Universidade de Lisboa
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE INFORMÁTICA

Uma Metodologia para Construção de Geo-Ontologias

Marcirio Silveira Chaves

DOUTORAMENTO EM INFORMÁTICA
ESPECIALIDADE ENGENHARIA INFORMÁTICA

2009

Universidade de Lisboa
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE INFORMÁTICA

Uma Metodologia para Construção de Geo-Ontologias

Marcirio Silveira Chaves

DOUTORAMENTO EM INFORMÁTICA
ESPECIALIDADE ENGENHARIA INFORMÁTICA

2009

Tese orientada pelo Prof. Doutor Mário Jorge Costa Gaspar da Silva
e pela Doutora Diana Maria de Sousa Marques Pinto dos Santos

Abstract

Existing textual information is knowledge-rich and grows exponentially in time. To be useful to systems performing some type of intelligent processing, this knowledge needs to be automatically handled. Ontologies have emerged as the preferred unifying semantic representation of knowledge in texts and databases. However, the construction of ontologies requires the level-headed use of techniques of Natural Language Processing (NLP).

This thesis proposes an information extraction and integration methodology for building geo-ontologies from knowledge extracted from multiple sources. Initially, we collect geographic data from selected information sources which are integrated into a geographic knowledge base. This knowledge is then enriched from texts aided by a geographic information extraction and integration system. The results show that the geo-ontologies generated using the proposed methodology are useful to several applications. The geographic information extraction system reaches results at the same level of state of art systems in the task of recognition of places and its relationships in Portuguese.

Keywords: Methodology, knowledge representation, ontology, information extraction, recognition of geographic relationships.

Resumo

A informação textual existente é rica em conhecimento e cresce no tempo em ritmo exponencial. Esse conhecimento precisa poder ser manipulado automaticamente, de modo a tornar-se útil para sistemas que realizem algum tipo de processamento inteligente. As ontologias têm surgido como uma representação semântica unificada do conhecimento disponível tanto em textos quanto em bases de dados. No entanto, a construção de ontologias requer o uso criterioso de técnicas de Processamento de Linguagem Natural (PLN).

Esta tese propõe uma metodologia de extracção e integração de informação para construção de geo-ontologias a partir de conhecimento extraído de múltiplas fontes. Inicialmente, coletamos dados geográficos de diversas fontes de informação que integramos em uma base de conhecimento geográfico. Esse conhecimento é posteriormente enriquecido a partir de textos com o auxílio de um sistema de extração e integração de informação geográfica. Os resultados mostram a utilidade das geo-ontologias geradas a partir da metodologia proposta e colocam o sistema de extração e integração de informação geográfica ao nível do estado da arte em reconhecimento de locais e seus relacionamentos em português.

Palavras Chave: Metodologia, representação de conhecimento, ontologia, extração de informação, reconhecimento de relacionamentos geográficos.

Agradecimentos

Muitos são os envolvidos nesta tese. Agradeço:

- À minha pérola Cristiane Drebes Pedron pelo amor e apoio incondicional e, até mesmo, transatlântico.
- Aos meus orientadores Mário J. Silva e Diana Santos pelo profissionalismo, dedicação, diligência e entusiasmo com que conduziram essa jornada. "Se eu vi mais longe foi porque me apoiei nos ombros de gigantes" - Isaac Newton.
- Ao Bruno Martins, Leonardo Andrade, Nuno Cardoso, Sérgio Freitas e David Cruz pelas críticas e troca de idéias geo-referenciadas. Obrigado por terem tornado úteis as geo-ontologias desenvolvidas nessa tese.
- Ao Francisco Couto pela amizade e disponibilidade cotidiana no departamento. Ao Paulo Carreira, pela amizade. A Catarina Rodrigues pela introdução aos conceitos "exóticos" da geografia física. Aos demais membros do XLDB que, direta ou indiretamente, participaram desse trabalho.
- Aos colegas do LaSIGE pelas discussões técnicas e filosóficas. Em especial àqueles muito à frente: André Santos, João Antunes, Mônica Dixit, Giuliana Santos, Simão Fontes e Wagner Dantas. Crescemos bastante com o intercâmbio cultural e científico entre Portugal, Brasil e Rio Grande do Sul.
- Aos colegas da Linguateca e ao ex-colega Luis Sarmento pelas discussões técnicas e científicas. O modelo de trabalho na Linguateca foi um exemplo e os simpósios doutorais foram inesquecíveis.
- Às minhas ex-orientadoras Renata Vieira e Vera Lúcia Strube de Lima que foram as primeiras a acreditar e investir tempo e dedicação em mim.

0. AGRADECIMENTOS

- À minha família, em especial aos meus pais Alcino de Oliveira Chaves e Reinalda Silveira Chaves por fornecerem a base que me permitiu fazer escolhas ao longo da vida. Aos meus sogros Antônio Romeu Ilha Pedron e Wanda Drebes Pedron. Minha sogra, obrigado por teres vindo presenciar uma realidade diferente.

- A Cássia Trojahn dos Santos, Mirian Oliveira e Greise Korte.

- À Fundação para a Computação Científica Nacional (FCCN) e à Linguateca pelo suporte financeiro, através da bolsa POSI/PLP/43931/2001 da Fundação para a Ciência e Tecnologia (FCT), co-financiada pelo POSI. Ao Laboratório de Sistemas Informáticos de Grande Escala (LaSIGE), minha segunda casa nesses últimos anos.

E, finalmente mas não menos importante, quero registrar o nome de algumas pessoas com quem interagi presencial ou virtualmente durante o período de construção deste documento e que me auxiliaram a desenvolver algumas das idéias descritas nos próximos capítulos: Asunción Gómez-Pérez, James Hendler, Luke McDowell, Paul Buitelaar, Philipp Cimiano e Yorick Wilks. Obrigado pela vossa atenção e consideração.

Conteúdo

Lista de Figuras

LISTA DE FIGURAS

Lista de Tabelas

Lista de Algoritmos

Nomenclatura

ABox *assertion component*

ACE *Automatic Content Extraction*

ANMP Associação Nacional de Municípios Portugueses

BT *Broader Term*

CHAVE Coleção de textos dos jornais Público e Folha de São Paulo dos anos de 1994 e 1995 com os julgamentos de relevância para tópicos do CLEF.

CTT Correios, Telégrafos e Telefones

DCE Detecção e Caracterização de Eventos

DDE Detecção e Despiste de Entidades

DDR Detecção e Despiste de Relacionamentos

DDRL Descrições Diretas de Referência para Locais

DIRL Descrições Indiretas de Referência para Locais

DNS *Domain Name System*

EC Entidade Candidata

EG Entidade Geográfica

EI Extração de Informação

NOMENCLATURA

EIG Extração de Informação Geográfica

EM Entidades Mencionadas

ETL *Extraction, Transformation and Loading*

FCCN Fundação para a Computação Científica Nacional

Geo-Net-PT Primeira Ontologia Geográfica de Portugal

GeoCLEF *Evaluation of multilingual Geographic Information Retrieval systems*

GKB *Geographic Knowledge Base*

GNIS *USGS Concise Gazetteer*

GOG *GKB Ontology Generator*

GREASE *Geographic Reasoning for Search Engines*

HAREM Avaliação conjunta de sistemas de Reconhecimento de Entidades Mencionadas

HTML *HyperText Markup Language*

ICG Integração de Conhecimento Geográfico

IETF *Internet Engineering Task Force*

IGeoE Instituto Geográfico do Exército

IGP Instituto Geográfico Português

IMAR Instituto de Pesquisa da Marinha

INE Instituto Nacional de Estatística

IP *Internet Protocol*

KML *Keyhole Markup Language*

LN Linguagem Natural

MUC *Message Understanding Conference*

NT *Narrower Term*

NUT Nomenclatura de Unidade Territorial

OMG *Object Management Group*

OWL *Web Ontology Language*

PCM Produção do Cenário do Modelo

PLN Processamento da Linguagem Natural

PMI *Pointwise Mutual Information*

RACER *Renamed Abox and Concept Expression Reasoner*

RAP Resposta Automática a Perguntas

RCO Resolução de Co-referências

RDF *Resource Description Framework*

REM Reconhecimento de Entidades Mencionadas

RFC *Request for Comments*

RIG Recuperação de Informação Geográfica

SEI-Geo Sistema de Extração e Integração de Conhecimento Geográfico

SN *Synonym*

TBox *terminological component*

TGN *Getty Thesaurus of Geographic Names*

TLD *Top Level Domain*

URL *Uniform Resource Locator*

WGO *World Geographic Ontology*

NOMENCLATURA

WPT 03 Coleção da Web Portuguesa do ano de 2003

WS Web Semântica

XML *eXtensible Markup Language*

Capítulo 1

Introdução

1.1 Motivação

A maior quantidade de conhecimento existente atualmente está disponível em textos na web. De acordo com Wilks (2008), 85% da informação disponível para ciência, empresas e aquela encontrada de modo informal na web estão no formato não estruturado (texto, a maior parte). Contudo, de modo a tornar-se verdadeiramente útil para sistemas que fazem algum tipo de processamento inteligente, esse conhecimento precisa ser manipulado automaticamente.

Neste ponto, surge a necessidade de estruturar o conhecimento que, nesta forma, é inteligível apenas para humanos. Para isso, é comum a utilização de técnicas de extração de informação, as quais partem de algum modelo pré-definido e tentam encontrar em textos estruturas de informação que encaixem nesse modelo.

Pelo fato de a web ser de grande dimensão e estar em permanente crescimento, é fundamental que abordagens para extração de informação sejam escaláveis. Agichtein e Gravano (2000), Cunningham et al. (2002), Cafarella et al. (2005), e Etzioni et al. (2005) descrevem sistemas que fazem extração de informação a partir de textos em grande escala. Na maior parte desses trabalhos utilizou-se conjuntos simples de padrões para tentar capturar a informação que deve ser extraída. Entretanto, a linguagem natural (LN) permite que alguém expresse determinado tipo de conhecimento de diversas formas. Por exemplo, a informação que a freguesia de Santa Isabel é parte do concelho de Lisboa pode ser expressa

como[1]: *... perto da freguesia de Santa Isabel, concelho de Lisboa.*, *... perto da freguesia de Santa Isabel pertencente ao concelho de Lisboa.*, *... perto da freguesia de Santa Isabel que está localizada no concelho de Lisboa.* e *... Santa Isabel é uma das freguesias do concelho de Lisboa.*, entre outras formas. É essa variedade com que a informação é expressa que faz com que muito conhecimento ainda não esteja presente de modo formal (legível por máquina) em estruturas de representação de conhecimento.

Por outro lado, parte desse conhecimento já está também disponível em bases de dados estruturadas, devendo procurar-se encontrá-lo já nessa forma em primeiro lugar. Antes de extraí-lo de textos, deve-se verificar se este conhecimento ainda não está disponível numa representação inteligível pelas máquinas. Uma vez detectada a existência de novo conhecimento, deve-se representá-lo de modo inteligível, numa representação comum a várias aplicações, como uma ontologia, por exemplo. Uma ontologia é composta por um conjunto de conceitos, relacionamentos e suas propriedades (a definição de ontologia usada nesse trabalho será dada no Capítulo 2, Seção 2.3.2).

Navigli e Velardi (2004) observam que as ontologias são reconhecidas como recursos cruciais para a Web Semântica (WS), mas na prática elas não estão disponíveis, e quando a disponibilidade ocorre, elas são raramente usadas fora de ambientes específicos de pesquisa. Além disso, um problema atual da WS é que muitas ontologias estão distribuídas mas sem interconexão entre elas. Ou seja, na prática, é raro ocorrer a reutilização do conhecimento formalizado nas ontologias. Tal ocorre, principalmente, devido às ontologias serem construídas sob o consenso de comunidades locais e, consequentemente ficarem sub-utilizadas. Outro fator que leva a esse cenário é o fato de a maioria dos termos (conceitos e propriedades) das ontologias não serem disponibilizados juntamente com suas ocorrências. Segundo Ding e Finin (2006), 95,1% dos termos usados em ontologias na WS não contêm ocorrências.

Nesse contexto torna-se essencial suportar o povoamento automático de ontologias. As abordagens que lidam com povoamento (semi-)automático de onto-

[1]Nesse documento, eu adoto a representação gráfica de conceitos, relacionamentos e propriedades em *typewriter*, enquanto ocorrências de conceitos e exemplos retirados de texto são representados em itálico.

logias tentam, geralmente preencher uma hierarquia de conceitos pré-definida. O conteúdo existente em bases de dados pode povoar parcialmente essas ontologias, deixando o desafio de integrar o conhecimento complementar identificado em textos como uma tarefa subsequente. À integração de conhecimento também deve ser dada a mesma atenção, uma vez que os resultados dessa tarefa facilitarão a realização de conexões entre ontologias.

De modo a concretizar melhor os problemas (extração de informação e integração de conhecimento) apresentados até aqui, eu concentro este trabalho no domínio geográfico. A ideia básica é aproveitar o conhecimento geográfico existente de forma dispersa em bases de dados publicamente disponíveis, integrá-lo e expandi-lo com conhecimento proveniente de textos.

As dificuldades envolvidas na extração, limpeza e integração de informação são diversas e incluem:

- O custo para obter e manter bases de dados geográficas é muito caro. As fontes de informação geográfica publicamente disponíveis são raras e a qualidade dos dados é frequentemente baixa. Tal implica um trabalho longo, tedioso e caro na limpeza desses dados, de forma a poder torná-los úteis para outras aplicações. Além disso, a informação fornecida, geralmente, não está suficientemente detalhada.

- A linguagem natural (LN) é vaga e ambígua, o que dificulta o processo de extração de conhecimento. Como conseqüência dessas características, intrínsecas de qualquer lingua, muito conhecimento relevante não é extraído porque não se consegue reconhecê-lo de forma adequada.

- As propriedades dos conceitos geográficos variam bastante. Um `rio` poderá ser naturalmente caracterizado pela `nascente`, `foz` e `comprimento`, enquanto uma `serra` tem `altitude` e uma `cidade` tem `população`. Identificar e classificar corretamente essas propriedades em textos é uma tarefa complexa, dada a grande diversidade de formas como podem estar descritas.

- O estado da arte dos sistemas de reconhecimento de entidades mencionadas que trabalham com textos em português evidenciam que a tarefa de reconhecer (identificar e classificar) entidades mencionadas (EM) geográficas em

textos em português ainda precisa ser melhor investigada. Uma definição de EM pode ser encontrada em (Cardoso e Santos, 2006).

Ao contrário das ontologias construídas a partir de textos de um domínio específico, a informação e o conhecimento que constituem ontologias geográficas estão distribuídos em textos pertencentes a praticamente todos os domínios, como o direito, o turismo e o acadêmico. Ou seja, a informação geográfica é transversal a inúmeros domínios de conhecimento. Por exemplo, em todas as seguintes frases (a) *O Nuno foi alvejado na Av. da República, perto do Campo Pequeno*, (b) *Lisboa é uma das capitais turísticas da Europa* e (c) *Muitas das faculdades da Universidade de Lisboa estão localizadas no Campo Grande*, existe informação geográfica relevante para figurar numa ontologia.

Brewster et al. (2003) já alertaram para o fato de textos de domínio específico serem muito restritivos para a construção de ontologias de domínio e da necessidade de utilização de fontes alternativas como textos da web mundial. Como a informação geográfica é transversal a diversos domínios de conhecimento, é provável que seu uso seja feito por um número maior de pessoas. Himmelstein (2005) encontrou um ou mais identificadores geográficos reconhecíveis e não ambíguos (p. ex. códigos postais) em pelo menos 20% das páginas da web mundial. McCurley (2001) encontrou código postal no padrão americano em 4,5% das páginas web (em inglês) e número de telefone em 8,5% das mesmas.

Borges (2006) detectou presença de informação geográfica em 14,77% do total de páginas de uma coleção da web brasileira. Em Chaves e Santos (2006) e Santos e Chaves (2006) apresentamos estudos preliminares sobre o dimensionamento do conteúdo geográfico em textos da web portuguesa, bem como a sobreposição (ambigüidade) de nomes geográficos com nomes de organizações e pessoas. Os resultados, detalhados no Capítulo 4, evidenciam que existe informação geográfica em quantidade significativa que não está presente em bases de dados públicas administrativas com informação geográfica sobre Portugal para suportar a construção e povoamento de uma ontologia geográfica (geo-ontologia, daqui em diante).

1.1.1 Problemas

As bases de informação geográfica são ricas em informação. Contudo, trazem intrinsicamente propriedades como incompletude, desconexão, contradição e variação ao longo do tempo, entre outras, como em qualquer outra base de informação.

Por um lado, quando a informação geográfica histórica é proveniente de bases de dados, essa geralmente não está integrada com informação contemporânea. Especificamente em Portugal, o conceito histórico de `província` dificilmente está integrado com os conceitos atuais de `distrito` e `concelho`.

Por outro lado, quando a fonte de informação é texto em linguagem natural (LN), os relacionamentos entre os conceitos geográficos estão presentes com maior diversidade. Entretanto, raramente são aproveitados (explorados) no processamento automático por sistemas de Processamento da Linguagem Natural (PLN). Ou seja, estruturas de representação de conhecimento geográfico ricas em relacionamentos provenientes da LN são praticamente inexistentes.

Uma das razões para a carência de trabalhos em representação de conhecimento geográfico presente em textos é a falta de definição de uma terminologia capaz de aproximar o conhecimento presente em texto àquele processável por aplicações da WS, por exemplo.

1.2 Objetivo e Contribuições

O objetivo geral dessa tese é propor uma metodologia para construção de geo-ontologias com informação integrada de múltiplas fontes de informação, desde bases de dados até coleções de documentos.

Para alcançar esse objetivo geral, as seguintes etapas específicas devem ser realizadas:

- Encontrar, limpar e integrar informação proveniente de bases de dados geográficas com informações complementares umas das outras;
- Caracterizar a geograficidade presente em textos em português;

- Reconhecer o conhecimento disponível em textos e gerar uma representação formal desse conhecimento (Extração de Informação e Representação de Conhecimento);

- Integrar a ontologia gerada por um sistema de extração e integração de conhecimento geográfico numa ontologia existente (Integração de Informação e Povoamento e Extensão de Ontologias).

Nesse trabalho, o conhecimento geográfico é representado em geo-ontologias compostas por conceitos, relacionamentos, propriedades, axiomas e ocorrências. Esse conhecimento é proveniente tanto de bases de dados quanto de textos. A integração de conhecimento é realizada com a informação proveniente das bases de dados e com a informação relevante extraída dos textos.

Os principais resultados deste trabalho são:

- uma metodologia para construção de bases de conhecimento geográfico com informação integrada de múltiplas fontes de informação. Essa metodologia permite a geração de geo-ontologias que são utilizadas por aplicações da Web Semântica.

- um sistema de gestão de conhecimento geográfico. Esse sistema permite o armazenamento de conteúdo geográfico em um único repositório, integrando esse conteúdo de diferentes fontes de informação e disponibilizando o conhecimento integrado como geo-ontologias que são utilizadas por aplicações da Web Semântica (Chaves et al., 2005a,b);

- um sistema para extração e anotação de conhecimento geográfico de textos e integração desse conhecimento em geo-ontologias, utilizando textos em português. Esse sistema obteve o segundo melhor resultado na tarefa de classificação semântica de locais e o primeiro lugar no cenário de reconhecimento de relacionamentos de inclusão entre locais numa avaliação conjunta que participou (Chaves, 2008);

- o dimensionamento da geograficidade da web portuguesa. Esse dimensionamento é feito através da verificação da quantidade de locais em textos, bem como da sobreposição dos nomes desses locais com nomes de pessoas

e organizações. Numa amostra de 32.000 documentos, 31% dos nomes de pessoas e 23,43% dos nomes de organizações continham pelo menos um nome geográfico incluído na Geo-Net-PT (ver último item desse bloco). Nessa mesma amostra, 98,4% dos documentos continham locais (não necessariamente presentes em geo-ontologias). Uma coleção web continha somente 10% dos locais presentes numa geo-ontologia (Chaves e Santos, 2006; Santos e Chaves, 2006);

- a construção e disponibilização pública e gratuita de uma geo-ontologia de Portugal (Geo-Net-PT - `http://xldb.fc.ul.pt/geonetpt`) com conhecimento integrado de diversas fontes de informação, as quais são complementares, incompletas, contêm informação histórica e contemporânea e provenientes de entidades com diferentes graus de autoridade. Essas fontes incluem bases de dados com informação geográfica sobre Portugal, almanaques, textos jornalísticos e da web portuguesa. Até julho de 2009, a Geo-Net-PT teve mais de 35 pedidos provenientes de pessoas e grupos de pesquisa de 11 países.

Uma das inovações dessa tese é a utilização de textos indiscriminados e não somente um subconjunto de textos de um domínio específico na integração de informação em ontologias com conteúdo extraído de textos. Estes são muitas vezes criteriosamente selecionados, como por exemplo em (Celjuska e Vargas-Vera, 2004; Navigli e Velardi, 2004; Szulman et al., 2002; Velardi et al., 2001; Zong et al., 2005). Outros trabalhos (Dill et al., 2003; Etzioni et al., 2005) utilizam toda a web para extrair fatos, mas não apresentam nenhuma metodologia para formalizar e integrar o conhecimento extraído àquele existente.

As dificuldades na utilização de textos indiscriminados envolvem:

- a maior probabilidade de ocorrer ambigüidade, uma vez que o número de domínios dos textos é aberto;

- os nomes geográficos e de conceitos são mencionados com uma diversidade de denominações maior do que em textos de domínio específico (p. ex. o conceito de `cidade` pode ser mencionado como `município`, `cidadezinha` e

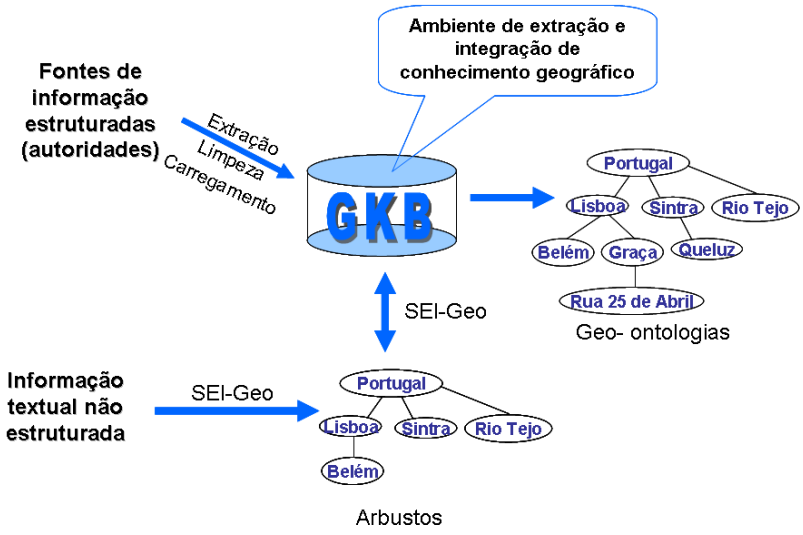


Figura 1.1: Arquitetura global do sistema de gestão de conhecimento geográfico.

até mesmo o termo `cidadela` deve ser considerado, mesmo sabendo-se que ele é erroneamente empregado nesse contexto);

- a qualidade das páginas é muito variável (Ringlstetter et al., 2006). As páginas de blog, os textos jornalísticos e de enciclopédias possuem características que variam desde a escrita informal, com gírias e neologismos até a escrita formal baseada em regras de estilo e sem erros ortográficos.

Nesse contexto se enquadram outras contribuições desta tese, a extração e integração em ontologias de conhecimento geográfico extraído de textos. O povoamento é realizado pela busca de nomes e relacionamentos geográficos em texto para associar aos conceitos previamente definidos em uma geo-ontologia.

A Figura 1.1 apresenta uma visão panorâmica da arquitetura global do sistema de gestão de conhecimento geográfico desenvolvido nessa tese, a *Geographic Knowledge Base* (GKB). A GKB é um ambiente de extração e integração de conhecimento geográfico que contém informações provenientes de fontes de dados administrativas semi-estruturadas de autoridades junto com um conjunto de regras para integração de informação. A expansão do conhecimento contido na GKB ocorre com informação proveniente de textos. Esses textos são a entrada de informação para o Sistema de Extração e Integração de Conhecimento Geográfico

(SEI-Geo), que é o responsável por gerar uma representação estruturada (em forma de arbustos) do conhecimento geográfico extraído e integrá-lo no repositório da GKB. *Scripts* para geração de ontologias exportam o conhecimento armazenado nesse repositório.

1.3 Metodologia Seguida Nesta Tese

A metodologia utilizada para desenvolver e avaliar a metodologia proposta nesta tese é principalmente experimental.

Diversos programas foram implementados para medir a geograficidade em textos da web e para verificar a sobreposição entre o conhecimento proveniente das fontes de autoridades e os textos. Após verificar que existe informação geográfica suficiente para expandir uma base de conhecimento geográfico, foi desenvolvido um sistema para extrair e integrar conhecimento geográfico de textos. Esse sistema foi avaliado através da participação em um evento de avaliação conjunta de sistemas de reconhecimento de entidades mencionadas em português.

A avaliação da metodologia proposta passou pela avaliação dos componentes envolvidos: as aplicações que utilizaram e utilizam as geo-ontologias geradas a partir da base de conhecimento geográfico, os resultados obtidos pelo sistema de extração e integração de informação geográfica na avaliação conjunta que participou, bem como na sua capacidade de expandir o conhecimento existente em geo-ontologias.

1.4 Contexto

O contexto dessa tese envolve os seguintes projetos e avaliações:

Linguateca: é um centro de recursos distribuído para o processamento computacional da língua portuguesa. A Geo-Net-PT, um dos recursos desenvolvidos nessa tese, é distribuída pela Linguateca.

GREASE e GREASE II: o projeto *Geographic Reasoning for Search Engines* (GREASE) pesquisa métodos, algoritmos e arquiteturas de software para atribuir âmbitos geográficos (a região geográfica, se ela existe, onde a

maioria das pessoas pensa ser mais relevante para caracterizar o contexto geográfico de uma página, sítio ou domínio da web) para recursos da web e para recuperar documentos usando entidades geográficas. O projeto GREASE II é uma extensão do GREASE. No âmbito desses projetos foi desenvolvido um sistema de recuperação de informação geográfica (RIG) que utilizou as geo-ontologias produzidas nessa tese.

CLEF e GeoCLEF: O *Cross Language Evaluation Forum* (CLEF[1]) é um fórum para avaliação de sistemas de informação mono-língues e multi-língues que utilizem linguagens europeias. O GeoCLEF[2] é uma das pistas do CLEF que avalia sistemas de RIG mono-língues e multi-língues (Gey et al., 2005). O sistema de RIG desenvolvido no âmbito do projeto GREASE participou nas quatro edições do GeoCLEF.

HAREM: é uma avaliação de sistemas de reconhecimento de entidades mencionadas em português em diversas categorias (p. ex. Pessoa, Local, Tempo), seus tipos (p. ex. Povo, Físico e Tempo_Calend, respectivamente) e subtipos no caso de Local e Tempo (p. ex. Ilha, Hora). O sistema SEI-Geo, produzido por mim, participou nessa avaliação reconhecendo locais e relacionamento entre locais.

1.5 Estrutura da Tese

Esta tese está organizada como segue: o Capítulo 2 apresenta os conceitos e a terminologia utilizada, bem como as principais estruturas de representação de conhecimento, seguidos de uma análise comparativa entre essas estruturas. São descritos trabalhos que constituem o estado da arte em extração e integração de informação geográfica bem como uma comparação entre os mesmos, na qual o sistema proposto nessa tese é enquadrado.

O Capítulo 3 descreve a metodologia para a construção de uma base de conhecimento geográfico a partir de dados semi-estruturados. O processo de extração, limpeza e integração das fontes de informação e a geração de geo-ontologias é

[1] www.clef-campaign.org
[2] www.uni-hildesheim.de/geoclef

detalhado juntamente com estatísticas descritivas dessas geo-ontologias. Por fim, são apresentadas as aplicações que utilizam as geo-ontologias produzidas a partir da base de conhecimento.

O Capítulo 4 introduz uma caracterização da geograficidade em textos em português e uma caracterização das geo-ontologias utilizadas nessa tese. Esse capítulo quantifica a presença de informação geográfica em textos em português bem como a ambigüidade existente com nomes de pessoas e organizações. Além disso, é descrita a presença do conhecimento de geo-ontologias nos textos.

O Capítulo 5 apresenta o formato utilizado para extrair conhecimento geográfico de textos e a arquitetura do Sistema de Extração de Informação e Integração de Conhecimento Geográfico – SEI-Geo. A seguir, os algoritmos de extração de informação e integração de conhecimento são descritos detalhadamente.

O Capítulo 6 descreve a avaliação dos métodos e algoritmos propostos. A avaliação do SEI-Geo foi feita com expansão de geo-ontologias e com a participação num evento de avaliação conjunta de reconhecimento de entidades mencionadas em português.

O Capítulo 7 apresenta a síntese da metodologia para a construção de geo-ontologias proposta nessa tese.

O Capítulo 8 encerra a tese com as considerações finais, as limitações das soluções propostas e apresenta algumas ideias para a continuidade desse trabalho entre as reflexões finais.

Capítulo 2

Conceitos e Trabalhos Relacionados

2.1 Introdução

O objetivo desse capítulo é apresentar os conceitos fundamentais que sedimentam essa tese juntamente com as principais estruturas de representação de conhecimento, as quais são descritas e comparadas. Em seguida, o capítulo apresenta e compara sistemas sobre extração de informação de textos e integração de informação geográfica. O capítulo encerra com as principais avaliações de sistemas de extração de informação e as conclusões.

2.2 Terminologia

As definições utilizadas nessa tese são inspiradas pela norma ISO19109 (2006) e constituem uma proposta de normalização da terminologia usada em Recuperação de Informação Geográfica (RIG) e Extração de Informação Geográfica (EIG) em português, considerando todas as variantes dessa língua.

- **Nome de Entidade (NE):** é definido como todo o nome próprio que refere-se a um local geográfico. É uma designação equivalente a Nome de Local.

- **Nome de Tipo de Entidade (NTE):** é um conceito geográfico (p. ex. país, cidade e rio).

- **Referência Ontológica (RO):** é uma entidade geográfica definida sem ambiguidade, por um único identificador numa geo-ontologia. Ou seja, é um objeto qualquer descrito em texto com um identificador numa estrutura de representação de conhecimento geográfico.

A partir desses três conceitos atômicos, definem-se os seguintes conceitos (os colchetes '[]' representam listas):

- **Ontologia Geográfica ou Geo-ontologia:** apoiado na definição de ontologia dada por Gruber (1993), eu adoto a seguinte definição de geo-ontologia neste trabalho. Uma geo-ontologia é um conjunto de conceitos geográficos e relacionamentos geográficos definidos formalmente e sem ambigüidade. Esses conceitos e relacionamentos fazem parte do vocabulário do domínio (neste caso, o geográfico) no qual está sendo usada a ontologia. Exemplo de conceitos geográficos incluem país, cidade e rio, enquanto adjacência, equivalência e meronímia são exemplos de relacionamentos entre conceitos.

- **Designação de Entidade (`DE`):** é um NE identificado em texto através de heurísticas ou que está presente em geo-ontologias. Pode ser precedido ou sucedido por um NTE. É representada por um par `<NTE,NE>` (p. ex. `<concelho,Porto>`), sendo `NTE` opcional.

- **Esboço de Entidade (`EE`):** representado pelo par de listas `<[DE],[RO]>`, onde qualquer uma das listas pode ter 0..n elementos. Se a lista `RO` não é vazia, a lista `DE` contém obrigatoriamente pelo menos um elemento. Por exemplo, `<[<concelho,Sintra>],[GEO_284]>`, `<[<vila,Nova>],[]>` e `<[<rua, Vasco da Gama>],[GEO_4118,GEO_5142,...,GEO_407509]>`. Existem pelo menos 207 ruas com o nome `Vasco da Gama` em Portugal.

- **Entidade Geográfica (`EG`):** é um objeto com significado no domínio do discurso (correspondente à *feature* na (ISO19109, 2006)). No domínio geográfico, a 'província do Algarve', o 'concelho de Évora' e a 'freguesia de Santa Isabel' são exemplos de tais entidades numa ontologia. Essas entidades geográficas devem ter uma referência numa ontologia, e essa referência fornece o significado no domínio geográfico. Formalmente, uma

entidade geográfica é um EE que refere apenas uma referência ontológica (|[RO]|=1) (p. ex. <[<concelho,Évora>],[GEO_346]>).

- **Tipo de Entidade (TE):** corresponde aos conceitos ou classes numa estrutura de representação de conhecimento (p. ex. 'região', 'estado' e 'cidade'). Cada TE corresponde a uma RO. Tipo de Entidade correspondente à *feature type* na (ISO19109, 2006).

- **Relacionamento (R):** é representado como um padrão geográfico através de uma cadeia de caracteres que indica presença de informação geográfica em textos. Um padrão geográfico pode ser de vários tipos: métrico (p. ex. 'km de' e 'minutos de'), direcional (p. ex. 'atrás de' e 'em frente a'), verbo (p. ex. 'localizado' e 'situado'), orientação (p. ex. 'norte' e 'leste'), fuzzy (p. ex. próximo, antes e acima) ou padrões de Hearst expandidos (p. ex. 'é um concelho' e 'é a aldeia'). Esses padrões também podem ser vistos como expressões gramaticais com termos que podem ser precedidos ou sucedidos por informação geográfica. Um relacionamento pode ser representado por tipos de relacionamentos (p. ex. 'parte-de' e 'adjacente'). Os relacionamentos também são utilizados para associar TE.

- **Relacionamento Espacial:** é a ligação entre pelo menos dois pontos no espaço.

- **Tipo de Relacionamento:** é o nome dado ao relacionamento conforme o tipo de relação existente entre as entidades geográficas.

- **Associação entre Entidades (AE):** é uma tripla definida por $<EE_1,R,EE_2>$, onde $EE_1 \neq EE_2$.

- **Associação entre Tipos de Entidades (ATE):** é uma tripla definida por $<TE_1,R,TE_2>$, onde $TE_1 \neq TE_2$. Por exemplo, <Estado,parte-de,País>.

- **Padrões léxicos-sintáticos:** são aqueles compostos por palavras e categorias gramaticais, tal como substantivo. Também é frequente o uso de sintagmas nominais, que são expressões que têm um substantivo como núcleo. Os substantivos na maioria dos casos são restritos a nomes próprios.

- **Âmbito Geográfico:** a região geográfica, se ela existe, onde a maioria das pessoas pensa ser mais relevante para caracterizar o contexto geográfico de uma página, sítio ou domínio da web. Por exemplo, o âmbito geográfico do sítio web da Câmara de Lisboa (*www.cm-lisboa.pt*) é o concelho de Lisboa.

- **Ocorrência Geográfica (ou objeto geográfico):** é equivalente ao termo *instance*, frequentemente utilizado pela comunidade de Inteligência Computacional.

2.3 Representação de Conhecimento Geográfico

A noção de geografia de senso comum[1] reflete o conhecimento que as pessoas têm sobre o mundo geográfico ao seu redor (Egenhofer e Mark, 1995). Por senso comum entende-se um pensamento instintivo ou espontâneo. Tezuka et al. (2004) introduzem a noção de nomes de lugares significantes, os quais são nomes de lugares bem conhecidos com alta significância cognitiva. A significância cognitiva pode ser genérica ou espacial, conforme proposta em Tezuka e Tanaka (2005). Objetos que são bem conhecidos em um sentido genérico, mas não no sentido espacial contém uma significância genérica. Universidades e empresas são exemplos de tais objetos algumas vezes, como em 'A Universidade de Lisboa está lançando um novo curso de Tecnologia de Informação'.

Por outro lado, pontos de referência, regiões turísticas e arruamentos contêm uma significância espacial. As pessoas conhecem suas localizações. São objetos que os residentes locais podem facilmente localizar em um mapa. Por exemplo, em 'O Bairro Alto está próximo da Baixa Chiado', qualquer cidadão local pode localizar em um mapa os locais citados.

Uma conceitualização comum do mundo geográfico é baseada na ideia de objetos e campos propostos em Couclelis (1992). O modelo de objetos representa o mundo como uma superfície ocupada por entidades discretas e identificáveis que possuem uma representação geométrica. Exemplos de entidades nesse modelo são rodovias, edifícios e outros pontos de referência. O modelo de campos vê a

[1] Do inglês, *naive geography*.

realidade geográfica como um conjunto de distribuições espaciais sobre o espaço geográfico. Campos incluem clima, vegetação e mapas geológicos.

Um problema com os modelos de objetos e campos é sua generalidade, o que leva a uma carência de suporte semântico para os diferentes tipos de dados espaciais. Uma solução possível para esse problema, passa pelo uso de ontologias, conforme sugerem Fonseca et al. (2002). As ontologias são flexíveis e capazes de representar tanto o modelo objeto quanto o modelo campo, permitindo uma representação semântica específica para os tipos de dados geográficos. Fonseca et al. assumem que um objeto geográfico é somente uma representação de uma entidade, mas é visto diferentemente por diferentes pessoas. Uma cidade pode ser vista como um par de coordenadas, um conjunto de pontos de referência ou um conjunto de sub-entidades administrativas, entre outras visões. Os objetos referenciados por pares de coordenadas estão presentes, por exemplo, em almanaques digitais.

2.3.1 Almanaques Digitais

Um almanaque digital é um dicionário geográfico que contém informação específica sobre topônimos, os quais estão normalmente dispostos de forma hierárquica. A cada topônimo estão associados, frequentemente, as suas coordenadas geográficas.

O *Getty Thesaurus of Geographic Names*[1] (TGN) é um vocabulário, estruturado com nomes e informação associada sobre locais atuais e históricos no mundo. Cada local no TGN é identificado com um único identificador. A cada identificador são associados nomes (históricos, comuns, alternativos e em diferentes línguas), posição na hierarquia, outros relacionamentos, coordenadas geográficas, notas, fontes para os dados e os tipos dos locais (p. ex. cidade, estado ou província).

Leidner (2007) apresenta uma lista com mais oito almanaques digitais disponíveis na web (p. 52). Manguinhas et al. (2008) também apresentam uma lista mais extensa de almanaques públicos e privados.

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

2.3.2 Ontologias

As ontologias são estruturas de representação de conhecimento apropriadas para a organização de informação não estruturada. O termo ontologia é interdisciplinar e tem-se tentado defini-lo distintamente conforme a área de conhecimento na qual ele está inserido. Guarino (1997) apresenta e discute diversos conceitos. Contudo, ainda não existe consenso sobre o que é uma ontologia em Ciência da Computação, conforme opiniões encontradas em diversos ensaios em Brewster et al. (2004). Por outro lado, uma definição amplamente utilizada é dada por Gruber (1993): ontologia é uma especificação explícita e formal de uma conceitualização compartilhada. Fensel (2001) descreve esse conceito em partes, afirmando que uma 'conceitualização' refere-se a um modelo abstrato de algum fenômeno no mundo que identifica conceitos relevantes daquele fenômeno. Guarino (1997) ainda comenta que uma 'conceitualização' explica o significado pretendido dos termos usados para indicar relações relevantes. 'Explícito' significa que os tipos de conceitos usados e as restrições para esses conceitos são definidos explicitamente. 'Formal' refere-se ao fato de que uma ontologia deve ser legível para as máquinas. Uma ontologia pode ser expressa em diferentes graus de formalidade, por exemplo usando as linguagens CycL[1], OWL-Lite, OWL-DL e OWL-Full[2]. A implicação dos graus de formalidade está no poder de expressividade de cada linguagem. 'Compartilhada' reflete a noção de que uma ontologia captura o conhecimento consensual, isto é, o conhecimento não é restrito a algum indivíduo, mas aceito por um grupo.

Apoiado na definição de ontologia dada por Gruber (1993), eu adoto a seguinte definição de neste trabalho:

> Uma geo-ontologia é um conjunto de conceitos geográficos e relacionamentos geográficos definidos formalmente e sem ambigüidade. Esses conceitos e relacionamentos fazem parte do vocabulário do domínio (neste caso, o geográfico) no qual está sendo usada a ontologia. Exemplo de conceitos geográficos incluem país, cidade e rio, enquanto

[1] http://www.cyc.com/cycdoc/ref/cycl-syntax.html
[2] http://www.w3.org/TR/owl-features

adjacência, equivalência e meronímia são exemplos de relacionamentos entre conceitos.

Freqüentemente, uma geo-ontologia é composta por ocorrências geográficas. Uma ocorrência geográfica[1] é definido por um nome e um tipo <N,T>, por exemplo, <Liberdade,rua>. Uma geo-ontologia quando dividida em geografia administrativa e física também possui relacionamentos inter-domínio, tal como um rio ser parte de uma ou várias cidades, e até mesmo países).

2.3.2.1 Usos de Ontologias

Uma ontologia possui os usos mais diversificados. A seguir são listados alguns deles:

Organização de sites e suporte à navegação: uma estrutura taxonômica em alto nível capaz de guiar usuário e desenvolvedor facilitando o acesso ao conteúdo do site.

Estrutura comum como ponto de partida para se estender conteúdo: uma organização taxonômica em alto nível sobre produtos e serviços é uma estrutura padrão que pode ser estendida de modo a suportar novos termos na taxonomia. Um exemplo de tal taxonomia é apresentado em www.unspsc.org.

Anotação de páginas da web: uma ontologia pode ser utilizada para fornecer suporte à anotação de páginas da web, fornecendo etiquetas semânticas que podem ser exploradas por motores de busca.

Suporte à pesquisa: as ontologias podem fornecer suporte à pesquisa de informação de duas formas:

Expansão de consulta: combinar um termo inserido pelo usuário com os termos alternativos constantes na ontologia.

[1] O termo ponto de referência também foi usado como sinônimo de ocorrência geográfica em (Tezuka e Tanaka, 2005).

Desambiguação e restrição do espaço de busca: a partir do termo inserido pelo usuário no motor de busca, obter uma identificação mais refinada sobre o termo que ele está procurando: por exemplo, em Portugal, o termo 'rua da Liberdade' ocorre em 278 concelhos. Com o suporte da ontologia, o usuário pode refinar sua consulta, explicitando em qual dos 278 concelhos está a 'rua da Liberdade' que ele está à procura.

Em ambos os casos, expansão de consulta e desambiguação e restrição do espaço de busca, um sistema deve explorar o uso de conceitos e relacionamentos presentes na geo-ontologia. A concretização dessas tarefas coloca em prática a visão de Web Semântica geo-espacial, inicialmente introduzida por Egenhofer (2002). Martins et al. (2006b) descrevem como consultas geográficas podem ser manipuladas juntamente com uma geo-ontologia. Uma consulta geográfica é dividida em três partes <o quê,relacionamento,onde>. As partes `relacionamento` e `onde` são projetadas sobre uma geo-ontologia que permite a expansão da consulta através de sub-regiões e locais adjacentes, quando for o caso. No caso da tarefa de desambiguação e restrição do espaço de busca, um sistema pode utilizar um conjunto de heurísticas (eg. população, área e freqüência do termo geográfico em textos) para desambiguar o termo geográfico e restringir a busca a uma determinada área geográfica.

Aplicações que podem fazer uso de geo-ontologias incluem sistemas de recuperação de informação conscientes da geografia, reconhecedores de entidades mencionadas e também aplicações para redução de junções espaciais em bancos de dados geográficos (Bogorny, 2006). Junções espaciais são operações realizadas para computar relacionamentos (p. ex. 'toca' e 'cruza') entre duas entidades espaciais. Esse tipo de junção é computacionalmente cara em bases de dados geográficas e pode ser reduzida com a utilização de geo-ontologias.

2.3.3 Outras Estruturas de Representação de Conhecimento

Além de ontologias, outras estruturas são utilizadas para representar conhecimento.

Vocabulário Controlado: um vocabulário controlado é uma lista de palavras e expressões utilizadas para etiquetar unidades de informação. Usa-se para facilitar a recuperação de informação. Um vocabulário controlado geralmente é registrado por uma autoridade. Todos os termos nesse vocabulário devem ter uma definição não redundante e sem ambigüidade. Se um termo é utilizado para representar diferentes conceitos, então seu nome deve ser explicitamente qualificado para resolver tal ambigüidade. Por outro lado, se múltiplos termos são usados para definir um mesmo conceito, um dos termos deve ser identificado como preferido e os outros são listados como sinônimos ou nomes alternativos. Exemplos de vocabulários controlados incluem a terminologia utilizada em páginas amarelas e lista de assuntos para indexação de banco de dados bibliográficos.

Tesauro: um tesauro é uma coleção de termos de um vocabulário controlado conectada na forma de um grafo. Um tesauro é uma linguagem fechada restrita normalmente por três relacionamentos: equivalência (*Used For* and *USE*), hierarquia (*Broader Term*/*Narrower Term*) e associatividade (*Related Term*) (ISO, 2002a). Outras definições para a estrutura de representação de conhecimento chamada tesauros podem ser encontradas em (Gonzalez, 2001). Exemplos de tesauros são o Eurovoc Thesaurus[1], o Tesauro de Folclore e Cultura Popular Brasileira[2] e o Tesauro em Ciência da Informação[3].

