Programa de Doutoramento em Informática
Departamento de Informática
Faculdade de Ciências - Universidade de Lisboa

Uma metodologia para construção de ontologias e integração de conhecimento geográfico

Proposta de Tese para a Prova de Qualificação

Marcirio Silveira Chaves
mchaves@di.fc.ul.pt

Orientadores:
Mário J. Silva e Diana Santos

11 de Janeiro de 2007

Conteúdo

1 Motivação e contexto

A maior quantidade de conhecimento existente atualmente está disponível em textos na web. De acordo com [Wilks, 2005], 85% da informação disponível para ciência, empresas e aquela encontrada de modo informal na web está no formato não estruturado (texto, a maior parte). Entretanto, esse conhecimento precisa ser manipulado automaticamente, de modo a tornar-se verdadeiramente útil para sistemas que fazem algum tipo de processamento inteligente.

Neste ponto, surge a necessidade de estruturar o conhecimento que até o momento é legível 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 informação ainda não incorporada 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, Etzioni et al., 2005] descrevem trabalhos que fazem extração de informação a partir de textos em grande escala.

A maior parte desses trabalhos utiliza conjuntos muito 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, fazendo com que muito conhecimento ainda não esteja presente de modo formal (legível por máquina).

Por outro lado, alguma parte desse conhecimento já está disponível em bases de dados estruturadas, sendo de mais fácil representação e interpretação automática. Antes de extraí-lo de textos, deve-se verificar se este conhecimento ainda não está disponível de forma inteligível pelas máquinas. Uma vez que se detecte que o novo conhecimento ainda não é inteligível pelas máquinas, deve-se representá-lo de modo formal, em 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 na seção 2).

[Navigli e Velardi, 2004] observam que ontologias de domínio 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 ligação 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 SW não contêm ocorrências.

Nesse contexto torna-se essencial o povoamento de ontologias, por exemplo. Abordagens que lidam com povoamento (semi-)automático de ontologias, geralmente utilizam ontologias compostas somente por uma hierarquia de conceitos e tentam povoá-las. 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 conhecimento relevante e integração de conhecimento) apresentados até aqui, eu concentro este trabalho no domínio geográfico. A idéia básica é aproveitar o conhecimento geográfico existente em bases de dados publicamente disponíveis, integrá-lo e expandi-lo com conhecimento proveniente de textos da web. [Chaves e Santos, 2006] apresentam diversas características dos modos como o conhecimento geográfico pode ser encontrado em textos.

As dificuldades envolvidas na extração, limpeza e integração de informação, para prover um grau mínimo de qualidade a uma base de conhecimento são diversas e incluem:

- 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 LN é vaga e ambígua, o que dificulta o processo de extração de conhecimento. Como conseqüência dessas características, intrínsecas de qualquer língua, muito conhecimento relevante não é extraído porque não se consegue reconhecê-lo de forma adequada.
- As propriedades dos conceitos geográficos variam bastante. Um `rio`[1] poderá ser naturalmente caracterizado pela `nascente`, `foz`, `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 com que elas podem estar descritas.
- Dentro da disciplina da geografia (e de muitas outras) o conhecimento pode ser incerto, incompleto e contraditório [Gahegan e Pike, 2006]. Mesmo o conhecimento completo muitas vezes precisa ser desambiguado. Tratar computacionalmente essas características dentro do domínio geográfico permanece um desafio.
- 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. Um definição de EM pode ser encontrada em [Cardoso e Santos, 2006].

Ao contrário das ontologias de domínio, 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. 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 constituir uma

[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.

ontologia. [Himmelstein, 2005] encontrou um ou mais identificadores geográficos reconhecíveis e não ambíguos (e.g. códigos postais) em pelo menos 20% das páginas da web mundial. Em [Chaves e Santos, 2006, Santos e Chaves, 2006] nós 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 evidenciam que existe informação geográfica suficiente (e que não está presente em bases de dados administrativas) para suportar a construção e povoamento de uma ontologia geográfica.

Uma das inovações dessa proposta de tese é a utilização de textos de uma web inteira 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 [Velardi et al., 2001, Szulman et al., 2002, Celjuska e Vargas-Vera, 2004, Navigli e Velardi, 2004, Zong et al., 2005]. Outros trabalhos [Dill et al., 2003, Etzioni et al., 2005] utilizam toda a web para extrair fatos, tal como aqui proposto, mas não apresentam nenhuma metodologia para integrar o conhecimento extraído àquele existente.

As dificuldades na utilização de toda a web envolvem:

- a maior probabilidade de ocorrer ambigüidade, uma vez que o número de domínios dos textos é maior;
- os nomes geográficos e de conceitos são mencionados com uma variabilidade maior do que em textos de domínio específico (e.g. o conceito de `cidade` pode ser mencionado como `município`, `cidadezinha` e 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].

Nesse contexto se enquadram outras contribuições desta tese, a extração e integração em ontologias de conhecimento geográfico extraído de textos em linguagem natural. O povoamento será realizado pela busca de ocorrências em texto que possam ser objetos das classes previamente identificadas em uma ontologia geográfica.

A metodologia que será proposta nesta tese visa minimizar a distância entre o conhecimento informal descrito nos textos e o conhecimento formal expresso nas ontologias. Essa aproximação entre o conhecimento informal e o formal será realizada através do povoamento de ontologias e integração de conhecimento geográfico de diversas fontes. Adicionalmente, nos resultados desta tese é esperada uma caracterização da presença de EMs geográficas em textos na língua portuguesa.

Aplicações que podem fazer uso de ontologias geográficas 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 (e.g. `toca` e `cruza`) entre duas *features* espaciais. Uma *feature* é um objeto com significado em um domínio do discurso selecionado [ISO19109, 2006]. Junções espaciais são computacionalmente caras em bases de dados geográficas e podem ser reduzidas com a utilização de ontologias geográficas.

Figura 1: Áreas de conhecimento diretamente relacionadas a esta tese.

2 Conceitos e estado da arte

O trabalho descrito nessa proposta de tese engloba concretamente cinco áreas de conhecimento dentro da Informática (ver Figura 1), nas quais o domínio geográfico está sendo utilizado para testar a metodologia que será descrita nessa proposta.

Povoamento e extensão de ontologias são áreas mais recentes nas quais uma definição consensual ainda não existe. A tarefa de povoamento de ontologias (também chamada de Extração de Informação orientada à ontologia) tem como objetivo a extração e classificação de ocorrências de conceitos e relacionamentos definidos em uma ontologia.

Por outro lado, na tarefa de expansão de ontologias são adquiridos novos conceitos e relacionamentos a partir do texto e posteriormente acrescentados a uma ontologia existente. É importante observar que a expansão ocorre tanto em nível conceitual (novos conceitos e relacionamentos detectados em texto) quanto em nível de termos, os quais se tornarão ocorrências dos conceitos adicionados à ontologia existente.

