Pontifícia Universidade Católica do Rio Grande do Sul
Faculdade de Informática
Pós-Graduação em Ciência da Computação

Mapeamento e Comparação de Similaridade entre Estruturas Ontológicas

Marcirio Silveira Chaves

Dissertação apresentada como requisito parcial à obtenção do grau de mestre em Ciência da Computação.

Orientador: Prof[a]. Dr[a]. Vera Lúcia Strube de Lima

Porto Alegre, janeiro de 2004.

Dedico esta dissertação aos meus pais Alcino e Reinalda e, especialmente, à minha Cris.

Agradecimentos

Uma pós-graduação em nível de mestrado não é realizada apenas em dois anos, e sim é um projeto de vida no qual inúmeras pessoas participam dando suas contribuições. Gostaria de agradecer:

À Deus, por iluminar meu caminho ao colocar nele todas as pessoas especiais citadas a seguir.

À profª. Drª. Vera, pela confiança, desde o processo de seleção até a entrega deste volume, pelas oportunidades proporcionadas com grupos de pesquisa no Brasil e no exterior, pelo auxílio na participação do I Workshop em Tecnologia da Informação e Linguagem Humana em São Carlos-SP e, finalmente, pela dedicação nas correções em nível lexical. Aprendi a ler a vírgula da bibliografia :-). Esses e outros ensinamentos foram de extrema importância, aos quais procurarei seguir sempre que possível.

A todos aqueles professores que passaram pela minha vida, especialmente à profª. Drª. Renata Vieira por nortear os primeiros passos na iniciação científica. Ao profº. Drº. Sérgio Crespo e ao profº. MSc. Sandro Rigo por confiarem em mim. Ao profº. MSc. Marco Gonzalez pelas idéias nos artigos e no volume final da dissertação. Ao profº. Drº. Duncan Ruiz pelas oportunidades e ensinamentos durante o estágio docente.

À profª. Drª. Nieves Brisaboa por proporcionar minha participação nas Jornadas Iberoamericanas de Informática em Cartagena de Indias na Colômbia, pela experiência no Laboratório de Base de Dados da Universidade da Coruña na Espanha, em intercâmbio que ocorreu pela *Red Iberoamericana de Tecnologías del Software para la década del 2000* - RITOS2 incluída no subprograma VII (*Electrónica e Informática Aplicadas*) de CYTED (*Ciencia y Tecnología para el Desarrollo*) e por possibilitar a visita à Universidade Nacional de Educação à Distância em Madrid na Espanha. Ao pessoal do Laboratório de Base de Dados, especialmente à profª. Drª. Angeles Saavedra Places e ao doutorando Francisco Javier Rodríguez Martínez (Fran).

Ao HP-CPAD pelo financiamento da bolsa de mestrado.

Ao corpo docente do PPGCC pela dedicação na difícil tarefa de construir o conhecimento.

Ao pessoal da secretaria, sempre muito atencioso.

Aos atuais colegas e amigos(as). Ao Leonardo Langie pelo companherismo e troca de idéias cotidianas, ao Afonso pelas "tips" do Latex (valeu a pena), ao Brenner, ao João, ao Leandro, à Cláudia, à Sabrina, à Simone e ao Osmar. Aos ex-colegas e amigos(as) INFOUNI. Aos mestrandos do curso de Letras Gabriel de Ávila Othero e Isa Mara da Rosa Alves pela colaboração na etapa da análise humana da similaridade dos termos. Ao bolsista Vinícius Emanuel Wobeto, pelo auxílio na fase de preparação dos termos para a análise. Espero sempre ter boas notícias de vocês todos.

À minha Cris, pela eterna (assim espero) paciência em todos os momentos durante esta dissertação.

À família, em especial aos meus criadores :-), Alcino e Reinalda, pela compreensão e por entenderem a importância de um mestrado no processo de formação de uma pessoa. Mais dois agradecimentos especiais, à Morgana e ao Lorenzo, meu querido afilhado.

À minha outra família, seu Antonio, dona Wanda e a Cecília. Os almoços de domingo serão mais freqüentes :-).

"Ninguém está proibido de fazer melhor do que eu."

Martinho Lutero, teólogo alemão (10/11/1483-18/2/1546).

Resumo

Este trabalho apresenta um estudo e contribuições no contexto do mapeamento entre termos pertencentes a Estruturas Ontológicas (EOs) distintas. Para enfocar esse mapeamento são utilizadas medidas de similaridade, heurísticas e relações semânticas entre termos. As principais contribuições consistem da aplicação e extensão da medida de similaridade "Combinação de Caracteres" (CC) de Maedche e Staab para os termos pertencentes a EOs das línguas inglesa e portuguesa.

Os estudos relativos à língua portuguesa incluem a proposta de uma medida específica de similaridade, aqui denominada "Similaridade Lexical" (SL), que faz uso de um algoritmo de *stemming* e está baseada na medida CC. A medida SL foi validada através da realização de experimentos, nos quais foi possível refinar essa medida para utilizá-la na fase de avaliação, feita à luz de uma verificação humana de similaridade.

São ainda relatados experimentos com a relação semântica de sinonímia, de modo a detectar similaridade entre termos com o mesmo significado, porém com baixa similaridade lexical.

Palavras-chave: Reuso de informação, Estruturas Ontológicas (EOs), mapeamento e similaridade entre EOs, medidas de similaridade.

Abstract

This work presents a study and contributions on the context of mapping among terms of distinct Ontological Structures (OS). In order to achieve this mapping we use similarity measures, heuristics and semantic relations among terms. The main contributions consist of the application and extension of the similarity measure called String Matching (SM) proposed by Maedche and Staab for English and Portuguese ontologies.

The studies concerning Portuguese language include the proposal of a specific similarity measure, called Lexical Similarity (LS), that uses a stemming algorithm and is based on SM measure. LS measure was validated through several experiments, so that it was refined. During the evaluation phase, the LS measure was evaluated based on human evaluation of similarity.

Finally, we carried out experiments with synonymy semantic relation to detect similarity among terms with the same meaning, but with low lexical similarity.

Key-words: Reuse of information, Ontological Structures (OSs), mapping and similarity of OSs, similarity measures.

Sumário

Lista de Tabelas

Lista de Figuras

Lista de Símbolos e Abreviaturas

EO	Estrutura Ontológica	1
XML	*eXtensible Markup Language*	1
PLN	Processamento da Língua Natural	5
CYC	*enCYClopedia Project*	5
GUM	*Generalized Upper Model*	5
PS	*Post Script*	12
TCP/IP	*Transfer Control Protocol / Internet Protocol*	12
HTTP	*HyperText Transfer Protocol*	12
MARC	*MAchine Readable Cataloging*	12
HTML	*HyperText Markup Language*	12
PDF	*Portable Document Format*	12
ANSI	*American National Standards Institute*	12
RI	Recuperação de Informação	14
DAML-OIL	*DARPA Agent Markup Language - Ontology Interchange Language* ou *Ontology Inference Layer*	14
RDF	*Resource Description Framework*	14
W3C	*World Wide Web Consortium*	14
RDF(S)	RDF + RDF *Schema*	15
URI	Uniform Resource Identifiers	15
OIL	*Ontology Inference Layer* ou *Ontology Interchange Language*	16
LDs	Lógicas de Descrição	17
XOL	*XML-based Ontology-exchange Language*	17
DAML	*DARPA Agent Markup Language*	18
OWL	*Ontology Web Language*	18
FCA-Merge	*Formal Concept Analysis-Merge*	22
HICAL	*HIerarchical Concept ALignment system*	25
MAFRA	*MApping FRAmework for Distributed Ontologies*	27

OBSERVER	*Ontology Based System Enhanced with Relationships for Vocabulary hEterogeneity Resolution*	28
ONION	*ONtology compositION*	30
DE	Distância de Edição	33
CC	Combinação de Caracteres	33
SL	Similaridade Lexical	45

Capítulo 1

Introdução

"Alguns homens vêem as coisas como são, e dizem 'Por quê?' Eu sonho com as coisas que nunca foram e digo 'Por que não?' ".
George Bernard Shaw, escritor irlandês (1856-1950)

1.1 Motivação e Contexto do Trabalho

Entende-se por Estrutura Ontológica (EO) um conjunto de termos previamente definidos, associados de forma explícita por relações semânticas, em formato legível por humanos e por máquinas, aí incluindo-se coleções de vocabulários ou de conceitos.

Se duas pessoas desejam comunicar-se, ambas utilizam uma linguagem comum. Entretanto, se uma dessas pessoas desconhece a(s) linguagem(ns) empregada(s) pela outra, faz-se necessária a intervenção de um mediador externo que possibilite um vocabulário comum, de forma que a comunicação possa ocorrer com um mínimo de entendimento. Este mesmo princípio pode ser utilizado para estruturas ontológicas.

EOs contemplam a troca de dados não somente em nível sintático (como, por exemplo, a marcação de dados em XML), mas também em nível semântico compartilhado [1]. Em nível semântico é possível codificar dados de maneira formal, de modo a permitir que o conhecimento armazenado seja compreensível tanto por seres humanos quanto por máquinas.

Com o grande número de EOs disponíveis na *web* nos últimos anos, surgiram preocupações, tais como, estendê-las, adaptá-las e compará-las. Engenheiros, biólogos e usuários de EOs freqüentemente possuem uma EO (que passaremos a chamar EO-base), na qual navegam ou sobre a qual realizam consulta, mas é importante ter o conhecimento da similaridade entre a EO-base e as EOs recuperadas em outras aplicações. Um especialista de um domínio que deseja determinar a correlação entre duas EOs deve encontrar todos os conceitos que são similares entre as duas EOs, e registrar um mapeamento entre as EOs, para referência futura [2]. Neste contexto surge a necessidade de investigar medidas de similaridade que se proponham a estabelecer uma correspondência consistente entre termos pertencentes a EOs distintas.

Atualmente, as pesquisas que envolvem o mapeamento entre EOs incluem uma grande quantidade de trabalho manual, sendo as propostas mais avançadas [3, 4, 5, 6, 7] caracterizadas como semi-automáticas, pois ainda não se conta com técnicas que permitam automatizar completamente esse processo.

Noy e Musen [3] afirmam que o trabalho de mapear, unir ou alinhar EOs é realizado, na maior parte das vezes, à mão, sem qualquer ferramenta para automatização total ou parcial. Esse mapeamento manual é lento [8], tedioso e suscetível a erro [3, 9, 6, 2, 10]. Noy e Musen

ainda acrescentam que é um processo difícil de repetir e, simplesmente, não é prático, para certas aplicações. Para Doan [6], o mapeamento manual não é escalável no contexto da *web*. Uschold [8] comenta que o mapeamento automático é difícil e constitui um problema de pesquisa desafiador. Conforme Ding e Foo [11], "a necessidade de intervenção manual nos aspectos de *geração*, *mapeamento* e *evolução* de EOs atesta a natureza complexa da pesquisa nesta área e os problemas associados ainda não solucionados". A afirmação de Ding e Foo explicita a natureza abstrata dos três aspectos de pesquisa relacionados às EOs, a saber, geração, mapeamento e evolução, confirmada pelo fato de as EOs serem desenvolvidas para sistemas distintos, por seres humanos com visões de mundo diferentes.

Na literatura foram encontrados poucos trabalhos que se destinam a tratar o tema do mapeamento entre EOs (e os trabalhos encontrados estão voltados notadamente à língua inglesa [3, 6, 2, 12] e à língua alemã [1]), mas nenhum que demonstre preocupação em detectar termos similares no contexto de EOs da língua portuguesa.

Dessa forma apresentamos o tema desta dissertação, que foi detectado a partir de estudos que evidenciaram os problemas intrínsecos da comunicação entre sistemas projetados independentemente, e permitiram chegar às seguintes questão e hipótese de pesquisa:

Questão: Como mapear conceitos similares entre estruturas ontológicas diferentes?

Hipótese: Existe um grau de similaridade entre estruturas ontológicas projetadas independentemente, que pode ser detectado, de modo a permitir um mapeamento.

Em nossa hipótese consideramos a similaridade entre termos de domínios distintos, entretanto estamos cientes de que o ambiente ideal para realizarmos o trabalho é utilizando termos pertencentes ao mesmo domínio.

1.2 Objetivos e Método

Para estudar o problema do mapeamento entre EOs, este trabalho concentra-se na utilização de medidas de similaridade entre termos. Essas medidas permitem detectar termos equivalentes entre EOs e, geralmente, têm sido utilizadas como uma primeira etapa, em um processo de integração de informação.

Para prover um mapeamento entre EOs, apresentamos os seguintes objetivos específicos:

- aplicar a medida de similaridade conhecida como "Combinação de Caracteres" [1] a EOs das línguas inglesa e portuguesa, e avaliar seus resultados;
- prover meios para facilitar o mapeamento de EOs, de forma que o mesmo não seja realizado exclusivamente de forma manual;
- propor, com base nos resultados obtidos por avaliação das demais medidas estudadas, uma medida de similaridade para tratar termos pertencentes a EOs da língua portuguesa;
- validar e avaliar a medida de similaridade proposta.

São tratadas EOs pertencentes às línguas inglesa e portuguesa. Optamos por trabalhar, num primeiro momento, com EOs da língua inglesa, pelo fato de estas se encontrarem em maior número na literatura. Aplicamos uma medida de similaridade lexical já disponível na literatura para prover mapeamentos entre os termos dessas EOs. Posteriormente, concentramos nossos estudos em EOs da língua portuguesa, que possuem características distintas das EOs da língua inglesa quanto ao formato e à terminologia. Finalmente, uma medida de similaridade útil para auxiliar o mapeamento de termos em nível lexical é proposta, validada e avaliada à luz de uma comparação com a análise humana.

1.3 Organização desta Dissertação

Neste trabalho são relatados experimentos com medidas de similaridade em nível lexical, além de heurísticas que permitem melhorar o resultado do mapeamento gerado automaticamente.

Esta dissertação está organizada em 7 capítulos, precedidos desta introdução. O capítulo 2 apresenta o referencial teórico sobre o qual está baseada a dissertação. No capítulo 3 são descritos os trabalhos correlatos sobre mapeamento e similaridade entre EOs. No capítulo 4 é apresentado um estudo inicial com EOs da língua inglesa incluindo as medidas de similaridade utilizadas e as heurísticas propostas. O capítulo 5 descreve o tratamento das EOs da língua portuguesa e apresenta a medida de similaridade proposta nesta dissertação, incluindo os experimentos realizados na fase de validação. A avaliação dessa medida, a análise humana da similaridade entre os termos e os experimentos com a relação semântica de sinonímia, são descritos no capítulo 6. Finalmente, no capítulo 7, são apresentadas as conclusões, as limitações do trabalho, os trabalhos futuros e as considerações finais.

Capítulo 2

Referencial Teórico

"Life can only be understood backwards, but it must be lived forwards".
Soren Aabye Kierkegaard, filósofo (1813–1855)

2.1 Preâmbulo

Este capítulo apresenta o referencial teórico sobre o qual está fundamentado o presente trabalho. A seção 2.2 introduz o conceito de Estrutura Ontológica (EO), ou seja, como nós entendemos EO neste trabalho. A seguir, apresentamos as abordagens de mapeamento de EOs que têm sido utilizadas na literatura. Essas abordagens buscam minimizar a heterogeneidade semântica encontrada quando se deseja que sistemas se comuniquem. Heterogeneidade semântica é o assunto tratado na seção 2.3. Ao minimizar-se a heterogeneidade semântica estamos promovendo a interoperabilidade semântica, que é apresentada na seção 2.4.

Conceitos, em EOs, estão vinculados por meio de relações semânticas. As principais relações semânticas tratadas no Processamento da Língua Natural (PLN) são apresentadas na seção 2.5. Finalmente, são apresentadas as linguagens de marcação semântica mais amplamente encontradas na literatura.

2.2 Estrutura Ontológica (EO)

De acordo com Stumme e Maedche [9], não existe uma definição formal que especifique o que é uma EO. Neste trabalho consideramos Estrutura Ontológica (EO) um conjunto de termos previamente definidos, associados de forma explícita por meio de relações semânticas, em formato legível por humanos e por máquinas, aí incluindo-se coleções de vocabulários ou de conceitos.

A diversidade de sistemas e aplicações existentes na *web*, com linguagens de codificação e marcação distintas, faz com que a comunicação torne-se bastante complexa. Isso ocorre, principalmente, devido à inexistência de algum tipo de mecanismo capaz de promover um entendimento entre esses sistemas. Uma EO captura o conhecimento consensual, isto é, o conhecimento não restrito a algum indivíduo, mas aceito por um grupo.

Entretanto, a definição de conceitos em uma EO varia bastante de domínio para domínio, conforme apresentado por Chandrasekaran *et al.* [13] e retratado na figura 2.1.

A figura 2.1 apresenta quatro EOs definidas com a classe mais abstrata *Thing*. Contudo, o que deve ser enfatizado nesta figura é a variação existente nas classes que ocupam o próximo nível na hierarquia. Em CYC[1], *Thing* tem as subclasses *individual object*, *intangible* e *represented.*

[1]CYC (de *enCYClopedia*) Project - Disponível em *http://www.cyc.com/cyc-2-1/cover.html*

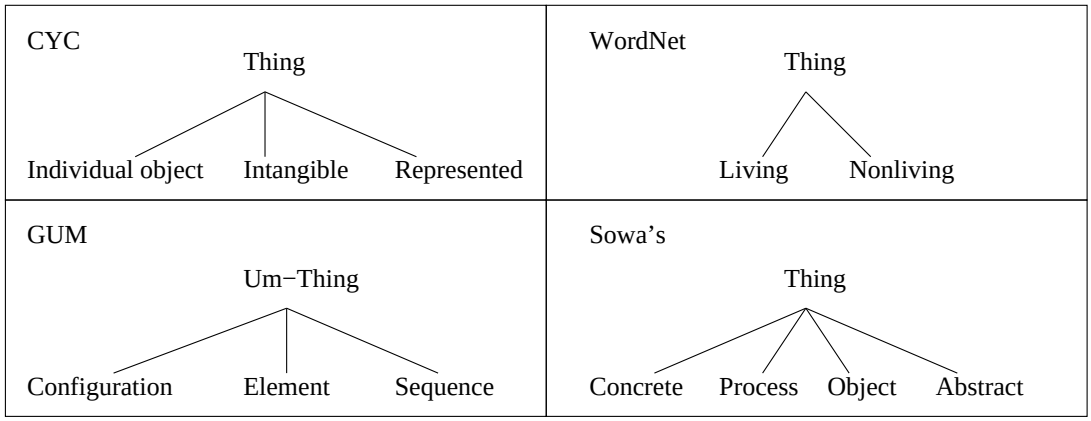

Figura 2.1: O modo como as EOs diferem, em sua análise de conceitos mais gerais [13]

No caso do GUM[2], *Um-Thing* tem as subcategorias *configuration*, *element* e *sequence*. *Thing*, no WordNet[3] [14], possui as subclasses *living thing* e *nonliving thing*. Por último, Sowa declara as subcategorias *concrete*, *process*, *object* e *abstract*, para *Thing*.

As diferenças entre as subclasses da classe *Thing* nas quatro EOs podem ser justificadas pelo fato de a construção de uma EO depender da intuição humana. Além disso, as EOs foram construídas para domínios de conhecimento distintos.

Apesar das diferenças apresentadas na figura 2.1, Chandrasekaran *et al.* [13] descrevem propriedades referentes às EOs, em torno das quais existe uma concordância geral da comunidade acadêmica:

- existem `objetos` no mundo;
- objetos têm `propriedades` ou `atributos`, que possuem `valores`;
- objetos podem relacionar-se com outros objetos através de `relações`;
- propriedades e relações podem mudar ao longo do `tempo`;
- existem eventos que ocorrem em diferentes `instantes de tempo`;
- existem `processos` dos quais os objetos participam e que ocorrem ao longo do tempo;
- o mundo, e seus objetos, podem estar em diferentes `estados`;
- eventos podem `causar` outros eventos, ou estados, como conseqüência;
- objetos podem ter `partes`.

De acordo com Chandrasekaran *et al.* [13], a EO de um domínio constitui o "coração" de qualquer sistema de representação de conhecimento para aquele domínio. Em síntese, uma EO é um modo de explicar o significado pretendido da informação. Uma EO identifica classes - cada uma caracterizada pelas propriedades que todos os elementos compartilham dentro daquela

[2] *Generalized Upper Model* - Disponível em *http://www.darmstadt.gmd.de/publish/komet/gen-um/newUM.html*

[3] É uma EO cujo objetivo é modelar o conhecimento lexical da língua inglesa. Está disponível em *http://www.cogsci.princeton.edu/∼wn/* para acesso *online*.

classe - e as organiza hierarquicamente [15]. De acordo com Gruber [16], uma EO pode ser definida como uma especificação formal e explícita de uma conceitualização. Fensel [17] descreve esse conceito em partes, afirmando que uma conceitualização se refere a um modelo abstrato de algum fenômeno no mundo, que identifica conceitos relevantes daquele fenômeno. Guarino [18] ainda comenta que tal conceitualização explica o significado pretendido dos termos usados para indicar relações relevantes. Por ser uma especificação formal, uma EO pode também dar origem a bases legíveis por máquina. Por ser uma especificação explícita, os tipos de conceitos usados e as restrições entre esses conceitos são definidos explicitamente.

Uma EO é um tipo de base de conhecimento que descreve conceitos por meio de definições, que são suficientemente detalhadas para capturar a semântica do domínio. Assim, a principal contribuição de uma EO é identificar classes de objetos e relações específicas existentes em um domínio.

Neste momento é importante explicitar que uma EO é freqüentemente equacionada com definições de classes e relações de inclusão, mas EOs não precisam estar limitadas a essas formas de definição [19]. Outras relações semânticas também podem ser incluídas de forma que o conhecimento do domínio esteja declarado explicitamente.

Tendo apresentado o conceito de EO, e tendo destacado que são as EOs um mecanismo para prover interoperabilidade, apresentamos, na seção que segue, as abordagens de mapeamento entre EOs.

2.2.1 Mapeamento entre Estruturas Ontológicas

De acordo com Ding e Foo [11], a tarefa de mapeamento entre EOs busca a reutilização de EOs existentes, expandindo-as e combinando-as de algum modo, e capacita a integração de uma grande quantidade de informação e conhecimento em diferentes domínios, para suportar uma nova comunicação e uso.

Para Prasad, Peng e Finin [12], mapear uma EO-base para outra EO-alvo consiste basicamente de, para cada conceito na EO-base, encontrar um conceito correspondente na EO-alvo, com semântica igual ou similar. Caso não exista correspondência na EO-alvo, o conceito não é mapeado.

A importância do mapeamento entre EOs pode ser melhor entendida através das seguintes demandas citadas em [11]:

- Reutilização das EOs existentes

Pelo fato de uma EO representar o conhecimento de um determinado domínio de forma explícita e não-ambígua, seu uso é indicado quando entidades pertencentes ao mesmo domínio precisam interoperar. Antes de desenvolver uma EO a partir do zero, é importante verificar as EOs já existentes, de modo a prover um mapeamento que permita a verificação da sobreposição entre as EOs. O objetivo desse mapeamento é o de reutilizar EOs já desenvolvidas.

- Expansão e combinação das EOs

No contexto dos serviços de recuperação de informações, à medida que vão sendo encontrados termos similares entre EOs, estes serviços podem, por exemplo, navegar através das EOs combinadas, de forma a otimizar o resultado de uma busca.

Considerando que uma EO-base nasce composta por um determinado conjunto inicial de termos e, ao ser utilizada, passa por um processo de mapeamento, os termos identificados em outras EOs como similares podem ser incorporados à EO-base. Dessa forma, cada termo da

EO-base que possui um termo similar em outra EO pode dar origem a um conjunto de termos, tais como, por exemplo, os *synsets* utilizados na WordNet. Através da combinação com termos similares, a EO-base passa por um processo de expansão.

O mapeamento direto entre EOs pode-se dar de duas formas. Em primeiro lugar, por meio da conexão entre as EOs e as fontes de informação, conforme a figura 2.2. Neste caso, as fontes de informação geralmente utilizadas são bases de dados. Por outro lado, o mapeamento pode ser realizado diretamente entre EOs na tentativa de identificar quais termos são de semântica similar, conforme a figura 2.3. Nosso trabalho não faz uso de fontes de informação mas, sim, busca alcançar o mapeamento entre os termos das EOs por meio da detecção da similaridade em nível lexical.

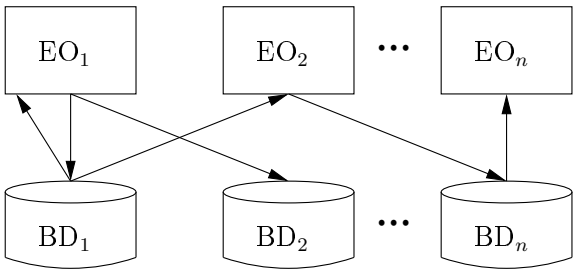

Figura 2.2: Mapeamento de Estruturas Ontológicas fazendo uso de bases de dados

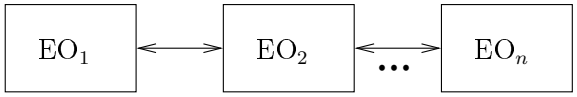

Figura 2.3: Mapeamento direto entre Estruturas Ontológicas

O mapeamento entre EOs (figura 2.3) tem sido tratado na literatura por meio de duas abordagens principais: união e alinhamento. Estas abordagens serão descritas nas próximas subseções.

2.2.1.1 União de Estruturas Ontológicas

Uma das primeiras abordagens para unir EOs encontrada na literatura é apresentada por Hovy em [20], e nela são descritas várias heurísticas para identificar conceitos correspondentes em EOs diferentes.

Para Noy e Musen [21], a união consiste na criação de uma EO coerente que inclui informação de todas as (pelo menos duas) EOs-base. Porém, na união de EOs, os termos que não combinam não são detectados nem tratados, conforme afirmam Hakimpour e Geppert [22]. Stumme *et al.* [23] ainda acrescentam que, no processo de união de EOs, a criação da nova EO ocorre de modo semi-automático, por meio do uso de conceitos de ambas as EOs, e pela identificação de similaridade entre alguns deles.

Conforme Sowa *apud* [24], união de EOs é o processo encarregado de encontrar associações entre uma EO-base e uma EO-alvo diferentes e derivar uma nova EO (que passaremos a chamar de EO_U) que facilite a interoperabilidade entre sistemas computacionais que estão baseados na EO-base e na EO-alvo, respectivamente. A EO_U pode substituir a EO-base ou a EO-alvo, ou ainda pode ser usada como um intermediário entre um sistema baseado na EO-base e outro baseado na EO-alvo.

A abordagem de união de EOs que cria uma EO unificada não é escalável, e é cara [25, 26]. A escalabilidade é comprometida pelo fato de os algoritmos serem de complexidade quadrática proporcionalmente ao número de nodos que serão percorridos na estrutura de dados. O alto custo é justificado porque o processo deve ser repetido quando surgem novas fontes, fazendo com que um especialista seja acionado a cada processo, e comprometendo a manutenção do sistema.

2.2.1.2 Alinhamento de Estruturas Ontológicas

"O alinhamento de EOs é muito relevante no contexto da *Web* Semântica. A *Web* Semântica será formada por muitas EOs de domínio específico com acesso gratuito. Para formar uma 'rede semântica' real - que permita aos computadores combinar e inferir conhecimento implícito - as EOs separadas devem ser alinhadas" [27].

O processo de alinhamento entre EOs é definido em [28] como sendo o mapeamento de conceitos e relações, entre uma EO-base e uma EO-alvo, que preserva em ambas a ordem parcial por subtipos. Dois conceitos ou relações são ditos equivalentes se um alinhamento mapeia um conceito ou relação x em uma EO-base, para um conceito ou relação y em uma EO-alvo. Caso não exista correspondência entre todos os conceitos ou relações, o mapeamento é dito parcial. A única alteração permitida antes de as EOs serem alinhadas é a inclusão de novos subtipos ou supertipos de conceitos ou relações em ambas as EOs, para prover alvos adequados para o alinhamento [28]. Entretanto, para realizar estas alterações, faz-se necessário um conhecimento prévio, por parte do engenheiro do conhecimento, sobre as EOs sendo alinhadas. Nosso trabalho busca evitar este tipo de intervenção manual antes da realização do mapeamento, fazendo com que o alinhamento ocorra de forma automática entre os termos existentes nas EOs. Apenas a etapa final, de verificação da consistência do mapeamento automático, requer intervenção manual.

Para Stumme *et al.* [23], alinhar EOs significa definir um mapeamento entre duas EOs que traduz conceitos da EO-base para a EO-alvo. Noy e Musen [21] afirmam que, no processo de alinhamento, as EOs devem estar consistentes e coerentes umas com as outras, mas mantidas separadamente. Nesse caso, as EOs persistem com ligações estabelecidas entre elas. O alinhamento geralmente ocorre quando as EOs cobrem domínios que são complementares uns aos outros [3]. Por exemplo, no domínio da educação, em uma universidade cada faculdade cria sua própria EO e, mais adiante, essas EOs precisarão ser mapeadas, por meio de alinhamento, para possibilitar a comunicação entre as mesmas.

A figura 2.4 ilustra a diferença entre união e alinhamento de EOs.

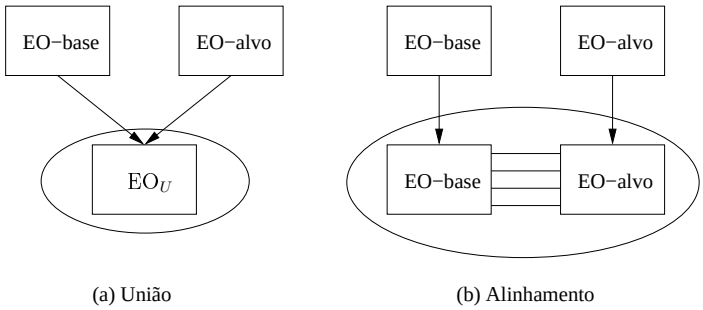

Figura 2.4: União e Alinhamento de Estruturas Ontológicas (adaptado de [3])

No processo de união de EOs, como se pode observar na figura 2.4 (a), a EO-base e a EO-alvo passam a ser representadas por uma única EO, ao passo que, no processo de alinhamento, figura

2.4 (b), a EO-base e a EO-alvo permanecem, e são criadas ligações entre os termos e relações nas mesmas.

Em um estudo mais profundo sobre alinhamento de EOs, Stumme *et al.* [23] propõem algumas questões que são expostas a seguir:

- Quais propriedades uma função deve ter que descrevam o alinhamento entre duas EOs?
- Deve este mapeamento ser uma função completamente definida? Ou esta função pode ser definida apenas parcialmente, quando os conceitos da EO-base não possuem um correspondente na EO-alvo?
- O mapeamento deve preservar a ordem da relação `é um`?
- O mapeamento deve ser injetivo (mapeamento 1-1) ou deve permitir mapear dois conceitos diferentes da EO-base para o mesmo conceito na EO-alvo?
- Como tais decisões influenciam *recall*[4], precisão[5] e o entendimento do resultado de um processo de recuperação de informações usando esta função de alinhamento?

Essas questões, além de permitirem ao leitor refletir sobre algumas dificuldades encontradas durante o processo de alinhamento, também apresentam o estado atual da arte na área da pesquisa do mapeamento automático entre termos de EOs diferentes.

2.2.1.3 União e Alinhamento de Estruturas Ontológicas

De acordo com Hovy [29, 20], na tarefa de união ou alinhamento de EOs, três situações podem surgir:

1. Dois termos são exatamente equivalentes; neste caso eles podem ser alinhados diretamente.
2. Um termo é mais geral do que o outro; neste caso o termo mais específico (e seus subordinados e possivelmente seus irmãos) podem ser integrados abaixo do termo mais geral.
3. Os termos são incompatíveis, neste caso:
   (a) um dos termos deve ser rejeitado e não incorporado, ou
   (b) um dos termos, e seus termos dependentes, devem ser redefinidos, ou
   (c) uma forma de representação separada deve ser criada, na qual os termos e todos os seus dependentes, existam em paralelo, ou ainda
   (d) uma versão "mais fraca" do termo em questão pode ser incorporada, sem definições ou relações que causem inconsistência.

As principais abordagens encontradas na literatura para as situações descritas acima incluem a comparação de termos ou conceitos em nível lexical e em nível semântico-estrutural.

As abordagens de união e alinhamento buscam reduzir os problemas de heterogeneidade semântica encontrados quando se deseja que sistemas se comuniquem. O problema da heterogeneidade semântica é descrito na próxima seção.

[4]Métrica utilizada para avaliar sistemas de RI. O valor dessa métrica representa a proporção de documentos relevantes que são realmente recuperados em uma coleção de documentos.

[5]Métrica utilizada para avaliar sistemas de RI. O valor dessa métrica representa proporção de documentos recuperados em um sistema de RI que são realmente relevantes.

2.3 Heterogeneidade Semântica

Heterogeneidade semântica[6] representa um "mal-entendido" nos significados, quando sistemas projetados independentemente são integrados. Esses "mal-entendidos", que geram diferentes interpretações para o mesmo dado, causam inconsistências semânticas entre *nomes*, *estruturas ou esquemas*, *atributos* e *granularidade de valores*, entre outras [22].

De acordo com Mitra e Wiederhold [26], as principais origens da heterogeneidade entre fontes de informação são:

1. fontes diferentes usam diferentes formatos de dados e diferentes linguagens de modelagem para representar seus dados e metadados;

2. fontes usando o mesmo formato de dados diferem, na organização estrutural e semântica da terminologia utilizada.