Taxonomia: uma taxonomia é um conjunto de termos organizados em uma estrutura hierárquica. Existem diferentes tipos de relacionamentos entre pais e filhos em uma hierarquia, por exemplo hiper/hiponímia e meronímia. Contudo, apenas uma relação está presente em uma taxonomia. Exemplos de aplicações utilizando taxonomias são sites de comércio eletrônico, bibliotecas e sistemas de RI.

Classificação Facetada: a classificação facetada provê uma estrutura para classificar documentos. Uma classificação facetada procura fornecer os

[1] http://europa.eu/eurovoc
[2] www.cnfcp.com.br/tesauro/index.html
[3] www.inf.pucminas.br/ci/tci

diferentes eixos nos quais um documento pode ser representado. Exemplos de facetas incluem tempo, espaço e matéria, entre outros. Essas facetas não possuem relacionamento direto entre si e também não são representadas como uma hierarquia, ao contrário de outras formas de representação de conhecimento. Aplicações usando classificação facetada disponibilizam para o usuário diferentes dimensões da mesma informação. Por exemplo, num site de venda de joias, o usuário pode ter na interface opção de procura por tipo de joia (p. ex. anel, colar, brinco) ou por material (p. ex. ouro, prata e bronze).

Mapa de tópicos: um mapa de tópicos é uma técnica de classificação baseada em tópicos (ou assuntos). Três construtores são utilizados nesse tipo de classificação: nomes (assuntos), associações (relacionamentos) e ocorrências. Um mapa de tópicos é uma norma ISO/IEC 13250:2003 (ISO, 2002b) para representação de conhecimento. Apesar de se conseguir representar o conhecimento com bastante flexibilidade utilizando um mapa de tópicos, tarefas de raciocínio são limitadas, uma vez que a formulação de regras é complexa. Sistemas utilizando mapa de tópicos incluem portais semânticos e sistemas de RI entre outros listados em sítio Ontopia[1].

Folksonomia: é um conjunto de termos obtidos seguindo uma metodologia colaborativa de anotação de objetos (p. ex. documentos, fotos e sítios) que utiliza etiquetas fornecidas por seres humanos (geralmente, não especialistas em determinado domínio). Essas etiquetas fornecidas são livres de estrutura e não possuem hierarquia, ou seja, a rede de etiquetas formada é rasa. O produto final de um sistema que utilize essa metodologia é um conjunto de termos que representa o conhecimento do ponto de vista do usuário. Ao contrário dos vocabulários controlados, tesauros, ontologias e taxonomias, a folksonomia permite a utilização de qualquer termo de forma livre. Por exemplo, usuário pode usar termos com variação de número (singular e plural) para representar o mesmo conhecimento e ainda utilizar um termo homônimo e/ou polissêmico para representar coisas distintas. Essa característica da abordagem pode trazer limitações para as aplicações que

[1] www.ontopia.net/solutions/oks-applications.html

Tabela 2.1: Comparação entre as diferentes estruturas de representação de conhecimento. TR: tipo de relacionamento; GF: grau de formalidade; ED: requer um especialista no domínio.

	TR	GF	ED
Classificação facetada	Livre	Baixo	Sim
Mapa de tópicos	Livre	Baixo	Não
Vocabulário controlado	-	Baixo	Sim
Taxonomia	É um/Parte de	Alto	Sim
Tesauro	BT/NT SN USE/Used For	Alto	Sim
Meta-modelo	Livre, porém limitado	Baixo	Sim
Folksonomia	Rasa	Baixo	Não
Ontologia	Livre	Alto	Sim

a utilizam. Sistemas que utilizam folksonomia são sítios da Web 2.0, tais como o gestor de *bookmarks* Delicious[1], o gestor de fotos Flickr[2] e o gestor de apresentações Slide Share[3].

2.3.4 Análise Comparativa das Estruturas de Representação de Conhecimento

A Tabela 2.1 apresenta um resumo comparativo das diferentes formas de representação de documentos, considerando os tipos de relacionamentos entre os conceitos utilizados em cada estrutura de representação de conhecimento bem como o grau de formalidade existente. Por grau de formalidade deve-se entender a capacidade de se realizar raciocínio automático com a estrutura em consideração. Outro aspecto avaliado é a necessidade de o conhecimento ser introduzido por um especialista de domínio. Esse aspecto é importante porque tem uma implicação direta no custo da utilização de qualquer uma das estruturas.

No contexto geográfico, a iniciativa de utilizar ontologias ao invés de estruturas rasas pode ser justificada pelos seguintes argumentos também descritos em (Manov et al., 2003):

[1] http://delicious.com
[2] www.flickr.com
[3] www.slideshare.net

- a ontologia possui informação extra, especialmente a relação transitiva *subRegiãoDe* que pode ser utilizada para desambiguar e realizar raciocínio.
- a localização de entidades geográficas no texto pode ser reconhecida no nível correto de granularidade que a aplicação-alvo deseja obter. (p. ex. distrito, concelho, localidade, etc.).
- a ontologia pode ser modificada pelo usuário e qualquer alteração é refletida imediatamente no resultado de uma consulta em um sistema de RI ou EI.

É importante notar que no último item, outras estruturas de representação de conhecimento também podem ser alteradas pelo usuário. Entretanto, o que varia de caso a caso é o grau de conhecimento exigido do usuário para fazer a alteração. Por exemplo, é requerido um conhecimento muito especializado do usuário para alterar um vocabulário controlado, ao passo que uma folksonomia pode ser alterada com conhecimentos mínimos sobre o domínio.

2.4 Processamento de Informação Geográfica

Essa seção descreve o processamento de informação geográfica, o qual é apresentado em duas tarefas: extração e integração de informação.

2.4.1 Extração de Informação Geográfica

Extração de Informação (EI) é a tarefa de localizar informação específica de um documento em linguagem natural. Cowie e Wilks (2000) definem EI como qualquer processo que seletivamente estrutura e combina dados explícita ou implicitamente declarados em um ou mais textos. Mais especificamente, Dowdall et al. (2004) afirmam que EI diz respeito ao reconhecimento das propriedades e relações que são mencionadas em um ponto particular de um texto. Para Cunningham (2006. ISBN 0-08-044299-4), EI é o processo de obtenção de dados quantificáveis desambiguados a partir da linguagem natural, para servir a alguma necessidade de informação precisa e pré-especificada.

No domínio geográfico, a extração de informação geográfica em texto pode ser dividida em dois subprocessos (Densham e Reid, 2003): *geo-parsing* e *geo-coding*,

onde o primeiro refere-se a identificação de nomes de locais, ao passo que o último diz respeito à classificação (desambiguação e marcação).

Um dos requisitos-chave para tornar a EI uma tecnologia útil é desenvolver a habilidade para produzir rapidamente sistemas de EI sem utilizar todos os recursos de pesquisa em PLN (Cowie e Wilks, 2000).

Nos últimos anos diversos esforços têm sido realizados no sentido de representar essa vasta quantidade de conhecimento de forma a permitir um acesso mais fácil e legível aos conteúdos. Assim, as estruturas de representação de conhecimento atrás referidas (vocabulários controlados, tesauros, taxonomias, classificações facetadas, ontologias e mapas de tópicos) têm sido utilizadas.

Com o objetivo de alcançar os requisitos de representação de conhecimento propostos por Berners-Lee et al. (2001) quando denominaram a próxima geração de conteúdo da web como Web Semântica, muitas aplicações da web passaram a utilizar ontologias. Essa forma de representação permite construir uma web com dados em formato legível tanto pelas máquinas quanto pelos humanos.

A extração de entidades geográficas em textos tem sido realizada com a ajuda de almanaques. Para Malouf (2002) sua utilização não auxilia a melhoria dos resultados, enquanto Mikheev et al. (1999) encontram bons resultados utilizando almanaques. Carreras et al. (2003) apresentam resultados melhores com o uso de almanaques. Mikheev et al. (1999) também comprovaram que a utilização de almanaques é necessária para identificar nomes de locais.

Os almanaques utilizados em trabalhos encontrados na literatura variam bastante quanto ao conteúdo da informação presente. Por exemplo, Southall (2003) utiliza um almanaque com informação de localização relativa de cada entrada, ao contrário dos trabalhos apresentados em (Pouliquen et al., 2004), (Fu et al., 2003) e (Li et al., 2002) entre outros, em que cada entrada é associada com informação de localização direta, isto é, um valor para latitude, um para longitude e outro para altitude (quando for o caso).

As referências geográficas são fáceis de identificar por humanos, enquanto que para as máquinas existe uma dificuldade associada, uma vez que a maior parte delas está na forma de texto 'cru ou plano' (sem marcação sintática ou semântica) necessitando um conhecimento relacionado para solucionar a ambigüidade

semântica (Hiramatsu e Reitsma, 2004). Rauch et al. (2003) ainda acrescentam que referências geográficas são com frequência pouco específicas.

Alguns dos problemas tratados nesse domínio referem-se a ocorrência de ambigüidade no nome das entidades. Existem dois principais tipos de ambigüidade em geo-referências: ambigüidade por referente e multiplicidade de referências.

A primeira ocorre quando o mesmo nome é utilizado para identificar mais de um local ou referir-se a outra categoria semântica. Um exemplo de identificação de mais de um local por um nome é 'Castelo Branco', que de acordo com dados da Geo-Net-PT, identifica um distrito, um concelho, uma localidade e uma rua, entre outros. Quando um nome de local refere-se a outras categorias (p. ex. pessoas e organizações), pode-se mencionar o mesmo exemplo. 'Em Castelo Branco encontram-se vários personagens literários.', neste caso 'Castelo Branco' é classificado como pessoa. Ou ainda, 'A restaurante Castelo Branco está sempre lotado, chegue cedo.'. Aqui, 'Castelo Branco' é classificado como organização.

A multiplicidade de referências ocorre quando o mesmo local possui mais de um nome (p. ex. 'Grande Lisboa' e 'região metropolitana de Lisboa') referem-se para o mesmo local no espaço. Outra situação nesse mesmo caso de ambigüidade é o uso de um mesmo termo de diversas formas (p. ex. 'Rua Nossa Senhora da Boa Viagem', 'Rua N. Senhora da Boa Viagem' ou 'Rua N. Sra. da Boa Viagem').

Para tentar reduzir a ambigüidade entre tipos de entidades geográficas a tarefa de normalização de localizações[1] tem sido utilizada. Normalização de locais é uma aplicação especial da desambiguação do sentido das palavras (Li et al., 2003).

Além das entidades geográficas extraídas dos textos, outra questão em aberto é a determinação de quais os relacionamentos espaciais devem ser codificados entre as entidades geográficas. Fu et al. (2003) utilizam os relacionamentos *parteDe*, *contém*, *adjacente* e *sobreposição*.

Na extração de informação geográfica de páginas web, Arikawa et al. (2004) consideram dois tipos de descrições geográficas.

- Descrições Diretas de Referência para Locais (DDRL). Neste caso, incluem-se as coordenadas geográficas latitude e longitude.

[1]Do inglês *location normalization*.

- Descrições Indiretas de Referência para Locais (DIRL). Exemplos deste caso são endereços, nomes de pontos turísticos, códigos postais, nomes de ruas e de localidades.

Qualquer sistema que utilize estes tipos de informação pode optar por trabalhar com ambas as descrições ou com a conversão entre essas descrições. Na conversão de DIRL para DDRL os problemas de ambigüidade da linguagem natural têm sido os maiores. Já a conversão de DDRL para DIRL é um processo mais fácil, uma vez que não existem fisicamente dois lugares no espaço que ocupem o mesmo ponto geográfico.

2.4.2 Integração de Informação Geográfica

Brewster e Wilks (2004) afirmam que o principal problema na construção de ontologias não é construir uma hierarquia, mas sim, assumir que os termos existem e determinar a natureza da relação entre eles.

A integração de dados provenientes de fontes de informação distintas, heterogêneas, complementares e autônomas permanece um problema desafiador, embora diversos trabalhos já tenham fornecido estratégias de integração de dados capazes de minimizar o problema (Cohen, 1997; Cohen et al., 2003; Gravano et al., 2003; Levenshtein, 1966; Winkler, 1995).

Um algoritmo de integração de dados consiste geralmente em comparar dois termos A e B, aplicando uma medida de similaridade sobre eles e como resultado, a saída do algoritmo é um valor (geralmente normalizado entre 0 e 1) que permite inferir se os termos comparados são similares ou não. Tipicamente, é adotado um limiar que permite distinguir termos similares e distintos.

A similaridade entre termos é um problema que vem sendo estudado ao longo de muitos anos, sendo das contribuições iniciais mais relevantes a de Levenshtein (1966). Depois disso, surgiram diversas outras métricas com abordagens ao mesmo tempo distintas e complementares, cujo objetivo comum é verificar se dois termos são lexicalmente similares ou não. Exemplos de métricas de similaridade são a distância de Levenshtein, o coeficiente de Dice, a métrica de Jaro Winkler e a similaridade do co-seno entre outras (Baeza-Yates e Ribeiro-Neto, 1999). Uma lista dessas e outras métricas está disponível em `http://www.dcs.shef.`

ac.uk/~sam/stringmetrics.html. Especificamente para o português, a medida Similaridade Lexical foi proposta em (Chaves, 2004; Chaves e Lima, 2004), a qual é baseada nos radicais dos termos comparados.

2.5 Sistemas de Extração e Integração de Informação Geográfica

2.5.1 Snowball

O sistema Snowball recebe um conjunto de tuplas (p. ex. 'organização,local') definidas manualmente (Agichtein e Gravano, 2000). A partir dessas tuplas o sistema procura segmentos de texto em que ambas ocorrem e tenta identificar padrões nessas ocorrências. O Snowball utiliza um etiquetador de entidades mencionadas (*MITRE Corporation's Alembic Workbench*). A todos os padrões identificados em Snowball é associado um grau de confiança. É possível aceitar ou rejeitar um padrão conforme o número de tuplas extraídas. No Snowball o usuário fornece exemplos de tuplas para treino do sistema bem como uma expressão regular genérica cujas entidades devem combinar. Por exemplo, '<PT Comunicações, Lisboa>' e a expressão '<texto1> localizado em <texto2>'.

Os experimentos realizados foram sobre o corpus jornalístico *North America News Text Corpus*, o qual é composto por artigos do *Los Angeles Times*, *The Wall Street Journal* e *The New York Times*, de 1994 a 1997. A avaliação dos resultados do Snowball foi realizada manualmente com 100 pares de `organização,local` escolhidos aleatoriamente. Os erros foram divididos em três categorias: local etiquetado erroneamente, organização etiquetada erroneamente e relacionamento errado (p. ex. 'Torre do Tombo localizada em Évora').

O método com pior desempenho foi o *Baseline*, que detecta nomes de organizações e locais que co-ocorrem na mesma sentença. *Baseline* extraiu 25 pares corretos e 75 incorretos. Por outro lado, o método Snowball com grau de confiança ajustado para 0,8 obteve o melhor desempenho, com 93 pares corretos e sete incorretos.

2.5.2 OntoLearn

Navigli e Velardi (2004) desenvolveram um método para extração de ontologias de domínio a partir de sítios da web no domínio específico do turismo. O método é composto por três fases:

Extração de terminologia: baseado na análise gramatical feita nos documentos, o OntoLearn extrai listas com termos marcados com as seguintes etiquetas: SN[1], adjetivo-SN e preposição-SN.

Interpretação semântica: é o processo para determinar o conceito correto (sentido) para cada componente de um termo complexo e então identificar as relações semânticas existentes entre os componentes do conceito para construir um conceito complexo. Um exemplo de um conceito complexo, que utiliza a abordagem de inclusão de cadeia de caracteres e é formado após o procedimento de desambiguação (consulta a base de dados lexical Wordnet (Miller et al., 1990)) é 'serviço de transporte público' (p. ex. serviço- >serviço de transporte->serviço de transporte público).

Integração na ontologia: a floresta de conceitos gerada pelo OntoLearn é usada para 'aparar' e atualizar o WordNet, criando uma ontologia de domínio. O processo ocorre como segue:

- após as árvores de conceitos de domínio serem anexadas (manual e automaticamente) aos nodos apropriados do WordNet, todos os ramos que não contêm um nodo do domínio são removidos da hierarquia do WordNet.
- um nodo intermediário do WordNet é aparado sempre que as seguintes condições são encontradas:
  1. não tem nodos irmãos;
  2. tem somente um hipônimo direto;
  3. não é a raíz da árvore de conceitos do domínio;

[1] Sintagma nominal.

4. não está a uma distância ≤ 2 de um nodo raiz da WordNet (para preservar uma ontologia de topo mínima).

Num ensaio, o OntoLearn extraiu 14.383 termos candidatos (fase de extração de terminologia) de textos retirados de sites sobre turismo. Desses, o sistema derivou 3.840 conceitos que foram avaliados por especialistas do domínio.

2.5.3 KnowItAll e KnowItNow

O sistema KnowItAll usa uma abordagem livre de treino que permite a extração de grandes coleções de fatos, conceitos e alguns relacionamentos de modo não supervisionado, dependente ou independente de domínio e escalável (Etzioni et al., 2005). Seu objetivo é melhorar a abrangência sem sacrificar a precisão.

KnowItAll atribui uma probabilidade a cada fato extraído. A entrada do sistema é um conjunto de predicados e a saída são informações na forma de fatos. Por exemplo, o predicado `filme` pode ser recebido pelo sistema e este pode gerar o fato `Filme(Dança com Lobos)`. Um dos processos importantes do KnowItAll é o *Bootstrapping*, o qual consiste de um conjunto de regras que são carregadas com predicados. Por exemplo, a regra:

```
Predicado: Class1
Padrão: NP1 'tal como' NPList2
Restrições: head(NP1)=plural(label(Class1))
            NomeProprio(head(each(NPList2)))
```

Pode ser carregada com:

```
Predicado: Cidade, município
Padrão: NP1 'tal como' NPList2
Restrições: head(NP1)='cidades', 'municípios'
            NomeProprio(head(each(NPList2)))
```

Assim, o motor de busca recebe apenas os padrões:

```
cidade tal como
cidades tais como
município tal como
municípios tais como
```

Esse tipo de regra permite extrair ocorrências pertencentes aos conceitos dentro de uma determinada ontologia, permitindo aumentar a população da ontologia.

O KnowItAll também possui dois módulos principais: `Extractor` e `Assessor`. O `Extractor` cria uma consulta a partir das palavras-chave em cada regra, envia a consulta para o motor de busca, aplica a regra para extrair a informação das páginas resultantes, testa o núcleo de cada NP para verificar se este é um nome próprio e, finalmente, extrai o núcleo (*head*). Por exemplo, no resultado 'cidades tais como Porto, Braga e Guimarães' os nomes são extraídos como nomes de cidades, ao passo que no resultado 'mapas detalhados e informação para várias cidades tais como mapas de aeroportos, centro das cidades, etc.', *aeroportos* e *centro das cidades* não são nomes próprios, logo, não são extraídos.

O módulo `Assessor` consiste de discriminadores (validadores) tais como 'cidade', 'concelho' e 'distrito'. Esses discriminadores submetem os extratos gerados pelo `Extractor` ao motor de busca. Por exemplo, 'Cidade Lisboa' ou 'Sintra Concelho de Lisboa'. Os resultados são validados com base em *Pointwise Mutual Information* (PMI) (Turney, 2001). Se os resultados estão acima de um determinado limiar, são considerados válidos e o sistema insere na base de conhecimento. Os discriminadores ignoram pontuação, espaço em branco e etiquetas HTML.

Outro método utilizado para aumentar a abrangência do KnowItAll é o aprendizado de padrões que consiste das seguintes fases: submissão de consultas (padrões) com os conceitos utilizadas na fase de *bootstrapping* juntamente com suas ocorrências. Por exemplo, o conceito <cidade> e as ocorrências (p. ex. 'Lisboa', 'Vila Nova de Gaia', 'Coimbra'). Para cada documento retornado pelo motor de busca, é armazenado um contexto com quatro palavras à direita e quatro palavras à esquerda do padrão. O contexto mais o padrão formam novos padrões candidatos que são avaliados. Exemplos de padrões candidatos incluem:

```
localizado em <cidade>
<cidade> é
Universidade de <cidade>
```

Outros métodos usados por KnowItAll para coletar fatos na web são:

Extração de subconceitos: o processo de extração de subconceitos pode ser dependente ou independente de contexto. Neste caso, o termo contexto refere-se para um conjunto de palavras-chave fornecidas pelo usuário que sugere um domínio de conhecimento de interesse.

Extração de listas: os métodos apresentados nas seções anteriores atuam sobre texto não estruturado, enquanto o método extração de listas parte da premissa que muitos sites da web são gerados automaticamente a partir de bases de dados. Dessa forma, é possível encontrar uma estrutura regular nesses tipos de páginas web. Assim, o uso de etiquetas HTML passa a ser considerado nesse método. A entrada desse método consiste de um nome de um conceito mais as 'Sementes Positivas' associadas a esse conceito. Por exemplo, 'Rio + Tejo, Douro'. Esta entrada pode produzir o seguinte conjunto de termos candidatos ao conceito Rio: 'Tejo, Douro, Mondego, Ave, Cávado'. Este método é bastante semelhante ao utilizado no Google Sets[1], o qual utiliza um conjunto de palavras como entrada e retorna uma lista máxima de 100 termos que são encontrados em listas na web.

Uma das limitações de KnowItAll é a utilização de um motor de recuperação de informação que limita a quantidade de consultas disponíveis. Para ultrapassar essa limitação Cafarella et al. (2005) desenvolveram uma arquitetura incorporada no sistema KnowItNow - um sistema de EI projetado para ser rápido e escalável. Ele utiliza um motor de busca próprio chamado *Bindings Engine* (BE) que faz uso de variáveis pré-definidas, tais como *nounphrase* e funções de processamento de caracteres que incluem *head(X)* ou *ProperName(X)*. BE utiliza uma estrutura de índices que mantém no índice os vizinhos à esquerda e à direita de um termo. Esses vizinhos são cadeias de caracteres adjacentes que satisfazem um reconhecedor para uma das variáveis pré-definidas. O processamento caro de um analisador sintático ou processamento sintático raso é desempenhado somente uma vez no momento da construção do índice, não sendo necessário em tempo de consulta.

[1] labs.google.com/sets

2.5.4 OntoSyphon

O OntoSyphon é um sistema de extração de informação conduzido por ontologia (McDowell e Cafarella, 2006, 2008). O OntoSyphon recebe como entrada uma ontologia e usa-a para especificar pesquisas na web que identificam possíveis ocorrências e relacionamentos. Para auxiliar a identificação de ocorrências e relacionamentos, o OntoSyphon também usa os padrões de Hearst (1992).

A combinação dos conceitos da ontologia recebida com os padrões de Hearst gera uma consulta em um motor de pesquisa na web. Para cada par <*ocorrência, conceito*> no conjunto de resultados obtidos, é atribuído um grau de confiança ou probabilidade baseado no número de repetições do par no corpus. Essas repetições também são normalizadas sobre o número de repetições da ocorrência sem o conceito associado, conforme a Equação 2.1, onde *o=ocorrência, c=conceito, p=padrão*.

$$Peso_{normalizado}(o,c) = \frac{\sum_{p \in P} \; conta(o,c,p)}{max(num_resultados(o), Num\_resultados_{25})} \qquad (2.1)$$

A métrica $Peso_{normalizado}$ atribui um valor de confiança para o par (o,c). Por exemplo, na métrica $Peso_{normalizado}$ o numerador contabiliza quantas vezes o par '<Portel,concelho>' está presente no padrão (p) 'é um concelho'. O denominador é dividido em duas partes: a) 'num_resultados(o)' é o número de ocorrências de 'Portel' sozinho na coleção de documentos. b) 'Num_resultados$_{25}$': imagine que exista 100 candidatos para ocorrência de 'concelho'. Ao calcular '(num_resultados(o)' para os 100 candidatos, existem 100 valores. Ordene eles, e então escolha o 25º menor valor e assuma ele como 'Num_resultados$_{25}$'. O resultado do denominador é o valor máximo entre ('(num_resultados(o)' e 'Num_resultados$_{25}$').

A intuição por trás do 'Num_resultados$_{25}$' é que 'o' algumas vezes é muito rara ou uma palavra com erro de grafia. Usando o valor cru 'num_resultados(o)' leva para um resultado do denominador muito alto, pois está-se dividindo por um valor muito pequeno. O 'max' com 'Num_resultados$_{25}$' está na equação para garantir que o valor usado no denominador é pelo menos algum valor mínimo, de

forma que nenhum termo tenha um resultado elevado somente porque ele é muito raro. A métrica $Peso_{normalizado}$ foi uma das métricas usadas por OntoSyphon que gerou melhores resultados.

2.5.5 OnLocus e Endereçamento

O interesse sobre a informação geográfica presente em textos em português tem ganho atenção apenas nos últimos anos. Na variante brasileira da língua os trabalhos de Delboni (2005) e Borges (2006) foram pioneiros. Borges (2006) propôs uma ontologia de lugar (onLocus) que possui conceitos geográficos norteados pela divisão administrativa do Brasil. Onlocus foi mais explorada na parte de endereços postais e telefones no trabalho de Borges (2006).

Numa amostra de 75.413 páginas da web brasileira, Borges (2006) encontrou 57% delas com presença de endereços que foram detectados de acordo com um conjunto de padrões pré-definidos. Os seis principais tipos de padrões utilizados são:

- Telefone
- EndereçoBásico+CidadeEstado+CEP
- EndereçoBásico+Telefone
- EndereçoBásico+CidadeEstado
- EndereçoBásico+CEP
- CEP

Esses padrões foram aplicados à coleção WBR05 (Modesto et al., 2005), extraindo 2.137.601 endereços de 603.798 páginas, o que representa 14,77% do total de páginas dessa coleção.

Delboni (2005) também apresenta um conjunto de expressões de posicionamento desenvolvidas para detectar nomes geográficos em textos em português. Essas expressões são baseadas quatro tipos de relações espaciais: fuzzy (p. ex. perto, depois e acima), direcionais (p. ex. em frente, ao lado, atrás), métricas (p. ex. quilômetros, minutos, quadras) e topológicas (p. ex. dentro de, no

coração de, na praça de alimentação). Os experimentos realizados por Delboni (2005) indicam que as relações direcionais e, principalmente, as métricas são predominantemente utilizadas no contexto de uma expressão de posicionamento (informação geográfica), enquanto os demais tipos de relações são empregados em outros contextos.

2.5.6 Comparação de Sistemas de Extração e Integração de Informação Geográfica

O Sistema de Extração e Integração de Conhecimento Geográfico (SEI-Geo) usa as estratégias propostas em trabalhos anteriores, concentra-se em textos em português, promove a integração do conhecimento adquirido e é aplicado ao domínio geográfico. A abordagem do SEI-Geo é fundamentalmente baseada em padrões e geo-ontologias. A proposta de extrair ocorrências relacionadas de uma sentença, considerando os diversos tipos de relacionamentos existentes na LN é outro fator inovador nessa tese. A explicação detalhada do funcionamento do SEI-Geo e sua avaliação são apresentadas nos Capítulos 5 e 6, respectivamente.

A Tabela 2.2 apresenta uma comparação entre os trabalhos relacionados com o SEI-Geo, no domínio da extração de informação. As características apresentadas em cada coluna não refletem necessariamente limitações dos trabalhos, mas servem principalmente para enquadrar as contribuições dessa tese.

Os critérios utilizados para fazer a comparação entre esses trabalhos mais correlacionados com essa tese são o uso de padrões (PAD) e o uso de ontologias para apoiar a extração de informação (Onto). Todos os trabalhos fazem extração de entidades mencionadas. Além disso, outro fator a considerar é a integração do conhecimento adquirido durante a extração de informação àquele existente (ICA). Caso o sistema faça processamento de informação geográfica (Geo), isso também é indicado na tabela. Finalmente, o último parâmetro de comparação é o fato de os sistemas processarem textos escritos na língua portuguesa (PT).

A maior parte dos trabalhos utilizam padrões léxico-sintáticos, extraem EM e fatos. O sistema OntoLearn faz integração do conhecimento adquirido no WordNet, explorando as definições em LN, os relacionamentos de hiperonímia, meronímia e outros relacionamentos léxico-sintáticos (Navigli e Velardi, 2004).

Tabela 2.2: Comparação entre os trabalhos correlatos.

	PAD	Onto	ICA	Geo	PT
Snowball	✓	x	x	✓	x
OntoLearn	x	✓ (WordNet)	✓	x	x
KnowItAll/KnowItNow	✓	x	x	x	x
OntoSyphon	✓	✓	x	x	x
OnLocus e Endereçamento	✓	✓(ênfase endereços)	x	✓	✓
SEI-Geo	✓	✓	✓	✓	✓

Contudo, os textos utilizados pelo sistema OntoLearn são criteriosamente selecionados de sites de domínio específico.

Delboni (2005) e Borges (2006) usaram um conjunto de padrões para extrair informação geográfica de textos, mas a informação extraída não foi integrada em bases de dados ou ontologias previamente existentes.

2.6 Avaliação de Sistemas de Extração de Informação

As avaliações de sistemas de EI têm sido realizadas desde o final da década de 80, com a primeira *Message Understanding Conference* (MUC) em 1987. Mais recentemente, a *Automatic Content Extraction* (ACE) (Doddington et al., 2004) tem ganho atenção, com a definição de tarefas mais complexas. Por exemplo, a taxonomia das entidades definidas na ACE é mais granular do que na MUC, a interpretação de metonímia, os múltiplos domínios utilizados e as fontes de informação distintas acrescentam maior dificuldade para os sistemas participantes nessa avaliação.

A Tabela 2.3 apresenta as tarefas propostas pela MUC e pela ACE. As tarefas de REM e RCO da MUC foram unidas na tarefa de DDE[1] da ACE, assim como a tarefa de CME e CMR foi unida na tarefa de DDR[2].

Um seguinte exemplo auxilia na explicação de cada uma das tarefas da MUC. 'O largo do Marquês de Pombal foi ponto de encontro dos adeptos após a vitória

[1]Do inglês, *Entity Detection and Tracking*.
[2]Do inglês, *Relation Detection and Tracking*.

Tabela 2.3: Tabela comparativa entre as tarefas propostas para a MUC e a ACE.

MUC	ACE
Reconhecimento de Entidades Mencionadas (REM) Resolução de Co-referências (RCO)	Detecção e Despiste de Entidades (DDE)
Construção do Modelo de Elementos (CME) Construção do Modelo de Relacionamentos (CMR)	Detecção e Despiste de Relacionamentos (DDR)
Produção do Cenário do Modelo (PCM)	Detecção e Caracterização de Eventos (DCE)

de Portugal na última terça-feira. Ele foi primeiro-ministro de Portugal e um dos principais personagens na reconstrução de Lisboa após o terremoto de 1755.'

Na tarefa de REM o sistema reconhece que as EM presentes são: o 'largo do Marquês de Pombal', 'Portugal', 'Lisboa' e 'terremoto de 1755'. A seguir, na RCO, identifica quais entidades e referências (pronomes, por exemplo) se referem para a mesma coisa. No exemplo, reconhece que 'Ele' se refere a 'Marquês de Pombal'.

A tarefa de CME adiciona informação descritiva para os resultados gerados em REM (usando RCO). No exemplo, reconhece que 'Marquês de Pombal' é um 'largo' (no modelo de elementos, Marquês de Pombal preencheria o campo largo, que geralmente são pontos de referência nas cidades). Em seguida, na CMR encontra relacionamentos entre as EM com informação descritiva. No exemplo, reconhece que 'Marquês de Pombal' foi 'primeiro-ministro'.

Finalmente, na PCM, analisa os resultados produzidos na CME e na CMR e reconhece o cenário e as entidades mencionadas envolvidas. No exemplo, reconhece que 'Portugal' venceu um jogo e como consequência os adeptos foram para junto ao 'largo do Marquês de Pombal'.

Na ACE, as tarefas são mais complexas, pois a taxonomia das entidades é mais granular, a interpretação de metonímia (análise semântica dos textos) é levada em consideração e múltiplos domínios de conhecimento e fontes de informação são utilizados. Além disso, todo o software desenvolvido pelo lado da organização é público, ao contrário da MUC. Contudo, a avaliação da ACE não é pública, sendo restrita apenas aos participantes.

2.6.1 Avaliação de Sistemas de Reconhecimento de Entidades Mencionadas em Português - HAREM

A avaliação de sistemas de reconhecimento de entidades mencionadas em português ganhou mais atenção em 2005 com o evento HAREM, que permitiu a comunidade de Processamento de Linguagem Natural ter uma noção da qualidade dos sistemas existentes para essa tarefa. O HAREM desafia os sistemas a reconhecer entidades mencionadas em texto em diversas categorias (p. ex. Pessoa, Local, Tempo), seus tipos (p. ex. Povo, Físico e Tempo_Calend, respectivamente) e sub-tipos no caso de Local e Tempo (p. ex. Ilha, Hora).

Uma das principais diferenças do HAREM para as demais avaliações conjuntas é a identificação e o reconhecimento de EMs em contexto. Por exemplo, na frase 'Brasil condenou a ação da Portugal', ambos os países devem ser reconhecidos como Organização do tipo Administração, e não Local como algumas pessoas consideram. Para uma comparação mais detalhada entre a MUC, a ACE e o HAREM, ver seção 4.3 do Capítulo 4 do livro (Santos e Cardoso, 2007).

No Segundo HAREM, evento que ocorreu em 2008, foram criadas mais duas pistas: Tempo e ReRelEM. A primeira trata do reconhecimento de expressões de tempo nos textos, ao passo que a última desafia os sistemas a identificar relacionamentos entre EMs. O SEI-Geo participou na tarefa ReRelEM identificando relacionamentos de inclusão (p. ex. O Brasil é o país mais emergente da América do Sul. 'Brasil' 'incluído' na 'América do Sul'.) entre locais.

2.7 Metodologias para Construção de Ontologias

Essa seção descreve as principais metodologias para o desenvolvimento de ontologias, cujo objetivo é apresentar os principais processos envolvidos em cada metodologia.

O método de Grüninger e Fox (1995) segue a abordagem dos sistemas baseados em conhecimento utilizando lógica de primeira ordem. Grüninger e Fox (1995) sugerem que no início os principais cenários (possíveis aplicações que utilizarão a ontologia) sejam intuitivamente identificados. Em seguida, um conjunto de questões, denominadas questões de competência informais, são usadas para

determinar o âmbito da ontologia. Nessa fase são definidas as questões que a ontologia deverá suportar para fornecer respostas adequadas às aplicações. A partir dessas questões e suas respostas são extraídos os principais conceitos e suas propriedades, relacionamentos e axiomas da ontologia. Todos esses componentes da ontologia são expressos em lógica de primeira ordem, fazendo desse método muito formal e tirando vantagem robustez da lógica clássica.

O método de Uschold e King (1995) é composto por quatro fases: identificação do propósito da ontologia, construção, avaliação e documentação. Durante a fase de construção, Uschold e King (1995) propõem a coleta de informação, codificação e integração de outras ontologias dentro da ontologia corrente. Três estratégias são propostas para identificar os principais conceitos: *top-down*, *bottom-up* e *middle-out*. A estratégia *middle-out* consiste em identificar os principais conceitos do domínio e, subsequentemente, os conceitos mais genéricos e específicos.

O método Sensus (Swartout et al., 1996) propõe uma estratégia *top-down* para derivar ontologias de domínio específico a partir de ontologias de grande dimensão. Swartout et al. (1996) sugerem a identificação de um conjunto de termos "sementes" que são relevantes para um domínio particular. Esses termos são ligados manualmente a uma ontologia de cobertura ampla (neste caso, a ontologia Sensus, que contém mais de 50.000 conceitos). Em seguida, todos os conceitos no caminho entre os termos semente e a raiz da Sensus são incluídos. Se um termo que poderia ser relevante no domínio não foi selecionado, ele é manualmente adicionado e o passo anterior é executado novamente até que nenhum termo esteja em falta. Finalmente, para aqueles nodos que têm um grande número de caminhos entre eles, a sub-árvore completa daquele nodo na Sensus é adicionada na ontologia de domínio específico. A ideia nessa fase é que se muitos nodos em uma sub-árvore são relevantes, então outros nodos na mesma sub-árvore provavelmente serão relevantes também. Consequentemente, essa abordagem promove o compartilhamento do conhecimento, pois a ontologia base (Sensus) é utilizada para desenvolver ontologias de domínios particulares.

METHONTOLOGY (López et al., 1999) é uma metodologia que permite a criação de ontologias a partir "do nada", reutilizando outras ontologias como elas são ou pelo processo de reconstrução delas. O ciclo de vida é baseado em protótipos que evoluem ao longo do tempo e em técnicas particulares que

realizam cada atividade. O processo de desenvolvimento da ontologia identifica quais tarefas devem ser desempenhadas durante a construção das ontologias (agendamento, controle, garantia de qualidade, especificação, aquisição de conhecimento, conceitualização, integração, formalização, implementação, avaliação, manutenção, documentação e gestão de configuração). O ciclo de vida identifica as fases que a ontologia passa ao longo do seu desenvolvimento bem como as interdependências com o ciclo de vida de outras ontologias. Finalmente, a metodologia especifica as técnicas utilizadas em cada atividade, os produtos que cada atividade gera e como eles têm que serem avaliados.

A metodologia On-To-Knowledge (Staab et al., 2001), assim como o método de Grüninger e Fox (1995), é baseada na análise de cenários de uso para a ontologia. As etapas propostas em On-To-Knowledge são: arranque (*kick-off*), na qual os requisitos da ontologia são capturados e especificados e as questões de competência identificadas, as ontologias com potencial reuso são estudadas e uma versão preliminar da ontologia é construída; refinamento, na qual uma ontologia madura e orientada à aplicação é produzida; avaliação, na qual os requisitos e as questões de competência são avaliados e a ontologia é testada no ambiente da aplicação; finalmente, a última etapa é a manutenção da ontologia.

As metodologias apresentadas nessa seção são comparadas em Corcho et al. (2003), onde os autores concluíram que nenhuma das metodologias estava suficientemente madura comparadas com metodologias de Engenharia de Software e Engenharia de Conhecimento. Outra conclusão de Corcho et al. (2003) é que as metodologias não são unificadas, sendo que cada grupo de pesquisa segue sua própria abordagem. Essa unificação não é encontrada ainda hoje.

Apesar de as metodologias apresentadas acima poderem ser úteis para construção de geo-ontologias, nenhuma delas foi aplicada para esse domínio. No que diz respeito a metodologias para construção de geo-ontologias existe uma carência na literatura, comparado com a quantidade de metodologias genéricas existentes.

2.7.1 Metodologias para Construção de Geo-Ontologias

Fu et al. (2005) construíram uma geo-ontologia a partir dos requisitos de quatro aplicações: interface do usuário, extração de metadados, sistema de ordenação de

resultados de um motor de busca e o motor de busca. As fontes de informação são o *Seamless Administrative Boundaries of Europe* (SABE) e o TGN. A integração de múltiplas bases de dados através de quatro medidas de similaridade baseadas em emparelhamentos de (1) nomes de locais, (2) conceitos, (3) hierarquias e (4) coordenadas geográficas. Os resultados dos dez experimentos realizados por Fu et al. (2005), os quais combinam as estratégias de similaridade, revelaram que a combinação entre emparelhamentos de nomes de locais e hierarquias apresentou melhores resultados. O emparelhamento de nomes de conceitos não é uma medida significante e a similaridade entre coordenadas geográficas pode ser utilizada para complementar os resultados de (1) e (2). A ontologia construída por Fu et al. (2005) não está publicamente disponível e foi utilizada somente no âmbito do projeto *Spatially-Aware Information Retrieval on the Internet*(SPIRIT - www.geo-spirit.org).

Outros recursos com informação geográfica que estão sendo amplamente utilizados são o TGN, já mencionado na Seção 2.3.1, e o GeoNames (http://www.geonames.org). As fontes de informação que alimentam o TGN são bases de dados privadas, cujos dados já estão curados. O processo de integração de informação usa regras para emparelhar nomes de locais, conceitos e coordenadas geográficas. Os dados são fornecidos no formato XML e numa base de dados relacional.

O GeoNames é uma base de dados de nomes geográficos pública e gratuita e também um dos recursos mais abrangentes em termos de quantidade de informação (mais de 8 milhões de nomes) atualmente. Suas fontes de informação mais importantes são: a *National Geospatial-Intelligence Agency's* (NGA) e o *U. S. Board on Geographic Names*, do qual são provenientes a maior parte dos nomes, exceto para EUA e Canadá, o *U. S. Geological Survey Geographic Names Information System* (GNIS) e o sítio www.geobase.ca, o qual fornece os nomes geográficos do Canadá. O processo de criação dessa base de dados utiliza somente fontes públicas que possam ser acessadas livremente quando geradas no formato de geo-ontologia. A fase de limpeza de dados não é proposta nessa metodologia e a integração de dados é realizada principalmente através de similaridade entre nomes de locais. Os formatos disponíveis da geo-ontologia são XML, RDF e OWL.

2.7.2 Comparação entre Metodologias para Construção de Geo-Ontologias

As metodologias para construção de ontologias já foram comparadas em Corcho et al. (2003) e essa seção apresenta uma comparação entre metodologias para construção de geo-ontologias. A Tabela 2.4 apresenta o enquadramento dessas metodologias aos critérios estabelecidos na primeira coluna. GKB é a referência para a metodologia proposta nessa tese, a qual está descrita nos próximos capítulos e sintetizada no Capítulo 7.

Quanto à limpeza de dados, o TGN descreve as directrizes usadas para tratar os dados, mas não os procedimentos seguidos internamente (conforme o documento "*TGN2 General Guidelines*", no sítio do instituto Getty (Paul, 2009)). O SPIRIT não propõe estratégias para limpeza de dados assim como o GeoNames. Por outro lado, na GKB descreve em detalhe todos os processos envolvidos nessa fase.

A integração de conhecimento no SPIRIT e no GeoNames é fortemente baseada em similaridade (uso da distância de Levenshtein (1966), por exemplo), enquanto TGN e GKB utilizam uma abordagem baseada em regras. O SPIRIT e a GKB são as metodologias que apresentam detalhes de como o processo de integração é realizado, enquanto o TGN e o GeoNames não o explicitam.

Os formatos utilizados nas quatro metodologias são facilmente importados por aplicações que processem XML (a maior parte, atualmente). Todos as metodologias utilizam dados multi-língua, exceto a metodologia proposta pelo SPIRIT, cujos dados estão somente em inglês.

A respeito do controle de versões, o TGN lança novas versões anualmente em arquivos licenciados e no sítio da web a cada mês, mas o SPIRIT não

Tabela 2.4: Tabela comparativa entre metodologias para construção de geo-ontologias.

	TGN	SPIRIT	GeoNames	GKB
Limpeza de dados	proposta	não proposta	não proposta	descrita em detalhe
Integração de conhecimento	regras	similaridade	similaridade	regras
Formato	XML	XML, RDF e OWL	XML, RDF e OWL	XML, RDF e OWL
Multi-língua	sim	não	sim	sim
Versionamento	mês/ano	N/D	variável	variável
Documentação	informal	formal	informal	formal

apresenta tal política. A periodicidade das versões do GeoNames é variável e, até esta data, foram lançadas seis versões. A primeira versão da geo-ontologia foi disponibilizada no dia 14 de outubro de 2006 e a última versão dia 2 de abril de 2007. Contudo, o GeoNames possui uma atualização diária das novas ocorrências que são adicionadas à base de dados. A GKB lança uma nova versão somente quando há alteração no modelo de dados (TBox).

Finalmente, um critério fundamental quando um Gestor deseja adotar uma metodologia, é a documentação. O TGN disponibiliza suas diretrizes em linguagem natural numa página da web. O SPIRIT formaliza alguns processos através de equações (p. ex. medidas de de similaridade) descrevendo detalhes da metodologia. Já o GeoNames possui alguns detalhes no seu próprio sítio e também pode-se encontrar apresentações e entrevistas do seu Gestor, podendo-se considerar uma documentação informal. A GKB apresenta seus processos descritos detalhadamente através de algoritmos, diagramas UML e regras em Lógicas de Descrição.

2.8 Discussão e Conclusões

As estruturas de representação de conhecimento descritas nesse capítulo evidenciam a diversidade e o grau de formalidade com que pode ser representado o conhecimento humano. A estrutura na forma de ontologia parece ser a mais adequada para ser adotada nesse trabalho, uma vez que permite a sistemas que utilizam informação geográfica manipular essa informação facilmente.

No que refere a extração e integração de informação é possível perceber que os sistemas ainda fazem pouco reuso de conhecimento, um dos princípios básicos de Engenharia de Software. Tal fato talvez seja uma das causas de os mesmos não expandirem o conhecimento presente nas estruturas de representação de conhecimento disponíveis.