Na área de representação de conhecimento, utilizam-se ontologias para tornar acessível por máquina o conhecimento geográfico existente. Neste trabalho deve-se entender ontologia como uma especificação explícita e formal de uma conceitualização compartilhada [Gruber, 1993]. [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. "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.

Conceitos na ontologia são representados por termos com significado importante no domínio geográfico (e.g. `província`, `distrito` e `concelho`). Tais conceitos são interligados através de relacionamentos (e.g. `parte-de` e `adjacência`). Os conceitos também são constituídos por propriedades (e.g. `comprimento`, `altitude` e `população`).

Outras formas de representar conhecimento, como tesauros, mapa de tópicos,

taxonomia, etc. serão discutidos no documento final da tese.

Para se conseguir representar conhecimento, é necessário, em primeiro lugar, identificá-lo e extraí-lo de fontes de informação, sejam elas bases de dados ou textos.

2.1 Extração de informação

Extração de Informação (EI) é uma sub-área do processamento da linguagem natural e diz respeito ao reconhecimento das propriedades e relações que são mencionadas em um ponto particular de um texto [Dowdall et al., 2004]. Neste trabalho, por propriedades deve-se entender as EMs geográficas e as relações são os conteúdos que indicam ligação entre EMs (e.g. uma `aldeia` é `parte de` uma `freguesia` ou de um `concelho`). O ponto particular de interesse aqui são os extratos de texto que mencionam informação geográfica.

Para [Cunningham, 2006] 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. Nesta proposta de tese, a necessidade de informação precisa é a informação geográfica sobre Portugal (num primeiro momento) e a pré-especificação se dá através dos conceitos, atributos e relacionamentos dentro de uma ontologia geográfica.

Eu selecionei alguns trabalhos relevantes que implementam EI e utilizam alguma estrutura formal para representar conhecimento.

O sistema Snowball [Agichtein e Gravano, 2000] recebe um conjunto de tuplas (e.g. [organização,localização]) definidas manualmente. 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>*.

O sistema KnowItAll recorre a uma abordagem livre de treino [Etzioni et al., 2005]. O KnowItAll permite a extração de grandes coleções de fatos, conceitos e alguns relacionamentos de modo não supervisionado, dependente ou independente de domínio e escalável.

A Tabela 1 apresenta uma comparação entre os trabalhos relacionados, no domínio da extração de informação, com o proposto nesta tese. 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 à essa tese são o uso de padrões (PAD), o uso de ontologias para apoiar a extração de informação (Onto) e a extração de entidades mencionadas (EEM). Além disso, outro fator a considerar é a integração do conhecimento adquirido durante a extração de informação àquele existente (ICA). Finalmente, o último parâmetro de comparação é o fato de os sistemas processarem textos escritos na língua portuguesa (PT).

Tabela 1: Comparação entre os trabalhos correlatos.

	PAD	Onto	EEM	ICA	PT
Snowball	✓	x	✓	x	x
KnowItAll/KnowItNow	✓	x	✓	x	x
[Alfonseca e Manandhar, 2002]	x	✓ (WordNet)	✓	✓	x
[Uryupina, 2003]	✓	✓ (dicionário geográfico)	✓	x	x
OntoLearn	x	✓ (WordNet)	✓	✓	x
[Borges, 2006]	✓	x	✓	x	✓

A maior parte dos trabalhos utilizam padrões léxico-sintáticos, extraem EMs e fatos. Padrões léxico-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.

No trabalho de [Alfonseca e Manandhar, 2002] é utilizada uma ontologia para apoiar a extração de ocorrências dos conceitos dessa ontologia. Esse trabalho integra o conhecimento adquirido com aquele existente utilizando uma ontologia lexical (WordNet) na língua inglesa. Uma ontologia lexical contém um subconjunto de palavras no vocabulário de uma linguagem natural. Essas palavras são definidas através de seus sentidos em qualquer domínio e seus relacionamentos com outras palavras. O sistema OntoLearn é outro sistema que também 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]. O OntoLearn faz a extração de ontologias de domínio, mas não propõe nenhum método para povoação de ontologias.

[Borges, 2006] usou 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.

2.2 Integração de dados e de informação

Após ter a informação extraída é necessário integrá-la com o conhecimento já existente. Essa tarefa é de extrema importância, uma vez que deve-se evitar cair no problema atual da WS, o qual consiste de diversas ontologias distribuídas mas sem ligação (relacionamentos entre elas). Nesse sentido, a tarefa de integração de informação será realizada como um complemento fundamental da tarefa de extração de informação.

A tarefa de integração de informação deve ser distingüida da tarefa de integração de dados. A integração de informação considera a posição hierárquica dos conceitos numa estrutura de representação de conhecimento (ver seção 2.2.2), enquanto a integração de dados normalmente investiga diferentes formas de similaridade entre cadeias de caracteres, sendo registros de bases de dados utilizados como estudos de caso, como por exemplo em [Cohen et al., 2003].

2.2.1 Integração de dados

A integração de dados provenientes de fontes de informação distintas, heterogêneas, complementares e autônomas permanece sendo um problema desafiador, embora diversos trabalhos já tenham fornecido estratégias de

integração de dados capazes de minimizar o problema [Levenshtein, 1966, Winkler, 1995, Cohen, 1997, Cohen et al., 2003, Gravano et al., 2003].

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, uma medida de similaridade lexical foi proposta em [Chaves e Lima, 2004, Chaves, 2004].

A utilização de vetores para representar documentos é uma abordagem clássica em RI [Baeza-Yates e Ribeiro-Neto, 1999]. [Gravano et al., 2003] utilizam vetores compostos por átomos de termos (i.e. palavras) com peso atribuído (projeção dos termos na forma de vetores) medindo-se o co-seno entre esses vetores. A intuição por trás dessa abordagem é que a magnitude de um componente (termo) de um vetor expressa a importância relativa dos átomos correspondentes na tupla representada pelo vetor. Intuitivamente, dois vetores são similares se eles compartilham muitos átomos importantes. A importância dos átomos segue o esquema TF/IDF utilizado em sistemas de RI. Ou seja, se um átomo aparece com muita frequência em uma relação sua importância é reduzida. Entretanto, quando se lida com átomos frequentes que fazem parte dos nomes de locais, sua importância não deve ser reduzida. Por exemplo, os átomos *Alto* e *Ilha* são bastante frequentes em uma ontologia geográfica e sua importância não deve ser reduzida, pois se isso acontecer os nomes dos termos *Alto do Castelo* e *Ilha do Castelo* podem ser considerados semelhantes.

Quando aplicados na comparação de nomes de locais, essas funções/medidas/algoritmos tendem a gerar mais entropia do que auxiliar na busca de pares de locais similares. Por exemplo, os seguintes pares de nomes possuem uma grande similaridade lexical considerando qualquer métrica de similaridade, mas são locais completamente distintos.