Essas fontes de heterogeneidade são o resultado da natureza autônoma das EOs e do fato de as mesmas serem construídas por diferentes pessoas, com diferentes objetivos em mente.

A falta de consistência na terminologia utilizada pelos diversos sistemas na *web* gera impactos, tais como, a dificuldade de interoperação entre sistemas e o retorno de respostas com baixa relevância para usuários que utilizam máquinas de busca. Neste trabalho, tratamos a heterogeneidade semântica sob o ponto de vista da interoperação entre EOs projetadas independentemente.

As EOs podem solucionar os problemas de interoperação, à medida que capacitam um sistema a oferecer equivalência entre "vocabulários" distintos. Por exemplo, para interoperar sistemas no domínio hospitalar que utilizem os termos `paciente` e `cliente`, pode ser utilizada uma EO que indique, para o sistema, que esses termos possuem o mesmo significado, associando o termo `paciente`, pela relação de sinonímia, ao termo `cliente`.

Michael Uschold [8] afirma que existem ao menos duas abordagens para melhorar a semântica nas EOs. A primeira é aumentar o grau de padronização, tanto das linguagens quanto do conteúdo das EOs. A segunda, aplicável aos casos onde a padronização não é possível, é o desenvolvimento de tecnologias para mapeamento e tradução entre duas ou mais EOs.

2.4 Interoperabilidade Semântica

Na apresentação deste tema é importante destacar que as diferenças entre interoperabilidade sintática e semântica nem sempre são claras, conforme afirmam Andreas Paepcke *et al.* [30]. Em termos gerais, sintaxe é o conjunto de regras ou padrões por meio dos quais as palavras são combinadas em sentenças, enquanto que semântica se refere ao significado/sentido dos termos - o modo como os termos relatam abstrações do mundo real.

Para esclarecer a diferença entre interoperabilidade sintática e semântica, Andreas Paepcke *et al.* [30] apresentam um exemplo simples, que passamos a relatar. Considere um componente descrevendo o fato de que qualquer pessoa pode chamar remotamente a função *Imprimir(String:autor, String:data, Float:preço, String:endereço)*. Assumindo uma invocação remota com a tecnologia apropriada, essa descrição produz interoperabilidade sintática. Qualquer pessoa pode chamar tal função sem provocar mensagens de erro. Para chegar à interoperabilidade semântica, este componente teria que tornar público, por exemplo, que a impressão seria em *720dpi* na impressora do laboratório *C* e que os parâmetros especificam um *livro* que deverá ser

[6] Pode-se encontrar na literatura a expressão "diferença ontológica" significando heterogeneidade semântica.

pago em *reais*, e que a saída impressa será na forma de um pedido, como requerido pelo procedimento padrão da empresa. Ilustrada no exemplo anterior, a interoperabilidade semântica é mais difícil de ser alcançada. Existem abordagens mais simplificadas de interoperabilidade que permitem a interoperabilidade sintática, tais como, protocolos (por exemplo, Z39.50) e formatos de dados (por exemplo, PDF, PS e XML).

A interoperabilidade é um problema complexo [31]. Esta complexidade deve-se, em parte, à falta de controle sobre o formato de armazenamento do conteúdo disponível na *web*. Alguns fatores que dificultam a interoperabilidade são apresentados por Moen [32]:

- Sistemas operacionais distintos, tais como, Linux e Windows.
- Múltiplos protocolos, em diferentes níveis de comunicação, como por exemplo TCP/IP, HTTP, Z39.50[7], etc.
- Múltiplos esquemas de metadados, tais como, Dublin Core e MARC.
- Múltiplos formatos de dados, entre eles HTML, PDF, XML, etc.
- Múltiplas linguagens e formatos de caracteres como, por exemplo, UNICODE e ANSI.
- Múltiplos vocabulários, ontologias e disciplinas definidos conforme a necessidade de cada domínio.

Jérôme Euzenat [33] define interoperabilidade semântica como a faculdade de interpretar o conhecimento importado de outras linguagens em nível semântico, ou seja, atribuir, a cada porção de conhecimento importado, sua interpretação correta.

Interoperabilidade semântica é a habilidade de um usuário acessar, consistentemente e coerentemente, classes similares de objetos e serviços digitais, distribuídos através de repositórios heterogêneos [34], com software de mediação que compense as variações "*site-a-site*" [35].

Através dos conceitos apresentados em [33, 34, 35], é possível observar que interoperabilidade semântica é um termo amplo, representando a necessidade de compartilhamento da interpretação semântica da informação nas mensagens transmitidas.

Andreas Paepcke *et al.* [30] afirmam que a interoperabilidade é um problema complexo e cuja solução está em desenvolvimento. Atualmente, diversos esforços têm sido feitos para promover a interoperabilidade semântica entre sistemas. No ano de 2001, no IJCAI[8], realizou-se um Workshop[9] sobre "Ontologias e Compartilhamento de Informação". Mais recentemente, no ECAI-2002[10], foi realizado um Workshop[11] tratando especificamente sobre "Ontologias e Interoperabilidade Semântica".

Em nosso entendimento, a interoperabilidade semântica pode ser alcançada a partir de uma aproximação consistente da similaridade lexical dos termos presentes no domínio de um sistema. Esses termos são ligados por relações semânticas, assunto da próxima seção.

[7] Protocolo de comunicação entre computadores projetado para suportar busca e recuperação de informações.
[8] *International Joint Conference on Artificial Intelligence*
[9] Artigos disponíveis em *http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-47/*
[10] *European Conference on Artificial Intelligence*
[11] Descrição do *Workshop* disponível em *http://www.afia.polytechnique.fr/* e artigos disponíveis em *http://SunSITE.Informatik.RWTH-Aachen.DE/Publications/CEUR-WS/Vol-64/*

2.5 Relações Semânticas

Esta seção apresenta uma breve descrição das principais relações semânticas utilizadas em EOs. São descritas as relações de polissemia, homonímia, hiperonímia, hiponímia, meronímia e holonímia.

"Em razão da existência da polissemia e da sinonímia, uma palavra pode denotar mais de um conceito e um conceito pode ser representado por palavras distintas" [36].

Segundo Bechara [37], sinonímia "é o fato de haver mais de uma palavra com semelhante significação, podendo uma estar em lugar da outra em determinado contexto, apesar dos diferentes matizes de sentido ou de carga estilística". Por exemplo, `casa`, `lar`, `morada`, `residência`, `mansão`.

Hiponímia é a relação entre dois lexemas, na qual um denota subclasse do outro [38]. Por exemplo, a relação entre `caipirinha` e `bebida alcoólica` é uma hiponímia, em que `caipirinha` é um hipônimo de `bebida alcoólica`, já que `caipirinha` "é um tipo de" `bebida alcoólica`. A relação hiperonímia é o inverso da hiponímia, onde um termo mais genérico é hiperônimo de outro mais específico. Neste caso, `bebida alcoólica` é hiperônimo de `caipirinha`.

Suponhamos que uma EO utilize o lexema `veículo` e este seja utilizado no mapeamento com outras EOs. O sistema de mapeamento deve considerar EOs que usam os hipônimos `veículo espacial` e `carro` e, ainda, considerar que esses dois últimos possuem uma relação de sinonímia. Sinonímia, hiponímia e hiperonímia, portanto, são relações semânticas que devem ser tratadas no processo de mapeamento entre EOs.

Muitas vezes, em EOs, é necessário representar relacionamentos em que um termo é uma parte de outro. Para isso são utilizadas as relações de meronímia e holonímia. Essas relações constituem outra importante forma de organização dos termos em uma EO e, também, apresentam os termos em uma estrutura hierárquica.

As relações de meronímia e holonímia ocorrem entre dois lexemas quando um denota parte e o outro denota todo. O lexema parte é um merônimo do lexema todo; por exemplo, o termo `teclado` é um merônimo de `computador` (`teclado` é parte do `computador`), enquanto que o lexema todo é dito holônimo do lexema parte. Nesse caso, `computador` é um holônimo de `teclado` (`computador` possui `teclado`).

2.5.1 Homonímia

De acordo com a definição de Ferreira [39], homonímia é uma "qualidade do que é homônimo", ou ainda "identidade fonética entre formas de significado e origem completamente distintos, como entre `são`, presente do verbo `ser` e, `são`, referente a `santo`. Na escrita, palavras que têm a mesma pronúncia, e igual grafia (como `falácia`, do que é `falaz`, `enganador`, e `falácia`, `falatório`) ou grafia diferente (como `lasso`, `cansado`, e `laço`, `laçada`)".

Os itens que fazem parte da relação semântica homonímia são chamados homônimos. Homônimo "que, ou palavra que, em relação a outra, tem a mesma pronúncia, ou pronúncia aproximada, mas escrita diferente. Alguns exemplos são: `acerto` (`ajuste`), `asserto` (`afirmação`) e `afear` (`tornar feio`), `afiar` (`dar fio a`)".

Para Ferreira [39], homônimo "diz-se de, ou palavra que se pronuncia da mesma forma que outra, mas cujo sentido e escrita são diferentes, ou que se pronuncia e escreve do mesmo modo, mas cujo significado é diverso".

Tradicionalmente, homonímia é definida como uma relação entre palavras que têm a mesma forma com significados não relacionados [38]. É possível observar que a conceituação de Jurafsky e Martin [38] está incompleta de acordo com Ferreira [39], que define palavras homônimas como

podendo ter uma grafia diferente e não somente como palavras com a mesma forma.

Pelos conceitos apresentados, conclui-se que não existe um consenso na direção de uma definição precisa de homônimo. Entretanto, uma discussão mais profunda sobre este conceito não faz parte do escopo do presente trabalho. Para este estudo, são considerados termos homônimos aqueles com a mesma grafia, mas com significados diferentes.

2.5.2 Polissemia

Polissemia é a relação que se caracteriza por uma só forma (significante) com mais de um significado unitário pertencente a campos semânticos diferentes. Ou, em outras palavras, a polissemia leva a um conjunto de significados, cada um unitário, relacionados com a mesma forma. Portanto, não se pode ver a polissemia como "significados imprecisos e indeterminados", porque cada um desses significados é preciso e determinado [37]. Por exemplo, a palavra `pregar` pode significar o ato de realizar um sermão, o ato de costurar uma bainha da roupa ou ainda o ato de fixar um prego; a palavra `manga` pode ter dois sentidos diferentes (no domínio do vestuário, referindo-se a uma parte de uma camisa ou jaqueta, por exemplo e no domínio de alimentação, fazendo referência à fruta manga).

Na Moderna Gramática Portuguesa, Bechara [37] comenta que os lingüistas divergem na conceituação de polissemia e homonímia e que nem sempre é possível distinguir a polissemia da homonímia.

Polissemia e homonímia, portanto, podem ter o efeito de reduzir a precisão de sistemas, tanto na recuperação de informação (RI) quanto no processo de integração ou mapeamento de informações. Na RI, podem levar um sistema a retornar documentos irrelevantes para a necessidade de informação do usuário. Por exemplo, em uma consulta pelo elemento químico `Mercúrio`, o sistema de busca não deve retornar ocorrências do primeiro planeta do sistema solar. Já na integração ou mapeamento, termos homônimos ou polissêmicos de EOs de domínios diferentes têm sido motivo de preocupação, uma vez que induzem ao mapeamento incorreto (conforme o exemplo anterior). Pois, no contexto de EOs, o elemento químico `Mercúrio` não possui semelhança ou similaridade com o planeta `Mercúrio`.

Polissemia e sinonímia também são causa da heterogeneidade semântica, tratada na seção 2.3. Para minimizar os problemas gerados pela heterogeneidade semântica encontram-se disponíveis na literatura diversas linguagens de marcação semântica, sendo as principais delas descritas na próxima seção.

2.6 Linguagens de Marcação Semântica

Diversas linguagens de marcação semântica, tais como, RDF[12], OIL[13], DAML-OIL[14] e OWL[15] têm sido utilizadas nos últimos anos. O objetivo dessas linguagens é prover uma definição explícita e não ambígua dos conceitos e relacionamentos descritos nas ontologias ou nas EOs. A seguir apresentaremos tais linguagens juntamente com exemplos de suas sintaxes, de modo a ressaltar que o mesmo conjunto de conceitos ou relações pode ser formalizado em diferentes sintaxes.

[12] *Resource Description Framework* - http://www.w3.org/RDF/

[13] *Ontology Interchange Language* ou *Ontology Inference Layer* - http://www.ontoknowledge.org/oil/

[14] *DARPA Agent Markup Language* - *Ontology Interchange Language* ou *Ontology Inference Layer* - http://www.w3.org/TR/daml+oil-reference

[15] *Ontology Web Language* - http://www.w3.org/TR/owl-ref/

2.6.1 *Resource Description Framework* - RDF

O RDF é um aplicativo XML recomendado pelo W3C (*World Wide Web Consortium*) que usa notação XML como sintaxe de codificação para descrição de metadados[16]. A especificação de RDF contém duas partes principais:

1. RDF propriamente dita: define como descrever recursos em termos de suas propriedades e valores [40];

2. RDF Schema: define propriedades específicas que podem ser utilizadas para definir esquemas [41].

De acordo com Staab *et al.* [42] a união dessas definições é tratada por RDF(S). Para Fensel [43] e Lassila e Swick [40], um dos objetivos do RDF é especificar (ou adicionar) semântica formal a dados baseados em XML, de forma interoperável e padronizada. Conforme Pitts-Moultis e Kirk [44], o principal objetivo do RDF é o de facilitar o intercâmbio de informações (que podem ser interpretadas por máquinas) entre aplicativos via *web*.

Em relação às ontologias, RDF produz duas contribuições importantes: uma sintaxe padronizada para escrever ontologias e um conjunto de modelagem de primitivas padrão na construção de ontologias, como os relacionamentos *instanceOf* e *subclassOf*.

RDF Schema produz um tipo básico de esquema para RDF. Objetos, classes e propriedades podem ser descritos. Além das propriedades pré-definidas *instanceOf* e *subclassOf* para modelar relacionamentos, restrições de domínio e restrições de variação de valores de atributos também podem ser adicionados.

O esquema no apêndice A pode declarar uma classe "BibDigPUCRS" que armazena obras publicadas por pessoas. Essas pessoas podem ser do tipo "autor", por exemplo. Este esquema é codificado em XML e validado em *http://www.w3.org/RDF/Validator/*. Este exemplo explicita a sintaxe RDF(S) como uma forma de declarar recursos na *web*.

O apêndice A ainda declara as classes *Criador*, *Obra* e *BibDig* e as subclasses *Autor* (subclasse de *Criador*) e *Livro* (subclasse de *Obra*). As relações *cria* e *escreve* são expressas como propriedades por meio da etiqueta *<rdf:Property>*.

Por razões de espaço é apresentada apenas uma instância de um autor *DanielGoleman* que escreve um livro (*Inteligência Emocional*, descrito pelo recurso *http://www.europa.com/danielg/livros/IntelEmoc.pdf*) e disponibiliza o mesmo em uma biblioteca digital *http://www.pucrs.br/BD*.

2.6.1.1 Limitações do RDF(S)

Heflin e Hendler [45] relatam três limitações do RDF(S):

1. RDF não possui qualquer mecanismo para definição de axiomas[17] genéricos. Por exemplo, alguém poderia especificar que a propriedade *subclasse* é transitiva ou que as propriedades *paiDe* e *filhoDe* são inversas uma da outra.

[16] Metadados são dados sobre dados, informações sobre informações. Em um catálogo bibliográfico, por exemplo, o metadado **Título** é um "dado sobre o dado" **Similaridade entre EOs**.

[17] Regras que permitem adicionar raciocínio. Outros benefícios do uso de axiomas são: um axioma pode prover informação adicional que não está explicitamente declarada e, talvez ainda mais importante para sistemas distribuídos como a *web*, axiomas podem ser usados para projetar diferentes representações dos mesmos conceitos.

2. A tendência da *web*, envolvendo rápidas mudanças, pode ser outro problema para RDF. Apesar do RDF prover um modo para revisar esquemas, este método pode ser insuficiente. Essencialmente, a cada nova versão de um esquema, lhe é atribuído um novo URI e, portanto, pode ser pensado como um esquema distinto da versão anterior. Entretanto, a versão é apenas um esquema que estende a versão original. Por fim, se os esquemas não têm uma versão oficial associada a eles, não existe um modo de localizar as revisões de um esquema, exceto se na manutenção do esquema for usado um nome de esquema consistente para os URIs.

3. Uma omissão significativa em RDF é a impossibilidade de renomear propriedades e classes para um vocabulário local. Apesar das propriedades *rdfs:subClassOf* e *rdfs:subPropertyOf* poderem ser utilizadas para declarar que um nome é uma especialização de outro, não existe um modo de declarar uma equivalência. Isto pode ser problemático à medida que dois esquemas separados renomeiem uma propriedade. Neste caso, a informação de que as propriedades *rdfs:subClassOf* e *rdfs:subPropertyOf* eram iguais, é perdida. Assim, é necessário que um esquema venha a ter habilidade para renomear propriedades e classes para um vocabulário local, uma vez que alcançar um consenso para nomes de esquemas, em uma escala mundial, será impossível.

Apesar dessas limitações, uma ontologia que utilize somente conceitos, relações e instâncias pode ser facilmente definida em RDF(S). Entretanto, a semântica declarável em RDF ainda permanece muito restrita e, assim, é necessária a utilização de uma camada "sobre" RDF que forneça melhor consistência à semântica. Ou seja, a criação de uma linguagem que use a sintaxe RDF, mas também adicione novas classes e propriedades que tenham uma semântica específica. Algumas linguagens que tenham essa propriedade de adicionar semântica ao padrão RDF são apresentadas a partir da seção 2.6.2.

Para melhor compreensão das diferenças entre as sintaxes das linguagens de marcação semântica que utilizam RDF, são mostrados exemplos de extratos de ontologia codificados em cada linguagem, começando pela figura 2.5.

```
<rdfs:Class rdf:about=‘‘PhDProfessor’’>
    <rdfs:subClassOf rdf:resource=‘‘Person’’/>
    <restrictedBy rdf:resource=‘‘&id49’’/>
<rdfs:Class>
<Restriction rdf:about=‘‘&id49’’>
    <toClass rdf:resource=‘‘University’’/>
    <onProperty rdf:resource=‘‘doctoralDegreeFrom’’/>
</Restriction>
```

Figura 2.5: Exemplo de extrato de ontologia utilizando a sintaxe RDF

A figura 2.5 define o conceito `PhDProfessor` (construtor `rdfs:Class`) como um subconceito (construtor `rdfs:subClassOf`) do conceito `Person`. O conceito `PhDProfessor` possui a propriedade `doctoralDegreeFrom` que é referenciada na declaração do conceito `PhDProfessor` através do identificador `&id49`[18]. Essa propriedade recebe valores das instâncias do conceito `University`.

Nas próximas seções apresentaremos esses mesmos conceitos, expressos em outras linguagens de marcação semântica.

[18]Esse identificador reflete a noção de ponteiro na linguagem de programação C.

2.6.2 *Ontology Inference Layer* - OIL

Conforme Fensel [43], existem ao menos dois significados para o acrônimo OIL - *Ontology Inference Layer* ou *Ontology Interchange Language*. OIL é uma proposta para uma camada de representação e inferência para ontologias, baseada na *web*, que combina primitivas de modelagem amplamente utilizadas pelas linguagens baseadas em *frames*[19] com semântica formal e serviços de raciocínio produzidos pelas lógicas de descrição [48]. OIL é baseado na noção de um conceito e na definição de suas superclasses e atributos.

De acordo com Horrocks *et al.* [49], as três principais origens do OIL são:

- Lógicas de Descrição (LDs);
- Sistemas baseados em *frames*;
- Padrões *web*.

Lógicas de descrição descrevem o conhecimento em termos de conceitos e restrições que são utilizados para, automaticamente, derivar classificação na forma de taxonomias. Uma característica das LDs é que classes (geralmente chamadas conceitos) podem ser definidas em termos de descrições que especificam as propriedades que os objetos devem satisfazer, para pertencer a um conceito.

Sistemas baseados em *frames* refletem a noção de conceito e a definição de suas superclasses e atributos. Este tipo de abordagem é bastante intuitivo para a maioria dos usuários, o que proporciona um melhor entendimento na modelagem de um sistema. A abordagem LDs, descrita no parágrafo anterior, pode ser vista como uma extensão e generalização da idéia de *frames*, com *frames* estando relacionados a conceitos na LD e propriedades referindo-se a papéis na LD.

A terceira raiz da OIL são as linguagens para *web*. A sintaxe utilizada para etiquetar é XOL (*XML-based Ontology-exchange Language*), uma linguagem para troca de ontologias que é baseada em XML. Outras candidatas a formalizar a sintaxe da OIL são RDF e RDF(S), já descritas na seção 2.6.1.

Um exemplo de ontologia utilizando a sintaxe OIL é apresentado na Figura 2.6.

```
<oil:Class rdf:ID=‘‘PhDProfessor’’>
    <rdfs:subClassOf rdf:ID=‘‘Person’’/>
</oil:Class>
<oil:ObjectProperty rdf:ID=‘‘doctoralDegreeFrom’’>
  <oil:domain rdf:ID=‘‘PhDProfessor’’/>
  <oil:range rdf:ID=‘‘University’’/>
</oil:ObjectProperty>
```

Figura 2.6: Exemplo de extrato de ontologia utilizando a sintaxe OIL

Na Figura 2.6 são apresentados os mesmos conceitos descritos na figura 2.5. Cabe destacar a declaração da propriedade `doctoralDegreeFrom` com o construtor `oil:ObjectProperty`, diferente do padrão RDF que utiliza o construtor `onProperty`.

[19] Uma estrutura de dados contendo um número específico de subdivisões [46]. Um grupo de fatos e objetos que descrevem alguma situação ou objetos típicos, junto com estratégias de inferência específicas para raciocínio sobre a situação [47].

2.6.2.1 Limitações de OIL

Uma das limitações de OIL refere-se à impossibilidade de expressar sinônimos de classes ou propriedades, bem como a impossibilidade de expressar o mapeamento de diferentes estruturas que representam o mesmo conceito.

Apenas um número limitado de propriedades algébricas pode ser expresso em OIL, além de não existir uma forma de definir relações compostas. Também não existem meios de renomear, reestruturar e redefinir ontologias importadas.

Quanto a propriedades algébricas adicionais - OIL contém somente as propriedades `inversa`, `transitiva` e `simétrica`. Finalmente, em OIL não existe suporte para a definição de domínios concretos (por exemplo, inteiros e pontos flutuantes, entre outros). Assim, de acordo com [50] OIL é mais adequada como uma representação RDF de lógica de descrição do que como uma base para a *web* semântica.

2.6.3 *DARPA Agent Markup Language OIL* - DAML+OIL

A linguagem DAML+OIL emergiu da iniciativa DAML (*DARPA Agent Markup Language*) e foi desenvolvida em conjunto com membros do consórcio europeu que desenvolveram a OIL. DAML+OIL surgiu em março de 2001, como uma linguagem de representação de ontologias, não ambígua e interpretável por computadores, sendo extensão para XML e RDF.

Um exemplo de extensão e, ao mesmo tempo, uma vantagem sobre RDF Schema, oferecida pela linguagem DAML+OIL, refere-se aos tipos de dados definidos em um esquema XML. O esquema RDF, por si só, não provê este tipo de suporte.

DAML+OIL possui uma abordagem orientada a objeto com a estrutura do domínio sendo descrita em termos de "classes" e "propriedades" [51]. As ontologias criadas consistem de um conjunto de axiomas que declaram relacionamentos (de subsunção[20], por exemplo) entre classes e propriedades.

São encontrados em [53, 54, 55] alguns exemplos de aplicações que utilizam DAML+OIL para definir uma ontologia que auxilia sistemas de recuperação de informação.

A figura 2.7 apresenta um exemplo de extrato de ontologia utilizando a sintaxe DAML-OIL.

```
<daml_oil:Class rdf:ID=‘‘PhDProfessor’’>
    <rdfs:subClassOf rdf:ID=‘‘Person’’/>
</daml_oil:Class>
<daml_oil:ObjectProperty rdf:ID=‘‘doctoralDegreeFrom’’>
  <daml_oil:domain rdf:ID=‘‘PhDProfessor’’/>
  <daml_oil:range rdf:ID=‘‘University’’/>
</daml_oil:ObjectProperty>
```

Figura 2.7: Exemplo de extrato de ontologia utilizando a sintaxe DAML-OIL

No extrato apresentado pode-se observar a utilização do construtor `daml_oil:ObjectProperty` para definir a propriedade `doctoralDegreeFrom`. As definições de domínio e de alcance desta propriedade são similares ao padrão RDF e à linguagem OIL.

[20] Colocação de uma idéia particular sob a dependência de uma idéia geral [52].

2.6.4 *Ontology Web Language* - OWL

Recentemente, em 18 de agosto de 2003, a linguagem de marcação semântica OWL tornou-se uma candidata a recomendação pelo W3C. OWL está desenvolvida como extensão do vocabulário de RDF e é derivada da linguagem DAML-OIL [56].

Pelo fato de OWL ser uma extensão do vocabulário de RDF, qualquer grafo RDF forma uma ontologia OWL. A figura 2.8 apresenta um extrato de uma das ontologias utilizada em nosso trabalho com a sintaxe OWL.

```
<owl:Class rdf:ID=‘‘PhDProfessor’’>
    <rdfs:subClassOf rdf:resource=‘‘Person’’/>
    <rdfs:subClassOf>
        <owl:Restriction>
            <owl:onProperty rdf:resource=‘‘doctoralDegreeFrom’’/>
            <owl:allValuesFrom rdf:resource=‘‘University’’ />
        </owl:Restriction>
    </rdfs:subClassOf>
</owl:Class>
```

Figura 2.8: Exemplo de extrato de ontologia utilizando a sintaxe OWL

Na figura 2.8 tem-se a declaração da propriedade `doctoralDegreeFrom` como uma restrição e uma subclasse do conceito `PhDProfessor`. Essa propriedade é declarada com a utilização dos construtores de OWL, a saber: `owl:Restriction`, `owl:onProperty` e `owl:allValuesFrom`. O construtor `owl:allValuesFrom` equivale ao `range` da linguagem OIL e do padrão RDF.

Em [56], Harmelen *et al.* apresentam uma lista com as onze principais alterações que a linguagem OWL propõe, em relação à linguagem DAML-OIL. Essas alterações não serão detalhadas neste trabalho pois transcendem o escopo de interesse do mesmo.

2.7 Considerações sobre o Capítulo no Contexto da Dissertação

Este capítulo apresentou o conceito de EO, bem como as abordagens de mapeamento entre termos pertencentes a EOs diferentes. O problema da heterogeneidade semântica, que ocorre com freqüência quando se deseja promover a interoperabilidade entre EOs, também foi descrito.

O capítulo ainda descreveu as principais relações semânticas entre termos consideradas. Foram mencionadas e, brevemente descritas, com exemplos, linguagens de marcação semântica amplamente utilizadas na literatura. Apresentamos os mesmos conceitos e a mesma propriedade ao longo da descrição das linguagens, com o objetivo de mostrar que o mesmo conhecimento de mundo pode ser implementado através de diferentes sintaxes.

Especificamente, as linguagens de marcação semântica são utilizadas, neste trabalho, no tratamento de EOs da língua inglesa, que constitui o enfoque inicial de nosso estudo, conforme descrito no capítulo 4.

O próximo capítulo apresenta trabalhos voltados ao mapeamento semi-automático de EOs. Todos os trabalhos relatados utilizam alguma das abordagens de mapeamento apresentadas no presente capítulo. Alguns deles também utilizam as linguagens de marcação semântica aqui descritas.

Capítulo 3

Trabalhos Correlatos

"Idéias são como pulgas, saltam de uns para outros, mas não mordem a todos".
George Bernard Shaw, escritor irlandês (1856-1950)

3.1 Preâmbulo

O referencial teórico apresentado no capítulo anterior permitiu-nos oferecer uma base comum de conceitos utilizados direta e indiretamente nesta dissertação, e na área de ontologias. Dessa forma, foi construído o embasamento teórico para entendimento dos trabalhos relacionados com nosso estudo.

Neste capítulo será oferecida uma descrição dos trabalhos correlatos que demonstram preocupações da mesma natureza do estudo aqui realizado. O capítulo está dividido em duas grandes seções: trabalhos que realizam mapeamento entre EOs por meio das abordagens de união, alinhamento e ontologias articuladas (apresentadas na figura 3.1), e trabalhos que realizam o mesmo mapeamento, mas focados em medidas de similaridade.

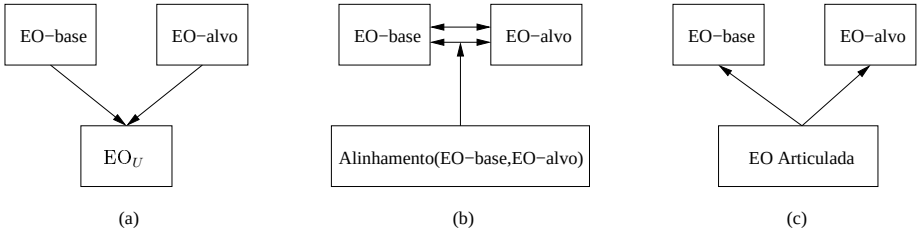

Figura 3.1: Ferramentas e algoritmos para mapeamento de ontologias (adaptado de [57, 58])

Na figura 3.1 as letras `a`, `b` e `c` representam as abordagens de mapeamento entre EOs. Essa figura permite ao leitor uma visão abrangente do capítulo e é retomada à medida que as ferramentas e algoritmos vão sendo descritos.

3.2 Trabalhos sobre Mapeamento entre Estruturas Ontológicas

3.2.1 Trabalhos que utilizam a Abordagem de União

Neste estudo, adotaremos a noção de união de EOs proposta por Noy e Musen [21]. União de EOs consiste na criação de uma EO coerente, que inclui informação de todas (pelo menos duas)

as EOs-base. Porém, na união entre EOs os termos que não combinam não são detectados nem tratados [22]. A abordagem de união de EOs é apresentada na figura 3.1a. A seguir discutiremos trabalhos que utilizam essa abordagem.

3.2.1.1 Chimaera

McGuinness *et al.* [59, 7] desenvolveram a ferramenta `Chimaera` para unir e também para detectar EOs diferentes. Essa ferramenta possui duas tarefas principais:

1. Unir dois termos semanticamente idênticos provenientes de EOs diferentes, de forma que eles sejam referenciados pelo mesmo nome na EO resultante. Na figura 3.1 a EO_U, que representa a união da EO-base com a EO-alvo, é composta pelos conceitos dessas duas últimas EOs.

2. Identificar termos que podem ser relacionados por subsunção e prover suporte para introduzir na EO esses relacionamentos. Neste caso, a ferramenta busca identificar termos que possam ser mapeados da EO-base para a EO-alvo como superclasse ou subclasse.

Para sugerir ao usuário termos que devam ser unidos entre as EOs, `Chimaera` também faz uso da similaridade lexical no nome das classes. Entretanto, não é apresentada em [59, 7] a medida de similaridade lexical utilizada na ferramenta. Conforme Doan *et al.* [6], `Chimaera` não lida com noções explícitas de similaridade. Ao contrário, são utilizadas diversas heurísticas para combinar elementos entre as EOs.

Nosso trabalho difere do trabalho de McGuinness *et al.* no sentido de que não realiza a união de EOs. Ao encontrar termos similares, os mesmos são apresentados ao usuário de forma que esse possa verificar a sobreposição entre as EOs. Contudo, em ambos os casos a similaridade lexical é considerada ao prover-se um mapeamento entre os termos das EOs.

3.2.1.2 FCA-Merge

Stumme e Maedche [9] desenvolveram um método para unir ontologias baseado em técnicas de PLN e análise formal de conceitos. O método trabalha com duas ontologias e um conjunto de documentos-instância (melhor explicados no texto que segue) em linguagem natural. A figura 3.2 apresenta o método `FCA-Merge`, acrônimo para *Formal Concept Analysis*.

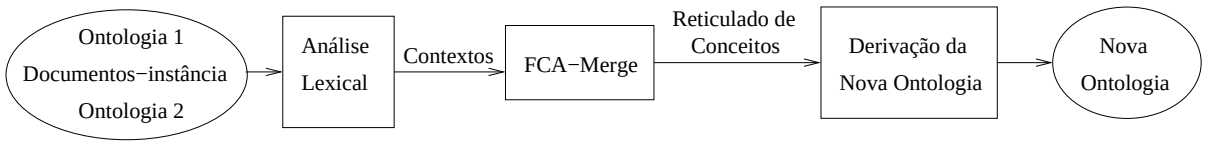

Figura 3.2: Método `FCA-Merge` (adaptado de [9])

Como pode ser observado na figura 3.2, o método `FCA-Merge` recebe como entrada duas ontologias e um conjunto de documentos-instância. Esses documentos passam por um processo de análise lexical que verifica quais termos, nos documentos, estão associados a quais conceitos, nas ontologias. Essa etapa gera dois contextos (um para cada ontologia), formados pelas instâncias dos conceitos que estão presentes nos documentos. Os contextos gerados servem de entrada para a geração de um reticulado de conceitos construído por meio da análise formal de conceitos. A seguir, os contextos são unidos e formam um único reticulado de conceitos para as duas

ontologias. De posse deste reticulado, para derivar a nova ontologia é necessária a intervenção humana.