Especificamente, na extração de informação geográfica, alguns sistemas usam almanaques nas fases de identificação e reconhecimento de locais. Contudo, nenhum deles faz a expansão do conhecimento existente nos almanaques com a informação nova reconhecida em textos.

As metodologias para construção de ontologias apresentam processos comuns mas variam na forma como são implementadas. As metodologias genéricas não foram aplicadas ao domínio geográfico, enquanto as metodologias específicas para esse domínio raramente detalham os processos recomendados para o desenvolvimento de geo-ontologias.

Este capítulo apresentou os principais conceitos envolvidos no tema dessa tese, bem como diversas estruturas de representação de conhecimento geográfico, juntamente com uma comparação entre as mesmas. Em seguida, sistemas que realizam extração e integração de informação geográfica foram descritos e comparados, de modo a enquadrar melhor o SEI-Geo. Na sequência, as principais avaliações de sistemas de EI foram mencionadas.

Os processos mais relevantes das metodologias para construção de ontologias foram apresentados fornecendo uma visão panorâmica sobre as atividades envolvidas no desenvolvimento de ontologias. A seguir, uma comparação entre as principais metodologias para construção de geo-ontologias foi descrita.

O próximo capítulo descreve a metodologia utilizada para a construção de uma base de conhecimento geográfica (GKB), bem como as geo-ontologias geradas a partir da conhecimento integrado na GKB e algumas das aplicações que as utilizam.

Capítulo 3

Uma Metodologia para a Construção de uma Base de Conhecimento Geográfico

3.1 Introdução

Este capítulo apresenta uma metodologia para recolher e manter conhecimento geográfico. O conhecimento pode ser exportado sob a forma de ontologias para aplicações da Web Semântica (WS). Os métodos propostos suportam a integração semântica de dados geográficos coletados de fontes de informação heterogêneas. São genéricos e a aplicação dos mesmos para outras geo-ontologias pode ser facilmente replicada.

O número de aplicações utilizando informação geográfica tem aumentado consideravelmente nos últimos anos. Tais aplicações incluem sistemas de informação geográfica, reconhecedores de entidades mencionadas e motores de busca geográficos, entre outras. Para suportar o uso de informação geográfica dessas aplicações, foi desenvolvida a Geographic Knowledge Base (GKB). A GKB integra dados e conhecimento provenientes de múltiplas fontes sob um esquema comum e cria um ambiente para derivar conhecimento e gerar geo-ontologias a partir da informação disponível.

A GKB mantém informação geográfica sobre entidades geo-administrativas e geo-físicas e também sobre os atributos geográficos de locais virtuais, tais como

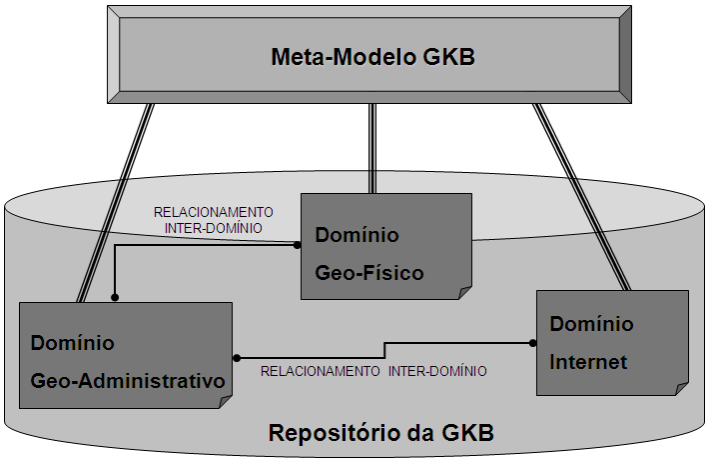


Figura 3.1: Arquitetura de informação da GKB.

sítios da web e domínios da internet.

O conhecimento presente na GKB é representado em Lógicas de Descrição, o formalismo adotado por aplicações da WS para esse propósito (Baader et al., 2003).

Além de um repositório comum, GKB inclui dois conjuntos de ferramentas:

Conversores: carregam dados de diversas fontes e realizam procedimentos de normalização para manter uma visão unificada de toda a informação.

Geradores: criam geo-ontologias, seguindo o padrão OWL (*Web Ontology Language*) (McGuinness e van Harmelen, 2004).

A GKB foi utilizada para criar e manter duas bases de conhecimento geográfico: a primeira restrita a Portugal e a segunda em nível mundial com informação em quatro línguas (*World Geographic Ontology* - WGO). Essas línguas foram escolhidas em razão das línguas utilizadas nos documentos do GeoCLEF. Contudo, o modelo de dados da GKB suporta a extensão para qualquer língua.

3.2 Projeto Conceitual da GKB

A Figura 3.1 apresenta a arquitetura da GKB. Os dados são organizados em

domínios de informação, cada um representando um conjunto de características relacionadas. Atualmente, existem três domínios de informação presentes na GKB: geo-administrativo (com informação sobre a geografia administrativa), geo-físico (com informação sobre a geografia física) e internet (com informação sobre endereços virtuais da web). A informação em cada domínio é estruturada de forma semelhante, sendo todos os domínios representados em termos de um meta-modelo comum.

A GKB é um repositório de conhecimento geográfico que contém informações provenientes de fontes de dados administrativas semi-estruturadas de autoridades junto com um conjunto de regras para integração de informação no repositório.

A GKB suporta a definição de relacionamentos ontológicos entre entidades de cada domínio. No domínio geográfico, a GKB fornece relacionamentos de meronímia, sinonímia e adjacência, entre outros. A GKB também suporta relacionamentos inter-domínios, os quais são associações entre entidades de domínios diferentes. Por exemplo, o âmbito geográfico de uma entidade do domínio de rede é representado como um relacionamento entre um sítio da web (entidade do domínio de rede) e uma região geográfica (uma entidade do domínio administrativo).

3.2.1 Classes de Fontes de Informação

A qualidade da informação presente em um repositório de conhecimento é dependente de suas fontes de informação. A seleção das fontes de informação depende do âmbito da ontologia a produzir como resultado da exportação do conhecimento armazenado no repositório. Nessa tese, eu desenvolvi duas geo-ontologias, uma de âmbito nacional (Geo-Net-PT) e outra de âmbito mundial (WGO).

A GKB integra dados provenientes de várias classes de fontes de informação. No domínio geográfico, são suportadas fontes das seguintes classes:

Administrativa: contém dados demográficos e administrativos, tal como as divisões territoriais. Para Portugal, esse tipo de informação é fornecido pelo Instituto Nacional de Estatística (INE) e também na enciclopédia on-line Wikipedia (http://pt.wikipedia.org). A Associação Nacional de

Municípios Portugueses (ANMP) fornece os relacionamentos de adjacência entre distritos e concelhos, nos seus respectivos níveis, ou seja, cada distrito é adjacente somente à outro distrito e não a concelhos. Para a WGO, a Wikipedia também foi utilizada como fonte de informação.

Postal: inclui informação usada para identificar endereços. Para Portugal, a base dados dos Correios, Telégrafos e Telefones (CTT) foi utilizada. Cada código-postal dessa fonte de informação foi associado com a área administrativa correspondente.

Almanaque: fornece coordenadas geográficas para as principais regiões de todos os países do mundo. O almanaque utilizado em 2004, na fase de carregamento de dados à GKB foi o fornecido pelo sítio `calle.com`. Atualmente, esse sítio foi transformado num sítio para reservas de hotéis, mas permanece usando o mesmo almanaque para suportar o sistema. Os dados carregados na GKB foram as coordenadas geográficas para as principais regiões de Portugal.

Física: contém dados sobre toda a geografia física. Para Portugal as fontes de informação utilizadas são: Instituto Geográfico do Exército (IGeoE), Instituto Geográfico Português (IGP), Instituto da Água (IA) e Instituto de Pesquisa da Marinha (IMAR). Para a WGO, a Wikipedia foi utilizada como fonte de informação.

Para obter dados sobre o domínio internet, podem ser usadas as seguintes classes:

Domínios da internet: dados com domínios da internet - *Domain Name System* (DNS). Para Portugal, foi usado banco de dados da Fundação para a Computação Científica Nacional (FCCN), a entidade responsável por gerenciar os domínios de nível de topo (TLD).

Sítios da web: endereços de sítios da web com seus endereços *Internet Protocol* (IP). Para Portugal, esta informação foi obtida do repositório de metadados da web Versus (Gomes et al., 2002), pertencente ao motor de pesquisa tumba!.

3.2.2 Modelo de Informação

O processo de modelagem descrito nessa tese segue as diretivas do *Object Management Group* (OMG)[1]. A GKB distingue para cada nome a entidade que representa. A noção de entidade usada nessa tese é a definida na norma ISO19109 (2006):

> um objeto com significado no domínio de discurso.

No domínio geográfico, países, cidades e municípios são exemplos de tais objetos. Na GKB, entidades e seus nomes são classes distintas e cada entidade está associada a um tipo de entidade. Por exemplo, a entidade 'concelho de Faro' está associada ao tipo de entidade 'concelho' e ao nome 'Faro'.

Como na ISO 19109, as entidades são classificadas segundo tipos de entidades com base em conjuntos de características ou propriedades comuns. Esta abordagem capacita a GKB a suportar relacionamentos 'vários para um' entre nomes e entidades. Esta flexibilidade também permite a incorporação de novos tipos de dados.

Um modelo é uma abstração de um fenômeno no mundo real, e um meta-modelo é ainda outra abstração, que serve para caracterizar conceitos utilizados na definição do modelo. A Figura 3.2 apresenta o meta-modelo base da GKB, o qual é suficientemente genérico para representar informação de qualquer domínio. O meta-modelo base constitui-se das principais classes usadas para modelar qualquer domínio de conhecimento que seja representado na GKB, ou seja, é o núcleo do modelo.

Uma entidade geográfica *Feature* é composta por um nome *Name* e um tipo *Type*. A classe *Feature* é associada com a classe *Type* (p. ex. a entidade `Douro` é um tipo de `rio`). A classe *Relationship-Type* captura relacionamentos suportados entre tipos e entidades (p. ex. `parte de` e `adjacência` entre outros de natureza geográfica). A classe *Type-Relationship* armazena os relacionamentos entre tipos (p. ex. um concelho é `parte de` um distrito). A classe *Feature-Relationship* armazena os relacionamentos entre entidades (p. ex. Sintra (concelho) é `parte de` Lisboa (distrito)).

[1] www.omg.org/

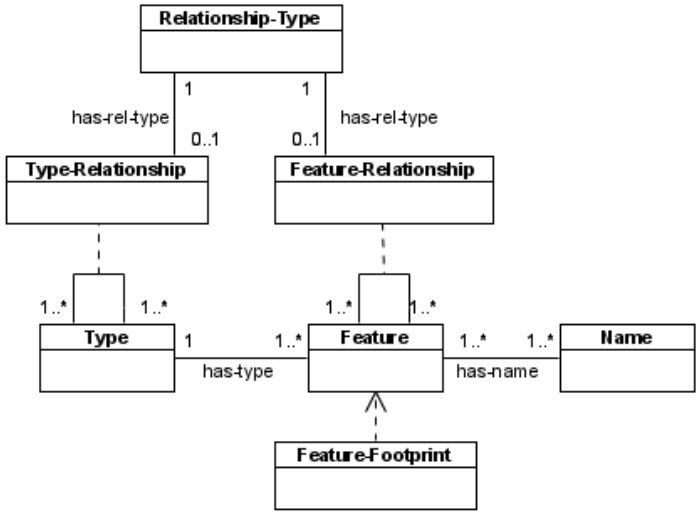


Figura 3.2: Meta-modelo base da informação na GKB.

As entidades podem ser especializadas por *Feature-Footprint*, a qual captura as coordenadas geográficas. As coordenadas podem identificar centroides, caixas limitadoras ou polígonos. A entidade `Serra da Estrela`, por exemplo, tem o centroide em `40°20N, 7°38W`.

3.2.2.1 Representação de Atributos e Nomes

A Figura 3.3 representa o modelo base da Figura 3.2 estendido com as classes usadas para representar atributos de tipos, entidades e nomes. Diferentes tipos geográficos têm diferentes atributos. Uma `cidade` tem `população`, um `rio` tem uma `nascente` e uma `montanha` tem uma `altitude`, por exemplo. As classes *Type-Attribute*, *Feature-Attribute* e *Name-Attribute* adicionam conjuntos de propriedades às classes *Type*, *Feature* e *Name*, respectivamente, do modelo base.

Cada nome capturado na classe *Name* está associado com uma língua e um código de país associado a ele (p. ex. PT-BR). O código da língua adotado segue o padrão de etiquetas de língua <código da língua“–”código do país> definido pela RFC 3066 da IETF (Alvestrand, 2001). Os nomes podem ser estendidos com conjuntos de atributos capturando preferências (p. ex. um nome é preferido

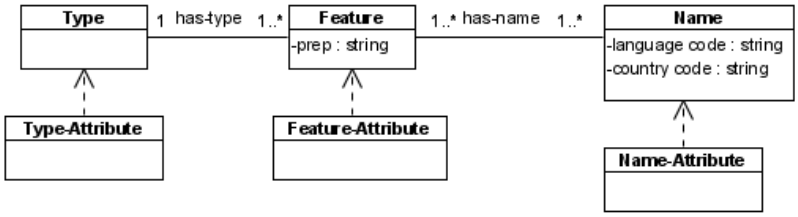


Figura 3.3: Representação de nomes e atributos na GKB.

ou alternativo), tempo (p. ex. histórico), uso (p. ex. raro, coloquial) e gentílicos. Esses atributos são armazenados na classe *Name-Attribute*.

Por exemplo, o rio `Tajo` tem seu equivalente em português `Tejo` e seu nome histórico `Tagus`, do Latin. O atributo histórico é capturado na classe *Name-Attribute*, enquanto a classe *Name* armazena os nomes (`Tajo`, `Tejo` e `Tagus`), as línguas (ES, PT e LA) e os códigos de país (ES e PT), respectivamente.

Finalmente, a classe *Feature* também armazena no atributo *prep* as palavras de ligação (p. ex. 'de', 'da' e 'dos') que compõem os nomes das entidades geográficas. Essas palavras de ligação são bastante relevantes para sistemas de REM, por exemplo, uma vez que auxiliam na identificação de nomes de locais em textos.

3.2.2.2 Relacionamentos Inter-Domínio

Relacionamentos inter-domínio são conexões entre diferentes domínios modelados na GKB. A Figura 3.4 apresenta o modelo representando os relacionamentos inter-domínio.

A informação armazenada nas classes apresentadas na Figura 3.4 está assim distribuída: a classe *Adm-Feature* contém informação do domínio administrativo, a classe *Phy-Feature* armazena informação do domínio físico e a classe *Net-Feature* captura informação do domínio internet.

Todos os tipos de relacionamentos inter-domínio são armazenados na classe *ID-Type-Relationship* (p. ex. `parte-de` e `adjacência`). As classes *ID-Feature-Relationship-Adm-Phy* e *ID-Feature-Relationship-Phy-Adm* capturam os relacionamentos entre entidades do domínio administrativo e físico. Quando uma entidade do domínio administrativo é parte-de uma entidade do domínio físico

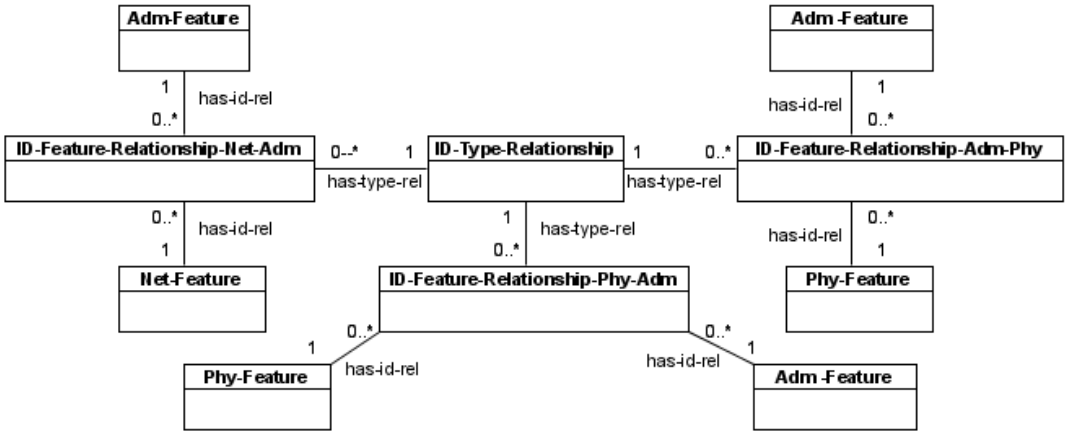


Figura 3.4: Relacionamentos inter-domínio na GKB.

o relacionamento é capturado na classe *ID-Feature-Relationship-Adm-Phy*, caso contrário utiliza-se a classe *ID-Feature-Relationship-Phy-Adm*. Relacionamentos do tipo adjacência podem ser armazenados em qualquer uma das classes. Um exemplo de relacionamento inter-domínio é: os concelhos de Lisboa e Setúbal (domínio administrativo) são `adjacentes` ao rio Tejo (domínio físico). Outros relacionamentos tais como `cruza`, `toca` e `intersecta` estão implícitos nas coordenadas dos dados e não são modelados na GKB.

Por outro lado, quando os domínios envolvidos são o administrativo e o de internet, os relacionamentos são capturados na classe *ID-Feature-Relationship-Adm-Net*. Esses relacionamentos permitem tornar explícito o âmbito geográfico de sítios e domínios da internet. A classe *ID-Feature-Relationship-Adm-Net* armazena os identificadores das entidades do domínio internet e do domínio geográfico administrativo com o relacionamento `tem-âmbito`.

3.2.2.3 Procedência dos Dados

Um dos requisitos da GKB é suportar o rastreamento da informação, ou seja, possibilitar a uma aplicação encontrar a proveniência dos dados. Tão importante quanto o dado em si, é a fonte de informação da qual esse dado foi extraído. No modelo da GKB as fontes de informação estão distribuídas por todo o modelo, conforme mostra a Figura 3.5. As fontes de informação são independentemente

associadas para cada atributo e relacionamento individual no modelo, bem como para cada associação entre entidade e seus nomes e tipos.

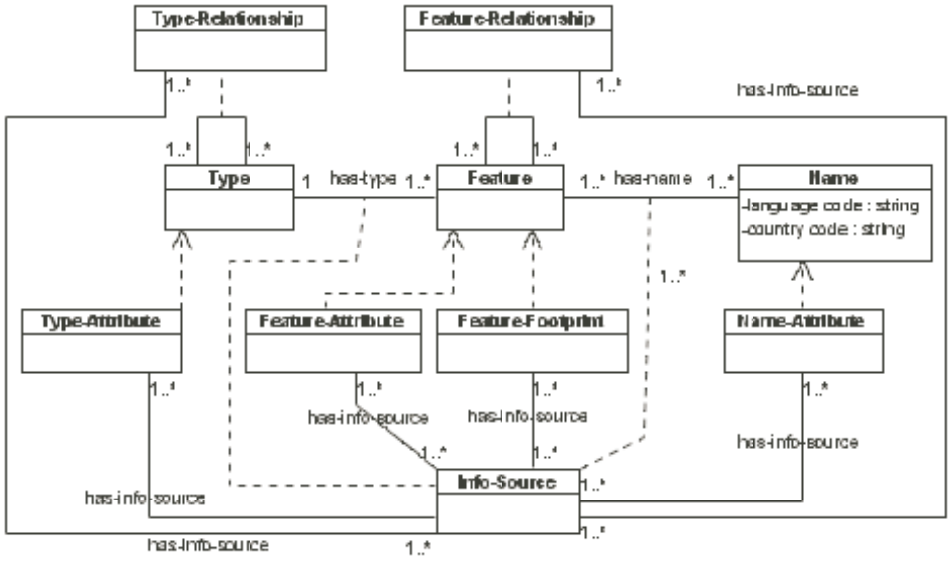


Figura 3.5: Modelo da distribuição das fontes de informação na GKB.

Esta abordagem também permite que as aplicações que usam a GKB façam raciocínio baseadas na credibilidade da informação, atribuindo pesos para cada parte dos dados baseado no nível de autoridade da fonte de informação associada. Por exemplo, um GIS pode necessitar utilizar apenas o conhecimento associado a coordenadas fornecidas por determinadas autoridades, especificamente, autoridades do Estado.

3.3 Integração de Dados e de Conhecimento

As fontes de dados possíveis de serem utilizadas pela GKB são desenvolvidas e mantidas independentemente com o objetivo de servir necessidades diferentes. Estes fatos originam redundância e uma grande heterogeneidade em termos de modelo de informação. Algumas fontes são complementares a outras, fornecendo informação adicional sobre uma entidade geográfica. Assim, a informação duplicada tem que ser eliminada e a informação complementar deve ser consolidada

para se alcançar uma visão consistente das entidades do mundo real. Sempre que uma nova fonte de informação é carregada na GKB, é realizado um procedimento para detectar se as novas entidades já estão definidas. Nesse caso, somente novos nomes e relacionamentos são adicionados ao conhecimento existente na GKB.

3.3.1 Limpeza de Dados

O processo de limpeza de dados é essencial para construir uma base de conhecimento consistente. Esse processo geralmente é realizado em três fases, conhecidas como Extração, Transformação e Carregamento (em inglês, ETL (*Extraction*, *Transformation* e *Loading*) (Rahm e Do, 2000). Rahm e Do (2000) classificam os problemas de limpeza de dados como problemas de fonte única e de múltiplas fontes. Na limpeza de dados geográficos, eu encontrei problemas em ambas as classes, os quais são detalhados a seguir.

3.3.1.1 Fonte Única

Os problemas mais comuns na limpeza dos dados de uma fonte são:

Erros de grafia: são inevitáveis em grandes fontes de informação contendo dados inseridos por humanos. A maior parte das fontes de informação usadas nas geo-ontologias produzidas com a GKB são curadas, mas erros ainda são comuns. A remoção de todos os erros de grafia é uma tarefa impossível. Quando detectados, tais erros são eliminados, mas alguns sempre permanecerão.

Códigos postais inválidos: quem registra domínios na internet frequentemente insere códigos postais inválidos nas bases de dados do domínio internet. Eu detectei eles quando pesquisei um código postal e não podemos validar ele. Os scripts da GKB podem ocasionalmente detectar e corrigir algum deles com o seguinte procedimento:

- identificar sequências de dígitos em campos de códigos-postais nas fontes de dados que estão sendo carregadas;

- converter os dígitos para o formato de códigos-postais (em Portugal, 4+3 dígitos, como em "1250-212");
- os dígitos são considerados um código postal válido se o código obtido é emparelhado com um código-postal proveniente da base de dados do CTT;

Inserção de nomes alternativos: dados recebidos de fontes do tipo almanaque contêm nomes de localidades com caracteres com e sem acento como alternativas. Quando a fonte de dados é o português, eu considero apenas os nomes com caracteres acentuados e assumimos que os outros caracteres representam alternativas para codificação de caracteres que não suportam acentos. Entretanto, geralmente é comum encontrar no almanaque nomes alternativos de locais. Por exemplo, *São João*, localizado em *Viana do Castelo*, tem os seguintes nomes alternativos: *Vila Chã* e *São João Baptista*. Esses nomes são armazenados como nomes alternativos associados ao nome preferido com um relacionamento de equivalência.

Correção de coordenadas geográficas: num almanaque, uma região é algumas vezes associada com mais de uma coordenada geográfica. Nesses casos, a abordagem seguida foi utilizar a média das coordenadas atribuídas.

3.3.1.2 Múltiplas Fontes

Eu também encontrei inconsistências quando integrando dados de múltiplas fontes de informação. Em geral, as fontes de dados geográficos organizam a informação de modos diferentes e, assim, é necessário lidar com as heterogeneidades estruturais também. Para resolver algumas dessas inconsistências, é possível atribuir um nível de autoridade a cada fonte de informação a integrar na GKB e usar essa autoridade para resolver as inconsistências nos dados. Por exemplo, aquando da construção da geo-ontologia de Portugal, ao emparelhar dados dos CTT com o almanaque `calle.com`, o programa que faz a integração de informação encontrou 11 distritos nos dados dos CTT, abrangendo duas regiões autônomas (Açores e Madeira) no almanaque. O programa atribuiu para os locais no almanaque o

correspondente nome do distrito nos dados dos CTT, considerando os CTT como uma autoridade mais importante.

Contudo, o problema geralmente é mais complexo, quando as fontes de informação não podem ser exaustivas ou autoritárias. Nessas situações, a GKB pode manter toda a informação recebida para carregamento, deixando para as aplicações consumidoras a possibilidade de rastrear a origem dos dados e suas fontes e tomar a decisão final sobre sua validade.

3.3.2 Normalização de Dados

Os nomes de EG devem ser escritos corretamente nas geo-ontologias, com a primeira letra de cada palavra, exceto preposições, em maiúscula. Apesar de essa tarefa parecer inicialmente simples e fácil, existe um conjunto de casos que são mais complexos de se encontrar uma solução, os quais são descritos a seguir:

- **Números romanos:** caracteres identificados como números romanos devem ser colocados em letra maiúscula. Solução: definição de duas tabelas de dispersão: uma com caracteres romanos e outra com exceções (p. ex. 'civil'). Todos os nomes que estão na primeira tabela e não estão na segunda são colocados em letras maiúsculas.

- **Preposições:** algumas preposições assumem a categoria gramatical de substantivo e, nesses casos, a primeira letra precisa ser colocada em maiúscula (p. ex. *Entre-Campos*). Solução: sempre que uma preposição aparece no início de um nome, sua primeira letra é capitalizada. Contudo, em expressões como 'entre As Ruas Alfredo Pinto e Alfredo Feio', a palavra 'entre' é uma preposição. Como eu não tenho nenhuma ferramenta para ajudar nesta tarefa, ainda é possível encontrar preposições em maiúscula.

- **Artigos:** alguns artigos devem ser colocadas em maiúsculas (p. ex. 'O' em Jornal 'O Povo de Cortega'). Solução: todos os artigos são colocados em maiúscula quando precedidos das palavras 'Jornal' e 'Revista' e no início de um nome (p. ex. 'O Algarve').

- **Caracteres Especiais:** aspas, parênteses e outros caracteres especiais devem ser considerados durante o processo de normalização de dados. Solução: a primeira letra da palavra após o caracter especial é colocada em maiúscula, pois esses caracteres não são seguidos por preposições nas bases de dados trabalhadas nessa tese.

- **Apóstrofes:** ocasionalmente, alguns nomes começam com a letra 'd' seguida pelo apóstrofe. Solução: a letra 'd' não é colocada em maiúscula, mas o primeiro caracter depois do apóstrofe é colocado em maiúscula.

- **Pontos:** letras seguidas por pontos nem sempre são fornecidas em maiúscula pelas fontes de informação: Solução: quando uma palavra é composta por uma letra e esta letra é seguida por um ponto, a mesma é colocada em maiúscula. Essa regra é utilizada automaticamente para colocar em maiúsculo os acrônimos que são digitados com pontos (p. ex.: 'A.E.P.' - Associação Empresarial de Portugal).

- **Acrônimos:** quando os acrônimos são digitados sem pontos separando as letras, esses não são identificados. Solução: criação de uma tabela de dispersão com os acrônimos mais comuns (p. ex. CP, CTT, EDP). As palavras nessa tabela de dispersão são sempre colocadas em maiúscula. Entretanto, essa solução não é exaustiva.

Essa seção descreveu os principais problemas relativos à normalização de dados no que tange aos nomes de EG. Apesar de os métodos funcionarem adequadamente para os casos descritos acima, ainda é possível encontrar caracteres em minúscula ao invés de maiúscula e vice-versa.

3.3.3 Integração de Conhecimento na GKB

A GKB recebe informação de múltiplas fontes, cada uma com conhecimento organizado diferentemente e representando a informação geográfica em diferentes níveis de abstração. Algumas fontes fornecem informação apenas sobre as principais regiões de um país, enquanto outras incluem entidades ao nível de arruamento e código-postal. Neste contexto, é necessário lidar com o conhecimento geográfico

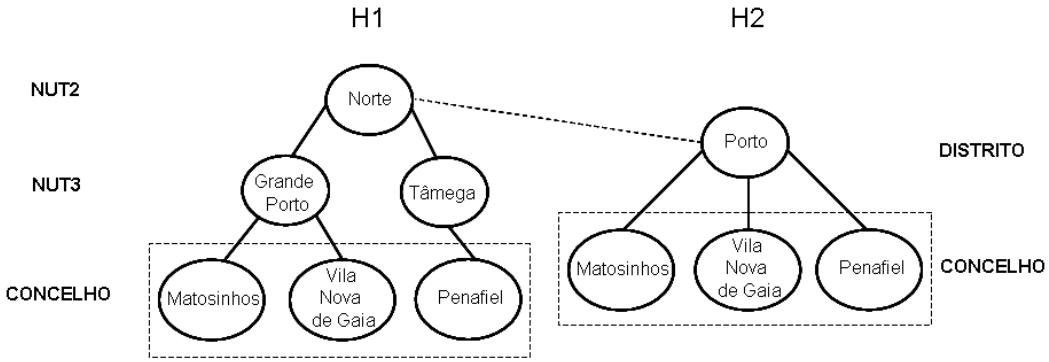


Figura 3.6: Hierarquias de diferentes fontes de informação na GKB.

de modo consistente. A Figura 3.6 apresenta um exemplo concreto de uma situação onde é necessário aplicar um procedimento de união de hierarquias na GKB. A hierarquia `H1` está carregada na GKB e outra hierarquia `H2` para ser carregada. `H1` tem três regiões de Portugal (na GKB, três tipos de entidade): duas NUT (Nomenclatura de Unidade Territorial) e um tipo mais específico (concelho). `H2` possui duas regiões de Portugal: `distrito` e `concelho`.

O algoritmo faz a união de hierarquias através dos seguintes passos (os exemplos dados entre parênteses referem-se à Figura 3.6): Em primeiro lugar, ele procura os tipos de entidades comuns no nível mais específico em ambas as hierarquias (concelho). Se encontra, ele identifica as ocorrências comuns entre as hierarquias (*Matosinhos*, *Vila Nova de Gaia* e *Penafiel*). Após as ocorrências comuns serem identificadas, o algoritmo sobe a hierarquia e procura o nó pai comum no nível mais baixo (*Norte* em `H1` e *Porto* em `H2`). Após esses passos, o algoritmo verifica a distância (em número de relacionamentos `parte-de`) entre as ocorrências comuns dos tipos de entidades e seus pais. O pai (`Porto`), que possui a menor distância até as ocorrências comuns, é unido através do relacionamento `parte-de` com o pai (`Norte`) na outra hierarquia. Os relacionamentos existentes em ambas as hierarquias são mantidos. A Figura 3.7 apresenta as hierarquias unidas.

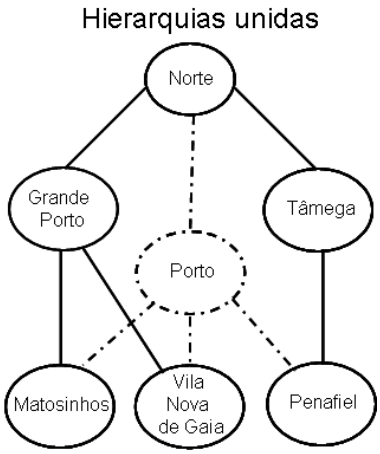


Figura 3.7: Hierarquias unidas na GKB.

3.3.4 Usando o Conhecimento Geográfico na GKB

A GKB gerencia além de entidades geográficas e relacionamentos, regras para integração de conhecimento. As regras podem ser adicionadas manualmente e usadas por programas da GKB para verificar regras de integridade do domínio de informação geográfica em tratamento na GKB e gerar novos relacionamentos. Para gerar relacionamentos, a GKB recebe dados geográficos e regras para produzir novos relacionamentos para serem adicionados à base de dados relacional.

Em geral, o nome dado a uma entidade é representado em diferentes modos, dependendo do domínio de informação considerado. Por exemplo, os nomes podem ser compostos por múltiplas palavras. Nos domínios geográficos, um caracter de espaço é o separador, mas no domínio internet esse caracter é inválido nos *Uniform Resource Locator* (URL).

A Figura 3.8 apresenta um extrato da descrição de mundo da GKB (ABox em Lógicas de Descrição). A descrição de mundo é composta por diferentes representações de nomes geográficos. Nomes de URL são usados no formato original, apenas decompostos pela divisão de domínio correspondente. Um nome geográfico codificado em uma URL não possui espaços, pode ter hífens substituindo eles ou ainda pode não ter preposições em seu nome. As diferentes representações do nome `Santiago do Cacém` (ver os valores do conceito atômico `geoFeatureName`) ilustram como eu represento o conhecimento geográfico em

```
geoFeatureName(270,‘‘santiagodocacem’’).
geoFeatureName(270,‘‘santiagocacem’’).
geoFeatureName(270,‘‘santiago-do-cacem’’).
geoFeatureName(270,‘‘santiago-cacem’’).
geoFeatureType(270,‘‘CON’’).
netSiteSubDomain(33684,‘‘www’’).
netSitePrefix(33684,‘‘cm’’).
netSiteDomainToken(33684,‘‘santiago-do-cacem’’).
netSiteTLD(33684,‘‘pt’’).
```

Figura 3.8: ABox em Lógicas de Descrição para a cidade de “Santiago do Cacém” (o valores numéricos 270 e 33684 correspondem aos identificadores da entidade da ocorrência na GKB).

Lógicas de Descrição. O valor do conceito atômico `geoFeatureType` corresponde ao tipo de entidade geográfica do nome e *270* é o identificador da entidade.

Para o domínio de rede, eu represento a URL dos sítios divididas em três conceitos atômicos: subdomínio, domínio e domínio de nível de topo. Além disso, eu crio o conceito atômico `netSitePrefix`, que indica o prefixo a ser usado em uma regra. Por exemplo, www.cm-santiago-do-cacem.pt é codificado como:

- `netSiteSubDomain(33684,‘‘www’’)`,
- `netSitePrefix(33684,‘‘cm’’)`,
- `netSiteDomainToken(33684,‘‘santiago-do-cacem’’)` e
- `netSiteTLD(33684,‘‘pt’’)`

onde `33684` é o identificador da entidade.

O novo conhecimento é incorporado na GKB através de regras, descritas no TBox (descrição da terminologia em Lógicas de Descrição). Em Portugal, muitos dos sítios da web das câmaras municipais estão hospedados com domínios cujos nomes contêm os prefixos “cm-” or “mun-”. Eu expresso esse conhecimento pela seguinte regra:

Concelhos: `hasScope(idN,idG)` $\equiv$ `∃netSiteDomainToken(idN,X)` $\sqcap$
`(∃netSitePrefix(idN,‘‘cm’’)` $\sqcup$ `∃netSitePrefix(idN,‘‘mun’’))` $\sqcap$
`∃geoFeatureType(idG,‘‘CON’’)` $\sqcap$ `∃geoFeatureName(idG,X).`

Existe um `netSiteDomainToken X` que tem o `netSitePrefixes` "cm" ou "mun" e um `geoFeatureType` "CON" com o `geoFeatureName X`. Quando é encontrado um emparelhamento os valores `X` de `netSiteDomainToken` e `geoFeatureName`, eu assumo que a entidade do domínio internet representada pelo identificador `idN` tem o âmbito geográfico a entidade representada pelo identificador `idG`.

A Tabela 3.1 apresenta estatísticas sobre os sítios para os quais eu criei regras como a do exemplo acima, juntamente com o número de sítios identificados para cada tipo e o número de emparelhamentos obtidos após a aplicação de regras. Por exemplo, em 2005, Portugal tinha 308 concelhos e 288 deles tinham sítios na web. Para esses, eu assumi um âmbito geográfico para 261. Este simples conjunto de regras podem atribuir âmbitos geográficos para 22% dos sítios considerados nas regras. Contudo, esses emparelhamentos nem sempre funcionam porque o nome do domínio para alguns sítios não deriva diretamente do nome da entidade correspondente. Por exemplo, o sítio `www.cm-ofrades.com` diz respeito ao concelho de `Oliveira de Frades`.

3.4 Geração de Geo-ontologias a partir da GKB

A informação armazenada no repositório da GKB pode ser exportada no formato OWL com uma ferramenta nomeada GOG (*GKB Ontology Generator*). A GOG permite selecionar partes de informação armazenada na GKB e gerar geo-ontologias com vários níveis de detalhe. Os repositórios da GKB tem atualmente cerca de meio milhão de entidades e o usuário raramente quer receber toda a

Tabela 3.1: Âmbitos baseados em regras da GKB atribuídos para sítios em Portugal.

Tipo de sítio	# de sítios	# âmbitos
distritos	33	17 (52%)
concelhos	288	261 (90%)
freguesias	300	124 (41%)
escolas primárias	1955	124 (6%)
centros de treinamento	152	55 (36%)
escolas secundárias	402	105 (26%)

Tabela 3.2: Estatística sobre as geo-ontologias geradas pela GKB.

Estatística	Geo-Net-PT	WGO
# de entidades	418.065	12.293
# de relacionamentos	419.867	12.258
# de relacionamentos parte-de	418.340 (99,83%)	12.245 (99.89%)
# de relacionamentos equivalência	395 (0,09%)	2.501 (20.40%)
# de relacionamentos adjacência	1.132 (0,27%)	13 (0,10%)
Média de entidades ascendentes por entidade	1,0016	1,07
Média de entidades descendentes por entidade	10,56	475,44
Média de entidades equivalentes por entidade equivalente	1,99	3,82
Média de entidades adjacentes por entidade adjacente	3,54	6,5
# de entidades sem ascendentes	3 (0%)	1 (0%)
# de entidades sem descendentes	374.349 (89,54%)	12.045 (97.98%)
# de entidades sem equivalentes	417.867 (99,95%)	11.819 (96.14%)
# de entidades sem adjacentes	417.739 (99,92%)	12.291 (99.99%)

informação.

A GOG exporta a informação no formato OWL, uma representação que estende o RDF[1] e, consequentemente, o XML. As geo-ontologias geradas estão conforme as regras definidas pelo formato RDF, sendo validadas pelo *RDF Validator*[2].

A Tabela 3.2 apresenta uma estatística descritiva das duas geo-ontologias criadas com recurso à GKB. Em ambas as geo-ontologias, a maioria dos relacionamentos são do tipo `parte-de`, enquanto relacionamentos de `equivalência` e `adjacência` são menos frequentes. A WGO é muito menor que a Geo-Net-PT, uma vez que a primeira contém uma granularidade maior, apenas incorpora cidades com mais de 100.000 habitantes, enquanto que a última possui informação geográfica até o nível de código-postal.

3.4.1 Geo-ontologia de Portugal - Geo-Net-PT

A geo-ontologia completa de Portugal (Geo-Net-PT) contém mais de 400.000 entidades, é um recurso público que foi desenvolvido no Pólo XLDB da Linguateca em colaboração com o projecto GREASE e está disponível a partir de `http://xldb.fc.ul.pt/geonetpt`.

A Figura 3.9 apresenta um excerto da geo-ontologia Geo-Net-PT, produzida pela ferramenta GOG a partir de dados geográficos integrados na GKB.

O excerto descreve o tipo de entidade `concelho` (codificado como `CON`), cujo

[1] `http://www.w3.org/TR/REC-rdf-syntax/`
[2] `http://www.w3.org/RDF/Validator/`

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<rdf:RDF xmlns:gn = "http://xldb.di.fc.ul.pt/geo_net_pt01.owl#">
<gn:Geo_Feature rdf:ID="GEO_238">
  <gn:geo_id>238</gn:geo_id>
  <gn:geo_name xml:lang="pt">Porto</gn:geo_name>
  <gn:geo_type_id rdf:resource="#CON"/>
  <gn:info_source_id rdf:resource="#INE"/>
  <gn:related_to>
      <rdf:Bag>
         <rdf:li>
            <gn:Geo_Relationship>
                <gn:rel_type_id rdf:resource="#PRT"/>
                <gn:geo_id>
                    <rdf:Bag>
                      <rdf:li rdf:resource="#GEO_130"/>
                      <rdf:li rdf:resource="#GEO_3967"/>
                    </rdf:Bag>
                </gn:geo_id>
            </gn:Geo_Relationship>
         </rdf:li>
         <rdf:li>
            <gn:Geo_Relationship>
                <gn:rel_type_id rdf:resource="#ADJ"/>
                <gn:geo_id>
                    <rdf:Bag>
                      <rdf:li rdf:resource="#GEO_127"/>
                      <rdf:li rdf:resource="#GEO_156"/>
                      <rdf:li rdf:resource="#GEO_162"/>
                      <rdf:li rdf:resource="#GEO_331"/>
                    </rdf:Bag>
                </gn:geo_id>
            </gn:Geo_Relationship>
         </rdf:li>
      </rdf:Bag>
  </gn:related_to>
  <gn:population>263131</gn:population>
</gn:Geo_Feature>
</rdf:RDF>
```

Figura 3.9: Um excerto da Geo-Net-PT.

nome é `Porto` e possui o identificador `GEO_238`. Esta entidade foi importada da fonte de informação `INE`, código para Instituto Nacional de Estatística. O `Concelho` do `Porto` tem dois tipos de relacionamento com outras entidades: `parte-de` (`PRT`) com a entidades `Grande Porto` e com o `Distrito` do `Porto`, identificados pelos códigos `GEO_130` e `GEO_3967`, respectivamente; `adjacência`

(ADJ) com as entidades Gondomar, Maia, Matosinhos e Vila Nova de Gaia, identificados pelos códigos GEO_127, GEO_156, GEO_162 e GEO_331, respectivamente. A população do Concelho Porto é de 263131 pessoas (esse dado refere-se ao ano de 2004, atualmente a população é de cerca de 240.000 pessoas).

3.4.2 Geo-ontologia Mundial - WGO

Além da geo-ontologia de Portugal, eu gerei uma geo-ontologia de nomes geográficos do mundo, WGO (*World Geographic Ontology*), obtida pela integração de fontes de dados públicas diretamente disponíveis da web. A Figura 3.10 apresenta um extrato dessa geo-ontologia, com a descrição da cidade de Guatemala. Esta entidade geográfica é identificada por GEO_170 e pelo seu tipo cidade-capital (CITY-CAP). Guatemala City tem quatro nomes comuns em inglês, português, espanhol e alemão. Guatemala City também possui um relacionamento do tipo *part-de* (PRT) com a entidade GEO_169, que está declarada em outro parte da geo-ontologia e tem o nome Guatemala. Esta informação foi obtida da fonte de informação Wikipedia (WIKI), de 10 de abril de 2005.

Pelo fato de os modelos geográficos não serem óbvios, eles obrigam que sejam tomadas decisões. No caso de capital-de, eu optei por modelar essa característica do modelo como um tipo. Contudo, o modelo da GKB também suporta a definição do atributo (capital-de). Apenas é necessário criar uma especialização da classe Feature ou acrescentar um atributo (capital-de) nessa classe.

3.5 Aplicações que Utilizam as Geo-ontologias Geradas a partir da GKB

As aplicações descritas nessa seção formam o software base usado no projeto GREASE. A GKB já foi utilizada por diferentes aplicações cujo objetivo é classificar e recuperar páginas da web de acordo com seu âmbito geográfico. Sistemas para reconhecimento de entidades mencionadas (REM), um classificador de documentos de acordo com seu âmbito geográfico e uma interface de recuperação de informação para consultas geográficas.

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<rdf:RDF xmlns:gn = "http://xldb.di.fc.ul.pt/wgo.owl#">
<gn:Geo_Feature rdf:ID="GEO_170">
  <gn:geo_id>170</gn:geo_id>
  <gn:geo_name xml:lang="en">Guatemala City</gn:geo_name>
  <gn:common_name>
    <rdf:Bag>
      <rdf:li>
        <gn:Geo_Name>
          <gn:geo_name xml:lang="es">Ciudad de Guatemala</gn:geo_name>
        </gn:Geo_Name>
      </rdf:li>
      <rdf:li>
        <gn:Geo_Name>
          <gn:geo_name xml:lang="de">Guatemala-Stadt</gn:geo_name>
        </gn:Geo_Name>
      </rdf:li>
      <rdf:li>
        <gn:Geo_Name>
          <gn:geo_name xml:lang="pt">Cidade da Guatemala</gn:geo_name>
        </gn:Geo_Name>
      </rdf:li>
    </rdf:Bag>
  </gn:common_name>
  <gn:geo_type_id rdf:resource="#CITY-CAP"/>
  <gn:related_to>
    <gn:Geo_Relationship>
      <gn:rel_type_id rdf:resource="#PRT"/>
      <gn:geo_id rdf:resource="#GEO_169"/>
    </gn:Geo_Relationship>
  </gn:related_to>
  <gn:info_source_id rdf:resource="#WIKI"/>
</gn:Geo_Feature>
```

Figura 3.10: Um excerto da WGO.

