouliquen, B., Kimler, M., Steinberger, R., Ignat, C., Oellinger, T., Blackler, K., Fuart, F., Zaghouni, W., Widiger, A., Forslund, A.-C., and Best, C. (2006). Geocoding multilingual texts: Recognition, disambiguation and visualisation. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006)*, pages 53–58, Genoa, Italy. Accessed: 2016-05-30.
- Pouliquen, B., Steinberger, R., Ignat, C., and Groeve, T. (2004). Geographical information recognition and visualization in texts written in various languages. In *Proceedings of the 2004 ACM Symposium on Applied Computing, SAC '04*, pages 1051–1058, New York, NY, USA. ACM.
- Purves, R. S., Clough, P., Jones, C. B., Arampatzis, A., Bucher, B., Finch, D., Fu, G., Joho, H., Syed, A. K., Vaid, S., and Yang, B. (2007). The design and implementation of SPIRIT: a spatially aware search engine for information retrieval on the Internet. *International Journal of Geographical Information Science*, 21(7):717–745.

- Purves, R. S., Clough, P., Jones, C. B., Hall, M. H., and Murdock, V. (2018). Geographic information retrieval: Progress and challenges in spatial search of text. *Foundations and Trends[®] in Information Retrieval*, 12(2-3):164–318.
- Rafiei, J. Y. and Rafiei, D. (2016). Geotagging named entities in news and online documents. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pages 1321–1330.
- Rauch, E., Bukatin, M., and Baker, K. (2003). A confidence-based framework for disambiguating geographic terms. In *Proceedings of the HLT-NAACL 2003 Workshop on Analysis of Geographic References - Volume 1, HLT-NAACL-GEOREF '03*, pages 50–54, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Ribeiro Jr., S. S., Davis Jr., C. A., Oliveira, D. R. R., Meira Jr., W., Gonçalves, T. S., and Pappa, G. L. (2012). Traffic Observatory: A system to detect and locate traffic events and conditions using Twitter. In *Proceedings of the 5th ACM SIGSPATIAL International Workshop on Location-Based Social Networks, LBSN '12*, pages 5–11, New York, NY, USA. ACM.
- Sanderson, M. and Kohler, J. (2004). Analyzing geographic queries. In Purves, R. and Jones, C., editors, *Proceedings of the Workshop On Geographic Information Retrieval, (SIGIR 2004)*, Sheffield, England. Department of Geography, University of Zurich. Accessed: 2016-05-30.
- Santos, J., Anastácio, I., and Martins, B. (2015). Using machine learning methods for disambiguating place references in textual documents. *GeoJournal*, 80(3):375–392.
- Santos, R., Murrieta-Flores, P., and Martins, B. (2018). Learning to combine multiple string similarity metrics for effective toponym matching. *International Journal of Digital Earth*, 11(9):913–938.
- Serdyukov, P., Murdock, V., and Van Zwol, R. (2009). Placing Flickr photos on a map. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 484–491.
- Shi, G. and Barker, K. (2011). Extraction of geospatial information on the web for GIS applications. In *10th IEEE International Conference on Cognitive Informatics Cognitive Computing, ICC*CC*, pages 41–48. IEEE.
- Silva, M. J., Martins, B., Chaves, M. S., Afonso, A. P., and Cardoso, N. (2006). Adding geographic scopes to web resources. *Computers, Environment and Urban Systems*, 30(4):378–399.

- Singleton, A. and Arribas-Bel, D. (2021). Geographic data science. *Geographical Analysis*, 53(1):61–75.
- Smith, D. A. and Crane, G. (2001). Disambiguating geographic names in a historical digital library. In *Proceedings of the 5th European Conference on Research and Advanced Technology for Digital Libraries*, ECDL '01, pages 127–136, London, UK, UK. Springer-Verlag.
- Smith, D. A. and Mann, G. S. (2003). Bootstrapping toponym classifiers. In *Proceedings of the HLT-NAACL 2003 Workshop on Analysis of Geographic References - Volume 1*, HLT-NAACL-GEOREF '03, pages 45–49, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Souza, L. A., Davis Jr., C. A., Borges, K. A. V., Delboni, T. M., and Laender, A. H. F. (2005). The role of gazetteers in geographic knowledge discovery on the web. In *Proceedings of the 3rd Latin American Web Congress*, LA-WEB '05, pages 157–, Washington, DC, USA. IEEE Computer Society.
- Speriosu, M. and Baldrige, J. (2013). Text-driven toponym resolution using indirect supervision. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1466–1476, Sofia, Bulgaria. Association for Computational Linguistics. Accessed: 2016-05-30.
- Teitler, B. E., Lieberman, M. D., Panozzo, D., Sankaranarayanan, J., Samet, H., and Sperling, J. (2008). NewsStand: A new view on news. In *Proceedings of the 16th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, GIS '08, pages 18:1–18:10, New York, NY, USA. ACM.
- Twaroch, F. A., Smart, P. D., and Jones, C. B. (2008). Mining the web to detect place names. In *Proceedings of the 2nd International Workshop on Geographic Information Retrieval*, GIR '08, pages 43–44, New York, NY, USA. ACM.
- Vaid, S., Jones, C. B., Joho, H., and Sanderson, M. (2005). Spatio-textual indexing for geographical search on the web. In *Proceedings of the 9th International Conference on Advances in Spatial and Temporal Databases*, SSTD'05, pages 218–235, Berlin, Heidelberg, Germany. Springer-Verlag.
- Vargas, R. N. P., Moura, M. F., Speranza, E. A., Rodriguez, E., and Rezende, S. O. (2012a). Discovering the spatial coverage of the documents through the SpatialCIM methodology. In *Proceedings of the 15th AGILE'2012 International Conference on*

