UNIVERSIDADE FEDERAL DA BAHIA

DISSERTAÇÃO DE MESTRADO

Um Método para o Povoamento de Ontologias: Extração de Textos da Web no Idioma Português

Fabio dos Santos Lima

Programa de Pós-Graduação em Ciência da Computação

Salvador
05 de Novembro de 2015

PGCOMP-Msc-2015

FABIO DOS SANTOS LIMA

UM MÉTODO PARA O POVOAMENTO DE ONTOLOGIAS: EXTRAÇÃO DE TEXTOS DA WEB NO IDIOMA PORTUGUÊS

Esta Dissertação de Mestrado foi apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal da Bahia, como requisito parcial para obtenção do grau de Mestre em Ciência da Computação.

Orientador: Laís do Nascimento Salvador

Salvador
05 de Novembro de 2015

Ficha catalográfica.

Lima, Fabio dos Santos

Um Método para o Povoamento de Ontologias: Extração de Textos da Web no Idioma Português/ Fabio dos Santos Lima– Salvador, 05 de Novembro de 2015.

75p.: il.

Orientador: Laís do Nascimento Salvador.
Dissertação de Mestrado– UNIVERSIDADE FEDERAL DA BAHIA, INSTITUTO DE MATEMÁTICA, 05 de Novembro de 2015.

1. Inteligência artificial. 2. Processamento de linguagem natural..
I. Salvador, Laís do Nascimento Salvador. II. UNIVERSIDADE FEDERAL DA BAHIA. INSTITUTO DE MATEMÁTICA. III Título.

¡NUMERO CDD¿

TERMO DE APROVAÇÃO

FABIO DOS SANTOS LIMA

UM MÉTODO PARA O POVOAMENTO DE ONTOLOGIAS: EXTRAÇÃO DE TEXTOS DA WEB NO IDIOMA PORTUGUÊS

Esta Dissertação de Mestrado foi julgada adequada à obtenção do título de Mestre em Ciência da Computação e aprovada em sua forma final pelo Programa de Pós-Graduação em Ciência da Computação da Universidade Federal da Bahia.

Salvador, 05 de Novembro de 2015

Prof. Dr. Frederico Araujo Durão
Universidade Federal da Bahia

Profa. Dra. Laís do Nascimento Salvador
Universidade Federal da Bahia

Prof. Dr. Renato de Freitas Bulcão Neto
Universidade Federal de Goiás

RESUMO

A produção e a disponibilização de informações não estruturadas (informações em formato textual) na Web aumentam diariamente. Essa abundância de informações na Web representa um grande desafio para a aquisição de conhecimento que seja processado por seres humanos e também por máquinas. Nesse sentido, diversas abordagens têm sido propostas para a extração automática de informações a partir de textos escritos em linguagem natural. Contudo, ainda existem poucos estudos que investigam a extração de informações a partir de textos escritos na língua portuguesa. Diante disso, o objetivo deste trabalho foi propor e avaliar um método não supervisionado para o povoamento de ontologias, utilizando a Web como fonte de informações no contexto da língua portuguesa. Além disso, este trabalho apresenta: (i) uma breve discussão sobre medidas de pontuação baseadas na PMI (Pontuação de Informação Mútua); (ii) novas medidas de pontuação com base na PMI e no cálculo de Desvio Padrão; (iii) uma avaliação das medidas discutidas no contexto de textos em português do Brasil extraídos da web. Os resultados obtidos com os experimentos realizados foram encorajadores e demonstraram que o método proposto obteve uma taxa de precisão média de 70% na extração de instâncias de classes ontológicas.

Palavras-chave: Ontologias, Povoamento de Ontologias, Extração de Informações

ABSTRACT

The production and availability of unstructured information (textual-format information) on the web increase daily. The abundance of this type of information represents a major challenge for the acquisition of knowledge that can be possessed by both human beings and machines. Therefore, several approaches have been proposed to automatically extract information from web texts written in natural language, many of them having English as the targeted language. However, there is still little research that investigates extraction of information from web texts written in the Portuguese language. This work aims to propose and evaluate an unsupervised method to populate ontologies using the web as a source of information in the Portuguese language context. Additionally, this work presents: (i) a brief discussion of score measures based on PMI (Score of Mutual Information); (ii) new score measures based on PMI and the standard deviation calculation; and (iii) an evaluation of the discussed measures in the context of web Brazilian Portuguese texts. The obtained results were encouraging and demonstrated that the proposed method obtained an average precision rate of 70% in extracting instances of ontological classes.

Keywords: Ontologies, Ontology Population, Information Extraction

SUMÁRIO

Capítulo 1—Introdução 1

1.1 Contextualização . . . 1
1.2 Motivação . . . 2
1.3 Problema . . . 3
1.4 Objetivos . . . 3
1.5 Metodologia . . . 4
1.6 Organização do Texto . . . 4

Capítulo 2—Revisão do estado da arte 7

2.1 Ontologias . . . 7
2.1.1 Estrutura de uma Ontologia . . . 8
2.1.2 Usos e Aplicações . . . 8
2.1.3 Linguagens para construção de Ontologias . . . 10
2.1.4 Aprendizado e Povoamento de Ontologias . . . 11
2.2 Extração de Informações - EI . . . 12
2.3 Extração de Informação Baseada em Ontologias(EIBOs) . . . 19
2.3.1 Classificação das Atuais Abordagens SEIBOs . . . 20
2.4 Trabalhos relacionados . . . 23

Capítulo 3—Método proposto 29

3.1 Método para Povoamento de Ontologias . . . 29
3.1.1 Etapa 1 - Coleta . . . 30
3.1.2 Etapa 2 - Extração . . . 32
3.1.3 Etapa 3 - Classificação . . . 35
3.1.4 Etapa 4 - Povoamento . . . 39
3.2 Validação do Método para Povoamento de Ontologias . . . 39
3.2.1 Etapa I - Coleta . . . 39
3.2.2 Etapa II - Extração . . . 42
3.2.3 Etapa III - Classificação . . . 43
3.2.4 Etapa IV - Povoamento . . . 44
3.2.5 Resultados e Discussões . . . 44
3.3 Experimentos . . . 46
3.3.1 Visão geral . . . 46
3.3.2 Experimento I . . . 49
3.3.2.1 Resultados e Discussões do Experimento I . . . 49

3.3.2.2 Conclusões do Experimento I 51
3.3.3 Experimento II 52
3.3.3.1 Resultados e Discussões do Experimento II 52
3.3.3.2 Conclusões do Experimento II 53
3.3.4 Experimento III 54
3.3.4.1 Resultados e Discussões do Experimento III 55
3.3.4.2 Conclusões do Experimento III 56
3.3.5 Comparativo entre os experimentos I, II e III 57

Capítulo 4—Conclusões e Trabalhos Futuros 61

4.1 Conclusões 61
4.2 Contribuições 62
4.3 Limitações do Método 63
4.4 Trabalhos Futuros 63

Apêndice A—Testes Estatísticos 65

LISTA DE FIGURAS

2.1 Processo de Povoamento de Ontologia adaptado de (PETASIS et al., 2011) 13
2.2 Taxonomia dos métodos de extração de informação 17

3.1 Visão geral do método para Povoamento de Ontologias 30
3.2 Fases da Etapa I - Coleta 31
3.3 Demonstração de uma consulta e as prévias retornadas pelo Bing 33
3.4 Exemplo de um Snippet armazenado em formato de texto 33
3.5 Visão geral da etapa de extração de dados - Extrator 33
3.6 Demonstração da ontologia no Protégé antes do povoamento 40
3.7 Demonstração da ontologia no Protégé após a fase de povoamento 40
3.8 Resultados do Experimento I no limiar Top 10. 50
3.9 Resultados do Experimento I no limiar Top 50. 50
3.10 Resultados do Experimento I no limiar Top 100. 50
3.11 Resultados do Experimento I no limiar Top 200. 51
3.12 Resultados do Experimento II no limiar Top 10. 52
3.13 Resultados do Experimento II no limiar Top 50. 53
3.14 Resultados do Experimento II no limiar Top 100. 53
3.15 Resultados do Experimento II no limiar Top 200. 54
3.16 Resultados do Experimento III no limiar Top 10. 55
3.17 Resultados do Experimento III no limiar Top 50. 56
3.18 Resultados do Experimento III no limiar Top 100. 56
3.19 Resultados do Experimento III no limiar Top 200. 57
3.20 Resultados Comparativos no limiar Top 10. 57
3.21 Resultados Comparativos no limiar Top 50. 58
3.22 Resultados Comparativos no limiar Top 100. 58
3.23 Resultados Comparativos no limiar Top 200. 59

LISTA DE TABELAS

2.1 Classificação do fenômeno da anáfora 16
2.2 Padrões linguísticos independentes de domínio 23
2.3 Trabalhos relacionados 27

3.1 Padrões linguísticos com a respectiva consulta formulada 32
3.2 Lista de candidatos a instância com os respectivos padrões que o extraíram 35
3.3 Lista de candidatos a instância com suas respectivas consultas 36
3.4 Quantidade de snippets coletados por classe - parte I 41
3.5 Quantidade de snippets coletados por classe - parte II 41
3.6 Percentagem documentos eliminados na Etapa I e II 42
3.7 Total de instâncias selecionadas para povoamento 43
3.8 Classes utilizadas na etapa de Classificação 47
3.9 Combinação linear utilizada no experimento III 55

A.1 P-valores do teste de Shapiro 66
A.2 P-valores do Teste-T-student 67

Capítulo 1

INTRODUÇÃO

1.1 CONTEXTUALIZAÇÃO

Com o crescimento da produção de informações em formato digital, disponibilizadas na Web, segundo Morais e Ambrósio (2007) é constante o interesse no desenvolvimento de técnicas automáticas capazes de recuperar, analisar e sumarizar esse grande volume de informações. Atualmente a Web pode ser considerada como o maior repositório de informações do mundo, contemplando os mais variados domínios do conhecimento, por exemplo, a biomedicina possui uma vasta literatura contendo informações sobre novas doenças e seus tratamentos, sintomas, microrganismos causadores de enfermidades, dentre outras informações importantes. De maneira similar, outros domínios, como o de notícias, possuem informações sobre os mais variados temas: política, esporte, economia, dentre outros.

Tais informações não são exploradas em todo o seu potencial devido à capacidade humana de processamento manual ser limitada (BEN-DOV; FELDMAN, 2010). Surge, assim, a necessidade da criação de sistemas computacionais que sejam capazes de analisar automaticamente o enorme volume de informações disponíveis na Web (MORAIS; AMBRóSIO, 2007). Entretanto, a maioria dessas informações encontradas na Web está representada em formato textual, escrita em linguagem natural, sendo destinada à consulta, à análise e à interpretação realizadas pelas pessoas (BEN-DOV; FELDMAN, 2010).

O armazenamento em formato textual não é a forma mais apropriada para o processamento computacional, uma vez que não é estruturada e não expressa explicitamente os aspectos semânticos de seu conteúdo. Para que o processamento automático dessas informações seja realizado de forma mais eficaz, é necessário que elas sejam armazenadas em um formato estruturado, ou seja, a representação dos dados permita uma melhor organização das informações armazenadas e possa também agregar um sentido semântico. Com isso, é possível que tanto pessoas quanto agentes computacionais possam analisar e extrair conhecimento relevante (W3C, 2015; FENSEL et al., 2000).

De modo a lidar com esta limitação, uma evolução da Web atual é esperada e planejada, originando o que se chama de web semântica. A visão proposta por BERNERS-LEE, HENDLER e LASSILA (2001) para a web semântica tem o propósito essencial

de explicitar o significado dos dados disponíveis na web, permitindo que as informações sejam operadas automaticamente por processos computacionais que passam a ter acesso à semântica dos dados. Para estes mesmos autores, a web semântica estende a web atual por meio da atribuição de significado bem definido à informação, permitindo que computadores e pessoas trabalhem melhor em cooperação.

Diversas técnicas podem ser aplicadas para atingir este objetivo, dentre elas, a criação manual ou automática de metadados, o uso de ontologias, a aplicação de regras lógicas (ANTONIOU e VAN HARMALEN, 2004) e a transformação de construções sintáticas em semânticas que pode ser auxiliada pela construção de bases de conhecimento.

A tarefa de construir Bases de Conhecimento (BC) é o processo de povoar um repositório de conhecimento com novos fatos extraídos a partir de uma ou mais fontes de informação. Esse processo requer o uso de técnicas de Extração de Informação (EI) e Processamento de Linguagem Natural (PLN) para analisar e transformar fontes de dados desestruturadas (textos) em um formato estruturado.