3.5.1 Sistemas de REM, de Reconhecimento de Locais e Módulos de Sistema de Recuperação de Informação Geográfica

As geo-ontologais geradas a partir da GKB têm sido utilizadas por diversos sistemas de REM e módulos do sistema de RIG da Universidade de Lisboa, que participou nas quatro edições do GeoCLEF.

CaGE: é um sistema de REM e de atribuição de âmbito geográfico a páginas

da web (Martins et al., 2007; Silva et al., 2006). O CaGE utiliza as geo-ontologias geradas a partir da GKB nas fases de identificação e desambiguação de locais (Cardoso et al., 2005). Martins et al. (2007) apresentam a arquitetura do CaGE, bem como a descrição detalhada do uso das geo-ontologias.

Faísca: é um sistema de reconhecimento de locais que usa nomes de locais e de tipos de locais contidos nas geo-ontologias geradas a partir da GKB (Cardoso et al., 2008). O Faísca não explora os relacionamentos existentes entre conceitos nas ontologias, mas utiliza os conceitos para desambiguar nomes de locais na fase de REM.

SEI-Geo: é um sistema de extração e integração de conhecimento geográfico e de enriquecimento de geo-ontologias que é descrito em detalhe no Capítulo 5.

QueOnde: é um módulo que utiliza as EG da geo-ontologia para dividir o tópico de uma consulta em três partes 'O que', 'Relacionamento espacial' e 'Onde'. Por exemplo, para o tópico 'tráfego marítimo nas ilhas portuguesas', QueOnde consulta a geo-ontologia e verifica que 'portuguesas' é um adjetivo referente a Portugal e que 'ilhas' é um conceito geográfico.

QuerCol: é um módulo que utiliza a geo-ontologia para fazer expansão de consultas. QuerCol interpreta uma consulta como duas partes 'O que' e 'Onde'. A geo-ontologia é usada para expandir o(s) termo(s) da parte 'Onde'. Por exemplo, na consulta 'regiões vinícolas em Portugal', o módulo QuerCol expande o nome Portugal para todos as províncias, distritos, concelhos e freguesias existentes na geo-ontologia.

3.5.2 Interface de Motor de Pesquisa Geográfica

A GKB é usada também na interface do protótipo Geo-Tumba, um sistema para recuperação de informação geográfica (ver Figura 3.11).

Figura 3.11: Exemplos de interfaces para RIG usando a GKB.

No campo `Local?` o usuário digita a região, a rua, o código postal ou outra entidade geográfica para reduzir o âmbito da consulta. Quando um nome geográfico ambíguo é detectado na consulta, Geo-Tumba apresenta as possíveis alternativas para desambiguação da mesma. Por exemplo, o nome "rua Castelo Branco" ocorre em cinco concelhos diferentes na Geo-Net-PT, os quais são apresentados no lado esquerdo superior da Figura 3.11. Além da consulta por texto, o usuário pode utilizar os mapas para definir o âmbito da consulta.

3.5.3 Interface para Consultas a Almanaques Geo-temporais

A Geo-Net-PT também é utilizada no projeto DIGMAP[1] (*Discovering our Past World with Digitized historical Maps*) (Borbinha et al., 2007), especificamente em uma interface web XML para consultas a almanaques geo-temporais (Manguinhas et al., 2008). Neste serviço, a Geo-Net-PT é integrada com outros almanaques existentes considerando a dimensão temporal juntamente com o conteúdo geográfico dos almanaques. A Figura 3.12 apresenta a interface do sistema.

[1] http://gaz.digmap.eu/

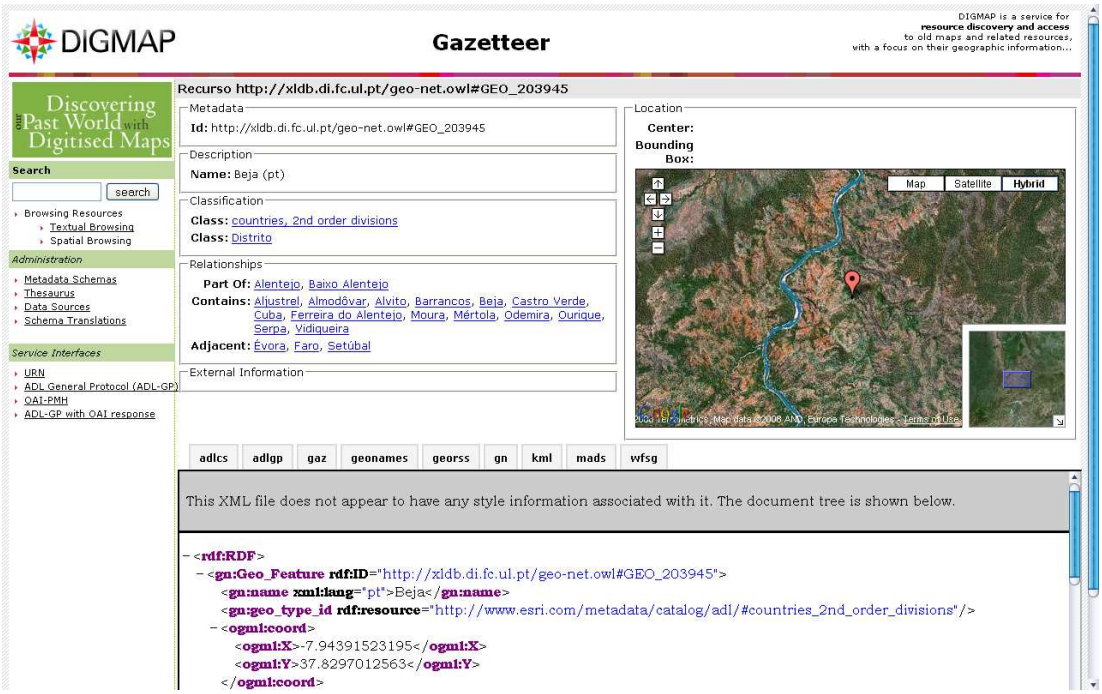


Figura 3.12: Interface para Consultas a Almanaques Geo-temporais.

Para cada local inserido pelo utilizador, o sistema de consultas a almanaques geo-temporais percorre os almanaques e apresenta o nome do local juntamente com seus metadados, relacionamentos e população, entre outras informações subjacentes a cada almanaque. A informação geográfica é apresentada em diversas linguagens (p. ex. XML, OWL e KML - *Keyhole Markup Language*), conforme o almanaque disponibiliza.

No exemplo da Figura 3.12, o sistema apresenta os metadados sobre o 'distrito de Beja', os quais incluem os relacionamentos de `parte-de`, `contém` e `adjacência`. Na parte inferior da figura, estão nove almanaques que contém informação sobre o 'distrito de Beja'. No canto superior direito, o 'distrito de Beja' é ilustrado no mapa.

3.6 Conclusões

Esse capítulo descreveu uma metodologia para construção de uma base de conhecimento geográfico a partir de múltiplas fontes de informação semi-estruturadas.

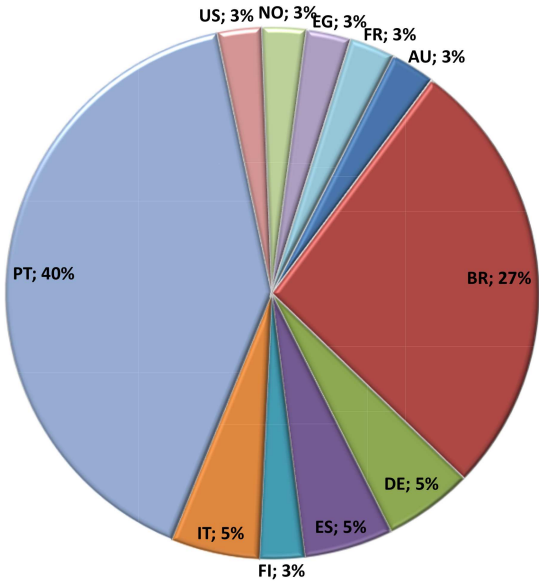


Figura 3.13: Distribuição geográfica dos pedidos da Geo-Net-PT por países.

Essa metodologia contém etapas de modelagem, limpeza de dados, integração de conhecimento e exportação do conhecimento armazenado na forma de geo-ontologias. Estatísticas sobre as geo-ontologias geradas a partir da GKB foram descritas com o objetivo de apresentar a dimensão da informação geográfica após integração de diversas fontes de informação.

As geo-ontologias geradas com a informação geográfica armazenada na GKB foram utilizadas em diversas aplicações, tais como sistemas de REM e de RIG. Uma dessas geo-ontologias, a Geo-Net-PT, tem sido requisitada por diversos grupos de pesquisa ao redor do mundo. A Figura 3.13 apresenta a distribuição geográfica dos pedidos por países. A Geo-Net-PT já foi requisitada por dezenas de investigadores evidenciando o interesse da comunidade em estruturas de representação de conhecimento geográfico.

As geo-ontologias geradas pela GKB têm sido usadas também em várias avaliações conjuntas, a saber: quatro edições do GeoCLEF, 2005 a 2008 (Cardoso et al., 2005; Martins et al., 2006a), HAREM e Mini-HAREM (Martins et al.,

2007). O sistema de RIG da Universidade de Lisboa que utilizou as geo-ontologias obteve o primeiro lugar na avaliação em 2006 nas tarefas monolíngue inglês e português. No GeoCLEF 2007 as geo-ontologias geradas com a GKB foram utilizadas pelos módulos do sistema de RIG QueOnde, Quer-Col e Faísca (Cardoso et al., 2008). Já em 2008, apenas os módulos QueOnde e Quer-Col utilizaram as geo-ontologias.

Os resultados obtidos pelo sistema de RIG nessas avaliações podem constituir indicações da qualidade das geo-ontologias geradas pela GKB assim como das suas lacunas. Um dos resultados foi o primeiro lugar na avaliação em 2006 nas tarefas monolíngue inglês e português. Por outro lado, uma das lacunas das geo-ontologias utilizadas é a falta de dados carregados na GKB sobre o domínio físico da geografia.

A GKB foi descrita anteriormente em publicações como o artigo de Chaves et al. (2005a) e um relatório técnico (Chaves et al., 2005b). Adicionalmente, a extensão do conteúdo da GKB com informação do domínio da geografia física foi realizada em trabalho conjunto com a aluna de mestrado Catarina Rodrigues descrito em Chaves et al. (2007).

Após a construção da GKB a partir de informação geográfica utilizada por autoridades administrativas em Portugal, deseja-se expandir o conhecimento integrado na GKB com informação geográfica presente em textos. Para isso é necessário medir sobre a presença de informação geográfica na web portuguesa, assunto do próximo capítulo, que também descreve uma caracterização da GeoNet-PT.

Capítulo 4

Caracterização da Geo-Net-PT e a Geograficidade da Web Portuguesa

4.1 Introdução

A estatística descritiva da Geo-Net-PT foi apresentada no capítulo anterior. Entretanto, a riqueza de dados dessa geo-ontologia precisa ser quantificada em maior detalhe de modo a fornecer uma visão mais profunda do seu conteúdo. Esse capítulo descreve uma caracterização da Geo-Net-PT explorando a ambiguidade dos termos e a formação lexical dos nomes contidos nela.

Além dessa caracterização o capítulo também apresenta uma panorâmica sobre o conteúdo geográfico presente em textos. *A priori* tem-se a noção de que textos são uma fonte rica em informação geográfica. Contudo, há poucos estudos que mensurem a informação geográfica em textos em português, tal como o de Delboni (2005).

Questões como a quantidade de informação geográfica ambígua com nomes de organizações e pessoas e a quantidade de tipos geográficos (p. ex. administrativo e físico) e de arruamentos que estão mais presentes nos textos ainda permanecem pouco exploradas em português.

Esse capítulo também descreve a geograficidade presente em uma amostra de um corpus da web portuguesa. A quantidade de informação geográfica em textos escritos em outras línguas já foi medida, enquanto para a língua portuguesa havia essa carência.

Para conduzir esse estudo, eu utilizei os seguintes recursos livremente disponíveis:

Geo-Net-PT: descrita no Capítulo 3.

WPT 03: é um corpus da web portuguesa de 2003, com 12 Gbytes e 3.7 milhões de páginas e 1.6 bilhões de palavras (`www.linguateca.pt/WPT03`) (Cardoso et al., 2007). Aproximadamente 68.6% dessas páginas estão em português e mais de 1.5 milhão de paginas distintas.

SIEMÊS: é um sistema de REM (Sarmento, 2006a). Na avaliação conjunta do Primeiro HAREM (Santos et al., 2006) SIEMÊS alcançou 70% de precisão e 75% de abrangência no REM da categoria local. Entretanto, a versão utilizada em nossos experimentos é uma versão com melhoramentos sobre aquela utilizada no Primeiro HAREM.

Baco: é uma base de dados implementada em MySQL que armazena os documentos do WPT 03 juntamente com tabelas de co-ocorrências entre as palavras contidas nos documentos (Sarmento, 2006b).

4.2 Caracterização da Geo-Net-PT

Essa seção tem como objetivo apresentar uma caracterização do conteúdo geográfico presente na Geo-Net-PT.

4.2.1 Descrição Quantitativa da Geo-Net-PT

Uma das formas de apresentar uma descrição do conteúdo presente na Geo-Net-PT é a composição das ocorrências dos conceitos por número de palavras, bem como a ambiguidade entre os nomes geográficos que constituem essa geo-ontologia.

A Tabela 4.1 apresenta uma descrição quantitativa da parte geo-administrativa da Geo-Net-PT para os 11 tipos de locais que estão na parte superior da hierarquia da geo-ontologia. Não estão contabilizados os arruamentos e os códigos postais.

Tabela 4.1: Descrição quantitativa da Geo-Net-PT para 11 tipos de locais: MP - multi-palavra, NUT - Nomenclatura de Unidade Territorial, T(otal) para combinação perfeita e P para parcial.

Tipo	# de termos distintos	# de palavras nos termos					# de termos MP	Total ambiguidade	1 grama ambiguidade	
		1	2	3	4	$\sum$ >4			T	P
NUT1	3	1	0	0	2	0	2	3	0	0
NUT2	7	5	0	0	2	0	2	7	5	0
NUT3	30	8	11	8	3	0	22	6	2	4
região	2	0	1	0	1	0	2	0	-	-
província	11	4	6	0	1	0	7	5	2	1
distrito	18	15	2	1	0	0	3	18	15	0
concelho	323	203	27	68	22	3	121	301	193	1
ilha	11	0	1	6	4	0	11	1	-	-
freguesia	3.597	2.133	336	764	287	77	1.462	2.799	1884	51
localidade	26.924	10.851	4.098	9.661	1.783	531	16.073	3.655	2388	607
zona	3.593	1.201	540	1.233	456	163	2.392	1.241	804	55
Total	34.519	14.421	5.022	11.741	2.561	774	-	-	-	-

4.2.2 Distribuição e Ambiguidade dos Termos na Geo-Net-PT

A Tabela 4.2 apresenta a ambiguidade existente baseada no número de palavras de cada termo na geo-ontologia. Por exemplo, um local nomeado 'Castelo' na linha 1, seria contabilizado como ambíguo de um local chamado 'Castelo' (todas palavras ambíguas) e de outro local nomeado 'Castelo Branco' ($\geq$ 1 palavra ambígua). Os termos formados por uma palavra que são ambíguos contabilizam 21,04%. Ou seja, 21,04% desses termos ocorrem como o mesmo nome de pelo menos uma outra entidade. Também é possível observar que 45,78% dos termos formados por uma palavra estão presentes no nome de outras entidades formadas por mais de uma palavra.

Os termos da parte administrativa da Geo-Net-PT que são ambíguos com outros termos da geo-ontologia somam 12,35%. Quando a ambiguidade é relaxada para o nível de palavra, a ambiguidade atinge 26,59% dos termos. Ou seja, 26,59% dos termos contêm palavras que formam o nome de outras entidades.

Este resultado pode ser comparado com os do almanaque *USGS Concise Gazetteer* (GNIS), que possui 37.479 entradas. Garbinand e Mani (2005) encontraram mais de 50% dos nomes de locais ambíguos nesse almanaque. Contudo, o GNIS inclui locais da geografia administrativa e física e é composto por nomes em língua inglesa.

Em ambas as Tabelas 4.1 e 4.2 o número de termos multi-palavra é superior ao

Tabela 4.2: Distribuição e ambiguidade dos termos na Geo-Net-PT por número de palavras.

# Palavras	Distinto	Todas palavras ambíguas (%)	≥ 1 palavra ambígua (%)
1	11.561	2.433 (21,04)	5.293 (45,78)
2	4.569	381 (8,34)	834 (18,25)
3	10.984	705 (6,42)	1.462 (13,31)
4	2.351	194 (8,25)	404 (17,18)
5	589	20 (3,39)	42 (7,13)
6	109	0	0
7	42	0	0
8	6	0	0
9	6	0	0
$\sum$	30.217	3.733 (12,35)	8.035 (26,59)

número de termos mono-palavra, mas a diferença não é significativa ao ponto de recomendar alguma aplicação a concentrar um maior esforço em tentar trabalhar com termos multi-palavra, por exemplo. Também é interessante observar que o número de termos formados por três palavras é mais do que o dobro dos termos formados por duas palavras. Note-se que os termos formados por três palavras incluem na sua grande maioria a 'de' ou contrações de 'de' na palavra do meio.

A Figura 4.1 ilustra a ambiguidade dos termos da Geo-Net-PT considerando os conceitos acima do nível de arruamento distribuídos por número de repetições. Há 5.755 EG distintas cujos nomes se repetem entre 2 e 10 vezes em outras EG. A partir de 10 repetições a ambiguidade é mais atenuada, conforme mostra o gráfico. Essa distribuição segue a lei de Zipf (1949).

4.3 Geograficidade em Textos

Nos próximos experimentos eu mensurei a ambiguidade dos nomes de locais com as categorias pessoa e organização definidas no HAREM e dimensionei os tipos de locais presentes numa coleção da web portuguesa e nos textos em português da coleção dourada do Primeiro HAREM.

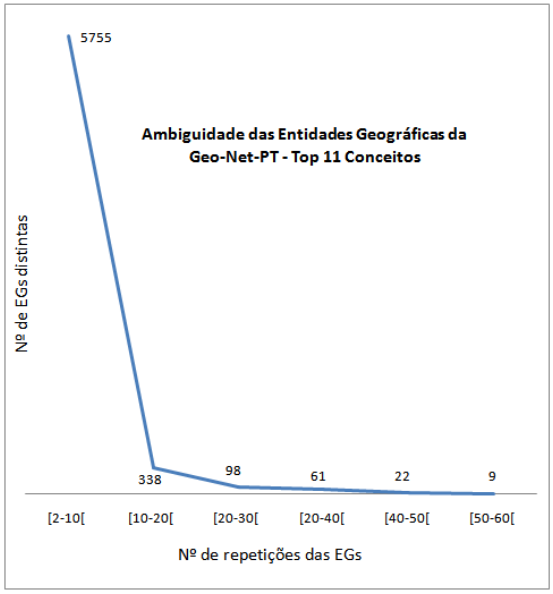


Figura 4.1: Ambiguidade das Entidades Geográficas da Geo-Net-PT por número de repetições.

4.3.1 Nomes de Pessoas e Organizações como Locais?

Eu selecionei aleatoriamente 32.000 documentos do WPT 03 (1.5 milhão de documentos distintos) que foram etiquetados pelo SIEMÊS que foi configurado para etiquetar EM das categorias pessoa, organização e local, uma vez que a maior parte da ambiguidade com locais ocorre nas duas primeiras categorias. A Tabela 4.3 apresenta os principais resultados.

Em português existem vários sobrenomes idênticos a nomes de locais (p. ex. 'Irene Lisboa', 'Camilo Castelo Branco'). Eu pretendo identificar a frequência com que um local é incluído no nome de uma organização para estimar quantos casos um local realmente reflete a localização física da organização. Por exemplo, a 'Universidade de Évora' está fisicamente em Évora, enquanto a localização da 'Associação de Amizade Portugal-Itália' e da 'Pastelaria Finlândia' não são refletidas nos seus nomes.

A Tabela 4.3 mostra que quase 1 milhão de EM, pertencentes às três categorias, foram identificadas, 30% das quais correspondem a locais. Para todas as

Tabela 4.3: EM detectadas numa amostra de 32.000 documentos do WPT 03: MP significa multi-palavra e GN significa Geo-Net-PT. EMD: Entidade Mencionada Distinta.

	# de EM (%)	# de EMD	# de EM MP (%)	# de EMD MP(%)	# de EMD MP com um nome na GN (%)	# de EMD na GN (%)
PEO	250,585 (26.48)	77,228	140,155 (55.93)	58,991 (76.39)	24,105 (31.21)	521 (0.67)
ORG	418,915 (44.27)	114,353	214,698 (51.25)	89,790 (78.52)	26,789 (23.43)	462 (0.40)
LOC	276,775 (29.25)	47,972	90,018 (32.52)	36,395 (75,87)	22,959 (47.86)	4,576 (9.53)
$\sum$	946,275 (100.00)	239,553	444,871 (47.01)	185,176(77.30)	73,853 (30.83)	5,559 (2.32)

categorias, mais de 75% das entidades mencionadas distintas são multi-palavra. Os nomes de organizações são os mais frequentes, tanto em número absoluto quanto no número de distintos. Considerando o rácio entre a quantidade de nomes distintos e repetidos, os nomes de locais são consideravelmente mais repetidos do que nomes de pessoas na amostra.

As duas últimas colunas da Tabela 4.3 medem a sobreposição total e parcial entre nomes de entidades e de locais: enquanto a ambiguidade com nomes de pessoas e organizações é menos de 1%, 31% das entidades mencionadas distintas da categoria pessoa e 23,43% das entidades mencionadas distintas da categoria organização contêm pelo menos um nome geográfico incluído na Geo-Net-PT Para esta comparação, eu utilizei todos os nomes na Geo-Net-PT (27.855), exceto nomes de arruamentos e códigos postais.

A coleção WPT 03 contém apenas 10% dos locais distintos presentes na Geo-Net-PT. Esse fato parece surpreendente, uma vez que todo o vocabulário geográfico administrativo está presente na Geo-Net-PT.

A explicação para grande quantidade de locais ausentes na Geo-Net-PT (além da sistemática sobre-geração[1] do SIEMÊS) consiste nas seguintes hipóteses:

- uma considerável parte dos locais presentes nos textos são de locais fora de Portugal;
- as pessoas tendem a escrever nomes de locais indiretamente (p. ex. 'perto da universidade' ou 'em frente ao Saldanha') sem mencionar o real nome do local ao qual estão se referindo. A grande maioria (senão todos esses nomes)

[1] A sobre-geração ocorre quando um sistema anota mais entidades mencionadas do que aquelas que realmente são e existem no texto.

não fazem parte do vocabulário geográfico de autoridades administrativas de Portugal.

4.3.2 Caracterização de Tipos de Locais em Documentos

As próximas duas seções apresentam a caracterização dos tipos de locais na amostra de 32.000 documentos da WPT 03 e na parte portuguesa da coleção dourada do Primeiro HAREM. Essa caracterização fornece mais detalhes sobre a formação das EM da categoria local.

4.3.2.1 Tipos de Locais numa Amostra da Web Portuguesa

Para investigar se o tipo de local ocorrendo em textos da web portuguesa tinham diferentes propriedades (granularidade, geografia física (rios, montanhas, etc.)), eu verifiquei os tipos das EM da categoria local definidos no Primeiro HAREM que o SIEMÊS encontrou após ser executado sobre a amostra. A Tabela 4.4 apresenta os resultados.

Tabela 4.4: Distribuição dos tipos contidos na categoria local na amostra da WPT 03.

Tipo	# de EMD(%)	# de EMD MP(%)
Nomes de locais habitados (POV)	33.827 (70,51)	24.037 (71,06)
End. completo (ENDRALAR)	3.505 (7,31)	3.313 (94,52)
Sociedade/Cultura (SOCCUL)	3.474 (7,24)	3.161 (90,99)
País (PAIS)	1.987 (4,14)	1.419 (71,41)
Religião (RLG)	1.197 (2,50)	1.113 (92,98)
Outro ($\sum$ 11 tipos)	3.982 (8,30)	3.352 (84,18)
$\sum$ total	47.972 (100)	36.395 (75,87)

Mais de 85% dos tipos de locais estão concentrados em apenas três tipos (POV, ENDRALAR e SOCCUL) e o mesmo ocorre quando contabilizamos somente os locais formados por nomes compostos por múltiplas palavras.

4.3.2.2 Tipos de Locais na Parte PT da Coleção do Primeiro HAREM

A Tabela 4.5 apresenta o emparelhamento dos tipos da categoria local definidos no Primeiro HAREM com a Geo-Net-PT e a WGO (uma geo-ontologia multi-língue

Tabela 4.5: Tipos de locais na parte PT da coleção dourada (CD) do Primeiro HAREM anotada manualmente.

Tipo HAREM CD	# Átomos	# Átomos distintos (%)	Geo-Net-PT		WGO	
			# Átomos (%)	# Átomos distintos (%)	# Átomos (%)	# Átomos distintos (%)
Administrativo	754	275 (36,47)	248 (32,89)	105 (38,18)	169 (22,41)	108 (39,27)
Alargado	110	97 (88,18)	34 (30,91)	18 (18,56)	9 (8,18)	3 (3,09)
Geográfico	65	52 (80)	30 (46,15)	16 (30,77)	18 (27,69)	8 (15,38)
Total	929	424 (45,64)	312 (33,58)	139 (32,78)	196 (21,1)	119 (28,07)

com 15.005 nomes em quatro línguas (751 (4.87%) em português. Primeiramente nomeada GKB-ML): somente 32,89% (22,41%) dos nomes de locais administrativos aparecem nas geo-ontologias. Tal fato se deve a diversos fatores: tipos de locais como aldeia (p. ex. 'Cetos' e 'Fontanelas') e bairros (p. ex. 'Bairro Alto' e 'Baixa Chiado') não estão incluídos nas geo-ontologias. Cidades importantes com menos de 100.000 habitantes (p. ex. 'Caçapava do Sul' no Brasil e 'Toledo' na Espanha) estão fora da WGO.

4.3.3 Distribuição dos Locais por Documentos de uma Amostra da Web Portuguesa

Uma das especificidades deste trabalho de análise da informação geográfica em textos é o fato dessa ser transversal a todos os tipos de texto e não somente a textos de domínio específico (p. ex. turismo e geografia). Eu medi o número de documentos com pelo menos uma EM: 31.489 (98,4% da amostra). As referências para pessoas estão presentes em 21.499 (67,18%) documentos, para organizações em 30.328 (94,77%) documentos e para locais em 24.468 (76,46%) documentos.

A Tabela 4.6 mostra que cada documento (contendo ao menos uma EM) contém em média cerca de 20 entidades mencionadas distintas das quais mais de sete são locais e cerca de 50% dos documentos com locais contém mais de três locais. Os valores da coluna 'Distintos' medem as entidades mencionadas distintas dentro de cada documento.

4.3.4 Co-ocorrências entre Tipos de Locais

Uma ontologia construída a partir de fontes de dados administrativos contém tipos de entidade que estão relacionados conforme a visão de mundo das autori-

Tabela 4.6: Distribuição das EM por documentos de uma amostra da web portuguesa.

	Total	Distinto		Total	Distinto
Média PEO por doc. com PEO	11,65	7,82	Mediana LOC	4	3
Média ORG por doc. com ORG	13,81	9,78	Desvio Padrão LOC	149,7	57,54
Média LOC por doc. com LOC	11,31	7,34	# docs. com 1 LOC	5.443	6.184
Média NE por doc. com NEs	30,04	20,47	# docs. > 3 LOC	12.913	11.640
# máx. de LOC em 1 doc.	20.594	6.472	# docs. > 30 LOC	1.483	713

Tabela 4.7: Co-ocorrências entre tipos de locais presentes na Geo-Net-PT. Pro=província; Reg=região; Dis=distrito; Ilh=ilha; Con=concelho; Frg=freguesia; Loc=localidade; Ald:aldeia; Cid=cidade; Vil=vila e Mun=município

	Pro	Reg	Dis	Ilh	Con	Frg	Loc	Ald	Cid	Vil	Mun
Pro											
Reg			138		163				131		
Dis					862	447					
Ilh											
Con						2136	125	269	216	176	
Frg								126	232		259

dades administrativas. Entretanto, esses mesmos tipos de locais podem estar co-relacionados de forma distinta em textos da web. A motivação desse experimento consiste em encontrar evidências da co-relação existente entre tipos de locais num texto. Por exemplo, quando uma pessoa menciona um tipo de local tal como 'freguesia', qual(is) o(s) tipo(s) de locais que também são mencionados próximos (na mesma sentença) desse tipo?

Nesse experimento o algoritmo verifica todas as combinações possíveis entre os tipos que co-ocorrem no WPT 03. Por exemplo, numa sentença contendo menção aos tipos distrito, concelho e aldeia, as co-ocorrências contabilizadas são [distrito, concelho], [distrito, aldeia] e [concelho, aldeia]. A Tabela 4.7 apresenta os resultados encontrados para o número de ocorrências acima de 100.

Os tipos que estão mais co-relacionados são 'concelho' e 'freguesia' com 2.136 ocorrências, seguido por 'distrito' e 'concelho' com 862 ocorrências. Dos tipos que estão fora do vocabulário administrativo, os conceitos de 'aldeia' e 'cidade' estão mais relacionados com concelho (269 e 216 ocorrências, respectivamente). Outro

tipo, o de 'município' aparece mais relacionado com 'freguesia' (259 ocorrências). Já o tipo de 'vila' ocorre com maior frequência com 'concelho' (176 ocorrências).

Outro resultado importante é o número de ocorrências da tripla 'distrito, concelho e freguesia' que é 341. Naturalmente, quando um humano está descrevendo um local mais específico tende a mencionar tipos mais abrangentes para referenciar o leitor no espaço.

4.4 Projetando a Geo-Net-PT sobre o WPT 03

Até essa seção, eu estava preocupado com a questão de quantos locais encontrados em texto (numa pequena amostra do WPT 03) estavam incluídos na Geo-Net-PT. Agora, eu vou na direção oposta, ou seja, quanto da informação presente na Geo-Net-PT está incluída na web portuguesa.

Existem duas listas de frequência do WPT 03: uma distingue entre maiúsculas e minúsculas enquanto a outra não (Cardoso et al., 2007) (p. 285). Eu usei a lista que faz distinção entre maiúsculas e minúsculas para evitar casos de ambiguidade com nomes comuns. Essa lista contém 7.751.473 palavras distintas, correspondentes a 1.529.758 documentos distintos.

A Geo-Net-PT contém 78.349 nomes de entidades distintos entre todos os tipos de entidades definidos nela, os quais 58.219 (74,3% dos nomes distintos) são multi-palavra. Restam 19.933 nomes formados por uma palavra para serem comparados diretamente com a lista de frequência.

Para os 187 nomes de concelhos formados por apenas uma palavra, existem 4.959.681 ocorrências no WPT 03. O concelho mais frequente é Lisboa com 856.911 ocorrências em 300.507 documentos, enquanto o concelho de Povoação é o menos frequente com 3.832 ocorrências em 2.578 documentos.

A Tabela 4.8 apresenta a frequência dos nomes na Geo-Net-PT por n-gramas. Essa estatística foi baseada nas relações que armazenam os n-gramas no Baco. Cada linha da relação possui as palavras que formam o termo, a frequência no WPT 03 e o número de documentos em que o termo ocorre. Cerca de 60% dos nomes presentes na Geo-Net-PT estão presentes no WPT 03. Aqueles compostos por quatro palavras são os menos frequentes, ao passo que os nomes formados por uma palavra atingem quase 80% de presença nesse corpus da web. Esses

Tabela 4.8: Frequência dos nomes acima dos tipos de arruamento da Geo-Net-PT no WPT 03.

# palavras	# nomes na GN	# no WPT 03	% da GN na WPT 03
1	11.164	8.761	78,48
2	4.193	2.121	50,58
3	9.985	4.678	46,85
4	1.925	691	35,90
Total	27.267	16.251	59,60

Tabela 4.9: Estatística descritiva dos nomes de entidades acima dos tipos de arruamento da Geo-Net-PT no WPT 03.

# palavras	média	mediana	desvio padrão	soma	máximo
frequência no WPT 03					
1	3.438,25	68	21.610,38	30.122.547	856.911
2	596,28	11	4.240,12	1.264.705	125.180
3	164,49	7	1.216,31	769.502	49.254
4	292,55	17	1.501,26	202.151	27.529
# de documentos					
1	1.821,67	50	8.913,04	15.959.685	300.507
2	336,49	9	1.767,59	713.700	32.969
3	108,58	6	746,24	507.933	26.116
4	187,07	14	825,58	129.262	11.542

resultados são mais um indício do volume de informação geográfica presente em texto.

A Tabela 4.9 apresenta a estatística descritiva dos nomes presentes na Geo-Net-PT, agrupados por número de palavras. A moda, tanto na frequência quanto no número de documentos, é um.

4.4.1 Distribuição de Arruamentos por Documentos da Web

Estudos que tenham mensurado a presença dos tipos de arruamento (p. ex. avenida, rua e travessa) em corpus da web em português são desconhecidos por mim. Nessa seção eu apresento uma caracterização da presença de todos os tipos de arruamentos presentes numa geo-ontologia num corpus da web.

A geo-ontologia utilizada foi a Geo-Net-PT que contém 146.422 ocorrências

de 45 tipos de arruamentos. Cada um desses 45 tipos foi projetado sobre a lista de frequências de palavras do WPT 03. A Tabela 4.10 apresenta os tipos de arruamento presentes na Geo-Net-PT projetados na lista de frequências do WPT 03.

Os arruamentos predominantes na geografia administrativa de Portugal são 'ruas' e 'travessas'. As ruas representam mais de 60% do total de arruamentos em Portugal, enquanto as travessas formam 12,40% dos arruamentos. 'Rua' também é o tipo de arruamento mais frequente no WPT 03, após o tipo ambíguo 'acesso'. Os tipo ambíguos incluem 'largo' (adjetivo), 'travessa' (substantivo), 'caminho' (verbo), 'quinta' (numeral), 'calçada' (verbo), 'vale' (verbo), 'passeio' (verbo, substantivo), 'canto' (substantivo, verbo) e 'via' (verbo) entre outros.

Por outro lado, as 'travessas' ocorrem com bem menos frequência no WPT 03, sendo apenas o 28º tipo de arruamento mais frequente. A média da frequência de arruamentos no WPT 03 é de 62.650 e a mediana é 20.010 ocorrências com um desvio padrão de 89.539.

4.5 Resumo e Conclusões

Esse capítulo apresentou uma caracterização da Geo-Net-PT, que inclui a descrição do número de palavras que compõem cada termo da Geo-Net-PT, bem como a ambiguidade existente entre os termos dessa geo-ontologia.

Além da ambiguidade entre locais, a ambiguidade de locais com nomes de pessoas e organizações também foi medida. Eu encontrei 31% das entidades mencionadas distintas da categoria pessoa e 23,43% das entidades mencionadas distintas da categoria organização contendo pelo menos um nome geográfico incluído na Geo-Net-PT.

Eu também encontrei mais de 85% dos tipos de locais concentrados em apenas três tipos (povoamento, endereço alargado e sócio-cultural) e o mesmo ocorre quando eu contabilizo somente os locais formados por nomes compostos por múltiplas palavras. Quando medindo a co-ocorrência entre os tipos de locais da Geo-Net-PT, eu verifiquei que os tipos 'concelho' e 'freguesia' estão mais correlacionados nos textos.

Ao projetar a Geo-Net-PT sobre o WPT 03, verifica-se que cerca de 60% dos nomes presentes nessa geo-ontologia estão também presentes no WPT 03. Tal

Tabela 4.10: Frequência dos tipos de arruamento presentes na Geo-Net-PT projetados na lista de frequências do WPT 03. * indica ambiguidade com outras categorias sintáticas.

Tipo	Ocorrências distintas na GN	Freq. WPT 03	# docs. WPT 03
rua	91310	288045	123865
travessa*	18150	11029	6259
largo*	7284	41298	24910
praceta	3749	5221	3376
avenida*	3630	39809	21974
beco	3426	2447	1763
estrada	2317	73494	42008
bairro	2009	45139	23546
caminho*	1450	93852	60028
praça*	1358	58412	37514
quinta*	1196	213643	60525
urbanização*	816	24387	12132
calçada*	712	7155	4680
canada	673	6057	4139
vereda	600	326	232
viela	427	512	319
azinhaga	340	1493	850
pátio*	324	6136	4100
escadinhas*	280	428	318
alameda	257	7605	5839
rampa*	189	4211	2434
escadas*	165	6226	4231
ladeira*	161	2264	1596
rotunda	153	4810	2931
quelha	130	137	104
parque	126	129609	56152
zona	107	194700	105690
monte*	103	62149	36906
passeio*	97	20010	13052
jardim*	93	70606	36828
sítio*	89	72354	51090
canto*	88	29557	14634
loteamento	79	19937	3548
terreiro*	73	5289	3513
vale*	68	129169	72560
via*	56	138988	82875
lugar*	47	237234	125326
campo*	46	181721	100465
ponte*	46	72197	31407
cais*	43	13449	8860
recanto*	42	1172	1044
adro	30	1991	1459
acesso*	24	410576	207576
carreira*	20	83906	46146
ruela	18	514	396

fato é um indício de que a construção de geo-ontologias somente a partir de textos não é suficiente para prover uma quantidade abrangente de locais para as mesmas. Assim, a utilização de fontes de informação semi-estruturadas, tais como bases de dados, é complementar ao uso de textos para o povoamento de geo-ontologias.

No que diz respeito à geograficidade da web portuguesa, os resultados evidenciam que existe informação geográfica suficiente em textos e que esses textos são uma fonte valiosa para complementar o conhecimento geográfico formal de geo-ontologias.

Parte dos resultados apresentados nesse capítulo foram publicados em dois artigos curtos (Chaves e Santos, 2006) e (Santos e Chaves, 2006). O próximo capítulo apresenta a descrição dos métodos para identificação e reconhecimento de entidade geográfica em textos bem como os métodos de integração de conhecimento em geo-ontologias.

Capítulo 5

Extração, Anotação e Integração de Conhecimento Geográfico

5.1 Introdução

O tratamento do conhecimento geográfico presente em textos ainda é uma área pouco explorada. Esse conhecimento varia bastante na forma como se exprime nos textos: terminologia, posição das expressões e relacionamentos entre as entidades geográficas são algumas das características a que sistemas de extração de informação se devem adaptar para traduzir o conhecimento informal para um formato legível por máquina.

Este capítulo apresenta o restante da terminologia utilizada para representar conhecimento geográfico de textos juntamente com um novo formato proposto para representação de conhecimento geográfico extraído a partir de textos. Esse formato é baseado em triplas RDF codificando explicitamente o conhecimento geográfico que é apenas inteligível por humanos na representação em linguagem natural.

Em seguida, descreve a arquitetura de extração, anotação e integração de conhecimento geográfico do sistema SEI-Geo (acrônimo de Sistema de Extração e Integração de conhecimento Geográfico), juntamente com os algoritmos usados para identificar e classificar o conhecimento geográfico de textos e integrá-lo em geo-ontologias existentes.

O SEI-Geo tem como objetivo reconhecer o conhecimento geográfico em

textos, gerar uma representação estruturada desse conhecimento e integrá-lo em geo-ontologias.

5.2 Representação de Conhecimento Geográfico Extraído de Texto

Uma das alternativas para representar o conhecimento presente em textos é através de conjuntos de triplas (p. ex. RDF). Ou seja, representam-se termos e relacionamentos do texto no formato de triplas 'sujeito-predicado-objeto'. Wilks (2008) argumenta que a estratégia de explorar o uso dos termos no texto e representá-los estruturadamente tem sido a tendência mais razoável para o real desenvolvimento da Web Semântica (WS). O mesmo autor ainda enfatiza que o Processamento de Linguagem Natural (PLN), ao utilizar a web como um corpus, será capaz de prover a base semântica da WS.

O objetivo dessa seção é descrever a estratégia adotada para concretizar a idéia de formalizar o conhecimento geográfico presente em textos. Pelo fato de a terminologia em português ainda não estar estabelecida, a próxima seção define o restante da terminologia utilizada nessa tese e que pode servir de embasamento teórico para futuros trabalhos na área de Recuperação de Informação Geográfica (RIG), Extração de Informação (EI) e representação de conhecimento geográfico.

5.2.1 Terminologia

Os conceitos e definições usados no domínio de extração e integração de informação geográfica podem ser interpretados de diferentes formas conforme seus usos. As definições apresentadas nessa seção estendem aquelas descritas na Seção 2.2. A partir de três conceitos atômicos, `nome de entidade (NE)`, `nome de tipo de entidade (NTE)` e `referência ontológica (RO)`, definem-se os seguintes conceitos (os colchetes '[]' representam listas):

- **Entidade Candidata (`EC`):** é um topônimo, um nome próprio (composto por pelo menos uma palavra). Exemplos de entidades candidatas incluem 'Grécia', 'Brasília' e 'concelho de Braga'. Formalmente, uma `EC` é um

Esboço de Entidade EE em que a cardinalidade da lista da esquerda é 1 e a da direita é 0, |[DE]| = 1 e |[RO]|=0. (p. ex. <[<‘ ’,Sintra>],[ ]>).

- **Padrão Geográfico (PG):** é um conjunto de expressões de vários tipos: métrico (p. ex. ‘km de’, ‘minutos de’), direcional (p. ex. ‘atrás de’, ‘em frente a’), orientação (p. ex. ‘norte’, ‘leste’), fuzzy (p. ex. ‘próximo’, ‘antes’, ‘acima’), verbo (p. ex. ‘localizado’, ‘situado’), substantivo (p. ex. ‘litoral’, ‘natural’), advérbio (‘cá’, ‘lá’) ou padrões do tipo Hearst expandidos (p. ex. ‘é um concelho’, ‘aldeias tais como’). Esses padrões também podem ser vistos como expressões gramaticais com termos que podem ser precedidos ou sucedidos por informação geográfica.

- **Associação entre Esboços de Entidades (AEE):** é uma tripla definida por <EE_1,R,EE_2>, onde $EE_1 \neq EE_2$. Por exemplo, <Canaviais,contida, concelho de Évora>.

- **Associação Candidata (AC):** uma associação onde um EE é EC e o outro EE é E. Por exemplo < <[<‘ ’,Azenhas do Mar>],[ ]>,parte-de, <[<concelho,Sintra>],[GEO_284]> >.

- **Arbusto (Ar):** um arbusto é composto por pelo menos duas entidades candidatas e um relacionamento. Não há número máximo de entidades e relacionamentos pré-definido. Essas entidades e os relacionamentos são procurados nas geo-ontologias que podem fornecer identificadores das entidades e dos relacionamentos já presentes nelas. Um exemplo de arbusto é <cidade de Oslo,parte-de,Noruega>. Portanto, um arbusto é definido mais formalmente como um conjunto de uma ou mais triplas <EC_1,R,EC_2>. Um arbusto é um conjunto {A_i} com cardinalidade 1..n, onde A_i é uma AE ou ATE. Os arbustos são formalizados no formato de triplas *Resource Description Framework* (RDF).

Existem relacionamentos entre tipos de entidades, entidades geográficas e entre tipo de entidade e entidade, conforme a Tabela 5.1.

As etiquetas apresentadas na Tabela 5.2 são utilizadas na representação dos arbustos em XML no SEI-Geo. Para ilustrar a representação de um arbusto extraído de uma frase, é dado o exemplo a seguir:

Tabela 5.1: Relacionamentos possíveis na (ISO19109, 2006)

		Exemplo
TE	TE	Uma freguesia é parte de um concelho.
E	E	A freguesia de Santa Isabel é parte do concelho de Lisboa.
E	TE	Santa Isabel é uma freguesia.

Tabela 5.2: Equivalências entre termos e etiquetas

Termo	Etiqueta
Ar	*shrub*
AE ou ATE	*triple*
E ou EE	*feature*
NE	*name*
TE	*type*
R	*relationship*
RO	*geo_id*

'Francisco nasceu na aldeia Vale do Souto que está localizada em Oleiros, distrito de Castelo Branco situado na província Beira Baixa'.