- *concelho de Lagoa* e *concelho de Lagos*
- *avenida D. Pedro I* e *avenida D. Pedro II*
- *aldeia de Rabaça* e *aldeia de Rabaçal*
- *Além da Fonte* e *Além da Ponte* (localidades em Portugal)
- *aldeia de Salgueirais* e *aldeia de Salgueiros*

Além disso, um dos objetivos da integração de informação nesse trabalho é expandir uma ontologia com nomes alternativos e abreviações dos nomes já existentes na mesma. Neste aspecto, o uso de medidas de similaridade lexical parecem não ser o método mais adequado.

Neste trabalho eu preciso adotar a noção de equivalência e não somente a de similaridade para integrar informação de fontes distintas. Para isso, eu manterei um algoritmo "conservador", no qual nomes de locais somente serão considerados equivalentes caso haja uma combinação perfeita entre dois nomes sendo comparados. Essa abordagem é contrária à implementada em [Cohen, 1997], no qual aquele autor acredita que a combinação exata de nomes não é uma abordagem confiável quando comparando pares de tuplas entre relações de bases de dados.

Entretanto, é importante destacar que esse procedimento somente será adotado para o tratamento de nomes próprios geográficos. Para nomes de conceitos, na maior parte dos casos, podem-se utilizar medidas de similaridade que permitam a dois termos serem considerados similares mesmo que não haja uma combinação perfeita entre seus caracteres. Por exemplo, os termos `aldeia`, `aldeias` e `aldeamento` podem ser considerados similares ao conceito de `aldeia`.

É importante ressaltar que antes da integração de informação ser finalizada, haverá uma etapa de eliminação da ambigüidade dos nomes para os casos de ambigüidade lexical. Por exemplo, a *aldeia de Parada* é um nome para dois locais diferentes em Portugal. Um é parte do *concelho de Alfândega da Fé* e outro parte do *concelho de Almeida*.

2.2.2 Integração de informação

A seção anterior descreveu a possibilidade de uso de medidas de similaridade lexical com nomes de locais bem como os diversos problemas que podem ser causados com seu uso. Assim, sugeriu-se por considerar similares aqueles conceitos que fazem uma combinação perfeita entre seus caracteres. Tão importante quanto a comparação em nível lexical é a comparação entre termos considerando suas posições numa estrutura conceitual, ou seja, o nível hierárquico em que se encontram.

[Rodríguez e Egenhofer, 2003] apresentam medidas de similaridade semântica entre ontologias que consideram a profundidade de cada conceito em uma ontologia bem como o conjunto de sinônimos ocorrendo com cada conceito. As medidas de similaridade propostas exploram a posição de cada conceito em cada hierarquia, bem como o conjunto de relacionamentos semânticos existentes no WordNet.

O conhecimento presente nos textos varia em abrangência e profundidade no que diz respeito ao conhecimento formal e estruturado presente em uma estrutura ontológica. Uma EM geográfica pode estar presente em uma frase junto a um ou mais tipos geográficos pertencentes a um mesmo nível em uma hierarquia (e.g. a *aldeia da Beira Baixa* ocorre com o *concelho de Castelo Branco* e com o *concelho do Fundão*). Neste caso, o tipo geográfico `concelho`, das ocorrências *Castelo Branco* e *Fundão*, está no mesmo nível da hierarquia.

Por outro lado, a *aldeia de São Miguel Outeiro* ocorre numa frase com o *concelho de Tondela* e em outra com o *distrito de Viseu*. Considerando que os tipos geográficos `concelho` e `distrito` estão em níveis diferentes na hierarquia, um algoritmo de integração de informação deve ser consciente dessa informação e integrar o novo conhecimento (neste caso `aldeia`) no nível mais específico de granularidade (neste caso `concelho`).

Outro problema na integração de informação geográfica são os fatos geográficos compostos por conceitos e relacionamentos inter-domínio. No

seguinte extrato de texto, retirado da Wikipedia, tem-se que: *O rio Douro (Duero, em castelhano) é um rio que nasce em Espanha, na província de Sória, nos picos da Serra de Urbião (Sierra de Urbión), a 2.080 metros de altitude e atravessa o norte de Portugal. A foz do Douro é junto à cidade do Porto. Tem 850 km de comprimento. Afluentes: Rio Paiva, Rio Sousa, Rio Tua.* É possível verificar diversas ocorrências de conceitos (`rios` e `serras`) que fazem parte da geografia física de Portugal bem como suas propriedades (`altitude` e `comprimento`). Neste caso, as conceitos e propriedades já estão definidas na ontologia e o povoamento é feito conforme a presença das ocorrências no texto.

Ainda no exemplo, pode-se observar que a *cidade do Porto* também é mencionada, mas essa é uma informação sobre a geografia administrativa que pode estar presente em uma ontologia geográfica já existente. Nesse caso, o desafio é integrar os fatos encontrados nos textos (neste caso, *rio(Douro)*, *comprimento(Douro,850 km)*, *cidade(Porto)*, *foz(Douro,Porto)*, etc.) com aqueles já existentes.

3 Objetivos

O objetivo geral desta tese é desenvolver uma metodologia para criação de ontologias geográficas a partir de bases de dados e de textos da web portuguesa. Para se alcançar esse objetivo três tarefas relevantes devem ser executadas (entre parênteses são mencionadas as áreas nas quais as tarefas se encontram):

1. Dimensionar a "geograficidade" presente em textos da web em português;
2. Reconhecer o conhecimento disponível em textos e gerar uma floresta de conceitos e ocorrências a partir de textos; (Extração de Informação e Representação de Conhecimento)
3. Integrar o conhecimento gerado na Geo-Net-PT. (Integração de Informação e Povoamento e Extensão de Ontologias)

Para executar as três tarefas acima será necessário responder às seguintes questões:

- Quais os conceitos, propriedades e relacionamentos geográficos presentes nos textos que podem ser representados numa ontologia e são ao mesmo tempo relevantes para aplicações de PLN?
- Quais são os conceitos (e.g. `cidades`, `rios` e `serras`) de ocorrências geográficas existentes nos textos da web em português?

3.1 Contribuições

Com as respostas para as questões acima, será possível propor uma metodologia para povoamento e integração de conhecimento geográfico em ontologias. Nesse trabalho, o conhecimento geográfico é representado em uma ontologia composta 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 tanto com a informação proveniente das bases de dados quanto com a informação relevante extraída dos textos.

Os principais resultados esperados deste trabalho são:

- uma arquitetura para um sistema de gerenciamento de conhecimento geográfico;
- um sistema para extração de conhecimento geográfico de textos e integração desse conhecimento em ontologias geográficas, utilizando textos da web em português;
- a construção e disponibilização pública e gratuita de uma ontologia geográfica de Portugal com conhecimento integrado de diversas fontes de informação, as quais são complementares, incompletas e provenientes de entidades com diferentes graus de autoridade. Essas fontes incluem todos os documentos em português da web portuguesa e bases de dados com informação geográfica sobre Portugal.