`FCA-Merge` é um método que une duas ontologias e gera uma terceira, caracterizando a abordagem de união. Um requisito deste sistema é que as ontologias-base não somente estejam ligadas a instâncias de documentos mas também compartilhem essas instâncias. Dessa forma, se as ontologias do usuário não possuem instâncias ou as instâncias não cobrem todos os conceitos declarados nas ontologias-base, `FCA-Merge` não será útil. Isto significa que o algoritmo é extremamente dependente do conjunto de instâncias disponíveis. Caso um programa (um agente de software, por exemplo) acesse a ontologia disponível em um sistema, não será possível que esse programa detecte se um determinado conceito é similar a outro conceito de uma ontologia externa ao sistema, sem conhecer as instâncias da ontologia disponível.

O fato de `FCA-Merge` utilizar documentos-instância, e não medidas de similaridade para verificar a correspondência entre os conceitos das ontologias, caracteriza as principais diferenças do mesmo com relação ao nosso trabalho.

3.2.1.3 O Trabalho de Hakimpour e Geppert

Hakimpour e Geppert [22] usam uma abordagem para integrar esquemas[1] de diferentes comunidades, na qual cada comunidade usa sua própria EO. A abordagem utiliza a união de diferentes EOs baseada nas relações de similaridade entre conceitos. Os níveis de similaridade são identificados por meio de quatro definições, que apresentam as relações de similaridade com alguns exemplos ilustrativos:

1. Definições disjuntas: esse nível tem o menor grau de similaridade. Dois conceitos são disjuntos se a conjunção de suas definições intencionais[2] implica valor verdade falso. Por exemplo, `pai` e `filho`, `gordo` e `magro`, `pequeno` e `grande`.

2. Definições sobrepostas: se a conjunção de duas definições intencionais não pode ser provada como falsa, então as definições se sobrepõem. Por exemplo, `empregado` e `estudante`, `colega` e `irmão`, `diretor` e `cunhado`.

3. Definições especializadas: se a definição intencional de um conceito C_i é uma implicação da definição intencional de um conceito C_j, então C_i é uma especialização de C_j. Alguns exemplos são `pessoa` e `esposa`, `disciplina` e `disciplina da graduação`, `veículo` e `carro`.

4. Definições iguais: esse nível possui o mais alto grau de similaridade. Se duas definições intencionais são equivalentes, então os conceitos definidos são iguais. Casos típicos que caracterizam definições iguais são os conceitos que possuem relação de sinonímia. Por exemplo, `bruxaria` e `feitiçaria`, `idoso` e `velho`, `farmácia` e `drogaria`.

Após detectadas as relações de similaridade descritas, as mesmas são utilizadas para compor um esquema global que será a união resultante das EOs.

Nosso trabalho utiliza as definições 3 e 4 citadas anteriormente. As definições de especialização são levadas em consideração quando detectamos similaridade entre os conceitos em nível semântico-estrutural, caso em que a posição dos conceitos na hierarquia auxilia a identificação de similaridade. Já as definições iguais são detectadas em nível lexical: identificam-se termos iguais por meio de medidas de similaridade lexical.

[1] No trabalho [22], o termo esquema refere-se à noção de esquema de Banco de Dados.

[2] Definições intencionais são definições de termos por axiomas lógicos [22].

3.2.2 Trabalhos que utilizam a Abordagem de Alinhamento

Entende-se por alinhamento a busca por termos similares entre EOs, de modo a prover correspondências entre esses termos. Neste caso, não é criada uma nova EO que represente as EOs sendo alinhadas. Essa abordagem é apresentada na figura 3.1b.

3.2.2.1 Anchor-Prompt

Noy e Musen [2] desenvolveram o algoritmo `Anchor-Prompt`, que utiliza como entrada um conjunto de combinações-âncora[3] identificadas previamente (de modo automático ou manual). A figura 3.3 apresenta uma representação esquemática do algoritmo `Anchor-Prompt`.

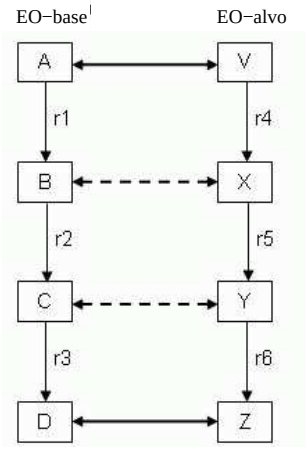

Figura 3.3: Os elementos do algoritmo `Anchor-Prompt` (adaptado de [4])

Na figura 3.3 os retângulos representam termos e os arcos rotulados representam relações semânticas entre os termos. A parte esquerda da figura representa os termos na EO-base e os termos no lado direito são os da EO-alvo. As setas contínuas conectam pares de termos-âncora, ao passo que as setas tracejadas indicam pares de termos relacionados.

Inicialmente, o algoritmo recebe como entrada os termos-âncora. Conhecido o comprimento do caminho entre esses termos, é atribuído um grau à similaridade entre os mesmos. Tomemos como exemplo as EOs na figura 3.3, nas quais os termos-âncora são `A` e `V` e `D` e `Z`. Na EO-base composta dos termos `A-B-C-D` o comprimento do caminho do nodo `A` até o nodo `D` é 3; na EO-alvo formada pelos termos `V-X-Y-Z` o comprimento do caminho entre `V` e `Z` também é 3. Neste caso, o grau de similaridade entre `B` e `X` e entre `C` e `Y` será mais elevado, pois esses termos estão nas mesmas posições relativas, no caminho que vai de `A` até `D`. Portanto, o resultado da similaridade entre os termos é cumulativo, ou seja, à medida que vão sendo encontrados termos similares no caminho, o grau de similaridade vai aumentando.

Apesar de prover mapeamentos consistentes, a abordagem baseada em âncoras possui uma forte limitação quando as ontologias são de profundidades muito diferentes, ou seja, quando uma ontologia é profunda (possuindo muitos níveis na hierarquia) e a outra ontologia é rasa (contendo poucos níveis na hierarquia). Neste caso, Noy e Musen [21] afirmam que o algoritmo não funciona bem. Para avaliar o algoritmo os autores elencaram ontologias provenientes da biblioteca do programa DAML, desenvolvidas independentemente uma da outra.

[3]Pares de termos relacionados.

Nosso trabalho não utiliza a abordagem de termos-âncora, pois não raras vezes temos encontrado ontologias com profundidades distintas e, conseqüentemente, optamos por utilizar uma abordagem que não esteja baseada somente na profundidade das hierarquias.

3.2.2.2 GLUE

Doan *et al.* [6] desenvolveram um sistema que emprega técnicas de aprendizagem de máquina para encontrar, semi-automaticamente, mapeamentos semânticos entre ontologias. Entretanto, antes de detectar tais mapeamentos, é utilizada uma medida de similaridade que mensura a proximidade dos termos nas ontologias sendo comparadas.

Como já pudemos mencionar, ontologias geralmente possuem instâncias associadas a conceitos em uma taxonomia, e cada conceito é visto como um conjunto de instâncias retiradas de um universo finito de instâncias. Neste caso, por exemplo, o conceito `empregado` é visto como o conjunto de todas as instâncias no universo do domínio que são `empregados`. Pela definição de taxonomia, as instâncias de um conceito são também instâncias de um conceito ancestral. Por exemplo, uma instância do conceito `professor` também é uma instância do conceito `pessoa`, pois todo professor é uma pessoa.

Doan *et al.* tratam um problema específico: dadas duas taxonomias e suas instâncias associadas, para cada nodo (um conceito, por exemplo) em uma taxonomia, deseja-se encontrar o nodo mais similar na outra taxonomia, por meio de uma medida de similaridade pré-definida.

A medida de similaridade utilizada para avaliar `GLUE` foi a medida de Jaccard [60]. No processo de avaliação foram criados mapeamentos 1-1 manuais para taxonomias do mesmo domínio. Os resultados apresentados nos experimentos relatados em [6] sugerem que `GLUE` pode trabalhar bem somente com uma quantidade modesta de dados (até 2000 instâncias em uma taxonomia).

Uma desvantagem da abordagem apresentada em [6] é o fato de que o mapeamento esteja baseado em um conjunto de instâncias. No momento em que as instâncias são alteradas, os resultados do mapeamento podem ser fortemente afetados.

Em nosso trabalho não fizemos uso de instâncias, contudo utilizamos medidas de similaridade para detectar conceitos similares entre EOs. Outras semelhanças entre o sistema `GLUE` e nosso trabalho são:

- a utilização de heurísticas próprias para melhorar o mapeamento. (no caso de `GLUE` o mapeamento é gerado com uma técnica de aprendizagem automática);
- a indicação de mapeamentos 1-1;
- o fato de, ao final do mapeamento, o sistema permitir ao usuário analisar as sugestões propostas, acrescentar mapeamentos não identificados automaticamente e alterar mapeamentos considerados incorretos.

3.2.2.3 HICAL

O sistema `HICAL` [61], acrônimo para *HIerarchical Concept ALignment system*, faz o alinhamento de instâncias baseado na similaridade entre as categorias[4] nas hierarquias. Para encontrar categorias similares, o algoritmo começa pelas categorias mais genéricas realizando uma abordagem *top to bottom*.

O algoritmo verifica a similaridade dos conceitos através de suas instâncias. Para determinar o grau de similaridade entre dois conceitos foi utilizado o método *k-statistic*. Esse método tem sido

[4]Entenda-se o termo categoria utilizado por Rytaro, Hideaki e Shinichi [61] como significando conceito em uma hierarquia.

utilizado para avaliar a similaridade entre dois critérios. Neste caso, os critérios são "pertencer" ou "não pertencer" a uma categoria.

O método *k-statistic* será explicado brevemente: vamos supor dois critérios de categorização C_1 e C_2. Pode-se decidir se uma instância particular pertence a uma categoria ou não. Conseqüentemente, as instâncias são divididas em quatro classes conforme a tabela 3.1.

Tabela 3.1: Classificação de instâncias em duas categorias

		Categoria C_1	
		pertence	não pertence
Categoria C_2	pertence	N_{11}	N_{12}
	não pertence	N_{21}	N_{22}

Os símbolos N_{11}, N_{12}, N_{21} e N_{22} denotam o número de instâncias satisfazendo as condições da tabela, para cada classe da ontologia. Por exemplo, N_{11} denota o número de instâncias que pertencem à categoria C_1 e à categoria C_2. Pode-se notar que, se as categorias C_1 e C_2 têm o mesmo critério de categorização, então N_{12}, N_{21} tornam-se próximos a zero, pois as mesmas instâncias provavelmente pertencem à categoria C_1 e à categoria C_2. Caso contrário, se as instâncias não pertencem à categoria C_1, então provavelmente não pertencerão à categoria C_2, pois as instâncias passaram pelo mesmo critério de classificação. Entretanto, se ambas as categorias possuem um critério diferente de categorização, então N_{11}, N_{22} tornam-se próximos a zero. O método *k-statistic* utiliza esse princípio para determinar a similaridade do critério de classificação.

Em `HICAL`, se uma categoria em uma hierarquia `A` conta com 50 instâncias, e outra categoria em uma hierarquia `B` recebe as mesmas 50 instâncias, então o algoritmo gera uma regra de alinhamento entre essas categorias.

Nosso trabalho difere de [61] pelo fato de que não utilizamos instâncias das ontologias e, conseqüentemente, não podemos utilizar o método *k-statistic*, que está baseado na comparação do número de instâncias de cada classe. Além disso, no sistema `HICAL` não existe intervenção de especialista humano para refinar o alinhamento automático. Dada a importância desta atividade, em nosso trabalho o usuário pode realizar alterações nos mapeamentos encontrados de modo automático pelo algoritmo ou, ainda, acrescentar mapeamentos que não tenham sido encontrados pelo algoritmo.

3.2.2.4 Heurística e Probabilidade

Prasad, Peng e Finin [12] desenvolveram um sistema para mapear EOs que apresenta as EOs ao usuário, e esse último indica os pontos (conceitos) equivalentes para mapeamento. Esses pontos são funcionalmente similares aos termos-âncora em [2].

Após a intervenção inicial do usuário, o mapeamento automático pode ser realizado por meio de duas abordagens: heurística ou probabilística. A abordagem heurística considera a porcentagem de filhos de um conceito, em uma hierarquia `A`, que podem ser mapeados para outro conceito, em uma hierarquia `B`. Por exemplo, se o conceito C_A possui dez filhos e seis (ou, 60%) desses filhos são similares[5] a um conceito C_B, pode-se concluir que C_A é similar a C_B. Já a abordagem de probabilidade é baseada nos resultados obtidos por um classificador[6] de documentos, que permite a identificação da similaridade entre conceitos. Essa identificação se

[5]Prasad, Peng e Finin [12] operam sobre uma taxa mínima de similaridade de 60%.

[6]Rainbow - http://www-2.cs.cmu.edu/~mccallum/bow/rainbow

dá através de cada conceito de uma EO-base que é mapeado para um ou mais conceitos de outra EO-alvo pela comparação dos documentos da EO-base com os documentos da EO-alvo.

Em [12] foram utilizadas somente EOs no formato DAML+OIL e não foram consideradas as propriedades relacionadas a cada conceito. O algoritmo de probabilidade, que apresentou, segundo os autores, melhores resultados, é fortemente dependente do conjunto de documentos disponíveis.

Nosso trabalho não utiliza documentos associados a conceitos. Faz uso de heurísticas próprias para realizar o mapeamento entre conceitos similares. A interface do usuário de nosso protótipo é bastante próxima à interface apresentada em [12], uma vez que ambas mostram as EOs na forma de uma hierarquia e permitem ao usuário realizar alterações nos mapeamentos gerados automaticamente. Em [12] o usuário pode mapear conceitos utilizando a opção heurística ou probabilística, ao passo que, no protótipo por nós desenvolvido, o usuário pode escolher realizar o mapeamento por meio de duas medidas diferentes de similaridade, ou utilizando as heurísticas oferecidas ou utilizando as duas alternativas, heurística e medidas de similaridade.

3.2.2.5 MAFRA

Maedche *et al.* [62] desenvolveram uma estrutura a qual nomearam `MAFRA` (*Mapping Framework for Distributed Ontologies*) para mapear EOs distribuídas na *web*. Essa estrutura é formada por cinco módulos horizontais e quatro módulos verticais, conforme apresentado na figura 3.4.

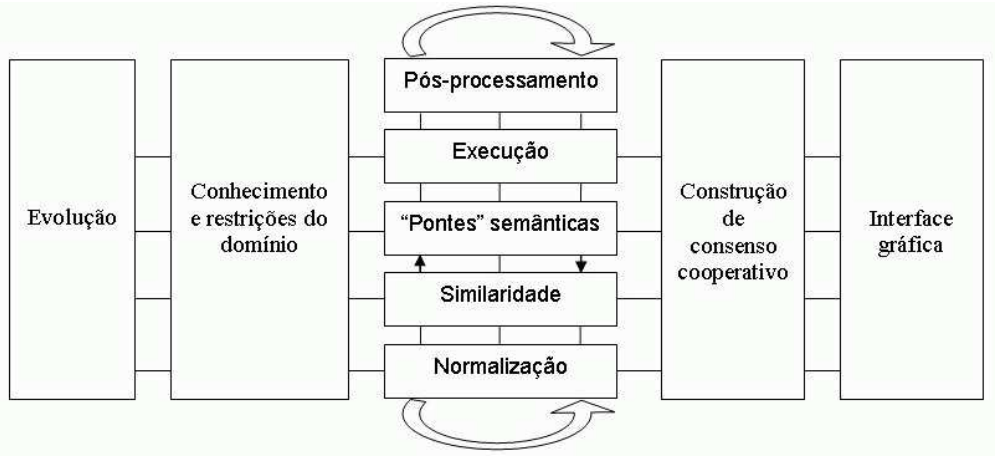

Figura 3.4: Arquitetura conceitual do `MAFRA` (adaptada de [62])

Os módulos horizontais são:

1. `Normalização`: lida com a heterogeneidade sintática, estrutural e de linguagem. Nesse caso, linguagem se refere às diferentes linguagens de marcação semântica, tais como, RDFS, DAML e OIL, entre outras.

2. `Similaridade`: trata a similaridade entre as entidades de uma EO-base e de uma EO-alvo.

3. `Ligações semânticas`[7]: estabelecem correspondências (ligações semânticas) entre os conceitos da EO-base e da EO-alvo. O papel das ligações semânticas é encapsular toda a informação necessária para transformar instâncias da EO-base em instâncias da EO-alvo.

[7] Do inglês, *semantic bridging*.

As ligações semânticas descritas em [62] são as mesmas apresentadas em [26], onde os autores utilizam regras de articulação[8] para capturar essas pontes e possibilitar a interoperação entre as EOs.

4. `Execução`: transforma instâncias da EO-base em instâncias na EO-alvo, avaliando as ligações semânticas entre as entidades da EO-base e da EO-alvo.

5. `Pós-processamento`: a partir dos resultados gerados pelo módulo `execução`, o módulo de `Pós-processamento` é responsável por analisar e melhorar a qualidade dos resultados. Os autores não descrevem como é realizado esse procedimento. O principal desafio deste módulo é reconhecer que duas instâncias representam o mesmo objeto do mundo real.

Quanto aos módulos verticais:

1. `Evolução`: mantém as ligações semânticas obtidas no módulo `ligações semânticas` à medida que vão ocorrendo mudanças na EO-base e na EO-alvo. Esse componente reutilizará as ligações semânticas existentes, adaptando-as aos novos requisitos do sistema enquanto esse evolui.

2. `Construção de consenso cooperativo`: responsável por estabelecer o consenso das ligações semânticas entre duas comunidades, durante o processo de mapeamento. Isto é um requisito quando se tem muitas alternativas de mapeamento. Este módulo busca reduzir a quantidade de pessoas envolvidas na tarefa de construir consenso entre as ligações semânticas.

3. `Conhecimento e restrições do domínio`: responsável por melhorar a qualidade das similaridades e das ligações semânticas detectadas, por introduzir o conhecimento e as restrições do domínio. Realiza a identificação de sinônimos e conceitos similares com auxílio de tesauros, por exemplo.

4. `Interface gráfica do usuário`: permite ao usuário atuar como condutor do processo de mapeamento, criando ligações semânticas, refinando-as, etc.

Na arquitetura apresentada na figura 3.4 os módulos horizontais descrevem as fases do processo de mapeamento. Os componentes verticais são utilizados durante todo o processo de mapeamento, interagindo com os módulos horizontais. Essa interação é representada através das duas setas em forma de arco.

`MAFRA` ainda não sofreu nenhum processo de avaliação, pois está em fase de desenvolvimento. A relação com nosso trabalho é estabelecida no módulo de `Similaridade`, no qual são tratadas as similaridades entre os termos nas EOs. O último módulo vertical, `Interface gráfica`, também é implementado em nosso trabalho.

3.2.2.6 OBSERVER

`OBSERVER`, acrônimo para *Ontology Based System Enhanced with Relationships for Vocabulary hEterogeneity Resolution* [63, 64], trata diferentes ontologias de domínio (por exemplo, ontologias sobre universidades, medicina, jornalismo, etc.) que são utilizadas para descrever informação similar através de domínios distintos. O uso de ontologias visa evitar que os usuários tenham que lidar com múltiplos repositórios de dados heterogêneos. A idéia é permitir que os usuários formulem consultas sobre as ontologias e o sistema gerencie a heterogeneidade e a distribuição

[8] Regras que expressam o relacionamento entre conceitos de EOs diferentes.

nos repositórios. Assim, o principal objetivo do sistema é proporcionar interoperabilidade através das ontologias de domínio. Para isso tais ontologias são expressas em lógicas de descrição.

O conjunto de ontologias utilizadas no OBSERVER foi construído sob três pontos de vista distintos: pesquisa em lingüística, pesquisa em representação do conhecimento e o ponto de vista individual de alguns grupos de pesquisa. Este aspecto caracteriza também uma das preocupações de nosso trabalho, que é investigar a similaridade entre termos de EOs projetadas por pessoas com diferentes pressupostos e conhecimento de mundo variado.

O módulo do OBSERVER relacionado ao nosso trabalho é o Gerenciador de Relacionamentos Interontologias (GRI), apresentado na figura 3.5.

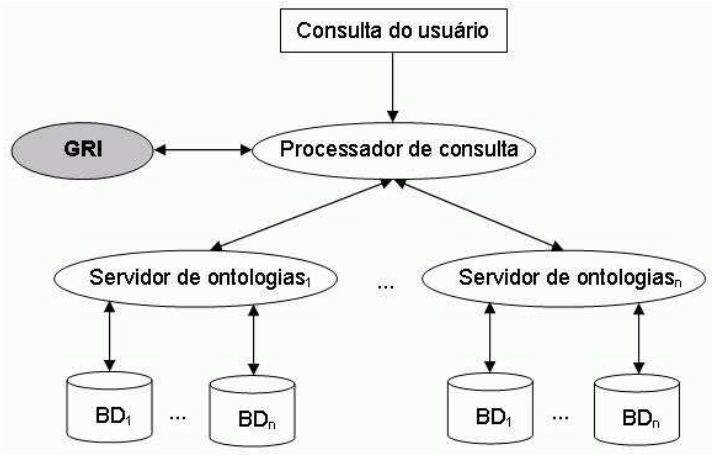

Figura 3.5: Arquitetura do OBSERVER (adaptado de [63])

A figura 3.5 apresenta um processo que inicia por uma consulta do usuário, que é processada pelo processador de consultas. Esse processador acessa os servidores de ontologias, que estão ligados a uma ou mais bases de dados. Após acessar os servidores de ontologias, o sistema ativa o módulo GRI. Esse módulo provê relacionamentos interontologias permitindo ao sistema expandir a consulta e retornar melhores resultados para o usuário.

O módulo GRI busca evitar:

- a projeção de uma ontologia global comum contendo todos os termos relevantes em um sistema de informação global;
- o investimento de tempo e energia no desenvolvimento de uma ontologia específica quando ontologias similares estão disponíveis.

O objetivo desse módulo é prover relacionamentos interontologias que são utilizados para:

- traduzir consultas do usuário de uma ontologia para outra;
- suportar o processamento de consulta que acessaria dados descritos por múltiplas ontologias.

No OBSERVER existe uma tarefa importante que é realizada manualmente. Para cada termo em uma nova ontologia, o administrador do sistema deve estabelecer quais tipos de relacionamentos semânticos existem (caso efetivamente existam) com as ontologias pré-estabelecidas. Nosso trabalho busca investigar as alternativas para automatizar esse tipo de tarefa, uma vez que, no

contexto da *web* e com ontologias maiores (acima de 100 nodos), tal tarefa não seria escalável: teria um alto custo de mão-de-obra de um especialista, e seria bastante tediosa.

Em nosso trabalho, por meio de medidas de similaridade, buscamos localizar termos que sejam lexicalmente similares entre as ontologias de domínio disponíveis.

3.2.3 A Abordagem EO Articulada

Entende-se por abordagem EO articulada, neste trabalho, um conjunto de regras de articulação entre termos de duas EOs-base. Essas regras de articulação indicam quais termos, nas EOs-base, estão relacionados. A seguir é apresentado um trabalho voltado para tal abordagem, que está representada na figura 3.1c.

3.2.3.1 ONION

Esta subseção reúne aspectos considerados relevantes, provenientes de dois artigos [4, 26] de Mitra e Wiederhold. Em [4] é descrito o sistema `ONION`, acrônimo para *ONtology compositION* e em [26] é apresentado o algoritmo que constituiu a base para prover interoperação entre ontologias do sistema `ONION`.

A abordagem de união de EOs, que consiste de uma EO representando todas as EOs-base, não tem sido considerada uma solução escalável, principalmente no contexto da *web*, pela dificuldade de centralizar o conhecimento distribuído em várias (centenas de) EOs em uma só EO. Mitra e Wiederhold adotaram uma abordagem distribuída que permite às fontes de informação serem atualizadas e mantidas de forma independente.

`ONION` é um modelo orientado a grafo que realiza mapeamento por meio de articulação entre ontologias. Uma regra de articulação indica quais termos, individualmente ou em conjunto, estão relacionados nas ontologias-base. O objetivo principal da criação de regras de articulação é facilitar a manutenção de um sistema e proporcionar um certo grau de escalabilidade. A ontologia articulada é composta pelos termos e relacionamentos contidos nas regras de articulação. Em `ONION` as regras de articulação são geradas semi-automaticamente, necessitando a intervenção de um especialista humano. Essas regras de articulação podem ainda ser utilizadas na formulação de consultas.

Mitra e Wiederhold utilizam o termo "ontologia unificada", mas essa ontologia não é uma entidade física. Em [4] as ontologias-base são mantidas independentes, e a articulação é que é fisicamente armazenada. A figura 3.6 apresenta uma arquitetura simplificada de `ONION`.

Na figura 3.6 tem-se uma máquina de busca que atua sobre uma ontologia unificada. Essa ontologia é representada pelas ontologias-base $\texttt{Ontologia}_1$ e $\texttt{Ontologia}_2$ e pela $\texttt{Ontologia Articulada}_1$. As bases de conhecimento podem estar ligadas a uma ou mais ontologias.

Para prover interoperação entre essas fontes, Mitra e Wiederhold [26] implementaram dois métodos para a combinação de termos usados em ontologias diferentes. Um dos métodos é baseado em corpus e o outro é baseado em tesauros. Esses métodos têm seu foco na similaridade lingüística dos termos utilizados nas ontologias.

Em [26] são utilizadas regras de articulação, as quais expressam o relacionamento entre dois ou mais conceitos pertencentes às ontologias que estão sendo interoperadas, e que indicam quais termos, individualmente ou em conjunto, estão relacionados. Os relacionamentos considerados em [26] são `SubClasseDe`, `ParteDe`, `AtributoDe`, `InstânciaDe`, `ValorDe`. A relação semântica `SubClasseDe` é a mesma relação de hiponímia, já a relação *ParteDe* equivale à relação de meronímia. Hiponímia e Meronímia foram apresentadas na seção 2.5.

As regras de articulação não indicam o relacionamento exato entre dois conceitos. A intervenção de um especialista humano é necessária para refinar, validar as combinações geradas

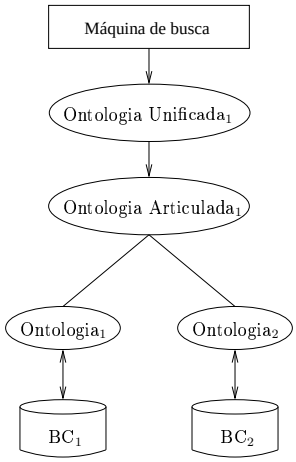

Figura 3.6: Arquitetura simplificada do sistema `ONION` (adaptado de [4])

automaticamente e gerar as regras que não foram detectadas. Ou seja, as regras de articulação são estabelecidas semi-automaticamente.

O cálculo de similaridade entre os termos é realizado com base em uma tabela de similaridade de palavras. Tabelas são geradas a partir de um tesauro e de um corpus. Os resultados e experimentos descritos em [26] mostram que o método baseado em corpus produziu melhores resultados do que o método baseado em tesauro. Isto pode ser atribuído ao fato de o corpus utilizado ser retirado de um *site*[9] de busca. Quanto ao modelo proposto em [4], nenhum processo de avaliação foi empregado para validá-lo. Um fato observado em [4] é a utilização dos seguintes relacionamentos semânticos: `SubClasseDe`, `AtributoDe`, `InstânciaDe` e `ImplicaçãoSemântica`. Já na seqüência de suas pesquisas em [26], nenhuma justificativa é dada pelos autores para a eliminação do relacionamento `ImplicaçãoSemântica` e para a adição dos relacionamentos `ParteDe` e `ValorDe`.

Nosso trabalho difere de [26], uma vez que não faz uso de corpus. A similaridade entre os termos das EOs é detectada por meio de medidas de similaridade lexical e heurísticas em nível semântico-estrutural.

3.3 Trabalhos sobre Mapeamento entre EOs focados em Medidas de Similaridade

Antes de iniciar esta seção faz-se necessário esclarecer a distinção entre medidas de similaridade e medidas de distância entre cadeias de caracteres. Cohen, Ravikumar e Fienberg [65] mencionam que a comparação de cadeias de caracteres se dá por meio de funções de distância e de similaridade. Funções de distância mapeiam duas cadeias de caracteres `f` e `d` para um número real `r`, onde quanto menor o valor de `r`, maior a similaridade entre `f` e `d`. As funções de similaridade apresentam um comportamento análogo, exceto que valores maiores indicam maior similaridade.

Dekang Lin [66] apresenta três "intuições" sobre similaridade, a saber:

- Intuição 1: A similaridade entre `A` e `B` está relacionada à associação entre essas cadeias. Quanto mais associados `A` e `B` estão, maior a similaridade entre elas.

[9] www.google.com

- Intuição 2: A similaridade entre `A` e `B` está relacionada às diferenças entre essas cadeias. Quanto mais diferenças existirem entre `A` e `B`, menor a similaridade entre elas.

- Intuição 3: A máxima similaridade entre `A` e `B` é alcançada quando A e B são idênticas.

Medidas de similaridade são utilizadas em aplicações, tais como, desambigüação (*word sense disambiguation*), sumarização e anotação de texto, extração e recuperação de informação, indexação automática, seleção lexical e correção automática de erros [67]. Diversas medidas de similaridade têm sido utilizadas na literatura, cada uma aplicada a uma situação específica.

As medidas de similaridade semântica de Resnik [68], Lin [66] e Jiang [69], por exemplo, são baseadas no conteúdo de informação de cada termo. Esse conteúdo é definido como o número de vezes que um termo, ou qualquer termo filho em uma mesma hierarquia, ocorre em um corpus. Atualmente, existe uma preocupação em relação às medidas de similaridade semântica no sentido de que as mesmas devam corresponder à noção intuitiva humana de similaridade [6, 70]. Especificamente, essas medidas devem depender somente do conteúdo semântico dos conceitos envolvidos na combinação, e não de sua especificação sintática.

Quando comparando duas hierarquias, a aproximação taxonômica ótima é aquela que busca a máxima sobreposição entre as hierarquias em questão. Neste caso, entende-se por sobreposição a intersecção composta pelos termos pertencentes às duas hierarquias.

A seguir serão apresentados trabalhos que fazem uso de medidas de similaridade aplicadas à verificação da similaridade entre conceitos pertencentes a hierarquias diferentes.

3.3.1 O Trabalho de Rodríguez e Egenhofer

Rodríguez e Egenhofer [70] desenvolveram um modelo para avaliar a similaridade semântica através de três verificações diferentes de similaridade. O modelo leva em consideração conjuntos de termos sinônimos e características distintivas[10] (partes, funções e atributos de um termo, por exemplo) dos termos dos seus vizinhos semânticos.

Em [70], conjuntos de sinônimos são entendidos como grupos de palavras semanticamente equivalentes ou muito similares. Nos conjuntos de sinônimos são considerados o número de palavras comuns e diferentes. Palavras comuns são consideradas palavras com alta similaridade lexical. Uma função toma como entrada essas palavras, e devolve um valor de similaridade, de acordo com a posição do termo na hierarquia de conceitos. O objetivo de comparar conjuntos de sinônimos é explorar a concordância no uso de palavras nas EOs e detectar palavras equivalentes através dessas EOs.

As características distintas dos conceitos são consideradas na fase de combinação das características. Essa fase tem por objetivo verificar a similaridade entre as partes, as funções e os atributos dos conceitos.

Por último, mas não menos importante, é verificada a similaridade entre os vizinhos semânticos de um termo. Vizinhos semânticos são conceitos que possuem relacionamentos semânticos de hiponímia e meronímia ligados a um termo `X` com o qual se deseja verificar a similaridade. Neste modelo a quantidade de relacionamentos a ser levada em consideração é delimitada por um raio pré-estabelecido, e somente os termos que estão dentro desse raio são detectados como vizinhos do termo `X`. Ou seja, o conjunto de termos vizinhos do termo `X` é formado pelos elementos (conceitos) que estão dentro do raio pré-estabelecido. Este modelo de similaridade é útil como uma primeira etapa no processo de união de EOs, pois ele pode detectar conceitos similares entre ontologias.

[10] Do inglês, *distinguishing features*.

Durante o processo de avaliação de similaridade, Rodríguez e Egenhofer [70] constataram que *recall* e precisão diminuíram drasticamente para os casos nos quais a combinação entre palavras é ignorada. Essa constatação de Rodríguez e Egenhofer é um indício de que a utilização da combinação de caracteres entre palavras, para detectar similaridade entre termos de EOs, deve ser levada em consideração.

Assim como em [70], nosso trabalho também utiliza medidas de similaridade entre cadeias de caracteres. Em nível semântico-estrutural Rodríguez e Egenhofer verificam a posição, na hierarquia, do conceito sendo comparado. Essa abordagem faz com que hierarquias com profundidades distintas possam ser integradas com mais facilidade. Em nosso trabalho, o aspecto semântico-estrutural é tratado com heurísticas e com a relação de sinonímia. Em [70] é utilizada a WordNet da língua inglesa. Em nosso trabalho, não utilizamos tesauros ou bases de dados lexicais.

3.3.2 O Trabalho de Maedche e Staab

Maedche e Staab [71, 1] não encontraram trabalhos que se destinam a comparar taxonomias, apenas abordagens que tratam da comparação entre dois conceitos em uma taxonomia comum. Diante disso, propuseram uma abordagem em duas camadas (lexical e conceitual) para medir a similaridade entre EOs.