Esse arbusto é representado em XML como segue:

```
<gn:shrub rdf:ID="GEO_15">
  <gn:triple>
      <gn:feature>
        <gn:name>Vale do Souto</gn:name>
        <gn:type rdf:resource="#aldeia"/>
        <gn:ADM_id rdf:ID="ADM_500001"/>
      </gn:feature>
      <gn:relationship rdf:resource="#parte-de"/>
      <gn:feature>
        <gn:name>Oleiros</gn:name>
        <gn:ADM_id rdf:ID="ADM_201"/>
        <gn:ADM_id rdf:ID="ADM_3763"/> ... <gn:ADM_id rdf:ID="ADM_96275"/>
      </gn:feature>
  </gn:triple>
  <gn:triple>
      <gn:feature>
        <gn:name>Oleiros</gn:name>
        <gn:ADM_id rdf:ID="ADM_201"/>
        <gn:ADM_id rdf:ID="ADM_3763"/> ... <gn:ADM_id rdf:ID="ADM_96275"/>
```

```
      </gn:feature>
      <gn:relationship rdf:resource="#parte-de"/>
      <gn:feature>
        <gn:name>Castelo Branco</gn:name>
        <gn:type rdf:resource="#distrito"/>
        <gn:ADM_id rdf:ID="ADM_180"/>
      </gn:feature>
  </gn:triple>
  <gn:triple>
      <gn:feature>
        <gn:name>Castelo Branco</gn:name>
        <gn:type rdf:resource="#distrito"/>
        <gn:ADM_id rdf:ID="ADM_180"/>
      </gn:feature>
      <gn:relationship rdf:resource="#situado"/>
      <gn:feature>
        <gn:name>Beira Baixa</gn:name>
        <gn:type rdf:resource="#provincia"/>
        <gn:ADM_id rdf:ID="ADM_185"/>
      </gn:feature>
  </gn:triple>
</gn:shrub>
```

O arbusto é armazenado dentro da etiqueta `<shrub>` que é composta por pelo menos uma etiqueta `<triple>`. Dentro de cada tripla do arbusto estão identificadas as entidades pelas etiquetas `<feature>` e os relacionamentos pelas etiquetas `<relationship>`. Uma tripla tem sempre duas etiquetas `<feature>` e uma `<relationship>` dentro dela.

Uma entidade é composta por pelo menos um nome `<name>`. As etiquetas `<type>` e `<geo_id>` são opcionais. No exemplo, o nome `Oleiros` é ambíguo, e após consultar uma geo-ontologia, identificaram-se nove entidades diferentes com esse nome (apenas três estão no exemplo por razão de espaço). Os identificadores devem ser desambiguados no módulo de integração de informação.

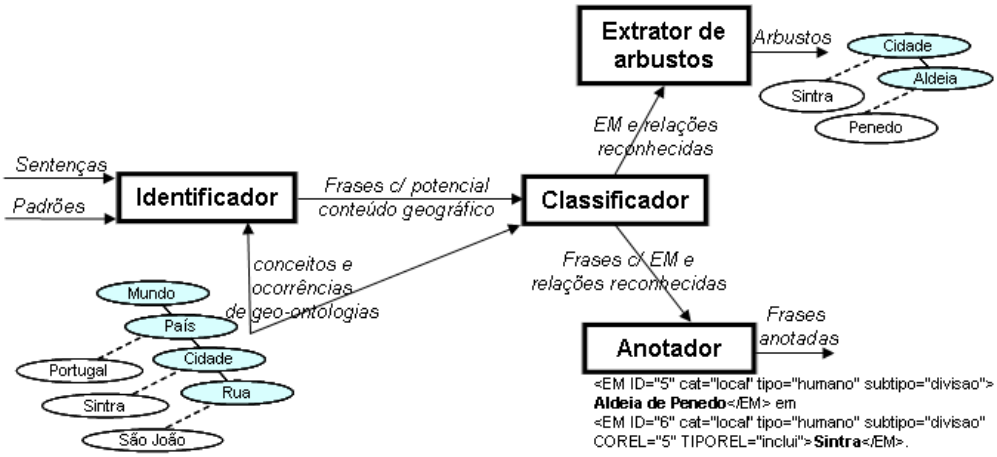


Figura 5.1: Arquitetura do módulo de extração de informação geográfica (EIG) do SEI-Geo.

5.3 Arquitetura de Extração e Integração de Conhecimento Geográfico - SEI-Geo

O SEI-Geo é composto por dois módulos principais: o de Extração de Informação Geográfica (EIG) e o de Integração de Conhecimento Geográfico (ICG). O módulo de extração tem como objetivo identificar e reconhecer o conhecimento geográfico em textos e representá-lo de forma estruturada. A Figura 5.1 apresenta a arquitetura do módulo de EIG. A seguir são descritas as funções de cada submódulo:

Identificador: recebe como entrada uma coleção de textos previamente segmentados em sentenças, mais um conjunto de padrões e conceitos e ocorrências de geo-ontologias. As frases com potencial conteúdo geográfico são a entrada do módulo Classificador.

Classificador: recebe as frases, consulta as geo-ontologias para fazer a desambiguação e identificar relacionamentos semânticos.

Extrator de arbustos: recebe os locais reconhecidos e constrói os arbustos. Um exemplo de arbusto na Figura 5.1 é <'Aldeia de Penedo','parte da','cidade de Sintra'>.

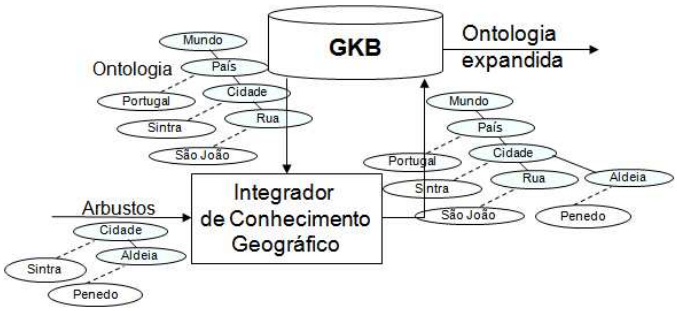


Figura 5.2: Arquitetura do módulo de integração de conhecimento geográfico (ICG) do SEI-Geo.

Anotador: recebe a sentença com locais e relacionamentos reconhecidos e faz a anotação no formato solicitado por qualquer aplicação. O anotador insere etiquetas com nomes de categoria semântica, tipo e subtipo.

A Figura 5.2 apresenta a interação entre o módulo de ICG com a GKB. O módulo de ICG utiliza também o conhecimento armazenado na GKB, faz a integração entre este e a informação nos arbustos e retorna para a GKB o conhecimento geográfico expandido.

A integração de conhecimento textual em geo-ontologias, nessa tese, concentra-se em encontrar informação geográfica complementar àquela existente nas geo-ontologias e integrar essa informação no nível de granularidade mais adequado na geo-ontologia. A integração de conhecimento geográfico ocorre quando novos fatos geográficos são descobertos em texto ou quando fontes de informação públicas fornecem seus dados.

A seguir são descritos problemas encontrados nos textos juntamente com as soluções implementadas no SEI-Geo.

- **Integração de Conhecimento Indireto:** considere-se, por exemplo, a seguinte sentença, retirada do WPT 03, 'Segundo informou a Protecção Civil à Lusa, as crianças de 6 e 4 anos, foram encontradas às 00h45 de hoje e cerca de meia hora depois foi detectado o corpo da mãe, muito próximo do local onde estavam soterrados os filhos, na aldeia da Azinheira, distrito de Vila Real.'. Verifica-se que existem duas entidades 'aldeia da Azinheira' e 'distrito de Vila Real' e, além disso, existe um relacionamento entre elas. Contudo, o NTE `aldeia` é parte de um NTE mais específico numa geo-

ontologia de Portugal. Uma `aldeia` é parte de uma `freguesia`, que por sua vez é parte de um `concelho`, que é parte de um `distrito`. Entretanto, o conhecimento disponível no texto apresenta um relacionamento direto entre uma `aldeia` e um `distrito`. **Solução:** preferencialmente, a integração acontece no nível mais específico da hierarquia; contudo, como isso nem sempre é possível, o conhecimento incompleto adquirido no texto não será desperdiçado, mas sim integrado conforme for encontrado no texto. Retornando ao exemplo acima, a integração se dará pela ligação entre a 'aldeia da Azinheira' e o 'distrito de Vila Real' com n nodos entre os dois NTE.

- **Integração de Domínios Físico e Administrativo:** considere-se, por exemplo, a seguinte sentença retirada da Wikipedia, 'A Serra da Peneda é a quinta maior elevação de Portugal Continental, com 1416 metros de altitude. Situa-se no Alto Minho, nas proximidades de Castro Laboreiro, fazendo parte do sistema montanhoso da Peneda-Gerês.'. Existem relacionamentos entre os NE 'Alto Minho' e 'Castro Laboreiro' do domínio da geografia administrativa e 'Peneda-Gerês' do domínio da geografia física. **Solução:** o SEI-Geo deve associar o NE da geografia física 'Serra da Peneda' ao NE da geografia administrativa 'Castro Laboreiro', freguesia existente na Geo-Net-PT. Para isso, o algoritmo faz uma consulta à Geo-Net-PT e detecta o NE 'Castro Laboreiro' como uma freguesia, mas não detecta 'Alto Minho', uma sub-região que, atualmente, está fora do vocabulário das autoridades administrativas. Quando o algoritmo procura o NTE da geografia física 'sistema montanhoso' e não encontra, ele adiciona esse novo NTE, juntamente com sua ocorrência 'Peneda-Gerês'. Por fim, o algoritmo adiciona o relacionamento entre as ocorrências 'Serra da Peneda' e 'Castro Laboreiro'.

O novo conhecimento integrado no existente na GKB pode ser extraído com o uso do programa *Geographic Ontology Generator* (GOG), o qual exporta o conteúdo da GKB para padrões internacionais de codificação de ontologias. O GOG tem sido usado e estendido para gerar a Geo-Net-PT e a WGO.

5.4 Extração de Informação Geográfica

5.4.1 Extração de Entidades Geográficas

O objetivo principal do algoritmo que concretiza o módulo de EI é extrair o máximo de informação geográfica presente em uma sentença, parágrafo ou texto. Esse algoritmo recebe como entrada um corpus, padrões e a lista de conceitos e ocorrências de geo-ontologias presentes na GKB. O resultado do algoritmo são arbustos geográficos extraídos das sentenças do corpus.

O algoritmo lê o corpus sentença a sentença e, se encontra um nome de local, verifica se o mesmo existe numa geo-ontologia. Caso exista, armazena o local numa variável temporária que será utilizada na formação da tripla que pode se formar. Caso o algoritmo encontre outro nome de local na sentença, forma-se uma tripla com os dois locais encontrados, e o relacionamento é definido conforme os seguintes casos:

- **Inferido a partir da geo-ontologia:** por exemplo, na sentença 'A freguesia de Vilar de Nantes em Chaves foi a mais atingida pelo mau tempo', o algoritmo procura os nomes 'Vilar de Nantes' e 'Chaves' na geo-ontologia e verifica que 'Vilar de Nantes' está incluída no concelho de 'Chaves'.

- **Indefinido:** caso não seja encontrado na geo-ontologia. Nessa situação, o relacionamento deve ser validado por um ser humano.

Os padrões usados como entrada no algoritmo foram sendo identificados a partir de experimentos realizados com versões preliminares do SEI-Geo. Os padrões utilizados pelo SEI-Geo nos experimentos descritos no Capítulo 6 são introduzidos a seguir juntamente com o algoritmo específico relacionado com cada padrão:

- **NTE ou nomes de conceitos geográficos:** nomes de conceitos de uma geo-ontologia existente mais conceitos complementares detectados pelo SEI-Geo, mas ausentes nas geo-ontologias. O algoritmo identifica todos os NTE definidos na geo-ontologia que estão presentes na sentença. Sempre que o algoritmo encontra um NTE, ele verifica se esse conceito é sucedido por um NE. Toda vez que ele encontra um NE, o par [NTE,NE] é armazenado

numa lista que irá formar um arbusto quando novos pares ou nomes de locais forem encontrados.

- **Padrões do tipo Hearst traduzidos para o português e estendidos:** são aqueles definidos em Hearst (1992) acrescentados de algumas variantes adaptadas ao português (p. ex. 'é o distrito', 'é um concelho' e 'é uma das cidades'). O algoritmo utiliza os padrões do tipo Hearst de duas formas '[NE] é um (d[eao]s?)? [NTE]' e '[NTE] tal(is) como [NE]'. Para cada padrão encontrado na frase, o algoritmo extrai locais na forma [NTE + NE].

- **Relacionamentos métricos, direcionais, *fuzzy* e orientação:** relacionamentos métricos descrevem proximidade (p. ex. 'km', 'minutos' e 'cerca de'), direcionais (p. ex. 'ao lado', 'atrás' e 'em frente'), *fuzzy* proximidade através da utilização de termos qualitativos e imprecisos (p. ex. 'próximo', 'perto' e 'acima') e orientação, que foram inicialmente definidos como direcionais no trabalho de Güting (1994), são expressos através de cardinais (p. ex. 'norte', 'sul', 'leste' e 'oeste'). Todos esses padrões são descritos em Delboni (2005) como aqueles que geram melhores resultados quando uma pessoa deseja expressar posicionamento.

- **Substantivos:** que frequentemente são sucedidos por nomes de locais incluem 'guerra', 'periferia' e 'herdade'.

- **Advérbios:** que são constantemente seguidos de nomes de locais incluem 'cá', 'aqui' e 'lá'.

- **Verbos:** que indicam presença de conteúdo geográfico em uma sentença incluem (p. ex. 'mora', 'nasce', 'localizad[ao]s?' e 'situad[ao]s?'). O padrão também inclui variações de tempo, gênero e número dos verbos.

- **Locativos:** que indicam presença de conteúdo geográfico em uma sentença incluem 'em', 'na' e 'no'.

- **Nomes de Entidades:** das geo-ontologias.

Exceto os padrões do tipo Hearst, todos os demais são imediatamente seguidos por preposição antes do nome de local.

Para cada padrão encontrado, o algoritmo procura NTE ou NE que sucedam esse padrão. Se encontra um NTE, o algoritmo passa para o próximo caso, pois locais com NTE já foram extraídos no primeiro caso (NTE ou nomes de conceitos da geo-ontologia). Se um NE é encontrado, o algoritmo armazena-o na lista na qual o arbusto está se formando. Após a análise de todas as sentenças pelo algoritmo, o resultado final consiste de arbustos que são enviados ao módulo de integração de informação.

A seguir são listados os termos utilizados nos padrões, os quais foram identificados a partir de observações junto aos dados extraídos por versões preliminares do SEI-Geo:

Advérbio: cá, aqui, lá e longe

Conceito: conceitos de geo-ontologias

***Fuzzy*:** antes, depois, acima, abaixo, próxima, próximo, perto e proximidades

Hearst: é um(a) 'Conceito', 'Conceito' tal(is) como

Locativo: em, na, nas, no, nos

Métrico: distante(s), distância, km(s), kilómetro(s), quilómetro(s), minuto(s), minuto(s), metro(s)

Orientação: norte, sul, leste, oeste, nordeste, noroeste, sudeste, sudoeste

Substantivo: água(s), afogada(s), afogado(s), beira(s), cabo(s), capital(ais), eleição(ões), favela(s), herdade(s), guerra(s), litoral(ais), margem(ns), natural(ais), penitenciária(s), periferia(s), prefeito(s), ex-prefeito(s), praia(s), procedente(s)

Verbo: chegar, falecer, ir, localizar, morar, morrer, mudar, nascer, ser, situar, sediar, realizar, viver, voltar e vir. O padrão dos Verbos inclui variações de tempo, gênero e número.

Os algoritmos apresentados a seguir são genéricos quanto ao uso de padrões e geo-ontologias. O Algoritmo 1 formaliza a fase de identificação de locais no módulo de EIG do SEI-Geo. A sintaxe de 'w[+1]' significa a palavra sucessora daquela que está sendo comparada no ciclo (*for*). Por exemplo, no seguinte excerto de uma sentença '... perto de Aveiro ...', se 'perto' é o padrão sendo comparado, 'w[+1]' é igual a 'de'.

Algoritmo 1 Algoritmo para reconhecimento de locais implementado no SEI-Geo.

```
1: O = {ocorrências de uma geo-ontologia}
2: P = {Adjetivo ∪ Adverbio ∪ Conceito geografico ∪ Fuzzy ∪ Hearst ∪ Metrico ∪
   Orientacao ∪ Substantivo ∪ Verbo}
3: S = {frases do texto a analisar}
4: Prep = {a,ao,de,da,das,do,dos,entre,na,nas,no,nos,em,à,para,pra}
5: L = ∅ //L = Locais identificados
6: for all s ∈ S do
7:    for all w ∈ s do
8:       if w ∈ P then
9:          EC = identificaEC(w[+1],s)
10:         if EC != null then
11:            L = L ∪ Algoritmo 2 (EC)
12:         end if
13:      else if w ∈ O then
14:         EC = identificaEC(w,s)
15:         if EC != null then
16:            L = L ∪ Algoritmo 2 (EC)
17:         end if
18:      end if
19:   end for
20: end for
21:
22: sub identificaEC(w,s) {
23: for all w ∈ s do
24:    if w ∈ {Prep ∪ [^ ([0-9]||[A-Z]) ∪ length(w) ≥ 2)} then
25:       EC += w
26:    end if
27:    if EC[0] ∈ Prep then
28:       EC = EC[1,length(EC)]
29:    end if
30:    if EC[-1] ∈ Prep then
31:       EC = EC[0,length(EC)-1];
32:    end if
33: end for
34: return EC
35: }
```

A fase de identificação de locais recebe como entrada ocorrências de geo-ontologias, padrões que incluem termos frequentemente utilizados ao redor de nomes de locais em textos e preposições que ocorrem em nomes de locais. Toda vez que um padrão é encontrado numa frase, o algoritmo invoca a função 'identificaEC' que identifica e retorna uma EC ou 'null', caso não seja um nome candidato a local. Essa função encontra os delimitadores da EC, ou seja, o início e o fim da mesma através de preposições e termos cuja primeira letra é maiúscula e seu comprimento é maior ou igual a dois.

Após encontrar uma EC, o Algoritmo 1 invoca a função de reconhecimento de locais[1], descrita no Algoritmo 2. Caso a palavra que está sendo comparada com os padrões não seja um padrão e sim um nome que está presente em geo-ontologias, o algoritmo verifica se a próxima palavra da sentença faz parte do nome. Se fizer, invoca a função 'identificaEC'. Senão, assume a palavra como nome de local e invoca a função de reconhecimento de locais, descrita no Algoritmo 2.

Algoritmo 2 Algoritmo para classificação de locais implementado no SEI-Geo.

1: EC = {nome extraído do texto = entidade candidata}
2: O_{adm} = {ocorrências do domínio administrativo de uma geo-ontologia}
3: O_{fis} = {ocorrências do domínio físico de uma geo-ontologia}
4: **if** EC $\in O_{adm}$ **then**
5: EG = {id do pai mais acima na hierarquia da O_{adm}}
6: **else if** EC $\in O_{fis}$ **then**
7: EG = {id do pai mais acima na hierarquia da O_{fis}}
8: **else**
9: EG = {novo id do domínio administrativo}
10: **end if**
11: return EG

A partir de uma EC o Algoritmo 2 consulta a geo-ontologia e, caso encontre o nome nessa geo-ontologia, verifica se esse nome está no domínio administrativo ou físico. Se encontra, atribui o identificador da entidade geográfica com conceito mais alto na hierarquia da geo-ontologia. Por exemplo, se encontra a EC 'França', atribui o conceito 'país' e não 'cidade' ou 'vila'. Se não encontra, tenta atribuir

[1]Essa função foi colocada num algoritmo a parte apenas por razão de espaço, pois dificultaria a leitura.

o identificador da entidade geográfica do domínio físico da WGO. Caso o nome não esteja na geo-ontologia, o algoritmo atribui um novo identificador do domínio administrativo para a EG.

Uma das vantagens de utilizar geo-ontologias na fase de reconhecimento de locais é o fato de se reconhecer um local com um nível mais específico de granularidade. Ao invés de reconhecer o local como administrativo ou físico, é possível reconhecê-lo como uma freguesia, localidade ou lago, por exemplo.

Um exemplo de extração de locais de uma sentença é: 'Belas é uma freguesia pertencente ao concelho de Sintra localizado a 30 km de Lisboa'. O algoritmo identifica que 'é uma freguesia' é um padrão do tipo Hearst e armazena o par 'freguesia Belas' na lista que formará o arbusto. Em seguida, verifica que 'concelho' é um conceito pertencente à geo-ontologia e que 'Sintra' é um NE e armazena o par 'concelho Sintra' na lista. Depois, identifica o padrão métrico 'km' e extrai o NE 'Lisboa'. Finalmente, o arbusto resultante na lista é composto por três EC: 'freguesia Belas', 'concelho Sintra', e 'Lisboa'. É importante observar que o nome 'Lisboa' é armazenado na lista sem um conceito associado. Na fase de integração de conhecimento geográfico, esse nome geográfico deve ser desambiguado de acordo com os conceitos associados a ele nas geo-ontologias.

Finalmente, é importante notar que os Algoritmos 1 e 2 são algoritmos genéricos que recebem geo-ontologias e reconhecem EG. Ao substituir a geo-ontologia (O) pelas geo-ontologias WGO e Geo-Net-PT tem-se esses algoritmos apresentados com as geo-ontologias utilizadas no SEI-Geo, tornando-os mais próximos do sistema real. A descrição dos Algoritmos 1 e 2 com a WGO e a Geo-Net-PT está publicada em Chaves (2008).

5.4.2 Detecção e Reconhecimento de Relacionamentos

A tarefa de Detecção e Reconhecimento de Relacionamentos (DRR) tem a cargo o relacionamento entre pares de entidades num texto. Nessa tese, eu concretizo essa tarefa no domínio geográfico com o objetivo de detectar e reconhecer relacionamentos entre pares de EG. Os tipos de relacionamentos são detectados com o auxílio de padrões.

A DRR envolvendo EG tem sido realizada geralmente com o apoio de entidades pertencentes à categoria pessoa (p. ex. 'Ana' nasceu em 'Évora'.)

e à categoria organização (p. ex. A 'PT' fica perto do 'Saldanha'.) (Agichtein e Gravano, 2000; Culotta e Sorensen, 2004).

Na avaliação de sistemas de REM ACE (*Automatic Content Extraction*[1]) essa tarefa foi definida com sete tipos de entidades distintas (p. ex. pessoa, agente-artefato e local). Além disso, o corpus foi composto por notícias de jornais em inglês.

Relacionamentos entre locais podem estar entre os mais frequentes em um corpus. Por exemplo, nas 1.437 sentenças presentes no conjunto de dados do *Text REtrieval Conference* (TREC), os relacionamentos mais freqüentes são entre locais, 406 ocorrências (Roth e Yih, 2004).

Já a tarefa de reconhecer relacionamentos entre EM em textos em português ganhou mais atenção no Segundo HAREM realizado em abril de 2008, com a tarefa ReRelEM (Reconhecimento de Relações entre Entidades Mencionadas), na qual participaram três sistemas. O SEI-Geo foi avaliado nessa tarefa e reconheceu relacionamentos de inclusão entre locais (ver Capítulo 6, seção 6.6.3.2).

Os Algoritmos 1 e 2 descrevem o processo de identificação e classificação de locais em texto. O reconhecimento de relacionamentos ocorre durante o processo de identificação, pois na fase de detecção de uma EC, o padrão que serviu como suporte para a identificação pode ser elevado a relacionamento. Por exemplo, na frase 'Setúbal é perto de Lisboa.' o Algoritmo 1 reconhece 'Setúbal' como uma EG e quando processa a palavra 'perto', verifica que é um átomo de um padrão catalogado na lista dos relacionamentos semânticos. Assim, armazena esse relacionamento e segue a procura de outra EC ou EG para completar uma tripla. Sempre que encontra mais uma EC ou EG, o algoritmo assume aquele relacionamento entre as EC ou EG reconhecidas.

A lista completa dos padrões lexicais utilizados pelo SEI-Geo na fase de integração de conhecimento que são transformados em relacionamentos semânticos é dada a seguir:

Direção: lado, atrás, defronte, frente.

Fuzzy: antes, depois, acima, abaixo, próxima, próximo, perto, proximidades.

Métrico: distante(s), distância, km(s), kilómetro(s), quilómetro(s), metro(s), minuto(s).

[1] http://nist.gov/speech/tests/summaries/2005/ace05rdr.htm

Orientação: norte, sul, leste, oeste, nordeste, noroeste, sudeste, sudoeste.

Substantivos: beira(s), margem(ns), periferia(s), capital(ais), litoral(ais).

No caso do padrão métrico, o relacionamento é composto pela palavra ou cadeia lexical (p. ex. 'poucos', '16' e 'cinco') que antecede o padrão mais a palavra do padrão quando esse padrão for 'km(s)', 'kilómetro(s)', 'quilómetro(s)', 'metro(s)' ou 'minuto(s)'.

O Algoritmo 3 formaliza o processo de reconhecimento de relacionamentos entre locais implementado no SEI-Geo. Esse algoritmo recebe como entrada uma geo-ontologia (O), triplas extraídas do texto (T), padrões léxico-sintáticos (P) e padrões que são transformados em relacionamentos (R). Para cada tripla, o algoritmo verifica se a mesma existe em O. Se existe, o algoritmo reconhece-a como existente. Caso contrário, o algoritmo verifica se o relacionamento entre as duas EG na tripla pertence ao conjunto dos padrões. Se existir, a tripla é considerada como nova.

Algoritmo 3 Algoritmo para reconhecimento de relacionamentos entre locais implementado no SEI-Geo.

```
1: O = {ocorrências de uma geo-ontologia}
2: T = {} //Triplas
3: P = {Adjetivo ∪ Adverbio ∪ Conceito geografico ∪ Fuzzy ∪ Hearst ∪ Metrico
   ∪ Orientacao ∪ Substantivo ∪ Verbo}
4: R = {} //Padrões que são transformados em relacionamentos
5: for all t ∈ T do
6:    if t ∈ O then
7:       t reconhecida como existente
8:    else if  R ⊂ P  then
9:       t nova
10:   end if
11: end for
```

5.5 Integração de Conhecimento Geográfico

Após a fase de extração de informação e formação de arbustos é necessária a realização de um processo de integração de conhecimento. O processo de

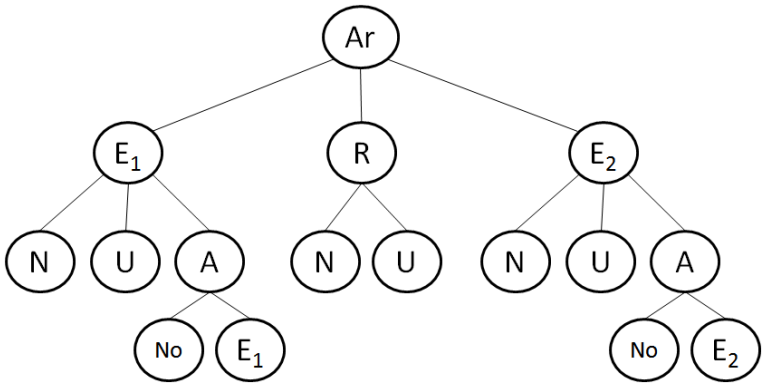


Figura 5.3: Espaço de possibilidades de integração de conhecimento geográfico.

integração de conhecimento permanece um problema em aberto na literatura. A maior parte das metodologias de extração de informação raramente alcançam o processo de integração de informação com o conhecimento disponível em estruturas formais (informação já compreensível por máquinas), fazendo com que a informação extraída fique sub-utilizada. Tão importante quanto extrair informação nova e útil de textos, é sua ligação ao conhecimento já existente de modo legível por máquina. Um dos objetivos do SEI-Geo é expandir o conhecimento presente em geo-ontologias com o conhecimento novo extraído de textos.

A Figura 5.3 ilustra um dos principais problemas que o algoritmo de integração de informação deve resolver. Um arbusto é composto por uma ou mais triplas. Cada tripla detectada no texto possui duas entidades e um relacionamento. Em relação a uma geo-ontologia, cada entidade pode ser nova (N), única (U) ou ambígua (A). Caso seja ambígua, a ambigüidade pode ocorrer somente (1) no nome (No), quando o texto fornece apenas um nome e esse é traduzido em vários identificadores de entidade com diferentes tipos de entidade ou (2) a entidade (tipo + nome) é traduzida em diferentes identificadores de entidades da geo-ontologia com o mesmo tipo encontrado no texto. Os relacionamentos podem ser novos (N) ou únicos (U).

Eu trato o problema da integração de conhecimento usando as seguintes estratégias.

Arbusto completo na geo-ontologia: nesse caso não existe nova informação a integrar, apenas o arbusto é confirmado como válido. Esse fato é usado

para avaliar a capacidade do algoritmo de extração de informação de extrair arbustos corretos.

Arbusto parcial na geo-ontologia: nesse caso há duas possibilidades:

1. Apenas uma entidade da tripla na geo-ontologia: a entidade que não está na geo-ontologia é validada por uma pessoa. Se realmente for uma entidade geográfica a tripla será integrada na geo-ontologia, caso contrário é descartada;
2. Ambas as entidades da tripla estão na geo-ontologia e o relacionamento está fora da geo-ontologia: adiciona um novo relacionamento entre as entidades existentes na geo-ontologia.

Arbusto fora da geo-ontologia: quando não há nomes de ocorrências comuns entre o arbusto e a geo-ontologia, ele pode pertencer a geografia portuguesa ou pode ser formado por nomes da geografia mundial. Nesses casos, é necessário a validação por uma pessoa. Se pertencer a geografia portuguesa, o arbusto deve ser inserido na geo-ontologia, caso contrário o arbusto deve ser inserido na geo-ontologia mundial.

5.5.1 Tratamento de Ambigüidade Geográfica

A ambigüidade é uma característica comum da LN. No contexto geográfico, substantivos como `local`, `área` e `região`, entre outros são utilizados de modo altamente ambíguo (Bennett e Agarwal, 2007). Existem dois principais tipos de ambigüidade na tarefa de reconhecimento de locais, conforme descrito no capítulo 2, na seção 2.4.1: ambigüidade por referente e multiplicidade de referências.

A ambigüidade por referente ocorre quando uma entidade ou um nome geográfico designa mais de um local. Nesse caso, a informação geográfica adicional a qualquer entidade geográfica, extraída de uma sentença, tem um papel fundamental no processo de desambiguação. Por exemplo, se 'Belém' é mencionada numa sentença com a 'cidade de Lisboa', é possível inferir que a 'Belém' em questão é a freguesia em Portugal e não a cidade no Brasil.

O algoritmo proposto para o tratamento dos casos ambíguos usa a informação geográfica complementar na mesma sentença mais os identificadores presentes em

geo-ontologias. Quando nenhum identificador é encontrado nas geo-ontologias, o arbusto ou a tripla extraída deve ser validada por uma pessoa.

O caso da multiplicidade de referências, no qual o mesmo local possui mais de um nome, não é tratado pelo SEI-Geo.

O SEI-Geo pode reconhecer nomes de organizações e pessoas multi-palavra que contenham nomes de locais formando o nome da organização ou do local. Essa capacidade de distinguir nomes de organizações e de pessoas de nomes de locais permite ao SEI-Geo aumentar a precisão em termos de EIG e RIG. A sobreposição (ambigüidade) entre nomes de organizações, pessoas e locais foi mensurada no Capítulo 4 e evidenciou que esses casos de ambigüidade ocorrem em quantidade considerável em textos em português e merecem ser tratados explicitamente por sistemas de EIG e RIG.

O algoritmo do SEI-Geo lê as sentenças átomo a átomo. Toda vez que o algoritmo encontra uma EM multi-palavra (p. ex. 'X Y Z') e não reconhece ela como local, ele passa para o próximo átomo do texto após o último átomo da EM multi-palavra, nesse caso após o Z. Assim, mesmo 'Y Z' ou 'Z' sendo nomes de locais, o algoritmo não os considera como locais. Essa estratégia de reconhecimento de EM permite que o algoritmo seja facilmente alterado para reconhecer nomes de locais no meio de nomes de organizações e pessoas.

O reconhecimento de nomes de locais como parte de nomes de organizações é útil para sistemas de REM. Por exemplo, a 'Farmácia Campo de Ourique' possui o nome do local onde ela está localizada e, nesse caso, o nome 'Campo de Ourique' passa a ser útil no processo de formação de arbustos a partir dos nomes de locais presentes no texto.

O Algoritmo 4 apresenta a formalização do módulo de integração de conhecimento do SEI-Geo.

Algoritmo 4 Algoritmo para integração de conhecimento geográfico extraído de texto em geo-ontologias.

```
1: P = {padrões} //todos os descritos na Seção 5.4.2
2: O = {ocorrências de geo-ontologias}
3: S = {sentenças do texto}
4: EG_s = {entidades geográficas extraídas das sentenças}
5: Ar = {} //Arbustos
6: T = {} //Triplas
7: for all s ∈ S do
8:    if |EG_s| ≥ 2 then
9:       if EG ∈ O then
10:         {ids} = consulta-geo-ontologia(EG_s)
11:         T = ids-EG_1,rel,ids-EG_2 //tripla existente numa geo-ontologia
12:      else if P ∈ s then
13:         T = EG_1,P,EG_2
14:      else
15:         T = EG_1,undef,EG_2
16:      end if
17:      Ar = Ar ∪ T
18:   end if
19: end for
20: for all T ∈ Ar do
21:    inseriu = falso
22:    for all EG ∈ T do
23:       if (EG ∈ O) and (! inseriu) then
24:          insere(T,O)
25:          inseriu = verdadeiro
26:       end if
27:    end for
28:    if ! inseriu then
29:       avaliacao-manual
30:    end if
31: end for
```

Esse algoritmo recebe como entrada os conceitos (C) da geo-ontologia que incluem aqueles já povoados e os vazios (sem ocorrências) além das ocorrências

dos domínios administrativo da WGO e da Geo-Net-PT e físico da WGO. Para cada sentença do texto, se existem pelo menos duas EG, uma tripla é formada. Se existe um padrão na sentença numa posição antes da segunda EG reconhecida, esse padrão é convertido num relacionamento (quando possível, conforme descrito na seção 5.4.2). Se não existe padrão, não há evidência na sentença para a inserção automática do relacionamento na tripla e o relacionamento é assumido como desconhecido. Se a tripla não está presente no arbusto, a mesma é inserida nele.

Finalmente, após todos os arbustos formados o algoritmo percorre um a um e faz a integração na geo-ontologia na qual exista pelo menos uma EG. Caso a tripla do arbusto não tenha nenhuma EG nas geo-ontologias, essa tripla é avaliada manualmente.

Nas triplas, cujas EG correspondem ao mesmo NTE (p. ex. duas aldeias), experimentos preliminares realizados por mim identificaram que a co-ocorrência de duas entidades do mesmo NTE em frases não refere normalmente qualquer relação entre elas. Tal fato dificulta sua inserção na geo-ontologia, fazendo com que esse tipo de tripla seja desde logo descartada.

5.6 Resumo e Conclusões

Um dos objetivos desse capítulo foi atacar o problema de identificação e classificação de informação geográfica em textos em português. Essa informação geralmente está relacionada nos textos e frequentemente os relacionamentos ficam subutilizados. O SEI-Geo contém algoritmos para tornar os relacionamentos geográficos úteis para processamento automático. O conhecimento geográfico presente nos textos pode ser formalizado através de arbustos ou ser anotado nos próprios textos utilizando XML.

O algoritmo de reconhecimento de entidades geográficas utiliza uma abordagem baseada no reconhecimento de padrões e ocorrências definidas em geo-ontologias, a qual pretende identificar a maior parte dos locais presentes em textos.

O algoritmo de reconhecimento de relacionamentos explora as ocorrências e relacionamentos existentes nas geo-ontologias e serve de suporte para anotar conhecimento geográfico em textos e expandir o conhecimento de geo-ontologias.

5. EXTRAÇÃO, ANOTAÇÃO E INTEGRAÇÃO DE CONHECIMENTO GEOGRÁFICO

O algoritmo de integração de informação concretiza a hipótese de estender o conhecimento existente em geo-ontologias. Esse algoritmo é simples e constitui uma das primeiras iniciativas para extensão de conhecimento ontológico com informação textual em língua portuguesa.

O próximo capítulo apresenta a avaliação de todos os algoritmos propostos e implementados no SEI-Geo juntamente com a participação do SEI-Geo no HAREM.

Capítulo 6

Avaliação dos Métodos Propostos

6.1 Introdução

Após mensurar a informação geográfica presente em textos e descrever os métodos de extração e integração de conhecimento geográfico, este capítulo apresenta uma avaliação dos métodos propostos no Capítulo 5.

Foram realizados experimentos preliminares com o objetivo de melhorar o reconhecimento do conteúdo geográfico dos textos e avaliar as limitações dos algoritmos. Os experimentos mais relevantes, descritos no início desse capítulo, evidenciam as principais dificuldades encontradas.

Um dos objetivos da avaliação foi verificar a capacidade do SEI-Geo em enriquecer geo-ontologias. A avaliação recorreu a diversos corpora que incluem um corpus de notícias, um corpus da web e dois corpora heterogêneos com textos de vários gêneros.

Finalmente, os módulos extrator e anotador de informação geográfica do SEI-Geo foram avaliados no HAREM. Esses módulos reconhecem locais e relacionamentos de inclusão entre locais num texto. Os resultados evidenciam que o SEI-Geo alcança resultados comparáveis a sistemas de REM que representam o estado da arte para textos em português. O capítulo termina com considerações sobre o estado da arte em extração e/ou integração de informação geográfica seguido pelas conclusões.

6.2 Caracterização e Análise da Informação Geográfica Relacionada Extraída de Textos

A avaliação do algoritmo de extração de informação geográfica utilizou dois gêneros de corpora, um jornalístico e outro com textos da web (WPT 03). O corpus jornalístico foi a coleção CHAVE composta pelos textos completos dos jornais Público e Folha de São Paulo dos anos de 1994 e 1995 (Santos e Rocha, 2004).

A coleção CHAVE é a mesma utilizada no GeoCLEF, uma avaliação internacional de sistemas de RIG. O corpus CHAVE-Público-94 contêm 51.754 documentos, o CHAVE-Público-95 contém 55.073 documentos, o CHAVE-Folha de São Paulo-94 contém 51.878 documentos e o CHAVE-Folha de São Paulo-95, 52.041 documentos.

A WPT 03 conta com 3.775.611 documentos, dos quais aproximadamente 68,6% (2.590.641 documentos) estão escritos em português (Cardoso et al., 2007; Gomes e Silva, 2005; Martins e Silva, 2004). Com a eliminação de documentos duplicados, a coleção possui 1.529.758 documentos em português.

É importante observar que a caracterização dos arbustos descrita nessa seção considera o conceito de arbusto mais relaxado, apenas com entidades geográficas e sem relacionamentos. Logo, um arbusto pode ser composto por um número ímpar de entidades geográficas. Ou seja, o conceito de triplas não é considerado nessa caracterização. Existe no arbusto um conjunto n de locais com potencial relacionamento entre si.

O SEI-Geo utilizado nos experimentos descritos nessa seção foi configurado para receber como entradas de dados os padrões orientação, fuzzy, verbo, métrico e direção e a Geo-Net-PT.

As Tabelas 6.1(a) e 6.1(b) apresentam o mapeamento existente entre os locais extraídos de corpus jornalístico e da web com o domínio administrativo da geo-ontologia Geo-Net-PT. Os valores percentuais correspondem ao número de locais extraídos em cada arbusto. Na primeira coluna das Tabelas 6.1(a) e 6.1(b) está o número de locais que compõem os arbustos de cada linha. Na terceira linha estão os resultados para todos os arbustos compostos por dois locais. Por exemplo, na Tabela 6.1(a) há 49 arbustos compostos por dois locais que são ambíguos (mapeamento 1→1..n) (basta que um dos locais seja

Tabela 6.1: Caracterização dos nomes de locais presentes nos arbustos distintos extraídos de coleções de texto em relação aos seus mapeamentos na Geo-Net-PT.

(a) Extraídos do CHAVE-Público

	Mapeamentos encontrados em número e porcentagem			
Nº EC no arbusto	map. 1↔1	map. 1→1..n	s/ map.	total
2	734 (45,20)	49 (3,02)	841 (51,79)	1.624
3	52 (37,68)	7 (5,07)	79 (57,25)	138
4	97 (53,89)	7 (3,89)	76 (42,22)	180
5	18 (40,00)	1 (2,22)	26 (57,78)	45
6	16 (53,33)	6 (20,00)	8 (26,67)	30
7	16 (57,14)	2 (7,14)	10 (35,71)	28
8+	45 (88,24)	1 (1,96)	5 (9,80)	51
$\sum$	978	73	1.045	2.096

(b) Extraídos do WPT03

	Mapeamentos encontrados em número e porcentagem			
Nº EC no arbusto	map. 1↔1	map. 1→1..n	s/ map.	total
2	6.826 (53,61)	890 (6,99)	5.016 (39,40)	12.732
3	2.553 (65,51)	234 (6,00)	1.110 (28,48)	3.897
4	687 (62,68)	83 (7,57)	326 (29,74)	1.096
5	282 (56,97)	58 (11,72)	155 (31,31)	495
6	188 (68,12)	17 (6,16)	71 (25,72)	276
7	76 (45,24)	12 (7,14)	80 (47,62)	168
8+	485 (53,30)	97 (10,66)	328 (36,04)	910
$\sum$	11.097	1.391	7.086	19.574

ambíguo para estar nessa coluna da tabela), 734 estão na Geo-Net-PT com apenas uma ocorrência (mapeamento 1→1) e 841 arbustos não possuem mapeamentos nessa geo-ontologia. Em ambos os corpora o número de locais não ambíguos (mapeamento 1↔1) é maior do que o número de locais ambíguos. No caso do corpus WPT 03, os mapeamentos 1↔1 superam também o número de locais sem mapeamento, com exceção dos arbustos formados por sete locais.

Tabela 6.2: Caracterização dos arbustos distintos extraídos de coleções de textos em relação aos seus mapeamentos na Geo-Net-PT.

(a) Extraídos da CHAVE-Público

	Mapeamentos encontrados em número e porcentagem			
Nº EC no arbusto	todas EC mapeadas	parc. mapeadas	não mapeadas	total
2	267 (32,88)	249 (30,67)	296 (36,45)	812
3	6 (13,04)	24 (52,17)	16 (34,78)	46
4	12 (26,67)	23 (51,11)	10 (22,22)	45
5	1 (11,11)	4 (44,44)	4 (44,44)	9
6	1 (20,00)	4 (80,00)	0 (0,00)	5
7	1 (25,00)	2 (50,00)	1 (25,00)	4
8+	1 (1,96)	45 (88,24)	5 (9,80)	51
$\sum$	289	351	332	972

(b) Extraídos da WPT 03

	Mapeamentos encontrados em número e porcentagem			
Nº EC no arbusto	todas EC mapeadas	parc. mapeadas	não mapeadas	total
2	2.694 (42,32)	2.328 (36,57)	1.344 (21,11)	6.366
3	586 (45,11)	593 (45,65)	120 (9,24)	1.299
4	104 (37,96)	142 (51,82)	28 (10,22)	274
5	26 (26,26)	62 (62,63)	11 (11,11)	99
6	18 (39,13)	23 (50,00)	5 (10,87)	46
7	3 (12,50)	17 (70,83)	4 (16,67)	24
8+	17 (18,89)	62 (68,89)	11 (12,22)	90
$\sum$	3.448	3.227	1.523	8.198

As Tabelas 6.2(a) e 6.2(b) apresentam uma caracterização dos arbustos extraídos de dois corpora distribuídos por número de locais em cada arbusto. Os valores percentuais correspondem ao número de locais extraídos em cada arbusto. Em ambos os corpora o número de mapeamentos parciais é maior do que os mapeamentos totais, independente do número de locais que compõem o arbusto. No corpus WPT 03, 2.694 arbustos possuem todos os locais com identificadores na geo-ontologia, 2.328 estão parcialmente na geo-ontologia e 1.344 estão fora da geo-ontologia, ou seja, não contém nenhum identificador na geo-ontologia. O algoritmo extraiu 6.366 arbustos compostos por dois locais, ou seja, mais de 77% dos arbustos extraídos são compostos por dois locais, enquanto mais de 15% são compostos por três locais. O restante do conteúdo das Tabelas 6.2(a) e 6.2(b)

Tabela 6.3: Estatística descritiva dos arbustos extraídos do corpus jornalístico Público e do corpus da web WPT 03 em relação às suas presenças em uma geo-ontologia.

Descrição	Público	WPT 03
# médio de locais por doc. com local	2,12	2,36
Variância	1,28	1,29
Desvio padrão	1,13	1,13
# total de locais	2.209	19.406
# total de arbustos	1.040	8.222

contém os valores para os arbustos compostos de três a oito ou mais locais.

A Tabela 6.3 apresenta estatísticas dos arbustos extraídos de texto. A média de locais por documento com local é de pouco mais de dois locais em ambos os corpora. O baixo desvio padrão evidencia que o número de locais na maioria dos arbustos está bastante próximo de dois.

