UNIVERSIDADE CATÓLICA DE PELOTAS
CENTRO POLITÉCNICO
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

Uma Abordagem para o Enriquecimento de Gazetteers a partir de Notícias visando o Georreferenciamento de Textos na Web

por

Cleber Gouvêa

Dissertação apresentada como
requisito parcial para a obtenção do grau de
Mestre em Ciência da Computação

Orientador: Prof. Dr. Stanley Loh

Co-Orientador: Prof. Dr. Luís Fernando F. Garcia

DM-2009/1-001

Pelotas, Abril de 2009.

À minha mãe Diva
e ao meu avô João.

AGRADECIMENTOS

Ao meu pai por seu apoio incondicional em todos os momentos e por me inspirar a cada dia com seu exemplo de dedicação e disciplina ao trabalho.

À minha mãe que por ser apoio incondicional a toda a família auxiliou meu pai em todos os momentos e tornou possível tudo o que somos.

À minha irmã por ser a guardiã fiel do meu passado e companheira leal no presente.

Aos meus avós por me incentivarem e motivarem com seu jeito simples e honesto de viver.

Aos meus orientadores Stanley e Luís Fernando pela confiança, incentivo, profissionalismo e critério na orientação dessa dissertação e durante todo o mestrado, resultando assim na qualidade do presente trabalho.

Aos professores Clodoveu Davis Jr., Marilton de Aguiar e Miguel Fornari por auxiliarem no aperfeiçoamento dessa dissertação com seus conselhos valiosos.

À CAPES, que na condição de órgão de fomento à pesquisa subsidiou o desenvolvimento desta dissertação.

À todos os amigos, parentes e desconhecidos, agora sempre próximos, pelo apoio inestimável demonstrado nos momentos difíceis enfrentados com a perda da minha mãe e do meu avô. OBRIGADO.

Não sei pois nasci para isso e aquilo
E o enguiço de tanto querer. Carpinteiro
do universo inteiro eu sou, assim
No final, carpinteiro de mim
-- RAUL SEIXAS

Gate Gate Paragate Parasamgate Bodhi Svaha

SUMÁRIO

LISTA DE ABREVIATURAS E SIGLAS

ACR	Ambigüidade da Classe do Referente
API	Application Programming Interface
ARC	Ambigüidade da Referência
ART	Ambigüidade do Referente
CLEF	Cross-Language Evaluation Forum
EM	Entidade Mencionada
GeoCLEF	Geographic Cross-Language Evaluation Forum
GeoRSS	Geographic Really Simple Syndication
GKB	Geographic Knowledge Base
GML	Geography Markup Language
GPS	Global Positioning System
GREASE	Geographic Reasoning for Search Engines
IL	Indicador de Localidade
IP	Internet Protocol
HAREM	Avaliação de Reconhecimento de Entidades Mencionadas
KL	Kullback-Leibler
KML	Keyhole Markup Language
MBR	Minimum Bounding Rectangle
MI	Mutual Information
NAACL	North American Chapter of the Association for Computational Linguistics
NP	Nome Próprio
ODP	Open Directory Project
REM	Reconhecimento de Entidades Mencionadas
RI	Recuperação de Informações
RIG	Recuperação de Informações Geográficas
RLM	Retângulo de Limite Mínimo
RT	Resolução de Topônimos
SIG	Sistema de Informações Geográficas
SPIRIT	Spatially-Aware Information Retrieval on the Internet
TF-IDF	Term Frequency–Inverse Document Frequency
TGN	Thesaurus of Geographic Names
WPT	colecção da Web PorTuguesa
WSD	Word Sense Disambiguation

LISTA DE FIGURAS

LISTA DE TABELAS

RESUMO

Com o advento da Internet e o crescente número de informações disponíveis torna-se necessária a definição de estratégias especiais que permitam aos usuários o acesso rápido a informações relevantes. Como a Web possui grande volume de informações principalmente com o foco geográfico torna-se necessário recuperar e estruturar essas informações de forma a poder relacioná-las com o contexto e realidade das pessoas através de métodos e sistemas automáticos. Para isso uma das necessidades é possibilitar o georreferenciamento dos textos, ou seja, identificar as entidades geográficas presentes e associá-las com sua correta localização espacial. Nesse sentido, os topônimos (ex: nomes de localidades como cidades, países, etc.), devido à possibilidade de identificar de forma precisa determinada região espacial, apresentam-se como ideais para a identificação do contexto geográfico dos textos. Essa tarefa, denominada de Resolução de Topônimos apresenta, no entanto, desafios importantes principalmente do ponto de vista lingüístico, já que uma localidade pode possuir variados tipos de ambigüidade. Com relação a isso a principal estratégia para superar estes problemas compreende a identificação de evidências que auxiliem na identificação e desambiguação das localidades nos textos. Para essa verificação são utilizados geralmente os serviços de um ou mais dicionários toponímicos (*Gazetteers*). Como são criados de forma manual eles apresentam, no entanto deficiência de informações relacionadas principalmente a entidades que podem identificar, embora de forma indireta, determinados tipos de lugares como ruas, praças, universidades etc., as quais são definidas como Indicadores de Localidade. O presente trabalho propõe uma abordagem para a recuperação dessas entidades aproveitando para isso o caráter geográfico das informações jornalísticas. Para ilustrar a viabilidade do processo diferentes tipos de corpora de notícias foram testados e comparados pela habilidade de criação de *Gazetteers* com os Indicadores recuperados, sendo os *Gazetteers* avaliados então pela capacidade de identificação das cidades relacionadas às notícias testadas. Os resultados demonstram a utilidade da abordagem para o enriquecimento de *Gazetteers* e consequentemente para a recuperação de Indicadores de Localidade com maior simplicidade e extensibilidade que os trabalhos atuais.

Palavras-chave: Recuperação de Informações Geográficas, Resolução de Topônimos, Georreferenciamento de Textos, *Gazetteers*.

TITLE: "ENRICHMENT OF GAZETTEERS FROM NEWS TO IMPROVE TEXT-BASED GEOREFERENCING ON THE WEB"

ABSTRACT

Georeferencing of texts, that is, the identification of the geographical context of texts is becoming popular in the Web due to the high demand for geographical information and due to the raising of services for query and retrieval like Google Earth (geobrowsers). The main challenge is to relate texts to geographical locations. These associations are stored in structures called gazetteers. Although there are gazetteers like Geonames and TGN, they fail in coverage, lacking information about some countries, and they also fail by weak specialization, lacking detailed references to locations (fine granularity) as for example names of streets, squares, monuments, rivers, neighborhoods, etc. This kind of information that acts as indirect references to geographical locations is defined as "Location Indicators".

This dissertation presents an approach that identifies Location Indicators related to geographical locations, by analyzing texts of news published in the Web. The goal is to enrich create gazetteers with the identified relations and then perform geo-referencing of news. Location Indicators include non-geographical entities that are dynamic and may change along the time. The use of news published in the Web is a useful way to discover Location Indicators, covering a great number of locations and maintaining detailed information about each location. Different training news corpora are compared for the creation of gazetteers and evaluated by their ability to correctly identify cities in texts of news.

Keywords: Geographical Information Retrieval, Toponym Resolution, Georeferencing of Texts, Gazetteers.

1 INTRODUÇÃO

Com o surgimento da Web e com a enorme quantidade de informações não estruturadas ou semi-estruturadas (Abiteboul et al., 2000 *apud* Borges et al., 2003) disponíveis atualmente, surge a necessidade do uso de tecnologias semânticas para a obtenção de informações relevantes visando evitar assim problemas como a sobrecarga de informações (*information overload)*. (Perry et al., 2007)

Na internet, a recuperação de informações proporcionada pelos sistemas de busca tem ajudado a indexar e representar os fluxos de informação, facilitando e agilizando sua recuperação. Apesar de iniciativas recentes, esses mecanismos, no entanto, apresentam deficiência na recuperação de conteúdos semânticos (Baeza-Yates et al., 2008), como por exemplo, informações geográficas relacionadas ao contexto do usuário (Jones e Purves, 2008).

Atualmente, através da Web Semântica (Berners-Lee et al., 2001) e mais recentemente da Web Semântica Geoespacial (Egenhofer, 2002), iniciativas têm surgido visando auxiliar à representação e estruturação de conhecimento na Web (Berners-Lee et al., 2001). Do ponto de vista geográfico, alguns dos principais exemplos compreendem a definição de padrões específicos para a estruturação de informações geográficas, como o GeoRSS, o GML e o KML. Atualmente estes padrões têm ganhado grande popularidade, fato esse motivado principalmente pela popularização de serviços de navegação geográfica na Web (*geobrowsers*) os quais utilizam as informações contidas nesses arquivos para a integração descentralizada e contextualização dessas informações por meio de mapas (ex: Google Maps[1]).

Para povoar esses arquivos torna-se necessária, no entanto, a adoção de técnicas especiais para a identificação do contexto geográfico das informações. Isto ocorre já que a anotação das informações nestes arquivos é realizada de forma manual, o que despende tempo e desestimula a popularização desses formatos, demandando então a utilização de métodos automáticos para a recuperação geográfica das informações.

Com isso e devido à falta da identificação do contexto geográfico das informações na Web como um todo, a Recuperação de Informações Geográficas (RIG), área surgida, segundo (Larson, 1996) a partir da demanda por uma pesquisa integrada

[1] http://maps.google.com/

entre os Sistemas de Informação Geográfica e a Recuperação de Informações tradicional, tem sido alvo de intensa pesquisa. O foco central é lidar com todos os problemas da recuperação de informações que contenham algum tipo de consciência espacial (*spatial awareness),* ou seja, que incluam referências geográficas (georreferências) (Lana-Serrano et al., 2007), visando auxiliar dessa forma na identificação e contextualização das informações de acordo especificamente com seu contexto geográfico.

O processo de identificação do contexto geográfico de textos é denominado de *geotagging* (Amitay et al., 2004) e envolve duas etapas principais, o *geoparsing* e o *geocoding* (ou geocodificação). (McCurley, 2001)

A fase de *geoparsing* compreende a identificação das entidades geográficas presentes no texto por meio da análise do seu conteúdo ou de informações relativas ao servidor em que ele se encontra armazenado. As técnicas mais utilizadas empregam a análise do texto, para isso baseiam-se em heurísticas ou em técnicas de Inteligência Artificial como o Reconhecimento de Entidades Mencionadas (REM) e/ou o Aprendizado de Máquina. Já a fase de geocodificação tem o objetivo de reconhecer a localização espacial correta das localidades identificadas, associando a elas identificadores específicos como, por exemplo, coordenadas geográficas (ex: latitude e longitude).

Como os topônimos (ex: nomes de localidades como cidades, países, etc.) podem ser usados para identificar corretamente determinada região espacial, possuindo assim propriedades geográficas distintas (ex: geometria, topologia) (Hu e Ge, 2007), sua identificação torna-se ideal para a verificação do contexto geográfico de textos, apresentando, no entanto desafios específicos principalmente do ponto de vista lingüístico para o seu reconhecimento e desambiguação nos textos.

Com isso, não basta apenas identificar os topônimos pelo seu nome nos textos, visto que cada localidade pode apresentar variados tipos de ambigüidade, as quais de acordo com (Clough et al., 2004) podem ser: com outra localidade homônima (ambigüidade do referente), com outro tipo de entidade não-geográfica, por exemplo, nomes de pessoas ou organizações (ambigüidade da classe do referente) ou com nomes sinônimos, por exemplo, siglas da cidade (ambigüidade da referência). A área que

estuda a identificação e desambiguação de topônimos é definida por (Leidner, 2007) como Resolução de Topônimos.

Para viabilizar a Resolução de Topônimos e a superação destas ambiguidades uma das alternativas é utilizar técnicas relacionadas à Desambiguação Lexical de Sentido (*Word Sense Disambiguation*) (Li et al., 2003) onde termos co-ocorrentes (*collocations*) relacionados às localidades são identificados em uma coleção de textos de treinamento, utilizando pra isso algoritmos de Aprendizado de Máquina (*Machine Learning*). Para armazenar essas entidades e auxiliar assim nos processos de georreferenciamento de textos são utilizados dicionários toponímicos (*Gazetteers*), os quais têm o objetivo, portanto de armazenar informações diversas relacionadas às localidades (como o nome, sinônimos, o tipo e outras entidades) incluíndo também coordenadas geográficas, as quais visam identificar a correta extensão espacial das localidades representadas (Hill, 2000).

Estas técnicas buscam assim obter informações detalhadas sobre as localidades e superar os problemas dos *Gazetteers* atuais, como o Geonames[2] e o TGN[3], os quais embora consigam cobrir localidades de todo mundo apresentam carência de informações detalhadas sobre estas localidades, como por exemplo, nomes de ruas, praças, monumentos, rios, bairros, e outras entidades relacionadas a elas (Leveling e Hartrumpf, 2006a) demandando com isso atualizações constantes de seus índices para garantir a qualidade dos processos de georreferenciamento, já que conforme verificou (Leidner, 2004) os *Gazetteers* globais sofrem cerca de 20.000 modificações por mês.

Estes tipos de entidades que, embora de forma indireta, auxiliam na identificação das localidades referenciadas nos textos, são definidas por (Leveling et al., 2007) como Indicadores de Localidade (*Location Indicators),* podendo relacionar-se a entidades geográficas relacionadas às localidades (como nomes de ruas, rodovias, aeroportos, etc.), assim como a entidades não-geográficas, como pessoas importantes (ex: prefeito, vereadores, etc.), eventos históricos ou mesmo situações temporárias (ex: nomes de pessoas envolvidas em algum tipo de acontecimento), ilustrando assim a necessidade da adoção de métodos automáticos para a verificação correta desses Indicadores levando em conta sua característica dinâmica.

---

[2] http://www.geonames.org/

[3] http://www.getty.edu/research/conducting_research/vocabularies/tgn

O presente trabalho busca auxiliar nesse processo, conforme ilustra a próxima seção.

1.1 Objetivo e Contribuições

Embora já existam trabalhos que buscam recuperar Indicadores de Localidade, estes apresentam problemas, os quais se relacionam principalmente à necessidade de anotação manual de corpora de treino para a inferência das entidades e a utilização de técnicas úteis apenas para idiomas específicos. O presente trabalho busca superar estes problemas através da sugestão de uma abordagem para a identificação de Indicadores de Localidade a partir de notícias, aproveitando para isso o caráter geográfico das informações jornalísticas. A idéia é possibilitar a partir destes Indicadores a criação e enriquecimento dinânico de *Gazetteers,* mantendo informações detalhadas sobre as localidades (particularmente cidades) e auxiliando assim na sua identificação e desambiguação nos textos.

Para garantir isso a abordagem proposta por este trabalho realiza a extração de Indicadores de Localidade a partir das notícias baseado na co-ocorrência entre nomes próprios nos textos, sem a necessidade de anotação manual de corpora de treino e pra qualquer tipo de localidade e linguagem (desde que seja possível identificar nomes próprios no respectivo idioma e desde que haja notícias relacionadas às localidades). Busca-se também qualificar estes Indicadores de acordo com sua relevância (através de fórmulas específicas que determinam o peso das relações) às localidades levando em conta que um mesmo indicador pode estar relacionado a mais de uma localidade.

Para testar a qualidade da abordagem sugerida foram criados *Gazetteers* com os Indicadores recuperados, sendo utilizado pra isto diferentes corpora de notícias (ex: com quantidade e período temporal distintos) visando avaliar qual o tipo de corpora mais adequado para a extração e qualificação dos Indicadores de Localidade. Os *Gazetteers* foram então avaliados pela capacidade de corretamente identificar às cidades relacionadas às notícias utilizadas para teste. Os resultados foram então comparados com um *Gazetteer* baseline (contendo apenas ruas e bairros relacionados às cidades avaliadas), o qual é utilizado também para a identificação da variedade dos Indicadores recuperados.

A partir disso, o presente trabalho busca, portanto responder as seguintes

questões:

a) A abordagem proposta para identificação de Indicadores de Localidade pode gerar *Gazetteers* com qualidade? Esta questão será verificada a partir da comparação dos resultados do georreferenciamento de notícias utilizando dois *Gazetteers*: o primeiro enriquecido utilizando a abordagem proposta pelo trabalho e o segundo contendo apenas ruas e bairros relacionados às cidades avaliadas (as quais foram capturadas a partir de bases de dados governamentais públicas do Brasil).

b) O período temporal das notícias influencia na qualidade dos *Gazetteers* gerados? Para verificar essa questão será realizada a comparação dos resultados referentes ao georreferenciamento de notícias, utilizando pra isso *Gazetteers* enriquecidos com notícias de períodos temporais distintos.

c) O volume de notícias analisado influencia na qualidade dos *Gazetteers* gerados? Para essa verificação será realizada a comparação dos resultados do georreferenciamento de notícias, utilizando *Gazetteers* enriquecidos com diferentes quantidades de notícias para o corpus de treino.

d) A utilização de notícias com apenas uma cidade no texto melhora a qualidade dos *Gazetteers* gerados? Para analisar essa questão será realizado o georreferenciamento de notícias, utilizando *Gazetteers* enriquecidos a partir de notícias com apenas uma cidade no texto, sendo comparado o resultado com um *Gazetteer* gerado utilizando notícias diversas.

e) A fórmula de peso definida para classificar os Indicadores de Localidade de acordo com sua relevância às Localidades apresenta resultados superiores a não utilização de pesos específicos paras as relações? Para verificar essa questão será realizado o georreferenciamento de notícias, utilizando os *Gazetteers* criados pela abordagem com os Indicadores e os pesos definidos para classificá-los de acordo com sua relevância às cidades, sendo comparados os resultados com um *Gazetteer* onde estas mesmas relações não apresentam um peso específico.

O trabalho é particularmente útil, portanto para a desambiguação de topônimos em textos, tendo como foco a utilização da abordagem proposta na superação das ambiguidades do referente/referência, assim como para o georreferenciamento de textos

que não possuem nenhum nome de localidade em seu interior, visando superar com isso a *ambiguidade indireta da referência* conforme definida por este trabalho.

1.2 Estrutura do Trabalho

O trabalho está dividido da seguinte forma: primeiramente a seção 2 apresenta o referencial teórico do trabalho, descrevendo a área relacionada à Recuperação de Informações Geográficas e mais especificamente a área associada à Resolução de Topônimos, apresentando os principais desafios envolvidos (ex: os tipos de ambiguidades) com ênfase na importância e nos problemas principais dos *Gazetteers* atuais e das estratégias envolvendo a recuperação de Indicadores de Localidade. A seção 3 apresenta a abordagem proposta pelo trabalho para auxílio na superação destes problemas, bem como seus diferenciais para os trabalhos atualmente disponíveis. Já a seção 4 apresenta os experimentos realizados para avaliação do trabalho, os resultados encontrados e uma discussão relacionada a eles. Por fim a seção 5 apresenta um resumo dos principais resultados e trabalhos futuros pretendidos.

2 REFERENCIAL TEÓRICO

Para melhor contextualizar o problema alvo do trabalho torna-se necessário compreender o processo para a identificação do contexto geográfico dos textos.

Com isso o presente capítulo busca analisar a área de Recuperação de Informações Geográficas (seção 2.1) com foco na área relacionada à Resolução de Topônimos (seção 2.2), buscando compreender com isso as características e desafios associados ao georreferenciamento de textos particularmente relacionados à identificação das localidades referenciadas por eles com o apoio de estruturas denominadas de dicionários toponímicos (ou *Gazetteers*), os quais serão melhor ilustrados (com sua importância e problemas) na (seção 2.3).

2.1 Recuperação de Informações Geográficas

A recuperação de informações na *Web* se dá por meio dos mecanismos de busca, que consultam os sites e através da análise do seu conteúdo desenvolvem métodos próprios de classificação de propósito geral como o PageRank (Page *et al.*, 1999) ou focados em conteúdos específicos, como por exemplo, o CiteSeer (Giles *et al.*, 1998).

No entanto, os sistemas de busca tradicionais não apresentam suporte para informações contextuais, analisando o conteúdo da página ou as ligações entre seus *hiperlinks*, mas não possibilitando a verificação de informações semânticas, como por exemplo, a localidade referenciada pelos textos, impedindo assim a análise com precisão de informações dentro ou próximo a determinadas regiões geográficas (Buyukkokten et. al., 1999) e seu correspondente ranqueamento de acordo com a relevância para o usuário (Jones et al., 2001).

Uma larga proporção da informação presente na Web pode ser incluída dentro do espaço geográfico e, como conseqüência, muitos usuários desejariam especificar nomes geográficos de lugares como parte das consultas (*queries*) (Jones et al., 2001). De acordo com (Sanderson et al., 2004 *apud* Borges, 2006) consultas que incluem pelo menos um termo relacionado à geografia, como nomes de lugar e feições naturais (ex., “praia”, “serra”), são hoje um subconjunto significativo das pesquisas submetidas aos mecanismos de busca. Com isso Web sites que contém, por exemplo, informações sobre restaurantes, teatros e cinemas são mais interessantes para usuários vizinhos dessas localidades (Buyukkokten et. al., 1999). Variadas informações como as jornalísticas

(tempo, condições de tráfego) também são mais úteis se diretamente relacionadas com a região em que ocorrem.

O caráter semi-estruturado da Web dificulta, no entanto, o acesso a informações geográficas. (Jones et al., 2006) relaciona as principais dificuldades no uso da Web como fonte de informações geográficas:

- O contexto geográfico é incluído junto das descrições via linguagem natural.
- Nomes de lugares são ambíguos e confundidos com nomes de organizações, pessoas, construções e ruas.
- Dependência de presença e relação com os termos do texto.
- Interpretação das relações espaciais (“próximo”, “ao oeste”, etc.).
- Construção de *ranking* específico para definição da relevância geográfica.