Para a realização do processo de construção de uma base de conhecimento é necessária a existência de uma estrutura básica que possa representar conceitos, relações e propriedades de um ou mais domínios. Esse alicerce pode ser representado utilizando uma ou mais ontologias. A base de conhecimento, então, instância os elementos presentes na ontologia para um determinado domínio.

O termo *Ontologia*, no contexto da Ciência da Computação, pode ser definida como especificações explícitas e formais de uma conceitualização compartilhada (STUDER et al., 2001). Uma Ontologia representa um domínio por meio de seus conceitos (classes), propriedades, relações, axiomas, hierarquia de conceitos (taxonomia de conceitos) e hierarquia de relações (taxonomia de relações).

Portanto, devido o crescimento da *Web* e da existência de poucos Sistemas de Extração de Informação Baseados em Ontologias (SEIBOs) voltados para a língua portuguesa, para tentar suprir essas demandas com as possibilidades permitidas pelo uso da semântica e da ontologia, propõe e avalia-se, por meio deste trabalho, um método não supervisionado para o povoamento de ontologias utilizando a Web como grande corpus de trabalho. Este método utiliza uma ontologia para guiar o processo de entrada, o qual, define quais os conceitos devem ser povoados e um conjunto de padrões linguísticos usados para extrair e classificar termos candidatos a instâncias.

1.2 MOTIVAÇÃO

Uma das principais tarefas para a construção e manutenção de bases de conhecimento é o *Povoamento de Ontologias* (MAYNARD; LI; PETERS, 2008). Povoamento de Ontologias é o processo de inserção de novas instâncias de classes, propriedades e/ou relações em uma ontologia existente(PETASIS et al., 2011). Além disso, essa tarefa permite relacionar o conhecimento descrito em linguagem natural com ontologias, auxiliando o processo de geração de conteúdo semântico (WIMALASURIYA, 2010). Por fim, a ontologia povoada pode ser usada em diversas aplicações, como gerenciamento de conteúdo, recuperação de informação, raciocínio automático, dentre outras.

Diante disso, surgem os sistemas para auxiliar tanto no processo de Anotação Semântica

quanto na Construção de Ontologias por meio de bases não estruturadas, os denominados Sistemas de Extração de Informação Baseados em Ontologias (SEIBOs), que visam extrair informações relevantes utilizando a ontologia para armazenar e guiar o processo de extração de informação de um domínio (WIMALASURIYA, 2010).

1.3 PROBLEMA

De acordo com BERNERSLEE, HENDLER e LASSILA (2001), para suprir os objetivos da Web Semântica, os computadores devem ter acesso a um conjunto de ontologias que permitam representar e compartilhar o conhecimento de diferentes domínios. Neste sentido, também é necessário que exista um processo de mapeamento entre essas ontologias e o conteúdo presente na Web (Anotação Semântica). E neste caso, conforme Monllaó (2011), a falta de anotação semântica do conteúdo da grande maioria das informações presentes na Web tem que ser superada.

Para automatizar os processos de Anotação Semântica e Construção de Ontologias, como afirma McDowell e Cafarella (2008), é crescente o interesse no desenvolvimento de abordagens semiautomáticas ou automáticas para extração de conteúdo semântico a partir de fontes de dados não estruturadas.

Ainda nesse sentido, observa-se que poucos trabalhos abordam textos no domínio da língua portuguesa extraídos da Web com utilização de ontologias. Tão pouco há estudos focados na língua portuguesa quando se consideram os métodos para classificar instâncias de classes ontológicas.

1.4 OBJETIVOS

O objetivo geral dessa dissertação foi o de elaborar e avaliar um método não supervisionado para o povoamento de ontologias utilizando a Web como grande fonte de informações, explorando as técnicas existentes de EI e o preenchimento de ontologias com enfoque nas ferramentas de PLN para a língua portuguesa. Para aumentar a acurácia dos dados extraídos, também foram propostas e avaliadas a alteração do normalizador das medidas de classificação baseadas na PMI para a inclusão do cálculo de desvio padrão. Além disso, são efetuadas comparações entre as principais medidas da PMI propostas e as disponíveis na literatura.

Para isso, os seguintes objetivos específicos foram definidos:

1. Elaborar um método independente de domínio para a extração de instâncias de classes ontológicas a partir de textos escritos em linguagem natural expressos em língua portuguesa presentes na Web;

2. Elaborar uma Medida de Confiança baseada na PMI que obtenha uma taxa de precisão na extração de instâncias de classes ontológicas melhor do que as medidas PMI identificadas na literatura;

3. Desenvolver um Sistema de Extração de Informação Baseado em Ontologias (SEIBO) utilizando o método proposto neste trabalho;

4. Avaliar todas as etapas do método proposto executando experimentações com as medidas heurísticas para: (i) identificar a melhor configuração para a fase de classificação do método; (ii) elencar automaticamente os candidatos classificados à instância das classes da ontologia.

1.5 METODOLOGIA

Para desenvolvimento dos objetivos propostos na seção anterior, este trabalho foi dividido nas seguintes etapas:

1. A primeira etapa foi composta por pesquisa bibliográfica sobre os trabalhos recentes a respeito da definição, utilização e preenchimento de ontologias, bem como da sua construção. Foram pesquisadas abordagens baseadas em mecanismos automáticos ou semiautomáticos. As fontes pesquisadas foram livros, teses, dissertações, artigos publicados em eventos e periódicos coletados das seguintes fontes de busca: IEEE Digital Library, ACM Digital Library, Google Acadêmico e Springer;

2. Com o objetivo principal de identificar na literatura os principais trabalhos disponíveis, suas características, arquiteturas, ferramentas e técnicas, a segunda etapa envolveu a análise das técnicas utilizadas em Sistemas de Extração com uso e suporte de Ontologias. Nela, foram também testadas, avaliadas e definidas algumas ferramentas de apoio ao desenvolvimento do método estabelecido neste trabalho, tais como frameworks de desenvolvimento de aplicativos que envolvam processamento de linguagem natural e linguagens de representação de ontologias;

3. Na terceira etapa foi definida uma metodologia para Extração de Informação Baseada em Ontologias. Diante do estudo efetuado na segunda etapa pôde-se definir um método independente de domínio capaz de extrair instâncias de classes ontológicas a partir de textos escritos em linguagem natural expressos em língua Portuguesa encontrados na Web;

4. Na quarta etapa foi constituído um ambiente de processamento linguístico no qual o método pudesse ser implementado e as diversas etapas e técnicas utilizadas pudessem ser constituídas, configuradas e analisadas;

5. Na quinta e última etapa foram realizados diversos experimentos que tiveram como principal objetivo avaliar as etapas do método proposto e verificar a acurácia das métricas de classificação analisadas e propostas.

1.6 ORGANIZAÇÃO DO TEXTO

Esta dissertação está organizada em quatro capítulos. No capítulo dois é apresentada uma revisão do estado da arte, no qual aborda-se conceitos de Ontologias, Processamento de Linguagem Natural (PLN), Extração de informações e seus relacionamentos com ontologias. Além disso, são abordados os trabalhos correlatos com extração de informação baseada em ontologias. No capítulo três é apresentada a proposta da sistemática com os

estágios de Coleta, Extração, Classificação e Povoamento. Neste capítulo, apresenta-se ainda a validação da sistemática, os resultados encontrados e os experimentos das medidas de classificação. No capítulo quatro são listadas as principais contribuições deste trabalho, além dos trabalhos futuros e as limitações.

Capítulo

2

REVISÃO DO ESTADO DA ARTE

Com o objetivo de esclarecer questões relacionadas ao conceito, caracterização, aplicação e linguagens de representações das ontologias, este capítulo aborda as definições propostas por diversos pesquisadores da área.

2.1 ONTOLOGIAS

Ontologia é um termo cunhado originalmente no campo da Filosofia e etimologicamente pode ser interpretado como o estudo (-logia) da existência ou do ser (onto-). A ontologia tem seu conceito aprimorado pela metafísica e passa a ser definida como a ciência que estuda o ser e suas propriedades, utilizando de suas categorias para classificar qualquer objeto sobre qualquer outro (CORCHO; LOPEZ; PEREZ, 2006).

Segundo Gruber (1993), uma ontologia é uma especificação formal explícita de uma conceituação compartilhada de um domínio de interesse. Conceituação refere- se a um modelo abstrato de algum fenômeno do mundo. Explícito significa que o tipo de conceitos utilizados e as limitações do seu uso, são explicitamente definidos. Formal refere-se ao fato de que a ontologia deve ser legível por máquina. Compartilhada reflete a noção de que uma ontologia captura o conhecimento consensual, isto é, não é privada de algum indivíduo, mas aceita por um grupo.

Uma definição mais pragmática é dada por Maedche e Staab (2001). Para os autores, ontologias são esquemas para representação de metadados e apresentam um vocabulário controlado para especificação de conceitos. Este esquema seria representado de forma a ser processável por sistemas computacionais.

Segundo Sowa (2000) o objetivo de uma ontologia é classificar as coisas existentes ou as que podem existir em um domínio D utilizando uma linguagem de representação R. Para Guarino (1998), uma ontologia é uma teoria lógica que compreende o significado de um vocabulário natural e o transforma em formal. Na Ciência da Computação, as ontologias são tratadas como um artefato computacional composto de um vocabulário de conceitos, suas definições e suas possíveis propriedades.

O processo de construção de uma ontologia é uma tarefa complexa e custosa, devido ao fato das ontologias serem criadas através de um consenso de diferentes visões em relação a uma área de conhecimento (Azevedo, 2008). Além disso, é necessário que um ou mais especialistas do domínio de conhecimento em questão validem os conceitos, propriedades, relações, instâncias e axiomas representados (Freitas, 2003).

Diante disso, a construção de ontologias de maneira semiautomática ou automática a partir de grandes bases de dados é uma alternativa que vem sendo investigada nos últimos anos (Cimiano et al., 2004; Etzioni et al., 2004a; Cimiano et al., 2005a; Zhou, 2007; McDowell e Cafarella, 2008). No capítulo 2, são apresentados diversos trabalhos que abordam a utilização de ontologias como processo de captação de informação para disponibilização do conhecimento na forma estruturada.

2.1.1 Estrutura de uma Ontologia

De acordo com (CIMIANO, 2006), uma ontologia é uma estrutura composta por quatro conjuntos: conceitos, relações entre conceitos, atributos de conceitos e tipos de dados.

1. Conceitos

   Um conceito é a representação de uma entidade do mundo real que foi modelada na ontologia e pode ser definido pelas suas características intrínsecas (por exemplo, uma cidade pode ser definida como uma área geograficamente delimitada com suas pessoas e histórias). Dependendo do contexto e das ferramentas utilizadas, um "conceito" pode ser referido também como uma "classe", é o exemplo de cidade, país, inseto, peixe e futebol.

2. Relações entre conceitos

   Uma relação exprime a ligação entre dois conceitos. Dentro da estrutura da ontologia, entende-se por relação toda relação binária entre conceitos, com exceção das relações taxonômicas ou hierárquicas. Exemplos: autor-de-obra (relação entre os conceitos "autor" e "obra" que denota a autoria da obra), nascido-em-cidade (relação entre os conceitos "pessoa" e "cidade" que denota a naturalidade da pessoa).

3. Atributos de conceitos

   Atributos são as características que definem um conceito. Exemplos: data de nascimento e número de páginas de livro.

4. Tipos de dados

   Tipos de dados definem como os atributos são representados. Exemplos: strings (cidade), datas (dia de nascimento) e inteiros (número de páginas de livro).

2.1.2 Usos e Aplicações

De acordo com Maedche e Staab (2001), Uschold et al. (1996), é possível relacionar a Gestão do Conhecimento, Processamento de Linguagem Natural (PLN), Comércio

Eletrônico e a Web Semântica, como algumas das principais áreas de aplicação de ontologias.

1. Gestão do conhecimento

   De acordo com Sowa (2000), a utilização de ontologias possibilita o desenvolvimento de sistemas mais inteligentes, pois o conhecimento é representado formalmente de modo a suportar extração explícita e, mais importante, extração implícita de conhecimento.

   Chandrasekaran et al. (1999) afirmam que sistemas de recuperação de informação, bibliotecas digitais, integração de fontes de informação heterogêneas e máquinas de busca na Internet são outros exemplos de segmentos que estão usufruindo de ontologias para organizar o conhecimento de um domínio e proporcionar o desenvolvimento de sistemas com melhores funcionalidades.

   O uso de ontologias pode auxiliar no fornecimento da estrutura para a construção de bases de conhecimento, visando prover o armazenamento e o conhecimento de informações com valor agregado, que resultam no desenvolvimento de estratégias que apoiam a geração de ideias e a tomada de decisão.

   Sowa (2000), Kim (2000), Fensel et al. (2001) afirmam que as ontologias servem para representar e auxiliar na gerência do conhecimento e melhorar o funcionamento de sistemas de gestão de conhecimento e comércio eletrônico.