4 Trabalho realizado e em andamento

Antes de apresentar o trabalho realizado por mim até o momento, é necessário apresentar alguns recursos que estão sendo utilizados mas que não foram desenvolvidos por mim:

WPT 03: A WPT 03 é a coleção da web portuguesa de 2003 com aproximadamente 12 GB, e foi recolhida com os batedores (*crawlers*) do motor de pesquisa Tumba!. A WPT 03 conta com 3.775.611 documentos, dos quais aproximadamente 68,6% (2.590.641 documentos) estão escritos em português [Martins e Silva, 2004, Gomes e Silva, 2005, Cardoso et al., to appear]. Com a eliminação de documentos duplicados, a coleção possui 1.529.758 documentos em português.

BaCo: O BaCo (acrônimo para *Base de Co-ocorrências*) é uma base de dados construída a partir do WPT 03 que inclui, além das frases e da identificação do documento em que se encontram, várias tabelas de N-gramas que permitem testar rapidamente co-ocorrências e padrões em toda a WPT 03) [Sarmento, 2006b].

4.1 GKB 1.0

A primeira fase de desenvolvimento deste trabalho incluiu a criação de uma arquitetura para um sistema de gerenciamento de conhecimento geográfico. Essa arquitetura implementa a metodologia proposta e tem como componente principal a GKB - *Geographic Knowledge Base* - [Chaves et al., 2005b,a] um ambiente de extração e integração de conhecimento geográfico que é composto por conceitos e ocorrências de múltiplas fontes de informação geográfica administrativa (cada uma no seu formato específico) mais ocorrências da web, tais como nomes de sítios e domínios em Portugal. A GKB foi desenvolvida no âmbito do projeto GREASE [Silva et al., 2006], associado com o motor de busca tumba! [Silva, 2003], no pólo XLDB da Linguateca. O conteúdo da GKB é exportado no formato de ontologias (OWL), conforme indicação do W3C [Bechhofer et al., 2003]. Uma dessas ontologias é denominada Geo-Net-PT e se encontra publicamente disponível em `http://xldb.fc.ul.pt/geonetpt`.

A Geo-Net-PT está sendo utilizada em aplicações dentro do projeto GREASE. Na interface do Geotumba [Freitas et al., 2006], um motor de busca geográfico, pode-se verificar a presença da Geo-Net-PT para desambiguação de

Figura 2: Arquitetura do sistema de gerenciamento de conhecimento geográfico.

termos geográficos. O sistema de índice e ordenação geográfica do Geotumba [Martins et al., 2005] também faz uso da Geo-Net-PT, uma vez que a ordenação dos documentos é baseada no âmbito geográfico atribuído. O CAGE (*CApturing Geographic Entities*) [Silva et al., 2006], um sistema de REM geográficas, faz uso da Geo-Net-PT para identificar e classificar EMs geográficas em textos. Nesse contexto, percebe-se que a Geo-Net-PT tornou-se um componente fundamental dentro do sistema Geotumba. Além disso, essa ontologia é utilizada pelos participantes do Geo-CLEF (`http://ir.shef.ac.uk/geoclef`), uma avaliação conjunta de sistemas de recuperação de informação (RI) com diversas pistas, em que uma delas é dirigida a sistemas de RI geográficos.

A Figura 2 apresenta a arquitetura do sistema de gerenciamento de conhecimento geográfico que está em desenvolvimento. A GKB é um ambiente de extração e integração de conhecimento geográfico que, até o momento, 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 será realizada com a informação proveniente de textos da web. Esses textos fornecerão fatos geográficos que serão integrados no ambiente da GKB.

É importante ainda mencionar que em [Chaves et al., 2005b] é descrito também o desenvolvimento da GKB-ML, uma extensão para suportar várias funcionalidades não disponíveis no GKB. Entre essas funcionalidades pode-se citar nomes e relacionamentos geográficos em quatro línguas e gentílicos (adjetivos pátrios) associados a todos os países. A GKB-ML foi desenvolvida no âmbito do projeto GREASE para apoiar a participação em várias avaliações conjuntas, a saber: duas edições do Geo-CLEF, 2005 e 2006, HAREM e Mini-HAREM.

4.2 GKB 2.0

À medida que as ontologias geradas pela GKB foram sendo utilizadas, novos requisitos começaram a surgir, fazendo com que esse ambiente tivesse de ser estendido. Os principais requisitos são listados a seguir:

Suporte para relacionamentos entre tipos: A GKB 1.0 suportava somente relacionamentos entre *features*, como por exemplo, o *concelho do Gaia* é `parte do` *distrito do Porto*. A GKB 2.0 também deve suportar

relacionamentos entre tipos de *features* (conceitos), como por exemplo, os `rios` são `parte dos continentes`.

Suporte para propriedades genéricas: Todas as principais classes (`feature`, `type` and `name`) podem agora incluir atributos arbitrários, cujo suporte é definido no meta-modelo. Por exemplo, atributos para um rio incluem `fonte`, `foz` e `comprimento`, ao passo que um solo tem um `nível de PH` e uma montanha tem `altitude`. A GKB 2.0 também fornece uma especificação mais detalhada aos nomes de *features*. Atributos de nomes geográficos, como a língua na qual um nome é dado, podem ser capturados. Além disso, a GKB 2.0 pode armazenar outros atributos de nomes geográficos, tais como época e gentílicos. Por exemplo, *Olissipo* é um nome `histórico` de *Lisboa* e *lisboeta* é um dos `gentílicos` dos habitantes de *Lisboa*.

Melhor controle das fontes de informação: O controle sobre as fontes de informação foi muito simples e estava somente ao nível de *feature* na GKB 1.0. Por exemplo, toda informação relacionada ao *concelho de Lisboa* era associada a uma fonte de informação. Na GKB 2.0, cada nome, tipo e relacionamento pode ser independentemente associado a uma fonte de informação distinta. Esta extensão permite-nos saber, por exemplo, que dois tipos são fornecidos por fontes de informação distintas e o relacionamento entre eles é derivado de uma terceira fonte.

Disponibilização do conhecimento geográfico orientado à aplicação: Uma das limitações da GKB 1.0 era a disponibilização do conhecimento geográfico centrado no conceito de *feature*. A GKB 2.0 deve suportar a geração de diferentes representações sobre o mesmo conhecimento armazenado nela. Por exemplo, uma visão centrada em nomes geográficos, a qual pode ser muito útil para aplicações como reconhecedores de entidades mencionadas. Outras aplicações que estejam interessadas somente nos nomes e tipos de *features*, e não nos relacionamentos geográficos também podem querer fazer uso dessa representação.