Visando solucionar esses problemas e trazer à Web todas as vantagens relacionadas à descrição semântica e geográfica das informações, técnicas têm sido desenvolvidas tanto em ambiente acadêmico como comercial visando acessar recursos com base em seu contexto geográfico (Gey, 2005 e Jones, 2003).

Segundo (Larson, 1996) a área de Recuperação de Informação Geográfica (RIG) pode ser vista como um ramo da área de Recuperação de Informação tradicional, incluindo todas suas áreas de pesquisa, mas enfatizando a recuperação e indexação de informações geográficas e espaciais. O objetivo é lidar com todos os problemas da recuperação de informações que contenham algum tipo de consciência espacial (*spatial awareness),* ou seja, que incluam referências geográficas (georreferências) as quais são essenciais para o significado da consulta, por exemplo: “encontre-me hotéis próximos a Madrid” (Lana-Serrano et al., 2007).

De acordo com (Santos e Chaves, 2006) a RIG pressupõe o seguinte:

- a possibilidade de associar à coleção de documentos informações geográficas.
- a existência ou a possibilidade de criação de repositórios semânticos (*Gazetteers*) que permitam a inferência geográfica (*geographical reasoning*).

Com isso segundo (Larson, 1996) a RIG está relacionada com a recuperação de informações determinística (por exemplo, para encontrar todos os documentos relacionados à determinada coordenada geográfica) e com a recuperação de informações

probabilística (por exemplo, para encontrar todas as cidades próximas a um determinado rio).

O processo de georreferenciamento de textos possui de forma geral os componentes apresentados na figura 1.

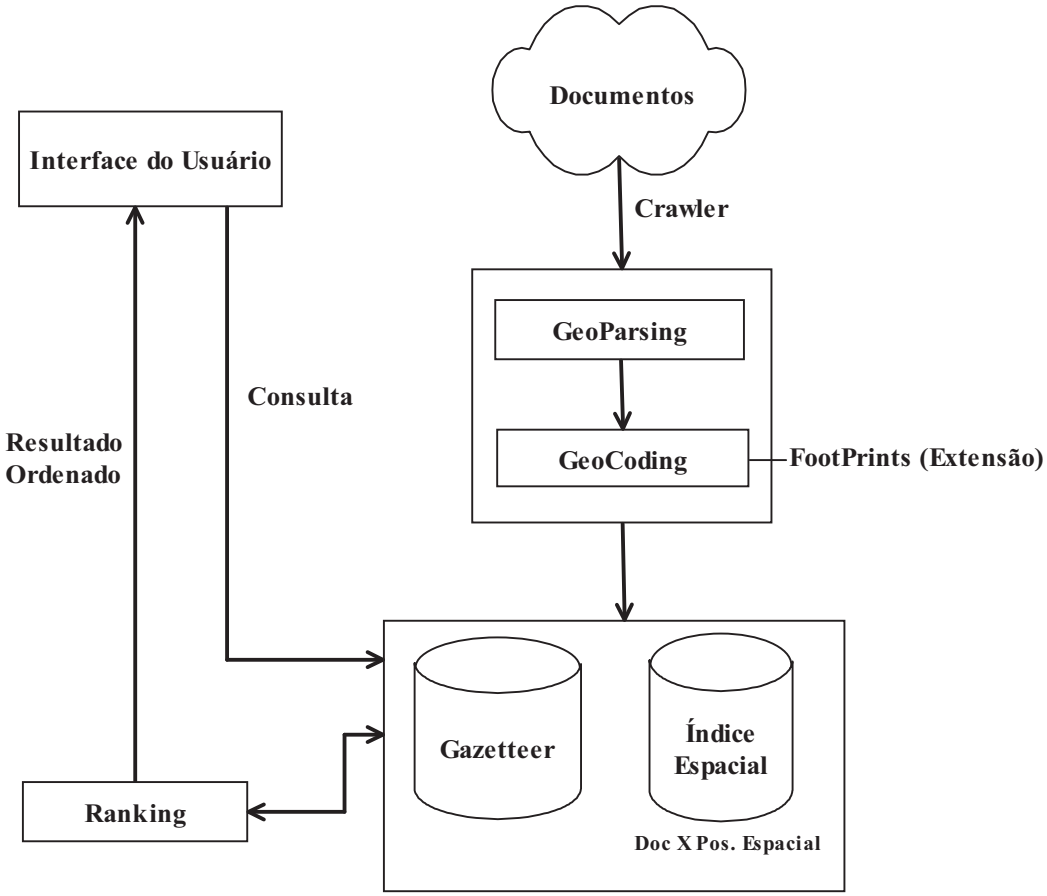

Figura 1. Principais Componentes da RIG

Diferente dos modelos de recuperação de informação tradicionais como o modelo Vetorial (Salton, 1989) que representa os documentos como índices de termos (por exemplo, com a posição e frequência das palavras nos documentos) utilizando o conteúdo parcial ou integral do texto, na RIG os termos extraídos como índices devem ser relacionados a descrições espaciais, ou seja, a entidades geometricamente definidas e localizadas no espaço (De Floriani, 1993 apud Larson, 1996).

Os processos de reconhecimento do contexto geográfico de textos e a definição de coordenadas espaciais (latitude, longitude, etc.) são definidos respectivamente como

geoparsing e *geocoding* (ou geocodificação) (McCurley, 2001). Os índices dos arquivos podem também ser estruturados conforme uma lista invertida (Wang et al., 2006 e Andrade e Silva, 2006), com a diferença que a lista agora apresenta as relações espaciais (associando documentos com seu contexto geográfico) em vez de relacionar apenas as palavras com os documentos em que elas ocorrem. O processo de *ranking* por sua vez tem o objetivo de ordenar os resultados de acordo com o grau de associação com a localidade espacial exposta na consulta.

O componente mais importante é o *Gazetteer*, o qual busca estruturar relacionamentos semânticos relacionados às localidades, sendo organizado geralmente de forma hierárquica com o tipo, a localização (através de coordenadas geográficas) e outras descrições (Hill, 2000) relacionadas a elas, sendo que dependendo da complexidade de sua estrutura, pode ser visto como uma geo-ontologia ou ontologia geográfica. A partir da utilização de *Gazetteers* torna-se possível, por exemplo, reconhecer palavras e frases que se relacionam a alguma localidade, auxiliando também na desambiguação desses termos (superando ambigüidades e reconhecendo a localidade exata referenciada).

Um dos problemas é que para essa correta identificação/desambiguação é necessário a utilização de *Gazetteers* dinâmicos e que abranjam com especificidade as localidades. A seção 2.3 apresenta em detalhes esses desafios.

A identificação do contexto geográfico de páginas *Web* tem ajudado na evolução da Web Semântica, onde, através do acesso a informações estruturadas pelas máquinas, torna-se possível o desenvolvimento de aplicações automáticas e mais inteligentes (Berners-Lee et al., 2001), levando-se em conta também os bancos de dados existentes e os dados contidos neles (Berners-Lee, 2006). Com isso, e devido à popularização dos serviços de navegação geoespacial na Web, juntamente com o surgimento de formatos semânticos para distribuição de informações geográficas, a Web Semântica Geo-espacial tem surgido (Egenhofer, 2002), permitindo assim que aplicações distintas possam compartilhar e integrar este tipo de informação com nível maior de interoperabilidade.

As principais etapas do georreferenciamento de textos, juntamente com uma seção especial explicitando as principais estratégias e métodos utilizados para sua avaliação são apresentados nas seções seguintes.

2.1.1 GeoParsing

Para entender como é possível extrair o contexto geográfico de textos a partir da análise de seu conteúdo é necessário primeiramente analisar como as relações espaciais são determinadas para especificar lugares ou espaços geográficos na Terra. As relações espaciais são comumente agrupadas segundo (Egenhofer e Franzosa, 1991) em três categorias:

- Topológicas: descrevem os conceitos de vizinhança, incidência, sobreposição sem variar com escalas e rotações (ex. dentro de).
- Métricas: são consideradas em termos de direções (ex: orientações no espaço = ao leste, ao oeste...) e distâncias (ex: dependem de definições métricas = perto de).
- De Ordem: expressam a ordem, total ou parcial, dos objetos espaciais (ex: em frente a, acima de).

Raciocínio espacial (*spatial reasoning*) é a expressão usada para denotar inferências sobre relacionamentos espaciais entre objetos no espaço, usando um subconjunto conhecido de relações espaciais. Este tipo de raciocínio permite fazer predições e diagnósticos. O raciocínio espacial pode ser classificado como quantitativo ou qualitativo, dependendo do tipo de informação usada no processo de raciocínio (Rodríguez, 2002 *apud* Borges, 2006).

O "raciocínio espacial" quantitativo diferencia diversas relações espaciais, por exemplo, relações topológicas e métricas, e é tipicamente formalizado usando um sistema de coordenadas geográficas e álgebra vetorial. Este tipo de processamento da informação é claramente distinto da forma como as pessoas interpretam relações espaciais. Assim, processos de raciocínio qualitativo tornam-se necessários em, por exemplo, sistemas especialistas espaciais e SIGs. O "raciocínio espacial" qualitativo é amplamente utilizado por seres humanos para entenderem e analisarem um ambiente espacial quando a informação disponível está na forma qualitativa, como ocorre em documentos textuais (Borges, 2006).

As pessoas pensam e se comunicam a respeito do mundo em termos de conceitos vagos, que são imprecisos ou probabilísticos, como, por exemplo, “centro da cidade”, “perto de”, “nos arredores de” (Montello, 2003 *apud* Borges, 2006). Elas raramente dizem “o restaurante está a 35,93 metros a oeste”, por outro lado fornecem algumas

instruções qualitativas como “o restaurante está à direita, a duas quadras da rodovia” (Borges, 2006).

Dentro deste contexto, (Egenhofer e Mark, 1995) apresenta também a geografia do cotidiano (*naïve geography*), a qual é uma disciplina com o objetivo de “capturar e refletir o jeito que as pessoas pensam e raciocinam a respeito do espaço e tempo geográfico, tanto consciente como subconscientemente”.

Nos *Gazetteers* (seção 2.3) estes termos devem ser associados ou relacionados à *features* (ou entidades) as quais eles representam. A ISO 19109[4] define *feature* como “um objeto com significado em um domínio selecionado do discurso”. No contexto geográfico, países, cidades e ruas são exemplos desses objetos. Exemplos de features são “Rua Bento Gonçalves” e “Arroio São Lourenço”, os quais seriam representados respectivamente pelas *features* de nome “Bento Gonçalves” e “São Lourenço” e pelas *features* de tipo “Rua” e “Arroio”.

No entanto, embora seguindo determinados padrões relacionados às referências espaciais, outra característica é que as entidades geográficas podem estar também mencionadas no texto de forma implícita (quando apresenta relação topológica, por exemplo, a partir da inclusão do estado relacionado à cidade) ou explícita (quando aparece de forma direta no texto). (Wang et al., 2006)

Já (Borges et al., 2003) define as informações espaciais que não possuem associação direta com coordenadas geográficas (CEPs, telefones, etc.) como referências espaciais indiretas. Recentemente (Leveling et. al., 2007) definiu e conceituou de forma mais ampla estes tipos de entidades que auxiliam a identificar determinada localidade (ex: nomes de ruas, bairros, praças, aeroportos, etc.) como **Indicadores de Localidade** (*Location Indicators*), demonstrando através de experimentos a sua importância para a identificação de cidades e consequentemente do contexto geográfico dos textos, conforme ilustra a seção 2.2.2.

Para a recuperação automática de entidades geográficas é necessário, portanto levar em conta todas essas características, atentando também para o problema da desambiguação das localidades que possuam características semelhantes (ex: com o mesmo nome) visto que a vasta maioria dos nomes encontrados na Web apresenta

[4] http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=39891

algum tipo de ambigüidade (Smith et. al., 2001 *apud* Amitay et al., 2004), conforme ilustra a seção 2.2.1.

O reconhecimento do contexto geográfico em páginas da Web é realizado através de duas abordagens principais (Amitay et al., 2004). A primeira abordagem, chamada de Geografia da Fonte (*Source Geography*) é relacionada com a origem das páginas e utiliza os elementos de infra-estrutura da Internet para obtenção de informações sobre a localização física dos servidores onde esses documentos estão hospedados. A segunda abordagem, denominada Geografia do Alvo (*Target Geography*), é baseada no conteúdo das páginas, utilizando elementos nelas contidos para deduzir uma ou mais localizações encontradas em seus textos. Esses elementos consistem em nomes de lugar, coordenadas geográficas, códigos postais e endereços que são usados para classificar e indexar as páginas (Borges, 2006).

Os trabalhos publicados geralmente focam na geografia do alvo ou apresentam estratégias híbridas evolvendo ambos os tipos de geografia.

Alguns exemplos de uso da geografia de fonte pode ser encontrada nos seguintes trabalhos (Wang et al., 2005) (Pyalling et al., 2006), (Zhang, 2006) os quais recuperam a localização geográfica de documentos por dados relacionados ao IP do servidor (utilizando serviços como o GeoIp[5], os quais relacionam endereços IP à sua localização geográfica) e outras informações contidas em órgãos como a Internic[6], utilizando também regras relacionadas ao nome ou posição do site dentro do contexto geográfico do conjunto de sites de seu domínio.

Um dos problemas dessas técnicas é que nem sempre o conteúdo da página condiz com a localização do servidor que a hospeda, o que traz desvantagens com relação à geografia do alvo.

Como a tarefa de recuperar referências em documentos é diferente de recuperar documentos a partir de consultas (*queries*), visto que as consultas são geralmente curtas e não constituem nomes próprios, alguns trabalhos visam também identificar e relacionar documentos de acordo com as consultas informadas, dentre eles (Martins et al., 2006b e Graupmann e Schenkel, 2006).

---

[5] http://www.maxmind.com/app/ip-location
[6] http://www.internic.net/

2.1.2 GeoCoding

Após a verificação das entidades geográficas contidas nos textos é necessário reconhecer o correto contexto geográfico representado por elas. Para (Davis Jr. et al., 2003) a fase de *geocoding* (ou geocodificação) compreende a localização de pontos na superfície da terra a partir de informações alfanuméricas, envolvendo três etapas: o tratamento do endereço alfanumérico semi-estruturado (*parsing*), o estabelecimento de uma correspondência entre o endereço estruturado e o banco de dados (*matching*) e a atribuição das coordenadas geográficas com a extensão geográfica (*footprint*) da entidade que está sendo alvo da geocodificação (*locating* ou *grounding*).

O *footprint* é mais especificamente uma representação geométrica da extensão do conteúdo geográfico do objeto sendo descrito, sendo expresso em coordenadas geográficas (latitude, longitude). A localização referida por um *footprint* conforme ilustra a figura 2 é geralmente definida segundo (Larson e Frontiera, 2004) por:

- Pontos: onde se mantém um senso geral da localização sem extensões ou formas.
- Polígonos: onde ocorre a identificação da localização, extensão e forma com grau variável de precisão. O retângulo envolvente mínimo é a representação espacial poligonal mais usada em sistemas RIG. Contudo, segundo (Papadias et al., 1995) o retângulo envolvente mínimo apresenta algumas limitações, as quais relacionam-se principalmente às representações diagonais, irregulares, desconexas ou de regiões multirrepresentadas.

Figura 2. Representação de uma Localização Usando Pontos e Polígonos (Larson e Frontiera, 2004)

Como mencionado para a conversão dos lugares em coordenadas geográficas

torna-se necessário a utilização de *Gazetteers* que relacionem aos nomes das localidades sua respectiva posição espacial. Associados aos *Gazetteers* há a necessidade de adoção de estratégias específicas visando a desambiguação das localidades e a definição do seu grau de associação com o texto alvo do georreferenciamento, já que um mesmo documento pode conter referência a mais que uma localidade. O ranqueamento das informações torna-se, portanto necessário, sendo descrito na seção seguinte.

2.1.3 Ranking

O ranqueamento das informações relevantes[7] em um sistema de RIG está intimamente relacionado às técnicas de indexação empregadas, podendo envolver métodos e características específicas de acordo com o tipo de consulta realizada.

Para viabilizar a formação do *ranking* é necessário, portanto associar esses conceitos a métricas específicas que possam abranger de forma ampla a necessidade de informação relacionada à consulta.

As métricas para definição da relevância para a RIG são ilustradas por (Silva et al., 2005) e compreendem as distâncias espacial e semântica. A distância espacial ou geográfica (relacionada exclusivamente ao contexto espacial) compreende a verificação da sobreposição entre regiões e as relações topológicas (como por exemplo, o número de relações separando lugares em um *Gazetteer*). Já a distância semântica ou temática (relacionada a informações que contextualizam ou expandem o contexto espacial) considera a análise de características especiais como linguagem, a população ou outras relações não geográficas para a definição de relevância. (Cai, 2002 e Silva et al., 2005).

Segundo (Worboys, 1996 *apud* Silva et al., 2005) um bom motivo para usar a similaridade semântica é que o espaço Euclidiano não parece ser capaz de modelar a proximidade geográfica corretamente, visto que o conceito de proximidade espacial de um lugar para outro pode ser relativo para cada pessoa (ex: distância ao longo de ruas e estradas).

Para satisfazer as características relacionadas à RIG alguns autores, no entanto, propõem sistemas híbridos os quais levam em conta tanto a distância espacial como a

[7] Determinar se um documento é relacionado com a necessidade de informação do usuário ou não é um tópico subjetivo e controverso da mesma forma que a noção de relevância, podendo ser definido de muitas maneiras. (Saracevic, 1975 *apud* Bucher et al., 2005)

semântica.

(Cai, 2002) em seu trabalho representa os documentos e consultas de acordo com o escopo geográfico (GS_d , GS_c,) e temático (TS_d , TS_c). O grau de relevância de um documento para uma determinada consulta (*query)* é dado pela fórmula (1).

$$Rel(d, q) = f(\text{SimG}(GS_d, GS_c), \text{SimT}(TS_d, TS_c)) \quad (1)$$

Onde SimG(*) mede a similaridade entre o *escopo geográfico* do documento e da consulta, SimT(*) mede o grau de similaridade entre o *escopo temático* do documento e o da consulta e f(*) é uma função para combinação das medidas de relevância das dimensões temática e geográfica.

Já (Wang et al., 2006) apresenta um sistema baseado em dois índices, o *focus-index*, que utiliza uma lista invertida de termos para representar cada documento, e o *grid-index* que divide a superfície da terra em uma grade visando representar os documentos nas células dessa grande de acordo com seu contexto geográfico. O processo baseia-se em recuperar os dois índices e então combiná-los, visando obter a relevância total a partir da associação entre relevância textual (*focus-index*) e da relevância geográfica (*grid-index*), seguindo para isso a fórmula (2).

$$R_{combined} = R_{text} \times \alpha + R_{geo} \times (1- \alpha) \quad (2)$$

Onde R_{text} é o escore da relevância textual, R_{geo} é o escore da relevância geográfica e α é o peso estabelecido para cada tipo de escore, sendo que a relevância textual foi definida com mais peso, já que experimentos mostraram que ela deveria possuir maiores escores do que a relevância geográfica.

Para avaliar a qualidade do algoritmo de ranqueamento e das estratégias de georreferenciamento como um todo, torna-se necessária, então, a adoção de métricas e avaliações específicas visando legitimar assim a qualidade das abordagens, conforme descreve a próxima seção.

2.1.4 Avaliação

De forma geral, o objetivo da avaliação é medir a performance do trabalho de recuperação de informações, por exemplo, comparando e medindo o desempenho de diferentes sistemas ou de componentes próprios relacionados a um mesmo mecanismo.

Com a popularização dos sistemas de RIG surge a necessidade da adoção de *frameworks* específicos que possam realizar sua análise individual permitindo também a comparação com outros sistemas. Estas análises podem ser centradas no sistema (através de avaliações automáticas) ou no usuário (por meio de determinada avaliação individual) (Bucher et al., 2005).

Contudo, de acordo com (Martins et al., 2005) um sistema completo de RIG envolve diferentes componentes, os quais influenciam um ao outro, podendo se beneficiar de uma avaliação separada. O autor reúne com isso as principais necessidades relacionadas a esse tipo de avaliação:

1) Construção de uma ontologia geográfica para dar suporte a RIG;
2) Verificar referências geográficas nos textos;
3) Definir contextos geográficos para documentos;
4) Ranquear documentos de acordo com a relevância geográfica;
5) Construir interfaces visuais para a RIG.

Tabela 1: Classificação Binária dos Problemas na RI (Martins et al., 2005)

	Itens Relevantes	Itens Irrelevantes
Considerados Relevantes	Positivos Verdadeiros (pv)	Falsos Positivos (fp)
Considerados Irrelevantes	Falsos Negativos (fn)	Verdadeiros Negativos (vn)

(Martins et al., 2005) apresenta também as principais métricas de avaliação utilizadas na RI tradicional as quais são também largamente utilizadas na RIG, seguindo a classificação apontada pela tabela 1. Algumas das medidas mais populares são a abrangência (*recall*) ($r = \frac{pv}{pv + fn}$) que corresponde à razão dos documentos relevantes recuperados pelo número total de documentos relevantes na coleção, a precisão (p = $\frac{pv}{pv + fp}$) que é a razão dos documentos relevantes recuperados pelo total de

documentos recuperados e a f1(p,r) = $\frac{2pr}{p+r}$ que é a média harmônica entre a precisão e a abrangência.