- Geographic Information Science, Avignon, France, April*, pages 181–186. Accessed: 2016-05-30.
- Vargas, R. N. P., Moura, M. F., Speranza, E. A., Rodriguez, E., and Rezende, S. O. (2012b). The SpatialCIM methodology for spatial document coverage disambiguation and the entity recognition process aided by linguistic techniques. In *Geospatial Information and Documents. Pacif-Asia Conference on Knowledge Discovery and Data Mining, 16. (Geo Doc 2012), PAKDD Workshop*, Kuala Lumpur, Malaysia. Accessed: 2016-05-31.
- Vasardani, M., Winter, S., and Richter, K.-F. (2013). Locating place names from place descriptions. *International Journal of Geographical Information Science*, 27(12):2509–2532.
- Volz, R., Kleb, J., and Mueller, W. (2007). Towards ontology-based disambiguation of geographical identifiers. In Bouquet, P., Stoermer, H., Tummarello, G., and Halpin, H., editors, *Proceedings of the WWW2007 Workshop i3: Identity, Identifiers, Identification*, Banff, Canada. CEUR Workshop Proceedings. Accessed: 2016-05-30.
- Wang, C., Xie, X., Wang, L., Lu, Y., and Ma, W.-Y. (2005). Detecting geographic locations from web resources. In *Proceedings of the 2005 Workshop on Geographic Information Retrieval, GIR '05*, pages 17–24, New York, NY, USA. ACM.
- Woodruff, A. G. and Plaunt, C. (1994). GIPSY: Automated geographic indexing of text documents. *Journal of the American Society for Information Science - Special Issue: Spatial Information*, 45(9):645–655.
- Yadav, A. K., Yadav, J. K. P. S., and Yadav, D. (2020). Spatial ambiguities optimization in GIR. *EAI Endorsed Transactions on Scalable Information Systems*, 7(28).
- Zhang, Q., Jin, P., Lin, S., and Yue, L. (2012). Extracting focused locations for web pages. In *Proceedings of the 2011 International Conference on Web-Age Information Management, WAIM'11*, pages 76–89, Berlin, Heidelberg. Springer-Verlag.
- Zong, W., Wu, D., Sun, A., Lim, E.-P., and Goh, D. H.-L. (2005). On assigning place names to geography related web pages. In *Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries, JCDL '05*, pages 354–362, New York, NY, USA. ACM.
- Zubizarreta, A., Fuente, P., Cantera, J. M., Arias, M., Cabrero, J., García, G., Llamas, C., and Vegas, J. (2008). A georeferencing multistage method for locating geographic

context in web search. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management, CIKM '08*, pages 1485–1486, New York, NY, USA. ACM.

Appendix A

Results for G1 Minas Gerais sub-section news

This appendix has individual tables with results for all G1 *Minas Gerais* sub-sections news. In all tables, *PR*, *RC*, and *F1* means, respectively, Precision, Recall and F1 Score. The *N* column shows the news id, and the next five large columns show the values of precision and F1 score using, respectively, the gazetteers: Geonames, country-focused (*Brazil*), region-focused (*Sudeste*), state-focused (*Minas Gerais*), and sub-state-focused (*Belo Horizonte e Região*). As recall values are constant, only the column Geonames presents these values.

In Table A.1 the column *CO* means *Centro-Oeste* G1 sub-section.

Table A.1: Centro-Oeste sub-state news evaluation

N	Geonames			Brazil		Southeast		MG		CO	
	PR	RC	F1	PR	F1	PR	F1	PR	F1	PR	F1
81	29%	50%	37%	67%	57%	67%	57%	67%	57%	67%	57%
82	25%	75%	38%	60%	67%	75%	75%	75%	75%	75%	75%
83	23%	100%	37%	43%	60%	50%	67%	50%	67%	67%	80%
84	44%	100%	61%	80%	89%	100%	100%	100%	100%	100%	100%
85	30%	100%	46%	43%	60%	60%	75%	60%	75%	60%	75%
86	33%	100%	50%	100%	100%	100%	100%	100%	100%	100%	100%
87	33%	80%	47%	80%	80%	100%	89%	100%	89%	100%	89%
88	30%	75%	43%	75%	75%	100%	86%	100%	86%	100%	86%
89	50%	86%	63%	80%	80%	100%	89%	100%	89%	100%	89%
90	18%	50%	26%	33%	40%	50%	50%	50%	50%	50%	50%