A desenvolvimento da GKB 2.0 está no momento em curso. Além dos novos requisitos, a GKB 2.0 será estendida com conhecimento do domínio físico da geografia. Essa extensão implica a definição de conceitos e relacionamentos dentro desse domínio, bem como relacionamentos inter-domínio.

4.3 Medições iniciais

Até recentemente muito pouco se sabia sobre o conteúdo geográfico dos textos da web em português. [Delboni, 2005, Borges, 2006] apresentam estudos realizados com uma amostra da web brasileira e que descrevem os primeiros resultados caracterizando o conteúdo geográfico dos textos. Essa amostra foi composta por 75.413 documentos, nos quais foram encontrados 893.260 endereços em 57% (43.121) dos documentos. Tal fato também evidencia que a web brasileira contém muita informação geográfica. Apenas utilizando um reconhecedor de endereços, [Delboni, 2005, Borges, 2006] encontraram informação geográfica em mais de 50% dos documentos. Provavelmente, uma procura por nomes de estados, cidades e bairros (fora do contexto de endereços), por exemplo,

aumentaria mais esse percentual. Deve-se notar também, que esse conhecimento geográfico presente em textos da web brasileira ainda não está formalizado em uma estrutura de representação de conhecimento que permita a realização de raciocínio automático por parte de agentes computacionais.

Em [Chaves e Santos, 2006, Santos e Chaves, 2006], nós apresentamos medições iniciais sobre o conteúdo geográfico dos textos na web portuguesa utilizando a Geo-Net-PT. Os objetivos dessas medições incluem a detecção do conteúdo das EMs geográficas em texto, do grau de ambigüidade intracategorial (dentro do Geo-Net-PT e dentro das classificações de EM do SIEMÊS e do HAREM (e.g. a sobreposição entre nomes de pessoas e nomes de locais) e a sobreposição entre a Geo-Net-PT e os textos da web.

Nesse estudo foi utilizada uma amostra de 32.000 documentos etiquetados por um reconhecedor de EMs (SIEMÊS, [Sarmento, 2006a]). As categorias elencadas para o estudo foram: pessoa, organização e local. As duas primeiras foram utilizadas para verificar a ambigüidade existente com nomes de locais. Nós encontramos que 31% das EMs distintas reconhecidas como pessoa e 23% das EMs distintas reconhecidas como organização continham um nome geográfico incluído na Geo-Net-PT. Os locais constituem 30% do total de EMs identificadas. Considerando apenas os locais distintos, 75% são EMs multi-palavra.

Além das categorias, os tipos da categoria local também foram estudados. A maior parte (70%) dos tipos reconhecidos são locais com população, seguidos por endereços postais (7,3%) e tipos sócio-culturais (e.g. *Centro Cultural de Belém*, *Biblioteca Nacional*) (7,2%). Nós também identificamos a distribuição dos locais reconhecidos nos textos. Em 76% dos documentos da amostra foram reconhecidos pelo menos um local. Em média, foram reconhecidos sete locais distintos por documento contendo locais e a mediana foi igual a três locais.

Finalmente, nós concluímos que existe muita informação geográfica em texto que ainda não está presente na Geo-Net-PT e pode complementar essa ontologia.

5 Sistema de Extração e Integração de Conhecimento Geográfico – SEI-Geo

Esta seção descreve o SEI-Geo (acrônimo de Sistema de Extração e Integração de Conhecimento Geográfico), um sistema para extração e integração de conhecimento geográfico que tem como objetivo reconhecer o conhecimento geográfico disponível em textos, gerar uma representação estruturada desse conhecimento e integrá-lo na GKB. O sistema é composto por dois módulos: o de extração de informação geográfica (EIG) (ver Figura 3) e o de integração de conhecimento geográfico (ICG) (ver Figura 4).

O módulo de extração tem como objetivo detectar o conhecimento geográfico disponível em textos web e representá-lo de forma estruturada. A Figura 3 apresenta a arquitetura do módulo EIG.

Esse módulo recebe como entrada um conjunto de frases extraídas a partir de consultas à base de dados (BaCo). Esse módulo deve conter uma quantidade bastante abrangente de regras que indicam a presença de conceitos e relacionamentos nas frases. Tais frases, juntamente com conceitos da ontologia geográfica, são a entrada do módulo extrator de conteúdo geográfico que extrai

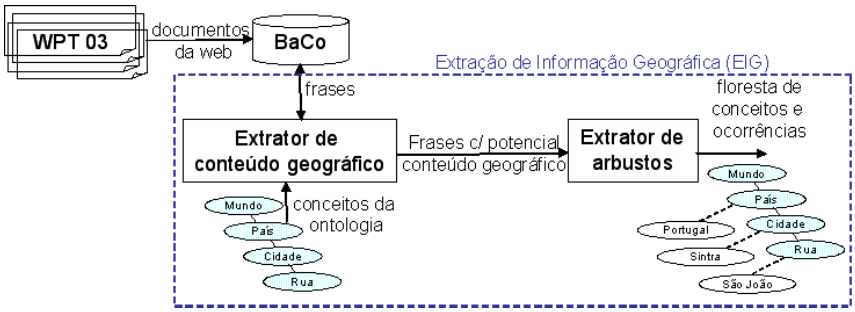

Figura 3: Arquitetura do módulo de extração de informação geográfica (EIG) do SEI-Geo.

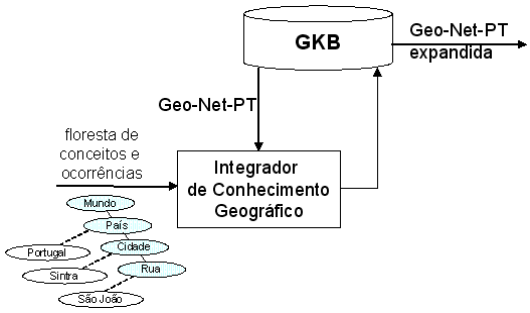

Figura 4: Arquitetura do módulo de integração de conhecimento geográfico (ICG) do SEI-Geo.

frases com potencial conteúdo geográfico. Essas frases são processadas pelo módulo extrator de arbustos que detecta ocorrências geográficas a partir dos conceitos fornecidos pela ontologia. O extrator de arbustos também tem uma função de filtro, na qual o conteúdo geográfico, duplicado ou sobreposto, é eliminado. O resultado desse processo é um conjunto de arbustos não necessariamente ligados, os quais são utilizados como entrada no módulo ICG. Em síntese, a saída do módulo de EIG será uma floresta de conceitos e ocorrências do domínio geográfico.

A Figura 4 apresenta a interação entre o módulo de integração de conhecimento geográfico com a GKB. O módulo de integração de conhecimento geográfico recebe também o conhecimento armazenado na GKB, faz a integração e retorna para a GKB o conhecimento geográfico expandido.