6.3 Contribuição dos Padrões para Extração de Informação e Formação de Triplas

O SEI-Geo usa vários padrões para identificar locais em textos. Os locais identificados e reconhecidos são anotados como entidades geográficas (EG) e, em seguida, passam pelo processo de formação de triplas, o qual verifica se as EG formam uma tripla. Nesse sentido, é relevante mensurar a contribuição de cada padrão, ou seja, quantas EG formarão triplas. Essa informação é relevante para mensurar se EG presentes em uma mesma sentença estão relacionadas. O objetivo dos experimentos descritos a seguir é analisar a contribuição de cada padrão sobre diferentes corpora.

As Tabelas 6.4 e 6.5 apresentam os resultados para as coleções com textos em português utilizadas no GeoCLEF.

Tabela 6.4: Contribuição dos padrões para identificação e reconhecimento de locais na coleção CHAVE-Folha de São Paulo.

(a) CHAVE-Folha de São Paulo-1994

Padrão	EC	EG	Tripla	EG/EC(%)	Tripla/EG
Advérbio	599	595	528	0,99	0,89
Conceito	20.752	20.518	13.054	0,99	0,64
Fuzzy	2.073	2.070	1.858	1,00	0,90
Métrico	869	847	785	0,97	0,93
Ocorrência	130.477	128.004	40.705	0,98	0,32
Orientação	3.811	3.754	3.325	0,99	0,89
Locativo	93.538	93.202	27.064	1,00	0,29
Substantivo	3.733	3.735	3.076	0,99	0,80
Hearst	270	267	267	0,99	1,00
Verbo	7.231	7.212	5.607	1,00	0,78
Total padrões	263.353	260.184	96.269	0,99	0,37

(b) CHAVE-Folha de São Paulo-1995

Padrão	EC	EG	Tripla	EG/EC(%)	Tripla/EG
Advérbio	546	543	491	0,99	0,90
Conceito	19.812	19.606	12.920	0,99	0,66
Fuzzy	1.975	1.970	1.760	1,00	0,89
Métrico	1.291	1.267	1.138	0,98	0,90
Ocorrência	129.932	127.396	41.648	0,98	0,33
Orientação	4.604	4.538	3.829	0,99	0,84
Locativo	94.802	94.444	27.797	1,00	0,29
Substantivo	3.417	3.405	2.941	1,00	0,83
Hearst	254	250	250	0,98	1,00
Verbo	7.355	7.333	5.797	1,00	0,79
Total padrões	263.988	260.752	98.571	0,99	0,38

Tabela 6.5: Contribuição dos padrões para identificação e reconhecimento de locais na coleção CHAVE-Público.

(a) CHAVE-Público-1994

Padrão	EC	EG	Tripla	EG/EC(%)	Tripla/EG
Advérbio	441	429	402	0,97	0,94
Conceito	29.500	29.166	20.643	0,99	0,71
Fuzzy	4.732	4.720	4.209	1,00	0,89
Métrico	951	941	900	0,99	0,96
Ocorrência	224.838	219.015	88.313	0,97	0,40
Orientação	2.660	2.611	2.416	0,98	0,93
Locativo	150.879	150.000	53.888	0,99	0,36
Substantivo	6.142	6.108	5.296	0,99	0,86
Hearst	322	314	314	0,98	1,00
Verbo	8.715	8.651	7.306	0,99	0,84
Total padrões	429.180	421.955	183.687	0,98	0,44

(b) CHAVE-Público-1995

Padrão	EC	EG	Tripla	EG/EC(%)	Tripla/EG
Advérbio	413	411	391	1,00	0,95
Conceito	38.068	37.677	25.925	0,99	0,69
Fuzzy	4.935	4.913	4.454	1,00	0,91
Métrico	1.150	1.138	1.062	0,99	0,93
Ocorrência	243.684	237.475	97.554	0,97	0,41
Orientação	3.217	3.180	2.906	0,99	0,91
Locativo	159.389	158.479	58.820	0,99	0,37
Substantivo	6.292	6.259	5.458	0,99	0,87
Hearst	342	334	334	0,98	1,00
Verbo	9.571	9.495	8.071	0,99	0,85
Total padrões	467.061	459.361	204.975	0,98	0,45

Os resultados dessas tabelas evidenciam que os padrões mais produtivos (i.e. os que geram mais EC) são locativo, ocorrência e conceito em todas as coleções. Desses padrões mais produtivos, as ocorrências das geo-ontologias são aquelas que geram mais EG.

Por outro lado, os padrões advérbio, fuzzy, métrico e Hearst apresentam valores elevados na precisão, ou seja, a maior parte das EC identificadas com esses padrões é realmente uma EG. Considerando os resultados das quatro coleções o número médio de EG que forma uma tripla (última coluna das tabelas) varia de 0,37 a 0,45.

O padrão locativo é um dos mais produtivos padrões na identificação de EG que formam triplas, ou seja, em número absoluto só gera menos triplas do que as ocorrências. O padrão Hearst é o menos produtivo, conforme já esperado, uma vez que raramente ocorre em textos de domínio específico. Brewster et al. (2003) realizaram um pequeno experimento com padrões Hearst aplicados ao domínio da Biologia e também constataram que sua ocorrência é rara, não sendo suficiente para encontrar 10 pares de termos de uma ontologia de domínio.

Os altos valores percentuais para o quociente entre EC e EG podem ser explicados pelo fato de que o procedimento que verifica se uma EC é uma EG apenas detecta se a EC não faz parte de uma lista negra de termos (ver Apêndice B) e se possui um comprimento maior que um.

6.4 Análise dos Nomes, Entidades Geográficas e Triplas dos Arbustos

Essa seção tem como objetivo apresentar resultados da aplicação do SEI-Geo ao corpora do GeoCLEF. O resultados das Tabelas 6.6 e 6.7 contabilizam ocorrências distintas dentro de cada tripla, mas não entre arbustos. Por exemplo, se o nome do país 'Peru' está em dois arbustos distintos, ele é contabilizado duas vezes. A Tabela 6.6 apresenta a distribuição dos nomes de locais encontrados nos quatro corpora e sua distribuição pelas geo-ontologias.

Tabela 6.6: Total de nomes de locais nos arbustos da coleção CHAVE.

	Total de Nomes			
Coleção	WGO Adm	WGO Fis	Geo-Net-PT	Fora das Onto.
CHAVE-FSP-94	154.676	3.481	18.881	62.086
CHAVE-FSP-95	157.620	3.523	16.635	61.621
CHAVE-Público-94	238.335	10.168	51.448	89.628
CHAVE-Público-95	245.064	9.370	68.848	95.889

A Tabela 6.7 apresenta o número de features e triplas encontradas nos corpora. Os corpora do jornal Público mostram-se mais ricos em informação sobre a geografia física. Em todos os quatro corpora o SEI-Geo encontrou conhecimento geográfico fora das geo-ontologias.

Tabela 6.7: Total de EG e triplas nos arbustos da coleção CHAVE.

	Entidades Geográficas				Triplas	Arbustos
Coleção	Única	Ambígua	Fora das Onto.	Total		
CHAVE-FSP-94	42.985	5.543	45.654	94.182	47.091	33.752
CHAVE-FSP-95	44.046	5.834	46.670	96.550	48.275	34.678
CHAVE-Público-94	92.604	10.033	105.573	208.210	104.105	68.727
CHAVE-Público-95	102.114	12.958	115.830	230.902	115.451	76.023

Em ambas as Tabelas 6.6 e 6.7 o SEI-Geo encontrou mais conhecimento geográfico nos corpora do jornal Público, o que é evidenciado pelo número total de arbustos, que alcança praticamente duas vezes o valor encontrado no CHAVE-Folha de São Paulo.

6.5 Expansão de Geo-ontologias com o SEI-Geo

Nessa seção, o SEI-Geo recebe como entrada um corpus e geo-ontologias e devolve como resultado um conjunto de arbustos com as geo-ontologias enriquecidas com novos locais e relacionamentos reconhecidos no corpus. Esses locais podem ou não estar presentes na geo-ontologia. Se o SEI-Geo encontra uma ocorrência de um conceito e essa ocorrência já está na geo-ontologia, o resultado permite validar a ocorrência e a geo-ontologia não é expandida.

Para os experimentos de expansão de geo-ontologias usou-se a versão do SEI-Geo descrita no Capítulo 5 com a alteração da remoção do padrão locativo. Após análise de resultados preliminares, concluiu-se que o mesmo é bastante útil para o reconhecimento de locais, mas na geração de arbustos, os locativos geram mais triplas incorretas (falsos positivos) do que corretas. As causas principais são as EG da tripla serem reconhecidas corretamente, mas sem relacionamento real entre as mesmas e o reconhecimento de organizações ao invés de locais.

6.5.1 Expansão de Geo-ontologias com Corpus Jornalístico e da Web

A primeira avaliação foi realizada no corpus CHAVE-Público-95 e com o SEI-Geo utilizando ambas as geo-ontologias Geo-Net-PT e WGO. De um total de 50.495 arbustos, foi selecionada aleatoriamente uma amostra de 100 arbustos compostos por 143 triplas. Cada tripla dessa amostra foi avaliada manualmente de acordo com os seguintes critérios:

Integrável (I): quando as duas EG da tripla forem realmente locais e o relacionamento entre elas estiver correto.

Integrável com Assistência (IA): quando duas EG forem corretas e não existir relacionamento explícito no texto ou o algoritmo não conseguiu identificar. Nesse caso o avaliador deve inserir o relacionamento correto.

Existente (E): quando as EG e o relacionamento reconhecido entre essas EG já está em pelo menos uma das geo-ontologias.

Falso (F): quando no máximo uma EG da tripla é um local ou duas EG não possuem relacionamento no mundo real.

Das 143 triplas, duas foram avaliadas como Integráveis, 61 como Integráveis com Assistência, 19 como Existentes e 61 como Falsas. Dado o pequeno número de arbustos avaliados, já era esperado que poucos arbustos fossem integráveis diretamente nas geo-ontologias. Ainda resta um número elevado de casos com triplas falsas.

As duas triplas integráveis foram reconhecidas em arbustos distintos e são apresentadas a seguir. Ambas as triplas expandem o conhecimento presente na

WGO. Embora as EM já existissem nessa geo-ontologia, os relacionamentos que associam elas não estavam presentes.

```
<geo:shrub rdf:ID="SH_20629">
   <geo:triple>
     <geo:feature>
        <geo:name>Puerto Montt</geo:name>
        <geo:type rdf:resource="#PLACE"/>
        <geo:geo_id rdf:ID="WGO_ADM_6121"/>
     </geo:feature>
     <geo:relationship rdf:resource="#sul"/>
     <geo:feature>
        <geo:name>Santiago</geo:name>
        <geo:type rdf:resource="#ISO-3166-2"/>
        <geo:geo_id rdf:ID="WGO_ADM_9207"/>
     </geo:feature>
   </geo:triple>
</geo:shrub>
<geo:shrub rdf:ID="SH_12382">
   <geo:triple>
     <geo:feature>
        <geo:name>Porto</geo:name>
        <geo:type rdf:resource="#ISO-3166-2"/>
        <geo:geo_id rdf:ID="WGO_ADM_11264"/>
     </geo:feature>
     <geo:relationship rdf:resource="#norte"/>
     <geo:feature>
        <geo:name>Espinho</geo:name>
        <geo:type rdf:resource="#ADM_DIV"/>
        <geo:geo_id rdf:ID="WGO_ADM_12358"/>
     </geo:feature>
   </geo:triple>
  ...
</geo:shrub>
```

Os casos avaliados como integrável com assistência merecem uma análise pormenorizada e incluem: cidades com menos de 100.000 habitantes Entidades geográficas com relacionamentos inter-domínio. Por exemplo, uma tripla composta pelas EG Porto (domínio administrativo) e Douro (domínio físico), mas sem relacionamento explícito.

Alguns dos erros cometidos pelo SEI-Geo no reconhecimento de EM incluem: 'Mosteiro', 'Freitas' e 'Silveira' reconhecidos como freguesias, 'Mesa' é um local no Arizona, USA e 'Globo' reconhecido como canal da geografia física quando de fato é uma emissora de TV. Nomes de equipes de futebol que são homônimos às suas cidades sedes como 'Milan' e 'Barcelona' também são identificados como uma tripla.

Após essa singela avaliação do módulo de ICG do SEI-Geo, as próximas seções apresentam a participação do SEI-Geo em eventos de avaliação conjunta. Esses eventos além de avaliarem o SEI-Geo, permitem uma comparação com sistemas que constituem o estado da arte em REM.

6.6 Avaliação do SEI-Geo no HAREM

O REM em português é uma tarefa que tem ganho mais atenção nos últimos três anos com a criação dos eventos HAREM, que são avaliações conjuntas de sistemas de REM. Reconhecer EM em textos não é uma tarefa simples nem para humanos. Mani et al. (2008) descrevem um acordo entre anotadores humanos com medida F de 0,77 para textos em inglês. Em português ainda não temos essa medida, mas pelo inglês já é possível ter uma dimensão das dificuldades impostas pela língua natural para os sistemas de REM.

Para participar no HAREM o SEI-Geo utilizou o módulo anotador de locais e relacionamentos entre locais em textos, ao invés de simplesmente extrair conhecimento geográfico dos mesmos. O resultado do módulo de extração de informação, nesse caso, é o texto anotado com locais e relacionamentos entre locais, conforme já descrito na Seção 5.3.

A avaliação do módulo de extração de informação do SEI-Geo foi realizada com as coleções dos eventos HAREM. O SEI-Geo não participou do Primeiro HAREM nem do Mini-HAREM, mas reutilizou a coleção dourada e os programas de avaliação para avaliar o reconhecimento de locais em textos. Já no Segundo HAREM, o SEI-Geo foi um dos sistemas participantes com quatro corridas. As seções subsequentes descrevem as avaliações em cada evento.

6.6.1 Avaliação do SEI-Geo com a Coleção Dourada do Primeiro HAREM

Após a finalização da implementação do módulo de extração de informação do SEI-Geo, foi realizada uma avaliação do processo de identificação e reconhecimento de locais na coleção dourada do Primeiro HAREM para o cenário seletivo Local. Essa coleção possui 128 documentos que foram etiquetados com locais pelo SEI-Geo.

Essa avaliação foi realizada com base na arquitetura de avaliação do Primeiro HAREM apresentada no Capítulo 19 de (Santos e Cardoso, 2007) e o procedimento executado com os módulos para a obtenção dos resultados está no Apêndice C. A Tabela 6.8 apresenta os principais resultados obtidos.

Tabela 6.8: Resultado do SEI-Geo com a CD do Primeiro HAREM.

	Classificação Semântica		
	Categorias	Tipos	Plana
Total EM classificadas na CD	818	523	818
Total EM classificadas pelo SEI-Geo	598	527	598
Total Corretos	512,7	488	488,31
Espúrios	71	32	103
Em Falta	295	30	325
Precisão	0,857	0,92	0,82
Abrangência	0,63	0,93	0,60
Medida F	**0,72**	**0,93**	**0,68**
Sobre-geração	0,12	0,06	0,17
Sub-geração	0,36	0,05	0,39

O SEI-Geo alcançou um excelente resultado na classificação semântica de tipos, superando o sistema que obteve o melhor resultado nas avaliações do Primeiro HAREM (0.89) e do Mini-HAREM (0.92). A classificação semântica por categoria e plana está abaixo dos melhores sistemas do HAREM e Mini-HAREM. Os resultados indicam que deve-se melhorar a abrangência do SEI-Geo, uma vez que a medida F ficou prejudicada pela abrangência na classificação semântica por categorias e plana.

Outras corridas foram testadas com a CD do Primeiro HAREM, variando o conjunto de nomes das geo-ontologias de entrada que o SEI-Geo recebe. Contudo, os resultados não apresentaram mudanças significativas.

6.6.1.1 Análise dos Resultados do SEI-Geo com a Coleção Dourada do Primeiro HAREM

Após analisar os resultados da anotação do SEI-Geo na coleção do Primeiro HAREM, detectaram-se os seguintes casos nos quais o sistema apresenta um desempenho menos satisfatório.

- **Diferenças entre as variantes do português:** alguns nomes de locais são escritos com diferente grafia nas variantes do português europeu e brasileiro (p. ex. 'Egipto' e 'Egito', 'Irão' e 'Irã'). Como as geo-ontologias utilizadas ainda contém muito poucos nomes na variante brasileira, os nomes de locais nessa variante não são anotados pelo SEI-Geo quando eles não são precedidos ou sucedidos por padrões.
- **Tipos geográficos:** tipos geográficos que não estão presentes nas geo-ontologias também não são reconhecidos pelo SEI-Geo (p. ex. 'Barra de São Miguel', 'Alpes Dináricos', 'Floresta Amazônica' e 'Selva Amazônica').
- **Tipos de vegetação:** não são reconhecidos pelo SEI-Geo. Exemplos incluem: 'Cerrado', 'Pampa', 'Caatinga', 'Mata Atlântica' e 'Pantanal'. Esses tipos de entidades fazem parte da geografia física e só são tratados pelo sistema quando ocorrem próximos aos padrões.
- **Casos espúrios:** casos em que nomes de locais são reconhecidos com categorias diferentes de locais. A habilidade do SEI-Geo para reconhecer locais em contexto ainda tem limitações, tal fato faz com que o sistema anote EM de outras categorias como locais.
- **Locais da geografia administrativa pouco povoados:** o SEI-Geo usa a WGO, que contém locais da geografia administrativa com população superior a 100.000 habitantes. Locais com população inferior a 100.000 habitantes que não são precedidos ou sucedidos por padrões, não são reconhecidos.

Tabela 6.9: Correspondência entre os tipos de locais nas duas edições do HAREM.

Primeiro HAREM	Segundo HAREM
Administrativo	Humano
Geográfico	Físico
Alargado	-
Correio	-
Virtual	Virtual

- **Erros de grafia:** casos como 'rioTapajós', escrito sem espaço entre o tipo e o nome; 'Vietna' sem tilde na letra 'a' e 'America Latina' e 'Suecia' sem acento agudo na letra 'e' são exemplos de locais não reconhecidos pelo SEI-Geo devido a erros de ortografia. Em todos esses casos, os locais não estavam precedidos ou sucedidos por padrões.

Esses experimentos com a CD do Primeiro HAREM foram úteis para melhorar o SEI-Geo para o Segundo HAREM. A análise dos resultados permitiu minimizar os principais problemas detectados. Cabe destacar que o módulo de reconhecimento de relacionamentos semânticos não foi testado nesses experimentos, uma vez que não existia essa tarefa no Primeiro HAREM.

6.6.2 Alterações do Primeiro para o Segundo HAREM

Algumas alterações foram realizadas do Primeiro para o Segundo HAREM. A Tabela 6.9 apresenta as principais diferenças. O SEI-Geo não foi concebido para identificar locais pertencentes à categoria Virtual em nenhuma das edições.

O Segundo HAREM também incluiu o reconhecimento de subtipos para algumas categorias. O SEI-Geo reconheceu os subtipos descritos na Tabela 6.10, os quais foram retirados das diretivas do HAREM (Mota e Santos, 2008).

Tabela 6.10: Classificação de tipos e subtipos de locais reconhecidos pelo SEI-Geo no Segundo HAREM.

Tipo Humano	
Subtipo	Descrição
País	países, principados, e uniões de países, como é, por exemplo, o caso da União Europeia
Divisão	agregados populacionais como metrópoles, cidades, aldeias, vilas ou freguesias, assim como outras divisões administrativas tais como estados (Brasil), concelhos, distritos, províncias (Portugal), continentes, ou bairros fiscais
Região	localização cultural ou tradicional, sem valor administrativo, tal como a Baixa, o Grande Porto, o Médio-Oriente, o Terceiro Mundo ou o Nordeste (brasileiro)
Construção	inclui todo o tipo de construções, desde edifícios, aglomerados de edifícios ou zonas específicas de um edifício (por exemplo, sala, galeria, jardim ou piscina), a pontes, barragens, portos, etc.
Rua	inclui todo o tipo de arruamentos, como ruas, avenidas, estradas, travessas, praças, pracetas, becos, largos, etc.
Outro	outro subtipo que não esteja contemplado nos subtipos acima
Tipo Físico	
Subtipo	Descrição
Aguacurso	inclui rios, ribeiros, riachos, afluentes, quedas de água, etc.
Aguamassa	inclui lagos, mares, oceanos, golfos, estreitos, canais, bacias, barragens, etc.
Relevo	inclui montanhas, cordilheiras, montes, serras, planícies, planaltos, vales, etc.
Planeta	inclui todos os corpos celestes
Ilha	inclui ilhas e arquipélagos
Região	designa uma região geográfica/natural, tal como o Bósforo, os Bálcãs, a Meseta Ibérica, a região do Amazonas, o Deserto do Sahara, ou os continentes vistos como região da geografia física
Outro	outro subtipo que não esteja contemplado nos subtipos acima.

6.6.3 A Participação do SEI-Geo no Segundo HAREM

O SEI-Geo participou do Segundo HAREM com quatro corridas, as quais permitem verificar o quanto o sistema é dependente de geo-ontologias e se o uso de várias geo-ontologias beneficia o desempenho do mesmo.

6.6.3.1 Descrições das Corridas do SEI-Geo

As variações realizadas nas quatro corridas do SEI-Geo, são as geo-ontologias de entrada do sistema e o âmbito dos relacionamentos a serem reconhecidos. Apesar de o Primeiro HAREM ter confirmado que a combinação de geo-ontologias gera melhores resultados, eu quis confirmar se isso foi por acaso ou se realmente noutra colecção (a do Segundo HAREM) isso já poderia ser verificado novamente.

O SEI-Geo participou no Segundo HAREM com quatro corridas:

- **Corrida 1 (Geo-Net-PT):** utilizou somente a Geo-Net-PT até o nível de localidade, ou seja, conceitos e entidades geográficas acima do conceito de localidade inclusive.
- **Corrida 2 WGO - Relacionamento com Âmbito no Documento):** utilizou apenas a WGO com nomes geográficos de todo o mundo, incluindo nomes de países, cidades capital, principais regiões administrativas e cidades com mais de 100.000 habitantes. Na tarefa de ReRelEM essa corrida anota relacionamentos existentes entre locais ao longo de todo o documento.
- **Corrida 3 (Duas Ontologias - Relacionamento com Âmbito na Sentença):** o âmbito dos relacionamentos foi restrito ao nível de sentença, conforme o SEI-Geo foi projetado originalmente. Na tarefa de ReRelEM a corrida 3 anota relacionamentos existentes somente para locais que estejam na mesma sentença.
- **Corrida 4 (Duas Ontologias - Relacionamento com Âmbito no Documento):** o âmbito dos relacionamentos foi o documento completo, o que caracteriza a proposta original da tarefa ReRelEM.

Sempre que o algoritmo encontra um mesmo nome em ambas, a opção é feita pela geo-ontologia WGO, uma vez que a mesma possui conceitos que estão na

parte superior da hierarquia das ontologias. Por exemplo, 'França' é uma freguesia do 'concelho de Bragança' e um país, como país está acima na hierarquia das ontologias, o SEI-Geo assume que o nome 'França' num texto refere-se ao país e não à freguesia, a não ser que esteja precedido pelo conceito 'freguesia'.

Casos de ambiguidade entre nomes do domínio administrativo e físico, o algoritmo usa uma heurística que prioriza o domínio administrativo. Por exemplo, se um mesmo nome se refere a uma cidade e a um lago e não possui nenhum discriminador (conceito geográfico) no texto, o algoritmo assume que o nome refere à cidade. Essa abordagem também é usada em (Volz et al., 2007).

As corridas 3 e 4 utilizaram as geo-ontologias WGO e Geo-Net-PT, essa mutilada no nível de localidades, ou seja, os nomes de localidades não foram incluídos nessas corridas. Os resultados das corridas 3 e 4 para o HAREM clássico são muitos próximos, uma vez que a única variação feita foi ao âmbito dos relacionamentos. Nessa tese apresentamos somente os resultados da corrida 3.

6.6.3.2 Avaliação do SEI-Geo

Essa seção descreve a primeira participação do SEI-Geo no Segundo HAREM. Uma das funcionalidades do SEI-Geo é a extração de relacionamentos semânticos entre entidades geográficas. Uma das pistas do Segundo HAREM propõe o desafio de reconhecer relacionamentos entre EM. A participação do SEI-Geo nessa tarefa restringiu-se ao reconhecimento de relacionamentos entre entidades pertencentes à categoria local, que é um dos problemas tratados pelo SEI-Geo.

A abordagem utilizada pelo SEI-Geo na tarefa de reconhecimento de relacionamentos foi baseada em geo-ontologias. Todos os locais encontrados num documento são projetados sobre geo-ontologias com o objetivo de encontrar relacionamentos de inclusão (inclui/incluído) entre eles. Caso encontre algum relacionamento, o SEI-Geo anota o mesmo no documento. Essa abordagem permite testar até que ponto um algoritmo de reconhecimento de relacionamentos geográficos consegue ser preciso e abrangente só com o uso de geo-ontologias.

Um fator importante a destacar é o âmbito no qual um relacionamento pode ocorrer. O SEI-Geo foi desenvolvido originalmente para relacionar locais dentro de uma mesma sentença. Entretanto, de acordo com as diretivas do Segundo HAREM, relacionamentos devem ser identificados ao nível do documento.

A Tabela 6.11 apresenta os resultados alcançados no cenário seletivo 5 considerando a classificação com avaliação relaxada de ALT[1], uma vez que o SEI-Geo não usa a opção de marcação com a etiqueta ALT. A última linha da Tabela 6.11 apresenta os resultados dos melhores sistemas para cada medida.

Tabela 6.11: Resultados do cenário seletivo 5 considerando a classificação com avaliação relaxada de ALT.

	Classificação Semântica Combinada			Identificação		
Corrida	P	A	F	P	A	F
2	**0,6821**	0,5182	0,5890	0,7109	0,5346	0,6102
3	0,6801	0,5377	**0,6006**	**0,7075**	0,5552	0,6222
4	0.6726	**0.5413**	0.5999	0.7009	**0.5595**	**0.6223**
Melhor sistema	0,7105	0,7126	0,6325	0,8332	0,8134	0,7939

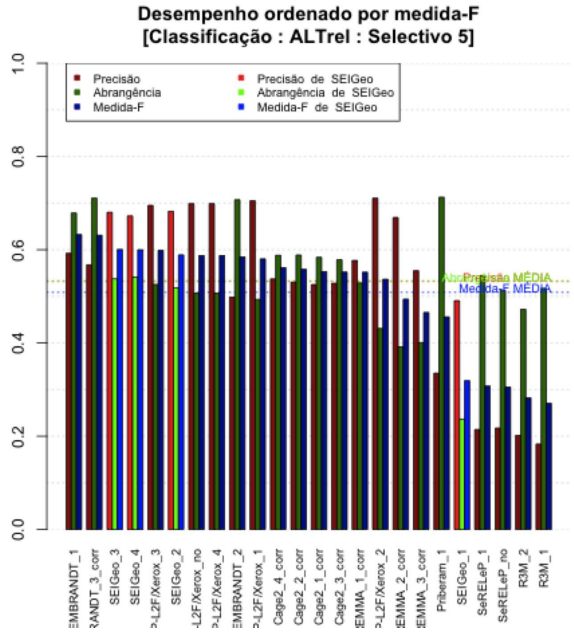
As Figuras 6.1(a) e 6.1(b) apresentam um comparativo dos resultados do SEI-Geo comparados com os demais sistemas participantes. A Figura 6.1(a) mostra que o SEI-Geo, nas duas melhores corridas (3 e 4), conseguiu atingir resultados acima da média dos sistemas em todas as medidas: precisão, abrangência e medida F. No que diz respeito à identificação, a Figura 6.1(b) apresenta o SEI-Geo com valores de precisão e medida F acima da média dos sistemas, mas com abrangência inferior, o que evidencia uma das limitações do SEI-Geo.

Além desses resultados, o SEI-Geo também alcançou o primeiro lugar na medida de precisão nos cenários Total, 2, 3, 4 e 6 para a tarefa de identificação e identificação com avaliação relaxada de ALT. Os valores da precisão nesses cenários variam de 0,86 à 0,91.

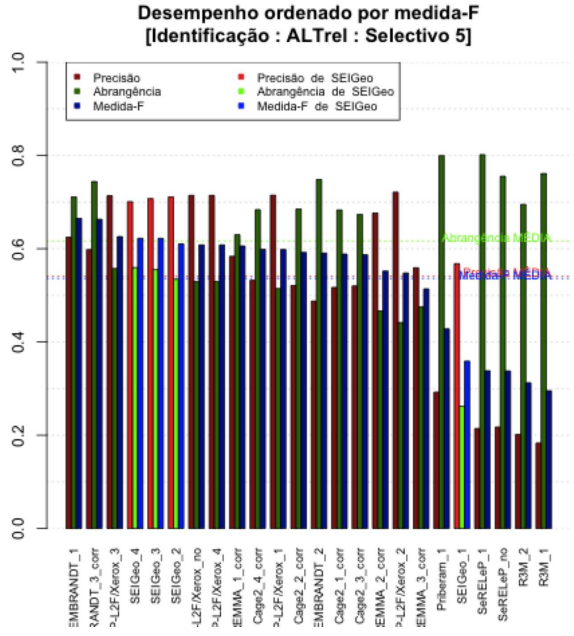
Nos resultados por categoria, a Tabela 6.12 indica que o SEI-Geo aproxima-se bastante do melhor sistema nas medidas de precisão e medida F na tarefa de classificação semântica. No que diz respeito à identificação, o SEI-Geo é o sistema mais preciso entre os concorrentes e alcançou um valor próximo ao melhor sistema na medida F.

A Tabela 6.13 apresenta os resultados do SEI-Geo no HAREM Clássico distribuídos por subtipos da categoria Local. O SEI-Geo apresenta os melhores resultados para os subtipos Pais, Aguacurso e Aguamassa. Embora carecendo de informação sobre a geografia física na geo-ontologia, o SEI-Geo ainda alcança resultados competitivos por meio do uso de padrões para os subtipos físicos.

[1]Cenário de avaliação que não considera anotação com etiquetas alternativas (ALT) para uma EM.



(a) Resultado da **classificação** ordenado pela medida F.



(b) Resultado da **identificação** ordenado pela medida F

Figura 6.1: Resultados da participação dos sistemas no Cenário Seletivo 5 do Segundo HAREM.

Tabela 6.12: Resultados da categoria Local considerando a classificação com avaliação relaxada de ALT.

	Classificação Semântica Combinada			Identificação		
Corrida	P	A	F	P	A	F
2	**0,6830**	0,5029	0,5793	**0,7121**	0,5175	0,5994
3	0,6810	0,5215	**0,5906**	0,7087	0,5375	0,6113
4	0,6736	**0,5252**	0,5902	0,7020	**0,5416**	**0,6115**
Melhor sistema	0,6928	0,7015	0,6078	**0,7121**	0,7982	0,6376

Tabela 6.13: Avaliação dos subtipos da categoria Local.

	Precisão	Abrangência	Medida F
País	0,8488	0,6518	**0,7503**
Divisão	0,6384	0,3818	0,5101
Região	1,0000	0,0448	0,5224
Construção	0,3636	0,0220	0,1928
Rua	0,4615	0,1818	0,3216
Outro	0,0408	0,0625	0,0517
Aguacurso	0,7143	0,6250	**0,6697**
Aguamassa	0,8889	0,4444	**0,6666**
Relevo	0,5714	0,4000	0,4857
Planeta	0,3333	0,3333	0,3333
Ilha	0,3333	0,1111	0,2222

Os resultados da classificação semântica, nas Tabelas 6.11 e 6.13, evidenciam que o SEI-Geo é um sistema que pode alcançar os melhores sistemas no reconhecimento de locais para o português.

A Tabela 6.14 apresenta os resultados das principais corridas do SEI-Geo na tarefa de ReRelEM. Apesar do sistema identificar corretamente os relacionamentos que se propõe a identificar, sua abrangência ainda é muito baixa, comparada ao melhor sistema nessa medida. De acordo com a Tabela 6.14, o SEI-Geo foi o melhor sistema no reconhecimento de relacionamentos de inclusão

Tabela 6.14: Resultado da participação do SEI-Geo na tarefa de ReRelEM do Segundo HAREM - Avaliação de Relacionamentos - Cenário Total - Inclusão.

Corrida	P	A	F	Espúrios	Falta	Tot. CD	Tot. id.	Tot. correc. id.
3	1,0	0,0769	0,1428	0	72	78	6	6
2	**0,9166**	**0,2973**	**0,4490**	2	52	74	24	22
4	0,9166	0,2820	0,4314	2	56	78	24	22
Melhor sist.	1,0	0,4231	**0,4490**	0	52	74	24	22

para a categoria Local. O SEI-Geo não reconheceu relacionamentos de identidade (p. ex. USA=EUA) nos textos. O reconhecimento desse tipo de relacionamento foi deixado para trabalho futuro.

Os resultados da Tabela 6.14 são provenientes de uma coleção de 12 documentos com 579 EM e 603 relacionamentos, que após expansão totalizam 5.716. Um exemplo de expansão é: o relacionamento inclui(Lisboa, Odivelas) implica o seu inverso incluido_em(Odivelas, Lisboa). Freitas et al. (2008) descrevem os detalhes sobre a expansão de relacionamentos na avaliação da pista ReRelEM.

6.6.3.3 Considerações sobre a Participação do SEI-Geo no HAREM Clássico

Após a análise dos resultados da participação do SEI-Geo no Segundo HAREM é possível concluir que a combinação das geo-ontologias WGO e Geo-Net-PT produziu os melhores resultados. A contribuição da Geo-Net-PT ainda é mínima, mas o suficiente para ser um diferencial quando os resultados são comparados com os outros sistemas participantes.

É importante notar que a Geo-Net-PT foi mutilada no nível de localidade. Os nomes de localidade inserem muitos falsos positivos no processo de reconhecimento de EM. O uso de nomes de localidade da Geo-Net-PT (p. ex. 'Caracol', 'Namorados' e 'Nabo') implica numa sobre-geração de EM reconhecidas, além de gerar mais EG ambíguas.

Outra constatação a mencionar é o fato do tamanho da geo-ontologia utilizada. O SEI-Geo utilizou menos de 18 mil EG obtendo o segundo melhor resultado no cenário seletivo 5 com medida F 0,6006, enquanto o sistema sistema CaGE (Martins, 2008a,b) utilizou mais de 15 milhões de EG e obteve uma medida F de 0,5615 na sua melhor corrida. Embora a abrangência do CaGE (0,5879) seja melhor que a do SEI-Geo (0,5413), a medida de precisão é bastante inferior (0,5374 contra 0,6801 do SEI-Geo). Tal fato é um indício de que o número de EG utilizadas para reconhecer locais em textos em português deve ser limitado à ordem de grandeza dos milhares e não dos milhões.

A principal limitação do SEI-Geo está na medida de abrangência. Tal fato pode ser justificado pela simplicidade do sistema, uma vez que não há análise sintática do texto, o conjunto de padrões é limitado e as geo-ontologias são desprovidas de locais físicos (apenas a WGO contém locais físicos customizados

para as participações do sistema da Universidade de Lisboa nas quatro edições do GeoCLEF, de 2005 a 2008). Para aumentar a abrangência do SEI-Geo pode-se inserir no domínio administrativo da geo-ontologia WGO cidades com menos de 100 mil habitantes e no domínio físico mais conceitos e ocorrências.

Por outro lado, o SEI-Geo apresentou resultados satisfatórios para a medida de precisão nas corridas 3 e 4, obtendo o melhor resultado no cenário seletivo total na tarefa de identificação de locais.

Na avaliação do cenário seletivo da categoria Local, para a classificação semântica, o SEI-Geo atingiu o segundo melhor resultado com medida F de 0,6006.

6.6.3.4 Considerações sobre a Participação do SEI-Geo na Tarefa de ReRelEM

Além da identificação e reconhecimento de locais, o SEI-Geo também foi avaliado na tarefa de reconhecimento e relacionamentos semânticos entre locais. Os relacionamentos identificados são dependentes das geo-ontologias utilizadas e dos resultados do SEI-Geo no HAREM Clássico. Relacionamentos fora das geo-ontologias não são identificados e os locais não reconhecidos no HAREM Clássico nunca fazem parte de um relacionamento.

A tarefa de reconhecimento de locais, conforme proposta inicialmente, propõe a identificação de relacionamentos dos seguintes tipos:

- `identidade` (sem TIPOREL ou TIPOREL = 'ident');
- `inclusão` (TIPOREL = 'inclui' ou TIPOREL = 'incluido');
- `localização` (TIPOREL = 'ocorre_em' ou TIPOREL = 'sede_de');
- `outra` (TIPOREL = 'outra').

O SEI-Geo reconheceu apenas o relacionamento de inclusão na sua participação nessa tarefa. Os resultados indicam que a abordagem e as geo-ontologias utilizadas auxiliam bastante, mas não são suficientes para reconhecer relacionamentos entre locais em textos, apesar de o SEI-Geo ter sido o melhor sistema no reconhecimento de relacionamentos de inclusão.

Tabela 6.15: Tempos de processamento das corridas submetidas ao Segundo HAREM.

Corrida	1	2	3	4
Minutos	27	76	30	101

Finalmente, uma nota sobre o custo computacional do SEI-Geo. A Tabela 6.15 apresenta os tempos de processamento das corridas submetidas ao Segundo HAREM. Esses tempos foram obtidos em um servidor com sistema operacional Linux, processador Intel(R) Xeon(TM) CPU 3.20GHz e 8GB de memória. Mais detalhes sobre a participação do SEI-Geo no Segundo HAREM estão em Chaves (2008).

A comparação entre os resultados obtidos pelo SEI-Geo nas duas edições do HAREM pode ser tendenciosa uma vez que as medidas de avaliação foram alteradas do Primeiro para o Segundo HAREM. Além disso os subtipos dos tipos Humano e Físico do Segundo HAREM não têm correspondência no Primeiro HAREM, pois os tipos Administrativo e Geográfico não continham sub-tipos naquele evento.

6.7 Considerações sobre Sistemas do Estado da Arte

Uma comparação direta com os sistemas que representam o estado da arte (os sistemas mais relevantes foram descritos no Capítulo 2) em extração e integração de informação é uma tarefa bastante complexa devido ao grande número de variáveis envolvidas. Essas variáveis incluem as diferenças na língua, o propósito dos sistemas e as medidas de avaliação, entre outras.

O sistema Snowball foi concebido para extrair relacionamentos entre organizações e locais, logo uma comparação com o SEI-Geo, que extrai relacionamentos somente entre locais, não é aplicável.

Os sistemas KnowItAll/KnowItNow e OntoSyphon, por exemplo, são sistemas de extração de informação em grande escala que não fazem integração da informação extraída em estruturas de representação de conhecimento. A melhor aproximação seria implementar esses sistemas adaptando-os para processar textos

em português e assim fazer a comparação com o SEI-Geo somente no que diz respeito à parte de extração. Contudo, esperamos que venha a ser possível, talvez no âmbito de uma tese de mestrado, realizar esta comparação.

O sistema OntoLearn faz extração e integração de informação, mas utiliza a WordNet em inglês para a realização dos testes e da avaliação. Como ainda não há uma WordNet em português, a comparação entre o OntoLearn e o SEI-Geo torna-se bastante difícil.

Todos os sistemas descritos até aqui foram avaliados com textos em inglês e, portanto, teriam que ser adaptados para processar textos em português.

O sistema OnLocus identifica e reconhece endereços completos e, embora aplicado a textos em português, há pouca sobreposição com o SEI-Geo, que não reconhece endereços completos, somente reconhece os locais presentes no endereço separadamente. Por exemplo, no endereço 'Travessa de Santa Quitéria, 36 3E 1250-212 LISBOA', o SEI-Geo reconhece 'Travessa de Santa Quitéria' e 'Lisboa' separadamente como locais distintos. A parte de reconhecimento de endereços completos é uma extensão que pode ser implementada no SEI-Geo como trabalho futuro.

Finalmente, cabe destacar que o módulo de extração e anotação de informação e a parte de reconhecimento de relacionamentos do SEI-Geo foram avaliados e comparados com sistemas do estado da arte que processam textos em português no Segundo HAREM.

6.8 Conclusões

Esse capítulo apresentou uma caracterização da informação geográfica em textos e a avaliação dos métodos implementados no SEI-Geo.

A avaliação do módulo de integração de conhecimento do SEI-Geo evidenciou a dificuldade de se expandir conhecimento com informação textual. A avaliação do SEI-Geo no HAREM, tanto na tarefa do HAREM Clássico quanto na tarefa e reconhecimento de relacionamentos entre locais, foi bastante positiva alcançando resultados próximos aos melhores sistemas, o que evidencia sua qualidade. Os resultados dessa avaliação confirmam que a metodologia com uso de padrões e geo-ontologias é útil para reconhecer locais, mas limitada para reconhecer relacionamentos entre locais.

6. AVALIAÇÃO DOS MÉTODOS PROPOSTOS

O próximo capítulo sintetiza a metodologia proposta nessa tese.

Capítulo 7

Metodologia para Construção de Geo-Ontologias

7.1 Introdução

Nos capítulos precedentes foram apresentados os algoritmos e sistemas usados nas várias etapas do processo de construção de geo-ontologias, bem como a descrição das geo-ontologias criadas e a sua utilidade para diversas aplicações.

O desenvolvimento de geo-ontologias é um processo lento, longo, suscetível a erros e que exige conhecimento especializado de todos os intervenientes. Os processos envolvidos durante esse desenvolvimento carecem de uma definição detalhada das atividades a serem executadas. Esse capítulo descreve uma metodologia para construção de geo-ontologias de modo a colmatar essa carência na literatura.

Inicialmente, a criação de geo-ontologias deve ser motivada pela necessidade de alguma aplicação. Quando uma ou mais aplicações apresentam requisitos de uso de informação geográfica, procede-se ao início da construção das geo-ontologias.

Contudo, antes de se construir uma geo-ontologia é necessário verificar se há informação geográfica suficiente em formato digital. Preferencialmente, deve-se começar a pesquisa por fontes semi-estruturadas ao invés de textos em linguagem natural, pois é mais fácil encontrar informação completa (dados relacionados em todos os níveis de informação, desde código-postal até o nome da cidade, país ou continente, por exemplo) em bases de dados semi-estruturadas do que em

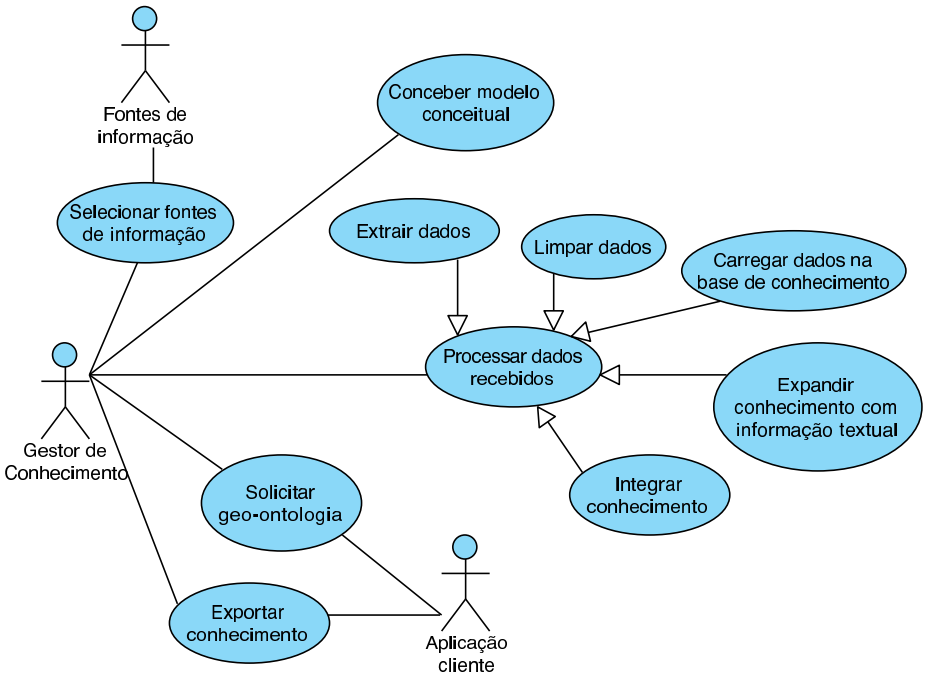


Figura 7.1: Casos de uso para a construção de geo-ontologias.

textos. Uma vez confirmada a existência de fontes de informação suficientes para a construção de uma geo-ontologia, procede-se a criação de um sistema de gestão de conhecimento geográfico (p. ex. GKB). Esse sistema deve ter um conjunto de funcionalidades que incluem a capacidade de limpeza de dados e integração e exportação de conhecimento, entre outras. No momento que um sistema de gestão de conhecimento geográfico é dotado de tais funcionalidades, pode-se iniciar o processo de geração de geo-ontologias.

É importante enfatizar que a metodologia proposta nessa tese pressupõe a existência de um sistema de gestão de informação para suportar a implementação dos processos de extração, limpeza e carregamento de dados, bem como integração e exportação de conhecimento. A metodologia descrita nessa tese é ilustrada através de diagramas UML. O diagrama de casos de uso apresentado na Figura 7.1 descreve a interação dos atores externos com o sistema de gestão de conhecimento geográfico. Esse diagrama apresenta os processos dessa metodologia.