Table A.1 continued from previous page

N	Geonames			Brazil		Southeast		MG		CO	
	PR	RC	F1	PR	F1	PR	F1	PR	F1	PR	F1
91	18%	100%	31%	33%	50%	100%	100%	100%	100%	100%	100%
92	30%	100%	46%	50%	67%	100%	100%	100%	100%	100%	100%
93	40%	100%	57%	100%	100%	100%	100%	100%	100%	100%	100%
94	20%	67%	31%	50%	57%	100%	80%	100%	80%	100%	80%
95	42%	78%	55%	50%	61%	54%	64%	82%	80%	88%	83%
96	50%	100%	67%	67%	80%	100%	100%	100%	100%	100%	100%
97	27%	60%	37%	50%	55%	75%	67%	75%	67%	75%	67%
98	20%	50%	29%	50%	50%	67%	57%	67%	57%	67%	57%
99	18%	43%	25%	27%	33%	50%	46%	60%	50%	60%	50%
100	21%	100%	35%	60%	75%	75%	86%	75%	86%	75%	86%
101	10%	33%	15%	15%	21%	38%	35%	60%	43%	100%	50%
102	25%	75%	38%	60%	67%	75%	75%	75%	75%	75%	75%
103	33%	43%	37%	60%	50%	100%	60%	100%	50%	100%	50%
104	25%	67%	36%	67%	67%	100%	80%	100%	80%	100%	80%
105	32%	86%	47%	86%	86%	86%	86%	100%	92%	100%	92%
106	22%	67%	33%	50%	57%	67%	67%	67%	67%	67%	67%
107	33%	75%	46%	75%	75%	100%	86%	100%	86%	100%	86%
108	18%	67%	28%	33%	44%	50%	57%	50%	57%	50%	57%
109	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
110	20%	67%	31%	50%	57%	67%	67%	67%	67%	67%	67%
111	25%	75%	38%	60%	67%	75%	75%	75%	75%	75%	75%
112	23%	50%	32%	38%	43%	38%	43%	38%	43%	60%	55%
113	22%	67%	33%	50%	57%	67%	67%	67%	67%	67%	67%
114	60%	67%	63%	86%	75%	100%	80%	100%	77%	100%	67%
115	13%	67%	22%	18%	28%	29%	40%	67%	67%	67%	67%
116	18%	67%	28%	33%	44%	50%	57%	50%	57%	50%	57%
117	38%	75%	50%	50%	60%	75%	75%	75%	75%	75%	75%
118	27%	67%	38%	43%	52%	55%	60%	67%	67%	83%	71%
119	55%	100%	71%	83%	91%	100%	100%	100%	100%	100%	100%
120	30%	75%	43%	75%	75%	100%	86%	100%	86%	100%	86%
121	22%	67%	33%	50%	57%	67%	67%	67%	67%	67%	67%
122	27%	75%	40%	50%	60%	75%	75%	100%	86%	100%	86%

Table A.1 continued from previous page

N	Geonames			Brazil		Southeast		MG		CO	
	PR	RC	F1	PR	F1	PR	F1	PR	F1	PR	F1
123	30%	75%	43%	75%	75%	100%	86%	100%	86%	100%	86%
124	21%	100%	35%	27%	43%	60%	75%	100%	100%	100%	100%
125	27%	75%	40%	50%	60%	75%	75%	100%	86%	100%	86%
126	40%	100%	57%	67%	80%	100%	100%	100%	100%	100%	100%
127	26%	83%	40%	45%	58%	71%	77%	100%	91%	100%	89%
128	18%	67%	28%	33%	44%	50%	57%	67%	67%	67%	67%
129	33%	80%	47%	67%	73%	100%	89%	100%	89%	100%	89%
130	23%	75%	35%	33%	46%	43%	55%	60%	67%	60%	67%
131	22%	100%	36%	50%	67%	67%	80%	67%	80%	67%	80%
132	15%	100%	26%	20%	33%	29%	45%	67%	80%	67%	80%
133	22%	100%	36%	50%	67%	67%	80%	67%	80%	67%	80%
134	27%	75%	40%	50%	60%	75%	75%	100%	86%	100%	86%
135	45%	83%	58%	73%	76%	80%	80%	80%	80%	80%	80%
136	20%	67%	31%	50%	57%	67%	67%	67%	67%	67%	67%
137	22%	67%	33%	50%	57%	67%	67%	67%	67%	67%	67%
138	22%	67%	33%	50%	57%	67%	67%	67%	67%	67%	67%
139	36%	80%	50%	80%	80%	100%	89%	100%	89%	100%	89%
140	33%	43%	37%	75%	55%	100%	60%	100%	60%	100%	60%
141	14%	67%	23%	20%	31%	33%	44%	40%	50%	40%	50%
142	22%	67%	33%	50%	57%	67%	67%	67%	67%	67%	67%
143	50%	50%	50%	50%	50%	67%	57%	67%	57%	67%	57%
144	22%	67%	33%	50%	57%	67%	67%	67%	67%	67%	67%
145	50%	100%	67%	100%	100%	100%	100%	100%	100%	100%	100%
146	23%	100%	37%	44%	61%	58%	73%	60%	75%	75%	86%
147	33%	83%	47%	45%	58%	62%	71%	71%	77%	71%	77%
148	33%	75%	46%	75%	75%	100%	86%	100%	86%	100%	86%
149	25%	75%	38%	60%	67%	75%	75%	75%	75%	75%	75%
150	11%	67%	19%	20%	31%	40%	50%	67%	67%	67%	67%
151	36%	80%	50%	80%	80%	100%	89%	100%	89%	100%	89%
152	33%	100%	50%	67%	80%	75%	86%	75%	86%	75%	86%
153	46%	100%	63%	75%	86%	100%	100%	100%	100%	100%	100%
154	25%	75%	38%	60%	67%	75%	75%	75%	75%	75%	75%