A avaliação correta de sistemas de RIG é requisito fundamental para a evolução da qualidade desses sistemas. Atualmente uma das iniciativas para avaliação e comparação de sistemas relacionados à informação geográfica é o GeoCLEF[8] o qual surgiu em 2005 no CLEF[9] e é o primeiro fórum de avaliação de sistemas de RIG. O principal objetivo é disponibilizar o *framework* necessário visando avaliar sistemas de RIG para buscas envolvendo os aspectos geográficos em várias línguas (atualmente são suportados o Português, o Alemão e o Inglês). Para isso são disponibilizadas coleções específicas de documentos (geralmente notícias) em cada linguagem e várias tarefas envolvendo seu georreferenciamento (como a resolução de ambigüidades).

Outra fonte de dados bastante utilizada é o ODP[10] o qual é um sistema de busca que organiza as informações em diretórios os quais são administrados por voluntários em todo o mundo. Uma das seções (Regional) busca a organização de sites baseado em seu foco geográfico (como sites sobre algum lugar, sobre alguma empresa em determinada cidade, etc.) sendo cada página organizada de forma hierárquica entre seções de acordo com seu contexto geográfico (Amitay et al., 2004). Dessa forma a avaliação pode ser realizada através da análise do conteúdo das páginas em cada categoria pelo algoritmo de RIG a ser testado e a posterior comparação dos resultados com o nome da categoria correspondente.

2.2 Resolução de Topônimos

Devido ao fato de poder identificar de forma precisa um determinado espaço geográfico, a utilização de técnicas para identificação correta de topônimos (particularmente cidades, estados e países) apresenta-se como possibilidade importante para a inferência do contexto geográfico abordado pelos textos.

Com a correlação entre espaço geográfico e linguagem textual tornam-se

[8] http://ir.shef.ac.uk/geoclef/

[9] http://www.clef-campaign.org/

[10] http://www.dmoz.org/

possíveis variadas aplicações, dentre elas (Leidner, 2007) :

Agrupamento de informações pela localidade relacionada – Com o reconhecimento da localidade referenciada pelos textos é possível o agrupamento de informações levando em conta o caráter geográfico do texto, agilizando assim a consulta por informações relacionadas a determinado lugar.

Navegação espacial através de *geobrowsers* – Com a popularidade dos serviços de navegação espacial na internet (como Google Earth[11]) e outros serviços de mapas (como Google Maps[12], Yahoo Maps[13]) e a correspondente capacidade de integração de conteúdos diversos por meio de suas APIs abertas, torna-se possível associar informações diretamente com sua posição espacial relacionada.

Serviços Baseados em Localização (*Location-Based Services*) – Com o caráter pervasivo dos dispositivos móveis como os celulares e a inclusão neles de mecanismos capazes de reconhecer a localização do usuário (ex: GPS) juntamente com serviços capazes de associar a esses dados outros tipos de informações georreferenciadas (ex: notícias) torna possível o direcionamento de informações relevantes ao tempo e local do usuário.

No entanto, devido às características dos espaços geográfico e textual surge a necessidade de lidar com vários tipos de ambigüidade para a correta identificação das localidades. Por exemplo, no aspecto geográfico um topônimo como entidade geopolítica pode mudar de nome ou extensão ao longo do tempo, já no âmbito lingüístico lugares distintos na Terra podem compartilhar o mesmo nome. Segundo (Leidner, 2007) a área encarregada de estudar estes problemas é definida como Resolução de Topônimos (RT), a qual tem o objetivo, portanto de possibilitar o mapeamento correto das localidades referenciadas pelos textos a partir da resolução dos vários tipos de ambiguidades relacionadas a elas.

O presente capítulo busca com isso investigar os desafios envolvidos para a Resolução de Topônimos, relatando as principais características dos topônimos e tipos de ambigüidades envolvidas (seção 2.2.1), complementando com um novo tipo de

[11] http://earth.google.com/
[12] http://maps.google.com/
[13] http://maps.yahoo.com/

ambigüidade identificado e definido por este trabalho (seção 2.2.2) bem como com as principais estratégias e arquiteturas já utilizadas para a desambiguação de topônimos como um todo (seções 2.2.3, 2.2.4 e 2.2.5). O objetivo é ilustrar as particularidades do georreferenciamento de textos baseado em topônimos e indiretamente demonstrar as necessidades envolvidas para a atualização dinâmica de *Gazetteers* com as referências ou mais especificamente com os Indicadores de Localidade relacionados.

2.2.1 Características dos Topônimos

Nomes de lugares, como cidades ou mesmo países exibem suas próprias idiossincrasias, merecendo com isso tratamentos especiais (Leidner, 2007). Para entender como realizar a Resolução de Topônimos em textos é necessário inicialmente compreender como estes podem ser referenciados nos textos e quais tipos de ambigüidades podem estar relacionados.

De acordo com (Garbin e Mani, 2005) um topônimo é o nome de uma entidade geográfica na superfície da Terra que pode ser representada por alguma especificação geométrica em um SIG, por exemplo, como um ponto, linha ou polígono.

Como ilustrado na seção 2.1 há muitas diferenças entre a visão espacial (geográfica) da RIG e o jeito que um lugar é descrito em uma linguagem natural. O tipo de ambigüidade a ser solucionado vai depender, portanto do nível de georreferenciamento utilizado.

Com relação a isso um desafio é capturar os limites de expressões vagas (vernaculares), as quais se relacionam a lugares imprecisos onde a extensão espacial reflete uma percepção comum (Twaroch et al., 2008) não correspondendo a uma terminologia oficial e administrativa de um lugar (Jones et al., 2007), por ex “Sul do Rio Grande do Sul”, “Centro de Pelotas” ou “próximo a Rio Grande”. Contudo, em geral os trabalhos buscam realizar a resolução de topônimos considerando seu aspecto lingüístico e a extensão espacial padrão relacionada, não compreendendo assim esse nível de refinamento para o georreferenciamento.

Do ponto de vista linguístico, portanto as principais dificuldades relacionam-se a inferência correta das localidades referenciadas nos textos, visando com isso à superação das seguintes ambigüidades (Clough et al., 2004), as quais são exemplificadas na tabela 2:

Ambigüidade da Referência (ARC - *Reference Ambiguity*) – quando determinada localidade pode ser referenciada por vários nomes diferentes (ex: devido a outros nomes históricos e transliterações).

Ambigüidade do Referente (ART - *Referent Ambiguity*) – quando o nome pode ser usado para referenciar outras localidades (ex: cidade com o mesmo nome de outra).

Ambigüidade da Classe do Referente (ACR - *Referent Class Ambiguity*) – quando o nome pode ser usado para referenciar outros tipos de entidades (ex: nome de pessoas, nomes de empresas).

Tabela 2. Tipos de Ambigüidade para a Resolução de Topônimos

		Localização	**Ambiguidade**
Geo/Geo	ARC	Pelotas, RS Rio de Janeiro, RJ	Princesa do Sul Rio
	ART	Bom Jesus, RS Belém, PA	Bom Jesus, RN Belém, PB
Geo/Não Geo	ACR	Pelotas, RS Serra, ES	Rua Pelotas Serra (Governador de São Paulo)

(Amitay et al., 2004) divide essas ambigüidades em Geo/Geo (compreendendo a ARC e a ART) e Geo/Não Geo (compreendendo a ACR).

Outro problema é que um nome de lugar pode estar sendo utilizado no texto como metonímico de outro, ou seja, uma localidade ser referenciada a outra entidade que é relacionada a ela (Leveling e Hartrumpf, 2006b). Exemplos: 1) “Rio Grande prepara-se para a construção do dique seco”, 2) “Pelotas assinou ontem lei para a educação” nesses casos o topônimo não é referenciado no texto com o aspecto geográfico propriamente, mas sim se relaciona a pessoas localizadas na cidade (no caso 1 à população em geral e no caso 2 especificamente ao prefeito).

Como para o presente trabalho busca-se considerar o sentido padrão das localidades (não levando a metonímia envolvida) o decorrer desse e dos demais capítulos buscam abranger os principais desafios relacionados à superação dos tipos de ambigüidades padrão relacionadas à Resolução de Topônimos (conforme a tabela 2).

Alguns trabalhos têm proporcionado a mensuração dessas ambigüidades. (Smith e Mann, 2003) quantificou o tipo e o grau de ambigüidade dos topônimos examinando a ontologia geográfica TGN, conforme ilustra a tabela 3.

Tabela 3. Ambigüidade presente na TGN por (Smith e Mann, 2003)

Continente	% Lugares com Múltiplos Nomes (Sinônimos)	% Nomes com Múltiplos Lugares (Homônimos)
América do Norte e Central	11.5	57.1
Oceania	6.9	29.2
América do Sul	11.6	25.0
Ásia	32.7	20.3
África	27.0	18.2
Europa	18.2	16.6

Já (Chaves e Santos, 2006) verificou, a partir da análise de um corpus de exemplo da coleção WPT 03[14] na ontologia geográfica Geo-Net-PT-01, que 75% das entidades geográficas possuíam mais que uma palavra, sendo que 31.21% das entidades representavam pessoas e 23.43% organizações. Outra análise compreendeu o tipo de EM apresentada, com relação a isso as mais freqüentes entidades geográficas compreendiam cidades ou vilas, seguidas pelo nome de países.

Com relação a essas análises algumas técnicas podem ser úteis para medir a entropia (no caso o nível de ambigüidade entre topônimos) automaticamente para uma determinada coleção, auxiliando assim no processo de desambiguação. Algumas das principais são utilizadas por (Overell e Ruger, 2007) e compreendem a KL Divergência (Raghavan et al., 2004) e a Informação Mútua (*Mutual Information* (MI)) (Cover e Thomas, 1991), as quais buscam medir o nível de informações compartilhadas por determinadas coleções.

Outro tipo de ambigüidade particularmente identificada e atendida pelo presente trabalho, sendo definida como *Ambigüidade Indireta da Referência* é mais bem ilustrada a seguir.

2.2.2 Ambigüidade Indireta da Referência

A partir do desenvolvimento do trabalho e do estudo realizado em textos

[14] http://linguateca.di.fc.ul.pt/qWPT03/

jornalísticos pôde-se identificar um novo tipo de ambigüidade, a qual se relaciona aos textos que não possuem nomes de localidades explicitamente em seu conteúdo, embora possuam entidades que de forma indireta possibilitam a inferência das localidades.

Esse tipo de ambigüidade, definida aqui como *ambigüidade indireta da referência* pode ser mais bem entendida a partir da identificação dos tipos de localizações possíveis de serem inferidas por meio de determinado Web-site, as quais de acordo com (Wang et al., 2005) podem ser as seguintes:

- **Localização do Provedor do Conteúdo**: A localização geográfica atual relacionada à empresa dona do Web-site.
- **Localização do Conteúdo**: A localização representada através do conteúdo do Web-site.
- **Localização do Serviço**: O escopo geográfico da audiência que o site deseja alcançar.

Exemplificando, no caso do *Web Site* do jornal Folha de São Paulo[15] a localização do Provedor, ou seja, a localização física da empresa está localizada em São Paulo, a do Conteúdo pode ser identificada analisando as localidades referenciadas no texto de cada notícia e a do Serviço pode ser considerado o Brasil como um todo, mas com foco específico nos estados do Rio de Janeiro e de São Paulo (o site busca direcionar notícias de interesse do Brasil inteiro, mas tem atenção especial para notícias desses estados).

Dessa forma, como o escopo geográfico da audiência relaciona-se a cidades/estados específicos, as localidades referenciadas no conteúdo dos textos não necessitam serem apresentadas explicitamente, bastando pra isso a notícia possuir referência a determinadas entidades que são associadas a essas regiões. (Leveling et. al., 2007) constatou a importância desse tipo de entidade no auxílio à identificação do contexto geográfico dos textos, conceituando-os como Indicadores de Localidade (*Location Indicators*) e definindo seus tipos, os quais podem ser:

Adjetivos e Sinônimos = Rio de Janeiro => Cidade Maravilhosa, Rio de Janeiro => Rio.
Gentílico = Rio de Janeiro => Carioca.

[15] http://www.folha.com.br/

Códigos para Localidades = Brasília => BSB (Sigla Aeroportuária).
Abreviações e Acrônimos = S. Paulo => São Paulo, POA => Porto Alegre.
Variações Ortográficas de Idioma = São Paulo => San Pablo (Spanish).
Outras Entidades Associadas à Localidade = Praças, Ruas, Aeroportos, Autoridades (Prefeito), Viadutos, Rodovias.

Os Indicadores de Localidade incluem Entidades Não-Geográficas, as quais geralmente possuem caráter temporário, sendo relacionadas a uma situação ou acontecimento específico. Por exemplo, nomes de pessoas relacionadas ou naturais de determinada cidade que são alvo de notícias (estando envolvidas, por exemplo, em acontecimentos ou outras situações as quais tem relação direta com uma cidade específica).

Com isso, a efetividade do georreferenciamento depende da abrangência das entidades que podem ser reconhecidas nos textos. A utilização de dicionários toponímicos (*Gazetteers*) para a resolução desse tipo de ambigüidade demanda, portanto a correta representação dos Indicadores de Localidade e sua associação com as localidades identificadas por eles. Naturalmente a existência dessas estruturas e a abrangência de Indicadores tornam-se potencialmente úteis também para a desambiguação de outros tipos de ambiguidades, particularmente as ambiguidades da referência e do referente.

Para melhor compreender o processo de resolução de topônimos, a seção seguinte apresenta as principais estratégias utilizadas para a desambiguação dos vários tipos de ambigüidade já apresentados.

2.2.3 Estratégias para Resolução de Topônimos

A RT abrange duas etapas principais (identificação e desambiguação). Primeiramente as referências geográficas devem ser identificadas (superando assim a ACR) de acordo com o tipo, por exemplo, cidade ou país, para posteriormente serem desambiguadas, ou seja, descritas com uma extensão espacial única (superando a ART e a ARC), podendo então ser associadas a sua localização geográfica em um *Gazetteer*. Para ser completa a desambiguação deve abranger também, conforme definido por este trabalho, a ambigüidade indireta da referência, ou seja, ser capaz também de identificar e desambiguar textos que não possuem localidades explicitamente, mas contenham

algum tipo de Indicador de Localidade. Esta informação pode com isso ser utilizada em outras tarefas, tais como a indexação e recuperação de documentos de acordo com os seus âmbitos geográficos (Jones et al., 2004 *apud* Martins et al., 2006a).

Para cada uma dessas etapas são necessárias técnicas específicas. A primeira etapa (identificação) está principalmente relacionada ao Reconhecimento de Entidades Mencionadas (REM) do inglês *Named Entity Recognition* (Leidner, 2007), a qual segundo (Clough et al., 2004) representa o processo de assinalar a cada palavra ou grupo de palavras uma determinada categoria pré-definida. Para a RT a classificação limita-se consequentemente à categoria "Localização". Ao se lidar com referências geográficas o foco é a utilização dessas entidades em outras tarefas de recuperação da informação, tendo as referências obrigatoriamente que ser associadas a uma representação única para o conceito geográfico subjacente (Martins et al., 2006a).

Para a segunda etapa (desambiguação) utilizam-se técnicas principalmente inspiradas na WSD. De acordo com (Ide e Véronis, 1998) a Desambiguação Lexical de Sentido (*Word Sense Disambiguation* ou WSD) tem sido almejada desde o início do tratamento de linguagem em computadores nos anos 50 (1950), sendo que desde então os trabalhos buscam considerar tanto ambigüidades relacionadas à polissemia quanto a de homonímia (Specia e Nunes, 2004). Com isso a WSD consiste de forma geral na classificação de uma palavra ambígua a partir da avaliação do sentido dos candidatos (verificando por ex. os diferentes significados) de acordo com um contexto particular, por ex: "Banco de areia" ou "Banco do Brasil".

Portanto, devido a seus objetivos similares, os trabalhos relacionados à desambiguação de topônimos têm utilizado com sucesso várias das técnicas e heurísticas relacionadas à WSD. Embora similares, a desambiguação de topônimos apresenta desafios próprios. De acordo com (Li et al., 2003) através da WSD é possível identificar o topônimo como localidade, mas nem sempre é possível inferir o sentido correto relacionado a ele. Por exemplo, na frase: "A Casa Branca é localizada em Washington", a expressão "localizada em" pode somente determinar que "Washington" é um nome de lugar, mas não conseguiria decidir o atual sentido da localização (se é o estado ou a cidade), nesse caso torna-se necessário realizar outros tipos de processamento. Uma das estratégias é analisar termos que co-ocorrem no texto e são

relacionados a determinado estado, já que não é comum um mesmo estado possuir duas cidades com o mesmo nome.

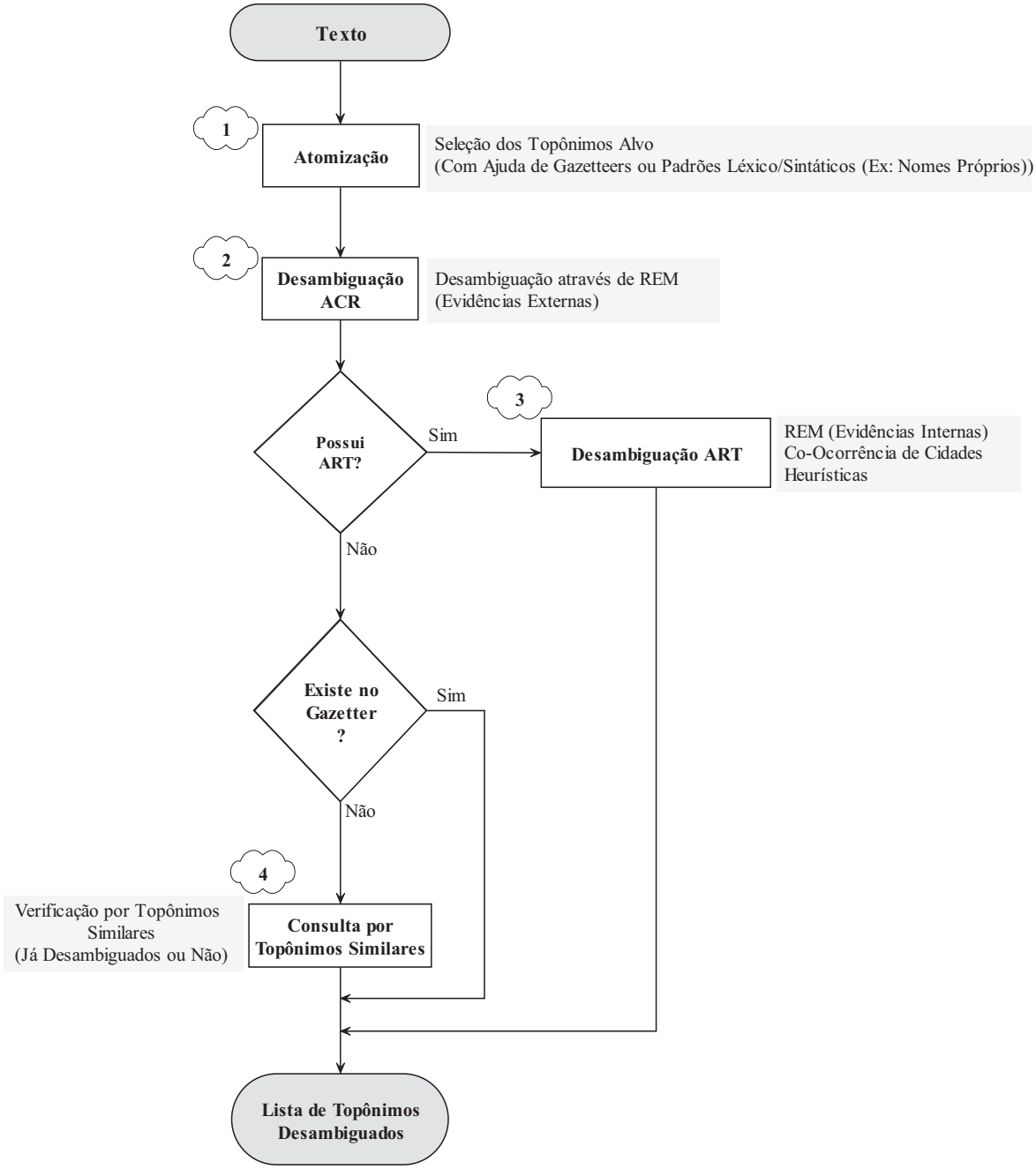

Figura 3. Procedimento Padrão para a Resolução de Topônimos

A partir da análise de trabalhos envolvendo à RT, o presente trabalho organiza e resume o seu procedimento padrão, conforme ilustra a figura 3. Na primeira etapa o texto é atomizado (as strings são separadas) e normalizado (são retirados caracteres inúteis para análise como *stopwords*, sinais de pontuação, etc.), visando recuperar os possíveis topônimos alvo da análise. Para isso basicamente são utilizadas duas

abordagens: a recuperação dos topônimos alvo através de um *Gazetteer* ou o reconhecimento por meio de análise léxico/sintática (ex: todas as palavras em maiúsculo, Nomes Próprios, etc.).

Na etapa 2 as *strings* recuperadas são comparadas com padrões e regras sintáticas visando separar as que se relacionam às localizações (desambiguação da classe do referente). Após na etapa 3 é necessário verificar se alguma localidade é homônima de outra (superando a ambigüidade do referente), para isso utiliza-se heurísticas (ex: considerar como sentido padrão a cidade com maior número de habitantes) e técnicas de REM.