Em nível lexical, os autores basearam-se na Distância de Edição (DE) de Levenshtein [72], que é implementada utilizando programação dinâmica. Essa medida considera as mudanças que devem ocorrer para transformar uma cadeia de caracteres `f` em outra cadeia `d`. O valor da distância é o custo da melhor seqüência de operações de edição para converter `f` para `d`. Operações de edição típicas são a inserção, deleção ou substituição de caracteres. A `DE` considera o custo igual a 1 para qualquer uma dessas três operações. Por exemplo, a `DE(automóvel, automóveis)` é 2, pois uma operação de alteração e uma operação de inserção mudam a cadeia de caracteres `automóvel` para `automóveis`. A contribuição de Maedche e Staab consiste na medida de Combinação de Caracteres (CC), dada pela equação 3.1.

$$CC(T_i, T_j) := max\left(0, \frac{min(\mid T_i \mid, \mid T_j \mid) \; - \; DE(T_i, T_j)}{min(\mid T_i \mid, \mid T_j \mid)}\right) \; \in \; [0, 1]. \tag{3.1}$$

A medida CC calcula a similaridade entre dois termos T_i e T_j. O comprimento do menor termo é representado por $min(\mid T_i \mid, \mid T_j \mid)$. Por exemplo, ao processar a similaridade entre os termos (`automóvel`, `automóveis`) o menor comprimento é 9 e o valor da $\texttt{DE}(T_i, T_j)$ é 2. Logo, tem-se os valores apresentados abaixo para $CC(\texttt{automóvel}, \texttt{automóveis})$:

$$CC(\texttt{automóvel}, \texttt{automóveis}) := max\left(0, \frac{9 \; - \; 2}{9}\right) = \frac{7}{9} \; \in \; [0, 1].$$

O menor comprimento é considerado tanto no numerador quanto no denominador da fórmula, o que permite ponderar o número de alterações obtidas com a `DE` em relação ao termo de menor comprimento. No exemplo, tem-se como resultado o valor 0,778 que corresponde à similaridade entre os termos (`automóvel`, `automóveis`). A medida CC sempre retorna um grau de similaridade entre 0 e 1, onde 1 indica uma combinação perfeita e zero indica ausência de similaridade.

Em nível semântico-estrutural, Maedche e Staab comparam as estruturas semânticas das EOs com o auxílio de um léxico. Neste nível a comparação é realizada entre as taxonomias sendo considerada a hierarquia dos conceitos. Maedche e Staab utilizam uma abordagem, a qual denominaram *semantic cotopy*, que leva em consideração a posição do termo na hierarquia. Para cada termo sendo tratado na hierarquia, são considerados todos os termos ascendentes e descendentes na mesma hierarquia.

Em [71, 1] são utilizadas EOs da língua alemã pertencentes ao domínio do turismo. Nosso trabalho não faz uso de léxico, mas utiliza a medida de similaridade CC aplicando essa medida a EOs das línguas inglesa e portuguesa.

3.4 Considerações sobre o Capítulo no Contexto da Dissertação

Neste capítulo foram apresentados trabalhos voltados ao mapeamento entre termos pertencentes a EOs projetadas independentemente. Como pôde ser observado, existem trabalhos que fazem uso de medidas de similaridade e trabalhos que utilizam outras alternativas de solução para mapear termos similares entre EOs.

A ferramenta `Chimaera`, o método `FCA-Merge` e o trabalho de Hackimpour e Geppert fazem o mapeamento de EOs por meio da abordagem de união, na qual as EOs-base geram uma única EO. `Anchor-PROMPT`, `GLUE`, `FCA-Merge`, o sistema que utiliza heurística e probabilidade, `HICAL`, `OBSERVER` e `MAFRA` realizam o mapeamento entre termos de EOs através de alinhamentos entre os termos. O método `FCA-Merge` está enquadrado tanto no caso de união quanto no caso de alinhamento de EOs pois, além da geração de uma EO a partir das EOs-base, esse sistema passa por uma etapa que apresenta as correspondências entre os termos nas EOs. Por fim, no sistema `ONION` a relação entre os termos das diferentes EOs é representada por regras de articulação.

As ferramentas `Chimaera`, `MAFRA` e `ONION` possuem interface gráfica que permite ao usuário comparar as EOs-base visualmente e, então, aceitar ou rejeitar os resultados do mapeamento automático gerado pela ferramenta.

Entre os trabalhos que têm seu foco na similaridade destacamos [1], do qual adotamos a medida CC aplicando-a a nossos experimentos nas EOs das línguas inglesa e portuguesa.

Nosso trabalho possui características distintas e complementares àqueles descritos neste capítulo. Procurou-se apresentar a intersecção (quando existia) desses trabalhos explicitando o que deles utilizamos e o que não utilizamos nesta dissertação.

O próximo capítulo descreve um primeiro experimento por nós realizado com EOs da língua inglesa, no qual fizemos uso de algumas das idéias apresentadas nos trabalhos correlatos.

Capítulo 4

Enfoque Inicial do Estudo: Tratamento de EOs da Língua Inglesa

"Nos campos da observação, o acaso favorece apenas as mentes preparadas".
Louis Pasteur, cientista (1822-1895)

4.1 Preâmbulo

Os trabalhos correlatos estudados permitiram-nos idealizar um experimento com EOs da língua inglesa, aplicando a medida de similaridade CC para mapear termos entre EOs.

Este capítulo aborda o experimento realizado com EOs da língua inglesa e descreve as heurísticas propostas e os resultados obtidos na realização do experimento. Além disso, desenvolvemos um protótipo que auxilia o usuário durante o processo de mapeamento entre EOs. Esse protótipo abstrai as diferentes sintaxes provenientes das linguagens de marcação semântica apresentando as EOs na forma de uma hierarquia de conceitos, tal como o protótipo descrito em [12].

Nosso trabalho, assim como em [2], utiliza EOs da biblioteca do programa DAML. O algoritmo desenvolvido trata os formatos OWL, RDF, OIL e DAML+OIL. Para o engenheiro do conhecimento, quando mapeando duas EOs, as diferentes sintaxes utilizadas pelos padrões OWL, RDF, OIL e DAML+OIL são um fator que dificulta a comparação entre as EOs. Neste sentido, nosso algoritmo abstrai para o usuário a sintaxe, apresentando a EO através de uma interface na forma de uma hierarquia com níveis e subníveis, tal como mostrado na figura 4.1. As EOs apresentadas nessa figura modelam os domínios das comunidades de pesquisa em *Web* Semântica[1] e dos departamentos acadêmicos de universidades[2], respectivamente.

[1] http://www.daml.org/ontologies/4
[2] http://www.daml.org/ontologies/64

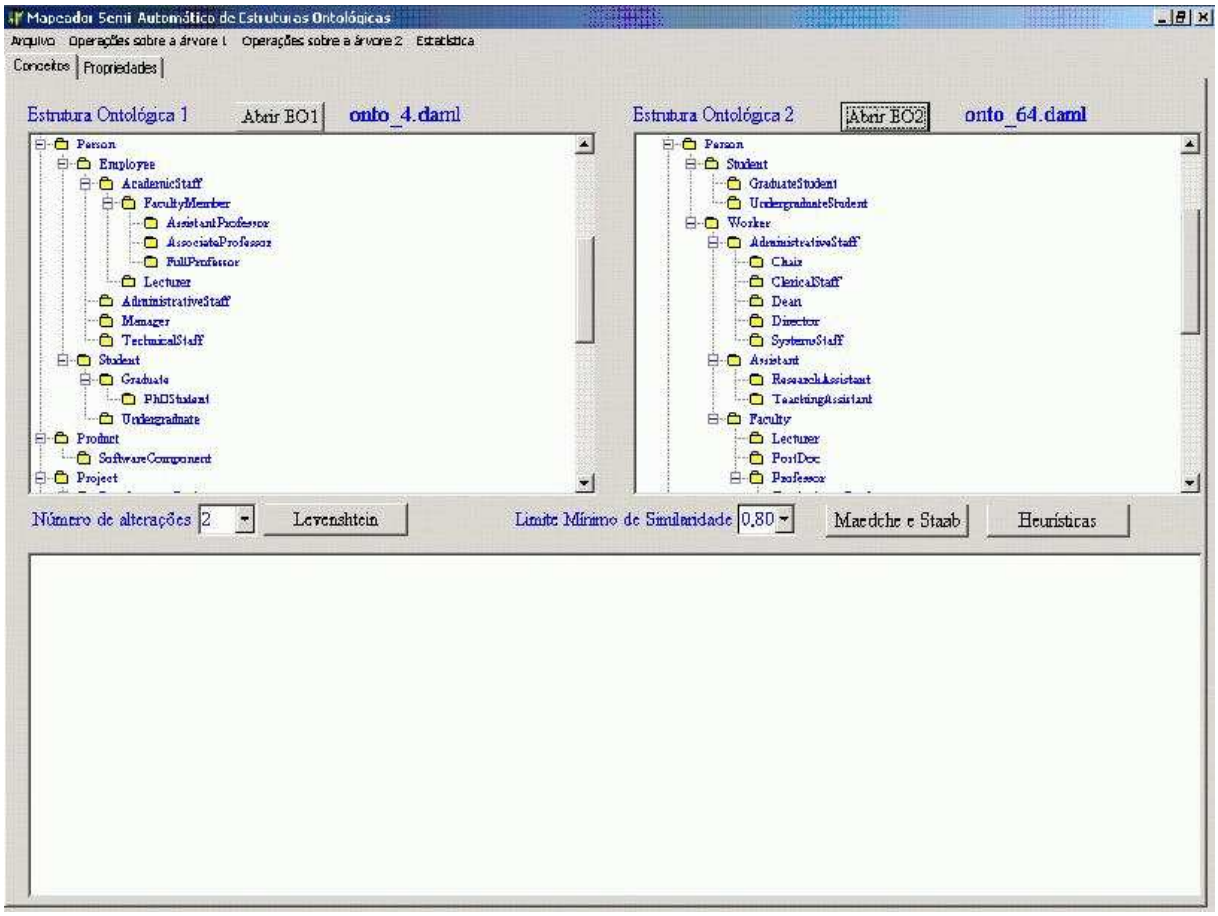

Figura 4.1: Interface do protótipo desenvolvido.

4.2 Mapeador Semi-Automático de Estruturas Ontológicas - Protótipo

O protótipo denominado `Mapeador Semi-Automático de Estruturas Ontológicas` foi desenvolvido no ambiente Borland Delphi 5 na plataforma Windows NT. A interface do protótipo apresentada na figura 4.1 permite ao usuário carregar os conceitos de duas EOs a serem mapeadas. Esses conceitos são apresentados na forma de uma hierarquia, permitindo ao usuário navegar entre eles. À medida que o usuário necessite conceitos mais específicos é possível ir descendo na hierarquia e expandindo a mesma.

O mapeamento pode ser realizado utilizando a Distância de Edição (DE) - botão `Levenshtein`, a medida Combinação de Caracteres (CC) (botão `Maedche e Staab`) ou ainda, no botão `Heurísticas` fazendo uso da medida CC combinada com as heurísticas que serão descritas nas subseções 4.4.1 e 4.4.2. Em todas essas ações o usuário deve especificar o limiar a ser considerado quando utilizando a DE ou a medida CC. Somente os mapeamentos com valores maior ou igual ao limiar estabelecido serão retornados na caixa de texto na parte inferior da interface.

Após a realização do mapeamento, os ícones dos conceitos que foram mapeados são alterados para facilitar a sua identificação pelo usuário. Já no menu `Estatística` (última opção no canto superior esquerdo) é possível verificar a quantidade de mapeamentos realizados, o percentual de mapeamentos encontrados e a contribuição das heurísticas no mapeamento, entres outras

informações. Maiores detalhes sobre as características do protótipo são descritos ao longo deste capítulo.

As principais abordagens encontradas na literatura para mapear EOs incluem a comparação entre termos ou conceitos em nível lexical e em nível semântico-estrutural. Nosso trabalho também segue estas abordagens, conforme descrito nas próximas seções.

4.3 Comparação Lexical

Mitra e Wiederhold [26] afirmam que uma combinação baseada somente no isomorfismo estrutural entre subgrafos das EOs, sem considerar o texto dos conceitos nessas EOs, apresenta um desempenho muito fraco. Com base nesta afirmação, nosso algoritmo inicialmente verifica a similaridade lexical entre conceitos, aplicando a medida de similaridade apresentada na equação 3.1.

A medida CC leva em consideração a DE de Levenshtein, a qual assume que, se dois termos possuem o mesmo comprimento e seus caracteres estão nas mesmas posições, então, bem provavelmente, esses termos serão equivalentes, exceto em caso de polissemia, que pode ser tratado por meio da abordagem semântico-estrutural (próxima subseção). Alguns resultados das combinações entre os termos pertencentes às duas EOs da língua inglesa são apresentados na tabela 4.1.

Tabela 4.1: Exemplos de valores de similaridade para termos da língua inglesa utilizando as medidas DE e CC

EO-base	**EO-alvo**	**DE**	**CC**
book	book	0	1
masterThesis	mastersThesis	1	0.92
book	work	2	0.5
book	booklet	3	0.25
employee	worker	7	0
academicStaff	faculty	11	0
facultyMember	professor	11	0

Se considerarmos DE $\leq$ `2` como limite aceitável para dois termos serem considerados similares, observa-se que serão gerados resultados inconsistentes para termos não equivalentes, tais como, `book` e `work`. Este limite pode ser modificado, pelo engenheiro do conhecimento, conforme o grau de consistência desejado no resultado do mapeamento.

Em nossa abordagem, adotamos a proposta de [1] e implementamos a medida CC. Essa medida parece ser melhor do que a DE de Levenshtein uma vez que leva em consideração o comprimento das palavras sendo combinadas. Entretanto, ainda é necessário investigar melhor seu desempenho para cadeias multipalavra.

Nosso algoritmo permite ao usuário estabelecer um grau mínimo de similaridade entre termos calculado pela medida CC, sendo que somente serão gerados mapeamentos acima desse mínimo. Quando esse valor for igual a `0.8`, por exemplo, combinações como `book` e `work`, não serão aceitas (ver tabela 4.1).

A medida CC diminui a influência de pseudo-diferenças[3] entre cadeias de caracteres em EOs distintas. Apesar de esta medida apresentar alguns resultados que induzem ao erro como, por

[3]Por exemplo, uso ou não de *underscore* ou hífens, uso de singular e plural ou uso de caracteres com marcação adicional.

exemplo, `lecture` e `lecturer`, a quantidade de combinações corretas contribuiu para o experimento realizado, sendo esta medida utilizada, também, em nível semântico-estrutural quando utilizamos duas heurísticas que consideram a posição dos termos na hierarquia.

No experimento realizado, ao mesmo tempo em que a medida proposta em [1] parecia gerar melhores resultados, algumas combinações bastante similares ou equivalentes ainda não estavam sendo mapeadas devido à baixa similaridade lexical.

Para completar o mapeamento entre termos não similares, a comparação foi realizada em nível semântico-estrutural.

4.4 Comparação Semântico-Estrutural

Após verificar se dois termos sendo combinados são equivalentes em nível lexical, verificam-se suas posições na hierarquia. Caso o termo mais genérico ou algum dos seus termos mais específicos sendo combinados sejam equivalentes, então sugere-se que os termos sendo combinados sejam equivalentes. Caso contrário, é bastante provável a ocorrência de polissemia.

Para realizar a comparação semântico-estrutural, fizemos uso de heurísticas e nos apoiamos na medida CC, pelo fato de a mesma ter produzido melhores resultados do que a medida DE em testes preliminares. Inicialmente, é sugerido o valor mínimo de similaridade 0.8, mas esse valor pode ser alterado pelo usuário.

Heurísticas têm sido utilizadas para auxiliar o mapeamento automático entre termos de EOs, conforme apresentado nos trabalhos correlatos [7, 6, 12].

4.4.1 Heurística 1: Normalização de Vocabulário

Pelo fato de as EOs serem projetadas por pessoas diferentes, com visões de mundo distintas, o texto dos conceitos pode ser diferente, mas sua semântica na EO é equivalente.

Levando em consideração os exemplos de EOs que estamos trabalhando, e o fato de tratar-se da língua inglesa, em alguns casos a terminação no texto de um conceito é igual àquela no texto do conceito pai. Por exemplo, os conceitos `GraduateStudent` e `UndergraduateStudent` encontram-se nessa situação.

Para tratar este tipo de situação, nosso algoritmo verifica se um conceito possui, na terminação, uma subcadeia do texto de seu conceito pai, e elimina essa terminação. No exemplo do parágrafo anterior, após essa normalização de vocabulário, `GraduateStudent` e `UndergraduateStudent` passam a ser tratados como `Graduate` e `Undergraduate` na comparação com outra EO. Quando detectado um grau de similaridade maior ou igual ao mínimo estabelecido pelo usuário, é identificado um mapeamento.

4.4.2 Heurística 2: Ancestral e Descendentes

Nos casos em que um conceito `X` em uma EO-base não possui equivalente lexical em uma EO-alvo, verifica-se o pai do conceito `X`. Se o pai do conceito `X` na EO-base é lexicalmente equivalente ao pai de um conceito `Y` na EO-alvo, então procura-se pela existência de similaridade entre os filhos dos conceitos `X` e `Y`. Se pelo menos um dos filhos de `X` é lexicalmente equivalente a um dos filhos de `Y`, então identifica-se um mapeamento entre `X` e `Y`. Por exemplo, na tabela 4.2 (a seguir), o termo `employee` pertencente à EO-base tem como pai o termo `person`, tal como o termo `worker` na EO-alvo. Seguindo na EO-base, verifica-se que `employee` possui o termo `administrativeStaff` como filho, o que também ocorre na EO-alvo para o termo `worker`. Assim, é identificado um mapeamento entre `employee` e `worker`.

Tabela 4.2: Exemplo de extratos de EOs da língua inglesa na forma hierárquica

EO-base	**EO-alvo**
Person	Person
Employee	Student
AcademicStaff	GraduateStudent
FacultyMember	UndergraduateStudent
AssistantProfessor	Worker
AssociateProfessor	AdministrativeStaff
FullProfessor	SystemStaff
Lecturer	Faculty
AdministrativeStaff	Lecturer
Student	PostDoct
Graduate	Professor
PhDStudent	AssistantProfessor
Undergraduate	AssociateProfessor
Event	FullProfessor
Lecture	VisitingProfessor

Após detectada a similaridade semântico-estrutural, esses termos passam a ser tratados como equivalentes para as combinações restantes. No exemplo anterior, os termos `employee` e `worker` auxiliam o algoritmo na identificação de mais dois mapeamentos. O primeiro mapeamento ocorre entre os termos `academicStaff` e `faculty`, pois ambos possuem pais equivalentes em nível imediatamente superior, e pelo menos um filho (`lecturer`) lexicalmente similar. Ainda, o segundo mapeamento ocorre entre os termos `facultyMember` e `professor`.

Finalizado o processamento automático, o especialista do domínio, utilizando o protótipo desenvolvido, pode modificar o grau desejado de similaridade lexical e processar o algoritmo novamente. A intervenção do especialista não se resume somente a identificar o grau de similaridade mais adequado para sua aplicação. O especialista pode ainda:

- aceitar ou rejeitar a combinação estabelecida pelo algoritmo;
- eliminar a combinação sugerida;
- detectar combinações irrelevantes à aplicação;
- indicar novas combinações que o algoritmo não detectou automaticamente.

Dessa forma, nosso protótipo caracteriza-se como semi-automático, tal como as propostas citadas na introdução deste trabalho.

4.4.3 Experimento e Resultados Preliminares

Após a descrição das medidas de similaridade e das heurísticas utilizadas, nesta subseção são apresentados o experimento realizado e os resultados obtidos.

Um total de 30 EOs nos formatos RDF, OWL, OIL e DAML+OIL foram processadas e exibidas na forma de uma estrutura hierárquica para o usuário. Elencamos duas EOs, para as quais alguns conceitos selecionados são apresentados na tabela 4.2. Essas EOs produziram os dados mostrados na tabela 4.3, após o processamento automático.

Tabela 4.3: Dados sobre as EOs da língua inglesa processadas

	Valor Absoluto		**Valor Percentual**	
	EO-base	EO-alvo	EO-base	EO-alvo
Quantidade Total de Conceitos	54	44	100%	
Quantidade de Mapeamentos (CC)	16		29,62%	36,36%
Quantidade de Mapeamentos (CC + Heurísticas)	22		40,74%	50%
Ganho obtido com as Heurísticas	6		11,12%	13,64%

A EO-base, pertencente ao domínio das comunidades de pesquisa em *Web* Semântica, possui 54 conceitos. Após o processamento utilizando a medida CC, o algoritmo encontrou 16 equivalências com os 44 conceitos pertencentes à EO-alvo, cujo domínio são os departamentos acadêmicos de universidades. Essas 16 equivalências constituem 29,62% dos 54 conceitos da EO-base e 36,36% dos 54 conceitos da EO-alvo.

Após o processamento usando apenas a medida CC, realizamos outro processamento para as mesmas EOs, utilizando as heurísticas propostas. Essas heurísticas permitiram a geração de mais `6` mapeamentos, o que corresponde a um ganho de `27,27%` sobre o resultado obtido com o uso somente da medida CC. A quantidade de mapeamentos detectados após a aplicação das heurísticas melhorou em 11,12% os resultados para a EO-base, o que corresponde a 40.74% de seus 54 conceitos. Para a EO-alvo, a utilização das heurísticas permitiu mapear 50% de seus 44 conceitos.

Os percentuais de conceitos mapeados utilizando as heurísticas (`40.74%` dos conceitos da EO-base e `50%` dos conceitos da EO-alvo) parecem ser significativos, considerando-se que uma verificação manual, sob o ponto de vista do usuário, não encontrou novas equivalências entre os conceitos sendo mapeados. Ou seja, os termos não mapeados possivelmente não possuem equivalência.

Neste experimento, o valor mínimo de similaridade para gerar um mapeamento foi igual a `0.75`, pois valores menores geraram mapeamentos incorretos, tal como o mapeamento entre os termos `AssistantProfessor` e `VisitingProfessor`. Nos casos em que o algoritmo encontrou mais de uma equivalência para um termo como, por exemplo, o termo `lecturer` (EO-alvo) que possui um valor alto de similaridade com os termos `lecture` e `lecturer` (EO-base), apenas aqueles termos mapeados cujo resultado possui maior valor são apresentados para o usuário.

Ao final do processamento, os mapeamentos encontrados pelo algoritmo são apresentados ao usuário juntamente com o valor de similaridade entre os termos mapeados. Além disso, o usuário pode conferir as posições dos termos mapeados na hierarquia, uma vez que seus ícones são marcados diferentemente.

Os resultados apresentados na tabela 4.3 permitem, ao engenheiro do conhecimento, estimar o grau de sobreposição existente entre as EOs. No caso de o engenheiro do conhecimento necessitar desenvolver uma EO para um determinado domínio, ele pode verificar as EOs já existentes, compará-las e decidir se é viável ou não o reuso das mesmas.

Na biblioteca do programa DAML existem, atualmente, mais de 270 EOs. A maior parte está descrita na língua inglesa e algumas na língua alemã. As EOs apresentam uma diversidade de tamanho e de domínios. Essa diversidade prejudicou nossa avaliação, já que não foi possível encontrar similaridade entre EOs de domínios distintos.

Atualmente, existe uma quantidade suficiente de ontologias na língua inglesa que estão marcadas com estas linguagens de marcação. Entretanto, ainda não dispomos de uma quantidade

suficiente de ontologias na língua portuguesa que utilizem tais linguagens.

4.5 Considerações sobre o Capítulo no Contexto da Dissertação

Este capítulo apresentou o experimento realizado com EOs da língua inglesa. Utilizamos EOs representadas em diferentes sintaxes pelas linguagens de marcação semântica. Foi desenvolvido um protótipo de modo a facilitar o mapeamento e a visualização gráfica das EOs.

Trabalhamos em dois níveis, a saber: lexical e semântico-estrutural. Em nível lexical, nos apoiamos na medida CC que já estava disponível na literatura, porém só havia sido aplicada a EOs da língua alemã. Em nível semântico-estrutural desenvolvemos duas heurísticas, cujo objetivo é detectar termos com baixa similaridade lexical, porém com semântica próxima.

O próximo capítulo apresenta o tratamento de EOs da língua portuguesa, no qual utilizamos a medida CC, propomos uma nova medida de similaridade e descrevemos a fase de validação dessa medida.

Capítulo 5

Tratamento de EOs da Língua Portuguesa e a Medida de Similaridade Proposta

"No próprio homem há uma maternidade carnal e espiritual; a sua criação também é uma maneira de dar à luz, pois criar com plenitude íntima é dar à luz".
Rainer Maria Rilke, escritor e poeta alemão (1875-1926)

5.1 Preâmbulo

Este capítulo trata da busca por termos similares em EOs da língua portuguesa. Inicialmente, aplicamos a medida CC para detectar termos similares em nível lexical. Esta medida apresentou diversos resultados inconsistentes nos motivando a desenvolver a medida "Similaridade Lexical".

Neste capítulo apresentamos, na seção 5.2, resultados da aplicação da medida CC para termos da língua portuguesa. Pelo fato de a medida "Similaridade Lexical" utilizar um algoritmo de *stemming*, a seção 5.3 discorre sobre esse assunto antes de apresentarmos nossa medida de similaridade.

Após a apresentação do algoritmo de *stemming*, a medida "Similaridade Lexical" é explicada em detalhes na seção 5.4. Com essa medida são realizados experimentos em duas fases, a saber: fase de validação (descrita na seção 5.5) e fase de avaliação. Este capítulo ainda apresenta resultados da fase de validação na seção 5.6, que permitiram refinar a medida para a fase de avaliação, assunto do capítulo 6. Finalmente, apresentamos uma heurística que nos auxiliou a detectar termos similares que não estavam sendo mapeados anteriormente.

5.2 Aplicação da Medida CC nas EOs da Língua Portuguesa

Após a realização dos experimentos com a língua inglesa, passamos a trabalhar com EOs na língua portuguesa. Essas EOs são provenientes de duas fontes distintas, sendo elas o Senado Federal, EO-base deste trabalho, e a USP, a EO-alvo. Ambas as EOs possuem diversos domínios do conhecimento, tais como agricultura, ciências políticas e enfermagem. Assim, os termos dessas EOs são comparados com termos do mesmo domínio (por exemplo, pertencentes ao domínio agricultura nas duas EOs) bem como de domínios distintos (termos pertencentes ao domínio ciências políticas comparados com termos pertencentes ao domínio de enfermagem, por exemplo).

Esta característica difere dos experimentos efetuados na bibliografia e daqueles realizados com a língua inglesa.

Os termos pertencentes às EOs da língua portuguesa também passaram por uma fase de pré-processamento, na qual foram convertidos para o mesmo formato das EOs da língua inglesa.

As EOs da língua inglesa utilizadas em nosso trabalho apresentam seus termos como uma única cadeia de caracteres, na qual são eliminados os espaços em branco, nos casos de termos multipalavra. Além da eliminação desses espaços, a primeira letra, a partir da segunda palavra de um termo multipalavra, é colocada como um caractere maiúsculo, permitindo identificar início e fim das palavras que compõe um termo multipalavra. Esse procedimento foi realizado para podermos comparar dados nos mesmos formatos.

Inicialmente, realizamos experimentos aplicando a medida CC aos termos da língua portuguesa, obtendo grande quantidade de resultados inconsistentes, alguns deles apresentados na tabela 5.1.

Tabela 5.1: Exemplos de termos mapeados entre as EOs da língua portuguesa utilizando a medida CC

EO-base	EO-alvo	CC
profissao	profissoes	0.67
religiao	religioes	0.62
comunicacaoDigital	comunicacoesDigitais	0.72
perversaoSexual	perversoesSexuais	0.67
bioetica	cinetica	0.75
amamentacao	lamentacao	0.80
criacaoDeEquino	criacaoDeSuinos	0.80
rendaPermanente	dentePermanente	0.80

A tabela 5.1 está dividida em duas partes. A primeira apresenta termos similares com variação de número, entretanto a medida CC aponta como resultado da similaridade valores abaixo do limiar 0.75. Esses resultados preliminares são um indício de que essa medida não seja adequada para tratar termos da língua portuguesa com a característica de variação de número. Por outro lado, na segunda parte da mesma tabela é possível notar termos dissimilares sendo considerados similares pela medida CC. Em ambas as partes evidencia-se resultados aparentemente inconsistentes tanto para termos formados por uma palavra quanto para termos multipalavra.

Tendo em vista os equívocos percebidos (ver tabela 5.1) nos resultados gerados pela medida CC quando aplicada a EOs da língua portuguesa, nosso trabalho apresenta uma medida alternativa. Baseados nos resultados preliminares retornados pela medida CC, acreditamos que se considerarmos somente o radical de cada palavra em um termo poderemos ter resultados mais consistentes. Para isso, nossa medida faz uso de um algoritmo de *stemming*, que é assunto da próxima seção.

5.3 Algoritmo de *Stemming*

Freqüentemente, o usuário especifica uma palavra em uma consulta, mas somente uma variante dessa palavra é apresentada em um documento relevante. Plurais, formas de gerúndio e sufixos de tempo passado são exemplos de variações que impedem uma perfeita combinação entre palavras sendo comparadas. Este problema pode ser parcialmente solucionado com a substituição de palavras por seus respectivos *stems* [73].

Stem é o conjunto de caracteres resultante de um procedimento de *stemming* [74]. Ele não necessariamente é igual à raiz lingüística, mas servirá como uma denotação mínima, não ambígua, do termo. De acordo com Spark Jones e Willet [75], o processo de *stemming* consiste em reduzir todas as palavras ao mesmo *stem*, por meio da retirada dos afixos[1] da palavra, permanecendo apenas a raiz. O propósito, segundo Honrado *et al.* [76], é chegar a um *stem* que capture uma palavra com generalidade suficiente para permitir sucesso na combinação de caracteres sem perder muito em detalhe e precisão. Um exemplo típico de um *stem* é `conect` que é o *stem* de `conectar`, `conectado`, `conectando`, etc.

Dois erros típicos que costumam ocorrer durante o processo de *stemming* são *overstemming* e *understemming*. *Overstemming* ocorre quando a cadeia de caracteres removida não é um sufixo, mas parte do *stem*. Por exemplo, a palavra gramática, após ser processada por um *stemmer*, é transformada no *stem* `grama`. Neste caso, a cadeia de caracteres removida eliminou parte do *stem* correto, a saber `gramát`. Já *understemming* ocorre quando um sufixo não é removido completamente. Por exemplo, este problema ocorre quando a palavra `referência` é transformada no *stem* `referênc`, ao invés do *stem* correto `refer`. Nos casos em que o algoritmo de *stemming* apresenta algum desses erros, os mesmos não costumam ser corrigidos manualmente. O procedimento de *stemming* utilizado não foi corrigido em nosso trabalho.

O algoritmo de *stemming* utilizado nesta dissertação foi desenvolvido especificamente para a língua portuguesa. Ele é chamado PortugueseStemmer e foi gentilmente fornecido a nosso grupo de pesquisa por Viviane Orengo [77]. Este algoritmo apresentou bons resultados quando comparado com o algoritmo de Porter[2] em [77] e quando comparado com outro algoritmo também desenvolvido especificamente para a língua portuguesa em [78].

5.4 A Medida "Similaridade Lexical"

Antes de apresentar nossa medida é importante retomar a medida CC, da qual nossa medida se origina.

$$CC(T_i, T_j) := max\left(0, \frac{min(\mid T_i \mid, \mid T_j \mid) \; - \; DE(T_i, T_j)}{min(\mid T_i \mid, \mid T_j \mid)}\right) \; \in \; [0,1]$$

A medida que propomos é por nós denominada Similaridade Lexical (deste ponto em diante será notada como SL) e está expressa na equação 5.1.

$$SL(T_i, T_j) = min\{\Delta^1_{ij}, \Delta^2_{ij}, \; \ldots, \Delta^k_{ij}\} \; \in \; [0,1] \tag{5.1}$$

Na medida SL os termos de cada EO, aos quais deseja-se atribuir um grau de similaridade, são representados por (T_i, T_j), onde o índice i refere-se aos termos da EO-base (neste caso a EO do Senado) e o índice j refere-se aos termos da EO-alvo (neste caso a EO da USP). Esses termos podem ser tanto monopalavra quanto multipalavra. SL, ao contrário da medida CC, leva em consideração somente o radical de cada palavra, e não a palavra com todos os seus caracteres. O símbolo Δ representa o cálculo realizado pela medida CC, que obedece às seguintes condições:

[1] Prefixos e sufixos.

[2] Algoritmo desenvolvido em 1980 por Martin Porter com o objetivo de tratar palavras da língua inglesa. Tem sido adaptado para várias línguas latinas, tais como espanhol e português.

$$\Delta_{ij}^{k} = \begin{cases} CC(Rad_i^k, Rad_j^k) & se\ DE = 0 \\ CC(Rad_i^k, Rad_j^k) - 0.1 & se\ DE = 1 \\ CC(Rad_i^k, Rad_j^k) - 0.2 & se\ DE = 2 \\ 0 & se\ DE \geq 3 \end{cases} \tag{5.2}$$

Na equação 5.2 o radical de uma palavra contida nos termos (T_i, T_j) é expresso por (Rad_i^k, Rad_j^k), onde o índice k indica a posição da palavra no termo. Quando os termos (T_i, T_j) possuem uma quantidade de palavras diferentes, o índice k varia até a quantidade de palavras do termo com menor número de palavras. A medida SL calcula a similaridade entre cada par de palavras pertencente aos termos (T_i, T_j) sendo mapeados.

O resultado final da medida SL é o menor valor gerado em Δ_{ij}^{k}, para os diferentes k. Esse valor é dependente do resultado de similaridade retornado pela distância de edição (DE). A DE verifica o número de inserções, alterações ou deleções necessárias para transformar o termo T_i em T_j. No caso da medida SL, o cálculo da DE é realizado sobre os radicais dos termos sendo comparados. Como o radical de um termo pode possuir carga semântica bastante forte, o resultado gerado pela DE ainda é decrementado conforme as condições apresentadas na equação 5.2.