2. Processamento de Linguagem Natural (PLN)

   De acordo com Silva et al. (2003), o PLN é uma subárea da Inteligência Artificial (IA) e tem o objetivo de desenvolver técnicas que possibilitem interpretar e gerar textos escritos em linguagem natural.

   Para efetuar alguma tarefa de PLN é muito importante ter uma compreensão coerente do texto e esse conhecimento pode ser representado por meio de uma ontologia. Nesse contexto, o uso de ontologias é importante para auxiliar na desambiguação durante o processo de interpretação do texto, proporcionar um dicionário de conceitos relevantes para o domínio do texto e também para facilitar a identificação de categorias semânticas envolvidas no universo de discurso de um determinado domínio (MAHESH, 1996).

3. Comércio Eletrônico

   Nesta área, especificamente no Business to Business (B2B), de acordo com Maedche e Staab (2001), para a automatização das transações é necessária, muitas vezes, além da padronização do formato de trocas sintáticas, a descrição formal dos processos. Isso provê um entendimento comum dos termos, permitindo melhor interoperabilidade e integração inteligente de informações, consequentemente recorrendo a recursos como a ontologia. Ainda nesse sentido, as ontologias podem auxiliar na área de Business to Consumer (B2C), de forma a solucionar as dificuldades existentes na construção de agentes de busca na Web e facilitar a visualização e recuperação automática de informação.

4. Web Semântica

   BERNERSLEE, HENDLER e LASSILA (2001) caracterizam a Web Semântica como uma evolução do cenário da atual Web, com o principal objetivo de tornar explícito a semântica dos dados disponíveis em seu conteúdo, de forma a criar ambientes em que agentes computacionais e usuários possam trabalhar de forma cooperativa. Essa nova visão possibilita que os agentes computacionais possam interpretar as informações e resolver problemas de recuperação de informação classificados como complexos.

   A Web Semântica é formada por uma arquitetura organizada em camadas, de forma que cada camada tenha um maior nível de expressividade e inferência (KOIVUNEN; MILLER, 2001). Uma das camadas fundamentais no desenvolvimento da Web Semântica é composta por ontologias, sendo elas responsavéis por fornecer a expressividade necessária à representação do conhecimento relevante sobre um domínio (FREITAS, 2003). Além disso, para se alcançar os objetivos da Web Semântica é necessário que os computadores tenham acesso a um conjunto de modelos que permitam representar o conhecimento de um domínio.

   Nesse sentido, BERNERSLEE, HENDLER e LASSILA (2001) partiu do princípio que todo recurso Web, ou seja, qualquer conteúdo publicado na Web, necessita de um Uniform Resource Identier (URI) único. Estas entidades, assim como são fundamentais para toda a Web, formam a base da Web Semântica, pois nomeiam univocamente todo e qualquer recurso da Web.

2.1.3 Linguagens para construção de Ontologias

Dependendo da tarefa requerida, diferentes tipos de conhecimentos precisam ser representados. Na próxima seção, demonstram-se diversas linguagens utilizadas para a criação de ontologias disponibilizadas na literatura, entre elas: RDFSchema, OWL e OWL2.

A escolha de qual linguagem será utilizada para criação de uma ontologia é uma importante decisão a ser tomada, já que ela afeta todo ciclo de vida dessa ontologia (YILDIZ; MIKSCH, 2007). De acordo com Grigoris e Harmelen (2004), os principais requisitos para a escolha de uma linguagem de representação de ontologias são sintaxe e semântica bem definidas, apoio ao raciocínio de forma eficiente e expressividade suficiente para a tarefa a ser desenvolvida.

A linguagem recomendada pelo W3C - World Wide Web Consortium e também a mais adotada (CARDOSO, 2007) para representação de ontologias é a OWL (Web Ontology Language) (W3C, 2004) e, foi a linguagem adotada neste trabalho devido suas características favorecerem sua manipulação pelas maiorias das ferramentas Opensource.

1. RDFSchema é uma linguagem declarativa para definição de esquemas em RDF (AMANN; FUNDULAKI, 1999). O modelo de dados do RDF(S), pelo fato de ser baseada em frames e oferecer mecanismos para a definição de relacionamentos entre propriedades e recursos, foi amplamente usada como formato de representação em muitas ferramentas e projetos.

2. OWL também é uma proposta do W3C para representação de ontologias (SMITH; WELTY; MCGUINNESS, 2004). Foi concebida para viabilizar a Web Semântica de forma a facilitar às máquinas o processamento e a integração de informações disponíveis na Web. Como parte dos requisitos básicos que orientaram essa concepção, a OWL foi desenvolvida para contemplar as camadas de linguagens da Web Semântica, incluindo XML e RDF.

   Através do uso de URIs (Universal Resource Identifiers) e RDF como base para nomeação e identificação de recursos, a OWL permite a distribuição de uma ontologia através de vários sistemas, e promove o equilíbrio entre o poder de expressividade da linguagem e a eficiência que ela oferece no suporte à inferência e raciocínio.

3. OWL2 é uma melhoria da linguagem OWL (SMITH; WELTY; MCGUINNESS, 2012). Foi recomentada pelo W3C em Dezembro de 2012 e agregou novas funcionalidades como: alteração da sintaxe da disjunção, união de classes na expressividade sintática referente as chaves, cadeias de propriedade, tipos de dados, intervalos de dados, restrições de cardinalidade e capacidades de anotação.

   A OWL2 contém três perfis diferentes a OWL2 EL, OWL2 QL e OWL2 RL. Cada perfil é definido como uma restrição sintática da OWL2 e como um subconjunto dos elementos estruturais que podem ser usados em uma ontologia. Assim, cada perfil também é mais restritivo do que o perfil OWL, pois os perfis oferecem diferentes aspectos de expressividade em troca de diferentes benefícios computacionais e de implementação (SMITH; WELTY; MCGUINNESS, 2012).

   - OWL2 EL: adequada para desenvolvimento de grandes ontologias, situações em que o poder expressivo pode ser trocado por garantias de desempenho permitindo a utilização de algoritmos de tempo polinomial para todas as tarefas de raciocínio padrão;
   - OWL2 QL: recomendada para aplicações de ontologias relativamente leves e para organizar um grande número de processos que permitem acessar os dados diretamente através de consultas relacionais (por exemplo, SQL);
   - OWL2 RL: também adequada para aplicações relativamente leves, mas, que auxiliam as operações sobre os dados no formato RDF.

2.1.4 Aprendizado e Povoamento de Ontologias

A construção e a manutenção de ontologias são atividades que requerem muito esforço e tempo. Além de cara, a modelagem feita à mão por um especialista humano pode apresentar erros e ser influenciada demasiadamente por sua experiência própria, podendo causar divergências quanto à interpretação predominante entre a maioria dos especialistas do domínio analisado.

Dadas essas dificuldades, é ainda bastante expressiva a quantidade de domínios para os quais não existem ontologias modeladas e reconhecidas como adequadas para a representação do conhecimento de uma área pelos seus respectivos especialistas.

Esses fatores constituem inibidores bastante poderosos para uma utilização mais massiva e ubíqua de ontologias. No intuito de tratar esses problemas, o campo da aprendizagem de ontologias surgiu com a finalidade de desenvolver técnicas e abordagens automáticas de construção de ontologias.

O termo Aprendizado de ontologias foi cunhado por Maedche e Staab (2001) justamente na época em que a Web Semântica ganhou status de tema de pesquisa e as ontologias despontaram como um potencial formalismo para representação de conhecimento neste ambiente.

Aprendizado de ontologias é a sua construção automática, evitando ou minimizando trabalho manual do especialista. Geralmente, esta automatização dá-se por meio de técnicas de Processamento de Linguagem Natural ou Aprendizado de Máquina (GóMEZ-PÉREZ; FERNáNDEZ-LóPEZ; CORCHO, 2007).

Já a definição dada por Gómez-Pérez, Fernández-López e Corcho (2007) é mais detalhada, incluindo diferentes técnicas de manipulação e criação destes artefatos. Para os autores, existem quatro principais formas de aprendizado de ontologias: (1) Aprendizado de ontologias com o uso de corpus em que os textos são lidos processados e extraídos os conceitos, instâncias e relacionamentos; (2) Aprendizado de ontologias a partir de instâncias; (3) Aprendizado de ontologias a partir de modelos, tais como Entidade-Relacionamento (ER); (4) Aprendizado de ontologias que visa à interoperabilidade e abrange o mapeamento das entidades, conceitos e instâncias entre duas ontologias.

O povoamento de ontologias é uma tarefa que pode ser vista de forma separada, descrita como sendo um processo no qual relações e conceitos são aprendidos (CIMIANO, 2006).

De acordo com Petasis et al. (2011), o processo de povoamento de ontologias não altera a estrutura da ontologia, modifica-se apenas o conjunto de instâncias, conceitos, relações e propriedades .

Observa-se na Figura 2.1 um processo de Povoamento de ontologias.

Na Figura 2.1, a ontologia de entrada será povoada a partir da saída do processamento de povoamento. Neste processo, tem-se como entrada a ontologia e o corpus. Na primeira etapa são extraídos as instâncias/classes e relações que servem como entrada da segunda etapa. Esta tem a função de efetuar a classificação das instâncias e enviá-las para a etapa final, que é responsável efetivamente pelo processo do povoamento da ontologia.

2.2 EXTRAÇÃO DE INFORMAÇÕES - EI

A EI identifica dados relevantes a um tema contido em um texto e os extrai convertendo-os para uma estrutura tabular. Esta estrutura tem o objetivo de sumarizar o conteúdo do assunto abordado no documento em uma forma legível, tanto para o usuário quanto para uma aplicação.

De acordo com Grishman (1997), as primeiras ideias de estruturação da informação em linguagem natural datam da década de 50, mas apenas no final da década de 80 a EI se destacou no meio científico por meio de desafios lançados pela Message Understand Conference (MUC). O desafio constava em entender o conteúdo de mensagens contidas em textos. Algoritmos propostos teriam de ser eficientes no entendimento do significado

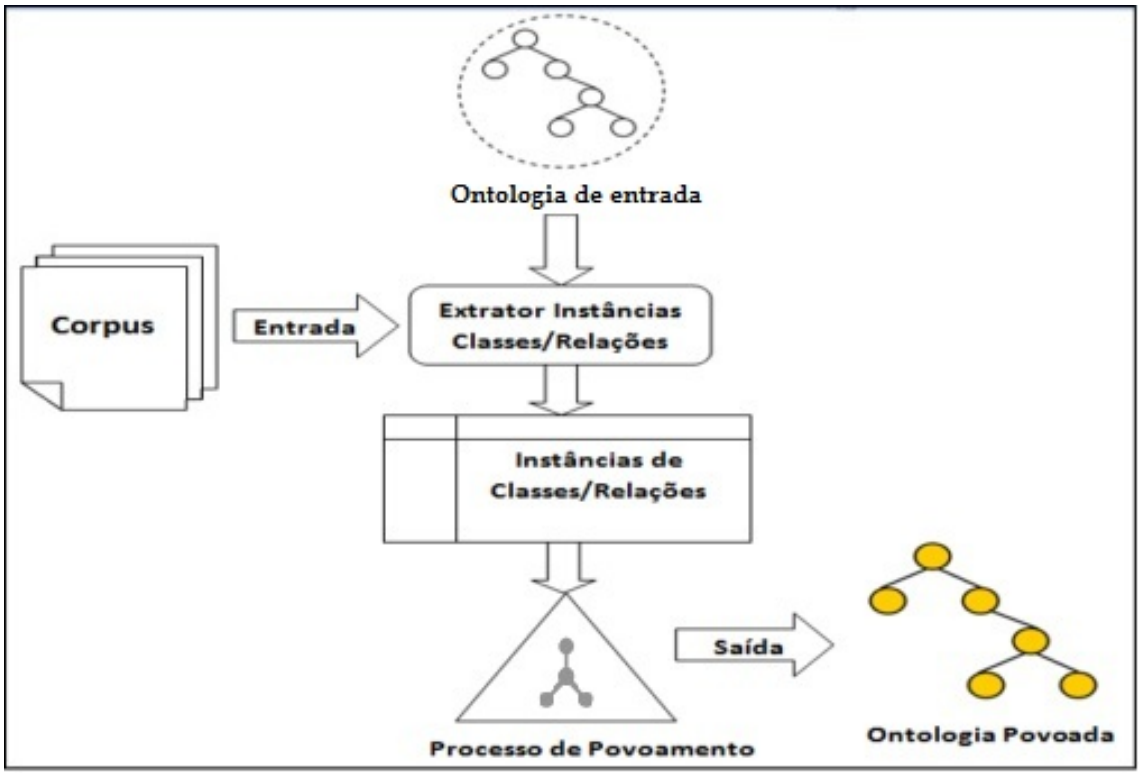

Figura 2.1 Processo de Povoamento de Ontologia adaptado de (PETASIS et al., 2011)

das mensagens.