Um procedimento também bastante importante é a verificação de similaridade (etapa 4), no caso da localidade não estar referenciada no *Gazetteer*, podendo esta ser realizada nos topônimos já desambiguados ou para todos os topônimos existentes no *Gazetteer*. A desambiguação relacionada à ambigüidade da referência está implícita nas etapas 1 e 3 (no caso da utilização de *Gazetteers* para a identificação dos topônimos alvo) e na etapa 4 (na representação geográfica final do topônimo desambiguado) visto que a desambiguação é realizada através de uma lista de alguns termos sinônimos (incluindo por ex: abreviaturas) relacionados aos topônimos. A ambigüidade indireta da referência pode ser vista por sua vez como um complemento à ambigüidade da referência com a diferença que para sua resolução o *Gazetteer* deveria abranger outros tipos de entidades (Indicadores de Localidade) conforme descreve a seção anterior.

Para a representação dos topônimos desambiguados as duas principais estratégias relacionam-se a utilização de coordenadas geográficas (ex. latitude e longitude) e estruturas hierárquicas administrativas (ex: Brasil/Rio Grande do Sul/Pelotas) (Hu e Ge, 2007).

É interessante notar que os trabalhos relacionados à RT geralmente apresentam estratégias com foco em ambiguidades específicas, não abrangendo com isso todas as ambiguidades citadas.

De forma geral essas estratégias são agrupadas dentro das seguintes categorias, podendo, no entanto compartilhar ambas (Clough et al., 2004; Overell e Ruger, 2007 e (Hu e Ge, 2007):

Baseada em Regras (*Rule-Based*) – baseiam-se em regras e heurísticas definidas manualmente (*knowledge-based*) visando reconhecer e desambiguar as entidades geográficas. Pode utilizar dois tipos de evidências (conforme exemplifica a tabela 4): "internas" as quais se baseiam no próprio nome da localidade (ex: verificar diretamente no texto todos os nomes de cidades, ou padrões relacionados a letras maiúsculas e/ou siglas de estados) e as "externas" as quais buscam identificar expressões relacionadas ao contexto no qual a localidade aparece (ex: expressões de contexto como *na cidade de, na localidade de,* etc.*)* (McDonald,1996). A tabela 5 apresenta uma relação de expressões de contexto utilizadas por (Martins et al., 2006a).

Tabela 4. Tipos de Evidências para a Desambiguação de Topônimos (Exemplos)

Internas (Pelo Próprio Nome incluindo ou não separadores)	(1ª Palavra de Alguma Frase) NOME CIDADE Ex: Pelotas a cidade do doce. PAL. MAIÚSCULO +{/,-}+Sigla de Algum Estado Ex: Pelotas/RS ou Pelotas-RS
Externas (Por Expressões de Contexto relacionadas ao Topônimo)	{em,de,na cidade de,no município de}+PAL.MAIÚSCULO Ex: na cidade de Pelotas {região metropolitana de}+PAL.MAIÚSCULO Ex: região metropolitana de Porto Alegre

Tabela 5. Expressões de Contexto em Português (Martins et al., 2006a)

Tipo de Expressão	**Expressão**
Identificadores de Contexto	cidade, município, distrito, rua, avenida, rio, ilha, montanha, vale, país, continente, zona, região, condado, freguesia, deserto, província, povoado, aldeia, monte, vila, república, península
Localização	fora de, nos arredores de, dentro de, entre, em, acima, ao longo, atrás, acima, ao lado, à esquerda, à direita
Distância Relativa	adjacente, longe de, perto de, próximo de
Orientação	leste, norte, sul, oeste, oriente, ocidente, sudeste, sudoeste, nordeste, noroeste
Outras Expressões	"cidades como", "e outras cidades", "cidades incluindo", "cidades especialmente", "uma das cidades", "cidades tais como"

O conjunto de regras é armazenado em um *Gazetteer*. Um problema da abordagem baseada em regras é que as regras são fixas e manualmente incluídas, o que

pode ser útil apenas para domínios específicos (Clough et al., 2004). Para uma maior abrangência outra estratégia relaciona-se a utilização de técnicas de análise estatística (ex: Aprendizado de Máquina) visando capturar automaticamente palavras e expressões que auxiliem na identificação e desambiguação das localidades, essa abordagem denominada de Guiada pelos Dados (*Data-Driven)* é ilustrada a seguir.

Outra estratégia também bastante comum envolvendo a abordagem Baseada em Regras compreende a inferência de heurísticas para a identificação e desambiguação das localidades (Amitay et al., 2004 e Overell e Ruger, 2007). Uma revisão das heurísticas utilizadas é apresentada na seção 2.2.4.

Guiada pelos Dados (*Data-Driven*) – aplica Análises Estatísticas e métodos de Aprendizado de Máquina (*Machine Learning*), onde um *corpus de treino* deve ter seus topônimos reconhecidos e desambiguados visando à identificação de regras e classificadores úteis para a desambiguação de topônimos vistos ou não durante o treino. Dependendo do esforço para a seleção do corpus de treino as técnicas podem ser divididas respectivamente em Supervisionada, Semi-Supervisionada e Não-Supervisionada, sendo que nessa última a seleção do corpus de treino é feita de forma automática não dependendo de anotação do corpus de treino.

Contudo, os esforços e custos requeridos para construir um corpus com qualidade e ampla cobertura para o corpus de treino podem ser muito significativos (Leidner, 2004 *apud* Hu e Ge, 2007). Uma solução parcial é utilizar métodos como os de *bootstrapping*, os quais podem usar corpora de treino reduzidos, combinando uma pequena quantidade de dados pré-classificados (com múltiplos exemplos de cada ambigüidade) com uma grande quantidade de dados não-classificados (Riloff e Jones, 1999 *apud* Clough et al., 2004).

A utilização da abordagem Guiada pelos Dados passa, portanto pelo tipo de domínio e esforço pretendidos. Embora úteis para domínios específicos a abordagem baseada em regras pode exigir muito tempo para a reunião dessas, para diferentes domínios a utilização de técnicas estatísticas apresenta-se com mais portabilidade e pouca intervenção manual (Clough et al., 2004).

Outro tipo de análise estatística envolve a verificação de termos que co-ocorrem no texto. Duas verificações comuns principalmente para identificação de sinônimos referem-se à análise de co-ocorrência e de *collocations*. De acordo com (Manning e

Schutze, 1999) uma *collocation* é uma expressão consistindo de duas ou mais palavras, em uma ordem particular, que correspondem a um jeito convencional de dizer certas coisas. Já co-ocorrência é menos restrito, representando simplesmente palavras que ocorrem no mesmo documento. Por exemplo, na frase: "A Recuperação de Informações é Importante na Web", o termo "Recuperação de Informações" poderia ser considerado uma *collocation* visto que é uma expressão popularmente conhecida, já as palavras "Recuperação Web" poderiam ser consideradas co-ocorrentes visto que não representam uma expressão propriamente, apenas apresentando co-ocorrência na frase. Nesse trabalho, no entanto não diferenciamos os termos, considerando ambos como "co-ocorrência".

É importante notar, portanto que ambas as estratégias auxiliam diretamente na desambiguação das ambigüidades do referente e da classe do referente. Contudo para auxiliar na separação das regras as análises estatísticas representam um importante aliado, podendo auxiliar na desambiguação de vários tipos de ambiguidades citados. O que acaba viabilizando com isso a partir da identificação de termos co-ocorrentes a seleção de palavras sinônimas (ajudando na resolução da ambigüidade da referência) e também a recuperação de outros Indicadores de Localidades (auxiliando na resolução da ambigüidade do referente e ambigüidade indireta da referência).

Essas entidades recuperadas são armazenadas então em *Gazetteers* para serem utilizadas posteriormente no processo de georreferenciamento. Com relação especificamente as estratégias que buscam recuperar Indicadores de Localidade alguns dos principais problemas são a grande dependência de anotação de corpora de treino para a verificação das entidades e o foco em idiomas específicos. Estes problemas acabam demandando com isso estratégias que possibilitem a atualização desses *Gazetteers* de forma ágil com abrangência de variados tipos de Indicadores (considerando pra isso o seu caráter dinâmico) conforme ilustra a seção 2.3.

Para apoiar a resolução de topônimos (principalmente a ART) são utilizadas também variadas heurísticas as quais acompanham as abordagens detalhadas acima. (Leidner, 2007) reúne todas elas, as principais são citadas na seção seguinte.

2.2.4 Heurísticas para a Resolução de Topônimos

Para apoiar a resolução de topônimos variadas heurísticas são utilizadas. (Leidner, 2007) apresenta uma relação das principais (as quais possuem foco principalmente na resolução das ambiguidades geo/geo):

1) Qualificador de Continência: Procurar padrões, por exemplo, "Pelotas/RS", "Pelotas, Rio Grande do Sul", relacionando cidade ao estado ou estado ao País visando assim à desambiguação.

2) Maior População: Prioriza o topônimo homônimo com maior população.

3) Uma Referência por Discurso: Assume que todos os Topônimos homônimos em um texto compartilham apenas um sentido, sendo necessário desambiguar apenas um (Gale et al.,1992). Por exemplo, o sentido da primeira ocorrência da localidade pode ser assumido como o sentido correto para as demais com o mesmo nome. Esta heurística é proveniente da WSD.

4) Minimalidade Geométrica (*Minimal Bounding Polygon*): Verifica a distância espacial mínima, analisando todas as localidades homônimas e também demais localidade no texto.

5) Priorizar Capital: Se a localidade homônima possuir o nome de uma capital, assumir ela como sentido.

6) Ignorar Cidades Pequenas: Limitar a verificação para Cidades Grandes (de acordo com sua população).

7) Foco numa área Geográfica: Ignorar localidades fora de determinada área Geográfica, podendo ser de forma estática (ex: limitando o *Gazetteer* para cidades de determinada área), ou dinâmica (ex: verificar o país relacionado a determina fonte de notícia e priorizar a resolução de topônimos para localidades relacionadas a ele).

8) Distância para Cidades não Ambíguas: Define como sentido o topônimo não-ambíguo geograficamente mais próximo do centróide das localidades homônimas. Os topônimos não-ambíguos relacionam-se a todos os que estiverem a até X palavras (onde X pode variar para cada trabalho) do nome da localidade a ser desambiguada.

9) Mais Freqüentes: Conceder mais relevância para topônimos mais freqüentes no texto. Particularmente útil para apoiar outras heurísticas e fórmulas ajudando a definir qual

topônimo priorizar para *rankings* de relevância similares (com cidades compartilhando pesos iguais ou parecidos).

10) Preferência Hierárquica: Priorizar sempre o nível superior no caso de topônimos homônimos em níveis hierárquicos diferentes (ex: em textos com a cidade "Rio de Janeiro" o sentido seria o do estado).

11) Desambiguação pelo tipo do Identificador: Busca padrões relacionados ao tipo do identificador do topônimo e elimina os candidatos à desambiguação que não possuem determinado identificador (ex: "cidade do Rio de Janeiro" eliminaria a verificação de ambigüidade com o estado).

12) Correlação entre Texto e Espaço: Assumir que os topônimos ocorrendo próximos no texto têm relação espacial próxima também.

13) Referente Padrão: Calcular qual o sentido do topônimo mais freqüente numa coleção de texto e assumir esse sentido como padrão para análises futuras.

14) Preferência do *Gazetteer*: Definir uma ordem de verificação no caso de utilização de mais de um *Gazetteer*.

É interessante notar que a maioria das heurísticas foram utilizadas por sistemas com suporte global a resolução de topônimos, algumas podendo não se aplicar a desambiguação com focos mais específicos. A utilização de cada uma vai depender, portanto da aplicação e do tipo de corpora utilizado. Por exemplo, no caso da desambiguação de topônimos em notícias com foco em cidades de um determinado país, a heurística 10 não teria utilidade, já que não seria preciso identificar relações hierárquicas, diferentemente da heurística 7 a qual poderia ser aplicada. A heurística 4 também não seria útil caso o sistema não associasse os topônimos a coordenadas geográficas.

2.2.5 Arquiteturas para a Resolução de Topônimos

A presente seção descreve em mais detalhes alguns trabalhos e abordagens relevantes que utilizam de alguma forma as estratégias mencionadas para resolução parcial ou integral dos tipos de ambiguidade relacionadas à RT.

(Li et al., 2003): InfoXtract – Apresenta uma abordagem híbrida para a desambiguação de localizações, consistindo de uma busca local de padrões e também análise de co-

ocorrência (levando em conta nomes de topônimos no texto). Utiliza um algoritmo de REM onde os topônimos encontrados são reconhecidos como localidade para após inferir o tipo (Cidade ou Estado) a partir de um módulo de desambiguação, o qual faz uso de heurísticas relacionadas ao sentido padrão dos topônimos. Verifica também padrões relacionados ao contexto local dos topônimos e utiliza um algoritmo de árvore geradora máxima (*maximum spanning tree*) visando desambiguar candidatos restantes.

O grafo relacionado à árvore geradora máxima é criado relacionando todos os topônimos ambíguos a partir de pesos específicos. Os pesos são calculados levando em conta o sentido dos *links* das localidades e as categorias de co-ocorrência relacionadas. Por exemplo, quando uma localidade com um sentido potencial para cidade co-ocorrer com uma localidade com sentido de estado e a cidade estiver dentro do estado o peso do estado é considerado alto, com peso definido para 3. Outros pesos são dados também para cidades e países.

O trabalho consegue uma acurácia de 96% para topônimos ambíguos, a heurística relacionada ao sentido padrão dos topônimos obteve 89,9%, já os padrões locais obtiveram somente 12% de acurácia analisando notícias da CNN[16].

(Amitay et al., 2004) Web-a-Where: Desenvolve um algoritmo visando identificar o foco geográfico da página, para isso utiliza um *Gazetteer* com informações de várias fontes com dados relacionados a topônimos (limitado a cidades com mais de 500.000 habitantes com estados, países e abreviações relacionadas) de todo o mundo, representando-os através de uma hierarquia geográfica (ex: País/Estado/Cidade) não utilizando nenhum conhecimento baseado em coordenadas geográficas.

Para a desambiguação geo/geo o sistema primeiro procura por todas as referências geográficas no texto (definidos como *spots*). Após, são utilizadas várias heurísticas associando confidências distintas para cada topônimo relacionado, por ex: procurando a sigla do estado seguida da cidade (peso = 0.95) ou a cidade ambígua comum compartilhada por topônimos sem esse tipo de identificação (peso entre 0.65 a 0.75 dependendo se o spot compartilha o significado padrão).

Após isso o foco é definido de acordo com esses pesos podendo ser relacionado à cidade, estado ou país (estado ou país no caso de existir muitas cidades ou estado

[16] http://www.cnn.com/

relacionados no texto), não sendo consideradas referências abaixo de determinado limiar (0.9).

Encontra um foco em 75% das páginas relacionadas ao corpus de teste utilizado (ODP), sendo 65% para estado ou cidade. Medidas de qualidade aumentam quando avaliam também cidades com mais de 5000 habitantes.

Para a desambiguação geo/não-geo o sistema analisa nomes ambíguos verificando a relação entre a população e a quantidade de ocorrências num corpus com 1200000 páginas, avaliando também a relação entre quantidade de ocorrências em maiúsculo e minúsculo, removendo manualmente exceções que não se encaixam nessas heurísticas. O trabalho é bastante referenciado também por ter definido os tipos de ambiguidades (geo/geo, geo/não-geo) e também os tipos de estratégias para o georreferenciamento (alvo e fonte).

(Garbin e Mani, 2005) Desambiguação de Topônimos em Notícias: Utiliza dois *Gazetteers* públicos (com ocorrência de países e cidades com mais de 500000 habitantes) como base para a resolução (ART e ACR). Devido à falta de identificadores de contexto nos textos (67.82% dos topônimos não os possuíam) adota um mecanismo de aprendizado não supervisionado que utiliza heurísticas para desambiguar topônimos aplicando depois esse processo automático para treinamento obtendo 78,5% de acurácia para ambigüidade da classe do referente. As heurísticas baseiam-se no sentido padrão (relacionado ao sentido mais freqüente relacionado aos topônimos). Após o corpus com os topônimos desambiguados é utilizado para o processo de aprendizado, o qual se baseia na verificação de termos próximos às cidades (foram testados entre 3 e 20 palavras distantes) reconhecidos através do teste de Informação Mútua (entropia) e também pela sua freqüência invertida (tfidf), outras evidências também são verificadas como letras maiúsculas para os topônimos. Os experimentos mostraram que aumentando a quantidade de palavras verificadas próximas às cidades somente há decréscimo da acurácia quando testado no mesmo corpus de treino. Aumentar o corpus de treino melhorou os resultados principalmente para distâncias maiores de palavras.

(Hu e Ge, 2007) Aprendizado Supervisionado para Desambiguação de Topônimos: Utiliza REM para desambiguação Geo/Não-Geo e Aprendizado de Máquina Supervisionado para desambiguação Geo/Geo (ambigüidade do referente), abrangendo cidades e estados da Austrália e recuperando o conhecimento a partir de *Gazetteers*

disponíveis online. O processo de REM segue as seguintes etapas: a primeira etapa verifica no texto a ocorrência de topônimos a partir do *Gazetteer* utilizado sendo retirados topônimos que possam representar outras entidades (definidas como *stopwords geográficas).* A etapa seguinte realiza o processo de reconhecimento de entidades usando um reconhecedor REM pré-treinado para três tipos de entidade: pessoa, localização e organização, após mantém-se somente entidades relacionadas à localização, realizando-se também um procedimento para reconhecer localidades que possam ser parte de uma localidade (exemplo no Brasil: Campos e São José dos Campos).

O sistema avalia a desambiguação da ambigüidade do referente apenas a partir da comparação de métodos baseline (ex: considerar a localização com maior ocorrência nos dados de treino) com técnicas de aprendizado de máquina, obtendo com elas resultados melhores que os definidos para o baseline. A desambiguação da referência é suportada através de siglas e sinônimos relacionados às localidades disponíveis nos *Gazetteers* utilizados.

2.3 Importância e Problemas dos *Gazetteers* Atuais

As ontologias são centrais para o desenvolvimento da Web Semântica e da Web Semântica Geo-espacial, pois elas especificam conceitos formais e suas relações, provendo o meio para criação de metadados semânticos para objetos (documentos, banco de dados, etc.) (Perry et al., 2007).

No campo da RIG, para a representação de nomes de lugares utilizam-se estruturas próprias denominadas de *Gazetteers*, os quais podem apresentar variedades de dimensões (ex: escopo, granularidade, abrangência) (Leidner, 2004). De acordo com a dimensão de sua estrutura, os *Gazetteers* podem variar de simples listas de lugares (*flat Gazetteers)* (Martins et al., 2005) a estruturas mais complexas, abrangendo relacionamentos entre as localidades e possibilitando variados tipos de inferência, sendo vistos nesse caso como geo-ontologias ou ontologias geográficas. Com isso os *Gazetteers* têm sido utilizados em vários campos, por exemplo, na desambiguação e expansão de consultas, na anotação de documentos, nos sistemas de *ranking* por relevância, além da identificação de conceitos geográficos em textos.

Na internet as primeiras tentativas para sua utilização relacionaram-se ao reconhecimento de localidades, utilizando para isso *flat Gazetteers*, em consultas enviadas para sistemas de busca, nesse caso cada lugar era associado à *map-grids* ou coordenadas geográficas. Essa identificação poderia ser utilizada então para consultas geográficas associando as informações retornadas ao mundo das coordenadas (Jones et al., 2001).

Para as tarefas relacionadas à identificação e desambiguação de entidades geográficas em textos (a partir de técnicas de REM e WSD) os *Gazetteers*, ou dicionários toponímicos (Hill, 2000) têm sido utilizados para armazenar os variados tipos de referências às localidades e sua posição espacial (utilizando pra isso coordenadas geográficas). Com isso eles podem integrar-se às geo-ontologias para apoiar inferências geográficas mais complexas (Souza, 2005) ou mesmo serem utilizados isoladamente visando auxiliar nos processos de georreferenciamento, conforme ilustraram as seções anteriores.

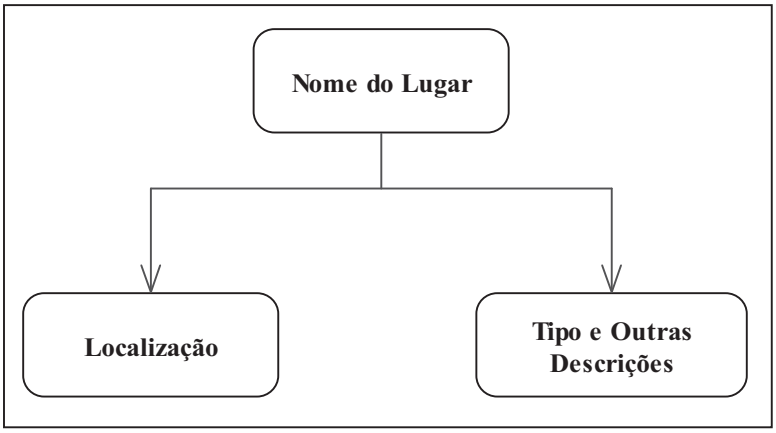

Figura 4: Principais Componentes de um *Gazetteer*

Embora os *Gazetteers* possam apresentar variados tipos de estrutura, grande parte das estratégias utilizam a organização proposta no trabalho de (Hill, 2000) a qual busca relacionar os componentes centrais de um *Gazetteer* (conforme a figura 4), os quais correspondem: a um nome de um lugar geográfico, a correspondente localização que ele abrange (*footprint*) e uma descrição sobre seu tipo.