Table A.2 continued from previous page

N	Geonames			Brazil		Southeast		MG		GM	
	PR	RC	F1	PR	F1	PR	F1	PR	F1	PR	F1
180	23%	75%	35%	30%	43%	35%	48%	31%	42%	75%	67%
181	9%	100%	17%	18%	31%	20%	33%	20%	33%	100%	100%
182	15%	100%	26%	40%	57%	50%	67%	67%	80%	100%	100%
183	16%	50%	24%	43%	46%	75%	60%	75%	60%	100%	67%
184	6%	100%	11%	9%	17%	11%	20%	11%	20%	50%	67%
185	22%	100%	36%	45%	62%	62%	77%	62%	77%	83%	91%
186	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
187	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
188	10%	100%	18%	14%	25%	20%	33%	33%	50%	100%	100%
189	8%	100%	15%	12%	21%	13%	23%	50%	67%	100%	100%
190	6%	100%	11%	17%	29%	25%	40%	0%	0%	0%	0%
191	10%	100%	18%	18%	31%	29%	45%	40%	57%	100%	100%
192	29%	100%	45%	45%	62%	50%	67%	56%	72%	80%	89%
193	14%	100%	25%	40%	57%	40%	57%	40%	57%	67%	80%
194	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
195	16%	75%	26%	33%	46%	100%	86%	100%	80%	100%	67%
196	25%	100%	40%	36%	53%	36%	53%	38%	55%	50%	67%
197	8%	100%	15%	10%	18%	10%	18%	14%	25%	33%	50%
198	4%	100%	8%	8%	15%	33%	50%	50%	67%	100%	100%
199	18%	67%	28%	22%	33%	67%	67%	100%	80%	100%	80%
200	18%	80%	29%	22%	35%	31%	45%	33%	46%	100%	86%
201	19%	100%	32%	27%	43%	75%	86%	75%	86%	100%	100%
202	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
203	12%	50%	19%	18%	26%	50%	44%	100%	57%	100%	57%
204	45%	83%	58%	62%	71%	80%	80%	100%	89%	100%	89%
205	9%	60%	16%	20%	30%	33%	43%	29%	37%	67%	57%
206	27%	88%	41%	54%	67%	86%	86%	86%	86%	100%	91%
207	11%	100%	20%	50%	67%	100%	100%	100%	100%	100%	100%
208	12%	60%	20%	16%	25%	50%	55%	100%	75%	100%	75%
209	6%	50%	11%	11%	18%	17%	25%	17%	25%	100%	67%
210	11%	50%	18%	18%	26%	12%	18%	25%	28%	50%	40%
211	8%	100%	15%	12%	21%	33%	50%	33%	50%	100%	100%

Table A.2 continued from previous page

N	Geonames			Brazil		Southeast		MG		GM	
	PR	RC	F1	PR	F1	PR	F1	PR	F1	PR	F1
212	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
213	8%	100%	15%	12%	21%	20%	33%	33%	50%	100%	100%
214	15%	100%	26%	31%	47%	83%	91%	67%	80%	100%	100%
215	14%	100%	25%	18%	31%	29%	45%	33%	50%	50%	67%
216	18%	80%	29%	36%	50%	50%	62%	80%	80%	100%	89%
217	6%	100%	11%	11%	20%	50%	67%	50%	67%	100%	100%
218	22%	100%	36%	39%	56%	53%	69%	58%	73%	75%	86%
219	35%	86%	50%	46%	60%	38%	50%	60%	67%	100%	86%
221	21%	67%	32%	36%	47%	75%	67%	100%	67%	100%	67%
222	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
223	12%	100%	21%	33%	50%	40%	57%	0%	0%	0%	0%
224	11%	71%	19%	31%	43%	83%	77%	100%	83%	100%	75%
225	14%	75%	24%	21%	33%	43%	55%	100%	86%	100%	86%
226	23%	100%	37%	38%	55%	75%	86%	100%	100%	100%	100%
227	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
228	12%	100%	21%	30%	46%	38%	55%	50%	67%	100%	100%
229	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
230	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
231	17%	100%	29%	33%	50%	33%	50%	33%	50%	50%	67%
232	15%	75%	25%	25%	38%	38%	50%	25%	33%	100%	67%
233	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
234	7%	100%	13%	10%	18%	20%	33%	20%	33%	50%	67%
236	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
238	14%	100%	25%	20%	33%	25%	40%	50%	67%	100%	100%
239	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
240	17%	100%	29%	38%	55%	60%	75%	67%	80%	67%	80%

In Table A.3 the column *SM* means *Sul de Minas* G1 sub-section.