Jacobs e Rau (1993) afimam que o objetivo inicial da Extração de Informações (EI) era a de povoar automaticamente bases de dados. Entretanto, sistemas de extração de informações também permitem melhorar o desempenho de sistemas de recuperação de informações por meio da integração e sintetização da informação, evitando a ocorrência de redundâncias em textos que tratam do mesmo assunto.

De acordo com Wives e Loh (1999), a EI foi uma evolução natural da área de recuperação de informações e cada dia aumenta a quantidade de informações disponíveis nos meios eletrônicos, sendo que uma porcentagem significativa das mesmas é composta de informações que, de alguma forma, podem ser estruturadas ou inter-relacionadas.

1. Estrutura dos Documentos

   Diante da diversidade e da quantidade de informações disponíveis em formato digital, surgiram também diferentes formas de estruturação desses documentos.

   A formação do conjunto de documentos, denominado corpus, é um importante item para os sistemas de extração, sendo eles utilizados como fonte de informações para iniciar o processo de extração.

   Os sistemas de extração buscam determinados elementos contidos nesses tipos de documentos para tentar auxiliar no processo de identificação de informações relevantes. Diante disso, as técnicas usadas para realizar o processo de identificação da

informação dependem diretamente da estrutura da fonte de informação processada. Maedche e Staab (2004) sugerem uma classificação dessas abordagens segundo a base utilizada na construção da ontologia e que pode ser composta por dados estruturados (como aqueles que se mantém em bases de dados), semiestruturados (como aqueles expressos em HTML) ou textos em linguagem natural.

(a) Fontes Estruturadas

Devido à estruturação dos dados, este tipo de fonte está associada com banco de dados relacionais, assim, observa-se que os significados dos dados podem ser facilmente atribuídos aos mesmos de forma que a extração de informação relevante torna-se uma tarefa simples. As informações armazenadas em arquivos escritos em XML são exemplos desse tipo de fonte (MAEDCHE; STAAB, 2004).

A criação de ontologias a partir de dados estruturados permite principalmente a extração de conceitos e relações contidos em bases de dados. Esse tipo de ontologia tem sido usada como uma forma de mediação ou integração entre bases de dados.

(b) Fontes Semiestruturadas

Neste tipo de fonte existem algumas estruturas nos documentos, como os indicadores para encontrar as informações desejadas. Exemplos dessas fontes são as páginas escritas em HyperText Markup Language (HTML). Este tipo de documento permite, através das etiquetas HTML, maior facilidade para localizar as informações desejadas (MAEDCHE; STAAB, 2004).

A criação de ontologias a partir de dados semiestruturados consiste na utilização de novos padrões para a publicação de documentos na web. Esses novos padrões têm alterado a maneira como se disponibilizam informações na web e aumentado a disponibilidade de dados semiestruturados, bem como as definições formais para esses dados, o que incorpora algum nível de expressividade semântica aos documentos.

(c) Fontes Desestruturadas

Diferentemente das outras fontes, a fonte desestruturada não possui algum tipo de estruturação e não contribui nem auxilia no processo de extração de informação. A criação de ontologias a partir de fontes não estruturadas envolve a captura de textos, utilizando-se de técnicas de processamento de linguagem natural. Esses textos apresentam vários níveis de informação que são representadas através de características e restrições sintáticas, morfológicas, semânticas e pragmáticas; atributos que convergem para expressar significados particulares. Ferramentas que aprendem ontologias a partir de linguagem natural utilizam a interação entre essas características e restrições para identificar conceitos e estabelecer relações entre eles (MAEDCHE; STAAB, 2004).

2. Funções de Apoio à Extração de Informações

Conforme CUNNINGHAM (2006), EI tem o objetivo de encontrar cinco tipos de informações nos textos:

a)Identificação e Classificação de Entidades

Para a classificação de entidades, a primeira etapa consiste na identificação dos nomes próprios que podem ser de pessoas, lugares e organizações. Em seguida, é atribuída a classe correta ao nome próprio identificado. Esta tarefa é também conhecida como Reconhecimento de Entidades Mencionadas (REM) (FERREIRA; BALSA; BRANCO, 2007) considerada uma das tarefas mais importantes no processo de EI, pois permite identificar e classificar instâncias do domínio de interesse.

A tarefa de identificação depende pouco do domínio que está sendo tratado, por exemplo, mudar de notícias sobre esportes para notícias financeiras, mas pode depender bastante do tipo de estrutura do texto, por exemplo, um texto científico pode exigir um tipo de extrator de entidades, enquanto um texto jornalístico pode demandar outro tipo. A tarefa de classificação depende do domínio, pois a estrutura de classificação reflete os conceitos específicos do domínio. Uma entidade dentro do contexto de EI é correspondente a um conceito (classe) da ontologia.

b)Identificação de Sintagmas Nominais

A unidade sintagmática é considerada um agrupamento intermediário entre o nível do vocábulo e o da oração. Desta maneira, um ou mais vocábulos se unem (em sintagmas) para formar uma unidade maior que é a oração. Os vocábulos que compõem a unidade sintagmática se organizam em torno de um núcleo, a depender do núcleo, é possível falar em sintagma nominal e sintagma verbal. Por exemplo, na oração “O especialista não respondeu todas as perguntas” são sintagmas nominais “O especialista” (sendo o núcleo “especialista”) e “todas as perguntas” (sendo o núcleo “perguntas”).

c)Etiquetagem de Papéis Semânticos

De acordo com Sánchez (2007), a etiquetagem de papéis semânticos é o processo de atribuição de representações na forma de papéis semânticos ou temáticos para as porções do discurso. Este processo tem o objetivo de identificar os argumentos associados a um verbo que funciona como âncora do processo de etiquetagem, tais como o agente que executa a ação descrita pelo verbo, o paciente que sofre a ação do verbo, os instrumentos utilizados para executar a ação e complementos (adjuntos) que caracterizam tempo, modo, lugar, entre outros.

d)Resolução de Anáforas

Anáfora é o fenômeno linguístico em que um trecho do discurso (referência) remete a um referente que pode já ter sido mencionado, que ainda será mencionado ou que é externo ao discurso (CHAVES; RINO, 2008). A posição do referente e existência explícita da referência determina a classe do fenômeno da anáfora, como mostrado na Tabela 2.1

Referente	**Referência Explícita**	**Tipo**	**Exemplo**
Ocorre antes da referência	Sim	Anáfora	Antonio Viajou. Ele foi de ônibus
Ocorre depois da referência	Sim	Catáfora	Achando que Antonio tinha ido embora. Aquiles retornou
Externo ao discurso	Sim	Exófora	Aquele carro é do Presidente
Ocorre antes da referência	Não	Elipse	Antonio chegou agora. (...) Estava muito cansado.

Tabela 2.1: Classificação do fenômeno da anáfora

A tarefa de resolução de anáforas divide-se em duas etapas: identificação das referências e resolução das anáforas, isto é, identificação dos referentes e associação às referências corretas.

A etapa de identificação das referências é responsável pela marcação de todos os trechos do texto que são referências anafóricas. Já a etapa de resolução de anáforas identifica candidatos a referentes e seleciona o mais provável para cada referência através de algoritmos de classificação.

e)Desambiguação do Sentido das Palavras

Dependendo do contexto uma mesma palavra pode ter significados diferentes, fenômeno este conhecido como polissemia, por exemplo, nas frases abaixo a palavra "manga" tem significados diferentes:

- ex1. "Preciso de uma camisa de manga curta (nesta frase "manga" é um substantivo, com significado de um componente de camisa);
- ex2. "Esta manga está madura (neste período "manga" é um substantivo, com significado distinto do anterior, de um tipo de fruta);
- ex3. "A moça manga do rapaz (neste contexto é um verbo, equivalente a "caçoar").

Para a EI é importante que a identificação do significado seja condizente com o contexto. Informações da classe gramatical da palavra no contexto e de termos que ocorrem conjuntamente no discurso são utilizadas para resolver a ambiguidade. No exemplo **ex3**, a classe gramatical seria suficiente para desambiguar os sentidos. Já nos casos **ex1** e **ex2** seria necessário identificar o contexto que está sendo utilizado, uma vez que os dois significados distintos são substantivos.

3. Taxonomia de Sistemas de Extração de Informações

   Sistemas de Extração de Informações (SEI) podem ser classificados de acordo com seu método de extração. Dois grupos principais podem ser definidos: sistemas de EI

baseados em conhecimento e sistemas de EI baseados em Wrappers. A taxonomia dos métodos de extração é apresentada na Figura 2.2

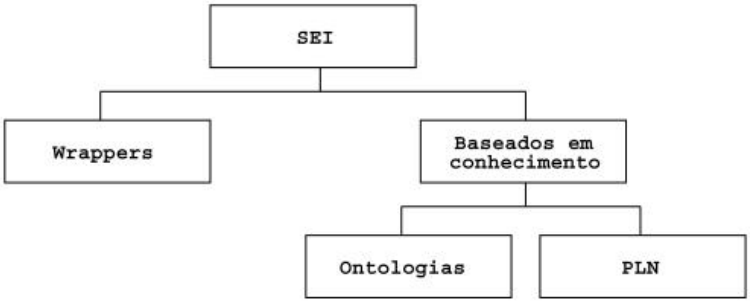

Figura 2.2 Taxonomia dos métodos de extração de informação

a)SEI baseados em conhecimento

Os sistemas de EI baseados em conhecimento relacionam informações do texto a bases de conhecimento e podem ser classificados da seguinte maneira:

- PLN - Processamento de Linguagem Natural
  Sistemas de EI baseados em PLN usam aprendizado para aquisição de conhecimento aplicando técnicas como filtragem, análise de classes de palavras (substantivos, verbos, predicados), análise léxica, semântica e relacionamento entre termos e sentenças. Uma desvantagem do PLN está na complexidade do desenvolvimento de regras e em sua manutenção, o que eleva os custos desses tipos de sistemas.
- Ontologias
  Silva et al. (2003), Embley et al. (1998) classificam o uso de um domínio de conhecimento como estruturados e declarativos, permitindo, assim, que seus conceitos possam ser descritos em ontologias representadas por quadros. Desta forma, o conhecimento é constituído por classes, subclasses, instâncias das classes e relacionamentos entre as classes, o que provê uma estrutura bem definida, possibilitando inferência de regras e diversos níveis de granularidade do conhecimento. Como consequência, ganha-se em reusabilidade de conhecimento e de regras, modularidade, abstração e herança (NOY; MCGUINNESS et al., 2001).

b) SEI Wrappers

Sem dependência da EI tradicional, esta abordagem é aplicada baseando-se nos marcadores da linguagem HTML e é considerada a mais usada na Web (MUSLEA; MINTON; KNOBLOCK, 1999). Wrappers são considerados sistemas de EI que atuam em textos semiestruturados que identificam dados de seu interesse e os mapeiam para um formato estruturado.

Um Wrapper consiste de um conjunto de regras e uma coleção de expressões para aplicar às regras. Geralmente, esta coleção contém marcadores da linguagem HTML

e extrai as informações contidas nos textos indicando seu conteúdo e sua posição. Este resultado não traz nenhuma conotação semântica e não tem boa restrição do domínio de aplicação, pois o mesmo é definido pela estrutura de formatação textual utilizada na construção das páginas.

4. Medidas de Avaliação de Sistemas de Extração de Informações

   Diante da importância de comparar a eficácia dos sistemas de extração de informação, surgiu, nas conferências MUC - Message Understanding Conferences (KAISER; MIKSCH, 2005) a necessidade da criação de medidas padronizadas. De acordo com Lavelli et al. (2008), dentre as várias medidas propostas destacam-se as medidas de Precisão e Cobertura.

   A primeira medida é utilizada para avaliar a corretude de um SEI e é calculada pela relação entre a quantidade de informações extraídas e o número total de informações extraídas (corretas + incorretas), já a Cobertura avalia a completude de um SEI, nesta calcula-se determinando a relação entre o total de informações corretas extraídas e o total de informações corretas presentes no corpus processado. Na Equação 2.1 e 2.2, pode-se observar o cálculo dessas medidas.

$$precisao = \frac{numero\_de\_elementos\_corretos}{numero\_de\_elementos\_extraidos} \tag{2.1}$$

$$cobertura = \frac{numero\_de\_elementos\_corretos}{numero\_de\_possiveis\_elementos\_no\_texto} \tag{2.2}$$

   Diante do fato de que a Cobertura e a Precisão são inversamente relacionais, torna-se complexo conseguir melhorar ambas as medidas ao mesmo tempo. Quando se prioriza a precisão, devem-se extrair poucas informações que possuam um maior grau de certeza, isso possibilita que diversas informações corretas sejam ignoradas no processamento do corpus. Já priorizando a cobertura, extrai-se uma grande quantidade de informações, que por sua vez, aumenta o grau de informações extraídas incorretamente.