Os nomes são expressos em linguagem natural e podem incluir nomes alternativos. A localização armazena as coordenadas geográficas e o tipo expressa as relações dos conceitos relacionados ao nome, sendo que a organização hierárquica tem

sido priorizada, visto que facilita a organização espacial das informações sendo considerada uma das melhores formas de se representar o mundo geográfico (Smith, 1995 *apud* Fonseca et al., 2000). Com relação à localização ou *footprint* descrito por meio de coordenadas geográficas, conforme ilustra a seção 2.1.2, ele pode ser representado através de um único ponto ou um centróide relacionado a real extensão da localidade, podendo também ser um polígono ou um conjunto de pontos (Jones et al., 2001).

Alguns dos *Gazetteers* disponíveis atualmente na Internet são o TGN (*Getty Thesaurus of Geographic Names),* o GKB (*Geographic Knowledge Base*) e o Geonames.

O TGN[17] é um dos mais conhecidos sendo utilizado em vários estudos e sistemas como é o caso da máquina de busca SPIRIT[18]. Referencia milhões de nomes de lugares em diferentes linguagens, incluindo entidades políticas (cidades, estados, etc.) e outros recursos físicos (rios, etc.), representando de forma hierárquica (devido aos múltiplos contextos pode ser considerada também polihierárquica) e associativa suas relações em conjunto com as coordenadas geográficas relacionadas.

A GKB foi desenvolvida no projeto GREASE[19] da Universidade de Lisboa e integra informações geográficas descrevendo entidades físicas e administrativas, além de integrar também informações relacionadas à Web sites e seus domínios. Uma das ontologias originadas da GKB é a Geo-Net-PT01, a qual contém informações geográficas relacionadas a Portugal.

Já a Geonames[20] é um *Gazetteer* contendo mais que 8 milhões de nomes geográficos e suas características possuindo também nomes alternativos aos lugares. Todos os nomes são categorizados em uma das nove classes disponíveis e subcategorizados em uma das 645 subclasses. Integra ao conhecimento geográfico dados como altitude, latitude/longitude, população e outras características, possuindo várias fontes de conhecimento, principalmente relacionadas a bases de dados públicas dos Estados Unidos, utilizando também a Wikipédia.

---

[17] http://www.getty.edu/research/conducting_research/vocabularies/tgn/
[18] http://www.geo-spirit.org/
[19] http://xldb.di.fc.ul.pt/grease/
[20] http://www.geonames.org/

Contudo, embora estes e outros *Gazetteers* consigam abranger localidades de todo o mundo, eles apresentam pouca cobertura relacionada a países específicos (como o Brasil) (Borges, 2006) possuindo carências na inclusão de Indicadores de Localidades (ex: nomes de ruas, praças presentes nas cidades, ou mesmo variações léxicas e semânticas como siglas, adjetivos e gerúndios relacionados a elas) e apresentando também pouca freqüência de atualização, já que são organizados manualmente. De acordo com (Delboni et al., 2007) a utilidade dos *Gazetteers* para algumas aplicações depende da sua constante atualização, sendo que *Gazetteers* globais sofrem cerca de 20.000 modificações por mês. (Leidner, 2004). Algumas dessas entidades úteis para a identificação das localidades (como ruas e bairros) já se encontram disponíveis em bases de dados específicas, no entanto grande parte das entidades não são acessíveis dessa forma, necessitando de métodos automáticos para a sua recuperação.

Com isso variados trabalhos têm surgido visando a recuperação automática de variados tipos de Indicadores de Localidade, possibilitando assim a atualização ou mesmo a construção dinâmica de *Gazetteers*. Nas seções seguintes são apresentadas as várias estratégias já utilizadas para esse fim (seção 2.3.1), bem como um resumo dos desafios ainda a serem superados (seção 2.3.2).

2.3.1 Estratégias para Identificação de Indicadores de Localidade

Os trabalhos analisados não consideram explicitamente a verificação de Indicadores de Localidade, embora ajudem a recuperar entidades compreendidas por eles.

Alguns trabalhos utilizam aprendizado supervisionado para a recuperação de entidades relacionadas a localizações geográficas a partir de corpora de treino anotado manualmente.

(Overell e Ruger, 2007) extrai termos relacionados a topônimos através da verificação de termos encontrados na descrição dos links que apontam para páginas de cidades na Wikipédia, possibilitando assim a identificação de sinônimos às localidades. (Popescu et al. 2008) por sua vez utiliza a Wikipédia[21] para extrair, a partir da página das localidades, referências às cidades, utilizando todos os links descritos com nomes

[21] http://www.wikipedia.org/

próprios, buscando também identificar coordenadas geográficas e o tipo das localidades (i. e. se é estado ou país), visando assim criar de forma automática um *Gazetteer* com todas as suas propriedades, conforme definido por (Hill, 2000).

(Buscaldi e Rosso, 2007) utiliza um método híbrido, analisando tanto a Wikipédia (para a identificação de termos relacionados a localidades) como a Wordnet[22] (para definição do seu tipo e para comprovar que páginas na Wikipédia são relacionadas exclusivamente a localidades), não capturando assim páginas de entidades não-geográficas homônimas, o objetivo é criar assim uma ontologia geográfica com as localidades organizadas de forma hierárquica de acordo com seu tipo.

(Rattenbury et al., 2007) descreve métodos para extrair *tags* semânticas (relacionadas a eventos e lugares) baseado nos padrões de uso delas no Flickr[23].

O problema é que tanto a Wikipédia quanto a Wordnet e o Flickr dependem de esforço humano para atualização, o que pode causar falta de cobertura (localizações sem informação) ou mesmo escassez de Indicadores.

Já o trabalho de (Borges et al., 2003) por sua vez obtém informações geográficas a partir de páginas da Web, utilizando pra isso um *wrapper* com exemplos selecionados, no entanto, de forma manual. O objetivo é aprimorar um sistema de informação geográfico urbano através da captura automática de referências indiretas como número de telefones, ceps e nomes de lugares.

Visando aproveitar o caráter dinâmico e geográfico dos textos jornalísticos, os quais abrangem uma grande variedade de Indicadores de Localidade e localizações, uma alternativa pra minimizar a falta de cobertura de Indicadores é utilizar notícias publicadas na Web como fonte de extração dessas entidades.

(Ferres et al., 2004) utiliza notícias em inglês e aplica técnicas de aprendizado de máquina semi-supervisionado para obter entidades correferenciadas (ex: Silva, José da Silva) e também pares de acrônimos (ex: USA, *United States of America*). (Maynard et al., 2004) utiliza técnicas parecidas mas foca na recuperação de nomes de pessoas levando em conta as particularidades apenas dos idiomas Hindu, Chinês e Arábico. Já (Kozareva et al., 2006) busca recuperar nomes de pessoas e de lugares utilizando

[22] http://wordnet.princeton.edu/

[23] http://www.flickr.com.br/

expressões de posicionamento (*positioning expressions*) para sua identificação, testando os resultados em um corpus de notícias da Espanha. Os autores sugerem a utilização dessa abordagem para outros idiomas, contudo não identificam a correlação entre os termos extraídos e os correspondentes topônimos relacionados. (Garbin e Mani, 2005) utiliza notícias para identificar relações (*collocations*) entre termos e localidades, não analisando, no entanto, relações em todo o texto (a janela para a análise é limitada a uma distância de 20 termos dos topônimos). Já (Smith e Mann, 2003), utiliza aprendizado de máquina semi-supervisionado (*bootstrapping*) visando identificar também *collocations* úteis para desambiguação das localidades, não considerando, no entanto o grau de importância ou peso das relações identificadas.

A seção seguinte apresenta um resumo desses trabalhos e dos problemas relacionados às suas abordagens.

2.3.2 Problemas das Abordagens Atuais para a Identificação de Indicadores de Localidade

De forma geral, os principais problemas dos trabalhos que buscam de alguma forma recuperar Indicadores de Localidade compreendem:

- A necessidade de seleção e preparação de um corpus de treino, com estratégias focadas em idiomas particulares.
- A análise das relações em janelas com distância limitada entre os termos (localidades e Indicadores).
- O uso de relações sem peso, não considerando a correspondente importância das relações entre Indicadores e localidades.

Portanto, embora as abordagens busquem recuperar termos que auxiliam na identificação das localidades, elas apresentam algumas desvantagens principalmente relacionadas ao tempo para anotação e preparação do corpus de treino, o corpus com dependência de atualização manual por voluntários (como a Wikipédia) e estratégias focadas em idiomas particulares.

As estratégias que utilizam notícias acabam por sua vez não levando em conta todo o potencial dessas informações e os Indicadores de Localidade que podem ser identificados (ex: praças, viadutos, rodovias, faculdades etc.) não sugerindo também métodos para identificar a importância do relacionamento dessas entidades com a

Tabela 6. Principais Trabalhos Abrangendo a Recuperação de Indicadores de Localidade

Trabalho	Tipo de Corpora	Abordagem	Problemas
(Overell e Ruger, 2007)	Wikipédia	Identificação de Termos Sinônimos às Localidades a partir da Descrição dos Links Associados às Páginas das Localidades	Corpus com Baixa Atualização Pouca Cobertura de Indicadores Não Qualificação dos Indicadores
(Popescu et al. 2008)	Wikipédia	Extrai termos relacionados às cidades a partir da identificação dos nomes próprios encontrados nas descrições dos links na página das Localidades	Não Qualificação dos Indicadores Corpus com Baixa Atualização
(Buscaldi e Rosso, 2007)	Wordnet e Wikipédia	Utiliza um algoritmo de reconhecimento de entidades mencionadas para extrair Indicadores às cidades a partir da Wikipédia, usando a Wordnet para identificar o tipo da localização (se é estado ou país) das páginas.	Não Qualificação dos Indicadores Corpus com Baixa Atualização
(Rattenbury et al., 2007)	Flickr	Análise estatística para identificar tags semanticamente relacionadas a lugares.	Não Qualificação dos Indicadores Corpus com Baixa Atualização
(Borges et al., 2003)	Páginas da Web	Wrapper pré-treinado manualmente para identificar termos relacionados às cidades.	Necessidade de Corpus de Treino Não qualificação dos Indicadores
(Smith e Mann, 2003)	Notícias	Verifica *collocations* (Indicadores e localidades) utilizando um algoritmo de aprendizado de máquina semi-supervisionado (técnica de *bootstrapping*)	Não Qualificação dos Indicadores Necessidade de Corpus de Treino
(Ferres et al., 2004)	Notícias	Algoritmo de aprendizado de máquina visando identificar nomes de entidades correferenciadas (ex: Smith => John Smith) e pares de acrônimos (ex: USA => United States of America)	Necessidade de Corpus de Treino Pouca Cobertura de Indicadores Não Qualificação dos Indicadores
(Maynard et al., 2004)	Notícias	Utiliza aprendizado de máquina para recuperação de nomes de pessoas focado nas línguas Hindu, Chinesa e Arábica.	Dependência de Idioma Necessidade de Corpus de Treino Não Qualificação dos Indicadores
(Garbin e Mani, 2005)	Notícias	Busca identificar collocations (Indicadores e localidades) a partir de um algoritmo de aprendizado de máquina não-supervisionado treinado com ajuda de heurísticas.	Não Qualificação dos Indicadores (Distância da Análise Limitada a Apenas 20 Termos das Localidades) Dependência de Idioma
(Kozareva et al., 2006)	Notícias	Busca recuperar nomes de pessoas e lugares utilizando expressões de posicionamento não identificando sua relação com os topônimos.	Não Qualificação dos Indicadores Dependência de Idioma Pouca Cobertura de Indicadores

respectiva localidade. Um resumo dos trabalhos citados anteriormente e os problemas relacionados à extração de Indicadores é apresentado na tabela 6.

3 ABORDAGEM PROPOSTA PARA A IDENTIFICAÇÃO DE INDICADORES DE LOCALIDADE

Levando em conta os principais problemas abordados pela seção anterior e apresentados pelos trabalhos atuais para a identificação de Indicadores de Localidade, o presente trabalho apresenta as seguintes estratégias e diferenciais para resolução desses problemas:

- Utilização de Notícias para a etapa de treino, isto é, para descobrir Indicadores de Localidade, sem a necessidade de seleção manual ou anotação de corpora de treino. O trabalho não discute como capturar as notícias, apenas sugere o uso de textos jornalísticos sem a necessidade de anotação manual.
- Uso de uma larga janela para a verificação das relações entre termos (localidades e Indicadores), utilizando pra isso nomes próprios e levando em conta todo o texto analisado (dando mais importância para as relações que ocorrem na mesma frase, mas considerando também as que ocorrem em frases diferentes). A utilização de nomes próprios para a identificação de Indicadores torna a abordagem proposta mais extensível a outros idiomas, bastando para isso que essas entidades sejam representadas na linguagem alvo.
- Uso de uma fórmula de peso específica para determinar a importância das relações entre localidades e Indicadores.

A abordagem proposta por este trabalho busca considerar, portanto o caráter geográfico das informações jornalísticas e consequentemente a grande abrangência de Indicadores de Localidade incluídos nesse tipo de informação, visando com isso desenvolver um método para a extração e qualificação dinâmica desses Indicadores levando em conta suas características específicas.

Busca-se com isso utilizar essas entidades identificadas para a construção e atualização de *Gazetteers*, auxiliando assim na superação dos principais tipos de ambigüidades linguísticas associadas à Resolução de Topônimos. É importante notar que o trabalho considera pra isso o sentido padrão das localidades não abrangendo, portanto a metonímia envolvida.

Com a identificação e associação dos Indicadores (ex: sinônimos e outras entidades) às suas respectivas localidades, torna-se possível inferí-las em textos que não possuam explicitamente a localidade referenciada em seu conteúdo (auxiliando na resolução da ambiguidade da referência e de forma mais abrangente na ambiguidade indireta da referência) e ajudando também na desambiguação de localidades que são homônimas a outras (auxiliando na resolução da ambiguidade do referente).

O principal desafio é, portanto recuperar automaticamente entidades que possibilitem a identificação de forma precisa de determinada localidade a partir delas, já que muitos desses Indicadores podem aparecer associados a mais que uma cidade, prejudicando assim o georreferenciamento de textos onde estejam presentes.

Para garantir a verificação de Indicadores específicos às localidades, a abordagem proposta pelo trabalho busca, portanto, analisar a relação entre localidades (especificamente cidades) e Indicadores em textos jornalísticos e definir métodos próprios para a qualificação de sua relevância, levando em conta para isso as seguintes estratégias:

Análise de Co-Ocorrência: Como os Indicadores e as localidades são representados em grande parte dos idiomas (ex: português, inglês, espanhol) por meio de Nomes Próprios (NPs) priorizou-se para as análises a verificação da correlação entre eles. Outro benefício dos NPs é que a partir da sua verificação torna-se possível a identificação de Indicadores que são representados por meio de nomes compostos, os quais conforme verificou (Chaves e Santos, 2006) compreendem grande parte dessas entidades (75% dos termos associados à ontologia geográfica *Geo-Net-PT-01)*, além disso, como constatou (Dias, 2003) em textos jornalísticos os nomes próprios apresentam grande dinamismo (considerando tanto siglas como outras entidades) o que garante a atualização e relevância dos Indicadores de Localidade verificados.

Força Local da Relação: Para a qualificação das relações adotaram-se pesos especiais, levando em conta a distância dos NPs (relacionados a localidades e Indicadores) localizados na mesma frase (distância interna) assim como dos NPs localizados em frases diferentes (distância externa).

Força Global da Relação: Para a qualificação dos Indicadores outro tipo de peso considerou a força global da relação, ou seja, a soma dos pesos das distâncias locais da relação levando em conta todos os corpora analisados.

A análise da força local e global parte do princípio que Indicadores mais próximos às localidades em determinado texto tendem a ser relacionados a elas. A soma da distância para o corpus inteiro busca normalizar Indicadores que, embora próximos, possam não ser necessariamente relacionados às localidades. O objetivo é com isso garantir que os Indicadores mais relevantes às localidades possuam as relações com maior peso. Os detalhes sobre estas estratégias e também as fórmulas para a identificação desses pesos são ilustrados na seção seguinte.

3.1 Estratégia para a Identificação e Qualificação de Indicadores de Localidade

Visando possibilitar a construção e atualização de *Gazetteers* com os Indicadores de Localidade relacionados às cidades, torna-se necessária a adoção de técnicas específicas visando não somente recuperá-los, mas também determinar de alguma forma sua relevância e especificidade para determinada cidade.

Para garantir isso, a primeira providência compreende a recuperação de textos jornalísticos na Web. Para isso, a abordagem sugere a seleção aleatória de notícias em sites jornalísticos sem nenhum tipo de filtragem e sem indicar um *Website* ou uma técnica específica para essa seleção, apenas recomendando o uso de sites que publiquem notícias com certeza. A sugestão é usar sites conhecidos e de fontes confiáveis. As técnicas para seleção dos textos não são o foco dessa dissertação.

Na internet a estrutura das informações jornalísticas em diferentes fontes de notícia segue um formato comum (Scanlan, 2008), geralmente utilizando a técnica denominada de pirâmide invertida, onde segundo (Canavilhas, 2006) a redação do texto começa pelos dados mais importantes (respondendo a perguntas como: o quê, quem, onde, como, quando e por que) seguido de informações complementares organizadas em blocos decrescentes de interesse. Umas das vantagens disso é que a verificação de Indicadores de Localidade pode ser realizada independente da fonte utilizada.

O trabalho baseia-se, portanto, no pressuposto que a maioria das notícias possui algum tipo de indicador geográfico no texto, já que é importante referenciar, para os leitores potenciais, os fatos jornalísticos de acordo com sua posição geográfica. O trabalho baseia-se, com isso, na idéia que uma análise estatística de notícias pode ser utilizada para o enriquecimento de *Gazetteers* com relações entre localizações

geográficas (particularmente cidades) e indicadores geográficos (definidos como Indicadores de localidade), auxiliando assim no georreferenciamento e na recuperação de notícias de acordo com sua localização.

A abrangência e especificidade das entidades e cidades abrangidas irão depender do escopo geográfico da audiência que o site deseja alcançar, sendo que fontes com foco nacional são adequadas para a recuperação de Indicadores relacionados a capitais e outras cidades grandes. Dessa forma, para cidades pequenas torna-se necessário utilizar fontes regionais que priorizem notícias dessas localidades.

Para ilustrar as características da abordagem proposta foi desenvolvido, portanto, um módulo responsável por identificar Indicadores de Localidade a partir de notícias e qualificá-los de acordo com sua associação com as cidades, visando assim permitir o enriquecimento de *Gazetteers* úteis para o georreferenciameto. As principais etapas são ilustradas na figura 5.

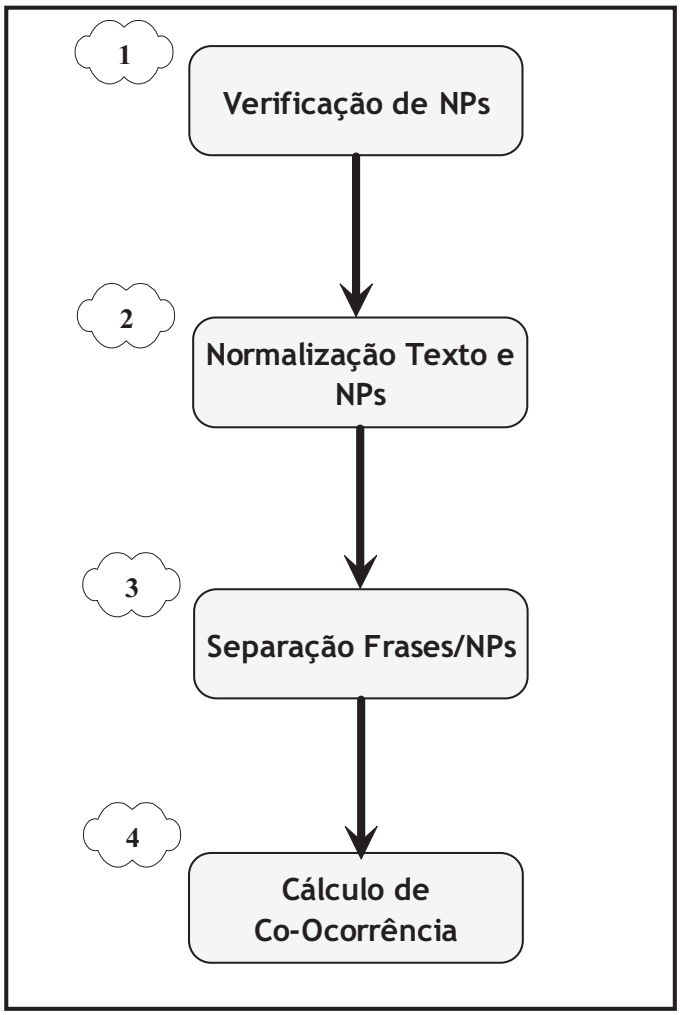

Figura 5. Etapas da Análise do Peso Local

A estrutura inicial do *Gazetteer* que irá ser enriquecido com os Indicadores de Localidade recuperados é composta por uma lista de nomes de cidades, este é o ponto inicial do processo de enriquecimento.

As relações entre nomes de cidades e Indicadores de Localidade são determinadas por um peso (um valor numérico representando a importância ou probabilidade da relação), o qual é calculado pela distância entre estes termos nos textos jornalísticos de uma coleção (o corpus de treino). A idéia é calcular o peso da relação em cada texto da coleção (peso local) e então utilizar toda a coleção para determinar o peso final (peso global).

Como os Indicadores de Localidade geralmente são representados através de Nomes Próprios (NPs) a primeira etapa para a análise do peso local compreende em analisar a notícia e extraí-los.