De acordo com a equação 5.2, quanto maior o resultado gerado pela DE, maior o valor de decremento utilizado. Os valores 0.1 e 0.2 foram utilizados devido à necessidade de se introduzir uma penalidade ao valor retornado pela medida CC quando aplicado ao radical da palavra. Esses valores foram detectados a partir de uma análise empírica de experimentos preliminares. Assumimos que, se a DE é ≥ 3, o valor de similaridade retornado pela medida CC em Δ_{ij}^{k} é igual a zero. Essa decisão pode ser justificada pelo fato de que três ou mais alterações no radical de uma palavra caracterizam um baixo grau de similaridade entre os termos sendo mapeados.

Suponha que se deseje verificar a similaridade entre os termos `amazoniaOriental` e `amazoniaOcidental`. Para calcular o valor resultante da medida CC, cada uma das palavras que compõem esses termos é processada pelo algoritmo de *stemming*. Logo, temos a medida SL como segue:

$$SL(amazoniaOriental, amazoniaOcidental) = min\{CC(amazon, amazon), CC(orient, ocident)\}$$

O processamento de $CC(amazon, amazon)$ resulta no valor 1, ao passo que o processamento de $CC(orient, ocident)$ assume o valor a seguir:

$$max\left(0, \frac{6\ -\ 2}{6}\right)\ =\ 0.67$$

Como a $DE = 2$, o valor 0.67 é decrementado em 0.2, resultando em 0.47. Portanto:

$$SL(amazoniaOriental, amazoniaOcidental) = 0.47$$

Não foi encontrado na literatura trabalhos que apresentem um estudo sobre o peso da carga semântica em termos multipalavra, ou seja, qual das palavras que compõem o termo representam melhor o seu significado. Em nossa proposta, conforme pode ser observado na medida SL, as palavras com menor similaridade lexical é que determinam o valor de similaridade entre os termos.

Após detalhar o funcionamento da medida SL, apresentaremos na próxima seção os experimentos realizados na fase de validação dessa medida.

5.5 Fase de Validação da Medida "Similaridade Lexical"

Os experimentos realizados com as EOs da língua portuguesa incluem uma fase de validação da medida `SL`, seguida de uma fase de avaliação. Para realizar os experimentos em duas fases, os termos das EOs foram divididos em dois grupos: termos monopalavra e termos multipalavra. Os termos da EO-base foram divididos pela metade, para cada fase. A EO-alvo permaneceu com todos os seus termos nas duas fases. Os termos que compõem as EOs foram colocados em ordem alfabética e percorreu-se os termos distribuindo um termo para cada fase. A tabela 5.2 apresenta a quantidade de termos das EOs e o modo como as mesmas foram divididas, para cada uma das fases. Apresenta também a quantidade total de termos utilizados em nossos experimentos. A EO do Senado possui 3647 termos compostos por uma palavra, o que representa 28% do total de 13049 termos. A EO da USP possui 7039 termos monopalavra, representando 29% dos total de 24025 termos. As campos da tabela 5.2 referentes à quantidade de termos da EO da USP estão em branco porque a mesma não foi dividida para a realização dos experimentos.

Tabela 5.2: Dados importantes das EOs da língua portuguesa

EO	Tipo Termo	Quantidade de Termos		Total Tipo Termo	%	Total de Termos
		Validação	Avaliação			
Senado	Monopalavra	1824	1823	3647	28	13049
	Multipalavra	4701	4701	9402	72	
USP	Monopalavra	-	-	7039	29	24025
	Multipalavra	-	-	16986	71	

Na fase de validação da medida `SL` utilizamos a quantidade de termos apresentada na coluna `Validação` da tabela 5.2.

O objetivo dos experimentos descritos neste capítulo é verificar o comportamento da medida SL quando aplicada a termos da língua portuguesa e ajustar a proposta, para que possa vir a ser estabilizada e, então, avaliada. Esses experimentos obedecem as diferentes condições, conforme mostrado na tabela 5.3.

Tabela 5.3: Casos tratados no experimento

Número do caso	Condição	
1º	$CC \geq 0.75$	$SL \geq 0.75$
2º	$CC \geq 0.75$	$SL < 0.75$
3º	$CC < 0.75$	$SL \geq 0.75$

A tabela 5.3 apresenta os casos possíveis de combinação entre as condições especificadas, para cada medida de similaridade. Esses casos serão explicados nos itens a seguir:

1. O primeiro caso diz respeito à concordância de similaridade entre as medidas, ou seja, busca-se verificar as características dos termos nos quais ambas as medidas detectam similaridade.

2. O segundo caso refere-se aos termos que são considerados similares pela medida CC e não são considerados similares pela medida SL.

3. O terceiro caso resulta em pares de termos considerados similares pela medida SL e não similares pela medida CC.

A tabela 5.4 apresenta um resumo dos resultados obtidos na fase de validação, para cada um dos casos descritos na tabela 5.3.

Tabela 5.4: Quantidade de pares de termos mapeados em cada caso

Caso	Monopalavra	Multipalavra	Total
1º	53	18	71
2º	1026	1608	2634
3º	45	10	55
Total	1124	1636	**2760**

De acordo com a tabela 5.4, um total de 2760 pares de termos foram considerados similares pela medida CC ou pela medida SL. Na análise desta tabela pode-se perceber que a maior parte (2634) dos pares de termos retornados como similares estão inclusos no segundo caso, no qual a medida CC acusa similaridade entre os termos e a medida SL refere-se a termos não similares. Em outras palavras, mais de 95% dos termos que participam da fase de validação das medidas CC e SL são considerados similares pela medida CC e não similares pela medida SL.

5.6 Considerações sobre os Resultados da Fase de Validação

No apêndice B são apresentadas as tabelas B.1, B.2, B.3 e B.4 com os resultados de cada caso em particular.

Ao se fazer uma análise caso-a-caso dessas tabelas, percebe-se algumas situações em que tanto a medida CC quanto a medida SL apresentam valores que consideram similares pares de termos não similares e vice-versa. Dessa forma, é importante a realização de análise humana sobre os pares de termos considerados similares pela medida CC ou pela medida SL, de modo que se possa fazer uma avaliação mais consistente sobre os resultados gerados automaticamente. Essa análise é apresentada no capítulo 6, e nela é verificada a precisão das medidas comparando-as com a análise humana.

Um extrato da tabela B.2 do apêndice B, no qual ambas as medidas consideram os mesmos pares de termos similares, é apresentado na tabela 5.5, na qual os termos apresentam a característica de variação de número. Nesta tabela são mostrados termos multipalavra.

A tabela 5.5 apresenta 13 (76,5%) dos 17 termos multipalavra com a característica de variação de número considerados similares pelas medidas CC e SL. Esses números são um indício de que as medidas CC e SL podem tratar termos multipalavra com variação de número de forma consistente, pois ambas detectam como similares termos realmente similares. Entretanto, para os termos monopalavra, o mesmo não ocorre, conforme a tabela 5.6.

Na tabela 5.6, o leitor pode constatar que a medida CC trata os termos monopalavra com variação de número com similaridade inferior ao limiar 0.75, enquanto a medida SL aponta um grau de similaridade acima de 0.75, mesmo havendo erro resultante do processo de *stemming*. Esses erros são devidos à troca do caractere `ç` por `c` e da eliminação do caracter $\sim$.

Outros erros do algoritmo de *stemming* que, dessa vez, prejudicaram o desempenho da medida SL são apresentados na tabela 5.7, que é um extrato da tabela B.4 do apêndice B. Estes termos são considerados similares em virtude dos erros do *stemmer*.

A tabela 5.7 apresenta as correções (realizadas de forma manual) nos casos em que o algoritmo de *stemming* não processou de forma correta os pares de termos. É possível observar que cinco, dos seis casos, são termos com semântica distinta; após corrigido o *stem*, passam a ser considerados não similares pela medida SL. Enquanto isso, os termos `profissao` e `profissoes`

Tabela 5.5: Termos multipalavra com variação de número considerados similares por CC e SL

EO-base	EO-alvo	CC	SL
acumulacaoDeAcoes	cumulacaoDeAcoes	0.94	0.77
bicho-da-seda	bichos-da-seda	0.92	0.82
competicaoEsportiva	competicoesEsportivas	0.79	0.79
condicoesEconomicas	condicaoEconomica	0.76	0.76
condicoesSanitarias	condicaoSanitaria	0.76	0.76
construcaoMetalica	construcoesMetalicas	0.78	0.79
criacaoDeCaracol	criacaoDeCaracois	0.88	0.76
descobertaEExploracao	descobertasEExploracoes	0.81	0.79
expedicaoCientifica	expedicoesCientificas	0.79	0.77
exposicaoInternacional	exposicoesInternacionais	0.77	0.77
instituicaoFinanceira	instituicoesFinanceiras	0.81	0.80
instituicaoPolitica	instituicoesPoliticas	0.79	0.80
religiaoPrimitiva	religioesPrimitivas	0.76	0.76

Tabela 5.6: Extrato dos termos da tabela B.4 com variação de número considerados similares por CC e SL

EO-base	EO-alvo	CC	SL
adivinhacao	adivinhacoes	0.73	0.80
caminhao	caminhoes	0.62	0.76
corporacao	corporacoes	0.70	0.79
embarcacao	embarcacoes	0.70	0.79
habitacao	habitacoes	0.67	0.77
profissao	profissoes	0.67	0.77
religiao	religioes	0.62	0.76

Tabela 5.7: Termos que apresentam erros ao final do processo de *stemming*

EO-base	EO-alvo	CC	SL	SL corrigida	*Stem* corrigido	*Stem* corrigido
empresario	emprestimo	0.70	0.76	0.73	empres	emprest
inflamaveis	inflacao	0.38	0.76	0.73	inflam	inflac
magistrado	magisterio	0.70	0.76	0.73	magistr	magist
metanol	metabolismo	0.29	0.76	0	metan	metabol
profissao	profissoes	0.67	0.77	1	profiss	profiss
responsabilidade	responsorio	0.27	0.76	0.51	respons	responsor

têm seu valor de similaridade elevado para 1 pela mesma medida. Realmente, esses últimos termos são bastante similares, havendo apenas variação de número, o que deve ser desconsiderado quando se deseja mapear termos entre EOs.

A partir da análise preliminar das tabelas no apêndice B, percebe-se que, em alguns casos, termos não similares estavam sendo considerados similares pela medida SL. Em muitos termos que possuem alta similaridade lexical, entretanto, apenas a primeira letra do par de termos sendo comparados é diferente. Considerando que, na língua portuguesa, a carga semântica da primeira letra de uma palavra é bastante forte, complementamos a medida SL com a heurística apresentada na seção 5.6.1.

5.6.1 Heurística "Primeira Letra"

Retomamos a medida SL apresentada na equação 5.1.

$$SL(T_i, T_j) = min\{\Delta_{ij}^1, \Delta_{ij}^2, \ \ldots, \Delta_{ij}^k\} \ \in \ [0,1]$$

Assumimos que:

$$Se\ Rad[1]_i^k \neq Rad[1]_j^k\ então\ CC(Rad_i^k, Rad_j^k) = 0$$

Seja o valor entre colchetes a posição da primeira letra, em um radical de uma palavra em um termo T. Caso esta primeira letra seja diferente entre quaisquer dois radicais Rad_i^k, Rad_j^k sendo comparados, o valor retornado na medida CC é igual a zero.

Esta heurística, após aplicada ao conjunto de termos da fase de validação, apresentou os resultados mostrados na coluna SLH da tabela 5.8.

Tabela 5.8: Resultados após a aplicação da heurística "primeira letra" na fase de validação

EO-base	EO-alvo	CC	SL	**SLH**	EO-base	EO-alvo
cartilha	partilha	0.88	0.76	0	cartilh	partilh
dolarizacao	polarizacao	0.91	0.80	0	dolarizaca	polarizaca
dolarizacao	solarizacao	0.91	0.80	0	dolarizaca	solarizaca
emigracao	imigracao	0.89	0.77	0	emigraca	imigraca
emigracao	migracao	0.88	0.76	0	emigraca	migraca
filiacao	afiliacao	0.88	0.76	0	filiaca	afiliaca
fitogenetica	citogenetica	0.92	0.77	0	fitogene	citogene
fitologia	citologia	0.89	0.76	0	fitolog	citolog
fitologia	litologia	0.89	0.76	0	fitolog	litolog
fitologia	mitologia	0.89	0.76	0	fitolog	mitolog
ginecologia	sinecologia	0.91	0.79	0	ginecolog	sinecolog
imigracao	migracao	0.88	0.76	0	imigraca	migraca
matrimonio	patrimonio	0.90	0.79	0	matrimoni	patrimoni
mobilizacao	imobilizacao	0.91	0.80	0	mobilizaca	imobilizaca
orizicultura	rizicultura	0.91	0.77	0	orizicult	rizicult
ovinocultura	bovinocultura	0.92	0.79	0	ovinocult	bovinocult
retencao	detencao	0.88	0.76	0	retenca	detenca
servidao	cervidae	0.75	0.76	0	servida	cervida
acumulacaoDeAcoes	cumulacaoDeAcoes	0.94	0.77	0	acumulacadeaco	cumulacadeaco
mercadoMobiliario	mercadoImobiliario	0.88	0.76	0	mercmobilia	mercimobilia

A maioria dos pares de termos apresentados na tabela 5.8 pode ser, em um primeiro momento, considerada de termos não similares. Após a aplicação da heurística, esses termos, que até então eram considerados similares, passam a ter como similaridade o valor zero e, conseqüentemente, não são mapeados entre as EOs.

Dos 20 pares de termos apresentados na tabela 5.8, apenas dois são termos multipalavra. Neste caso, a heurística atua sobre todas as palavras do termo. Por exemplo, nos termos `mercadoMobiliario` e `mercadoImobiliario` a heurística foi aplicada na segunda palavra do termo, ao passo que nos termos `acumulacaoDeAcoes` e `cumulacaoDeAcoes` a heurística foi aplicada na primeira palavra. Nesse último caso, cometendo-se um equívoco, pois os termos são, em um primeiro momento, semanticamente similares. Apesar desses termos não serem detectados como similares em nível lexical, eles ainda podem ser mapeados em nível semântico-estrutural, conforme suas posições nas hierarquias de conceitos e as relações semânticas envolvidas.

5.7 Considerações sobre o Capítulo no Contexto da Dissertação

Este capítulo apresentou resultados preliminares da aplicação da medida CC para EOs da língua portuguesa, bem como foi descrito o algoritmo de *stemming* utilizado na medida SL. A utilidade da medida SL foi, inicialmente, verificada através de uma fase de validação, na qual a mesma foi submetida a comparar pares de termos pertencentes a EOs da língua portuguesa.

Após a primeira análise dos resultados gerados pela medida SL verificou-se que pares de termos nos quais a primeira letra do radical, que possui alta carga semântica, é diferente, estavam sendo considerados similares quando seus significados eram distintos. Diante disso, desenvolvemos a heurística "primeira letra" para alcançar melhores resultados de similaridade.

As tabelas apresentadas ao longo deste capítulo buscam sintetizar os resultados preliminares de ambas as medidas - CC e SL - tentando unir pares de termos com características próximas (tais como variação de número, pares de termos nos quais o algoritmo de *stemming* apresenta erro, entre outros) na mesma tabela.

O próximo capítulo contempla a fase de avaliação da medida SL, na qual essa medida é comparada com a medida CC e com os resultados da análise humana.

Capítulo 6

Avaliação e Análise Crítica

"Se você não perguntar o porquê das coisas, logo estarão perguntado o porquê de você".
Autor desconhecido.

6.1 Preâmbulo

Após a realização dos experimentos da fase de validação da medida SL, é necessário que essa medida seja avaliada. O processo de avaliação foi realizado sobre um conjunto de termos diferente daqueles utilizados na fase de validação.

Este capítulo apresenta, em detalhes, os resultados da fase de avaliação da medida de similaridade SL. Para ocorrer essa avaliação, a medida SL é comparada com a medida CC e, principalmente, com os resultados de uma análise humana de similaridade.

Os resultados da avaliação foram divididos em grupos, de acordo com restrições de similaridade. Esses grupos são descritos na seção 6.3. Na seção 6.3.4 constituímos mais um grupo com todos os casos em que a medida SL detectou similaridade entre os termos. Após a análise desses grupos, solicitamos a outro indivíduo a tarefa de revisar a análise humana. Esta revisão é apresentada na seção 6.3.5. Na seção 6.4 já trabalhamos em nível semântico-estrutural, comentamos a aplicação da heurística "ancestral e descendentes" e apresentamos os experimentos realizados, levando em consideração a relação semântica de sinonímia.

Finalmente, na seção 6.5, é apresentada a análise crítica do que foi observado no comportamento da medida SL.

6.2 Avaliação de Similaridade

Não existe um modo padrão para avaliar medidas de similaridade semântica [66]. Uma abordagem razoável parece ser a comparação com a concordância expressa por humanos. Entretanto, é importante destacar que os indivíduos podem ter diferentes observações sobre os mesmos pares de termos avaliados pelo processamento automático derivado das medidas de similaridade.

No contexto dos modelos conceituais, a similaridade é difícil de medir e, freqüentemente, estabelecer uma medida de similaridade adequada é uma tarefa bastante subjetiva [62].

Para avaliar as medidas de similaridade estudadas neste trabalho, utilizamos a segunda metade do conjunto de termos que compõem as EOs, cuja composição foi apresentada na tabela 5.2. Inicialmente, esses termos foram processados de forma automática utilizando as medidas de similaridade. A tabela 6.1 apresenta a quantidade total de mapeamentos encontrados em cada um dos casos descritos na tabela 5.3.

Tabela 6.1: Quantidade de pares de termos mapeados em cada caso da fase de análise

Caso	Monopalavra	Multipalavra	Total
1º	73	21	94
2º	1149	1592	2741
3º	45	7	52
Total	1267	1620	**2887**

De modo análogo aos resultados da fase de validação, a quantidade maior de mapeamentos se concentra no segundo caso, onde a medida CC mostra termos similares, ao passo que a medida SL apresenta os mesmos pares de termos como não similares.

Conforme descrito no início deste capítulo, os resultados apresentados na tabela 6.1 foram confrontados com os resultados da análise humana.

Nosso trabalho contou com a colaboração de dois avaliadores humanos, ambos mestrandos do curso de Letras, aos quais foi solicitado que avaliassem 1267 pares de termos monopalavra e 1620 pares de termos multipalavra. Esses termos foram identificados como similares, pela medida CC ou pela medida SL[1], totalizando 2887 pares de termos. Cada avaliador humano recebeu uma metade dos termos monopalavra e uma metade dos termos multipalavra.

Na especificação da tarefa foi solicitado a ambos verificarem se os pares de termos apresentados possuem significado similar no contexto do mapeamento entre EOs. Os pares de termos foram enviados aos avaliadores em uma tabela no software Excel, na qual os avaliadores podiam detectar cada par de termos como similar, não similar, ou ainda, avaliá-lo como caso duvidoso.

Dada a grande quantidade de dados disponíveis, a saber, 2887 pares de termos avaliados por duas medidas de similaridade mais os dados da avaliação humana, estabelecemos grupos para facilitar a análise dos resultados. A comparação dos resultados da análise automática com a análise humana é apresentada na próxima seção.

6.3 Comparação entre Análise Automática e Análise Humana

Para analisar os 2887 pares de termos, buscamos dividi-los em grupos formados por restrições. Esses grupos são apresentados na tabela 6.2, onde a letra `G` representa a palavra `Grupo`.

Tabela 6.2: Formação dos grupos para a análise

	$CC \geq 0.75$ $SL \geq 0.75$	$CC \geq 0.75$ $SL < 0.75$	$CC < 0.75$ $SL \geq 0.75$
Termos considerados similares pelos humanos	G1	G2	G3
Termos considerados não similares pelos humanos	G4	G5	G6
Dúvida		G7	

O critério para divisão desses grupos foi o mesmo adotado na fase de validação. No entanto acrescentamos a análise humana. A análise individual de cada grupo é apresentada nas próximas subseções.

[1] Em um primeiro momento, a heurística "primeira letra" não foi utilizada junto com a medida SL.

6.3.1 Grupos de Termos Considerados Similares pelos Avaliadores Humanos

Nas próximas seções é descrita em detalhe a análise dos grupos G1, G2, G3 apresentados na primeira linha da tabela 6.2.

6.3.1.1 Análise do Grupo G1

Iniciamos a análise pelo grupo G1, o qual apresenta a concordância entre as medidas de similaridade e a análise humana. Na tabela 6.3 encontram-se todos os termos que foram detectados como similares no grupo G1, obedecendo o limiar 0.75.

Tabela 6.3: Termos considerados similares pelo analisador humano e pelas medidas CC e SL

EO-base	EO-alvo	CC	SL	EO-base	EO-alvo
agricultura	agricultor	0.80	0.76	agricult	agricul
bovinocultura	ovinocultura	0.92	0.79	bovinocult	ovinocult
cafeicultor	cafeicultura	0.82	0.77	cafeicul	cafeicult
cartografia	cartograma	0.80	0.79	cartograf	cartogram
elegibilidade	legibilidade	0.92	0.76	elegibil	legibil
epigrafia	epigrama	0.75	0.76	epigraf	epigram
eutroficacao	eutrofizacao	0.92	0.81	eutroficaca	eutrofizaca
interferon	interferons	0.90	0.80	interferon	interferom
lexiologia	lexicologia	0.90	0.77	lexiolog	lexicolog
teleinformatica	telinformatica	0.93	0.80	teleinforma	telinforma
aglomeracaoUrbana	aglomeracoesUrbanas	0.76	0.80	aglomeracaurban	aglomeracourban
condicoesDeTrabalho	condicaoDeTrabalho	0.83	0.76	condicodetrabalh	condicadetrabalh
criacaoDeBicho-da-seda	criacaoDeBichos-da-seda	0.95	0.82	criacadebicho-da-sed	criacadebichos-da-sed
dinamicaDePopulacao	dinamicaDePopulacoes	0.84	0.77	dinamdepopulaca	dinamdepopulaco
energiaHidroeletrica	energiaHidreletrica	0.95	0.79	energhidroeletr	energhidreletr
inovacaoTecnologica	inovacoesTecnologicas	0.79	0.76	inovacatecnolog	inovacotecnolog
instalacaoAgricola	instalacoesAgricolas	0.78	0.79	instalacaagricol	instalacoagricol
instalacaoEletrica	instalacoesEletricas	0.78	0.79	instalacaeletr	instalacoeletr
instituicaoReligiosa	instituicoesReligiosas	0.80	0.80	instituicareligi	instituicoreligi
maquina-ferramenta	maquinas-ferramenta	0.94	0.84	maquina-ferrament	maquinas-ferrament
pedraSemipreciosa	pedrasSemi-preciosas	0.82	0.81	pedrsemiprecios	pedrsemi-precios
redeDeTelecomunicacao	redeDeTelecomunicacoes	0.86	0.83	reddetelecomunicaca	reddetelecomunicaco
sistemaDeInformacao Contabil	sistemasDeInformacoes Contabeis	0.78	0.77	sistemdeinformaca contavel	sistemdeinformaco contabel
tituloMobiliario	tituloImobiliario	0.88	0.76	titulmobilia	titulimobilia
usinaHidroeletrica	usinasHidreletricas	0.83	0.79	usinhidroeletr	usinhidreletr

Na análise da tabela 6.3 é importante destacar que existe concordância total de similaridade em apenas 25 dos 2887 termos detectados como similares pela medida CC ou pela medida SL. Além disso, dos 94 termos considerados similares por ambas as medidas, apenas 25 foram considerados realmente similares pelo analisador humano. Esses números indicam que 69 (ou 73%) pares de termos considerados similares pelas medidas CC e SL não foram considerados similares pelo analisador humano. Desses 69 termos, 27 termos que o avaliador humano não considerou como similares também não são detectados como similares pela medida SL quando utilizada a heurística da primeira letra, restando 42 pares de termos mapeados de forma equivocada pela a medida SL de acordo com o avaliador humano. Considerando esse valor elevado, o usuário é motivado a aumentar o limiar. Quando usado o limiar 0.8, apenas 4 pares de termos permanecem sendo considerados similares pelas medidas CC e SL, e esses são apresentados na tabela 6.4.

A tabela 6.4 apresenta os 4 pares de termos restantes após o valor do limiar ser aumentado

Tabela 6.4: Resultado após alteração do limiar para o valor 0.8

EO-base	EO-alvo	CC	SL	EO-base	EO-alvo
descriminacao	discriminacao	0.92	0.82	descriminaca	discriminaca
interpelacao	interpolacao	0.92	0.81	interpelaca	interpolaca
macroeconomia	microeconomia	0.92	0.81	macroeconom	microeconom
microprocessador	macroprocessadores	0.81	0.82	microprocess	macroprocess

em 0.05. Assim, percebe-se que o ato de aumentar o limiar para minimizar os mapeamentos inconsistentes pode ocasionar que termos similares não sejam detectados, pois observando a tabela 6.3 nota-se que termos similares foram detectados com valores no intervalo [0.75,0.79].

Dos 25 termos apresentados na tabela 6.3 destacamos três casos:

1. `bovinocultura` e `ovinocultura`
2. `elegibilidade` e `legibilidade`
3. `tituloMobiliario` e `tituloImobiliario`

Através dos três casos apresentados, percebemos que o analisador humano não possui uma noção refinada quanto à semântica dos termos ou do grau de similaridade desejado.

Esses casos apresentam baixa similaridade semântica, mas foram considerados similares pelo analisador humano e pelas medidas CC e SL. Entretanto, utilizando a heurística "primeira letra", a medida SL considera os três pares de termos como não similares. Os demais casos da tabela 6.3 parecem estar mapeados de forma correta, tanto pelo analisador humano quanto pelas medidas de similaridade.

6.3.1.2 Análise do Grupo G2

O grupo G2 consiste de termos que são considerados similares pelo analisador humano e pela medida CC, porém a medida SL não os considera similares. Iniciamos a análise apresentando um extrato dos termos pertencentes a esse grupo, na tabela 6.5.

Os termos apresentados na tabela 6.5 parecem, em um primeiro momento, possuir baixa similaridade. Apesar de o analisador humano considerar esses termos como similares, no contexto do mapeamento entre EOs entendemos que os mesmos não devem ser mapeados, uma vez que possuem significados distintos.

As situações pertencentes a esse grupo comprovam a natureza abstrata do que se entende por similaridade, e comprovam, mais uma vez, a dificuldade de se fazer uma análise consistente sobre os resultados apresentados.

Ainda constitui outro fato a ser analisado, a ocorrência de termos que incluem preposição na sua formação, e não são considerados similares pela medida SL, conforme a tabela C.1 (ver apêndice C). Na fase de preparação dos dados optamos por não remover as preposições, pelo fato de que essas não são tão representativas no conjunto de 37.074 termos. Para os termos da tabela C.1, se removêssemos as preposições, a medida SL os consideraria similares.

6.3.1.3 Análise do Grupo G3

Os termos pertencentes ao grupo G3 são apresentados na tabela 6.6. Neste grupo encontram-se os termos considerados similares pela medida SL e pelo analisador humano e não similares pela medida CC.

Tabela 6.5: Resultados pertencentes ao grupo G2

EO-base	EO-alvo	CC	SL
agricultura	apicultura	0.80	0.51
ascendente	descendente	0.80	0.47
autor	autos	0.80	0.13
biologia	citologia	0.75	0.47
bovinocultura	equinocultura	0.77	0.00
bovinocultura	suinocultura	0.75	0
elegibilidade	inelegibilidade	0.85	0.55
emigrante	imigrante	0.89	0.70
equinocultura	ovinocultura	0.75	0
administracaoDaProducao	administracaoDaEducacao	0.78	0
administracaoDireta	administracaoIndireta	0.84	0.40
biologiaHumana	etologiaHumana	0.86	0.47
filhoLegitimo	filhoIlegitimo	0.85	0.70
filmeDeLongaMetragem	filmeDeCurtaMetragem	0.80	0
ontologiaSocial	odontologiaSocial	0.87	0.51

Tabela 6.6: Termos pertencentes ao grupo G3

EO-base	EO-alvo	CC	SL	EO-base	EO-alvo
aeronautica	aeronaves	0.44	0.76	aeronau	aeronav
bibliotecario	bibliotecas	0.73	0.79	biblioteca	bibliotec
cognicao	cognicoes	0.63	0.76	cognica	cognico
combustivel	combustao	0.56	0.77	combusti	combusta
comerciario	comerciantes	0.64	0.76	comercia	comerci
comerciario	comercio	0.63	0.76	comercia	comerci
comissoes	comissao	0.63	0.76	comisso	comissa
comunicacao	comunicacoes	0.73	0.80	comunicaca	comunicaco
construtor	construtivismo	0.40	0.76	constru	construt
edificacao	edificacoes	0.70	0.79	edificaca	edificaco
escorpiao	escorpioes	0.67	0.77	escorpia	escorpio
existencialismo	existencia	0.50	0.77	existenci	existenc
fotografo	fotogravura	0.56	0.77	fotograf	fotograv
fundacao	fundacoes	0.63	0.76	fundaca	fundaco
obrigacao	obrigacoes	0.67	0.77	obrigaca	obrigaco
organizacoes	organizacao	0.73	0.80	organizaco	organizaca
aeronauticaMilitar	aeronavesMilitares	0.61	0.76	aeronaumilit	aeronavmilit
bicho-da-seda,Criacao	bichos-da-seda	0.36	0.82	bicho-da-sed	bichos-da-sed
construcaoRural	construcoesRurais	0.67	0.79	construcarural	construcorural

No grupo G3 pode-se observar uma característica na qual a medida SL apresenta um comportamento consistente. Os termos que apresentam variação de número (singular e plural), e que neste grupo foram considerados similares pelo analisador humano, também são considerados similares pela medida SL. Por outro lado, fica explícita uma das deficiências da medida CC no tratamento da língua portuguesa: termos monopalavra no singular, quando comparados com

termos monopalavra no plural, apresentam resultados inconsistentes. Este fato já havia sido observado na fase de validação e pôde ser confirmado nesta fase de avaliação.

Termos com variação de número são freqüentes em EOs, conforme pode ser notado nas EOs utilizadas neste trabalho. Essa variação de número vai ao encontro da consideração feita por Noy e McGuinness [79], de acordo com a qual os seres humanos devem modelar o conhecimento utilizando um padrão: todos os termos no singular ou todos os termos no plural. Por exemplo, quando se deseja encontrar similaridade entre termos de uma EO-base construída com termos no singular, com uma EO-alvo construída com termos no plural, faz-se necessária a utilização de uma medida de similaridade que contemple de forma consistente esses casos.

6.3.2 Grupos de Termos Considerados Não Similares pelos Avaliadores Humanos

A análise dos grupos apresentados a seguir refere-se à segunda linha da tabela 6.2.

6.3.2.1 Análise do Grupo G4

O grupo G4 apresenta os termos que as medidas de similaridade apontam como similares e o analisador humano aponta como não similares. Neste grupo foram encontrados 54 pares de termos. Em um primeiro momento, pode-se considerar que as medidas de similaridade apresentam um desempenho fraco neste grupo, uma vez que ambas discordam da análise humana. Entretanto, alguns resultados da medida SL foram prejudicados em virtude de erros introduzidos pelo algoritmo de *stemming*. Quando esses erros são corrigidos, a medida apresenta resultados com valores mais baixos fazendo com que os termos não sejam considerados como similares e, conseqüentemente, concordando com o analisador humano. Um extrato desses termos, com seu valor de similaridade corrigido, é apresentado na tabela 6.7.

Tabela 6.7: Resultados pertencentes ao grupo G4 com os erros de *stemming* corrigidos

EO-base	EO-alvo	CC	SL	SL corrigida	EO-base	EO-alvo	Stem corrigido	Stem corrigido
cassacao	causacao	0.88	0.76	0.73	cassaca	causaca	cassac	causac
condicao	conducao	0.88	0.76	0.73	condica	conduca	condic	conduc
deteccao	detencao	0.88	0.76	0.73	detecca	detenca	detecc	detenc
fundacao	fundicao	0.88	0.76	0.73	fundaca	fundica	fundac	fundic
traducao	tradicao	0.88	0.76	0.73	traduca	tradica	traduc	tradic

A tabela 6.7 apresenta os pares de termos que são considerados não similares pela medida SL após a correção dos seus respectivos *stems*, ao passo que a medida CC continua considerando os mesmos como similares.

Os casos mostrados na tabela 6.7 possuem uma característica em comum: a terminação em "`ção`". Pelo fato de os termos estarem sem acentuação e a letra "`ç`" ter sido substituída por "`c`", os termos que apresentam essa característica retornam do processo de *stemming* de forma incorreta. Esses erros constituem casos de *understemming*, pois o sufixo do radical não é completamente removido.

Além dos erros de *stemming*, a medida SL detectou similaridade entre termos não similares sem a utilização da heurística da primeira letra. Após a inserção dessa heurística, tem-se os resultados apresentados na tabela C.2 (ver apêndice C). A abreviatura SLH significa a medida SL com a heurística "primeira letra".

Na tabela C.2, termos considerados não similares pelo avaliador humano e que eram considerados similares pela medida SL, quando utilizada a heurística "primeira letra", passam a ser

classificados como não similares.

Finalmente, apresentamos na tabela C.3 (ver apêndice C) aqueles casos nos quais ambas as medidas classificam, como similares, termos considerados não similares pelo avaliador humano. Nesses casos, nem a heurística "primeira letra" faz com que os termos não sejam considerados similares pela medida SL. Assim, encontram-se na tabela C.3 casos onde o desempenho da medida SL é fraco. Pode-se observar que os termos são compostos por sete ou mais letras e possuem uma alteração em uma letra que não é a primeira.