   Diante disso e da tarefa de extração desejada, faz-se necessário definir o aspecto a ser priorizado: a Completude ou a Corretude, essa problemática impulsiona a utilização da Medida-F que visa mensurar o impacto da Cobertura e da Precisão em uma única medida. Sua formulação pode ser observada na Equação 2.3.

$$MedidaF = \frac{2 * Precisao * Cobertura}{Precisao + Cobertura} \tag{2.3}$$

2.3 EXTRAÇÃO DE INFORMAÇÃO BASEADA EM ONTOLOGIAS(EIBOS)

O problema de Extração de Informação vem ganhando atenção há mais tempo que o campo das ontologias e a convergência destas duas áreas deu origem aos processos de Extração de Informação que usam ontologias.

As ontologias podem interagir com o processo de Extração de Informações de duas maneiras distintas, mas, que podem ser combinadas (LI; BONTCHEVA, 2007). A primeira é como repositório para armazenamento das informações extraídas e tal processo é conhecido como extração de informações orientada a ontologias. Neste caso a ontologia somente é usada como destino do processo de EI. Já na segunda maneira a ontologia é usada para armazenamento do processo de Extração de Informação e também como fonte de informações. De acordo com Yildiz e Miksch (2007), esta abordagem é chamada de extração de informações baseada em ontologias(EIBO).

De acordo com Buitelaar et al. (2008), Petasis et al. (2011), Sistemas de Extração de Informações auxiliam entre outros fatores nas tarefas de: (a) Aprendizagem de Ontologias; (b) Enriquecimento de Ontologias, ou seja, extensão de uma ontologia existente com novas classes e relações organizadas taxonomicamente; (c) Povoamento de Ontologias, nesta, existe a inserção de novas instâncias de classes, relações e propriedades; e (d) Anotação Semântica, nesse processo correlaciona-se o conhecimento representado por meio de ontologias a fragmentos encontrados em textos.

Quando as ontologias são utilizadas para descrever quais informações são relevantes para o domínio em análise e integrar o conhecimento extraído à ontologia de entrada, tem-se um processo de extração de informação guiado por ontologias, denominado Extração de Informação Baseada em Ontologias (EIBO). Estes podem envolver as ontologias no processo de EI com conceitos das áreas: (a) da Web Semântica (BERNERSLEE; HENDLER; LASSILA, 2001); (b) Extração de Informação (RILOFF, 1994), (c) Processamento de Linguagem Natural (SILVA et al., 2003); (d) Aprendizagem de Máquina (MITCHEL, 1997), dentre outras.

Neste sentido, os Sistemas de Extração de Informação Baseado em Ontologias (SEIBO) capturam e processam textos oriundos de fonte de dados desestruturado ou semiestruturado e utilizam de mecanismos direcionados por ontologias para extrair e apresentar essas informações.

Alguns trabalhos como os de Bontcheva et al. (2006), McDowell e Cafarella (2006) têm abordado o uso de ontologias tanto no armazenamento do processo de EI, como também para guiar a Extração. O uso de ontologias no processo de EI apresenta as seguintes diferenças para o processo tradicional de EI:

- O esquema da ontologia e suas instâncias substituem o uso de léxico e gazetteers lineares;
- A hierarquia de conceitos da ontologia é utilizada no processo de classificação das instâncias extraídas;
- Inferência sobre os conceitos representados na ontologia podem ser utilizados durante o processo de EI.

2.3.1 Classificação das Atuais Abordagens SEIBOs

Nessa seção, apresenta-se a classificação dos atuais Sistemas de Extração de Informação Baseado em Ontologias(SEIBO). De acordo com Wimalasuriya (2010), essa classificação aborda um longo conjunto de dimensões e visa proporcionar um melhor entendimento das técnicas, ferramentas e métodos propostos. Alguns deles já foram utilizados por tradicionais SEIs e inclusive aplicados nas tarefas de Aprendizagem de Ontologias, Enriquecimento de Ontologias, Povoamento de Ontologias e Anotação Semântica.

De acordo com Wimalasuriya (2010), seis métodos principais de extração são utilizados por SEIBOs:

1. Regras linguísticas representadas por expressões regulares

   Essa técnica permite a especificação de expressões regulares capazes de capturar certos tipos de informação, por exemplo, a expressão (assistido — visto) denota uma frase substantiva e pode capturar nomes de filmes (representados pelo sintagma nominal) em um conjunto de documentos.

   Através da especificação desse tipo de regras é possível extrair uma quantidade significativa de informação. O conjunto dessas expressões regulares podem ser implementadas utilizando o Transdutor de estados finitos que consiste em uma série de autômatos de estados finito e podem ser combinados com as ferramentas de PLN para formarem uma variedade de regras.

   Alguns exemplos de sistemas implementados com essa regra são: (a) o sistema FASTUS IE (APPELT; ISRAEL, 1999), implementado em 1993, apresenta-se como um dos primeiros sistemas a utilizar esse método; (b) A Arquitetura Geral de Engenharia de Texto (GATE) (CUNNINGHAM, 2002) que utiliza amplamente técnicas de PLN com expressões regulares com a finalidade principal de facilitar a utilização de técnicas de extração de informação; (c) o sistema OBIE de Embley (EMBLEY, 2004) que combina expressões regulares com elementos de ontologias, resultando em uma "extração de ontologias"; (d) o sistema Crystal (SODERLAND et al., 1995), visa identificar regras de extração e procura pelas generalizações mais específicas baseadas nos princípios do algoritmo de aprendizado indutivo.

2. Gazetteer List

   As palavras que devem ser reconhecidas são apresentadas para o sistema em forma de lista, denominadas "gazetteer list". É uma técnica amplamente utilizada para reconhecimento de entidade nomeada (REN) que pode ser vista como um complemento de extração de informação, e é utilizada na identificação individual de uma categoria particular que pode ser usada para reconhecer entidades, como, cidades de um determinado Estado, cidades do Brasil, rios, armas, entre outras.

   Existem sistemas que frequentemente usam "gazetteer lists" para preencher as instâncias das classes na ontologia, como o Sistema de Buitelaar e Siegel (2006) que captura detalhes sobre partidas de futebol e o de Saggion et al. (2007) utilizado para obter informações sobre países e regiões.

3. Técnicas de Classificação

   A classificação no contexto de um EIBO consiste basicamente no treinamento de classificadores para identificar diferentes componentes de uma ontologia, como as instâncias de classes, relações entre conceitos e valores de propriedades.

   Algumas características linguísticas como etiquetas de POS e palavras individuais são usadas como atributos para algoritmos de classificação, o que, de certa forma, possibilita a tarefa de extração de informação ser apresentada como um conjunto de tarefas de classificação.

   Alguns Sistemas de Extração de Informação utilizam diferentes técnicas de Aprendizagem de Máquina, como Máquina de Vetores de Suporte (Support Vector Machines SVM), Máxima Entropia (ME) (Maximum Entropy Models), Árvores de Decisão (Decision Trees), Cadeias de Markov (Hidden Markov Models) e Campos Aleatórios Condicionais (Conditional Random Fields).

   O sistema denominado The Kylin (WU; WELD, 2007) emprega duas técnicas de classificação. A primeira utiliza o modelo de entropia máxima para predizer quais são os valores dos atributos presentes em uma sentença. Já a segunda utiliza o modelo CRF para identificar atributos dentro da sentença. Outro exemplo pode ser observado no algoritmo de Hieron Large Margin (Grande Margem de Hieron) que faz extração de instâncias aplicando uma classificação com base na hierarquia da ontologia de entrada (LI; BONTCHEVA, 2007).

4. Construção de Árvores de Análise Parcial (Partial Parse Trees)

   Neste tipo de abordagem, a maioria dos SEIBOs são projetados para formar uma árvore semanticamente anotada para representar as relações entre as palavras do texto e o processo de Extração de Informação. Esta técnica tem o objetivo de fornecer uma análise minuciosa de cada sentença encontrada, além disso, demonstra a ocorrência das expressões regulares encontradas.

   Um exemplo desse tipo de SEIBO é Saarbücker Message Extracting System (SMES) (MAEDCHE; NEUMANN; STAAB, 2003) o qual utiliza uma ferramenta de PLN para o idioma alemão e retorna uma árvore de dependência que é basicamente uma árvore de análise parcial que pode ser utilizada no processo de extração de informação para a tarefa de Aprendizagem de Ontologias.

5. Analise de Etiquetas HTML/XML

   Da mesma forma que o Sistema de Extração de Informação, os SEIBOs podem explorar as etiquetas presentes em documentos semiestruturados para encontrar as informações desejadas. Na subseção 3.3.3.2, são discutidos os Wrappers, sistemas que utilizam dessa técnica para identificar tabelas presentes em páginas HTML.

   De acordo com Buitelaar e Siegel (2006) o SEIBO, denominado SmartWeb Ontology-Based Annotation (SOBA), é um exemplo desse tipo de sistema, ele explora informações de etiquetas presentes nas tabelas de páginas HTML e efetua automa-

ticamente o povoamento de uma base de conhecimento baseado nas informações extraídas.

6. Métodos Baseados em Pesquisas na Web

   Nesse tipo de abordagem, o princípio básico é a exploração da Web como fonte de informações. Para isso e de forma a auxiliar o processo de captura de informações relevantes, utilizam-se regras linguísticas de extração que são transformadas em consultas e aplicadas a algum mecanismo de busca na Web (WIMALASURIYA, 2010).

   Considerando que a Web pode ser considerada o maior repositório de informações disponíveis, ela é identificada como uma importante fonte de conhecimento, sendo útil para tarefas como extração de informação (BRILL, 2003). Contudo, essa característica impõe limitações a diversas técnicas clássicas de extração de informação, justamente pelo tamanho do corpus analisado.

   Outra característica importante presente na Web, de acordo com Brill (2003), Etzioni et al. (2004) é a alta redundância presente em seu conteúdo, considerado como uma importante propriedade já que essa quantidade de repetições de uma informação pode representar um medida de sua relevância. Diante dessas características, ultimamente a Web tem sido utilizada como corpus em diversos trabalhos (CIMIANO; HANDSCHUH; STAAB, 2004; ETZIONI et al., 2004; MCDOWELL; CAFARELLA, 2006; CARLSON et al., 2010; TOMAZ et al., 2012).

   Apesar disso, a utilização da Web como gerador de corpus ainda apresenta alguns desafios, como a falta de estrutura dos documentos disponíveis, informações não confiáveis, ruídos devido a representações visuais, além do problema da ambiguidade presente na linguagem natural (SáNCHEZ, 2007).

   O método proposto neste trabalho utiliza-se de mecanismos de extração tendo a Web como fonte de informações. Para tal, é aplicado um conjunto de padrões léxicos sintáticos para extrair e classificar candidatos a instâncias de classes de uma ontologia de domínio genérico.

   - Padrões Léxicos Sintáticos

     Hearst (1992) propôs padrões léxicos sintáticos para a descoberta de relações semânticas de hiponímia e de hiperonímia. Hiponímia é relação que denota que um conceito é subclasse de outro conceito, ou seja, um conceito mais específico é chamado de hipônimo. Por exemplo, a relação entre “mãe” e “pessoa” é uma hiponímia em que “mãe” é um hipônimo de “pessoa”. Hiperonímia é a relação inversa a de hiponímia, ou seja, denota que um conceito é uma generalização de outro conceito. O conceito mais genérico é um hiperônimo do mais específico. No exemplo dado “pessoa” é um hiperônimo de “mãe”.

     Os padrões léxicos sintáticos identificados por(HEARST, 1992) são utilizados na área da Extração de Informação e podem ser utilizados na área de Povoamento de Ontologias para a identificação de instâncias. Estes padrões são demonstrados na Tabela 2.2, na qual, observa-se na primeira coluna, seis padrões

linguísticos independentes de domínio adaptados de Hearst (1992) para o idioma inglês e na segunda coluna, a adaptação para o idioma português proposto por Baségio (2007) e utilizados no presente trabalho.