Uma característica que contribui na verificação é que os NPs geralmente são representados com letra maiúscula. Dessa forma a verificação baseou-se na recuperação de palavras em maiúsculo analisando também palavras na seqüência. A união ou separação das palavras foi decidida a partir da análise da ocorrência de padrões léxico/sintáticos utilizando pra isso expressões regulares específicas, por exemplo:

EXP1: ^([\"”"\(\)]?[A-ZÁÀÃÂÄÉÊÍÓÔÕÖÚÜÑ]{1})

EXP2 : ([\w-ÁÀÃÂÄÉÊÍÓÔÕÖÚÜÇÑáàãâäéêíóôõöúüçñ]+[,\.\):\"'%;“\?!]{1,2}$)

EXP3 : \b(da|das|de|do|dos)\b

Onde a EXP1 representa a expressão regular para identificação de palavras em maiúsculo, EXP2 por sua vez é utilizada para separação de termos (ex: quando uma palavra está em maiúsculo e possui uma vírgula após), já EXP3 representa a expressão para a identificação de caracteres de união, ou seja, se uma palavra em maiúsculo estiver seguida por algumas dessas preposições e após a preposição ocorrer outra palavra em maiúsculo estas serão unidas.

Devido à possibilidade da verificação de NPs recuperar termos que possam referenciar cidades em conjunto com outras palavras (ex: “Prefeitura de Porto Alegre”) na etapa 2 os NPs são normalizados visando identificar essas localidades (utilizando pra isso um *Gazetteer* contendo uma lista de cidades do Brasil), dessa forma levando em

conta o exemplo citado, o NP “Porto Alegre” também será adicionado à lista. Para facilitar sua identificação para a próxima etapa o texto também será atualizado com referência a localidade identificada, sendo esta incluída logo após o NP onde ela foi recuperada juntamente com um separador ex: “Prefeitura de Porto Alegre, Porto Alegre”.

Na etapa 3 todas as frases do texto são separadas mantendo referência somente aos NPs (sendo cada NP associado à posição de ocorrência na frase). Para a identificação dos NPs contidos em cada frase é necessário definir uma ordem para sua verificação visto que um texto pode conter NPs que aparecem separados e incluídos dentro de outros. Isso é muito comum, por exemplo, na referência a nomes de pessoas nos textos, sendo que a primeira referência relaciona-se ao nome completo e as demais somente ao sobrenome. Para identificar ambas as referências, os NPs são primeiramente ordenados em ordem decrescente de sua quantidade de palavras para então serem identificados nas frases.

Por fim, na etapa 4 a co-ocorrência entre os NPs relacionados aos nomes de Cidades e Indicadores de Localidade é analisada, sendo o peso de cada relação definido visando representar através de um valor numérico a importância ou probabilidade de cada relação. Para a identificação dos NPs relacionados a cidades foi utilizada uma base de dados simples contendo as cidades alvo dos experimentos desenvolvidos (seção 4).

Para a análise de co-ocorrência são considerados dois tipos de peso para as relações, os quais correspondem à análise da distância interna (para os Indicadores que co-ocorrem na mesma frase das localidades) e externa (para os Indicadores que co-ocorrem em frases diferentes). O peso para as relações internas (Ri) e externas (Re) entre cada cidade e indicador foi calculado através de fórmulas específicas levando em conta a respectiva distância (*d*) entre eles.

As fórmulas e os valores pré-definidos da constante *d* para ambos os pesos, e a corresponde opção pela escala decimal, foram definidos a partir de análises empíricas de notícias, onde se analisou, por exemplo, a distância média padrão entre cidades e Indicadores na mesma frase e em frases diferentes dos textos jornalísticos.

O objetivo é, com isso, dar mais relevância para relações próximas (dentro da mesma frase), mas considerar também as relações mais distantes (em frases diferentes do texto), conforme definido em (Gouvêa et al., 2008). Ou seja, se os Indicadores de

Localidade ocorrerem na mesma frase das cidades, o peso dessa relação será maior do que se ocorrer em frases diferentes (quanto mais frases de distância, menor será o peso da relação). Um valor mínimo é utilizado para definir o peso de relações muito distantes, as quais não influenciam diretamente nos resultados, sendo definidas então com um valor fixo a partir de uma distância (de palavras e frases) pré-definida. As fórmulas seguem, portanto, uma escala descendente, utilizando pra isso valores proporcionais.

Uma frase é um conjunto de termos ordenados separados por dois pontos finais. As fórmulas (3) e (4) apresentam o cálculo do peso interno $\boldsymbol{Pi_k}$ (para a frase $\boldsymbol{k}$) entre a cidade *c* e o Indicador de Localidade *r*.

$$\mathrm{Pi_k}(c,r) = \sum_{\substack{i=1 \\ d\leq 9}}^{n} \sum_{j=1}^{m} \frac{(10 - d_{c_i r_j})}{10} \tag{3}$$

$$\mathrm{Pi_k}(c,r) = \sum_{\substack{i=1 \\ d>9;d\leq 18}}^{n} \sum_{j=1}^{m} \frac{(19 - d_{c_i r_j})}{100} \tag{4}$$

Onde,
$\boldsymbol{d_{xy}}$ é o número de termos entre *x* e *y* na frase, sendo que *x* refere-se à cidade *c* e *y* relaciona-se ao indicador de localidade *r,*
k é a *k*-ésima frase no texto, onde os termos aparecem juntos,
i é o índice da *i*-ésima ocorrência do nome de *c* na frase**,**
j é o índice da *j*-ésima ocorrência do termo ***r*** na frase**,**
n é o número total de ocorrências de c na frase,
m é o número total de ocorrências de *r* na frase.

Para $\boldsymbol{d} > 18$, o peso ***Pi(c,r)*** tem o valor fixo de 0.01. O peso interno (*Pi*) deve ser calculado então para todos os pares de termos (cidades e Indicadores) que aparecem juntos dentro da frase.

A fórmula (5) por sua vez apresenta o cálculo do peso externo $\boldsymbol{Pe_t}$, para as relações entre determinada cidade *c* e o Indicador de Localidade *r* que ocorrem em diferentes frases do texto *t.*

$$\mathrm{Pe_t(c,r)} = \sum_{\substack{i=1 \\ d\leq 9}}^{n} \sum_{j=1}^{m} \frac{(10 - d_{c_i r_j})}{1000} \tag{5}$$

Onde,

d_{xy} é o número de frases entre x e y no texto t, sendo que x refere-se à cidade c e y ao Indicador de Localidade r.

i é o índice do i-ésima ocorrência do nome de c no texto t

j é o índice da j-ésima ocorrência do termo r no texto t,

n é o número total de ocorrências de c no texto t,

m é o número total de ocorrências de r no texto t,

t é o texto no qual o peso externo está sendo calculado.

Para $d > 9$, o peso $Pe(c,r)$ tem o valor fixo de 0.001. O peso externo (*Pe*) deve ser calculado para todos os pares de termos (nomes de cidades e Indicadores de Localidade) que aparecem no texto em frases diferentes.

O peso local (*Pl*) da relação entre *c* e *r* é calculado então por meio da soma entre o peso interno (*Pi*) e externo (*Pe*), para cada texto (um de cada vez), como expõe a fórmula (6). O peso local é calculado para a relação entre *c* e *r* em cada texto, devendo considerar a soma de todos os pesos internos (*Pi*) dessa relação, lembrando que o peso interno é calculado para cada frase.

$$\mathrm{Pl_t\ (c,r)} = \left[\sum_{k=1}^{n} Pi_{kt}(c,r) \right] + Pe_t(c,r) \tag{6}$$

Onde,

$Pl_t(c,r)$ é o peso local entre c e r para o t-ésimo texto na coleção,

t é o índice para todos os textos na coleção,

k é o índice de todas as frases no texto t onde c e r aparecem juntos,

n é o número total de frases dentro do texto t onde c e r aparecem juntos,

$Pi_k(c,r)$ é o peso interno entre c e r para cada frase k no texto t,

$Pe(c,r)$ é o peso externo entre c e r para o texto t.

Por exemplo, dado o seguinte texto:

"**Pelotas** se despede da **Fenadoce**. A 16ª edição da **Feira Nacional do Doce** (**Fenadoce**) terminou na noite deste domingo."

O Peso das Relações Internas (Pi) de "Pelotas" seria:

Fenadoce = 1ª Frase (d=4) → **0.6**

Já o peso das Relações Externas (Pe) de "Pelotas" seria:

Fenadoce = com 2ª Frase (d=1) → **0.009**
Feira Nacional do Doce = com 2ª Frase (d=1) → **0.009**

O Peso Final das Relações com a palavra "Pelotas", seria, portanto:

Fenadoce = 0.6+0.009 = **0.609**
Feira Nacional do Doce = **0.009**

A relação e seu peso final são incluídos então no *Gazetteer* seguindo a estrutura apresentada na figura 6.

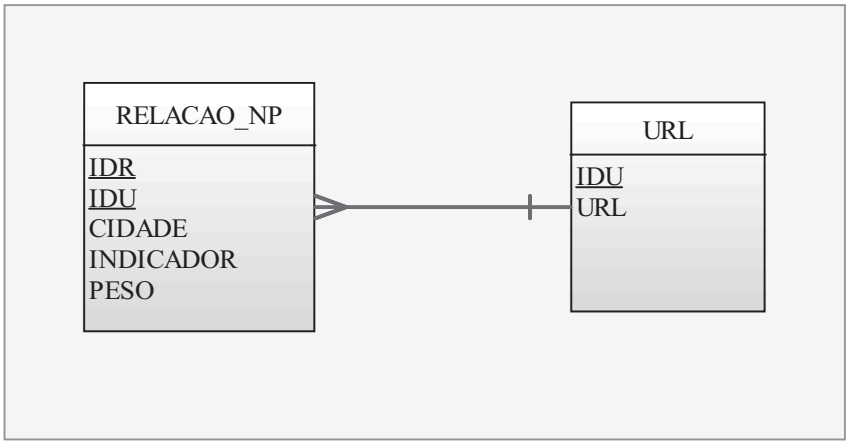

Figura 6. Estrutura do *Gazetteer* para as Relações

O *Gazetteer* é composto por um conjunto de nomes de cidades, cada uma com uma lista de Indicadores de Localidade (termos simples ou expressões). Entre a cidade e o indicador há um peso (o peso global) que representa a importância relativa da relação para a identificação dessa cidade em textos.

Visando aumentar a qualidade das relações foi necessário realizar algumas otimizações nas relações identificadas. Primeiramente, foram extraídas aquelas consideradas muito comuns ou pouco específicas às cidades. Como o peso local das relações não é por si só muito indicativo, foi necessário também identificar o peso

global da relação, ou seja, as relações mais associadas às cidades levando em conta todas as notícias analisadas.

As relações extraídas compreendem às associadas a outras cidades e estados, incluindo também NPs muito populares em minúsculo. Estes NPs muito populares em minúsculo, os quais não representam Indicadores de Localidade, são definidos por (Hu e Ge, 2007) como *stopwords geográficas*. Para identificá-los realizou-se uma análise estatística especial similar a (Amitay et al., 2004) onde se verificou a freqüência em maiúsculo e minúsculo de todos os NPs do corpus utilizado para a população do *Gazetteer*, sendo retirados todos os NPs que possuíam no mínimo três vezes mais freqüência em minúsculo do que maiúsculo, a qual foi a proporção melhor identificada (através de análises subjetivas) para a extração exclusiva de NPs não relacionados a Indicadores de Localidades. O anexo A apresenta uma lista de alguns dos nomes próprios removidos que apresentaram maior proporção de freqüência em minúsculo a partir dessa análise.

O peso local considera relações no interior de cada texto apenas considerando a qualidade das relações de forma isolada. Torna-se necessário, portanto analisar o peso obtido considerando toda a coleção de documentos, esse tipo de peso, definido como peso global é calculado conforme ilustra a fórmula (7).

$$\mathrm{Pg(c,r)} = \frac{\sum_{i=1}^{n} Pl_i(c,r)}{z} \qquad (7)$$

Onde,
$\boldsymbol{Pl_i}$ é o peso local entre *c* e *r*, considerando o texto *i,*
i é o índice dos textos da coleção de treino,
n é o número total de textos na coleção de treino,
z é o número total de cidades *c* que são relacionadas à *r* na coleção.

Esta formula busca normalizar o peso da relação (entre a cidade *c* e o indicador *r*) dividindo a soma de seus pesos locais pelo número de cidades que são relacionadas ao mesmo termo *r*, considerando que este pode ser relacionado a mais que uma cidade.

O objetivo é dar mais importância para termos que são relacionados a poucas cidades, atribuindo menor peso para termos mais gerais (que são relacionados a muitas cidades), os quais com a divisão irão receber um peso menor.

Por exemplo, levando em conta o exemplo apresentado anteriormente para o peso local, se outra notícia possuir a relação Pelotas → Fenadoce com peso local, por exemplo, de (0.6) e mais nenhuma cidade possuir a relação Cidade → Fenadoce, o peso global da relação será de (1.209/1=1.209). Essa divisão foi realizada visando à obtenção de relações mais específicas às cidades, desvalorizando assim relações que possuem foco mais geral e não conseguem identificá-las precisamente.

Para ilustrar o problema, o anexo B apresenta uma lista destas relações que são associadas a muitas cidades, as quais foram obtidas a partir dos experimentos desenvolvidos pelo trabalho. É interessante notar que algumas destas relações representam nomes de aeroportos do Brasil. Nesse caso, embora estas entidades apareçam associadas a várias cidades nas notícias, uma análise estatística de notícias demonstrou que a freqüência é mais alta para as cidades onde os aeroportos estão localizados, tornando possível, dessa forma, associá-las corretamente (dando maior relevância) com estas cidades a partir da utilização da abordagem proposta pelo trabalho.

A avaliação da qualidade da abordagem proposta e consequentemente das relações identificadas são ilustradas no capítulo seguinte.

Outra necessidade relacionada à identificação de Indicadores envolve a escolha do tipo de corpora utilizado para a verificação, para auxiliar nessa seleção o capítulo seguinte apresenta também experimentos utilizando corpora de notícias com características distintas.

4 EXPERIMENTOS

Levando em conta o objetivo do trabalho, o qual se relaciona a sugestão de uma abordagem para identificação de Indicadores de Localidade a partir de notícias, buscou-se testar a qualidade da abordagem proposta a partir da criação de *Gazetteers* com os Indicadores verificados. Os *Gazetteers* foram então avaliados pela habilidade de corretamente identificar às cidades relacionadas as notícias do corpus de teste. O objetivo é verificar com isso se as relações e os pesos estabelecidos são úteis para essa identificação.

Para melhor precisar a qualidade dessas relações foram testados métodos básicos tomados como referência para comparação (baseline), os quais não exigiram análises especiais, compreendendo apenas a organização de uma lista específica de ruas e bairros do Brasil relacionados às cidades analisadas buscando identificar assim se ruas e bairros são suficientes para a identificação das cidades nas notícias. Esse resultado foi considerado então como base para identificação da qualidade do método proposto.

Para a identificação das cidades através dos *Gazetteers* dois tipos de pesos foram adotados para qualificar as relações entre cidades e Indicadores de Localidade, os quais compreenderam a **freqüência simples** (sem peso especial apenas somando o número de relações encontradas associadas a cada cidade encontrada) e o **peso global** (somando o número de relações encontradas associadas a cada cidade a partir da fórmula do peso global desenvolvida pelo trabalho).

O objetivo é saber se as análises de co-ocorrência utilizando peso global tendem a trazer resultados mais precisos do que as de freqüência simples, ou seja, se o peso global ajuda na identificação única de cidades levando em conta aquelas identificadas com maior peso.

Para verificar as características envolvendo o tipo de corpora utilizado para identificação dos Indicadores, diferentes corpora de notícias foram analisados, sendo criados então *Gazetteers* específicos a partir dos Indicadores verificados com cada um, os quais compreenderam:

BASELINE – Apresenta apenas relações referentes a ruas e bairros associadas às cidades alvo do teste. Foi desenvolvido para testar a utilidade dessas entidades para a identificação das cidades e consequentemente compará-las com as relações extraídas

automaticamente. Para esse corpus não foi considerado peso especial para as relações (cada uma possuindo o peso padrão de (1)).

(C1) 3000 *Old* – Compreende a recuperação de relações a partir de um corpus com 3000 notícias do portal Folha Online[24] correspondentes a seção "Cotidiano", priorizando pra isso notícias antigas (entre os anos de 2001 e 2006).

(C2) 3000 *New* – Compreende a recuperação de relações a partir de 3000 notícias do portal Folha Online correspondentes a seção "Cotidiano", utilizando um corpus com notícias novas (anos de 2007 e 2008) visando identificar se o caráter temporal das notícias influencia na qualidade das relações.

(C3) 6000 – O mesmo tipo de análise da anterior só que aumentando o corpus (utilizando pra isso a união dos dois *Gazetteers* anteriores), visando identificar se a quantidade de notícias influencia a qualidade das relações.

(C4) 3000 – Utiliza um corpus com 3000 notícias recentes, também do Portal Folha Online com a diferença que as relações são recuperadas somente em notícias que possuem apenas 1 cidade no texto. As relações são apenas aquelas associadas às cidades incluídas em cada notícia, o objetivo é verificar se as notícias com apenas 1 cidade trazem relações mais específicas. A identificação das cidades foi realizada através de análise sintática, a qual levou em conta expressões de contexto populares percebidas nas notícias, como "cidade de", "em", referências a estados (ex: São Paulo (SP), São Paulo-SP, São Paulo – SP) e análise de expressões na frase da localidade referenciada (ex: cidade, município) caso não tenha identificado nenhum dos padrões anteriores.

A avaliação da qualidade da abordagem proposta, e consequentemente dos *Gazetteers* criados a partir dos Indicadores recuperados, é apresentada na seção seguinte.

4.1 Avaliação da Abordagem para a Identificação de Indicadores

Primeiramente, para avaliação foi necessário selecionar uma corpora de notícias úteis para identificar a qualidade da abordagem proposta. Para isso separou-se através de um módulo específico de análise sintática 1000 notícias do portal Folha Online, diferentes das utilizadas para a identificação dos Indicadores e relacionadas a 9 cidades

[24] http://www.folha.com.br/

do Brasil (Campo Grande, São Paulo, Belo Horizonte, Recife, Rio de Janeiro, Fortaleza, Porto Alegre, Florianópolis, Niterói), ou seja, que possuíam referência a somente uma dessas cidades no texto.

Após, cada notícia foi manualmente verificada, visando manter somente aquelas que possuíam pelo menos 1 Indicador de Localidade relacionado a alguma das cidades avaliadas, essa notícia foi posteriormente anotada então de acordo com essa localidade. Após essa verificação mantiveram-se 230 notícias para o processo de avaliação. Posteriormente esse processo foi realizado novamente e um novo corpus com mais 230 notícias foi também avaliado visando assim melhor legitimar os resultados.

Devido às características específicas do corpus para os experimentos e para a avaliação (ex: notícias com a presença de Indicadores de Localidade e publicadas em períodos diferentes), este trabalho utiliza corpora de treino e teste manualmente selecionados ao invés de utilizar corpora já existentes para tarefas envolvendo o georreferenciamento de textos, como é o caso, por exemplo, dos encontrados no GeoCLEF (conforme ilustra a seção 2.1.4) ou no HAREM[25]. O conjunto de notícias utilizado está, contudo disponíveis para outros autores[26].

Para a avaliação o objetivo é testar as relações identificadas nas notícias a partir da abordagem proposta visando responder com isso as questões levantadas na seção 1.1. Para isso realizou-se o georreferenciamento de notícias com o apoio dos *Gazetteers* criados com esses Indicadores, comparando os resultados com um *Gazetteer* criado relacionando apenas ruas e bairros das cidades avaliadas, visando analisar com isso a viabilidade desses *Gazetteers* para a identificação da cidade relacionada aos textos. Busca-se com isso verificar também a variedade de Indicadores recuperados por meio da comparação com essa lista de ruas e bairros.

O *Gazetteer* previamente criado com as relações para cada tipo de corpora foi, portanto limitado a somente relações associadas às 9 cidades alvo da avaliação.

Para cada tipo de teste o procedimento de avaliação compreendeu a identificação da cidade relacionada a cada uma das 230 notícias a partir das relações associadas às

---

[25] http://www.linguateca.pt/aval_conjunta/HAREM/

[26] http://gpsi.ucpel.tche.br/~cleber/mestrado/news_corpora/

cidades no *Gazetteer* para cada tipo de corpora e a posterior comparação da cidade identificada com a cidade correta anotada manualmente.

Para isso, primeiramente foi necessário identificar os Nomes Próprios nas notícias avaliadas e comparar eles com os Indicadores armazenados nos *Gazetteers* criados, lembrando que é possível que um indicador incluído no *Gazetteer* seja associado a mais que uma cidade. Para determinar a probabilidade de cada cidade ser associada à notícia analisada foi realizado então um raciocínio probabilístico, onde somente a cidade mais provável foi considerada relacionada à notícia.

O raciocínio probabilístico foi realizado conforme a figura 7.

1: **Para cada** Cidade *c* no Gazetteer **faça**
2: **Para cada** Termo *r* associado à Cidade *c* **faça**
3: **Se** *r* está contido na notícia *n* **então**
4: # Armazena o Peso *p* do Termo *r* como a Probabilidade da Cidade *c* ser Relacionada à Notícia
5: *Prob[c,n]= Prob[c,n]+p;*
6: **Fim_Se**
7: **Fim_Para**
8: **Fim_Para**
9: Escolha a cidade *c* com mais peso em *Prob[c,n]* como a cidade associada à notícia.

Figura 7. Algoritmo para o Georreferenciamento das Cidades a partir do *Gazetteer*

Após esse processo ser realizado para cada *Gazetteer* criado, a etapa final consiste então em comparar a cidade correta anotada manualmente relacionada à notícia com a cidade identificada a partir deste raciocínio probabilístico, sendo os resultados então analisados a partir de métricas específicas, conforme ilustra a seção seguinte.