Esse módulo é um dos mais complexos do sistema, uma vez que ele deve fazer a integração entre um arbusto (extraído dos textos) e um grafo (Geo-Net-PT). A integração ocorrerá em dois níveis: conceitual e ocorrências. O nível conceitual lida com conceitos e relacionamentos, ao passo que o nível de ocorrências trata da integração das ocorrências. Por exemplo, um novo conceito (e.g. `aldeia`, `cidade`) extraído do texto deverá ser conectado a um ou mais conceitos na ontologia. O nível hierárquico, no qual o conceito será integrado, pode depender da sua utilização nos textos. Caso uma `aldeia` ocorra frequentemente com o conceito de `freguesia`, provavelmente seja um indício de que o conceito de `aldeia` deva ser conectado ao conceito de `freguesia` com o relacionamento `parte de`. Consequentemente, a ocorrência do conceito `aldeia` será integrada

como `parte da` ocorrência do conceito `freguesia`.

Outro problema tratado por esse módulo são as divergências encontradas nos fatos extraídos dos textos. Por exemplo, o comprimento do rio Douro no texto da Wikipedia em português é de 850 km, enquanto na versão de língua inglesa é de 897 km. Ao mesmo tempo a enciclopédia Britannica apresenta 895 km como o comprimento do mesmo rio. Neste caso, uma solução possível é armazenar todos os fatos encontrados juntamente com as respectivas fontes de informação.

A integração de conhecimento da web em ontologias, nesse trabalho, envolve várias tarefas distintas, entre elas:

- Encontrar as várias formas (nomes alternativos e abreviações, por exemplo) de descrição de uma EM existente em texto e verificar a qual entidade do mundo real se referem. Na área de banco de dados tal problema tem sido denominado por *duplicados aproximados*. Entretanto, ao contrário do procedimento realizado por essa comunidade, nesse trabalho os duplicados serão aproveitados como formas alternativas para se referir a mesma EM do mundo real e, consequentemente, serão integrados na ontologia.

- Encontrar informação geográfica complementar àquela existente na Geo-Net-PT e integrar essa informação no nível de granularidade mais adequado na ontologia.

A integração de conhecimento geográfico ocorre quando novos fatos geográficos são descobertos em texto ou quando fontes de informação públicas fornecem seus dados. Em ambos os casos estratégias de integração de informação devem estar presentes no SEI-Geo.

Uma entidade da geografia física, por exemplo um rio, pode estender-se por diversas `cidades` ou até mesmo `países`. Como a Geo-Net-PT já contém conhecimento sobre a geografia administrativa a nova informação contendo nomes de `rios` e as `cidades` onde eles `cruzam`, `nascem` ou `desaguam` pode ser integrada na Geo-Net-PT baseada no conhecimento lá existente.

É importante considerar que a tarefa de integração de informação é bem mais complexa e que, além de ocorrências de conceitos e relacionamentos, novos conceitos devem ser integrados conjuntamente.

Nas seguintes sentenças extraídas 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.

O SEI-Geo deve integrar a entidade da geografia física *Serra da Peneda* ao conceito existente na Geo-Net-PT *Alto Minho*. Além disso, a Geo-Net-PT deve ser expandida com a informação adicional *sistema montanhoso da Peneda-Gerês*, ou seja, incluir o novo conceito `sistema montanhoso` e sua ocorrência *Peneda-Gerês*.

Outra situação a ser mencionada juntamente com o procedimento de solução é a presença de conhecimento indireto no texto. Por exemplo, na 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 EMs geográficas *aldeia da Azinheira* e *distrito de Vila Real* e, além

disso, existe um relacionamento entre elas. Contudo, uma EM do tipo `aldeia` é parte de um tipo mais específico numa ontologia geográfica 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`.

Casos como esse surgirão e devem ser integrados na ontologia existente. Preferencialmente, a integração deve acontecer no nível mais específico da hierarquia, contudo, como isso nem sempre é possível, o conhecimento incompleto adquirido no texto não será jogado fora, 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* diretamente.

Informações geográficas históricas também devem ser integradas na ontologia existente. 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` 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.

Finalmente, o conhecimento armazenado na GKB é extraído com o uso da aplicação *Geographic Ontology Generator* (GOG), o qual tem como objetivo exportar o conteúdo da GKB para padrões internacionais de formalização de ontologias. O GOG tem sido usado e estendido para gerar a Geo-Net-PT e a GKB-ML.

5.1 Formalizando o resultado do SEI-Geo

Conforme a Figura 3, o SEI-Geo recebe um conjunto de frases, às quais deve ser extraído seu conteúdo geográfico. Esse conteúdo é integrado na GKB e exportado (formalizado) como ontologias geográficas.

Essa seção apresenta um exemplo de um texto em linguagem natural e sua representação ontológica. O texto:

O rio Douro (Duero, em castelhano) é um rio que nasce em Espanha, na província de Sória, nos picos da Serra de Urbião (Sierra de Urbión), a 2.080 metros de altitude e atravessa o norte de Portugal. A foz do Douro é junto à cidade do Porto. Tem 850 km de comprimento. Afluentes: Rio Paiva, Rio Sousa, Rio Tua.

pode ser representado no formato OWL como segue:

```
------------------------------------------------------------
<gn:Geo_Feature rdf:ID="GEO_238">                           |
  <gn:names>                                                |
    <gn:name="Porto" xml:lang="PT-PT" gn:att="P" gn:is="INE"/>| Conhecimento
  </gn:names>                                               | existente
  <gn:geo_type_id rdf:resource="#CON"/>                     |
  ...                                                       |
</gn:Geo_Feature>                                           |
------------------------------------------------------------
<rdfs:comment>Novo conhecimento integrado</rdfs:comment>
<gn:Geo_Feature rdf:ID="GEO_169">
  <gn:names>
    <rdf:Bag>
      <rdf:li gn:name="Douro" xml:lang="PT-PT" gn:att="P" gn:is="IGeoE"/>
      <rdf:li gn:name="Duero" xml:lang="ES-ES" gn:att="A" gn:is="IGP"/>
    </rdf:Bag>
```

```
  </gn:names>
  <gn:geo_type_id rdf:resource="#RIO"/>
  <gn:spring_location rdf:resource="#GEO_120"/>
  <gn:outlet_location rdf:resource="#GEO_238"/>
  <gn:affluent>
      <rdf:Bag>
          <rdf:li rdf:resource="#400"/>
          <rdf:li rdf:resource="#401"/>
          <rdf:li rdf:resource="#402"/>
      </rdf:Bag>
  </gn:affluent>
  <gn:length xml:unit="km">850</gn:length>
  <gn:info_source rdf:resource="#texto_web"/>
</gn:Geo_Feature>

<gn:Geo_Feature rdf:ID="GEO_120">
  <gn:names>
    <rdf:Bag>
      <rdf:li gn:geo_name="Sória" xml:lang="ES-ES" gn:att="P" gn:is="texto_web"/>
    </rdf:Bag>
  </gn:names>
  <gn:geo_type_id rdf:resource="#PRO"/>
</gn:Geo_Feature>

<gn:Geo_Feature rdf:ID="GEO_400">
  <gn:names>
    <rdf:Bag>
       <rdf:li gn:geo_name="Paiva" xml:lang="PT-PT" gn:att="P" gn:is="texto_web"/>
    </rdf:Bag>
  </gn:names>
  <gn:geo_type_id rdf:resource="#RIO"/>
</gn:Geo_Feature>

<gn:Geo_Feature rdf:ID="GEO_758">
  <gn:name>
    <rdf:Bag>
      <rdf:li gn:geo_name="Serra de Urbião" xml:lang="PT-PT" gn:att="P"
              gn:is="texto_web"/>
      <rdf:li gn:geo_name="Sierra de Urbión" xml:lang="ES-ES" gn:att="P"
              gn:is="texto_web"/>
    </rdf:Bag>
  </gn:name>
  <gn:geo_type_id rdf:resource="#SERRA"/>
  <gn:altitude xml:unit="m">2080</gn:altitude>
</gn:Geo_Feature>
```