	Padrões Hearst	Tradução/Adaptação Baségio
1	NP such as (NP,)*(or|and) NP	SUB como (SUB,)*(ou|e) SUB SUB tal(is) como (SUB,)*(ou|e) SUB
2	such NP as (NP,)*(or|and) NP	tal(is) SUB como (SUB,)*(ou|e) SUB
3	NP , NP* , or other NP	SUB , SUB* , ou outro(s) SUB
4	NP , NP* , and other NP	SUB , SUB* , e outro(s) SUB
5	NP , including NP,*or|and NP	SUB , incluindo SUB,*ou|e SUB
6	NP , especially NP,*or—and NP	SUB , especialmente SUB,*ou|e SUB SUB , principalmente SUB,*ou|e SUB SUB , particularmente SUB,*ou|e SUB SUB , em especial SUB,*ou|e SUB SUB , em particular SUB,*ou|e SUB SUB , de maneira especial SUB,*ou|e SUB SUB , sobretudo SUB,*ou|e SUB

Tabela 2.2: Padrões linguísticos independentes de domínio

2.4 TRABALHOS RELACIONADOS

O processo de povoamento de ontologias de forma automática ou semiautomática depende diretamente do processo de aquisição do conhecimento. Várias propostas foram desenvolvidas, cada uma utilizando técnicas e métodos diferentes que permitem encontrar informações contidas em documentos e armazená-las em diversas formas, como por exemplo, em ontologias.

Neste cenário, diversos trabalhos foram propostos, como o PANKOW (CIMIANO; HANDSCHUH; STAAB, 2004), KnowItAll (ETZIONI et al., 2004), Never-Ending Language Learning (NELL) (CARLSON et al., 2010), UMOPOW (TOMAZ et al., 2012), OntoLP (XAVIER; LIMA, 2008), Poronto (ZAHRA; CARVALHO; MALUCELLI, 2013), A Semi-Automatic Method for Domain Ontology Extraction from Portuguese Language Wikipedia's Categories (XAVIER; LIMA, 2008), PSAPO (ALVES, 2013). Dentre esses, observa-se mais detalhes na Tabela 2.3, na qual:

Em Motta (2009) observa-se um processo para extrair informações estruturadas a partir de páginas da web. Seu objetivo é preencher uma ontologia de domínio de conhecimentos específicos que contenham afirmações declarativas em linguagem natural. Esse processo permite capturar informações semânticas contidas nos dicionários históricos textuais, ou seja, capturar entidades, relações e eventos e torná-los explícitos e disponíveis em uma ontologia.

Xavier e Lima (2008) apresentam um estudo sobre a extração de uma estrutura on-

tológica contendo relações de hiponímia (é uma) e localização a partir da Wikipédia em língua portuguesa. A abordagem visa capturar a estrutura de categorias de enciclopédias que contém um rico conteúdo semântico. As autoras fizeram um estudo de caso voltado para o domínio de Turismo e a proposta objetiva o mapeamento da estrutura taxonômica da ontologia e as relações de localização entre as instâncias, além da extração de instâncias.

O trabalho apresentado por Drumond e Girardi (2010) extrai estruturas taxonômicas a partir de textos, usando uma abordagem estatística, denominada Probabilistic Relational Hierarchy Extraction (PREHE) que faz a extração das estruturas através de reconhecimento de relações e outras técnicas de PLN. Ele também valida seu estudo no domínio de Turismo.

Baségio (2007) propõe uma abordagem para aquisição de estruturas ontológicas a partir de textos na língua portuguesa, mais especificamente, são extraídos conceitos e relações taxonômicas que servem como ponto de partida para o engenheiro de ontologia. Para a validação de sua proposta, o autor conduziu experimentos no domínio de Turismo.

Já Motta, Andreatta e Siqueira (2008), descrevem um ciclo do processo de extração de informações para preencher uma ontologia (conceitos e relações) de domínio utilizando textos disponíveis na internet com sites especificados previamente e que contenham afirmações declarativas em linguagem natural, agrupadas por entradas do dicionário, o qual descreve eventos biográficos e artísticos relacionados com as personalidades relevantes para o campo. O autor executou testes baseados no dicionário popular da música Brasileira.

No trabalho de Correia (2011), apresenta-se um processo para a aquisição automática de relações taxonômicas de uma ontologia baseada na aplicação de técnicas de Processamento de Linguagem Natural e Padrões heurísticos, que resulta na geração de uma taxonomia e utiliza como exemplo o domínio do Direito da família com corpus na língua inglesa.

Na mesma linha Zahra, Carvalho e Malucelli (2013), também apresentam uma ferramenta semi-automática para construção de ontologias a partir de textos em português na área da saúde(PORONTO). A ferramenta tem a principal finalidade de gerar uma Taxonomia sobre conceitos identificados no domínio da área de saúde da língua portuguesa.

Carlos et al. (2008), propõem a ferramenta OntoLP desenvolvida como um plug-in para o ambiente Protégé, que faz a análise de um corpus de domínio em língua portuguesa. Esse plug-in, na etapa inicial da construção de ontologias, sugere aos engenheiros de ontologias: a) os candidatos a conceitos; b) a extração e a organização hierárquica dos termos extraídos.

No seguimento de SEIBOS, Tomaz et al. (2012) propõem um método não supervisionado para o povoamento de ontologias a partir de textos escritos em inglês disponíveis na Web. Seu método extrai termos candidatos a instância utilizando

a Web como corpus e posteriormente combina diferentes medidas estatísticas e semânticas para classificar os termos extraídos.

Desta mesma forma, o método proposto baseia-se no trabalho de Tomaz et al. (2012) que foi desenvolvido originalmente para o idioma inglês. Nas investigações efetuadas durante o desenvolvimento dessa pesquisa não foram encontrados trabalhos relacionados que abordassem a aplicação de um processo de povoamento de ontologias similar ao proposto por Tomaz et al. (2012) para textos escritos em português. Essa lacuna motivou o desenvolvimento deste trabalho, no qual o método proposto se diferencia de outros métodos que lidam com textos em português por: (1) Utilizar a Web como grande corpus de trabalho para extração de termos candidatos a instâncias, (2) Avaliar diferentes medidas não supervisionadas para classificação de candidatos a instâncias, (3) Utilizar ontologias para guiar a extração e o armazenamento e (4) Utilizar classes que representem diferentes domínios ontológicos.

Guiado por ontologias	Corpus	Linguagem do corpus	Finalidade	Dominio estudado	Modo	Fases	Objetivo	Referência
não	Textos	Português	Povoamento de ontologias	História da música popular brasileira	Semiautomático	Preparação; Extração e preenchimento; Avaliação e Revisão.	Extração de entidades, relações e eventos.	(MOTTA, 2009)
não	Sites	Português	Povoamento de ontologias baseado em corpus e domínio especificado	Dicionário popular da música Brasileira	Semiautomático	1 - Analise corpus; 2 - Analise Ontologia; 3- Extração informação e povoamento; 4 - Avaliação; 5- Utilização.	Extração de conceitos e relações	(MOTTA; ANDREATTA; SIQUEIRA, 2008)
não	Site wikipedia	Português	Criação e povoamento de ontologias	BD da wikipedia	Semiautomático	1 - Extração taxonomia; 2- Identificação de classes, relações e instâncias; 3- Formatação linguística; 4 - Geração OWL Classes, relações e instâncias.	Extração de relações e instâncias	(XAVIER; LIMA, 2008)
não	Texto analisado morfossintaticamente, no padrão XCES, utilizando a ferramenta VISL	Português	Geração Taxonômica, extração de termos e relações taxonômicas.	Geral	Semiautomático	Não determinado	Extração de conceitos	(CARLOS et al., 2008)
não	Texto analisado morfologicamente, utilizando o TreeTagger	Português	Geração Taxonômica, extração de termos e relações taxonômicas	Área da saúde	Semiautomático	Não determinado	Extração de conceitos	(ZAHRA; CARVALHO; MALUCELLI, 2013)
não	Texto analisado morfologicamente(processo manual)	Português	Geração Taxonômica	Turismo	Semiautomático	1- Identificação de termos; 2 - Extração de relações taxonômicas; 3- Geração do código da estrutura ontológica.	Extração de termos e relações taxonômicas	(BASéGIO, 2007)
não	Textos	Inglês	Povoamento de ontologias	Direito de família	Semiautomático	Extração e Classificação de Instâncias; Representação de Instâncias	Extração de classes, relacionamentos e propriedades	(ALVES, 2013)

Guiado por ontologias	**Corpus**	**Linguagem do corpus**	**Finalidade**	**Dominio estudado**	**Modo**	**Fases**	**Objetivo**	**Referência**
não	Textos	Inglês	Povoamento de ontologias	Privacidade e responsabilização	Semiautomático	1 - Importação da ontologia e seleção de classes; 2- Expansão das classes; 3- Reconhecimento de entidades nomeadas e relações; 4- Geração de listas de entidades nomeadas e relações; 5- Geração da ontologia Owl com instâncias e propriedades;	Exração de Instâncias e propriedades	(BRUCKSCHEN, 2010)
não	Textos	Inglês	Geração Taxonômica	Direito de família	Semiautomático	1 - Marcação; 2 - Extração de Classes Candidatos; 3 - Identificação de Hiperônimos e Sinônimos; 4 - Identificação e Representação de Relações Taxonômicas.	Extração de termos e relações taxonômicas	(CORREIA, 2011)
sim	Textos escritos em linguagem natural (não estruturado) encontrados na Web	Inglês	Povoamento de ontologias	Independente	Semiautomático	Extração, Classificação, Povoamento, Extração de novos padrões e Classificação de novos padrões	Extração de instâncias de classes	(OLIVEIRA, 2013)
sim	Textos escritos em linguagem natural (não estruturado) encontrados na Web	Português	Povoamento de ontologias	Independente	Semiautomático	Extração, Classificação e Povoamento	Extração de instâncias de classes	(LIMA; OLIVEIRA; SALVADOR, 2015)

Tabela 2.3: Trabalhos relacionados

Capítulo 3

MÉTODO PROPOSTO

Baseado no modelo proposto por Oliveira (2013), Etzioni et al. (2004), neste capitulo são apresentadas as fases do Método para Povoamento de Ontologias. Na tentativa de automatizar o povoamento de ontologias escritas no idioma português, foi elaborado um método que utiliza técnicas de Processamento de Linguagem Natural (PLN) e Extração de Informação (EI) baseada em ontologias. Este, visa principalmente extrair instâncias de classes ontológicas a partir de textos escritos em linguagem natural disponíveis na Web e no idioma português (LIMA; OLIVEIRA; SALVADOR, 2015).

3.1 MÉTODO PARA POVOAMENTO DE ONTOLOGIAS

O principal objetivo do método proposto é executar o povoamento de uma ontologia independente do domínio e de forma mais automática possível, ou seja, com menor interferência humana. Para tanto, esse método inicia-se na etapa de definição da ontologia juntamente com as classes a serem instânciadas, única etapa com interferência humana, e prossegue até o povoamento da ontologia de entrada em formato OWL, contendo as instâncias classificadas, conforme é ilustrada na Figura 3.1.

Para a validação do método proposto, desenvolveu-se um sistema modular no qual cada módulo é responsável por um conjunto específico de tarefas: (1) Coleta dos documentos, (2) Extração dos documentos relevantes, (3) Classificação dos candidatos a instâncias e (4) Povoamento da ontologia com as instâncias resultantes do processamento. O sistema completo, sua arquitetura e funcionamento são detalhados a seguir.

- Etapa 1 - Coleta

  Nesta etapa efetua-se a coleta de documentos oriundos da Web, ou seja, recuperam-se documentos relevantes para formação do corpus de trabalho, de acordo com a ontologia independente do domínio e com os parâmetros de entrada.

- Etapa 2 – Extração

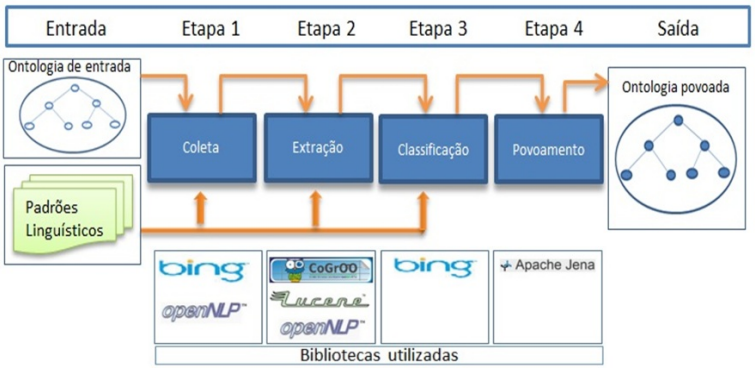

Figura 3.1 Visão geral do método para Povoamento de Ontologias

Executa-se o processamento do conjunto de documentos relevantes de forma a extrair apenas sentenças que possuem o padrão linguístico que originou o documento capturado. Além disso, aplicam-se filtros que visam validar os prováveis candidatos a instâncias as classes ontológicas.

- Etapa 3 – Classificação

  Nesta etapa executa-se um ranqueamento de cada candidato a instância atribuindo-lhe um valor absoluto, o qual proporciona um grau de confiança em relação a classe selecionada.