4.2 Resultados

Para determinar os resultados para cada análise foram utilizadas as métricas relacionadas à Precisão (fórmula 8), Abrangência (fórmula 9) e F1 (fórmula 10), as quais para os experimentos testados representam:

$$\text{Prec} = \frac{\text{Número de Cidades Corretamente Identificadas}}{\text{Número de Cidades Identificadas}} \quad (8)$$

$$\text{Abr} = \frac{\text{Número de Cidades Corretamente Identificadas}}{\text{Número Total de Cidades Corretas}} \quad (9)$$

$$\text{F1} = \frac{2 * (\text{Prec} * \text{Abr})}{(\text{Prec} + \text{Abr})} \quad (10)$$

Como cada uma das 460 notícias anotadas manualmente é relacionada a apenas uma cidade, o número máximo de cidades corretas é, portanto, igual a 1. Foram realizadas duas avaliações distintas cada uma com 230 notícias de teste, os resultados relacionados a cada uma delas são apresentados abaixo.

Como forma de comparar se o número de relações associados às cidades no *Gazetteer* influencia na sua identificação, a quantidade de relações recuperadas para cada tipo de corpora é apresentada na tabela 7. Uma lista dos Indicadores de Localidade relacionados com mais peso nos *Gazetteers* construídos a partir da abordagem é apresentada no anexo C.

Tabela 7. Número de Relações no *Gazetteer* para cada Tipo de Corpora

	Tipo de Corpora				
	C1	**C2**	**C3**	**C4**	**Baseline**
Nº de Relações	6945	5783	9159	4757	119184

Para melhor identificação, os resultados para todos os corpora utilizados são apresentados de acordo com o tipo de peso adotado, sendo o peso global e freqüência simples para ambas as avaliações apresentados respectivamente nas tabelas 9 e 10. Para comparação, primeiramente são apresentados os resultados médios (considerando as duas avaliações) relacionados ao método baseline (tabela 8) o qual utilizou a freqüência simples como padrão, nas demais tabelas é apresentada a diferença desses resultados para os métodos propostos pelo trabalho (comparando com a métrica F1). Os resultados médios relacionados ao peso global e a freqüência simples são apresentados nas tabelas 11 e 12. As figuras 7 e 8 representam esses resultados graficamente.

Tabela 8. Resultado Médio Baseline (Ruas ∪ Bairros) para as Duas Avaliações

Gazetteer	Prec	Abr	F1
BASELINE	94%	25%	39,5%

Tabela 9. Resultado para cada *Gazetteer* utilizando Peso Global

Gazetteer	Avaliação 1				Avaliação 2			
	Prec	Abr	F1	Dif. Bas.	Prec	Abr	F1	Dif. Bas.
(C1) 3000 *old*	100%	36%	52,9%	+34,0%	100%	53%	69,3%	+75,4%
(C2) 3000 *new*	100%	39%	56,1%	+42,0%	100%	51%	67,5%	+70,9%
(C3) 6000	100%	40%	57,1%	+44,6%	100%	56%	71,8%	+81,8%
(C4) 3000 (AS)	99,3%	35%	51,8%	+31,1%	100%	41%	58,2%	+47,3%
MÉDIA PESO GLOBAL	99,8%	37,5%	54,5%	+37,9%	100%	50,3%	66,7%	+68,9%

Tabela 10. Resultado para cada *Gazetteer* utilizando Freqüência Simples

Gazetteer	Avaliação 1				Avaliação 2			
	Prec	Abr	F1	Dif. Bas.	Prec	Abr	F1	Dif. Bas.
(C1) 3000 *old*	87%	37%	51,9%	+31,4%	88%	59%	70,6%	+78,7%
(C2) 3000 *new*	91%	38%	53,6%	+35,7%	92%	54%	68,1%	+72,4%
(C3) 6000	91%	42%	57,5%	+45,6%	92%	61%	73,4%	+85,8%
(C4) 3000 (AS)	90%	36%	51,4%	+30,1%	90%	49%	63,5%	+60,8%
MÉDIA F. SIMPLES	89,8%	38,3%	53,6%	+35,7%	90,5%	55,8%	68,9%	+74,4%

Tabela 11. Resultado Final Médio para cada *Gazetteer* utilizando Peso Global

Gazetteer	Prec	Abr	F1
(C1) 3000 *old*	100%	44,5%	61,6%
(C2) 3000 *new*	100%	45%	62,1%
(C3) 6000	100%	48%	64,9%
(C4) 3000 (AS)	99,6%	38%	55,0%
MÉDIA PESO GLOBAL	99,9%	43,9%	60,9%

Tabela 12. Resultado Final Médio para cada *Gazetteer* utilizando Freqüência Simples

Gazetteer	Prec	Abr	F1
(C1) 3000 *old*	87,5%	48%	62,0%
(C2) 3000 *new*	91,5%	46%	61,2%
(C3) 6000	91,5%	51,5%	65,9%
(C4) 3000 (AS)	90%	42,5%	57,7%
MÉDIA F. SIMPLES	90,1%	47,0%	61,7%

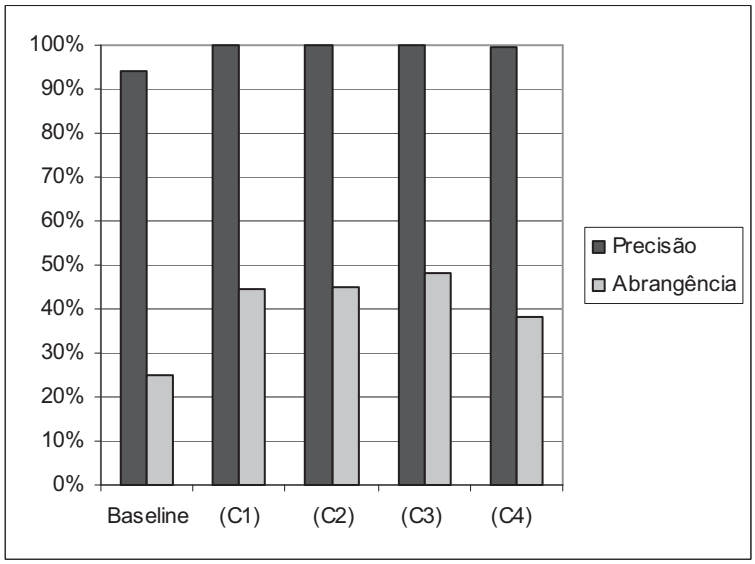

Figura 8. Resultados Médios Finais relacionados ao Peso Global

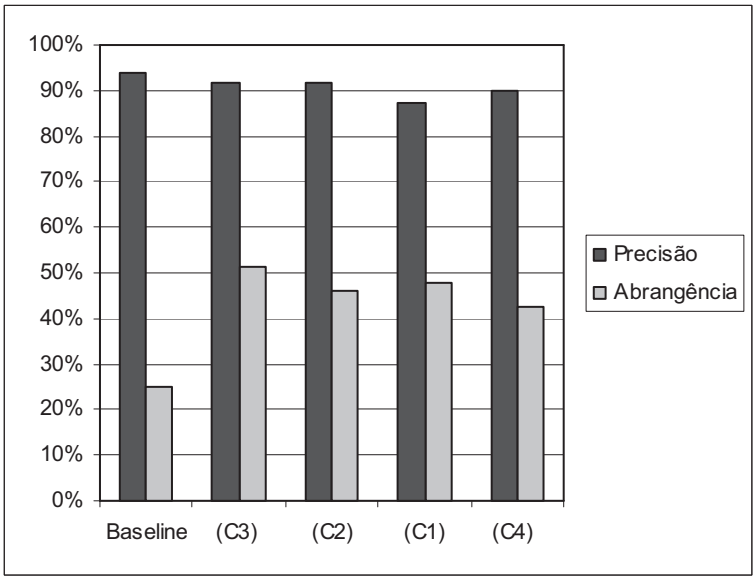

Figura 9. Resultados Finais Médios relacionados à Frequência Simples

4.3 Discussão

A partir da realização dos experimentos e da análise dos resultados é possível identificar a viabilidade e os desafios relacionados ao georreferenciamento de textos (particularmente a identificação de cidades) utilizando para isso os *Gazetteers* construídos e consequentemente os Indicadores de Localidade recuperados por meio da abordagem proposta. Esta seção apresenta e discute os principais resultados encontrados, buscando responder assim as questões levantadas na seção 1.1.

Com relação à variedade de Indicadores recuperados, uma análise da interseção de ruas e bairros com as 100 entidades recuperadas automaticamente que possuíam mais peso para cada cidade encontrou em média apenas 19 entidades iguais, retornando como complemento variados tipos de Indicadores relacionadas às cidades, dentre eles: nomes de pessoas alvos freqüentes de determinada notícia (ex: nome do prefeito e outras autoridades), nomes de hospitais, delegacias, museus, universidades, rodovias, parques, fundações etc. Demonstra-se, dessa forma, a utilidade dos métodos para a recuperação de Indicadores de Localidade com variedade além de ruas e bairros. O anexo C ajuda a ilustrar isso, apresentando uma lista dos Indicadores de Localidade relacionados com mais peso nos *Gazetteers* construídos pela abordagem.

Levando em conta os experimentos realizados envolvendo a identificação da cidade relacionada às notícias a partir dos *Gazetteers* criados automaticamente, pôde-se constatar a partir dos resultados médios encontrados (considerando ambas as avaliações realizadas) o seguinte: comparando os resultados utilizando os *Gazetteers* criados a partir da Análise de Co-Ocorrência com o *Gazetteer* Baseline pôde-se verificar que as técnicas para recuperação de Indicadores e as fórmulas para qualificar sua relação com as cidades (Peso Global e as demais otimizações) possibilitam a identificação da localidade com mais precisão e abrangência (ganhos respectivamente de 6% e 75% sobre a utilização de ruas e bairros).

Já com relação aos experimentos que buscaram identificar o tipo de corpora mais adequado para a recuperação de Indicadores de Localidade a partir da abordagem proposta (utilizando peso global), pôde-se verificar o seguinte: o período temporal das notícias demonstrou não influenciar significativamente na qualidade dos resultados, os corpus C1 e C2, embora de períodos diferentes, apresentaram precisão e abrangência semelhantes, com pequena vantagem para o *Gazetteer* com notícias mais recentes

(levando em conta a medida de abrangência). Embora resultados sugiram que as notícias mais recentes são melhores para a recuperação de Indicadores de Localidade relevantes, essa pequena diferença demonstra, no entanto que a recuperação não precisa ser realizada frequentemente (pode ser realizada apenas 1 vez por ano) já que o *Gazetteer* criado com notícias antigas (1 ano ou mais) apresentou performance relativamente parecida.

Com relação ao volume de notícias analisados e consequentemente ao número de relações obtidas, comparando os *Gazetteers* criados com diferentes quantidades (C3 vs C1 e C2), pode-se notar que o *Gazetteer* que utilizou maior quantidade de notícias (C3) obteve resultados ligeiramente superiores (5%) comparando com a média dos resultados encontrados por (C1 e C2) utilizando o peso global. Demonstra-se com isso que o tamanho do corpus utilizado é importante para a recuperação de Indicadores relevantes. Análises futuras poderão ser realizadas, contudo para identificar qual o tamanho de notícias suficiente para essa inferência.

Por sua vez, considerando o *Gazetteer* (C4), o qual fez uso de análise sintática para garantir a verificação de Indicadores de Localidade apenas em notícias com somente uma cidade no texto, pode-se notar que este tipo de análise não trouxe ganhos na performance (considerando tanto precisão como abrangência), apresentando o *Gazetteer* com o menor número de relações (desconsiderando o *Gazetteer* baseline). A idéia era obter ganhos relacionados à abrangência, contudo uma explicação provável para a baixa performance é que a quantidade de relações possivelmente influenciou nos resultados. De qualquer forma, como através de análise sintática torna-se possível garantir que apenas relações associadas a cidades sejam recuperadas, evitando assim a recuperação de Indicadores relacionados a estados com o mesmo nome de cidades (ex: Rio de Janeiro pode ter o sentido de cidade ou estado) espera-se que um aumento na quantidade de notícias analisadas garanta a melhoria desses resultados.

Os tipos de peso testados para qualificar os Indicadores de Localidade apresentaram resultados parecidos para a identificação das cidades com relação à média da precisão e abrangência para cada tipo de corpora. A diferença principal para os pesos relacionou-se a precisão, tendo as cidades encontradas com mais peso (utilizando o peso global) apresentado precisão 11% superior aos da freqüência simples. É interessante notar que as cidades encontradas corretamente apresentavam o maior peso na maioria

das vezes, ilustrando assim a importância da abordagem proposta e consequentemente da fórmula de peso estabelecida.

Considerando tudo isso, pode-se dizer que a utilidade dos métodos propostos depende da necessidade e especificidade dos resultados esperados. Com relação à abrangência ambos os tipos de peso apresentaram resultados semelhantes (com a freqüência simples apresentando ganhos médios de 7% para o peso global) e performance bem superiores aos métodos baseline, ilustrando assim o potencial da abordagem no auxílio aos processos de georreferenciamento, mais especificamente na identificação dos topônimos (no caso cidades) relacionados aos textos. Para resultados mais precisos os testes com peso global mostram-se particularmente úteis, podendo ser otimizados como se espera a partir do aperfeiçoamento das análises de co-ocorrência (ex: utilizando Análise Sintática para identificação das cidades, realizando a análise de similaridade entre NPs, etc.).

O principal desafio do ponto de vista da Resolução de Topônimos é, portanto obter resultados abrangentes e precisos, ou seja, garantir que somente as cidades realmente relacionadas aos textos sejam identificadas. A utilidade da abordagem proposta para identificação e qualificação de Indicadores de Localidade e consequentemente para a identificação das cidades nos textos pôde então ser constatada, podendo ser cada vez mais aperfeiçoada tendo como base as análises desenvolvidas e os trabalhos futuros sugeridos.

5 CONCLUSÃO

O presente trabalho trouxe como principal constatação a percepção do caráter geográfico das informações jornalísticas e consequentemente da grande variedade de Indicadores de Localidade presentes em seu conteúdo. Aproveitou-se então essa percepção para a sugestão de uma abordagem simples e extensível visando à recuperação abrangente dessas entidades viabilizando assim o desenvolvimento de *Gazetteers* com informações detalhadas e atualizadas sobre as localidades.

A partir dos experimentos realizados pôde-se perceber então que os *Gazetteers* criados a partir dos Indicadores de Localidade identificados podem aprimorar o processo de georreferenciamento (no caso desse trabalho à identificação das cidades), ajudando a superar assim ambiguidades importantes relacionadas à Resolução de Topônimos, as quais compreendem: a ambiguidade do referente (para cidades com o nome homônimo a outras), a ambiguidade da referência (para cidades que possuem nomes sinônimos), assim como para a superação como foi definido pelo trabalho, da ambiguidade indireta da referência (a qual se relaciona a textos que possuem apenas Indicadores de Localidade no conteúdo não possuíndo o nome de uma localidade explicitamente).

A abordagem proposta pelo trabalho pode ser utilizada consequentemente para a criação e atualização automática de *Gazetteers*, podendo abranger um grande número de localidades e possibilitando manter informações detalhadas sobre elas com esforço reduzido. Isso se torna possível devido ao fato das notícias possuírem uma grande variedade de Indicadores de Localidade, sendo mais acessíveis que bases de dados disponíveis que buscam armazenar e distribuir essas informações (ex: bases de ruas e bairros relacionadas às cidades), as quais são difíceis de ser encontradas, podendo inclusive não ser gratuitas. Outro problema é que essas bases, mesmo disponíveis, podem não suportar o caráter dinâmico dos Indicadores de Localidade, não disponibilizando assim informações atualizadas sobre estas entidades.

Contudo, é interessante lembrar que para viabilizar a identificação de Indicadores de Localidade a partir de notícias, as fontes analisadas devem abranger as cidades alvo da verificação. Devido à grande quantidade de sites jornalísticos na Web abrangendo cidades de variados tamanhos e compartilhando técnicas de redação jornalística semelhantes, esse problema pode ser com isso solucionado e mesmo cidades

pequenas podem ter seus Indicadores verificados pela abordagem sugerida, possibilitando consequentemente ter *Gazetteers* construídos cobrindo essas informações. Para garantir a confiabilidade dos Indicadores verificados a sugestão é capturar notícias de fontes bem conhecidas e com credibilidade, o trabalho, contudo não sugere técnicas específicas para captura de notícias, apenas indica a recuperação de informações em sites que publicam notícias com certeza.

Embora os experimentos tenham sido realizados utilizando notícias em português, outras linguagens podem ser usadas bastando pra isso que seja possível recuperar nomes próprios e que estes sejam utilizados para representar os Indicadores de Localidade na linguagem analisada. Para questões de otimização das relações outra necessidade é a existência de uma base de dados com o nome das cidades alvo da verificação. O restante da abordagem, incluindo as fórmulas mantém-se igual independente de idioma.

O trabalho analisou também o tipo de corpora de notícia mais adequado para a inferência dos Indicadores de Localidade e consequentemente para a construção automática de *Gazetteers*. A conclusão é que é importante manter *Gazetteers* atualizados ao longo do tempo, utilizando para isso notícias publicadas recentemente visando assim atualizar os Indicadores de Localidade e seus correspondentes pesos. Contudo, embora importante, essa atualização pode ser realizada apenas uma vez por ano, demandando menos esforço e custo para a manutenção dos *Gazetteers*. O volume de notícias demonstrou também influenciar na qualidade dos *Gazetteers* criados, ilustrando com isso que a performance da abordagem proposta cresce na medida em que se aumenta o número de notícias analisado. Para garantir que apenas relações associadas a cidades sejam recuperadas, a estratégia envolvendo a verificação de notícias com apenas uma cidade no texto pode ser aperfeiçoada, visando assim obter relações mais específicas.

Com relação à qualificação dos Indicadores, a partir da fórmula de peso desenvolvida pôde-se obter resultados com maior precisão que a não utilização de pesos especiais para as relações, demonstrando com isso a viabilidade da fórmula de peso proposta para a identificação das cidades nos textos.

Para complementar o presente trabalho alguns trabalhos futuros são sugeridos, conforme ilustra a próxima seção.

5.1 Trabalhos Futuros

Com relação a abordagem proposta, embora esta demonstra-se viável para a identificação de Indicadores de Localidade a partir de notícias e consequentemente para a Resolução de Topônimos (particularmente a identificação de cidades) a partir destes, alguns desafios ainda estão em aberto.

Um deles compreende a verificação em larga escala do conteúdo de textos jornalísticos. Como na Web as informações jornalísticas não estão disponíveis de uma maneira estruturada e facilmente acessível pelas máquinas, torna-se necessário então o desenvolvimento de mecanismos (habitualmente definidos como *wrappers*) para a recuperação dessas informações não estruturadas.

Embora os resultados tenham demonstrado que o volume de notícias influencia na qualidade dos *Gazetteers* e consequentemente dos Indicadores recuperados, não ficou claro ainda qual é o tamanho de corpora suficiente para essa verificação. Outro desafio é, portanto identificar com precisão qual o volume de notícias adequado que garanta ao mesmo tempo a construção de *Gazetteers* de qualidade e a performance da abordagem, não despendendo com isso recursos com análises desnecessárias. Análises utilizando variadas quantidades de notícias podem ser desenvolvidas visando possibilitar essa identificação.

Mesmo os resultados demonstrando que os Indicadores de Localidade mais relevantes tendem a possuir o maior peso, outro problema compreende em identificar e remover Indicadores não relacionados às localidades visando assim otimizar o *Gazetteer* e aprimorar os processos de georreferenciamento. Nesse sentido uma possibilidade é o estudo do peso definido para relações encontradas e a posterior definição de limiares para a retirada de Indicadores não relevantes às cidades.

Para a otimização do *Gazetteer* outra possibilidade compreende a definição de técnicas para verificação de similaridade dos nomes próprios recuperados visando com isso aperfeiçoar também os processos de georreferenciamento.

Dependendo do tipo de ambiguidade a ser solucionado torna-se possível também aplicar e complementar o presente trabalho de diferentes formas.

Com relação à *ambigüidade indireta da referência* e *ambiguidade do referente/referência*, são sugeridos os seguintes complementos:

- Utilização da página das cidades na Wikipédia para a verificação de relações co-ocorrentes às cidades visando compará-las com os resultados obtidos pelo trabalho atual.
- Utilização da abordagem proposta visando a inferência do contexto regional e nacional das notícias, conforme (Amitay et al., 2004), utilizando pra isso Indicadores relacionados ao estado e país das localidades.

Por sua vez a *ambiguidade da classe do referente* pode ser resolvida a partir das seguintes atividades:

- ❖ Verificação estatística das expressões de posicionamento mais comuns nas notícias do Brasil.
- ❖ Otimização da análise sintática já desenvolvida a partir da inclusão das expressões verificadas.

6 TRABALHOS PUBLICADOS

1. GOUVEA, Cleber, LOH, S., GARCIA, L. F. F., FONSECA, E. B., WENDT, I., da S. Discovering Location Indicators of Toponyms from News to Improve *Gazetteer*-Based Geo-Referencing. In: Simpósio Brasileiro de Geoinformática - GEOINFO, 2008, Rio de Janeiro, RJ. Anais do X Simpósio Brasileiro de Geoinformática - GEOINFO, 2008.