Table A.3: Sul de Minas sub-state news evaluation

N	Geonames			Brazil		Southeast		MG		SM	
	PR	RC	F1	PR	F1	PR	F1	PR	F1	PR	F1
241	52%	100%	68%	48%	65%	54%	70%	67%	80%	83%	91%
242	40%	55%	46%	67%	60%	50%	43%	50%	43%	60%	47%
245	33%	93%	49%	59%	72%	46%	60%	67%	75%	86%	86%
246	3%	100%	6%	6%	11%	11%	20%	17%	29%	0%	0%
247	52%	100%	68%	88%	94%	91%	95%	94%	97%	93%	96%
248	38%	85%	53%	55%	67%	65%	74%	77%	80%	77%	80%
249	47%	94%	63%	67%	78%	83%	88%	83%	87%	100%	94%
250	28%	78%	41%	50%	61%	64%	70%	70%	74%	78%	78%
251	20%	100%	33%	75%	86%	75%	86%	75%	86%	75%	86%
252	54%	88%	67%	73%	79%	77%	81%	81%	81%	81%	81%
253	85%	92%	88%	100%	95%	100%	94%	100%	92%	100%	91%
254	44%	100%	61%	100%	100%	100%	100%	100%	100%	100%	100%
255	27%	92%	42%	48%	63%	58%	71%	85%	88%	92%	92%
256	48%	77%	59%	67%	72%	79%	76%	74%	70%	67%	63%
257	23%	86%	36%	55%	67%	60%	71%	67%	73%	80%	80%
259	48%	91%	63%	75%	82%	89%	89%	100%	94%	100%	94%
261	40%	100%	57%	67%	80%	80%	89%	80%	89%	100%	100%
262	31%	92%	46%	58%	71%	79%	85%	85%	88%	78%	83%
263	48%	83%	61%	55%	66%	50%	60%	73%	73%	69%	69%
265	48%	88%	62%	80%	83%	100%	90%	100%	89%	100%	87%
266	32%	85%	46%	50%	63%	61%	71%	67%	73%	67%	73%
267	23%	80%	36%	38%	52%	57%	67%	60%	67%	67%	71%
268	19%	80%	31%	50%	62%	67%	73%	80%	80%	80%	80%
269	41%	88%	56%	64%	74%	70%	78%	70%	78%	88%	88%
270	17%	75%	28%	38%	50%	50%	60%	60%	67%	75%	75%
271	57%	95%	71%	74%	83%	76%	84%	75%	83%	79%	86%
273	33%	100%	50%	100%	100%	100%	100%	100%	100%	100%	100%
274	23%	100%	37%	75%	86%	75%	86%	67%	80%	100%	100%
275	29%	67%	40%	100%	80%	100%	80%	100%	80%	100%	80%
276	25%	100%	40%	75%	86%	100%	100%	100%	100%	100%	100%
277	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
278	26%	85%	40%	37%	52%	46%	60%	69%	75%	75%	78%

Table A.3 continued from previous page

N	Geonames			Brazil		Southeast		MG		SM	
	PR	RC	F1	PR	F1	PR	F1	PR	F1	PR	F1
279	26%	86%	40%	50%	63%	75%	80%	86%	86%	100%	92%
280	25%	83%	38%	50%	62%	71%	77%	83%	83%	83%	83%
281	14%	100%	25%	27%	43%	30%	46%	38%	55%	50%	67%
282	28%	78%	41%	38%	50%	50%	59%	56%	63%	62%	66%
283	23%	100%	37%	46%	63%	55%	71%	43%	60%	67%	80%
284	54%	100%	70%	67%	80%	76%	86%	87%	93%	87%	93%
285	50%	89%	64%	57%	69%	72%	78%	83%	83%	87%	84%
286	27%	100%	43%	46%	63%	50%	67%	56%	72%	71%	83%
287	40%	82%	54%	48%	61%	61%	70%	67%	74%	62%	69%
288	27%	81%	41%	46%	59%	50%	62%	55%	65%	57%	64%
289	54%	88%	67%	63%	73%	74%	79%	80%	82%	71%	73%
290	31%	62%	41%	40%	49%	44%	50%	44%	50%	58%	58%
291	43%	78%	55%	53%	63%	69%	73%	65%	70%	71%	73%
292	35%	84%	49%	41%	55%	54%	65%	65%	71%	64%	69%
293	32%	78%	45%	45%	57%	67%	71%	62%	64%	56%	56%
294	35%	89%	50%	41%	56%	53%	66%	55%	68%	58%	70%
295	43%	75%	55%	100%	86%	100%	86%	100%	86%	100%	86%
297	43%	75%	55%	100%	86%	100%	86%	100%	80%	100%	80%
298	20%	75%	32%	35%	48%	36%	48%	45%	55%	67%	67%
299	24%	83%	37%	43%	57%	47%	60%	47%	59%	55%	63%
301	44%	80%	57%	100%	89%	100%	89%	100%	86%	100%	86%
302	37%	77%	50%	42%	54%	60%	67%	64%	69%	69%	72%
303	56%	83%	67%	100%	91%	100%	91%	100%	89%	100%	89%
304	26%	92%	41%	52%	66%	85%	88%	91%	91%	100%	95%
305	29%	78%	42%	55%	63%	71%	71%	100%	83%	100%	83%
307	39%	89%	54%	50%	64%	67%	74%	77%	80%	100%	90%
308	14%	50%	22%	33%	40%	100%	67%	100%	67%	100%	67%
310	28%	73%	40%	70%	70%	100%	82%	100%	67%	100%	67%
311	17%	50%	25%	67%	57%	100%	67%	100%	67%	100%	67%
312	25%	75%	38%	50%	60%	62%	66%	83%	77%	83%	77%
313	49%	89%	63%	71%	79%	71%	77%	77%	80%	89%	84%
314	23%	71%	35%	33%	44%	44%	53%	43%	50%	75%	67%