- Etapa 4 – Povoamento

  Nesta última etapa, efetiva-se o povoamento da ontologia de entrada considerando principalmente um limiar mínimo que define quais candidatos a instâncias serão utilizados para povoar a ontologia.

3.1.1 Etapa 1 - Coleta

Etapa inicial e a única que necessita de intervenção humana. Nela, definem-se e preparam-se os parâmetros de entrada, como: (1) recursos e bibliotecas de suporte; (2) definição das classes para povoamento; (3) definição do conjunto de padrões linguísticos. A principal finalidade desta etapa é coletar automaticamente documentos de acordo com os parâmetros de entrada para formar o corpus de trabalho. As consultas são formuladas a partir da junção das classes da ontologia e dos padrões de busca, após essas formulações, essas consultas são aplicadas no mecanismo de busca na Web que recuperam um conjunto de documentos relevantes e armazena. Na Figura 3.2 são ilustradas as fases dessa etapa.

Como pode-se observar na Figura 3.2, os parâmetros de entrada são:

(1) Definição dos recursos e bibliotecas de suporte

As definições dos recursos servem para proporcionar certa facilidade para toda etapa do método proposto, destacam-se a API do Bing [1], permite manipular e formular consul-

[1] http://www.bing.com/

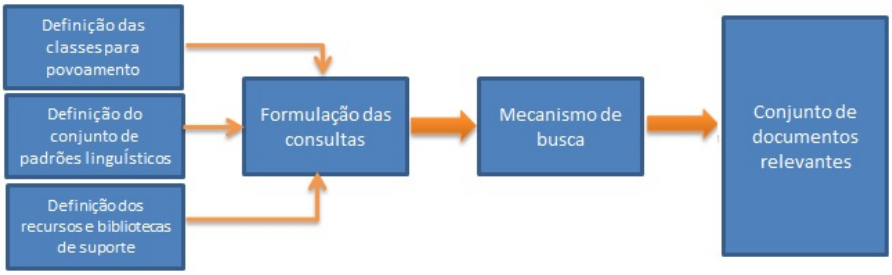

Figura 3.2 Fases da Etapa I - Coleta

tas e também executa as consultas geradas na máquina de busca; a biblioteca do Cogroo[2], entre outras tarefas, executa a tokenização, a divisão de sentenças e o reconhecimento de sintagmas nominais; a biblioteca Lucene[3] faz a radicalização dos tokens; a biblioteca OpenNLP[4] manipula os recursos em linguagem Natural; a biblioteca Apache Jena[5] fornece funções que permitem o preenchimento automático da ontologia.

(2) Definição das classes para povoamento

Nesta fase, seleciona-se a ontologia de entrada e definem-se as classes que serão povoadas.

(3) Conjunto de padrões linguísticos

Os padrões linguísticos servem para guiar o processo de extração de candidatos a instâncias e são utilizados juntamente com a classe da ontologia para a formulação das consultas. Neste método foi utilizado a proposta de Baségio (2007) que traduziu para a língua portuguesa os padrões linguísticos de Hearst (1992), como pode-se observar na Tabela 3.1, apresenta-se na 1ª coluna os padrões linguísticos adaptados para o idioma português e na 2ª coluna as consultas formuladas para a classe Cidade.

	Padrões Utilizados/Adaptados	Consultas formuladas
1	CLASSE(S) como CANDIDATOS CLASSE(S) tais como CANDIDATOS	cidades como cidades tais como
2	tais CLASSE(S) como CANDIDATOS	tais cidades como
3	CANDIDATOS ou outro(s) CLASSE(S)	ou outras cidades
4	CANDIDATOS e outro(S) CLASSE(S)	outras cidades
5	CLASSE(S) incluindo CANDIDATOS	cidades incluindo
6	CLASSE(S) especialmente CANDIDATOS CLASSE(S) principalmente CANDIDATOS CLASSE(S) particularmente CANDIDATOS CLASSE(S) em especial CANDIDATOS CLASSE(S) em particular CANDIDATOS CLASSE(S) sobretudo CANDIDATOS	cidades especialmente cidades principalmente cidades particularmente cidades em especial cidades em particular cidades sobretudo

[2]http://www.ccsl.ime.usp.br/cogroo
[3]http://www.lucene.apache.org
[4]http://www.opennlp.apache.org
[5]http://www.jena.apache.org

	Padrões Utilizados/Adaptados	Consultas formuladas
7	CANDIDATO é ART CLASSE	é uma cidade

Tabela 3.1: Padrões linguísticos com a respectiva consulta formulada

Os padrões linguísticos listados na Tabela 3.1, em geral, são precedidos ou seguidos de instâncias para a classe selecionada. O processo de formulação das consultas apresentadas na 2ª coluna da Tabela 3.1 é realizada da seguinte forma:

(1) o elemento *Classe* é substituído pelo rótulo da classe selecionada no singular, enquanto que o elemento *Classe(s)* é trocado pelo plural do rótulo da classe; (2) os elementos *Candidato* e *Candidatos* são removidos; e (3) o elemento ART é substituído pelos artigos indefinidos **um** ou **uma** de acordo com as regras gramaticais da língua portuguesa. Um detalhe importante é a presença de aspas nas consultas formuladas, elas indicam que a busca deve ser exata, ou seja, os documentos só devem ser recuperados se possuírem exatamente a consulta usada. Após a formulação das consultas, essas são aplicadas a um mecanismo de busca na Web para recuperação de n documentos relevantes para cada consulta apresentada na Tabela 3.1.

As consultas formuladas apresentam uma prévia da informação contida nos documentos recuperados pelos mecanismos de buscas na Web, Os snippets(fragmentos) são textos simples, que, em geral, possuem as palavras-chave que formam a consulta aplicada. Essas prévias (snippets), mesmo possuindo um tamanho reduzido são informativas o suficiente para extrair conhecimento relacionado com a consulta aplicada sem a necessidade de processar o documento inteiro coletado (MONLLAÓ, 2011).

Após as buscas esses Snippets são lidos e armazenados em formato de texto, identificados com a URL de cada Snippet.

Um exemplo de uma consulta utilizando um mecanismo de busca e um conjunto de documentos relevantes pode ser observado na Figura 3.3 e na Figura 3.4.

Na Figura 3.4 é apresentado um exemplo de um documento que foi coletado. Em seu conteúdo pode-se observar que foram adicionados os dados retornados pelo buscador, ou seja, um Snippet que contem a consulta formulada.

3.1.2 Etapa 2 - Extração

Após a Etapa I, cada um dos n documentos recuperados são processados, para isso, a ferramenta CoGrOO [6] é responsável pelas tarefas de tokenização, divisão de sentenças, etiquetagem das classes gramaticais e identificação de sintagmas nominais.

Como pode ser observado na Figura 3.5, esta fase funciona da seguinte maneira: recupera-se um documento(i) do conjunto de documentos relevantes; para cada documento recuperado, extraem-se sentenças que possuem o padrão linguístico que originou a consulta; para cada sentença encontrada, aplicam-se filtros que visam eliminar candidatos a instâncias que estejam duplicados e não válidos; cada candidato a instância encontrado contém, respectivamente, uma lista, sem repetição dos padrões linguísticos responsáveis

[6] http://ccsl.ime.usp.br/redmine/projects/cogroo

Figura 3.3 Demonstração de uma consulta e as prévias retornadas pelo Bing

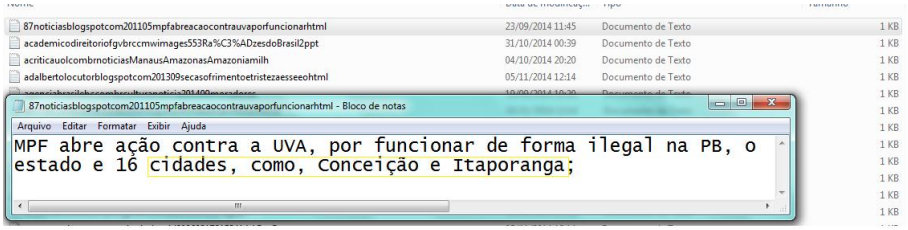

Figura 3.4 Exemplo de um Snippet armazenado em formato de texto

Figura 3.5 Visão geral da etapa de extração de dados - Extrator

por sua extração. Essa informação será usada posteriormente na fase de classificação dos candidatos a instâncias.

Um exemplo pode ser observado a seguir: na sentença *Cidades tais como Nova Iorque, Tóquio, Londres, Paris e Hong Kong são grandes pólos financeiros.* Os sintagmas nominais *Nova Iorque*, *Tóquio*, *Paris*, *Hong Kong* e *grandes pólos financeiros* são extraídos como candidatos a instâncias para a classe Cidade.

Usando sintagmas nominais é possível realizar a extração de palavras simples e compostas, aumentando, assim, a cobertura dos candidatos a instâncias extraídos. Outros trabalhos como (ETZIONI et al., 2004; MCDOWELL; CAFARELLA, 2008; TOMAZ et al., 2012) também utilizaram com sucesso sintagmas nominais como candidatos a

instâncias na tarefa de Povoamento de Ontologias.

Os filtros aplicados nesta fase visam eliminar candidatos a instâncias duplicados e não válidos, são eles: filtros de eliminação de duplicidade, eliminação de candidatos a instância sem valor semântico e filtro sintático.

a) Filtro de eliminação de duplicidade

Esse filtro objetiva remover candidatos a instâncias que já representam informações existentes na ontologia de entrada. Considerando-se que a classe *Cidade* é um possível candidato a instância *cidades*, *a cidade*, *uma cidade*, *na cidade*, entre outros. Esses candidatos por representarem a própria classe selecionada são removidos. Além disso, candidatos a instâncias que sejam instâncias já presentes na classe em povoamento ou que sejam instâncias de classes disjuntas também são eliminados. Para aumentar o alcance desse filtro utilizou-se o algoritmo de *stemming* que tem a finalidade de encontrar o radical das palavras.

b) Filtro de eliminação de candidatos a instância sem valor semântico

O objetivo geral deste filtro é remover candidatos a instâncias que não têm valor semântico, ou seja, são removidos candidatos que não possuem substantivos. Candidatos a instâncias, como *aqueles*, *a seguir*, *o quê* e *embora*, entre outros, são removidos nessa filtragem.

c) Filtragem Sintática

A lista de candidatos a instâncias produzida por essa etapa não deve possuir candidatos repetidos. Caso um candidato seja extraído mais de uma vez, esse é inserido em uma única vez na lista de candidatos a instâncias, sendo atualizada apenas a lista de padrões linguísticos distintos que o extraiu. Para melhorar o processo de identificação de redundâncias e evitar variações do singular e do plural, aplica-se o algoritmo de *stemming*. Além disso, candidatos que se diferem apenas pela presença de artigos, preposições e pronomes antes do primeiro substantivo são mapeados para apenas uma única forma. Por exemplo, os candidatos a instâncias *O cavalo*, *um cavalo*, *seu cavalo*, *cavalo* e *cavalos*, são identificados como um único candidato a instância. A decisão de qual representação deve permanecer é realizada mensurando o grau de coocorrência entre cada candidato a instância e a classe selecionada.

Para mensurar o grau de coocorrência é utilizada a medida PMI apresentada na Equação 3.1 e os padrões linguísticos listados na Tabela 3.1. Maiores detalhes sobre a medida PMI são apresentados na próxima seção. O candidato c com maior valor de coocorrência é inserido na lista de candidatos a instâncias e a lista de padrões linguísticos dos candidatos removidos são adicionadas na lista de padrões linguístico de c.

Após o processamento de todos os documentos coletados, ao final desta etapa, tem-se uma lista formada pelos candidatos a instâncias extraídos. Nesta, cada classe selecionada relaciona-se com seus respectivos padrões linguísticos. Na Tabela 3.2 observam-se exemplos de candidatos a instâncias extraídos para a classe *Cidade* e os respectivos padrões que os extraíram.