2. GOUVÊA, Cleber, LOH, S., GARCIA, L. F. F. Métodos para Seleção Automática de Tags para Descrição de Notícias. In: XIV Simpósio Brasileiro de Sistemas Multimídia e Web (Webmedia), 2008, Vila Velha, ES. XIV Simpósio Brasileiro de Sistemas Multimídia e Web (Webmedia), 2008.

3. GOUVÊA, Cleber, LOH, S., GARCIA, L. F. F. Tags Coletivas: Analisando Padrões de Uso para o Suporte a Sistemas de Folksonomia. In: Workshop de Aspectos da Interação Humano-Computador na Web Social, 2008, Porto Alegre. Workshop de Aspectos da Interação Humano-Computador na Web Social, 2008.

4. GOUVÊA, Cleber, LOH, S., GARCIA, L. F. F. Folksonomias: Identificação de Padrões na Seleção de Tags para Descrever Conteúdos. In: XIII Simpósio Brasileiro de Sistemas Multimídia e Web (Webmedia), 2007, Gramado. XIII Simpósio Brasileiro de Sistemas Multimídia e Web (Webmedia), 2007.

5. GOUVÊA, Cleber, LOH, S. Folksonomias: Identificação de Padrões na Seleção de Tags para Descrever Conteúdos. RESI. Revista Eletrônica de Sistemas de Informação, v. VI, p. 2, 2007.

7 REFERÊNCIAS BIBLIOGRÁFICAS

AMITAY E., HAR'EL N., SIVAN R., SOFFER A., Web-a-where: Geotagging Web Content. In Proceedings of the 27th SIGIR, pages 273–280, 2004.

ANDRADE L., M. J. SILVA, Indexing Structures for Geographic Web Retrieval. In Proceedings of the Conference on Mobile and Ubiquitous Systems (CSMU'06), Guimarães, Portugal, Junho 2006.

BAEZA-YATES R. A., CIARAMITA M., MIKA P., ZARAGOZA H., Towards Semantic Search. In Proceedings of Natural Language and Information Systems. n. 5039, p. 4-11, 2008.

BERNERS-LEE T.,. Isn't It Semantic?. 2006. Disponível em: (http://www.bcs.org/server.php?show=ConWebDoc.3337)

BERNERS-LEE, T., J. HENDLER, AND O. LASSILA,. The Semantic Web: a new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. The Scientific American, 284**:** 34-43. 2001.

BORGES K. A. V. Uso de uma Ontologia de Lugar Urbano para Reconhecimento e Extração de Evidências Geo-espaciais na Web. Belo Horizonte: Instituto de Ciências Exatas, 195p. 2006. (Tese, Doutorado em Ciência da Computação)

BORGES, K. A. V., LAENDER, A. H. F., MEDEIROS, C. B., SILVA, A. S., DAVIS JR., C. A., 2003, The Web as a data source for spatial databases, V Simpósio Brasileiro de Geoinformática - GeoInfo, Campos do Jordão (SP), 2003.

BUCHER B., CLOUGH P., JOHO H., PURVES P., SYED A., Geographic IR systems: requirements and evaluation. In Proceedings of the 22nd International Cartographic Conference, 2005.

BUSCALDI D.,ROSSO P. A Comparison of Methods for the Automatic Identification of Locations in Wikipedia. In Proceedings of the 2007 Workshop On Geographic Information Retrieval (GIR 2007), Lisboa, Portugal. 2007.

BUYUKKOKTEN O., CHO J., GARCIA-MOLINA H., GRAVANO L., SHIVAKUMAR N., Exploiting goegraphical location in-formation of Web pages. In Proceedings of the ACM SIGMOD Workshop on the Web and Databases, WebDB, 1999.

CAI G., GeoVSM: An Integrated Retrieval Model for Geographic Information, in: M. Egenhofer and D. Mark, Eds., Geographic Information Science—Second International Conference, GIScience 2002, Boulder, CO, vol. 2489, Lecture Notes in Computer Science, Springer, pp. 70-85. 2002.

CANAVILHAS, J. Webjornalismo: Da pirâmide invertida à pirâmide deitada. Universidade da Beira Interior, Portugal. 2006.

CHAVES, M.S., SANTOS, D., What kinds of geographical information are there in the portuguese Web? In Proc. of the 7th Workshop on Computational Processing of Written

and Spoken Portuguese, PROPOR 2006. Volume 3960 of Lecture Notes in Computer Science., Itatiaia, Rio de Janeiro, Brazil. 2006.

CLOUGH P., SANDERSON M., JOHO H., Extraction of semantic annotations from textual Web pages. Technical report, 2004.

COVER T., THOMAS J. Elements of Information Theory. Wiley, 1st edition, 1991.

DAVIS JR., C. A, FONSECA, F. T., BORGES, K. A. V. A Flexible Addressing System for Approximate Geocoding. In *Proceedings of the V Brazilian Symposium on Geoinformatics*. Campos do Jordão, SP, Brasil, 2003.

DELBONI, T.M., BORGES, K.A.V., LAENDER, A.H.F., DAVIS JR., C.A. Semantic Expansion of Geographic Web Queries Based on Natural Language Positioning Expressions. Transactions in GIS, 11(3): 377-397, 2007.

DIAS, M. P. L. Nomes próprios e siglas no texto jornalístico em séculos e anos consecutivos. Polifonia, Cuiabá, v. I, n. 07, p. 91-225, 2003.

EGENHOFER, M. J., MARK, D. M., Naïve Geography, Frank A. U., Kuhn, W.(Eds.): Spatial Information Theory: A theoretical foundation for GIS. Berlin, Springer Verlag (LNCS 998) pp. 1-15. 1995.

EGENHOFER, M. J., Toward the Semantic Geospatial Web. National Center for Geographic Information and Analysis. Department of Spatial Information Science and Engineering. Department of Computer Science. Main. 2002.

EGENHOFER, M. J., FRANZOSA, R. D. Point-set topological spatial relations. International Journal of Geographical Information Systems, London, v.5, n.2, p.161-174, 1991.

FERRES, D., MASSOT, M., PADRO, M., RODRIGUEZ, H. AND TURMO, J. Automatic Building Gazetteers of Co-referring Named Entities. Proceedings of the 4th International Conference on Languages Resources and Evaluation (LREC). Lisbon, Portugal. 2004.

FONSECA, F., EGENHOFER, M., BORGES K. A. V., Ontologias e Interoperabilidade Semântica entre SIGs. In: II Workshop Brasileiro em Geoinformática - GeoInfo2000, São Paulo, 2000.

GALE W., CHURCH K., YAROWSKY D. One sense per discourse. In Proceedings of the Fourth DARPA Speech and Natural Language Workshop, pages 233 237. Defense Advanced Research Projects Agency, Morgan Kaufmann, San Mateo, CA. 1992.

GARBIN E., MANI I., Disambiguating toponyms in news. In Proc. Human Language Technology Conference (HLT-EMNLP'05), pages 363–370, Vancouver, BC, October 2005.

GEY F.,LARSON R.,SANDERSON M., JOHO H. E P. CLOUGH. GeoCLEF: the CLEF 2005 cross-language geographic information retrieval track. In Working Notes

for the CLEF 2005 Workshop, 2005.

GILES, C. L., BOLLACKER, K., LAWRENCE, S. CiteSeer: An automatic citation indexing system. In: Digital Libraries 98 - The Third ACM Conference on Digital Libraries, Pittsburgh, PA. ACM Press. 1998.

GOUVÊA, Cleber, LOH, S., GARCIA, L. F. F., FONSECA, E. B., WENDT, I. da S. Discovering Location Indicators of Toponyms from News to Improve *Gazetteer*-Based Geo-Referencing. In: Simpósio Brasileiro de Geoinformática - GEOINFO, 2008, Rio de Janeiro, RJ. Anais do X Simpósio Brasileiro de Geoinformática - GEOINFO, 2008.

GRAUPMANN J., SCHENKEL R., GeoSphereSearch: Context-aware geographic Web search. In 3rd Workshop on Geographic Information Retrieval (GIR 2006), Seattle, WA, USA, 2006.

HILL, L., Core elements of digital Gazetteers: Placenames, categories and footprints Borbinha, J. and Baker, T. (Eds.) Research and Advanced Technology for Digital Libraries, proceedings. 2000.

HU Y., GE, L., A Supervised Machine Learning Approach to Toponym Disambiguation. In: Scharl, A., Tochtermann, K. (eds.): The Geospatial Web: How Geobrowsers, Social Software and the Web 2.0 are Shaping the Network Society. Springer, London, 3-14, 2007.

IDE, N., VÉRONIS, J., Introduction to the special issue on word sense disambiguation: the state ofthe art. Computational Linguistics, 24(1). March, 1998.

JONES, C. B., H. ALANI, TUDHOPE D., Geographical Information Retrieval with Ontologies of Place. In. Proceedings of COSIT 2001 International Conference on Spatial Information Theory, Morro Bay, CA, USA, September 19-23. D. R. Montello (eds.), Springer-Verlag: 322-335. 2001.

JONES C. B.,PURVES R.,RUAS,SANDERSON M.,SESTER M.,VAN KREVELD M. E WEIBEL R. Spatial information retrieval and geographical ontologies: An overview of the SPIRIT project. In Proceedings of SIGIR-02, the 25th Conference on Research and Development in Information Retrieval, 2002.

JONES C. B., PURVES R. S., Geographical information retrieval, International Journal of Geographical Information Science, v.22 n.3, p.219-228, 2008.

JONES. C. B., Geographical Information Retrieval. GeoInfo 2006. Disponível em: http://www.geoinfo.info/geoinfo2006/presentation/Christopher_Jones.ppt

JONES, C. B., PURVES, R. S., CLOUGH, P. D. AND JOHO, H. Modelling Vague Places with Knowledge from the Web.. International Journal of Geographical Information Science. 2007.

KOZAREVA, Z. Bootstrapping Named Entity Recognition with Automatically Generated Gazetteer Lists, In Proceedings of EACL student session (EACL), Trento, Italy. 2006.

LANA-SERRANO S., VILLENA-ROMÁN J., GOÑI-MENOYO J. M., Miracle at GeoCLEF Query Parsing 2007: Extraction and Classification of Geographical Information", 8th CLEF Workshop, September 2007.

LARSON R., FRONTIERA P., Geographic information retrieval (GIR): searching where and what, Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, July 25-29, Sheffield, United Kingdom. 2004.

LARSON, R.R., Geographical information retrieval and spatial browsing. Geographical Information Systems and Libraries: Patrons, Maps, and Spatial Information. pp. 81-124. 1996.

LEIDNER J. L., Toponym Resolution in Text - Annotation, Evaluation and Applications of Spatial Grounding of Place Names. Edinburgh: Institute for Communicating and Collaborative Systems, 287p. 2007. (Tese, Doutorado em Ciências da Comunicação)

LEIDNER J., Towards a reference corpus for automatic toponym resolution evaluation. In Workshop on Geographic Information Retrieval held at the 27th Annual International ACM SIGIR Conference (SIGIR'04), Sheffield, UK, 2004.

LEVELING, J., HARTRUMPF, S. AND VEIEL, D. Using semantic networks for geographic information retrieval. In Peters C., Gey F. C., Gonzalo J., Jones G. J. F., Kluck M., Magnini B., Muller H., de Rijke M., editors, Accessing Multilingual Information Repositories: *6th Workshop of the Cross-Language Evaluation Forum, CLEF*, Vienna, Austria, LNCS. Springer, Berlin. 2006a.

LEVELING J. AND HARTRUMPF S., On metonymy recognition for gir. In Proceedings of GIR-2006: 3rd Workshop on Geographical Information Retrieval. Seattle, USA. 2006b.

LEVELING, J. AND HARTRUMPF S. University of Hagen at GeoCLEF: Exploring location indicators for geographic information retrieval. In *Results of the Cross-Language System Evaluation Campaign, Working Notes for the CLEF Workshop*. Budapest, Hungary. 2007.

LI H., SRIHARI R. K., NIU C., LI W., InfoXtract location normalization: a hybrid approach to geographical references in information extraction. In Workshop on the Analysis of Geographic References, Edmonton, Canada, NAACL-HLT. May 2003.

MANNING C. D., SCHUTZE H. Foundations of Statistical Natural Language Processing. The MIT Press, Cambridge, Massachusetts. 1999.

MARTINS B., CHAVES M., SILVA M. J., "O sistema CaGE para Reconhecimento de referências geográficas em textos na língua portuguesa". Encontro do HAREM. Porto, Portugal, 2006a.

MARTINS B., SILVA M. J., FREITAS S. E AFONSO A. P., Handling Locations in Search Engine Queries, GIR, the Workshop on Geographic Information Retrieval at

SIGIR, 2006b.

MARTINS B., SILVA M. J.,CHAVES M. S., Challenges and resources for evaluating geographical ir. InProceedings of the 2005 Workshop On Geographic Information Retrieval (GIR 2005), Bremen, Germany,November 2005.

MAYNARD, D., BONTCHEVA, K. AND CUNNINGHAM, H. Automatic Language-Independent Induction of *Gazetteer* Lists. In Proceedings of 4th Language Resources and Evaluation Conference (LREC). 2004.

MCCURLEY, S.K. Geospatial mapping and navigation of the Web. In Proceedings of the Tenth International WWW Conference, Hong Kong, 1-5 May, 221-229. 2001.

MCDONALD D. Internal and external evidence in the identification and semantic categorization of proper names. In B. Boguraev and J. Pustejovsky, editors, Corpus Processing for Lexical Aquisition, chapter 2, pages 21–39. MIT Press, Cambridge, MA. 1996.

OVERELL, S. E. AND RUGER, S. Geographic Co-occurrence as a Tool for GIR. In Proceedings of the Workshop On Geographic Information Retrieval (GIR), Lisboa, Portugal. 2007.

PAGE, L.; BRIN, S.; MOTWANI, R.; WINOGRAD, T. The PageRank Citation Ranking: Bringing Order to the Web. 1999. Disponível em: <http://dbpubs.stanford.edu/pub/1999-66>.

PAPADIAS D., THEODORIDIS Y., SELLIS T., EGENHOFER M., Topological relations in the world of minimum bounding rectangles: a study with R-trees. Proceedings of the ACM SIGMOD Conference, San Jose, California, 1995.

PERRY M., SHETH A., ARPINAR I. B., Geospatial and Temporal Semantic Analytics. In Encylopedia of Geoinformatics, Hassan A. Karimi (Ed), Idea-Group Inc., 2007.

POPESCU, A., GREFENSTETTE, G. AND MOËLLIC, P. A. Gazetiki: automatic creation of a geographical *Gazetteer*. In Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries table of contents, Pittsburgh PA, PA, USA. 2008.

PYALLING, A., MASLOV, M., AND BRASLAVSKI, P., Automatic geotagging of Russian Web sites. In Proceedings of the 15th International Conference on World Wide Web (Edinburgh, Scotland, May 23 - 26). WWW '06. ACM Press, New York, NY, 965-966. 2006.

RATTENBURY, T., GOOD, N., NAAMAN M. Towards Automatic Extraction of Event and Place Semantics from Flickr Tags. In Proc. of SIGIR, Amsterdam, Netherlands. 2007.

SALTON. G., Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer. Addison-Wesley, Reading, Pennsylvania, 1989.

SANTOS, D., CHAVES, M.S., The place of place in geographical IR. In: Proc. of the 3rd Workshop on Geographic Information Retrieval, SIGIR'06, Seattle, USA 5–8. 2006.

SCANLAN C. Writing from the Top Down: Pros and Cons of the Inverted Pyramid. Disponível em: <http://www.poynter.org/column.asp?id=52&aid=38693>

SILVA M. J., MARTINS B., ANDRADE L., Indexing and ranking in Geo-IR systems. In Proceedings of the Workshop on Geographic Information Retrieval at CIKM 2005, 2005.

SMITH, D., MANN, G. Bootstrapping toponym classifiers. In Workshop on the Analysis of Geographic References, NAACL-HLT, Edmonton, Canada. 2003.

SOUZA, L. A. LOCUS: Um Sistema de Localização Geográfica Através de Referências Espaciais Indiretas. Programa de Pós-Graduação em Ciência da Computação, UFMG. Belo Horizonte. 67p. 2005. (Tese, Mestrado em Ciência da Computação)

SOUZA, L. A., DELBONI, T. M., BORGES, K. A. V., DAVIS JR., C. A., LAENDER, A. H. F. Locus: um Localizador Espacial Urbano. In Proceedings of the VI Brazilian Symposium on Geoinformatics, Campos do Jordão, SP, Brasil, 2004.

SPECIA L.,NUNES M. G. V., Desambiguação Lexical Automática de Sentido: Um Panorama. São Carlos. Núcleo Interinstitucional de Lingüística Computacional. 122p. 2004.

TWAROCH F. A., JONES C. B., ABDELMOTY A. I. A Comparison of Methods for the Automatic Identification of Locations in Wikipedia. In Proceedings of the first international workshop on Location and the Web. Beijing, China. 2008.

WANG C., XIE X., WANG L., LU Y., MA W., Web resource geographic location classification and detection. Special interest tracks and posters of the 14th international conference on World Wide Web - Chiba, Japan, 2005.

WANG, C., LI Z., XIE, X., WANG X. AND MA, W.-Y., Indexing implicit locations for geographical information retrieval. in Workshop on Geographic Information Retrieval, SIGIR. Seattle, USA. 2006.

ZHANG Q., XIE X., WANG L., YUE L., MA W., Detecting Geographical Serving Area of Web Resources, The 3rd International Workshop on Geographic Information Retrieval (GIR 2006), Seattle, USA, Aug. 2006.

ANEXO A - Exemplo de Nomes Próprios Removidos

causa
prisão
total
reportagem
têm
bairro
assessoria
responsável
conseguiu
acordo
processo
risco
valor
zona sul
feridos
principais
ocorrido
receber
maior
pagar
militares
avenida
entrou
entrar
problemas
menos
lentidão
encontrado
demais
morto
irá
caiu
passado
localizado
tentar
pessoas
medida
maioria
cabeça
teria
bem
casas
linha
volta
crime
resultado
pediu
trecho
tráfico de drogas
atingido
motivo
imprensa
avenidas
destino
aeronave
exemplo
tentou
vias
zona
dezenas

ANEXO B - Exemplo De Relações Associadas a muitas Cidades

estado
lula
justiça
pm
pf
polícia rodoviária federal
confins
polícia
tam
polícia civil
pms
defesa civil
prf
brasil
polícia federal
conac
galeão
uti
congonhas
santos dumont
anac
infraero
ministério público
jobim
iml
bra
guarda municipal
tom jobim
instituto médico legal
ibge
oceanair
união
estatística
sul
airbus-a320 da tam
conselho de aviação civil
ministério público federal
trf
nordeste
fgv
incêndio
tribunal regional federal
dp
defesa
polícia militar
corregedoria da polícia civil
batalhão
oms
legacy
pontifícia universidade católica
direitos humanos
corpo de bombeiros
oliveira
argentina
bombeiros
br-116
organização mundial de saúde
pt
boeing
sociedade brasileira de cirurgia plástica

ANEXO C – Exemplo de Indicadores de Localidade Recuperados com mais Peso

CIDADE = Rio de Janeiro

jornalista tim lopes
companhia de águas
marechal hermes
cesar maia
alexandre neto
josé milton rodrigues
segurança do rio de janeiro
cedae
parada de lucas
fernando villas-boas filho
cosme velho
superintendente da pf
dona marta
penha circular
polícia civil do rio de janeiro
caju
companhia docas do rio de janeiro
cidade de deus
cemitério são joão batista
senador camará
ronaldo cezar coelho
cdrj
vara da comarca de rio bonito
wagner victer
universidade estadual do rio de janeiro
defesa civil do rio de janeiro
bilhete do rio de janeiro
pilar de almeida oster
anthony garotinho
rogério lustosa
hospital da lagoa
vigário geral
instituto criminalística do rio de janeiro
carlos éboli
vara cível do rio de janeiro
arquidiocese do rio de janeiro
mercado são sebastião
aeroporto santos dumont
rodrigo de freitas
prefeitura do rio de janeiro
tom jobim
barra da tijuca
del castilho
transportes de passageiros do estado
aterro do flamengo
engenho novo
ouro
operação da pm
antonio carlos jobim
linha amarela
dia internacional do trabalho
instituto de psicologia da ufrj
polícia militar do rio de janeiro
rosinha matheus
maradona
anestor magalhães
linha vermelha
santos dumont

CIDADE = São Paulo

gilberto kassab
vila olímpia
museu de arte de são paulo
arquidiocese de são paulo
vila leopoldina
fiesp
rodrigo césar rebello pinho
marginal pinheiros
professor luís inácio de anhaia melo
universidade de são paulo
sindicato dos médicos de são paulo
vara criminal da justiça federal
secretaria de saúde da prefeitura
edson aparecido brandão
simesp
assembléia legislativa de são paulo
itaim paulista
órgão especial do colégio de procuradores
jardim sinhá
masp
congonhas
metrô de são paulo
mike tyson
vara federal de são paulo
galeria pagé
armazéns gerais de são paulo
rua dos pinheiros
punks
mooca
sindicato dos motoristas
bernardino de sena
vila nova curuçá
luiz carlos berrini
universidade federal de são paulo
arcebispo de são paulo
tj de são paulo
água branca
vila maria
fernanda teixeira taubemblatt
parque villa lobos
vara criminal de são paulo
ceagesp
cambuci
assis chateaubriand
justiça do estado de são paulo
sambódromo do anhembi
zoológico de são paulo
justiça de são paulo
barão de ladário
sifuspesp
governo de são paulo
são paulo transportes
companhia de entrepostos
agência usp
defesa da cidadania do estado
hospital penitenciário do carandiru
unifesp
tribunal do júri de são paulo