A representação OWL acima descreve o conhecimento existente na GKB, o concelho (#CON) do Porto identificado por *GEO_238*. Porto é o nome preferido (*gn:att="P"*) desse concelho, está em português, na variante de Portugal (*xml:lang="PT-PT"*) e foi fornecido pela fonte de informação Instituto nacional de Estatística (*gn:is="INE"*).

O conhecimento extraído do texto é representado logo a seguir. Um `rio` pode ter como atributo `afluentes`, que são outros rios menores que desaguam num rio principal. O *rio Douro* é banhado por dez rios afluentes (no exemplo é apresentado apenas um (*GEO_400, Paiva*) a título ilustrativo). A `nascente` de um `rio` geralmente é mencionada em texto referindo-se a uma entidade geográfica administrativa (e.g. `província`, `concelho`, `freguesia`). Neste caso, a `nascente` (*spring_location*)do *rio Douro* é a *província de Sória* (*GEO_120*). Enquanto a `foz` de um `rio` pode referir-se uma entidade geográfica administrativa (e.g. `cidade`) ou física (e.g. um `mar`, um `rio` e um

`oceano`). Neste exemplo, a `foz` (*outlet_location*) é a cidade do Porto *GEO_238*, considerando-se os conceitos de `cidade` e `concelho` como equivalentes.

5.2 Avaliação do SEI-Geo

O objetivo da avaliação é verificar se a metodologia proposta para construção de ontologias geográficas é produtiva e em que grau de qualidade. Por um lado será possível avaliar os módulos EIG e ICG do SEI-Geo como segue:

EIG Verificar quais conceitos são mais produtivos para ocorrências geográficas em textos. Avaliação baseada na quantidade de átomos retornados e quantos desses são realmente geográficos. Quão ricos em profundidade são os arbustos da floresta de conceitos e ocorrências extraídos?

ICG Quantos arbustos gerados pelo módulo EIG são completamente e parcialmente integrados na Geo-Net-PT?

Outra perspectiva de avaliação diz respeito à utilidade das ontologias geográficas para as aplicações. O sistema CaGE utilizou as ontologias geográficas geradas pela GKB em várias avaliações conjuntas, a saber: duas edições do Geo-CLEF, 2005 e 2006, HAREM e Mini-HAREM. Os resultados obtidos pelo sistema nessas avaliações podem constituir indicações da qualidade das ontologias geradas pela GKB assim como da suas lacunas.

6 Plano e Cronograma

As tarefas que serão realizadas nessa tese serão descritas nessa seção. Inicialmente, o domínio da geografia administrativa será utilizado para extração de arbustos e a integração dos mesmos na GKB. Em seguida, será feita uma comparação dos resultados gerados pelo SEI-Geo com aqueles existentes na GKB. Finalmente, o mesmo processo será realizado para o domínio da geografia física.

Ao final de todo o processo espera-se ter uma ontologia geográfica composta pelos domínios administrativo e físico com conhecimento proveniente de diversas fontes de informação, incluindo textos e bases de dados.

A Tabela 2 apresenta o cronograma delineado para o restante dessa tese.
1 = Escrita e defesa da proposta
2 = Criação da GKB 2.0
3 = Implementação do Extrator de Informação Geográfica (EIG) (domínio administrativo)
4 = Implementação do Integrador de Conhecimento Geográfico (ICG) (domínio administrativo)
5 = Comparação dos resultados extraídos pelo EIG e pelo ICG com os presentes na GKB 1.0
6 = Participação no Geo-CLEF 2007
7 = Realização das tarefas 3-5 para o domínio físico, sendo que na tarefa 5 será utilizada a GKB 2.0.
8 = Escrita de artigos
9 = Escrita da tese

Tabela 2: Cronograma de execução das tarefas planejadas ao longo da tese.

	2007												2008								
Nº tarefa	Jan	Fev	Mar	Abr	Mai	Jun	Jul	Ago	Set	Out	Nov	Dez	Jan	Fev	Mar	Abr	Mai	Jun	Jul	Ago	Set
1	x																				
2	x																				
3	x	x	x	x	x																
4					x	x	x	x													
5								x	x	x											
6			x	x																	
7											x	x	x	x	x	x					
8				x	x					x	x					x	x	x	x	x	x
9																	x	x	x	x	x

6.1 Marcos do plano

- Junho de 2007: lançamento da Geo-Net-PT02 expandida com informação da geografia administrativa e física.
- Abril de 2008: lançamento da Geo-Net-PT03 expandida com informação proveniente de textos.
- Maio de 2008: versão beta de um protótipo do sistema SEI-Geo.

Agradecimentos

Marcirio Silveira Chaves é membro do pólo XLDB da Linguateca financiado pelo POSI/PLP/43931/2001 da Fundação para a Ciência e Tecnologia, co-financiado pelo POSI.

Referências

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, USA, June, 2-7 2000.

Enrique Alfonseca e Suresh Manandhar. An unsupervised method for general named entity recognition and automated concept discovery. In *Proc. of the 1st International Conference on General WordNet*, Mysore, India, 21-25 January 2002.

Ricardo Baeza-Yates e Berthier Ribeiro-Neto. *Modern Information Retrieval*. New York, NY: Addison-Wesley, 513 p., 1999.

Sean Bechhofer, Frank van Harmelen, Jim Hendler, Ian Horrocks, Deborah L. McGuinness, Peter F. Patel-Schneider, e Lynn Andrea Stein. OWL Web Ontology Language Reference. Disponível em: http://www.w3.org/TR/owl-ref/, 2003.