Table A.3 continued from previous page

N	Geonames			Brazil		Southeast		MG		SM	
	PR	RC	F1	PR	F1	PR	F1	PR	F1	PR	F1
315	40%	100%	57%	55%	71%	67%	80%	80%	89%	100%	100%
316	13%	50%	21%	29%	37%	80%	62%	75%	55%	75%	55%
317	24%	82%	37%	28%	41%	38%	50%	36%	48%	40%	50%
318	30%	75%	43%	50%	60%	67%	67%	0%	0%	0%	0%
319	26%	71%	38%	38%	50%	80%	73%	100%	80%	100%	80%
320	13%	100%	23%	25%	40%	29%	45%	67%	80%	67%	80%

In Table A.4 the column *TM* means *Triângulo Mineiro* G1 sub-section.

Table A.4: Triângulo Mineiro sub-state news evaluation

N	Geonames			Brazil		Southeast		MG		TM	
	PR	RC	F1	PR	F1	PR	F1	PR	F1	PR	F1
241	52%	100%	68%	48%	65%	54%	70%	67%	80%	83%	91%
242	40%	55%	46%	67%	60%	50%	43%	50%	43%	60%	47%
245	33%	93%	49%	59%	72%	46%	60%	67%	75%	86%	86%
246	3%	100%	6%	6%	11%	11%	20%	17%	29%	0%	0%
247	52%	100%	68%	88%	94%	91%	95%	94%	97%	93%	96%
248	38%	85%	53%	55%	67%	65%	74%	77%	80%	77%	80%
249	47%	94%	63%	67%	78%	83%	88%	83%	87%	100%	94%
250	28%	78%	41%	50%	61%	64%	70%	70%	74%	78%	78%
251	20%	100%	33%	75%	86%	75%	86%	75%	86%	75%	86%
252	54%	88%	67%	73%	79%	77%	81%	81%	81%	81%	81%
253	85%	92%	88%	100%	95%	100%	94%	100%	92%	100%	91%
254	44%	100%	61%	100%	100%	100%	100%	100%	100%	100%	100%
255	27%	92%	42%	48%	63%	58%	71%	85%	88%	92%	92%
256	48%	77%	59%	67%	72%	79%	76%	74%	70%	67%	63%
257	23%	86%	36%	55%	67%	60%	71%	67%	73%	80%	80%
259	48%	91%	63%	75%	82%	89%	89%	100%	94%	100%	94%
261	40%	100%	57%	67%	80%	80%	89%	80%	89%	100%	100%
262	31%	92%	46%	58%	71%	79%	85%	85%	88%	78%	83%
263	48%	83%	61%	55%	66%	50%	60%	73%	73%	69%	69%

Table A.4 continued from previous page

N	Geonames			Brazil		Southeast		MG		TM	
	PR	RC	F1	PR	F1	PR	F1	PR	F1	PR	F1
265	48%	88%	62%	80%	83%	100%	90%	100%	89%	100%	87%
266	32%	85%	46%	50%	63%	61%	71%	67%	73%	67%	73%
267	23%	80%	36%	38%	52%	57%	67%	60%	67%	67%	71%
268	19%	80%	31%	50%	62%	67%	73%	80%	80%	80%	80%
269	41%	88%	56%	64%	74%	70%	78%	70%	78%	88%	88%
270	17%	75%	28%	38%	50%	50%	60%	60%	67%	75%	75%
271	57%	95%	71%	74%	83%	76%	84%	75%	83%	79%	86%
273	33%	100%	50%	100%	100%	100%	100%	100%	100%	100%	100%
274	23%	100%	37%	75%	86%	75%	86%	67%	80%	100%	100%
275	29%	67%	40%	100%	80%	100%	80%	100%	80%	100%	80%
276	25%	100%	40%	75%	86%	100%	100%	100%	100%	100%	100%
277	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
278	26%	85%	40%	37%	52%	46%	60%	69%	75%	75%	78%
279	26%	86%	40%	50%	63%	75%	80%	86%	86%	100%	92%
280	25%	83%	38%	50%	62%	71%	77%	83%	83%	83%	83%
281	14%	100%	25%	27%	43%	30%	46%	38%	55%	50%	67%
282	28%	78%	41%	38%	50%	50%	59%	56%	63%	62%	66%
283	23%	100%	37%	46%	63%	55%	71%	43%	60%	67%	80%
284	54%	100%	70%	67%	80%	76%	86%	87%	93%	87%	93%
285	50%	89%	64%	57%	69%	72%	78%	83%	83%	87%	84%
286	27%	100%	43%	46%	63%	50%	67%	56%	72%	71%	83%
287	40%	82%	54%	48%	61%	61%	70%	67%	74%	62%	69%
288	27%	81%	41%	46%	59%	50%	62%	55%	65%	57%	64%
289	54%	88%	67%	63%	73%	74%	79%	80%	82%	71%	73%
290	31%	62%	41%	40%	49%	44%	50%	44%	50%	58%	58%
291	43%	78%	55%	53%	63%	69%	73%	65%	70%	71%	73%
292	35%	84%	49%	41%	55%	54%	65%	65%	71%	64%	69%
293	32%	78%	45%	45%	57%	67%	71%	62%	64%	56%	56%
294	35%	89%	50%	41%	56%	53%	66%	55%	68%	58%	70%
295	43%	75%	55%	100%	86%	100%	86%	100%	86%	100%	86%
297	43%	75%	55%	100%	86%	100%	86%	100%	80%	100%	80%
298	20%	75%	32%	35%	48%	36%	48%	45%	55%	67%	67%