6.3.2.2 Análise do Grupo G5

O grupo G5 apresenta o maior número de termos similares, 907 termos monopalavra e 1211 termos multipalavra. Isso pode ser justificado pelo fato de a medida CC ser menos restritiva do que a medida SL. Neste grupo encontram-se aqueles termos que não são considerados similares pelo analisador humano nem pela medida SL, e são considerados similares pela medida CC.

Devido à grande quantidade de termos, optamos por apresentar os resultados deste grupo resumidos a duas tabelas. A primeira, a tabela C.4 (ver apêndice C), apresenta um extrato com termos monopalavra, e a segunda, a tabela C.5 (ver apêndice C), é formada por termos multipalavra. Em ambas as tabelas é possível observar que o desempenho da medida SL é bastante satisfatório, uma vez que existe concordância com o analisador humano. Para os pares de termos que possuem significados distintos o valor de similaridade é inferior ao limiar (0.75) estabelecido. Além disso, esses termos são considerados similares pela medida CC.

No sentido de prover mapeamentos mais consistentes neste grupo, seria possível elevarmos o limiar para 0.8. Contudo, os resultados gerados pela medida CC ainda permaneceriam inconsistentes para um grande número de mapeamentos.

Os pares de termos apresentados nas tabelas C.4 e C.5, em sua maioria, possuem a mesma cadeia de caracteres em seu final ou em seu sufixo. No caso dos termos multipalavra, pelo menos uma das palavras que compõem o termo possui a mesma terminação. A medida CC atribui o mesmo peso para alterações tanto no radical quanto nos caracteres fora do radical do termo. Como na língua portuguesa, o sufixo de um termo tem uma baixa carga semântica, e a medida CC não leva esta característica em consideração, são gerados mapeamentos inconsistentes por esta medida.

Pelo fato de este grupo representar a maioria dos termos detectados como similares por, ao menos, uma das medidas, é possível questionar se a medida CC é realmente adequada para o tratamento de termos da língua portuguesa.

Acreditamos que, no tratamento de termos multipalavra, a medida SL apresenta melhor desempenho do que a medida CC devido ao fato de tratar as palavras constituintes do termo de forma individual.

6.3.2.3 Análise do Grupo G6

No grupo G6 encontram-se os casos em que os termos não são considerados similares pelo analisador humano nem pela medida CC, ao passo que a medida SL os considera similares. Os termos pertencentes a esse grupo são apresentados na tabela 6.8.

Tal como na tabela C.3, o leitor pode observar que os termos na tabela 6.8, em sua maioria são compostos por 7 ou mais letras e possuem DE=1, sendo que o caractere diferente não é a primeira letra do termo. Neste sentido, a aplicação da medida SL assemelha-se aos casos onde a DE também apresenta um desempenho reduzido pois, se considerarmos o valor 2 como limiar, os mesmos termos considerados similares pela medida SL são também considerados similares pela DE.

Tabela 6.8: Resultados pertencentes ao grupo G6

EO-base	EO-alvo	CC	SL	EO-base	EO-alvo
supercondutor	supercondutividade	0.46	0.80	supercondu	supercondut
semicondutor	semicondutividade	0.42	0.79	semicondu	semicondut
servicoDeInformacao	servicosDeInformacoes Turisticas	0.32	0.79	servdeinformaca	servdeinformaco
sistemaDeInformacao	sistemasDeInformacoes Contabeis	0.37	0.79	sistemdeinformaca	sistemdeinformaco
sistemaDeInformacoes Gerenciais	sistemasDeInformacao	0.35	0.79	sistemdeinformaco	sistemdeinformaca
grafotecnica	grafoteca	0.67	0.77	grafotecn	grafotec
penitenciaria	penitencia	0.70	0.77	penitenci	penitenc
comportamento	composto	0.25	0.76	comport	compost
conflitos	confeitaria	0.44	0.76	conflit	confeit
consenso	condensadores	0.25	0.76	consens	condens
contrabando	contratos	0.44	0.76	contrab	contrat
descarga	descarnamento	0.25	0.76	descarg	descarn
emprestimo	empreitada	0.60	0.76	emprest	empreit
emprestimo	empreiteiro	0.70	0.76	emprest	empreit
emprestimo	empresarios	0.60	0.76	emprest	empresa
metabolismo	metanol	0.29	0.76	metabol	metanol
produtoIndustrializado	produtividadeIndustrial	0.45	0.76	produtindustri	produtindustr
protesto	progesterona	0.38	0.76	protest	progest
quadrilha	quadril	0.71	0.76	quadrilh	quadril
transito	transistores	0.50	0.76	transit	transis
utilitarismo	utilitarios	0.73	0.76	utilitar	utilita

Entretanto, se alterarmos o limiar para 0.8, apenas o par de termos `supercondutor` e `supercondutividade` permaneceria sendo considerado similar. O mesmo se aplica à medida CC quando alterado o limiar para 0.8.

Neste grupo existe a ocorrência de termos multipalavra formados por quantidade diferente de palavras (conforme a tabela 6.9), que foram considerados similares pela medida SL mas não similares pelo analisador humano.

Tabela 6.9: Termos multipalavra com número de palavras diferente pertencentes ao grupo G6

EO-base	EO-alvo	CC	SL
servicoDeInformacao	servicosDeInformacoesTuristicas	0.32	0.79
sistemaDeInformacao	sistemasDeInformacoesContabeis	0.37	0.79
sistemaDeInformacoesGerenciais	sistemasDeInformacao	0.35	0.79

Os termos apresentados na tabela 6.9 caracterizam uma relação semântica de hiponímia na qual, por exemplo, o termo `servicosDeInformacoesTuristicas` é um hipônimo de `servicoDeInformacao`. A medida SL contempla similaridade entre termos com quantidade distinta de palavras, entretanto o analisador humano não considerou tais casos como similares. Esse tipo de mapeamento, entre termos com quantidade distinta de palavras, pode ser útil quando se deseja realizar a união de EOs. Para isso, o termo com maior número de palavras é mapeado como hipônimo do termo com menor número de palavras.

6.3.3 Análise do Grupo G7

O grupo G7 representa todos os casos nos quais o analisador humano não conseguiu avaliar os termos como similares ou não similares, ou seja, o ser humano tem dúvida quanto à similaridade dos termos. Optamos por reunir todas as três combinações possíveis entre as medidas de similaridade em um único grupo, pelo fato de que, se o ser humano não possui clareza suficiente para avaliar a similaridade, não se pode exigir que uma medida de similaridade o faça. Dessa forma, torna-se uma tarefa complexa tecer afirmações quanto à precisão das medidas de similaridade nesse contexto em que nem o avaliador humano está seguro. No total, 290 termos compõem esse grupo, e um extrato dos termos aí contidos é mostrado na tabela C.6 (ver apêndice C).

6.3.4 Análise dos Casos Considerados Similares Utilizando a Medida SL

Nesta análise reunimos todos os casos (exceto aqueles nos quais o analisador humano possui dúvida) que a medida SL considera similares, ou seja, para os quais a medida apresenta valor de similaridade ≥ 0.75. Dessa forma, encontram-se nesta análise os termos pertencentes aos grupos G1, G3, G4 e G6.

No total, 122 termos compõem esta análise. Destes, 44 são considerados similares pelo analisador humano e pela medida SL, conforme apresentados na tabela 6.10.

Dos 44 termos, 25 são também considerados similares pela medida CC. Isso indica que 36% dos termos considerados similares pelo analisador humano são também considerados similares pela medida SL, ao passo que a medida CC alcança 20% de termos nessa situação.

Os pares de termos que não foram considerados similares pelo analisador humano somam 78. Destes 78, 14 termos apresentam erros de *stemming* que, se corrigidos, fazem com que os termos passem a ser tratados como não similares pela medida SL. Além disso, quando aplicada a heurística "primeira letra" à medida SL, mais 23 termos passam a não ser considerados similares. Assim, se corrigidos os erros de *stemming* e aplicada a heurística "primeira letra", a medida SL passa a considerar não similares 37 termos, o que representa 47% de concordância com o analisador humano. Os outros 41 pares de termos restantes, são considerados similares pela medida SL e não similares pelo analisador humano. Entretanto, se elevarmos o limiar para 0.8, ou seja, um aumento de 0.05, o número de pares de termos que o analisador humano não considera similares e a medida SL considera similares é reduzido para seis. Portanto, cabe ao usuário de EOs estabelecer o limiar mais adequado para seu caso.

No conjunto dos 41 pares de termos considerados similares pela medida SL e não similares pelo analisador humano, observa-se que a medida SL apresenta um desempenho mais fraco. Esses casos assemelham-se os casos do grupo G6, já analisados na seção 6.3.2.3.

Por outro lado, se analisarmos o desempenho da medida CC sobre os mesmos 41 termos, é possível verificar que 23 são considerados similares, ou seja, a medida CC apresenta mais de 50% de discordância do analisador humano.

Este fato explicita a dificuldade de se chegar a um acordo com a análise humana. Ao mesmo tempo, pode-se perceber que a medida CC também não apresenta um desempenho satisfatório nos 122 pares de termos analisados neste experimento.

6.3.5 Uma Revisão da Análise Humana

Devido à grande quantidade de termos analisados por cada avaliador humano, considera-se normal que alguns resultados tenham sido avaliados de maneira equivocada. Entretanto, o autor desse trabalho não poderia atuar como revisor da análise humana, pois seria difícil uma revisão

Tabela 6.10: Termos considerados similares pela medida SL e pelo analisador humano

EO-base	EO-alvo	CC	SL	EO-base	EO-alvo
criacaoDeBicho-da-seda	criacaoDeBichos-da-seda	0.95	0.82	criacadebicho-da-sed	criacadebichos-da-sed
energiaHidroeletrica	energiaHidreletrica	0.95	0.79	energhidroeletr	energhidreletr
maquina-ferramenta	maquinas-ferramenta	0.94	0.84	maquina-ferrament	maquinas-ferrament
teleinformatica	telinformatica	0.93	0.80	teleinforma	telinforma
eutroficacao	eutrofizacao	0.92	0.81	eutroficaca	eutrofizaca
bovinocultura	ovinocultura	0.92	0.79	bovinocult	ovinocult
elegibilidade	legibilidade	0.92	0.76	elegibil	legibil
interferon	interferons	0.90	0.80	interferon	interferom
lexiologia	lexicologia	0.90	0.77	lexiolog	lexicolog
tituloMobiliario	tituloImobiliario	0.88	0.76	titulmobilia	titulimobilia
redeDeTelecomunicacao	redeDeTelecomunicacoes	0.86	0.83	reddetelecomunicaca	reddetelecomunicaco
dinamicaDePopulacao	dinamicaDePopulacoes	0.84	0.77	dinamdepopulaca	dinamdepopulaco
usinaHidroeletrica	usinasHidreletricas	0.83	0.79	usinhidroeletr	usinhidreletr
condicoesDeTrabalho	condicaoDeTrabalho	0.83	0.76	condicodetrabalh	condicadetrabalh
pedraSemipreciosa	pedrasSemi-preciosas	0.82	0.81	pedrsemiprecios	pedrsemi-precios
cafeicultor	cafeicultura	0.82	0.77	cafeicul	cafeicult
instituicaoReligiosa	instituicoesReligiosas	0.80	0.80	instituicareligi	instituicoreligi
cartografia	cartograma	0.80	0.79	cartograf	cartogram
agricultura	agricultor	0.80	0.76	agricult	agricul
inovacaoTecnologica	inovacoesTecnologicas	0.79	0.76	inovacatecnolog	inovacotecnolog
instalacaoAgricola	instalacoesAgricolas	0.78	0.79	instalacaagricol	instalacoagricol
instalacaoEletrica	instalacoesEletricas	0.78	0.79	instalacaeletr	instalacoeletr
sistemaDeInformacaoContabil	sistemasDeInformacoesContabeis	0.78	0.77	sistemdeinformacacontavel	sistemdeinformacocontabel
aglomeracaoUrbana	aglomeracoesUrbanas	0.76	0.80	aglomeracaurban	aglomeracourban
epigrafia	epigrama	0.75	0.76	epigraf	epigram
comunicacao	comunicacoes	0.73	0.80	comunicaca	comunicaco
organizacoes	organizacao	0.73	0.80	organizaco	organizaca
bibliotecario	bibliotecas	0.73	0.79	biblioteca	bibliotec
edificacao	edificacoes	0.70	0.79	edificaca	edificaco
construcaoRural	construcoesRurais	0.67	0.79	construcarural	construcorural
escorpiao	escorpioes	0.67	0.77	escorpia	escorpio
obrigacao	obrigacoes	0.67	0.77	obrigaca	obrigaco
comerciario	comerciantes	0.64	0.76	comercia	comerci
cognicao	cognicoes	0.63	0.76	cognica	cognico
comerciario	comercio	0.63	0.76	comercia	comerci
comissao	comissoes	0.63	0.76	comissa	comisso
fundacao	fundacoes	0.63	0.76	fundaca	fundaco
aeronauticaMilitar	aeronavesMilitares	0.61	0.76	aeronaumilit	aeronavmilit
combustivel	combustao	0.56	0.77	combusti	combusta
fotografo	fotogravura	0.56	0.77	fotograf	fotograv
existencialismo	existencia	0.50	0.77	existenci	existenc
aeronautica	aeronaves	0.44	0.76	aeronau	aeronav
construtor	construtivismo	0.40	0.76	constru	construt
bicho-da-seda,Criacao	bichos-da-seda	0.36	0.82	bicho-da-sed	bichos-da-sed

imparcial. Dessa forma, foi solicitado a um revisor humano a tarefa de revisar a análise feita pelos dois primeiros avaliadores humanos.

O revisor humano recebeu todos os pares de termos analisados pelos avaliadores humanos. Sua tarefa era revisar essa avaliação. O objetivo dessa revisão é verificar se existe discordância

entre o revisor e os analisadores humanos. Essa discordância pode-se dar de duas formas:

1. se o par de termos considerado similar pelo analisador humano não é considerado similar pelo revisor;
2. se o par de termos considerado não similar pelo analisador humano é considerado similar pelo revisor.

No primeiro caso, onde o par de termos é considerado similar pelo analisador humano e não é considerado similar pelo revisor, foram detectadas 26 ocorrências para os termos monopalavra, conforme tabela C.7, e 106 ocorrências para os termos multipalavra, conforme as tabelas C.8 e C.9 (ver apêndice C). Esses 132 termos representam 4,5% dos 2887 desta fase de avaliação.

O maior número de termos pertencente ao conjunto de termos multipalavra é um indício de que os avaliadores julgaram termos similares apoiados sobre uma das palavras de um termo multipalavra, ou seja, se uma palavra de um termo multipalavra é similar a uma palavra do outro termo, o avaliador humano considerou os termos como similares.

Tanto para os termos monopalavra quanto para os termos multipalavra, todos os 132 termos considerados como não similares pelo revisor também são tratados como não similares pela medida SL e como similares pela medida CC (em ambos os casos considerando o limiar 0.75).

O revisor não encontrou nenhuma ocorrência de termos no segundo caso. Ou seja, os termos considerados não similares pelo analisador humano realmente não são similares de acordo com o revisor.

O revisor humano também manifestou seu posicionamento quanto aos termos em que os analisadores apontaram dúvida. Entretanto, optamos por não analisar esses casos pois, como já justificado, não são casos que os avaliadores têm segurança para analisar, embora os resultados do revisor tenham ido ao encontro dos valores apresentados pela medida SL.

Após a análise detalhada da medida SL em nível lexical, apresentamos na próxima seção os experimentos realizados em nível semântico-estrutural.

6.4 Nível Semântico-Estrutural

Em ambas as EOs estudadas nesta dissertação, a declaração de um termo é composta por outros termos que estão vinculados ao termo principal através de relações semânticas. A figura 6.1 apresenta um extrato da EO-base utilizada em nosso estudo.

```
<T ‘‘tribunal’’>
<SN ‘‘corte’’/>
<SN ‘‘corteDeJustica’’/>
<BT ‘‘sistemaJudiciario’’/>
<NT ‘‘conselhoDeGuerra’’/>
<NT ‘‘tribunalAdministrativo’’/>
<NT ‘‘tribunalMilitar’’/>
</T>
```

Figura 6.1: Exemplo de extrato da EO-base

A figura 6.1 utiliza a sintaxe XML e declara o termo `tribunal` sendo formado pelos sinônimos `corte` e `corteDeJustica`. A relação de sinonímia se estabelece pela declaração do elemento `SN`[2].

[2]Do inglês, Synonym.

O termo `tribunal` possui o termo mais amplo `sistemaJudiciario`, declarado pelo elemento `BT`[3]. Na formação de uma hierarquia, o termo `sistemaJudiciario` é chamado termo pai. Finalmente, os termos mais específicos `conselhoDeGuerra`, `tribunalAdministrativo` e `tribunalMilitar` são declarados pelo elemento `NT`[4]. Esses termos são chamados termos-filhos, em uma hierarquia.

Inicialmente, em nível semântico-estrutural, fizemos uso das relações BT e NT e aplicamos a heurística "ancestral e descendentes". Esse experimento é apresentado na próxima seção.

6.4.1 Heurística "Ancestral e Descendentes" aplicada a EOs da Língua Portuguesa

A heurística "ancestral e descendentes" utilizada nas EOs da língua inglesa também foi aplicada às EOs da língua portuguesa. A única alteração à mesma foi a inclusão da relação semântica de sinonímia. Tanto o termo da EO-base sendo mapeado, quanto seu termo sinônimo (quando existente), são considerados na detecção de similaridade com outros termos ou sinônimos na EO-alvo. Pelo fato de, neste nível, estarmos trabalhando com termos monopalavra e termos multipalavra, alteramos o índice k. Esse índice, que nos experimentos em nível lexical representava a quantidade de palavras do termo de menor comprimento, agora passa a representar a quantidade de palavras nos termos. Desse modo, a medida SL contempla a similaridade entre termos com o mesmo número de palavras. Neste experimento, é importante ressaltar que o algoritmo só aplica a heurística "ancestral e descendentes" se não existir nenhum termo ou sinônimo deste termo que lhe seja similar em nível lexical.

Realizamos dois experimentos, cada um utilizando uma medida de similaridade. No primeiro, utilizando a medida SL com a restrição $0.75 \leq SL < 1$, nenhum par de termos foi encontrado com a aplicação da heurística. No segundo experimento, com a restrição $0.75 \leq CC < 1$, apenas um par de termos foi encontrado, conforme a tabela 6.11.

Tabela 6.11: Resultado da aplicação da heurística ancestral e descendentes utilizando a medida CC

<T "artropodes"/>	<T "arthropoda"/>
<BT "invertebrado"/>	<BT "invertebrados"/>
<NT "aracnideo"/>	<NT "arachnida"/>
<NT "crustaceo"/>	<NT "crustacea"/>
<NT "inseto"/>	<NT "entomostraca"/>
</T>	<NT "malacostraca"/>
	<NT "myriapoda"/>
	<NT "merostomata"/>
	<NT "pentastomida"/>
	<NT "pycnogonida"/>
	<NT "symphyla"/>
	</T>

Na tabela 6.11 constam os termos `artropodes` e `arthropoda`, que possuem similaridade lexical menor do que 0.75, e puderam ser detectados com a utilização da heurística "ancestral e descendentes", pois seus ancestrais do primeiro nível (a saber, `invertebrado` e `invertebrados`) são similares, e ao menos um dos descendentes do primeiro nível, neste caso `crustaceo` e `crustacea`, também são similares.

[3]Do inglês, Broader Term.
[4]Do inglês, Narrower Term.

O fato de a heurística "ancestral e descendentes" produzir somente um mapeamento, utilizando a medida CC, destoa da expectativa inicial, pois esperava-se que essa heurística pudesse encontrar mais termos com o mesmo significado, porém com baixa similaridade lexical.

Diante desse fato, buscamos outra alternativa para detectar termos similares que possuam baixa similaridade lexical. Para isso, exploramos a relação de sinonímia. Os experimentos são apresentados na próxima seção.

6.4.2 Experimentos com a Relação Semântica de Sinonímia

Assim como no trabalho de Rodríguez e Egenhofer [70], que utiliza conjuntos de sinônimos, também fizemos uso da relação semântica de sinonímia. Neste momento é importante explicitar que as EOs da língua inglesa, tal como descritas, não ofereciam mecanismos para recuperar explicitamente a relação de sinonímia.

Foram realizados diversos experimentos com foco nessa relação. Aplicamos as duas medidas de similaridade CC e SL, utilizamos os limiares 0.75 e 0.8, combinamos as medidas de similaridade, considerando os casos em que um mapeamento é gerado por uma medida de similaridade e não é gerado pela outra. Além disso, realizamos experimentos considerando o valor de similaridade 1, no caso da medida CC, o que representa uma combinação perfeita das cadeias de caracteres sendo comparadas e, no caso da medida SL, indicando um alto grau de similaridade, mas não necessariamente uma combinação perfeita das cadeias de caracteres. Todas essas variáveis foram testadas e os resultados analisados de forma que pudéssemos apresentar ao leitor uma análise dos dados. A seguir apresentamos os principais resultados dos experimentos com foco na relação semântica de sinonímia.

6.4.2.1 Experimento com Sobreposição das EOs

A realização deste experimento tem por objetivo apresentar ao leitor a quantidade total de termos similares entre as EOs utilizadas neste trabalho. Em outras palavras, buscamos apresentar a sobreposição existente entre os termos das EOs. A idéia é verificar a porcentagem de termos que a EO-base possui como similares na EO-alvo. Este experimento, ao contrário dos anteriores, considera similaridade igual a 1, tanto pela medida CC quanto pela medida SL.

Nosso experimento foi conduzido da seguinte maneira: um mapeamento entre termos é realizado se um termo ou seu sinônimo, na EO-base, possui um termo similar (de acordo com alguma medida de similaridade), na EO-alvo. Esse termo na EO-alvo também pode estar declarado como um sinônimo. O limiar utilizado foi 0.75.

Tabela 6.12: Quantidade de termos similares das EOs

	EO-base	EO-alvo
Quantidade total de termos	8912	21859
Quantidade de termos similares pela medida CC	6009	
Quantidade de termos similares pela medida SL	5495	

De acordo com a tabela 6.12, utilizando a medida CC, encontram-se 67,40% dos termos da EO-base que possuem um correspondente na EO-alvo, ao passo que, utilizando a medida SL encontram-se 61,64% de termos correspondentes na EO-alvo.

Quando o valor de similaridade entre dois termos é igual a 1, não se pode fazer uma análise comparativa da medida CC com a medida SL. Adicionalmente, uma análise caso a caso dos mapeamentos gerados é tarefa bastante longa e tediosa. Nosso objetivo, além de detectar a máxima

similaridade entre termos pertencentes a EOs, é verificar o grau de precisão das medidas de similaridade, quando aplicadas a casos que não possuem uma combinação perfeita de caracteres, em nível lexical. Diante disso, realizamos mais experimentos, de modo a refinar os resultados gerados até então.

6.4.2.2 Comparação das Medidas CC e SL Utilizando a Relação de Sinonímia

Para compararmos as medidas CC e SL utilizando a relação semântica de sinonímia, realizamos dois experimentos. Retomamos o exemplo apresentado na 6.1, na qual o elemento `tribunal` é declarado com o elemento XML `T`. O objetivo dessa comparação é verificar a similaridade entre termos declarados com esse elemento e que possuam baixa similaridade lexical, mas com significado próximo. Assim, buscamos detectar similaridade entre os termos declarados com a relação semântica de sinonímia de modo a prover um mapeamento entre os termos declarados com o elemento `T`.

As condições para a realização dos dois experimentos apresentados nesta subseção são descritas a seguir:

- os termos declarados com o elemento `T` nas EOs não podem ser similares de acordo com os limiares estabelecidos às medidas de similaridade;
- a similaridade deve ocorrer:
  - entre um termo declarado com o elemento `T` na EO-base e um elemento `SN` da EO-alvo ou;
  - entre um termo declarado com o elemento `SN` da EO-base e o elemento `T` na EO-alvo ou;
  - entre um termo declarado com o elemento `SN` com outro termo declarado no elemento `SN`.

6.4.2.2.1 Experimento SN com a medida CC

Devido aos diversos resultados inconsistentes encontrados em nível lexical com o limiar 0.75, consideramos neste experimento o limiar 0.8. Não são considerados os casos onde existe combinação perfeita de caracteres, pois ambas as medidas tratam da mesma forma estes casos. Finalmente, verificamos se os termos são similares pela medida CC e não similares pela medida SL, considerando o limiar 0.8.

Este experimento gerou 425 mapeamentos. Para analisar esses resultados focalizamos aqueles mapeamentos considerados corretos e incorretos. O mapeamento considerado correto é apresentado na tabela 6.13.

A tabela 6.13 apresenta apenas um mapeamento considerado correto. A grande quantidade de mapeamentos incorretos é mais um indício de que a medida CC é inadequada para o tratamento de termos da língua portuguesa. Um extrato desses mapeamentos incorretos é apresentado na tabela D.1.

Na análise da tabela D.1 pode-se observar que a inconsistência de diversos mapeamentos ocorreu devido às palavras de comprimento (≤ 6) que são tratadas de forma equivocada pela medida CC. Alguns desses mapeamentos são apresentados na tabela 6.14.

É possível notar que os termos mostrados na tabela 6.14 apresentam somente uma letra diferente entre seus conjuntos de caracteres. Nestes casos, a medida CC assume valor maior ou igual a 0.8 detectando similaridade entre termos com significados distintos.

Tabela 6.13: Mapeamento gerado de forma correta com a medida CC utilizando a relação semântica de sinonímia: caso não mapeado pela medida SL (Limiar 0.8)

EO-base	EO-alvo
<T "atoProcessual"> <SN "atoDoProcesso"> <SN "**atoJudicial**"> <SN "atoJudiciario"> </T>	<T "**acaoJudicial**"> <BT "acoes"> </T>

Tabela 6.14: Mapeamentos inconsistentes gerados pela medida CC com limiar 0.8 utilizando a relação semântica de sinonímia (Termos que possuem comprimento ≤ 6)

EO-base	EO-alvo
solar	molar
limao	licao
pasto	parto
coito	conto
cafes	cames
cosmos	colmos

6.4.2.2.2 Experimento SN com a medida SL

De modo análogo ao experimento anterior, procedemos com a medida SL. Neste caso consideramos o limiar 0.75. Neste experimento também não são considerados os casos onde existe uma combinação perfeita das cadeias de caracteres sendo comparadas, entretanto termos que apresentam valor de similaridade igual a um, mas que não apresentam uma combinação perfeita das cadeias de caracteres, podem ser mapeados. Foram encontrados 106 mapeamentos. Todos esses mapeamentos não são detectados pela medida CC utilizando o mesmo limiar. Na análise desses resultados foram encontrados tanto mapeamentos corretos quanto incorretos.

A tabela D.2 (ver apêndice D) apresenta um extrato dos resultados considerados corretos utilizando a medida SL. Na análise desse extrato e dos mapeamentos restantes também considerados como corretos não se consegue detectar nenhum padrão para os casos mapeados como similares, exceto os casos de termos com variação de número, já analisados anteriormente.

A tabela D.3 (ver apêndice D) apresenta os resultados considerados incorretos pela medida SL. É importante lembrar que esses resultados não são mapeados pela medida CC quando utilizado o limiar 0.75. De fato, não existe uma característica única nos termos na tabela D.3 (ver apêndice D). Pode-se perceber que os termos com o mesmo radical, mas com significados distintos, tal como `coqueiro` e `coque`, são considerados similares pela medida SL.

Outros casos, como a similaridade dos termos `protesto` e `progesterona` (ver tabela D.3), caracterizam os radicais compostos por sete ou mais letras que possuem uma letra diferente, a qual não é a primeira.

6.5 Análise Crítica

Primeiramente, é importante destacar que cada par de termos detectado como similar pelas medidas de similaridade foi avaliado por um avaliador humano. Certamente, se apresentarmos os mesmos pares de termos para outros seres humanos, com outro conhecimento de mundo, teremos outros resultados. Entretanto, no contexto de uma dissertação de mestrado e dos recursos disponíveis, acreditamos que tenha sido possível realizar uma avaliação adequada a este tipo de pesquisa.

Durante a análise dos grupos, e através das tabelas apresentadas, foi possível observar as situações nas quais a medida SL apresenta melhor ou pior desempenho. Para os casos que apresentam erros do algoritmo de *stemming* (por exemplo, o *stem* `conduca` para o termo `condução` e o *stem* `traduca` para o termo `tradução`), a solução pode estar na alteração deste algoritmo.

Uma possível variação a ser aplicada à medida SL é alterar os valores, a saber: 0.1 e 0.2, das penalidades impostas às alterações nos radicais. Acreditamos que, se esses valores forem incrementados, termos similares podem não ser mapeados pela medida SL, ao passo que se esses valores forem decrementados, termos não similares podem ser considerados como similares. Observamos que a eliminação dessas penalidades já foi testada na fase de validação, gerando inúmeros resultados inconsistentes.

Experimentos com a aplicação da heurística da primeira letra para a medida CC também poderiam ter sido realizados; entretanto, devido a o foco deste trabalho ser a medida SL, não executamos todas as variações possíveis com a medida CC.

É importante destacar que foram realizados diversos experimentos além daqueles descritos neste capítulo. Devido à grande quantidade de dados resultantes, procurou-se refinar os mesmos de modo que fosse possível apresentar resultados de forma explícita para o leitor.

Um fato considerado relevante de ser mencionado foram os termos que, aparentemente, de acordo com o revisor humano, não são similares e que o avaliador humano considerou como similares. Isso é um indício de que o conceito de similaridade seja bastante subjetivo, demonstrando a dificuldade de se avaliar medidas de similaridade.

Em nível semântico-estrutural exploramos a relação de sinonímia e obtivemos mapeamentos entre os termos com baixa similaridade lexical, mas com mesmo significado. Procuramos apresentar tanto os mapeamentos corretos quanto os incorretos de modo a explicitar os pontos positivos e negativos das medidas de similaridade. Pôde-se notar que ambas as medidas apresentam resultados inconsistentes demonstrando que, mesmo com a utilização de relação semântica de sinonímia, o problema de mapear termos similares entre EOs ainda não está completamente solucionado.

Faz-se necessária uma análise mais profunda dos resultados gerados, de modo a realizar alterações na medida SL para que a mesma gere resultados mais consistentes.

Finalmente, gostaríamos de apresentar ao leitor dois casos que demonstram a complexidade de se detectar termos similares na língua portuguesa, conforme a tabela 6.15.

O primeiro caso apresentado na tabela 6.15 envolve o mapeamento entre os termos `equino` e `cavalinha` detectado como similares de forma equivocada pela medida SL. Esses termos, se colocados ao julgamento humano, talvez fossem considerados como similares. Entretanto, nas EOs estão em contextos distintos.

O segundo caso é um mapeamento detectado de modo correto pela medida CC entre os termos `metro` e `transporteFerroviario`. Esses termos são detectados como o auxílio da relação de sinonímia, pois os termos `transporteMetroviario` e `transporteFerroviario` são similares. Neste caso, apenas uma letra é diferente, e essa letra, ao contrário dos casos apresentados até aqui, é a primeira. Por ser a primeira letra diferente, os termos `transporteMetroviario` e

Tabela 6.15: Curiosidades

EO-base	EO-alvo
<T "equino"> <SN "**cavalo**"> <SN "equideo"> <BT "animalDomestico"> </T>	<T "**cavalinha**"> <BT "pisciculturaMarinha"> </T>
<T "metro"> <SN "metropolitano"> <SN "**transporteMetroviario**"> <SN "tremMetropolitano"> <BT "transporteDeMassa"> <NT "segurancaMetroviaria"> </T>	<T "**transporteFerroviario**"> <BT "transportes"> </T>

`transporteFerroviario` não são considerados similares pela medida SL.

No conjunto de termos que trabalhamos, consideramos a heurística da letra inicial bastante produtiva, pois evitou diversos mapeamentos considerados errôneos, conforme apresentados na fase de validação.

6.6 Considerações sobre o Capítulo no Contexto da Dissertação

Neste capítulo foi descrita a fase de avaliação da medida SL, na qual os resultados apresentados levam em consideração a avaliação humana. Devido ao grande número de dados, a análise dos resultados foi dividida em grupos. Procurou-se apresentar ao leitor o comportamento da medida SL sob diversas condições, de forma que a avaliação pudesse ser mais completa.

Além disso, apresentamos mais dois experimentos realizados levando em consideração a relação semântica de sinonímia.

Uma consideração importante, neste momento, é destacar a grande diversidade de termos pertencentes a domínios distintos sobre os quais foi aplicada a medida SL. Este fato introduz uma dificuldade para se fazer uma análise quantitativa dos dados levando em consideração o domínio no qual pertence cada termo. No contexto do ambiente de trabalho de um especialista do domínio que deseje mapear EOs do mesmo domínio, acreditamos que a medida SL possa obter um desempenho melhor.

O próximo capítulo apresenta a conclusão deste trabalho, incluindo as limitações do mesmo, bem como trabalhos futuros e considerações finais.

Capítulo 7

Conclusão

"O rio atinge seus objetivos porque aprendeu a contornar obstáculos".
Lao-Tsé, filósofo taoísta (604 a 527 a. C.)