O Gestor de Conhecimento (daqui em diante, Gestor) seleciona as fontes de informação e concebe o modelo conceitual no qual são definidas as classes

e seus relacionamentos. Esse modelo será utilizado para desenvolver a base de conhecimento. Em seguida, o Gestor inicia o processamento dos dados recebidos, o qual envolve tarefas de extração, limpeza e carregamento de dados. Outra tarefa do Gestor é a integração de conhecimento quando fontes de informação são fornecidas. Com os dados integrados na base de conhecimento, o Gestor controla a fase de exportação do conhecimento no formato de geo-ontologias, as quais são solicitadas por aplicações cliente. Finalmente, o Gestor expande o conteúdo presente na base de conhecimento com informação geográfica proveniente de textos.

O Gestor é o responsável pela construção da geo-ontologia, que é um trabalho de edição, tal como é o de produção de um jornal ou de qualquer outra coletânea de textos. As decisões tomadas pelo Gestor baseiam-se no estabelecimento de critérios bem fundamentados de inclusão e exclusão de elementos informativos. Esses critérios devem ser estabelecidos no princípio do desenvolvimento da geo-ontologia e nortearem as decisões futuras do Gestor.

Uma metodologia define uma sequência de etapas para ser mais facilmente compreendida e implementada. Após a descrição da metodologia, a seguir são listadas as fases que a compõem:

1. Determinar o âmbito da geo-ontologia, definindo os sub-domínios que serão criados. Nessa fase são definidos os conceitos mais importantes que formarão as geo-ontologias.

2. Reutilizar conhecimento: verificar as fontes de informação existentes e reutilizar aquelas com maior grau de autoridade.

3. Definir um modelo conceitual para construir a base de conhecimento.

4. Extrair, limpar e carregar as fontes de informação.

5. Integrar conhecimento.

6. Exportar conhecimento.

7. Reusar ou desenvolver um sistema de extração e integração de conhecimento geográfico proveniente de textos.

8. Voltar a etapa 6.

7.2 Concepção de um Modelo Conceitual

Após identificar a necessidade de uso de uma geo-ontologia e confirmar a existência de fontes de informação, é necessário estabelecer um modelo conceitual a partir do qual uma base de conhecimento geográfica possa ser construída. Esse modelo conceitual deve ser suficientemente abrangente para ser capaz de incorporar novos conceitos e sub-domínios dentro do domínio geográfico. O modelo conceitual apresentado e testado nessa tese é apenas um exemplo de outras variações que o mesmo pode ter conforme o contexto no qual as geo-ontologias serão geradas.

De qualquer forma, devem-se considerar os seguintes aspectos na concepção de um modelo conceitual para o domínio geográfico:

- **Proveniência da informação:** para qualquer dado existente no modelo, o mesmo deve permitir que uma aplicação seja capaz de rastreá-lo até sua fonte de informação. Ou seja, o modelo deve permitir que se responda a pergunta: de onde tal dado é proveniente? Pode-se referir nesse ponto que o Gestor deve adotar um critério de *rastreabilidade* dos dados. Na GKB, pode-se observar na Figura 3.5, que todas as classes estão associadas à classe *Info-Source*, a qual armazena todas as fontes de informação do modelo.

- **Definição de sub-domínios de conhecimento dentro do domínio geográfico:** no caso das geo-ontologias que foram construídas com a GKB, os sub-domínios foram separados em administrativo e físico (mais domínio de rede na Geo-Net-PT), mas o modelo proposto suporta múltiplos sub-domínios. No caso de uma geo-ontologia para um continente, por exemplo, pode-se particionar a informação em outros sub-domínios, como aquático e vegetação conforme a quantidade e relevância dos dados no contexto no qual as geo-ontologias estão sendo criadas. Os sub-domínios também podem ser definidos como resultado dos requisitos das aplicações. Por exemplo, caso haja um conjunto de aplicações cujo interesse prioritário seja informação hídrica, faz sentido criar um sub-domínio 'hidro' e adicionar os conceitos pertinentes a esse domínio. O mesmo raciocínio vale para outros sub-domínios. O Gestor deve tentar prever quais aplicações usarão as geo-ontologias e estar ciente de quão extensível é o modelo conceitual proposto, tentando obedecer ao critério de *extensibilidade*.

- **Definição de relacionamentos intra-domínio:** os relacionamentos entre conceitos dentro de um mesmo domínio devem ficar explicitamente definidos no modelo conceitual de modo a facilitar tarefas de raciocínio, por exemplo. Na GKB, as classes *Type-Relationship* e *Feature-Relationship* suportam a definição dos relacionamentos entre conceitos e entre ocorrências, respectivamente.

- **Definição de relacionamentos inter-domínio:** além dos relacionamentos entre entidades geográficas (EG) pertencentes ao mesmo sub-domínio, é necessário que o modelo preveja como serão tratados os relacionamentos inter-domínio. No caso da GKB, foram criadas relações específicas para armazenar esses tipos de relacionamentos. Por exemplo, para armazenar relacionamentos entre os domínios administrativo e físico utilizou-se a relação *ID-Feature-Relationship-Adm-Phy* e entre os domínios de rede e administrativo a relação *ID-Feature-Relationship-Net-Adm* (ver Figura 3.4).

- **Atributos das EG:** o modelo deve manter os atributos relacionados aos conceitos ou tipos geográficos e as suas ocorrências. Por exemplo, um país, uma cidade ou uma freguesia possuem o atributo população, um rio possui o atributo comprimento, uma montanha tem uma altitude, entre outros atributos. Todas as ocorrências desses conceitos podem conter valores para esses atributos quando as fontes de informação fornecem os mesmos.

- **Tratamento de variantes:** alguns nomes geográficos possuem variantes de um país ou região para outro (p. ex. 'Irã' no Brasil e 'Irão' em Portugal) ou nomes históricos (p. ex. 'Praça do Rio de Janeiro', antigo nome da atual 'Praça do Príncipe Real' em Lisboa). O Gestor deve armazenar todas as variantes disponíveis e no momento de exportar o conhecimento da base de conhecimento, deixar indicado através de um atributo, por exemplo, que determinado nome é de uma variante específica.

- **Representação de diferenças de opinião ou conflitos:** conflitos podem existir provenientes das fontes de informação. Um exemplo de conflito é a informação sobre o comprimento de um rio. Se duas fontes definem comprimentos diferentes, armazena-se ambos e, no momento da exportação de conhecimento para a geo-ontologia, deixa-se indicado através de um

atributo (p. ex. que ambos os comprimentos são pertinentes e que cada um é definido na fonte de informação X e Y). Essa abordagem é seguida devido ao fato de o Gestor não ser autoridade para definir qual o comprimento correto de um rio.

- **Informação a incluir ou descartar:** nesse ponto há que se estabelecer critérios. No caso da GKB, todas as fontes de informação são provenientes de autoridades administrativas. Entretanto, uma base de dados de uma empresa privada, sítios da web e a própria Wikipedia também podem ser úteis. No caso de fontes como a Wikipedia, é importante que seja armazenada a data da recolha dos dados na base de conhecimento, uma vez que o conteúdo dessa enciclopédia é volátil. O Gestor pode estabelecer diversos critérios que incluem autoridade, atualidade e gratuidade bem como uma combinação desses.

- **Controle de versões:** o Gestor deve ter em mente as aplicações que usam as geo-ontologias e decidir qual abordagem de controle de versões adotar. Pode ser preferível ter novas versões somente em casos de mudanças significativas, pois se o Gestor optar por lançar uma versão a cada inserção de novas ocorrências na base de conhecimento, arrisca-se a ter muitas versões num curto período de tempo, o que pode deixar insatisfeitos os utilizadores, pois serão obrigados a realizar muitas atualizações para ter a versão mais completa. Na GKB, uma nova versão só foi concebida após mudanças significativas (p. ex. a inserção de um novo domínio de conhecimento).

A definição do modelo conceitual é uma das fases mais relevantes da metodologia, pois todas as fases subsequentes estão apoiadas nas decisões tomadas nessa fase. O Gestor não deve ser visto como o "dono da verdade" e deve estar consciente de que sua autoridade no domínio geográfico pode frequentemente entrar em conflito com outros especialistas do mesmo domínio. Por exemplo, conceitos relevantes, para inclusão em uma geo-ontologia, para um geógrafo, podem não ser (e frequentemente não são) os mesmos que para um engenheiro informático.

7.3 Seleção e Limpeza de Fontes de Informação

Uma das primeiras etapas na construção de geo-ontologias é a seleção das fontes de informação geográfica com a informação pretendida pelo Gestor. Essa seleção deve considerar pelo menos os seguintes critérios:

- **Autoridade:** autoridade ou credibilidade da fonte selecionada no contexto (p. ex. 'cidade', 'país' e 'continente') ao qual está sendo criada a geo-ontologia. Esse critério pode variar entre autoridades científicas (p. ex. universidades ou fundações científicas), sociais (p. ex. Wikipedia), políticas (p. ex. institutos de estatística (*UNESCO Institute for Statistics*, INE em Portugal e IBGE no Brasil) e forças armadas (p. ex. exército, marinha e aeronáutica), entre outras. No caso da Geo-Net-PT, as fontes de informação predominantes foram provenientes de autoridades políticas e científicas, enquanto a WGO contém informação de autoridades sociais além dessas.

- **Custo de aquisição:** o valor a ser investido numa fonte de informação pode ser determinante na decisão de incluí-la numa base de conhecimento. No caso específico da construção da GKB, para todas as geo-ontologias, a maior parte das fontes de informação provém de instituições públicas e todas são fornecidas gratuitamente.

- **Tipo de licenciamento:** o tipo de licenciamento do recurso construído pode obedecer aos critérios da organização *Creative Commons* `http://creativecommons.org`. Esses critérios incluem o licenciamento que permite alteração do trabalho original, uso por empresas privadas ou permissão de uso somente sem fins lucrativos. No caso da Geo-Net-PT, o licenciamento é gratuito e os usuários podem estender o conteúdo da mesma.

- **Formato:** o formato que os dados são fornecidos pelas fontes de informação, muitas vezes, determina sua inclusão numa base de conhecimento. Os formatos podem ser tão simples quanto o *Comma Separate Value* (CSV) ou complexos no formato de uma base de dados particular ou alguma codificação especial de uma empresa privada. No caso da GKB, os dados recebidos estavam no formato semi-estruturado, em ficheiros Excel ou no formato CSV.

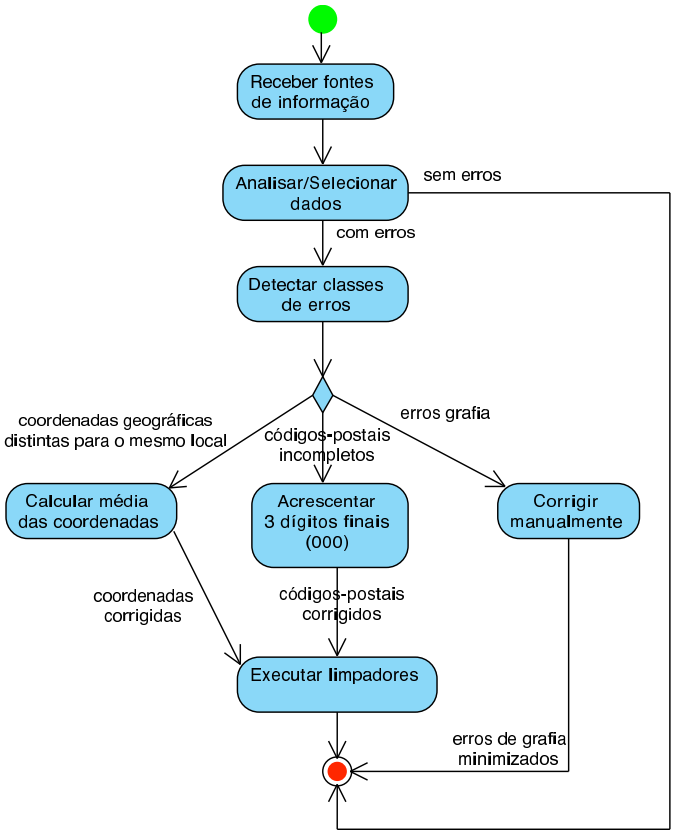


Figura 7.2: Diagrama de atividades: Limpar dados.

De posse dos dados, é necessário proceder à sua limpeza. A limpeza deve minimizar e, até mesmo, eliminar as inconsistências entre os dados das fontes de informação, além de corrigir erros de grafia, ajustes de coordenadas e códigos postais, por exemplo. Essa é uma fase importante, na qual deve-se prever um tempo considerável para sua execução, uma vez que os dados resultantes dessa limpeza serão a entrada dos processos de integração e exportação de conhecimento.

A Figura 7.2 apresenta o diagrama do processo Limpar dados. Limpar dados requer que o Gestor receba as fontes de informação, analise os dados e detecte as classes de erros existentes. Para os casos em que existe mais de uma coordenada associada a um local, é calculada a média das coordenadas. Note-se que esse processo também é uma estratégia de integração de dados. Se erros são encontrados em códigos-postais (p. ex. sem os últimos três dígitos), o algoritmo

deve acrescentar três zeros ao final dos códigos incompletos. Se há erros de ortografia, esses são normalmente corrigidos de forma manual, caso a caso. A detecção de tais erros muitas vezes depende do conhecimento do domínio por parte do Gestor. Finalmente, se a fonte de informação não contém erros, o processo termina diretamente.

No final desse processo deve haver uma verificação das convenções usadas na codificação dos dados disponíveis para o carregamento na base de conhecimento. Essa verificação deve atender aos seguintes casos:

- **Acentuação:** é comum haver acento grave no lugar de acento agudo, nomes sem tilde e sem cedilha, por exemplo.
- **Hifenização:** nesse caso recomenda-se incluir os nomes nas duas formas, de modo que as geo-ontologias possam facilitar e diminuir o trabalho das aplicações-cliente. De notar que os acordos ortográficos vão alterando as regras ao longo do tempo e a essa metodologia sugere que sejam incorporadas todas as grafias de escrita.
- **EG com nomes iguais:** verificar se o identificador de um nome está associado a identificadores diferentes para todas as suas EG (p. ex. o nome 'Lisboa' está associado a um distrito, um concelho e diversos arruamentos, entre outros tipos de EG).
- **Codificação dos caracteres:** verificar se a codificação dos dados fornecida pela fonte de informação é a mesma utilizada na base de conhecimento e na geração das geo-ontologias.

7.4 Integração de Conhecimento

A integração de conhecimento é realizada após a identificação de relacionamentos entre conceitos e suas ocorrências provenientes das fontes de informação. É muito difícil, antecipadamente, prever todos os conceitos que estarão presentes numa geo-ontologia durante sua existência. O procedimento adotado nessa metodologia é uma abordagem baseada em relevância, na qual primeiro são definidos os conceitos mais relevantes e depois os demais conceitos que vão sendo fornecidos pelas fontes de informação. Essa abordagem é diferente das clássicas *top-down*

e *bottom-up* e tem sido referenciada na literatura como *middle-out* (Fernández-López e Gómez-Pérez, 2002). Assim, após definidos os principais conceitos, a integração de conhecimento é realizada à medida que as fontes de informação vão sendo entregues ao Gestor.

Essa integração geralmente é feita através de regras que estabelecem a posição dos conceitos na hierarquia da geo-ontologia. Os conceitos mais atuais tendem a ser de mais fácil integração, enquanto a integração de conceitos históricos nos atuais, existentes à partida nas geo-ontologias, é uma tarefa mais complexa. No caso específico da Geo-Net-PT, houve a necessidade de integrar o conceito histórico 'província' com os conceitos mais atuais como 'distrito' e 'concelho', por exemplo. As regras para esse tipo de integração de conhecimento nem sempre são explícitas ou fornecidas pelas autoridades de um determinado país, fazendo com que o Gestor tenha que consultar outras fontes, como enciclopédias, por exemplo, para obter a informação necessária sobre como integrar conceitos históricos e atuais.

Após a limpeza dos dados, esses estão prontos para serem carregados na base de conhecimento. Nesse momento, começa a fase de integração de conhecimento, conforme ilustrada na Figura 7.3.

A integração de conhecimento na GKB ocorre a partir da identificação de conceitos ou ocorrências comuns entre geo-ontologias (geralmente representadas como hierarquias). Caso não haja ocorrências comuns, a geo-ontologia resultante é simplesmente a união das duas geo-ontologias. Se ocorrências comuns são detectadas, é necessário identificar o antecessor comum dessas ocorrências em nível mais específico na hierarquia. Caso não haja antecessor comum, a nova ocorrência também é integrada diretamente na raiz da geo-ontologia existente. Se há um antecessor comum, é necessário verificar a distância (em número de relacionamentos `parte-de`) entre as ocorrências comuns dos conceitos e seus antecessores. O antecessor que possuir a menor distância até as ocorrências comuns, é integrado através do relacionamento `parte-de` com o antecessor na outra hierarquia. Os relacionamentos existentes em ambas as hierarquias são mantidos.

Outra situação na fase de integração de conhecimento surge quando o conteúdo geográfico de textos em linguagem natural está disponível para ser integrado com o conteúdo já presente num sistema de gestão de conhecimento geográfico. Os

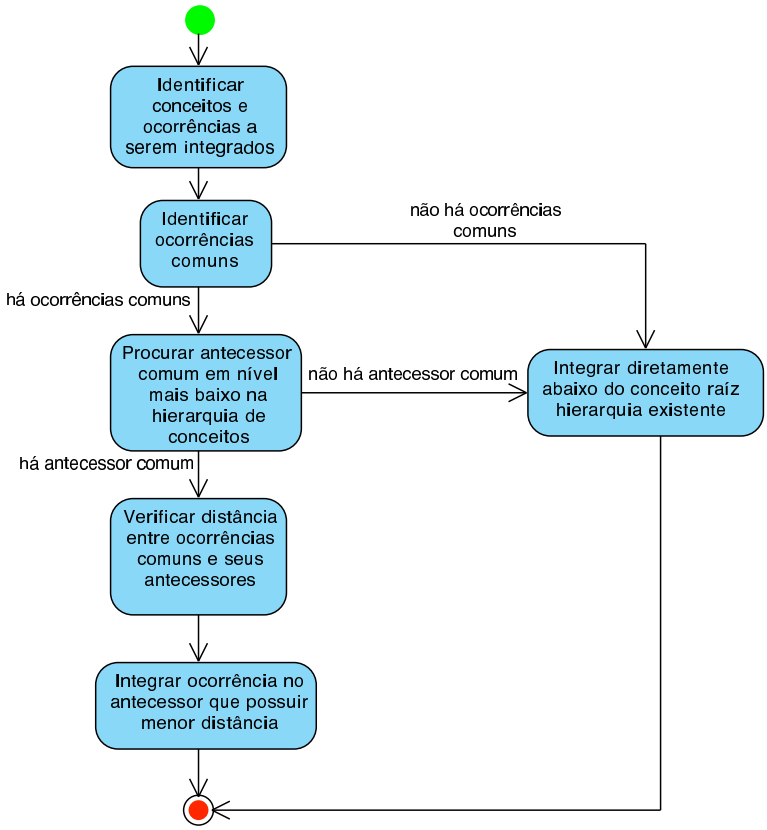


Figura 7.3: Diagrama de atividades: Exemplo de integração de conhecimento.

casos mais frequentes encontrados foram a integração de conhecimento indireto (conceitos que não estão imediatamente relacionados) e a integração inter-domínios. Na primeira situação recomenda-se que a integração aconteça no nível mais específico da hierarquia; contudo, como isso nem sempre é possível, o conhecimento incompleto adquirido no texto não será desperdiçado, mas sim integrado conforme for encontrado no texto. No segundo caso, os relacionamentos encontrados em texto entre EG dos domínios administrativo e físico são integrados diretamente na geo-ontologia. A Seção 5.3 apresenta exemplos desses casos.

A Figura 7.4 descreve o processo da integração de conhecimento proveniente de textos. Após selecionar um corpus com informação geográfica, executa-se um sistema de reconhecimento e integração de conhecimento (no caso da GKB, o SEI-Geo) e processa-se o resultado desse sistema. No diagrama, esse resultado é um conjunto de arbustos que são formados por triplas. Cada tripla é processada

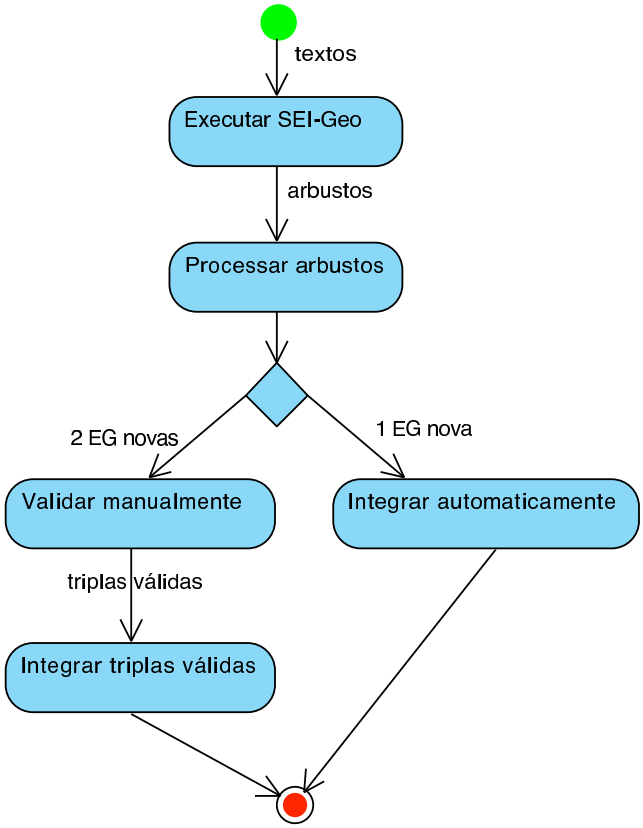


Figura 7.4: Diagrama de atividades: Exemplo de integração de conhecimento textual com o SEI-Geo e a GKB.

e pode originar dois casos: (1) possuir duas EG novas, e nesse caso a tripla é enviada para uma pessoa fazer a validação manual ou (2) possuir somente uma EG nova e nesse caso a integração pode ser feita automaticamente. Nesse último caso, supõe-se que a EG nova é uma EG reconhecida correctamente pelo SEI-Geo. Por outro lado, após a validação manual, as triplas consideradas válidas são integradas nas respectivas tabelas na base de conhecimento.

7.5 Exportação de Conhecimento e as Aplicações

Após a integração de conhecimento é necessário exportar a informação para as aplicações. O grau de formalidade no qual a geo-ontologia é expressa pode variar (p. ex. RDF, OWL-Lite e OWL-Full). Nesse ponto é importante que os progra-

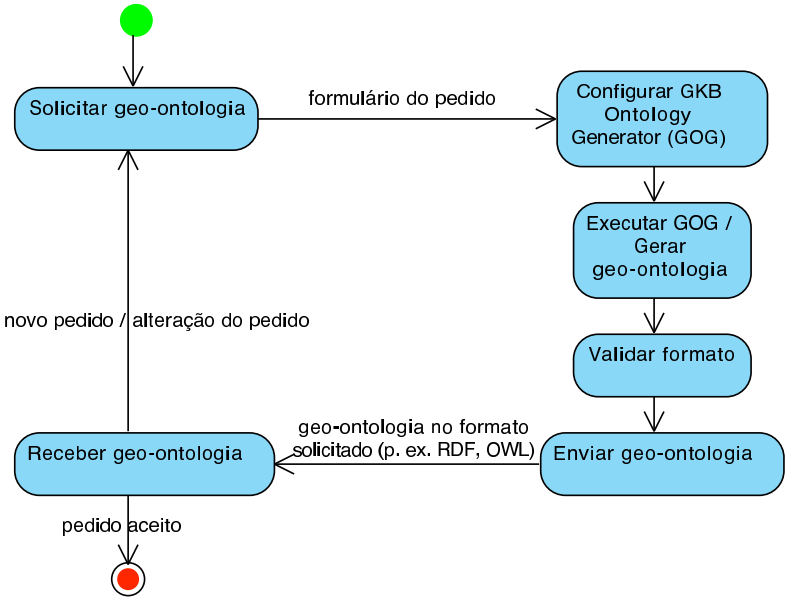


Figura 7.5: Diagrama de atividades: Exportar conhecimento.

mas geradores de geo-ontologias (na GKB, o *GKB Ontology Generator*) sejam flexíveis para entregar as geo-ontologias no formato solicitado pelas aplicações-cliente. Na GKB qualquer aplicação-cliente que processe dados no formato XML pode utilizar o contéudo geográfico armazenado na base de conhecimento.

Além disso, na fase de exportação de conhecimento, deve-se validar o formato no qual as geo-ontologias estão sendo formalizadas antes de entregar à aplicação-cliente. Na GKB, os validadores utilizados incluem um utilitário para Linux, o XMLValid, o RDF Validator (`http://w3.org/RDF/Validator`) e o OWL Validator (`http://mygrid.org.uk/OWL/Validator`).

Nessa etapa da metodologia, é necessário ter em mente que a diversidade de aplicações e suas necessidades variam muito. As geo-ontologias geradas pela GKB foram e são utilizadas por aplicações que vão desde uma interface web que lê os dados das geo-ontologias e os projeta na interface até um sistema de RIG que faz raciocínio automático sobre o conteúdo das geo-ontologias.

A última fase dos processos da GKB é a exportação do conhecimento integrado para o formato de geo-ontologias. A Figura 7.5 apresenta o diagrama de atividades.

O Gestor recebe um pedido através de um formulário. De acordo com o

conteúdo do formulário, o Gestor configura o GOG para gerar a geo-ontologia no formato solicitado. Após a geração da geo-ontologia, o Gestor valida o formato com um validador XML/RDF/OWL, por exemplo. Uma vez validado o ficheiro da geo-ontologia, esse é enviado à aplicação que o solicitou. Se o pedido é aceito o processo chega ao fim, caso contrário é feito um novo pedido através de outro formulário e o processo recomeça.

7.6 Avaliação de Metodologias para Construção de Ontologias

Após descrever detalhadamente a metodologia proposta nessa tese é fundamental apresentar critérios para avaliar uma metodologia para construção de ontologias. Tentando responder aos principais critérios para analisar metodologias propostos por Fernández-López e Gómez-Pérez (2002), podem-se fazer as seguintes considerações:

- **Herança da Engenharia de Conhecimento:** esse critério considera a influência da Engenharia de Conhecimento tradicional na metodologia proposta nessa tese. Pode-se considerar que essa herança é alta, uma vez que os processos de aquisição de conhecimento (obtenção de fontes de informação) e modelagem e codificação da base de conhecimento são claramente identificados. Contudo, a fase de avaliação da metodologia ainda precisa ser mais detalhada.

- **Detalhe da metodologia:** as fases e métodos para solucionar os problemas na construção de geo-ontologias são detalhadas através de algoritmos, diagramas UML, artigos e relatórios técnicos. Essa documentação provê uma explicação minuciosa dos problemas encontrados ao longo do processo bem como as estratégias utilizadas para solucioná-los.

- **Recomendação para formalização do conhecimento:** As geo-ontologias geradas na fase de exportação do conhecimento podem ter diversos formatos, variando o seu grau de formalidade. Recomenda-se o desenvolvimento de uma aplicação flexível que seja capaz de exportar conhecimento em diversos graus de formalidade. No caso da GKB, o

GKB Ontology Generator (GOG) foi desenvolvido para ser suficientemente flexível para permitir que o conhecimento armazenado no repositório seja representado numa ampla variedade de formatos, desde XML até OWL-DL.

- **Estratégia para construir ontologias:** esse critério é relativo a dependência da metodologia à aplicações. Uma metodologia dependente de aplicação é construída baseada nos requisitos da aplicação que fará uso da mesma, enquanto uma metodologia independente de aplicação é totalmente implementada sem considerar as aplicações que a utilizarão. A metodologia proposta nessa tese é semi-dependente de aplicação, ou seja, as geo-ontologias desenvolvidas com a metodologia proposta nessa tese consideram cenários de uso de aplicações na fase de especificação de requisitos. De qualquer forma, as geo-ontologias construídas podem ser utilizadas por qualquer aplicação que interprete os formatos XML, RDF ou OWL. Algumas dessas aplicações foram descritas nos Capítulos 3 e 5 .

- **Estratégia para identificar conceitos:** segue uma abordagem *middle-out*, isto é, primeiro são definidos os conceitos mais importantes e depois os mais abstratos e mais concretos. O critério de importância pode ser atribuído ao interesse das aplicações. No exemplo do sub-domínio 'hidro', referido acima, os conceitos mais importantes seriam aqueles mais relacionados ao sub-domínio (p. ex. 'rio', 'lago' e 'canal'), enquanto conceitos mais abstratos, como país e continente, seriam incluídos num segundo momento, se necessário.

- **Ciclo de vida recomendado:** é principalmente dependente dos requisitos das aplicações e do grau de maturidade da equipe de desenvolvimento. Com uma equipe de desenvolvimento experiente e madura, recomenda-se um ciclo de desenvolvimento em espiral, no qual tem-se um melhor controle dos riscos do projeto bem como uma flexibilidade maior para aceitar mudanças de requisitos, o que é bastante frequente atualmente. Por outro lado, se os requisitos de desenvolvimento estão bem explícitos e completos e a equipe de desenvolvimento é imatura tecnicamente, recomenda-se o ciclo de vida em cascata. Esse ciclo é linear, sendo que cada passo deve ser completado antes que o próximo passo possa ser iniciado.

- **Que ontologias têm sido desenvolvidas usando a metodologia e que sistemas têm sido construídos usando essas ontologias:** as geo-ontologias desenvolvidas com a metodologia proposta nessa tese são a Geo-Net-PT e a WGO, ambas utilizadas em diversas aplicações. O desenvolvimento de duas geo-ontologias com âmbitos diferentes e o uso de várias aplicações com objetivos distintos podem atestar um grau mínimo de maturidade para a metodologia proposta. Além disso, essas aplicações evidenciam a utilidade e relevância da metodologia para a prática.

Cabe ressaltar que a aplicação da metodologia na construção de mais de uma geo-ontologia e na extensão da Geo-Net-PT para o domínio físico permitiu aperfeiçoar os processos envolvidos. Em todas as fases da metodologia, o modelo conceitual e os algoritmos foram reutilizados e estendidos e nada foi criado a partir do zero. Tal evidencia que o reuso do conhecimento descrito na metodologia foi aplicado no desenvolvimento de geo-ontologias com âmbitos e domínios que estendem a Geo-Net-PT, a primeira geo-ontologia gerada com a metodologia proposta nessa tese. Note-se que nos últimos anos a cada edição do GeoCLEF a WGO era 'refrescada' num curto espaço de tempo (cerca de dois meses) e tornada disponível para o uso das diversas aplicações que formavam o sistema de recuperação de informação geográfica da Universidade de Lisboa.

7.7 Conclusões

A metodologia descrita nesse capítulo tem como objetivo sintetizar as principais fases do processo de construção de geo-ontologias bem como discutir e tentar responder questões recorrentes na área de representação de conhecimento. Essa metodologia ainda define um conjunto de marcos e entregas que devem ser verificados ao longo do desenvolvimento de geo-ontologias. Os principais são listados a seguir:

- Modelo conceitual definido.
- Base de conhecimento criada.
- Scripts de limpeza de dados desenvolvidos e testados.

- Fontes de informação limpas.
- Conhecimento integrado.
- Conversores desenvolvidos e testados.
- Geo-ontologia criada e validada.

O tempo necessário para ter cada um dos marcos atingidos depende do âmbito das geo-ontologias a serem criadas e da qualidade das fontes de informação. Esses marcos são suficientemente genéricos para serem utilizados em outros domínios de conhecimento. Durante todo processo recomenda-se a elaboração de documentação detalhada, que pode ser através de um relatório técnico.

Essa metodologia foi testada na construção de bases de conhecimento geográficas. As aplicações que utilizam e utilizaram as geo-ontologias produzidas foram descritas nos Capítulos 3 e 5. Um aspecto relevante da sua qualidade é a avaliação das geo-ontologias construídas com a metodologia proposta. A avaliação realizada foi através do seu uso pelas aplicações, cujo sucesso é condicionado pela qualidade das geo-ontologias.

O próximo capítulo apresenta as conclusões dessa tese, que incluem os principais resultados, as limitações, ideias para a continuidade do trabalho descrito e as reflexões finais.

Capítulo 8

Conclusões

8.1 Principais Resultados

O principal resultado desta tese é uma metodologia para construção de geo-ontologias a partir de dados de múltiplas fontes de conhecimento. Essa metodologia pode ser utilizada por qualquer empresa ou grupo de pesquisa que objetive construir um repositório de conhecimento geográfico a partir de fontes de informação complementares, incompletas e que contenham informação histórica e contemporânea, entre outras características. Um exemplo concreto é a necessidade de construir um geo-ontologia para um país (p. ex. Angola). A metodologia proposta nessa tese pode ser aplicada diretamente, desde a seleção de autoridades administrativas de Angola para fornecerem informação até a extração de conhecimento de textos que possuam informação geográfica sobre Angola.

A metodologia proposta nessa tese serviu para produzir geo-ontologias que foram úteis em diversas aplicações tais como sistemas de REM, módulos de um sistema de RIG, interface de motor de pesquisa geográfica e interface para consultas a almanaques geo-temporais. Além dessas aplicações, uma das geo-ontologias, a Geo-Net-PT, teve mais de 30 pedidos e também deve ter sido utilizada em sistemas que são desconhecidos por mim.

Esta tese reuniu numa base de conhecimento informação geográfica de Portugal que foi tornada pública e distribuída gratuitamente. A partir de um meta-modelo genérico foi possível armazenar informação geográfica proveniente de diversas fontes de informação, desde bases de dados semi-estruturadas até textos

em linguagem natural.

Inicialmente, a base de conhecimento (GKB) foi povoada com informação geográfica proveniente de autoridades portuguesas, sendo sucessivamente expandida com novas fontes de informação. O conhecimento geográfico que estava armazenado de forma distribuída, desconexa e 'escondida' em bases de dados, agora está acessível para qualquer aplicação da WS. Caso uma aplicação necessite somente dos nomes da GKB, o *GKB Ontology Generator* (GOG) permite que essa informação seja exportada em um formato orientado a nomes, mais simplificado.

Com a criação da GKB, a informação geográfica que antes estava distribuída e desconexa em várias fontes de informação, agora está reunida num único repositório e pode ser apresentada através de diversas visões do mesmo conteúdo.

Para complementar o conhecimento geográfico presente na GKB, eu desenvolvi o SEI-Geo, um sistema que identifica, reconhece e anota o conhecimento geográfico em textos e integra esse conhecimento em geo-ontologias. O SEI-Geo participou no Segundo HAREM com quatro corridas, o que permitiu avaliar seu desempenho tanto no reconhecimento de locais como no reconhecimento de relacionamentos entre os mesmos.

Durante o desenvolvimento e testes com o SEI-Geo verificou-se que o uso de conceitos e nomes das geo-ontologias tem que ser limitado em um determinado nível da geo-ontologia. Essa constatação vai ao encontro do que Mani et al. (2008) já haviam afirmado para textos em língua inglesa: quanto maior a ontologia ou almanaque utilizado, maior o grau de ambiguidade entre os locais. Nos experimentos realizados nessa tese verificou-se que com geo-ontologias maiores a ambiguidade aumenta tanto entre locais quanto entre nomes de locais com organizações e pessoas.

Outra contribuição dessa tese é um novo formato de representação de conhecimento geográfico extraído de textos, conforme descrito na Seção 5.2. Esse formato permite às aplicações que lidam com informação geográfica importar o conhecimento geográfico formalizado, ao contrário daquele encontrado em textos crus. A proposta do novo formato foi formalizar e relacionar o conhecimento geográfico presente em LN. Esse conhecimento dificilmente é aproveitado por aplicações informáticas dada a dificuldade de identificação e reconhecimento dos locais.

Apesar de o trabalho descrito nessa tese ter alcançado resultados significativos

em termos de impacto na comunidade de RIG e de PLN, essa última na área de reconhecimento de locais e de relacionamentos entre os mesmos, algumas limitações ficam evidentes.

8.2 Limitações

As limitações deste trabalho estão em ambas as áreas de bases de dados e PLN. Do lado da GKB, as limitações principais são as seguintes:

Raciocínio sobre a informação geográfica na GKB: a GKB ainda não possui um sistema de raciocínio que permita inferir fatos complexos. Contudo, as geo-ontologias produzidas pela GKB podem ser usadas por sistemas de raciocínio como *Renamed Abox and Concept Expression Reasoner* (RACER[1]) ou Jena[2], afim de suportar interrogações para aplicações clientes como sistemas de Resposta Automática a Perguntas (RAP), por exemplo.

Carência de informação da geografia física: a informação sobre a geografia física presente na GKB está restrita ao conhecimento exportado na WGO, o qual foi inserido na GKB a partir de informação criteriosamente selecionada da Wikipedia. Na Geo-Net-PT, a informação da geografia física ainda precisa ser integrada. Ressalta-se que o modelo conceitual já está pronto, sendo necessário apenas a inclusão das ocorrências.

Já o SEI-Geo, por ser um dos sistemas pioneiros para o português na tarefa de integração de conhecimento de texto para bases de conhecimento, está limitado no que diz respeito a:

Reconhecimento de relacionamentos limitado: o SEI-Geo reconhece poucos relacionamentos externos àqueles das geo-ontologias que recebe como entrada. Ou seja, poucos relacionamentos existentes em texto diferentes de inclusão são tratados pelo SEI-Geo.

Estratégia de integração de informação: essa estratégia deve ser tornada mais flexível de modo a facilitar a inclusão de novos conceitos na geo-ontologia. Para se inserir automaticamente um novo conceito geográfico

[1]http://www.sts.tu-harburg.de/~r.f.moeller/racer
[2]http://jena.sourceforge.net

aos conceitos existentes na geo-ontologia, ainda não existe nenhuma solução proposta na literatura que seja do meu conhecimento. Uma alternativa a ser testada é a co-ocorrência com outros conceitos previamente identificados. Contudo, essa abordagem ainda não permite identificar o tipo de relacionamento entre os conceitos que co-ocorrem.

Além dessas limitações, há exemplos de problemas para os quais ainda não tenho solução:

Na frase 'O Forte de Chaporá localiza-se no concelho de Bardez, no extremo norte de Goa, na costa oeste da Índia'. O 'concelho de Bardez' não é um novo concelho de Portugal, mas o SEI-Geo reconhece-o como sendo um novo concelho, apesar de não o conseguir relacionar com nenhuma entidade geográfica de Portugal.

Algumas organizações são reconhecidas como EG, por exemplo: 'Depois do Clube das Mulheres (em Portugal, Clube das Bigodudas) os peladões invadem os salões'. Forma um arbusto com as entidades 'Clube das Mulheres' e 'Portugal'.

Nomes de estádios de futebol compostos por nomes de conceitos são reconhecidos como locais com o nome do conceito como tipo, e não como construção. Exemplos dessa situação são 'Ilha do Retiro' e 'Parque Antarctica', construções que são reconhecidas como ilha e parque, respectivamente.

Outros casos de nomes de organização reconhecidos como locais são aqueles compostos por nomes de conceitos: 'Jardim de Infância das Galinheiras' e 'Parque de Ciência' são reconhecidos como local ao invés de organização.

8.3 Trabalho Futuro

Diversos são os trabalhos futuros que podem ser originados a partir da base deixada por esta tese. Especificamente, quanto às geo-ontologias e representação de conhecimento geográfico:

- **Criação de geo-ontologias nacionais:** qualquer país que deseje desenvolver uma geo-ontologia pode reutilizar a metodologia proposta nessa tese. As principais alterações devem ser realizadas na fase de limpeza de dados, que é dependente da língua na qual os nomes geográficos estão codificados.

- **Extensão da WGO:** o modelo da GKB já foi testado com sucesso para gerar uma pequena geo-ontologia em nível mundial, a qual precisa ser estendida. Essa geo-ontologia contém a maior parte dos nomes em inglês e há uma carência de nomes em espanhol e alemão, por exemplo, línguas que também têm sido amplamente utilizadas em avaliações de sistemas de RIG.

- **Criação de uma geo-ontologia mundial em português:** a metodologia utilizada para construir a Geo-Net-PT pode ser estendida para a geração de uma geo-ontologia mundial com nomes em português. Uma geo-ontologia em português dessa envergadura é uma carência atual para sistemas de PLN que utilizam informação geográfica.

- **Inclusão de uma dimensão temporal no modelo de dados:** um aspecto bastante importante ao lidar com informação geográfica é a dimensão temporal das entidades geográficas. Um nome ou conceito relevante num determinado período histórico deve ser representado com o período ao qual diz respeito. Por exemplo, 'Bona' foi a capital da 'Alemanha Ocidental' entre 1949 e 1990.

- **Uso de geo-ontologias por sistemas de Resposta Automática a Perguntas (RAP):** os sistemas de RAP devem cobrir uma gama muito ampla de padrões de perguntas e sustentar estratégias robustas para fornecer respostas minimamente próximas à expectativa dos usuários. Muitas vezes somente o uso de textos não é suficiente para responder às questões. Perguntas simples como: 'Quantos concelhos existem no distrito de Faro?[1]' ou perguntas mais complexas envolvendo conceitos históricos e atuais como: 'Quantos distritos têm todos os seus concelhos dentro de uma província portuguesa[2]?' podem ser facilmente respondidas quando geo-ontologias são utilizadas por tais sistemas. Um experimento interessante é aplicar um mesmo conjunto de perguntas envolvendo geografia em um sistema de RAP e comparar o desempenho do sistema com e sem o uso de geo-ontologias.

Quanto ao SEI-Geo:

[1] O distrito de Faro contém 16 concelhos.
[2] Nove distritos.

- **Testes de mutilação:** uma forma de avaliar a qualidade do SEI-Geo é através de testes de mutilação de geo-ontologias. Uma geo-ontologia pode ser mutilada em algum nível de sua hierarquia, com o intuito de permitir ao SEI-Geo repor o conhecimento mutilado. O SEI-Geo já realizou testes de mutilação com a WGO e a coleção CHAVE, mas apenas existem resultados preliminares.

- **Reconhecimento e integração de informação geográfica histórica:** por exemplo, na frase 'A freguesia de Anseriz, outrora pertencente ao concelho de Avô ...' o 'concelho de Avô', que já não existe, é mencionado junto com a 'freguesia de Anseriz'. Nesse caso, o 'concelho de Avô' deve ser integrado com um atributo caracterizando-o como 'histórico'. Contudo, ainda existe uma dificuldade a mais nesse aspecto, que é a detecção de quais as expressões em linguagem natural que descrevem nomes geográficos históricos.

- **Reconhecimento de endereços:** o SEI-Geo pode ser estendido para reconhecer endereçamentos postais com o objetivo de atualizar e/ou complementar os endereços presentes na GKB com informação proveniente dos textos e não só de bases de dados.

Um trabalho futuro comum à GKB e ao SEI-Geo é a integração conceitual entre conceitos desconhecidos encontrados em textos. A partir de uma geo-ontologia mundial definida com os conceitos utilizados por padrões e autoridades internacionais deve-se definir regras e heurísticas para integrar novos conceitos e ocorrências detectados em textos. Os resultados apresentados pelo SEI-Geo servem de parâmetro inicial para o desenvolvimento de um sistema mais robusto e que trate de casos mais complexos.

Outra questão a considerar como trabalho futuro é a desambiguação de nomes homógrafos. Por exemplo, o conceito de 'distrito' varia dependendo do país. Enquanto em Portugal é uma região administrativa no âmbito do país, no Brasil é uma região no âmbito de uma cidade ou município e na Índia é uma região que é parte de um estado. O desafio nesse ponto é identificar em texto os relacionamentos entre conceitos geográficos abstratos. A tarefa fica ainda mais complexa quando o objetivo é integrar conceitos atuais com conceitos históricos, pois as regras são escassas, quando existem.

Outros trabalhos futuros incluem:

- **Geo-referenciamento de recursos ontológicos:** geo-referenciamento automático de recursos ontológicos (p. ex. tesauros) ainda é uma ideia inexplorada considerando esses recursos em português. Sistemas de RIG e EIG podem usar recursos ontológicos geo-referenciados, uma vez que os termos com coordenadas geográficas atribuídas auxiliam no processo de desambiguação, por exemplo. Um tesauro do domínio de geo-ciências contém nomes de montanhas, relevos e penínsulas, entre outros. Atribuir coordenadas geográficas a esses nomes com auxílio de uma geo-ontologia amplia a gama de uso desse tesauro. Buscaldi e Rosso (2008) geo-referenciaram parte da WordNet em inglês, possibilitando a utilização desse recurso por aplicações de RIG e EIG que lidam com textos em inglês. O geo-referenciamento também pode ser aplicado para uma ontologia de eventos, por exemplo. Sistemas de pergunta e resposta e sistemas de RIG podem fazer uso de tal recurso, explorando a dimensão geográfica do recurso geo-referenciado.