Table A.4 continued from previous page

N	Geonames			Brazil		Southeast		MG		TM	
	PR	RC	F1	PR	F1	PR	F1	PR	F1	PR	F1
299	24%	83%	37%	43%	57%	47%	60%	47%	59%	55%	63%
301	44%	80%	57%	100%	89%	100%	89%	100%	86%	100%	86%
302	37%	77%	50%	42%	54%	60%	67%	64%	69%	69%	72%
303	56%	83%	67%	100%	91%	100%	91%	100%	89%	100%	89%
304	26%	92%	41%	52%	66%	85%	88%	91%	91%	100%	95%
305	29%	78%	42%	55%	63%	71%	71%	100%	83%	100%	83%
307	39%	89%	54%	50%	64%	67%	74%	77%	80%	100%	90%
308	14%	50%	22%	33%	40%	100%	67%	100%	67%	100%	67%
310	28%	73%	40%	70%	70%	100%	82%	100%	67%	100%	67%
311	17%	50%	25%	67%	57%	100%	67%	100%	67%	100%	67%
312	25%	75%	38%	50%	60%	62%	66%	83%	77%	83%	77%
313	49%	89%	63%	71%	79%	71%	77%	77%	80%	89%	84%
314	23%	71%	35%	33%	44%	44%	53%	43%	50%	75%	67%
315	40%	100%	57%	55%	71%	67%	80%	80%	89%	100%	100%
316	13%	50%	21%	29%	37%	80%	62%	75%	55%	75%	55%
317	24%	82%	37%	28%	41%	38%	50%	36%	48%	40%	50%
318	30%	75%	43%	50%	60%	67%	67%	0%	0%	0%	0%
319	26%	71%	38%	38%	50%	80%	73%	100%	80%	100%	80%
320	13%	100%	23%	25%	40%	29%	45%	67%	80%	67%	80%

In Table A.5 the column *VM* means *Vales de Minas* G1 sub-section.

Table A.5: Vales de Minas sub-state news evaluation

N	Geonames			Brazil		Southeast		MG		VM	
	PR	RC	F1	PR	F1	PR	F1	PR	F1	PR	F1
401	33%	80%	47%	67%	73%	100%	86%	100%	86%	100%	86%
402	8%	100%	15%	22%	36%	29%	45%	29%	45%	67%	80%
403	25%	75%	38%	50%	60%	62%	66%	62%	66%	100%	80%
404	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
405	18%	40%	25%	22%	28%	25%	31%	29%	34%	100%	57%
406	21%	60%	31%	30%	40%	60%	60%	60%	60%	100%	75%

Table A.5 continued from previous page

N	Geonames			Brazil		Southeast		MG		VM	
	PR	RC	F1	PR	F1	PR	F1	PR	F1	PR	F1
472	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
473	18%	71%	29%	27%	38%	75%	67%	67%	57%	100%	67%
474	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
475	12%	100%	21%	25%	40%	50%	67%	50%	67%	0%	0%
477	6%	100%	11%	11%	20%	25%	40%	25%	40%	33%	50%
478	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
479	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
480	13%	75%	22%	15%	25%	43%	55%	50%	60%	100%	80%

In Table A.6 the column *ZM* means *Zona da Mata* G1 sub-section.

Table A.6: Zona da Mata sub-state news evaluation

N	Geonames			Brazil		Southeast		MG		ZM	
	PR	RC	F1	PR	F1	PR	F1	PR	F1	PR	F1
481	16%	43%	23%	38%	40%	75%	55%	100%	60%	100%	60%
482	17%	75%	28%	23%	35%	30%	43%	40%	50%	40%	50%
483	16%	74%	26%	22%	34%	25%	37%	33%	43%	41%	45%
484	21%	58%	31%	35%	43%	67%	60%	62%	55%	71%	59%
485	17%	80%	28%	40%	53%	57%	67%	60%	67%	75%	75%
486	9%	100%	17%	18%	31%	27%	43%	40%	57%	44%	61%
488	24%	88%	38%	44%	59%	64%	74%	70%	78%	75%	80%
489	11%	100%	20%	15%	26%	25%	40%	36%	53%	71%	83%
490	38%	100%	55%	57%	73%	52%	68%	55%	71%	59%	74%
491	11%	80%	19%	17%	28%	24%	37%	29%	43%	57%	67%
492	8%	100%	15%	12%	21%	22%	36%	50%	67%	67%	80%
493	15%	75%	25%	38%	50%	75%	75%	75%	75%	75%	75%
494	32%	80%	46%	42%	55%	50%	61%	58%	67%	71%	71%
495	37%	89%	52%	41%	55%	44%	56%	36%	47%	43%	50%
497	24%	100%	39%	37%	54%	44%	61%	62%	77%	88%	94%
501	21%	88%	34%	30%	45%	58%	70%	75%	80%	75%	80%
502	38%	100%	55%	60%	75%	75%	86%	100%	100%	100%	100%