7.1 Sobre este Trabalho

Esta dissertação buscou identificar termos similares entre EOs projetadas independentemente, de modo a prover um mapeamento consistente entre as mesmas. Assim como alguns dos trabalhos correlatos apresentados, nosso trabalho fez uso de medidas de similaridade. Essas medidas foram, inicialmente, aplicadas a EOs da língua inglesa, para as quais foi desenvolvido um protótipo com o objetivo de auxiliar o usuário a realizar o mapeamento entre os termos das mesmas. Para desenvolver este protótipo foram estudadas diferentes linguagens de marcação semântica. Esse protótipo abstrai para o usuário as diferenças entre as sintaxes das linguagens de marcação semântica que têm sido utilizadas atualmente, e apresenta os conceitos de forma hierárquica. Adicionalmente, o usuário pode escolher as medidas de similaridade, bem como os limiares que deseja trabalhar, até encontrar os mapeamentos mais consistentes. Em nível semântico estrutural foram propostas duas heurísticas que permitiram detectar mapeamentos entre termos com mesmo significado, mas com baixa similaridade lexical.

Em seguida descrevemos o tratamento das EOs da língua portuguesa, para as quais aplicamos a mesma abordagem utilizada com as EOs da língua inglesa. Entretanto, ao aplicarmos a medida CC para os termos da língua portuguesa, surgiram diversos resultados inconsistentes, motivando-nos a desenvolver nossa própria medida de similaridade.

Por se tratar de uma proposta nova, foi necessário validar e avaliar a medida SL. Desse modo, nosso trabalho consistiu de uma fase de validação da medida SL, na qual foi possível analisar resultados preliminares e refinar a medida para a fase seguinte de avaliação.

A avaliação da medida SL ocorreu em um conjunto de termos diferente do utilizado na fase de validação. A esses termos também aplicamos a medida CC, para efeito de comparação. Todos aqueles pares de termos que foram identificados como similares pela medida SL ou pela medida CC, considerando o limiar 0.75, foram entregues a dois avaliadores humanos, sendo que cada um avaliou a similaridade de uma metade do conjunto total de termos.

A avaliação dos resultados foi apresentada no capítulo 6, no qual procuramos descrever os casos em que a medida SL apresenta um comportamento consistente ou satisfatório, bem como os casos em que a mesma gera resultados inconsistentes. Nossas conclusões nesta etapa do trabalho foram feitas sempre à luz do que o avaliador humano considerou como termos similares. Além disso, solicitamos a outro ser humano a tarefa de revisar a análise humana realizada inicialmente.

Essa revisão nos permitiu fazer mais algumas considerações sobre a medida SL.

Neste momento é importante mencionar que nós percebemos que a tarefa do revisor só tem sentido quando este possui mais conhecimento do domínio em relação ao avaliador, caso contrário acreditamos que o papel do revisor não possua grande contribuição para a análise dos resultados.

Em nível semântico-estrutural foi possível utilizar a relação de sinonímia para auxiliar a detecção de termos similares.

Como contribuição desta dissertação, pôde-se concluir que a medida CC não parece adequada para tratar termos da língua portuguesa. Para os casos da língua inglesa, não foram testadas quantidades de termos suficientes, por isso não podemos fazer nenhuma afirmação precipitada. Além disso, deixamos uma medida de similaridade nova na literatura, experimentos, resultados e análise da mesma para futuras referências. Uma das principais aplicações dessa medida de similaridade é como uma primeira etapa no processo de mapeamento entre termos pertencentes a EOs distintas. Esse mapeamento pode-se dar em diversos contextos, tais como, em agentes de software e sistemas de RI, entre outros.

Ao retomarmos a hipótese descrita no início deste trabalho, podemos confirmá-la quando aplicamos as medidas de similaridade e detectamos termos similares entre as EOs estudadas. Destacamos o desempenho consistente da medida SL para os termos com característica de variação de número, freqüentemente encontrados em EOs.

Finalmente, o leitor pôde constatar que o problema de se encontrar termos similares na língua portuguesa ainda é uma questão de pesquisa que não está completamente respondida e nem solucionada. Acreditamos que, com esta dissertação, avançamos ao apresentar uma alternativa de solução, por meio de uma medida de similaridade, e ao avaliá-la, fornecemos informações nas quais pode-se concluir onde seu desempenho é melhor ou pior.

7.2 Limitações

Dentro das limitações desta dissertação pode-se incluir a carência de EOs da língua inglesa pertencentes ao mesmo domínio de conhecimento. No caso da língua portuguesa, ainda não se conta com um número suficiente de EOs, codificadas com linguagens de marcação semântica, que permita a realização de uma pesquisa.

Outra limitação está relacionada ao protótipo desenvolvido para a língua inglesa, uma vez que esse trata somente os conceitos das EOs. A parte referente às propriedades dos conceitos está parcialmente implementada.

Quanto a medida SL, pode-se observar que em alguns casos a detecção de similaridade foi prejudicada devido ao desempenho do algoritmo de *stemming*. Esse fato não permitiu que a medida tivesse um melhor desempenho.

Ainda em relação às EOs da língua portuguesa, o tratamento do nível semântico-estrutural requer heurísticas mais eficientes, de forma que seja possível identificar todos os termos ancestrais e descendentes de um termo na hierarquia, e não somente o ancestral de nível anterior e os descendentes de primeiro nível.

A medida SL ainda não faz um tratamento que possa ser considerado eficiente para pares de termos formados com número diferente de palavras. Além disso, essas palavras devem estar na mesma posição no termo. O par de termos `industriaAlimenticia` e `agroindustriaDeAlimento`, por exemplo, é um caso de termos similares com mesmo significado que não são detectados como similares pela medida SL. Entretanto, as EOs que trabalhamos são compostas por relações de sinonímia, e quando a declaração de um termo é completa, ou seja, todos seus sinônimos são declarados, a medida SL pode detectar esses termos como similares, pois eles estarão declarados em ambas as EOs.

7.3 Trabalhos Futuros

Diante dos estudos realizados e dos resultados obtidos deixamos como continuidade os seguintes trabalhos futuros:

- o tratamento do formato de saída do protótipo desenvolvido para mapeamento de EOs da língua inglesa. Pode-se representar os mapeamentos resultantes do processamento do algoritmo, que inclui as medidas de similaridade e as heurísticas, nas linguagens de marcação semântica estudadas;

- a comparação da medida SL com outras medidas de similaridade a serem buscadas na literatura;

- a aplicação da medida SL em outros conjuntos de termos da língua portuguesa. Uma alternativa é a utilização somente de termos de um domínio específico do conhecimento em ambas as EOs;

- a utilização da medida SL em outros idiomas, tais como, espanhol e inglês. Quando aplicada a outros idiomas esta medida deve utilizar um algoritmo de *stemming* próprio de cada língua;

- a aplicação da medida SL para auxiliar um sistema de RI que utilize EOs, ou ainda, diretamente no sistema de RI para verificar a similaridade entre um termo consultado e os termos contidos nos documentos.

- a utilização de outros algoritmos de *stemming* para a língua portuguesa. A partir desse trabalho é possível realizar uma comparação do desempenho da medida SL com base em diferentes algoritmos de *stemming*;

- no contexto do sistemas multi-agentes os agentes de software que atuam sobre EOs precisam detectar termos similares para prover um mapeamento adequado e consistente entre esses termos. Neste sentido, a medida SL pode auxiliar um agente de software a encontrar termos similares entre EOs;

- a união das duas EOs da língua portuguesa. Por meio dessa abordagem é possível criar uma terceira EO que irá representar as duas EOs. Essa terceira EO poderá ser utilizada em sistemas de RI para auxiliar a tarefa de expansão de consultas;

- na área de Bancos de Dados, nossa medida de similaridade pode ser aplicada como uma primeira etapa no processo de integração de esquemas.

7.4 Considerações Finais

Durante o desenvolvimento deste trabalho obtivemos como resultado a produção e aprovação de 5 artigos em eventos científicos:

- Chaves, M. S.; Strube de Lima, V. L. Similaridade entre Estruturas Ontológicas. *XVI Brazilian Symposium on Computer Graphics and Image Processing - (SIBGRAPI). I Workshop em Tecnologia da Informação e Linguagem Humana*, São Carlos-SP, Brasil, 12 de Outubro de 2003.

Este artigo descreve o enfoque inicial de nosso estudo com os experimentos com EOs da língua inglesa. Na apresentação, além dos resultados obtidos com a língua inglesa, incluímos os experimentos com as EOs da língua portuguesa realizados até a fase de validação.

- Chaves, M. S.; Strube de Lima, V. L. Looking for Similarity among Ontological Structures. Technical Report, Departamento de Informática, Faculdade de Ciências da Universidade de Lisboa (DI-FCUL) TR-03-28, p. 15-18. *Tagging and Shallow Processing of Portuguese: Workshop notes of TASHA'2003*. António Branco, Amália Mendes e Ricardo Ribeiro (Eds.) Lisboa, Portugal, 2003.

  Este artigo é voltado aos experimentos realizados com a língua portuguesa. Nele são apresentados os resultados da fase de validação da medida SL sem a introdução das penalidades para as alterações nos radicais das palavras. Este artigo foi publicado como um relatório técnico. O mesmo, ao contrário dos outros três, não foi apresentado no evento.

- Chaves, M. S.; Strube de Lima, V. L. Looking for Similarity between Portuguese Ontological Structures. In. António Branco, Amália Mendes, Ricardo Ribeiro (Eds.). Edições Colibri, Lisboa, Portugal, 2004. (No prelo)

  Este artigo é uma versão expandida daquele descrito no item anterior. Aqui, nós apresentamos a medida SL com as penalidades introduzidas bem como uma análise mais completa dos resultados da fase de validação. O mesmo será publicado como um capítulo em livro.

- Chaves, M. S.; Strube de Lima, V. L. Em direção ao Mapeamento Automático entre Estruturas Ontológicas. *IX Jornadas Iberoamericanas de Informática*, Cartagena de Indias, Colômbia, 11-15 de agosto de 2003.

  Este artigo inclui um resumo dos experimentos realizados para a língua inglesa, bem como o protótipo desenvolvido.

- Chaves, M. S. Um Estudo e Apreciação sobre Algoritmos de Stemming para a Língua Portuguesa. *IX Jornadas Iberoamericanas de Informática*, Cartagena de Indias, Colômbia, 11-15 de agosto de 2003.

  Este artigo apresenta um estudo comparativo de dois algoritmos desenvolvidos especificamente para a língua portuguesa. Este estudo nos auxiliou a escolher o algoritmo que apresentou o melhor desempenho durante os testes realizados.

Referências Bibliográficas

[1] Alexander Maedche and Steffen Staab. Measuring Similarity between Ontologies. In *Proceedings of the European Conference on Knowledge Acquisition and Management - (EKAW-2002). Madrid, Spain, October 1-4*, pages 251–263, 2002.

[2] Natalya Fridman Noy and Mark A. Musen. Anchor-PROMPT: Using Non-Local Context for Semantic Matching. In *Proceedings of the Workshop on Ontologies and Information Sharing at the Seventeenth International Joint Conference on Artificial Intelligence (IJCAI-2001), Seattle, WA*, August 2001.

[3] Natalya Fridman Noy and Mark A. Musen. SMART: Automated Support for Ontology Merging and Alignment. In *Twelfth Banff Workshop on Knowledge Acquisition, Modeling, and Management - Banff, Alberta, Canada*, 1999.

[4] Prasenjit Mitra, Gio Wiederhold, and Martin Kersten. A Graph-Oriented Model for Articulation of Ontology Interdependencies. *Lecture Notes in Computer Science*, 1777:86, 2000.

[5] H. Chalupksy. OntoMorph: A Translation System for Symbolic Knowledge. In *Proceedings of the 17^{th} International Conference on Knowledge Representation and Reasoning (KR-2000)*, 2000.

[6] AnHai Doan, Jayant Madhavan, Pedro Domingos, and Alon Halevy. Learning to Map between Ontologies on the Semantic Web. In *Proceedings of the World-Wide Web Conference (WWW-2002), Honolulu, Hawaii, USA*, May 2002.

[7] Deborah L. McGuinness, Richard Fikes, James Rice, and Steve Wilder. The Chimaera Ontology Environment. In *Proceedings of the 17^{th} National Conference on Artificial Intelligence (AAAI)*, 2000.

[8] Michael Uschold. Where is the Semantics in the Semantic Web? In *Workshop on Ontologies in Agent Systems*, Montreal, Canada, May 2001.

[9] Gerd Stumme and Alexander Maedche. FCA-MERGE: Bottom-Up Merging of Ontologies. In *Proceedings of the 17^{th} International Joint Conference on Artificial Intelligence (IJCAI)*, pages 225–234, 2001.

[10] Erhard Rahm and Philip A. Bernstein. A Survey of Approaches to Automatic Schema Matching. *Very Large Data Base (VLDB) Journal*, 10(4), 2001.

[11] Ying Ding and Schubert Foo. Ontology Research and Development Part 2 - A Review of Ontology Mapping and Evolving. *Journal of Information Science*, 28(5):375–388, 2002.

[12] Sushama Prasad, Yun Peng, and Timothy Finin. Using Explicit Information To Map Between Two Ontologies. In *Proceedings of the 1st International Joint Conference on Autonomous Agents and Multi-Agent Systems - Workshop on Ontologies in Agent Systems (OAS) - Bologna, Italy. 15-19 July*, 2002.

[13] Balakrishnan Chandrasekaran, John R. Josephson, and V. Richard Benjamins. What Are Ontologies, and Why Do We Need Them? *IEEE Intelligent Systems*, 14(1):20–26, 1999.

[14] Christiane Fellbaum, editor. *WordNet: an electronic lexical database*. Massachusetts: The MIT Press, 1998. p.423.

[15] Dieter Fensel. Ontology-Based Knowledge Management. *IEEE Computer*, pages 56–59, 2002.

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

[17] Dieter Fensel. *Ontologies: Silver Bullet for Knowledge Management and Electronic Commerce*. Springer-Verlag, 2001.

[18] 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.

[19] Tom Gruber. What is an Ontology? Disponível em: <http://www.ksl.stanford.edu/kst/what-is-an-ontology.html>, Acesso em: 03 de novembro de 2003.

[20] Eduard H. Hovy. Combining and Standardizing Large-Scale, Practical Ontologies for Machine Translation and Other Uses. In *Proceedings of the 1st International Conference on Language Resources and Evaluation (LREC)*, Granada, Spain, 1998.

[21] Natalya Fridman Noy and Mark A. Musen. PROMPT: Algorithm and Tool for Automated Ontology Merging and Alignment. In *Proceedings of the 17th National Conference on Artificial Intelligence, (AAAI'00)*, 2000.

[22] Farshad Hakimpour and Andreas Geppert. Resolving Semantic Heterogeneity in Schema Integration: an Ontology Based Approach. In *Proceedings of the International Conference on Formal Ontology in Information Systems FOIS-2001*, 2001.

[23] Gerd Stumme, R. Studer, and Y. Sure. Towards an Order-Theoretical Foundation for Maintaining and Merging Ontologies. In *Verbundtagung Wirtschaftsinformatik 2000. F. Bodendorf, M. Grauer (eds.)*, pages 136–149. Shaker, Aachen, 2000.

[24] Helena Sofia Pinto, Asunción Gómez-Pérez, and João P. Martins. Some Issues on Ontology Integration. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI) Workshop on Ontologies and Problem-Solving Methods (KRR5), Stockholm, Sweden*, August 1999.

[25] Prasenjit Mitra, Gio Wiederhold, and J. Jannink. Semi-automatic Integration of Knowledge Sources. *Proceedings of Fusion´99, Sunnyvale, USA*, 1999.

[26] Prasenjit Mitra and Gio Wiederhold. Resolving Terminological Heterogeneity in Ontologies. In *Proceedings of the European Conference on Artificial Intelligence (ECAI) Workshop on Ontologies and Semantic Interoperability*, 2002.

[27] Michel Klein, Atanas Kiryakov, Damyan Ognyanoff, and Dieter Fensel. Finding and specifying relations between ontology versions. *European Conference on Artificial Intelligence (ECAI), Workshop on Ontologies and Semantic Interoperability*, 2002.

[28] John F. Sowa. Building, Sharing, and Merging Ontologies. Disponível em: <http://www.jfsowa.com/ontology/ontoshar.htm>, Acesso em: 03 de novembro de 2003.

[29] Eduard H. Hovy and Sergei Nirenburg. Approximating an Interlingua in a Principled Way. In *Proceedings of the DARPA Speech and Natural Language Workshop*, Arden House, NY, 1992.

[30] Andreas Paepcke, Chen-Chuan K.Chang, Terry Winograd, and Hector García-Molina. Interoperability for Digital Libraries Worldwide. *Special Issue on Digital Libraries, Communications of the ACM*, 41(4):33–43, 1998.

[31] Sergey Melnik and Stefan Decker. A Layered Approach to Information Modeling and Interoperability on the Web. Disponível em: <http://www-db.stanford.edu/∼melnik/pub/sw00/>, Acesso em: 03 de novembro de 2003.

[32] William E. Moen. Mapping the Interoperability Landscape for Networked Information Retrieval. In *Proceedings of First ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 50–52, Roanoke, VA, Junho 2001. ACM Press.

[33] Jérôme Euzenat. Towards a Principled Approach to Semantic Interoperability. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI) Workshop on Ontologies and Information Sharing*, pages 19–25, 2001.

[34] Clifford Lynch and Hector Molina-Garcia (Eds.). Interoperability, Scaling, and Digital Libraries Research Agenda. Technical report, IITA Digital Libraries Workshop, 1995.

[35] Hsinchun Chen. Semantic Research for Digital Libraries. *D-Lib Magazine*, 5(10), 1999.

[36] James Pustejovsky. *The Generative Lexicon*. Cambridge: The MIT Press, 1995. p.298.

[37] Evanildo Bechara. *Moderna Gramática Portuguesa*. Rio de Janeiro: Lucerna, 2001. p.672.

[38] Daniel Saul Jurafsky and James H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Upper Saddle River, NJ:Prentice Hall, 2000. p.934.

[39] Aurélio Buarque de Holanda Ferreira. *Aurélio Século XXI: o dicionário da língua portuguesa*. Rio de Janeiro: Nova Fronteira, 1999. p.2128.

[40] Ora Lassila and Ralph R. Swick. Resource Description Framework (RDF) Model and Syntax Specification. Disponível em: <http://www.w3.org/TR/REC-rdf-syntax/>, Acesso em: 03 de novembro de 2003.

[41] Dan Brickley and R.V. Guha. Resource Description Framework (RDF) Schema Specification 1.0. Disponível em: <http://www.w3.org/TR/2000/CR-rdf-schema-20000327/>, Acesso em: 03 de novembro de 2003.

[42] Steffen Staab, Michael Erdmann, Alexander Maedche, and Stefan Decker. An extensible approach for modeling ontologies in RDF(S). *First Workshop on the Semantic Web at the Fourth European Conference International Workshop on Research and Advanced Technology for Digital Libraries*, 2000.

[43] Dieter Fensel. The Semantic Web and Its Languages. *IEEE Intelligent Systems*, 15(6):67–73, 2000.

[44] Natanya Pitts-Moultis and Cheryl Kirk. *XML Black Book - Solução e Poder*. Makron Books, 2000.

[45] J. Heflin and Jim Hendler. Semantic Interoperability on the Web. In *Proceedings of Extreme Markup Languages*, pages 111–120, 2000.

[46] Colin Beardon, David Lumsden, and Geoff Holmes. *Natural Language and Computational Linguistics: An Introduction*. London: Ellis Horwood, 1991.

[47] James Allen. *Natural Language Understanding*. Redwood City: The Benjamin/Cummings Publishing Company, 1995.

[48] Dieter Fensel, Ian Horrocks, F. van Harmelen, Stefan Decker, Michael Erdmann, and Michel Klein. OIL in a nutshell. In *Knowledge Acquisition, Modeling and Management*, pages 1–16, 2000.

[49] Ian Horrocks, Dieter Fensel, J. Broekstra, Stefan Decker, Michael Erdmann, C. Goble, F. van Harmelen, Michel Klein, Steffen Staab, Rudi Studer, and E. Motta. The Ontology Inference Layer OIL. Technical report, On-To-Knowledge, 2000.

[50] J. Heflin. *Towards the Semantic Web: Knowledge Representation in a Dynamic, Distributed Environment*. PhD thesis, University of Maryland - College Park, 2001.

[51] Ian Horrocks. DAML+OIL: a Description Logic for the Semantic Web. *IEEE Bull. of the Technical Committee on Data Engineering*, 25(1):4–9, March 2002.

[52] Dicionário Universal - Língua Portuguesa. Disponível em: <http://www.priberam.pt/DLPO/>, Acesso em: 03 de novembro de 2003.

[53] José Saias and Paulo Quaresma. Construção Automática de Ontologias e sua Utilização em Sistemas de Recuperação de Informação em Texto. In *XIII SBIE2002, Workshop de Ontologias*, pages 605–607, 2002.

[54] Pablo Gamallo, Alexandre Augustini, Paulo Quaresma, and José Gabriel Pereria Lopes. Using semantic word classes in text information retrieval systems. In *XIII SBIE2002, Workshop de Ontologias*, pages 590–592, 2002.

[55] Sandro Rigo and Renata Vieira. Busca de Informações Auxiliada por Ontologia. In *XIII SBIE2002 Workshop de Ontologias*, pages 596–598, 2002.

[56] Sean Bechhofer, Frank van Harmelen, Jim Hendler, Ian Horrocks, Deborah L. McGuinness, Peter F. Patel-Schneider, and Lynn Andrea Stein. OWL Web Ontology Language Reference. Disponível em: <http://www.w3.org/TR/owl-ref/>, Acesso em: 03 de novembro de 2003.

[57] Natalya Fridman Noy and Mark A. Musen. Evaluating Ontology-Mapping Tools: Requirements and Experience. Disponível em: <http://km.aifb.uni-karlsruhe.de/eon2002/EON2002_Musen.ppt>, Acesso em: 03 de novembro de 2003.

[58] Natalya Fridman Noy and Mark A. Musen. Evaluating Ontology-Mapping Tools: Requirements and Experience. In *Proceedings of the Workshop on Evaluation of Ontology Tools at EKAW'02 (EON2002), Siguenza, Spain, October*, 2002.

[59] Deborah L. McGuinness, Richard Fikes, James Rice, and Steve Wilder. An Environment for Merging and Testing Large Ontologies. In *Proceedings of the Seventh International Conference on Principles of Knowledge Representation and Reasoning (KR2000)*, Breckenridge, Colorado, United States, April 2000.

[60] Christopher D. Manning and Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. Cambridge, MA : The MIT Press, 1999.

[61] Ichise Ryutaro, Takeda Hideaki, and Honiden Shinichi. Rule Induction for Concept Hierarchy Alignment. In *Proceedings of the Workshop on Ontology Learning at the 17th International Joint Conferences on Artificial Intelligence (IJCAI)*. Press, 2001.

[62] Alexander Maedche, B. Motik, N. Silva, and R. Volz. MAFRA - A Mapping Framework for Distributed Ontologies. In *Proceedings of the European Conference on Knowledge Acquisition and Management - (EKAW-2002). Madrid, Spain, October 1-4*, 2002.

[63] Eduardo Mena. *OBSERVER: An Approach for Query Processing in Global Information Systems based on Interoperation across Pre-existing Ontologies*. PhD thesis, Departamento de Informática e Ingeniería de Sistemas. Universidad de Zaragoza, 1998.

[64] Eduardo Mena, Vipul Kashyap, Amit P. Sheth, and Arantza Illarramendi. OBSERVER: An Approach for Query Processing in Global Information Systems based on Interoperation across Pre-existing Ontologies. *International journal on Distributed And Parallel Databases (DAPD)*, 8(2):223–272, April 2000.

[65] William W. Cohen, Pradeep Ravikumar, and 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)*, Acapulco, México, 9-10 August 2003.

[66] Dekang Lin. An Information-Theoretic Definition of Similarity. In *Proceedings of the 15th International Conference on Machine Learning*, pages 296–304. Morgan Kaufmann, San Francisco, CA, 1998.

[67] Alexander Budanitsky and Graeme Hirst. Semantic Distance in WordNet: An Experimental, Application-oriented Evaluation of Five Measures. In *Workshop on WordNet and Other Lexical Resources, in the North American Chapter of the Association for Computational Linguistics (NAACL-2000)*, Pittsburgh, PA, June 2001.

[68] Philip Resnik. Using Information Content to Evaluate Semantic Similarity in a Taxonomy. In *Proceedings of the XI International Joint Conferences on Artificial Intelligence (IJCAI)*, pages 448–453, 1995.

[69] Jay J. Jiang and David W. Conrath. Semantic Similarity based on Corpus Statistics and Lexical Taxonomy. In *Proceedings of the International Conference on Research in Computational Linguistics (ROCLING)*. Taiwan, 1997.

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

[71] Alexander Maedche and Steffen Staab. Comparing Ontologies - Similarity Measures and a Comparison Study. Technical report, Institute AIFB, University of Karlsruhe. Internal Report, 2001.

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

[73] Ricardo Baeza-Yates and Berthier Ribeiro-Neto. *Modern Information Retrieval*. New York, NY: Addison-Wesley, 1999.

[74] Gerda Ruge. *Combining Corpus Linguistics and Human Memory Models for Automatic Term Association: Natural Language Information Retrieval: Text, Speech and Language Technology*. Kluwer Academic Publishers, 1999.

[75] Karen Spark Jones and P. Willet, editors. *Readings in Information Retrieval*. Morgan Kaufmann Publishers, Inc., San Francisco, California, 1997.

[76] Asunción Honrado, Ruben Leon, Ruairi O'Donnell, and Duncan Sinclair. A Word Stemming Algorithm for the Spanish Language. In *Proceedings of the Seventh International Symposium on String Processing Information Retrieval (SPIRE'00) - September 27 - 29*, A Coruña, Spain, 2000.

[77] Viviane Moreira Orengo and Christian Huyck. A Stemming Algorithm for Portuguese Language. In *Proceedings of Eigth Symposium on String Processing and Information Retrieval (SPIRE-2001)*, pages 186–193, 2001.

[78] Marcirio Silveira Chaves. Um Estudo e Apreciação sobre Algoritmos de Stemming para a Língua Portuguesa. IX Jornadas Iberoamericanas de Informática. Cartagena de Indias - Colômbia (CD-ROM), 11-15 de Agosto 2003.

[79] Natalya Fridman Noy and Deborah L. McGuinness. Ontology Development 101: A Guide to Creating Your First Ontology. Technical report, Stanford Knowledge Systems Laboratory, Technical Report KSL-01-05 and Stanford Medical Informatics, Technical Report SMI-2001-0880, 2001.

Apêndice A

Exemplo de código RDF

```
<?xml version="1.0"?>
   <rdf:RDF xml:lang="en"
        xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
        xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
        xmlns="http://www.pucrs.br/BD/tests/bib.rdf#">
<!-- Definição das classes -->
   <rdfs:Class rdf:ID="Criador"/>
   <rdfs:Class rdf:ID="Obra"/>
   <rdfs:Class rdf:ID="BibDig"/>
   <rdfs:Class rdf:ID="Autor">
      <rdfs:subClassOf rdf:resource="#Criador"/>
   </rdfs:Class>
   <rdfs:Class rdf:ID="Livro">
            <rdfs:subClassOf rdf:resource="#Obra"/>
   </rdfs:Class>

<!-- Definição das propriedades -->
   <rdf:Property rdf:ID="cria">
        <rdfs:domain rdf:resource="#Criador"/>
        <rdfs:range rdf:resource="#Obra"/>
   </rdf:Property>
   <rdf:Property rdf:ID="disponível_pela">
         <rdfs:domain rdf:resource="#Obra"/>
         <rdfs:range rdf:resource="#BibDig"/>
   </rdf:Property>
   <rdf:Property rdf:ID="escreve">
         <rdfs:domain rdf:resource="#Autor"/>
         <rdfs:range rdf:resource="#Livro"/>
         <rdfs:subPropertyOf rdf:resource="#cria"/>
   </rdf:Property>

<!-- Definição das instâncias -->
   <Autor rdf:ID="DanielGoleman">
         <Prim_nome>Daniel</Prim_nome>
         <Sobrenome>Goleman</Sobrenome>
         <escreve>
           <Livro rdf:about="http://www.europa.com/danielg/
                             livros/IntelEmoc.pdf">
               <disponível_pela>
                   <BibDig rdf:about="http://www.pucrs.br/BD"/>
               </disponível_pela>
           </Livro>
        </escreve>
  </Autor>
</rdf:RDF>
```

Apêndice B

Extratos dos Experimentos da Fase de Validação

Tabela B.1: Termos monopalavra considerados similares pelas medidas CC e SL na fase de validação

EO-base	EO-alvo	CC	SL	EO-base	EO-alvo
agricultor	agricultura	0.80	0.76	agricul	agricult
apicultura	avicultura	0.90	0.76	apicult	avicult
caprinocultura	ciprinocultura	0.93	0.81	caprinocult	ciprinocult
cartilha	partilha	0.88	0.76	cartilh	partilh
composicao	compositae	0.80	0.79	composica	composita
concessao	concussao	0.89	0.77	concessa	concussa
condomino	condominio	0.89	0.77	condomin	condomini
confeitos	conceitos	0.89	0.76	confeit	conceit
confeitos	conflito	0.75	0.76	confeit	conflit
contrato	contralto	0.88	0.76	contrat	contralt
convenio	conventos	0.75	0.76	conveni	convent
cristologia	criptologia	0.91	0.79	cristolog	criptolog
cunicultura	canicultura	0.91	0.77	cunicult	canicult
distribuicao	distribuicoes	0.75	0.81	distribuica	distribuico
dolarizacao	polarizacao	0.91	0.80	dolarizaca	polarizaca
dolarizacao	solarizacao	0.91	0.80	dolarizaca	solarizaca
eletrocardiograma	eletrocardiografia	0.88	0.84	eletrocardiogram	eletrocardiograf
eletroencefalograma	eletroencefalografia	0.89	0.84	eletroencefalogram	eletroencefalograf
emigracao	imigracao	0.89	0.77	emigraca	imigraca
emigracao	migracao	0.88	0.76	emigraca	migraca
espanhois	espanhol	0.75	0.77	espanhoil	espanhol
filiacao	afiliacao	0.88	0.76	filiaca	afiliaca
fitogenetica	citogenetica	0.92	0.77	fitogene	citogene
fitologia	citologia	0.89	0.76	fitolog	citolog
fitologia	ficologia	0.89	0.76	fitolog	ficolog
fitologia	litologia	0.89	0.76	fitolog	litolog
fitologia	mitologia	0.89	0.76	fitolog	mitolog
fluorita	fluoreto	0.75	0.76	fluorit	fluoret
ginecologia	sinecologia	0.91	0.79	ginecolog	sinecolog
habitacao	habituacao	0.89	0.77	habitaca	habituaca
inquiricao	inquisicao	0.90	0.79	inquirica	inquisica
matrimonio	patrimonio	0.90	0.79	matrimoni	patrimoni
mobilizacao	imobilizacao	0.91	0.80	mobilizaca	imobilizaca
nefrologia	neurologia	0.90	0.77	nefrolog	neurolog
orizicultura	rizicultura	0.91	0.77	orizicult	rizicult
ovinocultura	bovinocultura	0.92	0.79	ovinocult	bovinocult
profissao	procissao	0.89	0.77	profissa	procissa
psicometria	psicrometria	0.91	0.79	psicometr	psicrometr
retencao	detencao	0.88	0.76	retenca	detenca
servidao	cervidae	0.75	0.76	servida	cervida
subversivo	subversao	0.78	0.76	subvers	subversa
telecomunicacao	telecomunicacoes	0.80	0.83	telecomunicaca	telecomunicaco
tipografia	topografia	0.90	0.77	tipograf	topograf
topografia	tomografia	0.90	0.77	topograf	tomograf
tradicao	traducao	0.88	0.76	tradica	traduca
veterinaria	veterinario	0.91	0.76	veterin	veterina

Tabela B.2: Termos multipalavra considerados similares pela medida CC e pela medida SL na fase de validação

EO-base	EO-alvo	CC	SL	EO-base	EO-alvo
acumulacaoDeAcoes	cumulacaoDeAcoes	0.94	0.77	acumulacadeaco	cumulacadeaco
administracaoSanitaria	admnistracaoSanitaria	0.95	0.81	administracasanit	admnistracasanit
bicho-da-seda	bichos-da-seda	0.92	0.82	bicho-da-sed	bichos-da-sed
competicaoEsportiva	competicoesEsportivas	0.79	0.79	competicaespor	competicoespor
condicoesEconomicas	condicaoEconomica	0.76	0.76	condicoeconom	condicaeconom
condicoesSanitarias	condicaoSanitaria	0.76	0.76	condicosanit	condicasanit
construcaoMetalica	construcoesMetalicas	0.78	0.79	construcametal	construcometal
criacaoDeCaracol	criacaoDeCaracois	0.88	0.76	criacadecaracol	criacadecaracoil
descobertaEExploracao	descobertasEExploracoes	0.81	0.79	descoberteexploraca	descoberteexploraco
expedicaoCientifica	expedicoesCientificas	0.79	0.77	expedicacientif	expedicocientif
exposicaoInternacional	exposicoesInternacionais	0.77	0.77	exposicainternac	exposicointernac
funcionalismoPublico	funcionarioPublico	0.83	0.76	funcionpubl	funcionapubl
instituicaoFinanceira	instituicoesFinanceiras	0.81	0.80	instituicafinanc	instituicofinanc
instituicaoPolitica	instituicoesPoliticas	0.79	0.80	instituicapoli	instituicopoli
integracaoSocial	interacaoSocial	0.93	0.77	integracasoc	interacasoc
mercadoMobiliario	mercadoImobiliario	0.88	0.76	mercmobilia	mercimobilia
partidoDemocratico	partidoDemocrata	0.81	0.76	partdemocra	partdemocrat
religiaoPrimitiva	religioesPrimitivas	0.76	0.76	religiaprimi	religioprimi

Tabela B.3: Termos multipalavra considerados não similares pela medida CC e similares pela medida SL na fase de validação