- **Anotação geográfica:** as geo-ontologias produzidas nesta tese podem ser úteis para sistemas que fazem anotação geográfica. Por anotação geográfica entende-se o processo de adicionar metadados com identificação geográfica a vários tipos de mídia (p. ex. imagens, sítios da web e fontes RSS). As ocorrências das geo-ontologias possuem coordenadas geográficas que podem ser atribuídas à mídia que uma aplicação usa. Para isso, basta que uma aplicação projete as entidades geográficas identificadas na mídia sobre aquelas nas geo-ontologias. Uma vez feito o emparelhamento, a mídia é anotada com as coordenadas provenientes das geo-ontologias.

8.4 Reflexões Finais

A área de pesquisa sobre construção de ontologias ainda pode ser considerada embrionária, apesar de já existir um número significativo de ontologias disponíveis para a WS. O principal desafio para o desenvolvimento da WS nos próximos anos concentra-se no aproveitamento do conteúdo existente na web atual, com ênfase no acesso à semântica da informação (Allemang e Hendler, 2008). Nesse

contexto, os sistemas de extração de informação ganharão ainda mais espaço tanto na academia quanto na indústria.

A maior parte (mais de 95%) das ontologias disponíveis para a WS encontram-se sem ocorrências associadas aos conceitos (Ding e Finin, 2006). Tal fato leva essas ontologias a ficarem sub-utilizadas. Para colmatar esse problema, ocorrências provenientes de bases de dadas e textos, principalmente da web, na sua maioria, devem ser utilizadas.

A tarefa de formalizar o conteúdo presente na LN para o formato de triplas (p. ex. RDF) permanece um problema aberto. Mas mais importante do que minerar a informação dos textos é reutilizar e assumir como ponto de partida a informação já existente de forma estruturada em bases de dados. A informação que `Berlim` é a capital da `Alemanha` está em qualquer base de dados geográficos, mas o fato de `Bona` ter sido a capital da `Alemanha Ocidental` entre 1949 e 1990, pode ser mais facilmente encontrado em textos. Nesse exemplo, ainda está presente a variável temporal, na qual uma base de dados geográficos pode ser estendida.

As fases mais trabalhosas dessa tese foram a limpeza de dados antes de inseri-los na GKB e a parte de integração de informação extraída de texto. Os problemas descritos no Capitulo 3 evidenciam as dificuldades encontradas para tratar informação geográfica introduzida por humanos. A limpeza de dados geográficos também permanece um problema aberto.

A complexidade envolvida na tarefa de integração de informação geográfica pode ser justificada em parte pela ambiguidade inerente da linguagem natural. Parte dos problemas foram resolvidos pelo SEI-Geo, mas os mais complexos ainda continuam por resolver.

A tarefa de reconhecimento de relacionamentos em textos em português é outro campo de pesquisa que está em fase embrionária, sendo que a primeira avaliação de tais sistemas ocorreu apenas agora em 2008. Essa tese é uma das primeiras a demonstram preocupação em atacar esse problema. Os resultados dos sistemas participantes no Segundo HAREM (HAREM Clássico, pista do Tempo e de ReRelEM) sugerem que essas tarefas ainda precisam ganhar mais atenção, pois o grau de dificuldade proposto pela organização não foi alcançado pelos sistemas de forma satisfatória. Uma vez que os sistemas de REM trabalhem adequadamente com relacionamentos entre EM as ontologias e as bases de conhecimento serão alimentadas de forma mais rápida e consistente.

Apêndice A

Padrões Utilizados no SEI-Geo

Advérbio cá, aqui, lá e longe.

Direção lado, atrás, defronte e frente.

Fuzzy antes, depois, acima, abaixo, próxima, próximo, perto e proximidades.

Locativo em, na, nas, no, nos.

Métrico distante(s), distância, km(s), kilómetro(s), quilómetro(s), minuto(s), minuto(s), metro(s).

Orientação norte, sul, leste, oeste, nordeste, noroeste, sudeste, sudoeste.

Verbos chegar, chega, chego, chegou, chegava, chegávamos, chegamos, chegaram, era, falecer, faleci, falecido, faleceu, falecemos, faleceram, localizado, localizados, localizada, localizadas, localiza-se, localizar-se, localizava-se, morar, mora, moro, morou, morava, morávamos, moramos, moravam, morrer, morri, morrido, morreu, morremos, morreram, mudar-se, muda-se, mudou-se, mudava-se, mudávamos, mudamos, mudaremos, mudarei, mudaram-se, nascer, nasci, nascido, nascida, nasceu, nascemos, nasceram, situado, situados, situada, situadas, situa-se, situar-se, situava-se, situamos, situavam-se, sediada, sediado, sediadas, sediados, realizada, realizado, realizadas, realizados, viver, vivo, vive, viveu, vivemos, vivia, vivíamos, viviam, voltar, volta, volto, voltou, voltava, voltávamos, voltamos, voltaram, ir, íamos, vou, vai, vais, vamos, vão, fui, foi, fomos, foram, vem, viemos, viriam.

A. PADRÕES UTILIZADOS NO SEI-GEO

Substantivos água(s), afogada(s), afogado(s), beira(s), cabo(s), eleição(ões), favela(s), herdade(s), guerra(s), margem(ns), penitenciária(s), periferia(s), prefeito(s), ex-prefeito(s), praia(s), capital(ais), litoral(ais), natural(ais), procedente(s).

Apêndice B

Lista Negra Utilizada pelo SEI-Geo

última, últimas, último, últimos, além, algum, alguma, algumas, alguns, anonymous, anos, antes, aqui, artigo, as, assembléia, assembleia, até, autor, aventura, banda, botão, calendário, capital, carvalho, central, civil, colégio, com, comando, combustíveis, combustível, como, conferência, conheça, conhecer, criança, depoimento, depois, dia, diferença, direitos, edição, embora, entrega, escolha, essa, essas, esse, esses, esta, este, f-1, faculdade, fiat, fiesta, figueiredo, figura, filmes, fiscal, foca, ford, foto, fotografia, fundada, governo, guerra, hino, história, igreja, ii, informações, internet, jornal, lei, liga, lista, localização, móveis, margarida, mas, melhor, mercedes, mil, mulher, mulheres, mundial, não, natal, nirvana, notícia, notícias, nota, obras, os, ordem, página, pela, pelas, pelo, pelos, penso, plano, planta, press, prémio, projecto, quer, rádio, responsáveis, responsável, resultado, resultados, senna, silva, sistemas, site, sites, sobre, ssh, substituir, taça, também, toda, todas, totais, total, transportes, tutorial, u$, um, unita, url, us$, veja, velocidade, vossa, vosso, xl

Apêndice C

Procedimento para Avaliar o SEI-Geo no Primeiro HAREM

Este apêndice apresenta o procedimento utilizado para avaliar o módulo de extração de informação do SEI-Geo no Primeiro HAREM. Os módulos utilizados do esquema de avaliação do Primeiro HAREM foram: Extrator, AlinhaEM, AvalIDa, Véus, EMIR, ALTinaSEM, Ida2Sem e Alcaide.

perl extrairCDdasSubmissoes.pl -in colecao-primeiro-harem-completa-tagged-ate-concelho.xml -out input-alinha.txt -cdids cdids2006

java -Dfile.encoding=ISO-8859-1 -cp ferramentas_HAREM_java.jar pt.linguateca.harem.Aligner -submissao input-alinha.txt -cd cd2006rules.txt > EMs.alinhado

java -Dfile.encoding=ISO-8859-1 -cp ferramentas_HAREM_java.jar pt.linguateca.harem.IndividualAlignmentEvaluator -alinhamento EMs.alinhado > EMs.alinhado.avalida

java -Dfile.encoding=ISO-8859-1 -cp ferramentas_HAREM_java.jar pt.linguateca.harem.AlignmentFilter -alinhamento EMs.alinhado.avalida -categoria 'LOCAL(GEOGRAFICO,ALARGADO,ADMINISTRATIVO)' -estilo harem > EMs.veu

java -Dfile.encoding=ISO-8859-1 -cp ferramentas_HAREM_java.jar pt.linguateca.harem.SemanticAlignmentEvaluator -alinhamento EMs.veu > EMs.emir

java -Dfile.encoding=ISO-8859-1 -cp ferramentas_HAREM_java.jar pt.linguateca.harem.SemanticAltAlignmentSelector -alinhamento EMs.emir > EMs.altinasem

java -Dfile.encoding=ISO-8859-1 -cp ferramentas_HAREM_java.jar pt.linguateca.harem.GlobalSemanticEvaluator -alinhamento EMs.altinasem > EMs-ate-con.ida2sem

Referências

ISO 2788:1986 Documentation - Guidelines for the Establishment and Development of Monolingual Thesauri, Acessado em: abril 2002a. URL http://www.nlc-bnc.ca/iso/tc46sc9/standard/2788e.htm. 21

ISO/IEC 13250:2003, Topic Maps. Internet, 19 de maio 2002b. URL y12.doe.gov/sgml/sc34/document/0322_files/iso13250-2nd-ed-v2.pdf. 22

Eugene Agichtein e Luis Gravano. Snowball: Extracting Relations from Large Plain-Text Collections. In *Proc. of the 5th ACM International Conference on Digital Libraries (ACM DL)*, páginas 85–94, San Antonio, Texas, EUA, 2-7 de junho 2000. 1, 28, 99

Dean Allemang e James Hendler. *Semantic Web for the Working Ontologist: Effective Modeling in RDFS and OWL*. Morgan Kaufmann, maio 2008. ISBN 0123735564. 157

Harald Alvestrand. Request for Comments: 3066 - Tags for Identification of Languages. http://www.ietf.org/rfc/rfc3066.txt, janeiro 2001. Acessado em janeiro de 2007. 50

Masatoshi Arikawa, Takeshi Sagara, Kouzou Noaki, e Hideyuki Fujita. Preliminary Workshop on Evaluation of Geographic Information Retrieval Systems for Web Documents. In *NTCIR Workshop 4 Meeting Evaluation of Information Access Technologies: Information Retrievaland Question Answering and Summarization - Tóquio, Japão*, 2-4 de junho 2004. 26

Franz Baader, Diego Calvanese, Deborah McGuinness, Daniele Nardi, e Peter Patel-Schneider, editores. *The Description Logic Handbook: Theory, Imple-*

mentation and Applications. Cambridge University Press, 2003. ISBN 978-0521781763. 46

Ricardo Baeza-Yates e Berthier Ribeiro-Neto. *Modern Information Retrieval*. New York, NY: Addison-Wesley, 513 p., 1999. ISBN 978-0201398298. 27

Brandon Bennett e Pragya Agarwal. Semantic Categories Underlying the Meaning of 'Place'. In Stephan Winter, Matt Duckham, Lars Kulik, e Benjamin Kuipers, editores, *COSIT*, volume 4736 of *Lecture Notes in Computer Science*, páginas 78–95. Springer, 2007. ISBN 978-3-540-74786-4. 102

Tim Berners-Lee, James Hendler, e Ora Lassila. The Semantic Web. A New Form of Web Content that is Meaningful to Computers will Unleash a Revolution of New Possibilities. *Scientific American*, 284(5):34–43, maio 2001. 25

Vania Bogorny. *Enhancing Spatial Association Rule Mining in Geographic Databases*. Tese de Doutorado, PPGC - Instituto de Informática - Universidade Federal do Rio Grande do Sul, outubro 2006. 20

José Luis Borbinha, Gilberto Pedrosa, Diogo Reis, João Luzio, Bruno Martins, João Gil, e Nuno Freire. DIGMAP - Discovering Our Past World with Digitised Maps. In László Kovács, Norbert Fuhr, e Carlo Meghini, editores, *ECDL*, volume 4675 of *Lecture Notes in Computer Science*, páginas 563–566. Springer, 2007. ISBN 978-3-540-74850-2. 67

Karla Borges. *Uso de uma Ontologia de Lugar Urbano para Reconhecimento e Extração de Evidências Geo-espaciais na Web*. Tese de Doutorado, PPGCC - Instituto de Ciências Exatas - Universidade Federal de Minas Gerais, 2006. 4, 34, 36

Paolo Bouquet, Heiko Stoermer, Giovanni Tummarello, e Harry Halpin, editores. *Proc. of the WWW2007 Workshop I^3: Identity, Identifiers, Identification, Entity-Centric Approaches to Information and Knowledge Management on the Web, Banff, Canadá, 8 de maio, 2007*, volume 249 of *CEUR Workshop Proceedings*, 2007. CEUR-WS.org. 180

Christopher Brewster e Yorick Wilks. Ontologies, Taxonomies, Thesauri: Learning from Texts. In Marilyn Deegan, editor, *Proc. The Use of Computational*

Linguistics in the Extraction of Keyword Information from Digital Library Content Workshop, Londres, UK, 5 e 6 de fevereiro 2004. Centre for Computing in the Humanities, Kings College. 27

Christopher Brewster, Fábio Ciravegna, e Yorick Wilks. Background and Foreground Knowledge in Dynamic Ontology Construction: Viewing Text as Knowledge Maintenance. In Y. Ding, K. van Rijsbergen, I. Ounis, e J. Jose, editores, *Semantic Web, Workshop held the 26th Annual International ACM SIGIR Conference*, Toronto, Canadá, 28 de julho a 1º de agosto 2003. 4, 114

Christopher Brewster, Kieron O'Hara, Steve Fuller, Yorick Wilks, Enrico Franconi, Mark A. Musen, Jeremy Ellman, e Simon Buckingham Shum. Knowledge Representation with Ontologies: The Present and Future. *IEEE Intelligent Systems*, 19(1):72–81, 2004. ISSN 1541-1672. 18

Davide Buscaldi e Paolo Rosso. Geo-WordNet: Automatic Georeferencing of WordNet. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odjik, Stelios Piperidis, e Daniel Tapias, editores, *Proc. of the Sixth International Language Resources and Evaluation, LREC-2008*, Marrakech, Moroccos, maio 2008. European Language Resources Association (ELRA). ISBN 2-9517408-4-0. 157

Michael J. Cafarella, Doug Downey, Stephen Soderland, e Oren Etzioni. KnowItNow: Fast, Scalable Information Extraction from the Web. In *Proc. of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, páginas 563–570, Vancouver, British Columbia, Canadá, outubro 2005. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/H/H05/H05-1071. 1, 32

Nuno Cardoso e Diana Santos. Directivas para identificação e classificação semântica na colecção dourada do HAREM. Relatório Técnico DI-FCUL TR 06-18, Faculdade de Ciências da Universidade de Lisboa, Dezembro 2006. 4

Nuno Cardoso, Bruno Martins, Marcirio Silveira Chaves, Leonardo Andrade, e Mário J. Silva. The XLDB Group at GeoCLEF 2005. In Carol Peters, Fredric C. Gey, Julio Gonzalo, Henning Müller, Gareth J. F. Jones, Michael

Kluck, Bernardo Magnini, e Maarten de Rijke, editores, *Accessing Multilingual Information Repositories, 6th Workshop of the Cross-Language Evalution Forum, CLEF 2005, Vienna, Austria, 21-23 de setembro de 2005, Revised Selected Papers*, volume 4022 of *Lecture Notes in Computer Science*, páginas 997–1006. Springer, 2005. ISBN 3-540-45697-X. 66, 69

Nuno Cardoso, Bruno Martins, Daniel Gomes, e Mário J. Silva. *WPT 03: a primeira colecção pública proveniente de uma recolha da web portuguesa*, capítulo 23. IST Press, 2007. ISBN: 978-972-8469-60-8. 72, 80, 108

Nuno Cardoso, David Cruz, Marcirio Silveira Chaves, e Mário J. Silva. Using Geographic Signatures as Query and Document Scopes in Geographic IR. In *Advances in Multilingual and Multimodal Information Retrieval: 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007*, volume 5152 of *Lecture Notes in Computer Science*, páginas 802–810, Berlin / Heidelberg, 2008. Springer. Revised Selected papers. 66, 70

Xavier Carreras, Llus Màrquez, e Llus Padró. Simple Named Entity Extractor using AdaBoost. In Walter Daelemans e Miles Osborne, editores, *Proc. of Conference on Computational Natural Language Learning (CoNLL), Edmonton, Canadá*, páginas 152–155. Morgan Kaufman, 31 de maio a 1º de junho 2003. 25

David Celjuska e Maria Vargas-Vera. Semi-Automatic Population of Ontologies from Text. In J. Paralic, G. Polzlbauer, e A. Rauber, editores, *Proc. of the Fifth Workshop on Data Analysis WDA-2004*, páginas 33–49, Tatranska Polianka, Slovak Republic, junho 2004. ISBN:80-89066-87-9. 7

Marcirio Silveira Chaves. *Geo-ontologias e Padrões para Reconhecimento de Locais e de suas Relações em Textos: o SEI-Geo no Segundo HAREM*, capítulo 13, páginas 231–245. In Mota e Santos (2008), 7 de setembro 2008. ISBN 978-989-20-1656-6. 6, 98, 130

Marcirio Silveira Chaves. Mapeamento e Comparação de Similaridade entre Estruturas Ontológicas. Dissertação de Mestrado, Pontifícia Universidade Católica do Rio Grande do Sul - Faculdade de Informática - Programa de Pós-Graduação em Ciência da Computação, 2004. 28

Marcirio Silveira Chaves e Vera Lúcia Strube de Lima. Applying a Lexical Similarity Measure to Compare Portuguese Term Collections. In Ana L. C. Bazzan e Sofiane Labidi, editores, *Lecture Notes in Artificial Intelligence Advances in Artificial Intelligence - Proc. of the 17th Brazilian Symposium on Artificial Intelligence (SBIA2004) - São Luis, Maranhão, Brasil*, volume 3171, páginas 194–203. Springer, 29 de setembro a 1º de outrubro 2004. ISBN 3-540-23237-0. 28

Marcirio Silveira Chaves e Diana Santos. What Kinds of Geographical Information Are There in the Portuguese Web? In Vieira et al. (2006), páginas 264–267. ISBN 3-540-34045-9. 4, 7, 84

Marcirio Silveira Chaves, Mário J. Silva, e Bruno Martins. A Geographic Knowledge Base for Semantic Web Applications. In Carlos Alberto Heuser, editor, *Proc. of the 20th Brazilian Symposium on Databases*, páginas 40–54, Uberlândia, Minas Gerais, Brazil, 3-7 de outubro 2005a. ISBN 85-7669-029-2. 6, 70

Marcirio Silveira Chaves, Mário J. Silva, e Bruno Martins. GKB - Geographic Knowledge Base. DI/FCUL TR 05–12, Departamento de Informática da Faculdade de Ciências da Universidade de Lisboa, julho 2005b. 6, 70

Marcirio Silveira Chaves, Catarina Rodrigues, e Mário J. Silva. Data Model for Geographic Ontologies Generation. In José Carlos Ramalho, João Correia Lopes, e Luís Carriço, editores, *XATA2007 - XML: Aplicações e Tecnologias Associadas*, páginas 47–58. Universidade do Minho, Fevereiro 2007. ISBN 978-972-99166-4-9. http://hdl.handle.net/1822/6234. 70

William W. Cohen. Knowledge Integration for Structured Information Sources Containing Text. In *Workshop on Networked Information Retrieval - SIGIR-97*, Filadélfia, PA, EUA, 31 de julho 1997. 27

William W. Cohen, Pradeep Ravikumar, e Stephen E. Fienberg. A Comparison of String Distance Metrics for Name-Matching Tasks. In *Proceedings of the XVIII International Joint Conferences on Artificial Intelligence (IJCAI) - Workshop on Information Integration on the Web (IIWeb)*, páginas 73–78, Acapulco, México, 9-10 de agosto 2003. 27

Oscar Corcho, Mariano Fernández-López, e Asunción Gómez-Pérez. Methodologies, tools and languages for building ontologies: where is their meeting point? *Data Knowl. Eng.*, 46(1):41–64, 2003. ISSN 0169-023X. doi: http://dx.doi.org/10.1016/S0169-023X(02)00195-7. 40, 42

Helen Couclelis. People Manipulate Objects (but Cultivate Fields): Beyond the Raster-vector Debate in GIS. In Andrew U. Frank, Irene Campari, e Ubaldo Formentini, editores, *Theories and Methods of Spatio-Temporal Reasoning in Geographic Space*, páginas 65–77, Nova Iorque, 1992. Springer-Verlag. ISBN 978-3-540-55966-5. 16

Jim Cowie e Yorick Wilks. *Information Extraction*. A Handbook of Natural Language Processing: Techniques and Applications for the Processing of Language as Text. Marcel Dekker Inc., Nova Iorque, EUA, 2000. 24, 25

Isabel F. Cruz, Stefan Decker, Dean Allemang, Chris Preist, Daniel Schwabe, Peter Mika, Michael Uschold, e Lora Aroyo, editores. *The Semantic Web - ISWC 2006, 5th International Semantic Web Conference, ISWC 2006, Atenas, GA, EUA, 5-9 de novembro de 2006, Proceedings*, volume 4273 of *Lecture Notes in Computer Science*, 2006. Springer. ISBN 3-540-49029-9. 171, 176

Aron Culotta e Jeffery Sorensen. Dependency Tree Kernels for Relation Extraction. In *Proc. of the 42rd Annual Meeting of the Association for Computational Linguistics (ACL'04)*, páginas 423–429. Association for Computational Linguistics, julho 2004. 99

Hamish Cunningham. Information Extraction, Automatic. Preprint, 18 de novembro de 2004, at http://gate.ac.uk/sale/ell2/ie/main.pdf. *Encyclopedia of Language and Linguistics, 2nd Edition, Elsevier*, 5:665–677, 2006. ISBN 0-08-044299-4. 24

Hamish Cunningham, Diana Maynard, Kalina Bontcheva, e Valentin Tablan. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. In *Proc. of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02)*, páginas 168–175, Filadélfia, EUA, julho 2002. 1

Tiago Marques Delboni. Expressões de posicionamento como fonte de contexto geográfico na web. Dissertação de Mestrado, Universidade Federal de Minas Gerais - UFMG, 2005. 34, 35, 36, 71, 94

Ian Densham e James Reid. System Demo: A Geo-coding Service Encompassing a Geo-parsing Tool and Integrated Digital Gazetteer Service. In András Kornai e Beth Sundheim, editores, *Proc. of the HLT-NAACL 2003 Workshop: Analysis of Geographic References*, páginas 79–80, Edmonton, Alberta, Canadá, 31 de maio 2003. Association for Computational Linguistics. 24

Stephen Dill, Nadav Eiron, David Gibson, Daniel Gruhl, Ramanathan V. Guha, Anant Jhingran, Tapas Kanungo, Sridhar Rajagopalan, Andrew Tomkins, John A. Tomlin, e Jason Y. Zien. SemTag and Seeker: Bootstrapping the Semantic Web via Automated Semantic Annotation. In *Proc. of the Twelfth International World Wide Web Conference (WWW2003)*, páginas 178–186, Nova Iorque, NY, USA, 20-24 de maio - Budapeste, Hungria 2003. ACM Press. ISBN 1-58113-680-3. doi: http://doi.acm.org/10.1145/775152.775178. 7

Li Ding e Tim Finin. Characterizing the Semantic Web on the Web. In Cruz et al. (2006), páginas 242–257. ISBN 3-540-49029-9. 2, 158

George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, e Ralph Weischedel. Automatic Content Extraction (ACE) Program - Task Definitions and Performance Measures. In *Prooc. of the Fourth International Conference on Language Resources and Evaluation - LREC 2004*, 24-30 de maio 2004. 36

James Dowdall, Jeremy Elleman, Michael Hess, Will Lowe, e Fabio Rinaldi. The role of MultiWord Terminology in Knowledge Management. In *Fourth International Conference on Language Resources and Evaluation - LREC2004*, 24-30 de maio 2004. 24

Max J. Egenhofer. Toward the Semantic Geospatial Web. In *GIS '02: Proc. of the 10th ACM international Symposium on Advances in Geographic Information Systems*, páginas 1–4, Nova Iorque, NY, EUA, 2002. ACM. ISBN 1-58113-591-2. doi: http://doi.acm.org/10.1145/585147.585148. 20

Max J. Egenhofer e David M. Mark. Naive Geography. In Frankand A. U. e Kuhnand W., editores, *COSIT'95: Conference on Spatial Information Theory*, páginas 1–15. Springer-Verlagand Lecture Notes in Computer Sciences Nº 988, 1995. 16

Oren Etzioni, Michael Cafarella, Doug Downey, Ana-Maria Popescu, Tal Shaked, Stephen Soderland, Daniel S. Weld, e Alexander Yates. Unsupervised Named-Entity Extraction from the Web: An Experimental Study. *Artificial Intelligence*, 165(1):191–134, 2005. 1, 7, 30

Dieter Fensel. *Ontologies: Silver Bullet for Knowledge Management and Electronic Commerce*. Springer Verlag, Berlin Heidelberg, 138 p., 2001. ISBN 978-3540003021. 18

Mariano Fernández-López e Asunción Gómez-Pérez. Overview and analysis of methodologies for building ontologies. *The Knowledge Engineering Review*, 17(2):129–156, 2002. ISSN 0269-8889. doi: http://dx.doi.org/10.1017/S0269888902000462. 142, 146

Frederico Fonseca, Max Egenhofer, Peggy Agouris, e Gilberto Câmara. Using Ontologies for Integrated Geographic Information Systems. *Transactions in GIS*, 3:231–257, 6 2002. 17

Cláudia Freitas, Diana Santos, Hugo Gonçalo Oliveira, Paula Carvalho, e Cristina Mota. *Relações Semânticas do ReRelEM: Além das Entidades no Segundo HAREM*, capítulo 4. In Mota e Santos (2008), 7 de setembro 2008. ISBN 978-989-20-1656-6. 128

Gaihua Fu, Alia Abdelmoty, e Christopher Jones. Design of a Geographical Ontology. Relatório técnico, D5 3101 - SPIRIT - Spatially-Aware Information Retrieval on the Internet, 2003. 25, 26

Gaihua Fu, Christopher B. Jones, e Alia I. Abdelmoty. Building a geographical ontology for intelligent spatial search on the web. In M. H. Hamza, editor, *Databases and Applications. IASTED International Conference on Databases and Applications, part of the 23rd Multi-Conference on Applied Informatics, Innsbruck, Áustria, fevereiro 14-16*, páginas 167–172. IASTED/ACTA Press, 2005. ISBN 0-88986-462-4. 40, 41

Eric Garbinand e Inderjeet Mani. Disambiguating Toponyms in News. In *Proc. of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, páginas 363–370, Vancouver, British Columbia, Canadá, outubro 2005. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/H/H05/H05-1046. 73

Fredric Gey, Ray Larson, Mark Sanderson, Hideo Joho, Paul Clough, e Vivien Petras. GeoCLEF: the CLEF 2005 Cross-Language Geographic Information Retrieval Track Overview. In *Sixth Workshop of the Cross-Language Evaluation Forum: Working Notes for the CLEF 2005 Workshop (CLEF'2005)*, página s/ pp., Viena, Áustria, 21-23 de setembro, 2005. 10

Daniel Gomes e Mário J. Silva. Characterizing a National Community Web. *ACM Transactions on Internet Technology*, 5(3):508–531, 2005. ISSN 1533-5399. doi: http://doi.acm.org/10.1145/1084772.1084775. 108

Daniel Gomes, João P. Campos, e Mário J. Silva. Versus: a Web Repository. In *WDAS - Workshop on Distributed Data and Structures 2002*, Paris, França, 20 a 23 de março 2002. 48

Marco Gonzalez. Thesauri. Relatório técnico, Pós-Graduação em Ciência da Computação, Faculdade de Informática - Pontifícia Universidade Católica do Rio Grande do Sul, 2001. 21

Luis Gravano, Panagiotis G. Ipeirotis, Nick Koudas, e Divesh Srivastava. Text Joins in an RDBMS for Web Data Integration. In *Proc. of the 12th International Conference on World Wide Web - WWW'03*, páginas 90–101, Nova Iorque, NY, EUA, 2003. ACM Press. ISBN 1-58113-680-3. doi: http://doi.acm.org/10.1145/775152.775166. 27

Tom Gruber. A Translation Approach to Portable Ontology Specifications. *Knowledge Acquisition*, 5(2):199–220, 1993. 14, 18

Michael Grüninger e Mark S. Fox. Methodology for the Design and Evaluation of Ontologies. In *IJCAI'95, Workshop on Basic Ontological Issues in Knowledge Sharing, April 13*, 1995. URL http://citeseer.ist.psu.edu/grninger95methodology.html. 38, 40

Nicola Guarino. Understanding, Building and Using Ontologies. A Commentary to "Using Explicit Ontologies in KBS Development". *International Journal of Human and Computer Studies*, 46(2–3):293–310, 1997. ISSN 1071-5819. 18

Ralf Hartmut Güting. An Introduction to Spatial Database Systems. *The VLDB Journal*, 3(4):357–399, 1994. ISSN 1066-8888. 94

Marti A. Hearst. Automatic Acquisition of Hyponyms from Large Text Corpora. In *Proc. of the Fourteenth Conference on Computational Linguistics, Nantes, França*, páginas 539–545, Morristown, NJ, EUA, 23-28 de julho 1992. Association for Computational Linguistics. 33, 94

Marty Himmelstein. Local Search: The Internet Is the Yellow Pages. *Computer*, 38(2):26–34, 2005. ISSN 0018-9162. doi: http://dx.doi.org/10.1109/MC.2005.65. 4

Kaoru Hiramatsu e Femke Reitsma. Georeferencing the Semantic Web: Ontology based Markup of Geographically Referenced Information. In *Joint EuroSDR/EuroGeographics workshop on Ontologies and Schema Translation Services*, Paris, França, 15-16 de abril 2004. 26

ISO19109. Geographic Information - Rules for Application Schema. `www.seegrid.csiro.au/twiki/pub/Xmml/FeatureModel/19109_DIS2002.pdf`, Acessado em novembro de 2006. xx, 13, 14, 15, 49, 88

Jochen Lothar Leidner. *Toponym Resolution in Text: Annotation, Evaluation and Applications of Spatial Grounding of Place Names*. Tese de Doutorado, School of Informatics - University of Edinburg, 2007. 17

Vladimir Levenshtein. Binary Codes Capable of Correcting Deletions and Insertions and Reversals. *Soviet Physics Doklady*, 10(8):707–710, 1966. ISSN 0038-5689. 27, 42

Huifeng Li, Rohini K. Srihari, Cheng Niu, e Wei Li. Location Normalization for Information Extraction. In *Proc. of the 9th International Conference on Computational Linguistics (COLING2002)- Howard International House and Academia Sinica, Taipei, Taiwan*. Association for Computational Linguistics, 24 de agosto a 1º de setembro 2002. 25

Huifeng Li, Rohini K. Srihari, Cheng Niu, e Wei Li. Infoxtract Location Normalization: A Hybrid Approach to Geographic References in Information Extraction. In A. Kornai e B. Sundheim, editores, *HLT-NAACL 2003 Workshop Analysis of Geographic References*, páginas 39–44, Edmonton, Alberta, Canadá, 31 de maio 2003. Association for Computational Linguistics. 26

Mariano Fernández López, Asunción Gómez-Pérez, Juan Pazos Sierra, e Alejandro Pazos Sierra. Building a chemical ontology using methontology and the ontology design environment. *IEEE Intelligent Systems*, 14(1):37–46, 1999. ISSN 1094-7167. doi: http://doi.ieeecomputersociety.org/10.1109/5254.747904. 39

Robert Malouf. Markov Models for Language-independent Named Entity Recognition. In Dan Roth e Antal van den Bosch, editores, *Proc. of the Sixth Conference on Natural Language Learning (CoNLL-2002)*, páginas 187–190, Taipei, Taiwan, 31 de agosto a 1º de setembro 2002. Morgan Kaufman. 25

Hugo Manguinhas, Bruno Martins, e José Borbinha. A Geo-Temporal Web Gazetteer Service Integrating Data From Multiple Sources. In *3rd IEEE International Conference on Digital Information Management*, Londres, Reino Unido, Novembro 2008. University of East London, IEEE. 17, 67

Inderjeet Mani, Janet Hitzeman, e Cheryl Clark. Annotating Natural Language Geographic References. In *LREC 2008, Workshop Methodologies and Resources for Processing Spatial Language*, páginas 11–15, 31 de maio 2008. 118, 152

Dimitar Manov, Atanas Kiryakov, Borislav Popov, Kalina Bontcheva, Diana Maynard, e Hamish Cunningham. Experiments with Geographic Knowledge for Information Extraction. In *Proc. Workshop on Analysis of Geographic References - Edmonton, Canadá*, 2003. 23

Bruno Martins. *O Sistema CaGE no Segundo HAREM*. In Mota e Santos (2008), 7 de setembro 2008a. ISBN 978-989-20-1656-6. 128

Bruno Martins. *Geographically Aware Web Text Mining*. Tese de Doutorado, Universidade de Lisboa, Faculdade de Ciências, Departamento de Informática, agosto 2008b. 128

Bruno Martins e Mário J. Silva. A Statistical Study of the WPT-03 Corpus. Relatório técnico, TR 04-04 Departamento de Informática da Faculdade de Ciências da Universidade de Lisboa, 2004. 108

Bruno Martins, Nuno Cardoso, Marcirio Silveira Chaves, Leonardo Andrade, e Mário J. Silva. The University of Lisbon at GeoCLEF 2006. In Carol Peters, Paul Clough, Fredric C. Gey, Jussi Karlgren, Bernardo Magnini, Douglas W. Oard, Maarten de Rijke, e Maximilian Stempfhuber, editores, *CLEF*, volume 4730 of *Lecture Notes in Computer Science*, páginas 986–994. Springer, 2006a. ISBN 978-3-540-74998-1. 69

Bruno Martins, Mário J. Silva, Sérgio Freitas, e Ana Paula Afonso. Handling Locations in Search Engine Queries. In Ross Purves e Chris Jones, editores, *GIR*. Department of Geography, University of Zurich, 2006b. 20

Bruno Martins, Mário J. Silva, e Marcirio Silveira Chaves. *O Sistema CaGE no HAREM - Reconhecimento de Entidades Geográficas em Textos em Língua Portuguesa.*, capítulo 8, páginas 97–112. In Santos e Cardoso (2007), 2007. ISBN: 978-989-20-0731-1. 66, 69

Kevin S. McCurley. Geospatial Mapping and Navigation of the Web. In *Proc. of the Tenth International Conference on World Wide Web (WWW'01), Hong Kong, Hong Kong*, páginas 221–229, Nova Iorque, NY, EUA, 1-5 de maio 2001. ACM. ISBN 1-58113-348-0. doi: http://doi.acm.org/10.1145/371920.372056. 4

Luke McDowell e Michael J. Cafarella. Ontology-Driven Information Extraction with OntoSyphon. In Cruz et al. (2006), páginas 428–444. ISBN 3-540-49029-9. 33

Luke K. McDowell e Michael Cafarella. Ontology-driven, Unsupervised Instance Population. *Web Semantics: Science, Services and Agents on the World Wide Web*, 6(3):218–236, 2008. ISSN 1570-8268. doi: http://dx.doi.org/10.1016/j.websem.2008.04.002. 33

Deborah L. McGuinness e Frank van Harmelen. OWL Web Ontology Language. http://www.w3.org/TR/2004/REC-owl-features-20040210/, Acessado em outubro de 2004. 46

Andrei Mikheev, Marc Moens, e Claire Grover. Named Entity Recognition without Gazetteers. In *Proc. of the Ninth International Conference of the European Chapter of the Association for Computational Linguistics (EACL'99)*, páginas 1–8, Bergen, Noruega, 1999. 25

George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, e Katherine Miller. Wordnet: An On-line Lexical Database. *International Journal of Lexicography*, 4(3):235–244, 1990. doi: 10.1093/ijl/3.4.235. 29

Marco Modesto, Álvaro R. Pereira Jr., Nívio Ziviani, Carlos Castillo, e Ricardo Baeza-Yates. Um Novo Retrato da Web Brasileira. In *Proc. of the XXXII Seminário Integrado de Software e Hardware - SEMISH*, páginas 2005–2017, São Leopoldo, Brasil, 2005. 34

Cristina Mota e Diana Santos, editores. *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*. Linguateca, Aveiro, 7 de setembro 2008. ISBN 978-989-20-1656-6. 121, 168, 172, 175

Roberto Navigli e Paola Velardi. Learning Domain Ontologies from Document Warehouses and Dedicated Web Sites. *Computational Linguistics*, 32(2):151–179, 2004. 2, 7, 29, 35

John Paul. TGN2 General Guidelines. Disponível em: www.getty.edu/research/conducting_research/vocabularies/guidelines/tgn_2_general_guidelines.pdf, Acesso em: agosto 2009. 42

Bruno Pouliquen, Ralf Steinberger, Camelia Ignat, e Tom De Groeve. Geographical Information Recognition and Visualization in Texts Written in Various Languages. In *Proc. of the 2004 ACM Symposium on Applied Computing*, páginas 1051–1058. ACM Press, Nicosia, Chipre 2004. ISBN 1-58113-812-1. 25

Erhard Rahm e Hong Hai Do. IEEE Bulletin of the Technical Committee on Data Engineering. *Data Cleaning: Problems and Current Approaches*, 23(4), 2000. 54

Erik Rauch, Michael Bukatin, e Kenneth Baker. A Confidence-based Framework for Disambiguating Geographic Terms. In A. Kornai e B. Sundheim, editores,

HLT-NAACL 2003 Workshop Analysis of Geographic References, Edmonton, Alberta, Canadá, 2003. Association for Computational Linguistics. 26

Christoph Ringlstetter, Klaus U. Schulz, e Stoyan Mihov. Orthographic Errors in Web Pages: Toward Cleaner Web Corpora. *Computational Linguistic*, 32 (3):295–340, 2006. ISSN 0891-2017. doi: http://dx.doi.org/10.1162/coli.2006.32.3.295. 8

Dan Roth e Wen-Tau Yih. A Linear Programming Formulation for Global Inference in Natural Language Tasks. In *Proc. of the Eighth Conference on Computational Natural Language Learning (CoNLL-2004)*, páginas 1–8. Boston, MA, EUA, 6-7 de maio 2004. 99

Diana Santos e Nuno Cardoso, editores. *Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área*. Linguateca, 2007. ISBN: 978-989-20-0731-1. 38, 119, 176

Diana Santos e Marcirio Silveira Chaves. The Place of Place in Geographical IR. In *Proc. of the 3rd Workshop on Geographic Information Retrieval, SIGIR'06*, páginas 5–8, Seattle, EUA, 10 de agosto 2006. 4, 7, 84

Diana Santos e Paulo Rocha. The Key to the First CLEF with Portuguese: Topics, Questions and Answers in CHAVE. In Carol Peters, Paul Clough, Julio Gonzalo, Gareth J. F. Jones, Michael Kluck, e Bernardo Magnini, editores, *CLEF*, volume 3491 of *Lecture Notes in Computer Science*, páginas 821–832. Springer, 2004. ISBN 3-540-27420-0. 108

Diana Santos, Nuno Seco, Nuno Cardoso, e Rui Vilela. HAREM: an Advanced NER Evaluation Contest for Portuguese. In *Proc. of the 5th International Conference on Language Resources and Evaluation (LREC'2006)*, páginas 1986–1991, Genova, Itália, 22-28 de maio 2006. ELRA. 72

Luis Sarmento. SIEMÊS - A Named Entity Recognizer for Portuguese Relying on Similarity Rules. In Vieira et al. (2006), páginas 90–99. ISBN 3-540-34045-9. 72

Luís Sarmento. BACO - A Large Database of Text and Co-occurrences. In Nicoletta Calzolari, Khalid Choukri, Aldo Gangemi, Bente Maegaard, Joseph

Mariani, Jan Odjik, e Daniel Tapias, editores, *Proc. of the 5th International Conference on Language Resources and Evaluation (LREC'2006)*, páginas 1787–1790, Genova, Itália, 22-28 de maio 2006b. ELRA. 72

Mário J. Silva, Bruno Martins, Marcirio Silveira Chaves, Nuno Cardoso, e Ana Paula Afonso. Adding Geographic Scopes to Web Resources. *CEUS - Computers, Environment and Urban Systems - Elsevier Science*, 30(4):378–399, julho 2006. ISSN 0198-9715. 66

Humphrey Southall. Defining and Identifying the Roles of Geographic References within Text. In András Kornai e Beth Sundheim, editores, *HLT-NAACL 2003 Workshop: Analysis of Geographic References*, páginas 69–78, Edmonton, Alberta, Canadá, 31 de maio 2003. Association for Computational Linguistics. 25

Steffen Staab, Rudi Studer, Hans P. Schnurr, e York Sure. Knowledge Processes and Ontologies. *IEEE Intelligent Systems*, Janeiro/Fevereiro:25–33, 2001. doi: 10.1109/5254.912382. 40

Bill Swartout, Ramesh Patil, Kevin Knight, e Tom Russ. Towards distributed use of large-scale ontologies. In *Proc. of the 10th. Knowledge Acquisition for Knowledge-Based Systems Workshop*, Banff, Canadá, 1996. URL `http://ksi.cpsc.ucalgary.ca/KAW/KAW.html`. 39

Sylvie Szulman, Brigitte Biébow, e Nathalie Aussenac-Gilles. Structuration de Terminologies à l'aide d'outils de TAL avec TERMINAE. *Revue Traitement Automatique des Langues*, 43(1):103–128, 2002. ISSN 1248-9433. 7

Taro Tezuka e Katsumi Tanaka. Landmark Extraction: A Web Mining Approach. In Anthony G. Cohn e David M. Mark, editores, *Conference On Spatial Information Theory - COSIT*, volume 3693 of *Lecture Notes in Computer Science*, páginas 379–396. Springer, 2005. ISBN 3-540-28964-X. 16, 19

Taro Tezuka, Yusuke Yokota, Mizuho Iwaihara, e Katsumi Tanaka. Extraction of Cognitively-Significant Place Names and Regions from Web-Based Physical Proximity Co-occurrences. In Zhouand X. et al., editor, *Web Information Systems – WISE 2004.*, páginas 113–124, Berlin, 2004. Lecture Notes in Computer Scienceand 3306. Springer. 16

Peter Turney. Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL. In Luc De Raedt e Peter A. Flach, editores, *Machine Learning: EMCL 2001, 12th European Conference on Machine Learning, Freiburg, Alemanha, 5–7 de setembro, Proceedings*, volume 2167 of *Lecture Notes in Computer Science*, páginas 491–502. Springer, 2001. ISBN 3-540-42536-5. 31

Mike Uschold e Martin King. Towards a methodology for building ontologies. In *Workshop on Basic Ontological Issues in Knowledge Sharing, held in conduction with International Joint Conferences on Artificial Intelligence (IJCAI), Montreal, Quebeque, Canadá*. Morgan Kaufmann, 20–25 de agosto 1995. 39

Paola Velardi, Michele Missikoff, e Roberto Basili. Identification of Relevant Terms to Support the Construction of Domain Ontologies. In *Proc. of the workshop on Human Language Technology and Knowledge Management*, páginas 1–8, Morristown, NJ, EUA, 2001. Association for Computational Linguistics. 7

Renata Vieira, Paulo Quaresma, Maria das Graças Volpe Nunes, Nuno J. Mamede, Claudia Oliveira, e Maria Carmelita Dias, editores. *Computational Processing of the Portuguese Language, 7th International Workshop, PROPOR 2006, Itatiaia, Brasil, 13-17 de maio, 2006, Proceedings*, volume 3960 of *Lecture Notes in Computer Science*, 2006. Springer. ISBN 3-540-34045-9. 169, 178

Raphael Volz, Joachim Kleb, e Wolfgang Mueller. Towards Ontology-based Disambiguation of Geographical Identifiers. In Bouquet et al. (2007). 124

Yorick Wilks. The Semantic Web: Apotheosis of Annotation, but What Are Its Semantics? *Intelligent Systems, IEEE*, 23(3):41–49, 2008. ISSN 1541-1672. doi: http://doi.ieeecomputersociety.org/10.1109/MIS.2008.53. 1, 86

William E. Winkler. *Business Survey Methods*, capítulo Matching and Record Linkage, páginas 355–384. Wiley-Interscience, 1995. 27

George Kingsley Zipf. *Human Behavior and the Principle of Least Effort*. Addison-Wesley, Reading, Massachussetts, 1949. 74

Wenbo Zong, Dan Wu, Aixin Sun, Ee-Peng Lim, e Dion Hoe-Lian Goh. On Assigning Place Names to Geography Related Web Pages. In *JCDL '05: Proc.*

of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries, páginas 354–362, Nova Iorque, NY, EUA, 2005. ACM Press. ISBN 1-58113-876-8. doi: http://doi.acm.org/10.1145/1065385.1065464. 7

Índice