Table A.6 continued from previous page

N	Geonames			Brazil		Southeast		MG		ZM	
	PR	RC	F1	PR	F1	PR	F1	PR	F1	PR	F1
503	21%	100%	35%	50%	67%	83%	91%	83%	91%	80%	89%
504	8%	100%	15%	15%	26%	25%	40%	33%	50%	67%	80%
505	33%	67%	44%	67%	67%	67%	67%	100%	80%	100%	80%
506	31%	83%	45%	42%	56%	56%	67%	56%	67%	56%	67%
507	32%	100%	48%	50%	67%	67%	80%	86%	92%	86%	92%
508	20%	100%	33%	33%	50%	75%	86%	75%	86%	100%	100%
509	50%	100%	67%	50%	67%	100%	100%	100%	100%	100%	100%
511	25%	50%	33%	33%	40%	67%	57%	100%	67%	100%	67%
512	19%	100%	32%	33%	50%	55%	71%	60%	75%	67%	80%
515	28%	100%	44%	56%	72%	83%	91%	83%	91%	100%	100%
516	11%	62%	19%	19%	29%	33%	42%	33%	42%	33%	42%
517	29%	83%	43%	42%	56%	50%	60%	50%	60%	75%	75%
520	17%	100%	29%	31%	47%	67%	80%	67%	80%	80%	89%
521	26%	100%	41%	56%	72%	71%	83%	60%	75%	50%	67%
522	29%	100%	45%	33%	50%	40%	57%	67%	80%	100%	100%
523	38%	75%	50%	60%	67%	100%	86%	100%	86%	100%	86%
524	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
525	38%	86%	53%	50%	62%	71%	77%	83%	83%	100%	89%
526	6%	50%	11%	12%	19%	100%	67%	100%	67%	100%	67%
527	29%	100%	45%	67%	80%	100%	100%	100%	100%	100%	100%
528	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
529	22%	100%	36%	40%	57%	67%	80%	67%	80%	67%	80%
530	22%	67%	33%	40%	50%	67%	67%	67%	67%	67%	67%
531	21%	75%	33%	43%	55%	60%	67%	75%	75%	100%	86%
532	12%	67%	20%	15%	25%	20%	31%	67%	67%	67%	67%
533	19%	43%	26%	43%	43%	67%	44%	67%	44%	100%	50%
534	25%	67%	36%	50%	57%	67%	67%	67%	67%	67%	67%
535	38%	60%	47%	60%	60%	100%	75%	100%	75%	100%	75%
536	18%	60%	28%	25%	35%	38%	47%	38%	47%	33%	40%
537	67%	100%	80%	80%	89%	100%	100%	100%	100%	100%	100%
538	80%	100%	89%	100%	100%	100%	100%	100%	100%	100%	100%
539	7%	50%	12%	15%	23%	40%	44%	50%	50%	100%	67%

Table A.6 continued from previous page

N	Geonames			Brazil		Southeast		MG		ZM	
	PR	RC	F1	PR	F1	PR	F1	PR	F1	PR	F1
540	22%	100%	36%	25%	40%	40%	57%	100%	100%	100%	100%
541	33%	100%	50%	67%	80%	67%	80%	100%	100%	100%	100%
542	40%	100%	57%	50%	67%	50%	67%	100%	100%	100%	100%
543	25%	100%	40%	29%	45%	40%	57%	100%	100%	100%	100%
544	33%	100%	50%	50%	67%	100%	100%	100%	100%	100%	100%
545	57%	100%	73%	67%	80%	100%	100%	100%	100%	100%	100%
546	75%	100%	86%	75%	86%	75%	86%	75%	86%	67%	80%
547	40%	100%	57%	40%	57%	67%	80%	100%	100%	100%	100%
548	33%	100%	50%	50%	67%	67%	80%	100%	100%	100%	100%
549	25%	67%	36%	40%	50%	67%	67%	67%	67%	67%	67%
550	25%	67%	36%	50%	57%	67%	67%	67%	67%	67%	67%
551	20%	100%	33%	67%	80%	67%	80%	100%	100%	100%	100%
552	25%	67%	36%	40%	50%	67%	67%	67%	67%	67%	67%
553	100%	90%	95%	100%	91%	100%	86%	100%	86%	100%	86%
554	26%	62%	37%	33%	43%	50%	55%	71%	66%	50%	44%
555	40%	100%	57%	80%	89%	100%	100%	100%	100%	100%	100%
556	25%	67%	36%	100%	80%	100%	80%	100%	80%	100%	80%
557	20%	67%	31%	40%	50%	67%	67%	67%	67%	67%	67%
558	62%	100%	77%	71%	83%	100%	100%	100%	100%	100%	100%
559	30%	100%	46%	33%	50%	43%	60%	100%	100%	100%	100%
560	22%	67%	33%	40%	50%	67%	67%	67%	67%	67%	67%