EO-base	EO-alvo	CC	SL	EO-base	EO-alvo
alimento,IndustriaEComercio	alimentosIndustrializados	0.60	0.76	alimentindustr	alimentindustri
auto-estrada	auto-estima	0.73	0.77	auto-estr	auto-est
comunicacaoDigital	comunicacoesDigitais	0.72	0.80	comunicacadigit	comunicacodigit
funcionalismoPublico	funcionarioPublicoEstadual	0.45	0.76	funcionpubl	funcionapubl
funcionalismoPublico	funcionarioPublicoFederal	0.50	0.76	funcionpubl	funcionapubl
funcionalismoPublico	funcionarioPublicoMunicipal	0.40	0.76	funcionpubl	funcionapubl
instalacaoEletricaPredial	instalacoesEletricas	0.50	0.79	instalacaeletr	instalacoeletr
partidoDemocratico	partidoDemocrataCristao	0.67	0.76	partdemocra	partdemocrat
perversaoSexual	perversoesSexuais	0.67	0.77	perversasex	perversosex
usinaHidroeletricaDeBaixaQueda	usinasHidreletricas	0.26	0.79	usinhidroeletr	usinhidreletr

Tabela B.4: Termos monopalavra considerados não similares pela medida CC e similares pela medida SL na fase de validação

EO-base	EO-alvo	CC	SL	EO-base	EO-alvo
adivinhacao	adivinhacoes	0.73	0.80	adivinhaca	adivinhaco
aeronave	aeronautica	0.38	0.76	aeronav	aeronau
anticoncepcional	anticoncepcao	0.69	0.81	anticoncepc	anticoncepca
artropodes	arthropoda	0.70	0.77	artropod	arthropod
biblioteca	bibliotecarios	0.60	0.79	bibliotec	biblioteca
bioenergia	bioenergetica	0.70	0.77	bioenerg	bioenerge
caminhao	caminhoes	0.62	0.76	caminha	caminho
combustao	combustiveis	0.44	0.77	combusta	combusti
comerciante	comerciarios	0.64	0.76	comerci	comercia
comercio	comerciarios	0.50	0.76	comerci	comercia
conflito	confeitaria	0.38	0.76	conflit	confeit
consultor	consultorios	0.67	0.76	consult	consulto
contradita	contraditorio	0.60	0.79	contradit	contradito
contrafe	contrabando	0.38	0.76	contraf	contrab
contrafe	contratos	0.62	0.76	contraf	contrat
contrato	contrabando	0.50	0.76	contrat	contrab
corporacao	corporacoes	0.70	0.79	corporaca	corporaco
desquite	mesquitas	0.62	0.76	desquit	mesquit
eletronica	eletrodo	0.50	0.76	eletron	eletrod
embarcacao	embarcacoes	0.70	0.79	embarcaca	embarcaco
empreitada	emprestimo	0.60	0.76	empreit	emprest
empresario	emprestimo	0.70	0.76	empresa	emprest
escolaridade	escolastica	0.55	0.76	escolar	escolas
estabilidade	estabulos	0.33	0.76	estabil	estabul
estilistica	estilosante	0.55	0.76	estilis	estilos
evangelho	evangelismo	0.67	0.76	evangelh	evangel
fotografia	fotogravura	0.70	0.77	fotograf	fotograv
funcionario	funcionalismo	0.73	0.76	funciona	funcion
habitacao	habitacoes	0.67	0.77	habitaca	habitaco
holografia	hologramas	0.70	0.77	holograf	hologram
inflamaveis	inflacao	0.38	0.76	inflama	inflaca
intermediario	intermedio	0.70	0.79	intermedia	intermedi
magistrado	magisterio	0.70	0.76	magistr	magiste
metanol	metabolismo	0.29	0.76	metanol	metabol
meteorito	meteoritica	0.67	0.76	meteorit	meteori
mosquito	mesquitas	0.62	0.76	mosquit	mesquit
municipalismo	municipio	0.56	0.76	municip	municipi
penitencia	penitenciarias	0.60	0.77	penitenc	penitenci
presidiario	presidios	0.56	0.76	presidia	presidi
profissao	profissoes	0.67	0.77	profissa	profisso
provincialismo	provincia	0.44	0.76	provinci	provinc
religiao	religioes	0.62	0.76	religia	religio
responsabilidade	responsorio	0.27	0.76	respons	responso
terapeutica	terapeutas	0.70	0.76	terapeu	terapeut
tolerancia	tolerantismo	0.60	0.77	toleranc	tolerant

Apêndice C

Extratos dos Experimentos da Fase de Avaliação

Tabela C.1: Resultados pertencentes ao grupo G2 com ocorrência de preposições diferentes nos termos

EO-base	EO-alvo	CC	SL
acidenteDoTrabalho	acidentesDeTrabalho	0.89	0.40
ambienteDoTrabalho	ambienteDeTrabalho	0.94	0.40
comunicacaoEmAdministracao	comunicacaoNaAdministracao	0.92	0
controleDeAdministracao	controleAdministrativo	0.77	0
controleDeProducao	controleDaProducao	0.94	0.40
controleDeQualidade	controleDaQualidade	0.95	0.40
declaracaoDaVontade	declaracaoDeVontade	0.95	0.40
desenvolvimentoTecnologico	desenvolvimentoDeTecnologia	0.85	0
exameDoDna	exameDeDna	0.90	0.40
execucaoDeTituloExtrajudicial	execucaoPorTituloExtrajudicial	0.90	0
guardaDosFilhos	guardaDeFilhos	0.86	0.40
industriaDeMadeira	industriaDaMadeira	0.94	0.40
liberdadeComercial	liberdadeDeComercio	0.78	0
liquidacaoDaSentenca	liquidacaoDeSentenca	0.95	0.40
prescricaoPenal	prescricaoDaPena	0.80	0
psicologiaDaAdolescencia	psicologiaDoAdolescente	0.83	0
remissaoDaPena	remissaoPenal	0.77	0
seguroAcidente	seguroDeAcidente	0.86	0
seguro-habitacao	seguroDeHabitacao	0.81	0
seguroTransporte	seguroDeTransporte	0.88	0
teoriaDeFila	teoriaDasFilas	0.75	0.40
trabalhoDeGrupo	trabalhoEmGrupo	0.87	0
tratamentoDaAgua	tratamentoDeAgua	0.94	0.40

Tabela C.2: Resultados pertencentes ao grupo G4: utilização da heurística da primeira letra

EO-base	EO-alvo	CC	SL	SLH	EO-base	EO-alvo
depressao	repressao	0.89	0.77	0	depressa	repressa
detencao	retencao	0.88	0.76	0	detenca	retenca
gastronomia	astronomia	0.90	0.77	0	gastronom	astronom
litologia	citologia	0.89	0.76	0	litolog	citolog
litologia	fitologia	0.89	0.76	0	litolog	fitolog
litologia	mitologia	0.89	0.76	0	litolog	mitolog
metrologia	petrologia	0.90	0.77	0	metrolog	petrolog
migracao	emigracao	0.88	0.76	0	migraca	emigraca
migracao	imigracao	0.88	0.76	0	migraca	imigraca
emigracaoRural	migracaoRural	0.92	0.76	0	emigracarural	migracarural
mitologia	citologia	0.89	0.76	0	mitolog	citolog
mitologia	fitologia	0.89	0.76	0	mitolog	fitolog
ontologia	antologia	0.89	0.76	0	ontolog	antolog
petrologia	metrologia	0.90	0.77	0	petrolog	metrolog
quitacao	equitacao	0.88	0.76	0	quitaca	equitaca
ranicultura	canicultura	0.91	0.77	0	ranicult	canicult
revolucao	evolucao	0.88	0.76	0	revoluca	evoluca
vitrificacao	nitrificacao	0.92	0.81	0	vitrificaca	nitrificaca

Tabela C.3: Resultados pertencentes ao grupo G4 em que ambas as medidas discordam do analisador humano

EO-base	EO-alvo	CC	SL	SLH	EO-base	EO-alvo
astrologia	artrologia	0.90	0.77	0.77	astrolog	artrolog
conflitos	confeitos	0.89	0.76	0.76	conflit	confeit
descriminacao	discriminacao	0.92	0.82	0.82	descriminaca	discriminaca
filologia	ficologia	0.89	0.76	0.76	filolog	ficolog
filologia	fitologia	0.89	0.76	0.76	filolog	fitolog
hepatite	hematita	0.75	0.76	0.76	hepatit	hematit
interdata	interdito	0.78	0.77	0.77	interdat	interdit
interpelacao	interpolacao	0.92	0.81	0.81	interpelaca	interpolaca
macroeconomia	microeconomia	0.92	0.81	0.81	macroeconom	microeconom
magnesita	magnesio	0.75	0.76	0.76	magnesit	magnesi
microeconomia	macroeconomia	0.92	0.81	0.81	microeconom	macroeconom
microprocessador	macroprocessadores	0.81	0.82	0.82	microprocess	macroprocess
mitologia	micologia	0.89	0.76	0.76	mitolog	micolog
ontologia	oncologia	0.89	0.76	0.76	ontolog	oncolog
repressao	regressao	0.89	0.77	0.76	repressa	regressa
revolucao	resolucao	0.89	0.77	0.77	revoluca	resoluca
tipologia	topologia	0.89	0.76	0.76	tipolog	topolog

Tabela C.4: Resultados pertencentes ao grupo G5: termos monopalavra

EO-base	EO-alvo	CC	SL	EO-base	EO-alvo
nacionalismo	racionalismo	0.92	0.73	nacion	racion
maternidade	paternidade	0.91	0.73	matern	patern
tuberculose	tuberculos	0.90	0.55	tuberculos	tubercul
imigrante	emigrante	0.89	0.70	imigr	emigr
terceiros	terreiros	0.89	0.65	terc	terr
atentado	atestado	0.88	0.70	atent	atest
corretor	corredor	0.88	0.65	corre	corr
dinheiro	pinheiro	0.88	0.65	dinh	pinh
estetica	estatica	0.88	0.65	este	esta
geologia	teologia	0.88	0.73	geolog	teolog
histeria	listeria	0.88	0.65	hist	list
indisponibilidade	biodisponibilidade	0.88	0.63	indisponibil	biodisponibil
moratoria	oratoria	0.88	0.65	morat	orat
oratoria	oratorio	0.88	0.65	orat	orato
vistoria	historia	0.88	0.65	vist	hist
assalto	asfalto	0.86	0.73	assalt	asfalt
acao	racao	0.75	0.57	aca	raca
alga	salga	0.75	0.57	alg	salg
alho	olho	0.75	0.57	alh	olh
alma	arma	0.75	0.57	alm	arm
alma	asma	0.75	0.57	alm	asm
amor	amora	0.75	0	am	amor
amor	ator	0.75	0.40	am	at
cloro	coro	0.75	0.57	clor	cor
coca	boca	0.75	0.57	coc	boc
convencao	invencao	0.75	0.51	convenca	invenca
desarmamento	desmatamento	0.75	0	desarm	desmat
erro	ferro	0.75	0.57	err	ferr
geologia	zoologia	0.75	0.47	geolog	zoolog
livreiro	letreiro	0.75	0.30	livr	letr
loja	soja	0.75	0.57	loj	soj
progresso	processo	0.75	0.51	progress	process
protesto	processo	0.75	0.51	protest	process
tecnologia	ecologia	0.75	0.47	tecnolog	ecolog
traducao	producao	0.75	0.51	traduca	produca

Tabela C.5: Resultados pertencentes ao grupo G5: termos multipalavra

EO-base	EO-alvo	CC	SL	EO-base	EO-alvo
produtividadeDoTrabalho	produtividadeNoTrabalho	0.96	0.40	produtdotrabalh	produtnotrabalh
violacaoDeComunicacao	violacaoDaComunicacao	0.95	0.40	violacadecomunicaca	violacadacomunicaca
cartaoDeCredito	cartaDeCredito	0.93	0.65	cartadecredit	cartdecredit
delitoFiscal	debitoFiscal	0.92	0.70	delitfiscal	debitfiscal
ensinoMedico	ensinoMedio	0.91	0.65	ensinmedic	ensinmedi
desenvolvimentoMental	desenvolvimentoFetal	0.90	0	desenvolvment	desenvolvfetal
direitoPenalEcologico	direitoPenalEconomico	0.90	0.47	direitpenalecolog	direitpenaleconom
irretroatividadeDasLeis	retroatividadeDasLeis	0.90	0.51	irretroatdaleil	retroatdaleil
policiaAdministrativa	politicaAdministrativa	0.90	0.65	policadministr	poliadministr
politicaDeImportacao	politicaDeExportacao	0.90	0.58	polideimportaca	polideexportaca
politicaInternacional	policiaInternacional	0.90	0.65	poliinternac	policinternac
prospeccaoGeoquimica	prospeccaoBioquimica	0.90	0.51	prospeccageoquim	prospeccabioquim
responsabilidadeLegal	responsabilidadePenal	0.90	0.40	responslegal	responspenal
restricaoAImportacao	restricaoAExportacao	0.90	0.58	restricaaimportaca	restricaaexportaca
administracaoDePessoal	administracaoDeRisco	0.75	0	administracadepessoal	administracaderisc
administracaoEstadual	administracaoEscolar	0.75	0	administracaestad	administracaescol
administracaoFederal	administracaoDeVendas	0.75	0	administracafeder	administracade
conferenciaInternacional	controversiaInternacional	0.75	0	conferencinternac	controversinternac
desenvolvimentoCultural	desenvolvimentoFetal	0.75	0	desenvolvcult	desenvolvfetal
direitoAVida	direitoAVoto	0.75	0.13	direitavid	direitavot
menorDelinquente	mulherDelinquente	0.75	0	mendelinqu	mulhdelinqu
policiaNaval	policiaCivil	0.75	0	policnaval	policcivil
producaoCultural	producaoColonial	0.75	0	producacult	producacolon
rioInternacional	bancoInternacional	0.75	0	riointernac	bancinternac
treinamentoNaval	treinamentoDaVoz	0.75	0	treinnaval	treinda

Tabela C.6: Resultados pertencentes ao grupo G7

EO-base	EO-alvo	CC	SL	EO-base	EO-alvo
aerossol	aerosol	0.86	0.76	aerossol	aerosol
auditoria	auditorios	0.78	0.70	audit	audito
bioecologia	sinecologia	0.82	0.58	bioecolog	sinecolog
biologia	axiologia	0.75	0.47	biolog	axiolog
biologia	ficologia	0.75	0.47	biolog	ficolog
biologia	litologia	0.75	0.47	biolog	litolog
colonialismo	comensalismo	0.75	0	coloni	comens
dermatologia	farmacologia	0.75	0	dermatolog	farmacolog
dermatologia	gerontologia	0.75	0	dermatolog	gerontolog
didatica	dinamica	0.75	0.30	dida	dinam
impressao	imprensa	0.75	0.51	impressa	imprens
invencao	invencoes	0.63	0.76	invenca	invenco
lagoa	lagos	0.80	0.57	lago	lag
administracaoFederal	administracaoSalarial	0.75	0	administracafeder	administracasalar
biologiaSocial	psicologiaSocial	0.79	0	biologsoc	psicologsoc
contabilidadeSocial	habilidadeSocial	0.75	0	contabilsoc	habilsoc
desenvolvimentoSustentado	desenvolvimentoSustentavel	0.88	0.76	desenvolvsustent	desenvolvsustenta
precoAoConsumidor	protecaoAoConsumidor	0.82	0	precaoconsum	protecaaoconsum

Tabela C.7: Termos monopalavra considerados similares pelo analisador humano e não similares pelo revisor

EO-base	EO-alvo	CC	SL	EO-base	EO-alvo
urbanizacao	reurbanizacao	0.82	0.60	urbanizaca	reurbanizaca
cineasta	videasta	0.75	0.51	cineast	videast
citricultura	cafeicultura	0.75	0	citricult	cafeicult
citricultura	floricultura	0.75	0	citricult	floricult
citricultura	silvicultura	0.75	0	citricult	silvicult
classificacao	classificados	0.77	0	classificaca	classific
comportamento	comportamentismo	0.77	0	comport	comportament
comunicacao	incomunicacao	0.82	0.60	comunicaca	incomunicaca
condecoracao	confederacao	0.75	0	condecoraca	confederaca
conhecimento	reconhecimento	0.83	0.47	conhec	reconhec
construtor	construtivismo	0.40	0.76	constru	construt
criminalizacao	descriminalizacao	0.79	0	criminalizaca	descriminalizaca
crustaceo	cretaceo	0.75	0.51	crustace	cretace
democracia	teocracia	0.78	0.55	democraci	teocraci
desburocratizacao	burocratizacao	0.79	0	desburocratizaca	burocratizaca
desnacionalizacao	nacionalizacao	0.79	0	desnacionalizaca	nacionalizaca
eletricidade	bioeletricidade	0.75	0	eletric	bioeletric
eletroencefalografia	magnetoencefalografia	0.75	0	eletroencefalograf	magnetoencefalograf
eletromecanica	aeromecanica	0.75	0	eletromecan	aeromecan
equinocultura	bovinocultura	0.77	0	equinocult	bovinocult
equinocultura	ovinocultura	0.75	0	equinocult	ovinocult
equinocultura	suinocultura	0.83	0.58	equinocult	suinocult
escritor	escritura	0.75	0.70	escri	escrit
filologia	fonologia	0.78	0.51	filolog	fonolog
fisioterapia	hidroterapia	0.75	0	fisioterap	hidroterap
geofisica	biofisica	0.78	0.47	geofis	biofis

Tabela C.8: Termos multipalavra considerados similares pelo analisador humano e não similares pelo revisor

EO-base	EO-alvo	CC	SL	EO-base	EO-alvo
acaoDeInvestigacaoDePaternidade	acaoDeInvestigacaoDeMaternidade	0.97	0.73	acadeinvestigacadepatern	acadeinvestigacadematern
administracaoContratada	administracaoCentralizada	0.83	0	administracacontrat	administracacentr
administracaoContratada	administracaoComparada	0.82	0	administracacontrat	administracacompar
administracaoDaProducao	administracaoDaEducacao	0.78	0	administracadaproduca	administracadaeducaca
administracaoDeEmpresas	administracaoDeCarreiras	0.78	0	administracadeempr	administracadecarr
administracaoDeEmpresas	administracaoDeCompras	0.82	0.30	administracadeempr	administracadecompr
administracaoDeEmpresas	administracaoDeImoveis	0.77	0	administracadeempr	administracadeimovel
administracaoDeFabricas	administracaoDeEmpresas	0.78	0	administracadefabr	administracadeempr
administracaoDeFabricas	administracaoDeMateriais	0.78	0	administracadefabr	administracademater
administracaoDeFabricas	administracaoDeServicos	0.78	0	administracadefabr	administracadeserv
administracaoDePessoal	administracaoDeCustos	0.76	0	administracadepessoal	administracadecust
administracaoDePessoal	administracaoDeVendas	0.76	0	administracadepessoal	administracadevend
administracaoDireta	administracaoIndireta	0.84	0.40	administracadiret	administracaindiret
administracaoFederal	administracaoFiscal	0.79	0	administracafeder	administracafiscal
administracaoFederal	administracaoFlorestal	0.75	0	administracafeder	administracaflorest
administracaoFederal	administracaoRegional	0.75	0	administracafeder	administracaregion
administracaoFederal	administracaoRural	0.78	0	administracafeder	administracarural
administracaoFiscal	administracaoRural	0.78	0	administracafiscal	administracarural
administracaoIndireta	administracaoDireta	0.84	0.40	administracaindiret	administracadiret
administracaoIntergovernamental	organizacaoIntergovernamental	0.76	0	administracaintergovernament	organizacaintergovernament
administracaoMilitar	administracaoPublica	0.75	0	administracamilit	administracapubl
ajustamentoSocial	ajustamentoEmocional	0.76	0.13	ajustsoc	ajustemoc
ajustamentoSocial	desajustamentoSocial	0.82	0	ajustsoc	desajustsoc
aliancaInternacional	politicaInternacional	0.80	0	aliancinternac	poliinternac
amazoniaOcidental	amazoniaOriental	0.88	0.47	amazonocident	amazonorient
ambienteDoTrabalho	higieneDoTrabalho	0.76	0	ambidotrabalh	higiendotrabalh
antropologiaCultural	antropologiaCriminal	0.75	0	antropologcult	antropologcrim
antropologiaCultural	antropologiaRural	0.76	0	antropologcult	antropologrural
antropologiaFilosofica	antropologiaBiologica	0.81	0	antropologfilosof	antropologbiolog
antropologiaFilosofica	antropologiaEcologica	0.76	0	antropologfilosof	antropologecolog
antropologiaFilosofica	antropologiaEconomica	0.76	0	antropologfilosof	antropologeconom
antropologiaFilosofica	antropologiaFisica	0.78	0	antropologfilosof	antropologfisic
antropologiaPedagogica	antropologiaBiologica	0.76	0	antropologpedagog	antropologbiolog
antropologiaPedagogica	antropologiaEcologica	0.76	0	antropologpedagog	antropologecolog
antropologiaPsicologica	antropologiaBiologica	0.86	0	antropologpsicolog	antropologbiolog
antropologiaPsicologica	antropologiaEcologica	0.86	0	antropologpsicolog	antropologecolog
antropologiaPsicologica	antropologiaEconomica	0.76	0	antropologpsicolog	antropologeconom
antropologiaPsicologica	antropologiaPolitica	0.75	0	antropologpsicolog	antropologpoli
antropologiaSocial	antropologiaRural	0.76	0	antropologsoc	antropologrural
antropologiaSocial	antropologiaVisual	0.78	0	antropologsoc	antropologvisual
aposentadoriaPorIdade	aposentadoriaPorInvalidez	0.76	0	aposentadporidad	aposentadporinvalid
armazenamentoDaInformacao	tratamentoDaInformacao	0.77	0	armazendainformaca	tratdainformaca
assentamentoUrbano	saneamentoUrbano	0.75	0	assenturban	saneurban
assistenciaEconomicaInternacional	politicaEconomicaInternacional	0.77	0	assistenceconominternac	polieconominternac
atoInternacional	tratadoInternacional	0.75	0	atointernac	tratinternac
auditoriaInterna	auditoriaExterna	0.88	0.47	auditintern	auditextern
autonomiaRegional	economiaRegional	0.81	0	autonomregion	economregion
bancoAgricola	precoAgricola	0.77	0	bancagricol	precagricol
bensImpenhoraveis	bensPenhoraveis	0.80	0.51	bemimpenhora	bempenhora
bensPenhorados	bensPenhoraveis	0.79	0.73	bempenhor	bempenhora
biologiaAnimal	citologiaAnimal	0.86	0.47	biologanimal	citologanimal
biologiaAnimal	ecologiaAnimal	0.86	0.47	biologanimal	ecologanimal
biologiaAnimal	embriologiaAnimal	0.79	0	biologanimal	embriologanimal
biologiaAnimal	fisiologiaAnimal	0.79	0	biologanimal	fisiologanimal
biologiaAnimal	histologiaAnimal	0.79	0	biologanimal	histologanimal
biologiaAnimal	virologiaAnimal	0.86	0.47	biologanimal	virologanimal
biologiaHumana	ecologiaHumana	0.86	0.47	biologhuman	ecologhuman
biologiaHmana	histologiaHumana	0.79	0	biologhuman	histologhuman
bombaNuclear	forcaNuclear	0.75	0	bombnucle	forcnucle

Tabela C.9: Termos multipalavra considerados similares pelo analisador humano e não similares pelo revisor. "continuação"

EO-base	EO-alvo	CC	SL	EO-base	EO-alvo
cambioInternacional	bancoInternacional	0.78	0	cambiinternac	bancinternac
cambioInternacional	comercioInternacional	0.79	0	cambiinternac	comerciinternac
cancaoFolclorica	dancaFolclorica	0.87	0.30	cancafolclor	dancfolclor
cancaoPopular	dancaPopular	0.83	0.30	cancapopul	dancpopul
cargoPublico	orgaoPublico	0.75	0	cargpubl	orgapubl
centroEsportivo	eventoEsportivo	0.80	0	centrespor	eventespor
circulacaoSanguinea	coagulacaoSanguinea	0.84	0	circulacasanguine	coagulacasanguine
cirurgiaEstetica	cirurgiaPlastica	0.75	0	cirurgeste	cirurgplas
civilizacaoAntiga	civilizacaoAssiria	0.76	0	civilizacaantig	civilizacaass
civilizacaoAntiga	civilizacaoGrega	0.75	0	civilizacaantig	civilizacagreg
civilizacaoAntiga	civilizacaoIndiana	0.76	0	civilizacaantig	civilizacaindi
civilizacaoAntiga	civilizacaoMinoica	0.76	0	civilizacaantig	civilizacamino
civilizacaoAntiga	civilizacaoVedica	0.76	0	civilizacaantig	civilizacavedic
codigoDeProcessoMilitar	codigoDeProcessoPenalMilitar	0.78	0	codigdeprocessmilit	codigdeprocesspenal
comercioExterno	comercioInterno	0.87	0.47	comerciextern	comerciintern
comercioInterior	comercioExterior	0.88	0.55	comerciinterior	comerciexterior
comportamentoAfetivo	comportamentoAgressivo	0.80	0	comportafet	comportagress
comunicacaoInternacional	cooperacaoInternacional	0.78	0	comunicacainternac	cooperacainternac
contratoUnilateral	contratoBilateral	0.88	0.51	contratunilater	contratbilater
criacaoDeOvino	criacaoDeBovinos	0.79	0.65	criacadeovin	criacadebovin
criacaoDeOvino	criacaoDeSuinos	0.79	0.30	criacadeovin	criacadesuin
criacaoDeSuino	criacaoDeEquinos	0.79	0.30	criacadesuin	criacadeequin
criacaoDeSuino	criacaoDeOvinos	0.79	0.30	criacadesuin	criacadeovin
culturaIndigena	pinturaIndigena	0.80	0	cultindigen	pintindigen
deficienciaAuditiva	deficienteAuditivo	0.78	0	deficiencaudi	deficiaudi
deficienciaVisual	deficienciaFisica	0.76	0	deficiencvisual	deficiencfisic
desequilibrioEconomico	equilibrioEconomico	0.84	0	desequilibeconom	equilibeconom
digestaoAnaerobia	digestaoAerobia	0.87	0.40	digestaanaerob	digestaaerob
direitoPenalInternacional	direitoInternacional	0.75	0	direitpenal	direitinternac
direitoProcessualPenalInternacional	direitoProcessualInternacional	0.83	0	direitprocesspenal	direitprocessinternac
ecologiaAquatica	biologiaAquatica	0.88	0.47	ecologaqua	biologaqua
emprestimoInterno	emprestimoExterno	0.88	0.47	emprestintern	emprestextern
estacaoDeRadio	edicaoDeRadio	0.77	0	estacaderadi	edicaderadi
estacaoDeRadio	estudioDeRadio	0.79	0	estacaderadi	estudideradi
faunaMarinha	floraMarinha	0.75	0	faunmar	flormar
filhoLegitimo	filhoIlegitimo	0.85	0.70	filhlegit	filhilegit
filmeDeLongaMetragem	filmeDeCurtaMetragem	0.80	0	filmdelongmetr	filmdecurtmetr
filosofiaClassica	mitologiaClassica	0.76	0	filosofclass	mitologclass
fontesNovasERenovaveisDeEnergia	fontesNaoRenovaveisDeEnergia	0.86	0	fontnoverenovade	fontnaorenovadeenerg
funcionarioPublicoEstadual	funcionarioPublicoFederal	0.76	0	funcionapublestad	funcionapublfeder
liberdadeDeExpressao	liberdadeDeImprensa	0.79	0	liberdaddeexpressa	liberdaddeimprens
regenciaDePedroI	renunciaDeD.PedroI	0.75	0	regencdepedri	renuncded.pedr
responsabilidadeTributaria	contabilidadeTributaria	0.78	0	responstribut	contabiltribut

Apêndice D

Extratos dos Experimentos com a Relação Semântica de Sinonímia

Tabela D.1: Mapeamentos gerados de forma incorreta com a medida CC utilizando a relação semântica de sinonímia: casos não mapeados pela medida SL (Limiar 0.8)

EO-base	EO-alvo
<T "casa"> <SN "mansao"> <SN "**solar**"> <NT "casaPropria"> </T>	<T "**molar**"> <BT "dente"> <NT "terceiroMolar"> </T>
<T "frutaCitrica"> <SN "**limao**"> <BT "fruta"> <NT "laranja"> </T>	<T "**licao**"> <BT "formaMusical"> </T>
<T "pastagem"> <SN "**pasto**"> </T>	<T "**parto**"> <BT "procedimentosCirurgicosObstetricos"> <NT "cesarea"> <NT "trabalhoDePartoInduzido"> </T>
<T "relacaoSexual"> <SN "**coito**"> <SN "relacoesSexuais"> </T>	<T "**conto**"> <BT "prosa"> </T>
<T "restaurante"> <SN "bar"> <SN "**cafes**"> <BT "estabelecimentoComercial"> </T>	<T "**cames**"> <BT "movimentosMecanicos"> </T>
<T "universo"> <SN "**cosmos**"> </T>	<T "**colmos**"> <BT "mudas"> </T>

Tabela D.2: Mapeamentos gerados de forma correta com a medida SL utilizando a relação semântica de sinonímia: casos não mapeados pela medida CC (Limiar 0.75)

EO-base	EO-alvo
<T "celulaFotovoltaica"> <SN "**celulaSolar**"> </T>	<T "**celulasSolares**"> <BT "energiaSolar"> </T>
<T "entorpecente"> <SN "alucinogeno"> <SN "drogaAlucinatoria"> <SN "estupefaciente"> <SN "narcotico"> <SN "psicotropico"> <SN "**toxico**"> <NT "cocaina"> <NT "heroina"> <NT "maconha"> <NT "opio"> </T>	<T "agenteToxico"> <BT "toxicologia"> <SN "**toxicantes**"> </T>
<T "estetica"> <SN "**beleza**"> <BT "antropologiaFilosofica"> </T>	<T "**belo**"> <BT "estetica"> </T>
<T "faunaSelvagem"> <SN "**animalSelvagem**"> <SN "animalSilvestre"> <SN "faunaDaSelva"> <BT "fauna"> <NT "macaco"> </T>	<T "**animaisSelvagens**"> <BT "animais"> </T>
<TM "**dentista**"> <SN "odontologiaComoProfissao"> <SN "odontologista"> <SN "odontologo"> <BT "pessoalDeSaude"> </T>	<TM "cirurgiao-dentista"> <BT "equipesDeSaudeBucal"> <SN "**dentista**"> <SN "odontologos"> </T>
<T "idoso"> <SN "**velho**"> <BT "grupoEtario"> </T>	<T "**velhice**"> <BT "adultos"> </T>
<T "imprensaOperaria"> <SN "**jornalOperario**"> <BT "imprensa"> </T>	<T "**jornalismoOperario**"> <BT "jornalismoEspecializado"> </T>
<TM "empresaHoteleira"> <SN "**hotelaria**"> <SN "industriaHoteleira"> </T>	<TM "**hoteis**"> <BT "edificiosResidenciais"> </T>
<TM "filhoAdotivo"> <SN "**criancaAdotada**"> <BT "filho"> </T>	<TM "**criancasAdotivas**"> <BT "membrosDaFamilia"> </T>
<TM "missoes"> <SN "**missaoReligiosa**"> </T>	<TM "**missoesReligiosas**"> <BT "religiao"> </T>

Tabela D.3: Mapeamentos gerados de forma incorreta com a medida SL utilizando a relação semântica de sinonímia: casos não mapeados pela medida CC (Limiar 0.75)

EO-base	EO-alvo
<T “aparelhoDigestivo”> <SN “**figado**”> <SN “pancreas”> <BT “anatomia”> </T>	<T “**figo**”> <BT “frutasDeClimaTemperado”> </T>
<T “coco”> <SN “**coqueiro**”> <BT “plantaOleaginosa”> </T>	<T “**coque**”> <BT “insumos”> </T>
<T “concubina”> <SN “**amante**”> <BT “concubinato”> </T>	<T “**amor**”> <BT “estadosEmocionais”> </T>
<T “dancaClassica”> <SN “**bale**”> <SN “baleClassico”> <BT “danca”> </T>	<T “**balas**”> <BT “confeitaria”> </T>
<T “eticaForense”> <SN “advocacia,EticaProfissional”> <SN “advogado,EticaProfissional”> <SN “**chicana**”> <SN “juiz,EticaProfissional”> <SN “magistrado,EticaProfissional”> <SN “magistratura,EticaProfissional”> <BT “eticaProfissional”> </T>	<T “**chicoria**”> <BT “hortalicasFolhosas”> </T>
<T “farmacia”> <SN “**drogaria**”> <BT “servicoDeSaude”> </T>	<T “**droga**”> <BT “vicio”> </T>
<T “inventarioJudicial”> <SN “**inventario**”> <BT “procedimentoEspecial”> <NT “arrolamento”> </T>	<T “**invencao**”> <BT “formaInstrumental”> <SN “invention”> </T>
<T “movimentoDeProtesto”> <SN “**protesto**”> <SN “protestoSocial”> <BT “movimentoPolitico”> </T>	<T “hormoniosProgestacionais”> <BT “hormonios”> <SN “**progesterona**”> </T>
<T “paisagem”> <SN “**vistas**”> </T>	<T “**vistoria**”> <BT “prova”> </T>
<T “peticaoInicial”> <SN “**contrafe**”> <BT “peticao”> </T>	<T “**contrabando**”> <BT “crimeContraAAdministracaoPublica”> </T>
<T “proverbio”> <SN “**ditos**”> </T>	<T “**ditador**”> <BT “governantes”> </T>