Vania Bogorny. *Enhancing Spatial Association Rule Mining in Geographic Databases*. Tese de doutorado, PPGC - Instituto de Informática - Universidade Federal do Rio Grande do Sul, October 2006.

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.

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, Canada, October 2005. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/H/H05/H05-1071`.

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.

Nuno Cardoso, Bruno Martins, Daniel Gomes, e Mário J. Silva. *WPT 03: Recolha da Web Portuguesa*, capítulo Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa. Diana Santos, to appear. (to appear).

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, June 2004. ISBN:80-89066-87-9.

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.

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, Brazil*, volume 3171, páginas 194–203. Springer, September 29 - October 1 2004. ISBN 3-540-23237-0.

Marcirio Silveira Chaves e Diana Santos. What kinds of geographical information are there in the portuguese web? In Renata Vieira, Paulo Quaresma, Maria da Graça Volpes Nunes, Nuno Mamede, Claudia Oliveira, e Maria Carmelita Dias, editores, *Proc. of the 7th Workshop on Computational Processing of Written and Spoken Portuguese, PROPOR 2006*, páginas 264–267, Itatiaia, Rio de Janeiro, Brazil, 13th - 17th May 2006. LNAI 3960 - Springer.

Marcirio Silveira Chaves, Mário J. Silva, e Bruno Martins. A Geographic Knowledge Base for Semantic Web Applications. In C. A. Heuser, editor, *Proc. of the 20th Brazilian Symposium on Databases*, páginas 40–54, Uberlândia, Minas Gerais, Brazil, October, 3–7 2005a.

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. URL `http://www.di.fc.ul.pt/tech-reports/05-12.pdf`.

William W. Cohen. Knowledge integration for structured information sources containing text. In *Workshop on Networked Information Retrieval - SIGIR-97*, Philadelphia, PA, USA, July 31 1997.

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 August 2003.

Hamish Cunningham. Information Extraction, Automatic. Preprint, 18th November 2004, at http://gate.ac.uk/sale/ell2/ie/main.pdf. *Encyclopedia of Language and Linguistics, 2nd Edition, Elsevier*, 5:665–677, November 2006.

Hammish 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, Philadelphia, July 2002.

Tiago Marques Delboni. Expressões de posicionamento como fonte de contexto geográfico na web. Dissertação de mestrado, Universidade Federal de Minas Gerais - UFMG, 2005.

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. ACM Press, May 20-24 - Budapest, Hungary 2003.

Li Ding e Tim Finin. Characterizing the Semantic Web on the Web. In Isabel F. Cruz, Stefan Decker, Dean Allemang, Chris Preist, Daniel Schwabe, Peter Mika, Michael Uschold, e Lora Aroyo, editores, *Proc. of the 5th International Semantic Web Conference*, volume 4273 do *Lecture Notes in Computer Science*, páginas 242–257, Athens, GA, USA, November 5-9 2006. Springer. ISBN 3-540-49029-9.

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*, 24th-30th May 2004.

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.

Dieter Fensel. *Ontologies: Silver Bullet for Knowledge Management and Electronic Commerce*. Springer Verlag, Berlin Heidelberg, 138 p., 2001.

Sergio Freitas, Ana Paula Afonso, e Mário J. Silva. Mobile Geotumba: Geographic Information Retrieval System for Mobile Devices. In *Proc. of the 4th MiNEMA Workshop*, páginas 83–87, Sintra, Portugal, July, 2-3 2006.

Mark Gahegan e William Pike. A situated knowledge representation of geographical information. *Transactions in GIS*, 10(5):727–749, November 2006. ISSN 1361-1682. doi: 10.1111/j.1467-9671.2006.01025.x.

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.

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, New York, NY, USA, 2003. ACM Press. ISBN 1-58113-680-3. doi: http://doi.acm.org/10.1145/775152.775166.

Tom Gruber. A Translation Approach to Portable Ontology Specifications. *Knowledge Acquisition*, 5(2):199–220, 1993.

Nicola Guarino. Understanding, Building and Using Ontologies. A commentary to "Using Explicit Ontologies in KBS Development". *International Journal of Human and Computer Studies*, 46:293–310, 1997.

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.

ISO19109. ISO 19109. `https://www.seegrid.csiro.au/twiki/pub/Xmml/FeatureModel/19109_DIS2002.pdf`, Acessado em novembro de 2006.

Vladimir Levenshtein. Binary Codes Capable of Correcting Deletionsand Insertions and Reversals. *Cybernetics and Control Theory*, 10(8):707–710, 1966.

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.

Bruno Martins, Mário J. Silva, e Leonardo Andrade. Indexing and ranking in Geo-IR systems. In *Proc. of the workshop on geographic information retrieval - GIR'05*, páginas 31–34, New York, NY, USA, 2005. ACM Press. ISBN 1-59593-165-1. doi: http://doi.acm.org/10.1145/1096985.1096993.

Roberto Navigli e Paola Velardi. Learning Domain Ontologies from Document Warehouses and Dedicated Web Sites. *Computational Linguistics*, 32(2):151–179, 2004.

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.

Andrea Rodríguez e Max Egenhofer. Determining Semantic Similarity Among Entity Classes from Different Ontologies. *IEEE Transactions on Knowledge and Data Engineering*, 15(2):442–456, 2003.

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, USA, August 10th 2006.

Luis Sarmento. SIEMÊS - a named entity recognizer for Portuguese relying on similarity rules. In Renata Vieira, Paulo Quaresma, Maria da Graça Volpes Nunes, Nuno Mamede, Claudia Oliveira, e Maria Carmelita Dias, editores, *Proc. of the 7th Workshop on Computational Processing of Written and Spoken Portuguese, PROPOR 2006*, páginas 90–99, Itatiaia, Rio de Janeiro, Brazil, 13-17 May 2006a. Springer.

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, Italia, 22-28 May 2006b.

Mário J. Silva. The Case for a Portuguese Web Search Engine. In *Proc. of the IADIS International Conference WWW Internet 2003*, páginas 411–418, Algarve, Portugal, November, 5-8 2003.

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, July 2006.

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.

Olga Uryupina. Semi-supervised learning of geographical gazetteers from the internet. In *Proc. of the HLT-NAACL 2003 Workshop on Analysis of Geographic References*, páginas 18–25, Morristown, NJ, USA, 2003. Association for Computational Linguistics.

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, USA, 2001. Association for Computational Linguistics.

Yorick Wilks. The Semantic Web as the apotheosis of annotation, but what are its semantics? In *Twentieth National Conference on Artificial Intelligence (AAAI'05)*, Pittsburgh, Pennsylvania, USA, July 9-13 2005.

William E. Winkler. *Business Survey Methods*, capítulo Matching and record linkage, páginas 355–384. Wiley-Interscience, 1995.

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, New York, NY, USA, 2005. ACM Press. ISBN 1-58113-876-8. doi: http://doi.acm.org/10.1145/1065385.1065464.