Candidato a Instância	Padrões Linguisticos que o Extraíram
Alexandria	[CANDIDATOS e -OUTRO- CLASSE | CLASSE sobretudo CANDIDATOS | CANDIDATO e -DET- CLASSE]

Candidato a Instância	Padrões Linguisticos que o Extraíram
Belo Horizonte	[CANDIDATOS ou -OUTRO- CLASSE | CLASSE tais como CANDIDATOS | CANDIDATOS e -OUTRO- CLASSE | CLASSE incluindo CANDIDATOS | CANDIDATO e -DET- CLASSE|CLASSE como CANDIDATOS]
Gramado	[CLASSE principalmente CANDIDATOS | CANDIDATOS e -OUTRO- CLASSE | CANDIDATO e -DET- CLASSE | CLASSE como CANDIDATOS]
São Paulo	[CLASSE principalmente CANDIDATOS | CLASSE especialmente CANDIDATOS | CLASSE em particular CANDIDATOS | CLASSE em especial CANDIDATOS | CANDIDATOS ou -OUTRO- CLASSE | CLASSE tais como CANDIDATOS | CANDIDATOS e -OUTRO- CLASSE | CLASSE incluindo CANDIDATOS | CLASSE sobretudo CANDIDATOS | CANDIDATO e -DET- CLASSE | CLASSE como CANDIDATOS | CLASSE particularmente CANDIDATOS]
Teresópolis	[CLASSE especialmente CANDIDATOS | CANDIDATO e -DET- CLASSE]
Viçosa	[CANDIDATO e -DET- CLASSE | CLASSE como CANDIDATOS]

Tabela 3.2: Lista de candidatos a instância com os respectivos padrões que o extraíram

3.1.3 Etapa 3 - Classificação

Após a extração dos candidatos a instâncias, é necessário avaliar a confiabilidade de cada candidato identificado. Esta etapa necessita de uma heurística de classificação que tem por objetivo, principalmente, avaliar cada candidato identificado pela etapa Extração e atribuir-lhe um grau de confiança. Após esse processamento, gera-se uma lista ordenada pela PMI dos candidatos a instâncias classificados.

Neste trabalho serão utilizadas as medidas estatísticas derivadas da Pointwise Mutual Information - (PMI) (TURNEY, 2001). A Pontuação de Informação Mútua caracteriza-se como uma medida estatística com objetivo de mensurar o grau de correlação entre dois termos.

O escore PMI calculado é o número de respostas (hits) para uma busca que combina o número de vezes que a instância aparece na sentença, dividido pelo número de hits para a instância sozinha, assim { PMI = Hits(Brasil é um país) / Hits(Brasil) }. Neste exemplo, supondo que o termo “Brasil” obteve o valor 4 de PMI, porque ele aparece 20 vezes na sentença e “Brasil é um país” apareceu 5 vezes sozinho.

Diante do fato que a medida PMI é tipicamente uma fração muito pequena, mesmo para instâncias positivas, seu cálculo não indica a probabilidade de uma instância ser membro da classe, apenas a probabilidade de ver o sintagma discriminador em webpages que contêm a instância (TURNEY, 2001; ETZIONI et al., 2004).

Tabela 3.3 Lista de candidatos a instância com suas respectivas consultas

(1) Padrões Linguísticos	(2) Consultas
Classe(s) tais como *Candidatos*	cidades tais como Salvador
tais *Classe(s)*) como *Candidatos*	tais cidades como Salvador
Candidatos ou outro(s) *Classe(s)*	Salvador ou outras cidades
Candidatos e outro(S) *Classe(s)*	Salvador e outras cidades
Classe(s) incluindo *Candidatos*	cidades incluindo Salvador
Classe(s) especialmente *Candidatos*	cidades especialmente Salvador
Candidato é ART *Classe*	Salvador é uma cidade

Cimiano, Handschuh e Staab (2004) afirmam que, em geral, as medidas estatísticas sofrem com o problema de esparsidade dos dados, assim, dependendo da fonte de informação utilizada, os dados disponíveis nem sempre apresentam um indicativo de sua relevância refletindo, então, uma baixa performance, principalmente, quando palavras relativamente raras são utilizadas.

Para resolver tal problema, Cimiano, Handschuh e Staab (2004), Etzioni et al. (2004), McDowell e Cafarella (2008), Tomaz et al. (2012) demonstraram que a utilização de medidas estatísticas explorando a grande quantidade de dados disponíveis na Web apresenta-se como uma solução viável.

Diante desse fato, neste trabalho a medida PMI objetivou mensurar o grau de coocorrência entre uma classe (c) e cada um dos seus candidatos a instâncias (ci) usando um conjunto de padrões linguísticos (P). Para isso, consultas foram formuladas utilizando c, ci e cada padrão linguístico $p \in \mathcal{P}$. Posteriormente, essas consultas foram aplicadas no mecanismo de busca na Web com o objetivo de obter a quantidade de ocorrências de cada consulta executada. São ilustradas na Tabela 3.3 as consultas formuladas para o cálculo da PMI utilizando a classe Cidade, o candidato a instância Salvador e os padrões linguísticos apresentados na Tabela 3.1.

a)Medidas de classificação de instâncias baseadas na PMI

Os trabalhos de Turney (2001), Cimiano, Handschuh e Staab (2004), Etzioni et al. (2004), McDowell e Cafarella (2008), Tomaz et al. (2012) são exemplos de estudos que utilizam variações de fórmulas aplicadas no cálculo da PMI como medida de classificação de instâncias de classes ontológicas. Em particular, McDowell e Cafarella (2008), Tomaz et al. (2012) apresentam estudos comparativos de diferentes variações para o cálculo da PMI aplicados no Povoamento de Ontologias na *Web* para textos escritos em inglês.

Nas equações seguintes, $\mathbf{c}$ representa uma classe da ontologia, $\mathbf{c}_i$ representa o candi-

dato a instância **i** de uma classe **c** enquanto **p** representa o padrão de Hearst (HEARST, 1992) usado na busca de termos na *Web*. Dessa mesma forma, **hits(c_i)** e **hits(c)** representam o total de ocorrências do candidato a instância $\mathbf{c}_i$ e da classe **c**, respectivamente, e **hits(c_i, c, p)** representa o número de ocorrências retornadas pelo mecanismo de busca na *Web* para a consulta formulada.

(I) **PMI Strength** é uma variação da PMI calculada pelo somatório de todas as coocorrências retornadas pela consulta *hits(c, ci, p)*. Para isso, consultas são formuladas para o par*(c, ci)* e cada padrão linguístico p contido em P listados na Tabela 3.1. Na Equação 3.1 é apresentado o cálculo desta variação da PMI.

$$PMI_{Strength} = \sum_{p \in P} hits(c, ci, p) \tag{3.1}$$

(II) Na medida **PMI Str-INorm-Thresh**, proposta porMcDowell e Cafarella (2008), o fator de normalização utilizado é o resultado do maior valor entre **hits(c_i)** do candidato à instância em classificação e o 25º percentil(Percen25) [7] da distribuição de **hits(c_i)** de todos os candidatos a instâncias para a classe **c** selecionada.

Na Equação 3.2 é demonstrado o cálculo desta variação da PMI.

$$PMI_{StrI-Norm}{}^{(c,c_i)} = \frac{\sum_{p \in P} hits(c, c_i, p)}{max(hits(c_i), Percen_{25})} \tag{3.2}$$

(III) A medida **PMI Str-ICNorm-Thresh** segue a mesma ideia da *PMI Str I-Norm*, mas com o fator de normalização considerando também o total de ocorrências da classe **c**. Na Equação 3.3 é especificado o cálculo desta variação da PMI.

$$PMI_{StrIC-Norm}{}^{(c,c_i)} = \frac{\sum_{p \in P} hits(c, c_i, p)}{max(hits(c_i), Percen_{25}) * hits(c)} \tag{3.3}$$

(IV) A medida proposta por Tomaz et al. (2012), **PMI Str I-Norm-Thresh-Hits0**, busca solucionar o problema de falsos candidatos a instâncias que possuem um alto valor de coocorrência para alguns padrões linguísticos e valor zero para outros. Os autores propuseram atribuir uma penalização para cada valor zero retornado pela consulta **hits(c, c_i, p)** usada para calcular o valor da PMI. Na equação 3.4 é apresentada a adaptação da variação da PMI Str I-Norm para inclusão do fator de penalização.

$$PMI_{StrI-Norm-Hits0}{}^{(c,c_i)} = \frac{\frac{\sum_{p \in P} hits(c,c_i,p)}{1+countHits0}}{max(hits(c_i), Percen_{25})} \tag{3.4}$$

b)Medidas PMI propostas

Com a finalidade de aumentar a assertividade na classificação dos candidatos a instâncias, as medidas propostas: *PMI Str I-Norm-Z, PMI Str IC-Norm-Z* e *PMI Str I-NormHits0-Z*, especificadas nas Equações 3.4, 3.7 e 3.8, utilizam como normalizador o cálculo de desvio padrão (LIMA; SALVADOR, 2015).

[7] 25º percentil(Percen25), é resultado da subtração entre o maior e o menor valor de hits(c_i) do conjunto de candidatos a instâncias, dividido por quatro (25%);

Como um baixo desvio padrão indica que os dados tendem a estar próximos da média e um alto indica que os dados estão bem dispersos, as medidas PMI propostas exploram essas questões e buscam calibrar a variação da média dos candidatos a instâncias, possibilitando ao normalizador da medida PMI variar dentro de um intervalo de confiança não disperso das médias de todos os candidatos a instâncias selecionados.

Na Equação 3.5 apresenta-se a fórmula do desvio padrão parametrizada com valor do hits do candidato a instância c_i. Nesta equação, obtem-se o valor do hits(c_i) de cada candidato a instância subtraindo-se da média de todos os hits(c_i) do conjunto de candidatos a instâncias.

$$\sigma = \sqrt{\sum_{i=1}^{n} \frac{((hits(c_i)) - (avg(hits(c_1), hits(c_2), ...hits(c_n))))^2}{n-1}} \tag{3.5}$$

As medidas propostas nas Equações 3.6, 3.7 e 3.8 são derivadas das Equações 3.2, 3.3 e 3.4, da seguinte forma: o numerador dessas equações continua representando o somatório de todas as ocorrências utilizando-se cada padrão linguístico **p** pertencente ao conjunto de padrões linguísticos **P**, contudo, os denominadores das equações propostas são alterados da seguinte forma.

I) Calcula-se o hits(c_i) para cada candidato a instância pertencente ao conjunto de candidatos a instância extraídos para a classe em povoamento;

II) Ordena-se esse conjunto de candidatos a instâncias com base nos valores dos hits(c_i);

III) Substitui-se o cálculo do 25^{o} percentil pelo valor do desvio padrão (σ) oriundo da Equação 3.5.

IV) Por último, com intuito de aproximar-se ao máximo do valor do hits do candidato a instância, em vez de escolher como fator de normalização o maior valor entre o hits(c_i) do candidato a instância em avaliação e o Percen25, escolhe-se o menor valor entre o hits(c_i) e o desvio padrão (σ). No caso de candidatos a instâncias que obtiverem o desvio padrão igual a zero, considera-se no fator de normalização apenas o total de ocorrências do candidato à instância, **hits(c_i)**.

$$PMI_{StrI-Norm-}Z^{(c,c_i)} = \frac{\sum_{p \epsilon P} hits(c, c_i, p)}{min(hits(c_i), \sigma)} \tag{3.6}$$

$$PMI_{StrIC-Norm-}Z^{(c,c_i)} = \frac{\sum_{p \epsilon P} hits(c, c_i, p)}{min(hits(c_i), \sigma) * hits(c)} \tag{3.7}$$

$$PMI_{StrI-Norm-Hits0-}Z^{(c,c_i)} = \frac{\frac{\sum_{p \in P} hits(c,c_i,p)}{1+countHits0}}{min(hits(c_i), \sigma)} \tag{3.8}$$

c)Medida heurística baseada em Números de Padrões Extraíram (NPE)

A heurística de Número de Padrões que Extraíram (NPE), proposta por Tomaz et al. (2012), é baseada na hipótese de quanto mais padrões linguísticos distintos forem responsáveis pela extração de um candidato a instância, mais forte é a evidência de que esse candidato seja realmente uma instância válida para a classe em povoamento. Baseada

nessa hipótese, a heurística de NPE mensura quantos padrões linguísticos distintos extraíram o candidato à instância avaliado. O intervalo de valores que essa heurística pode assumir varia de 1 ao total de padrões linguísticos utilizados pelo método para a classe selecionada. Por exemplo, na Tabela 3.2, o candidato a instância Alexandria possui NPE = 3, já que foi extraído por 3 padrões linguísticos, enquanto que o candidato a instância Viçosa possui NPE = 2.

3.1.4 Etapa 4 - Povoamento

Nesta última etapa, o importante é decidir quais candidatos a instâncias serão promovidos a instâncias da classe selecionada. Para isso, é necessário escolher um limiar, principalmente, levando em consideração que essa escolha impacta diretamente na taxa de acerto e cobertura do método.

Em geral, a escolha do limiar é realizada empiricamente, sendo promovida a instância apenas candidatos que possuem um valor maior do que o limiar escolhido. Conforme apontado por Tomaz et al. (2012), a medida PMI possui valores variando em ordem de grandeza diferentes dependendo da classe selecionada, do candidato a instância em avaliação e dos padrões linguísticos utilizados. Diante disso, estimar um valor para o limiar tornou-se inviável. Ao invés disso, assim como Tomaz et al. (2012), optou-se por promover a cada iteração os n melhores candidatos a instâncias (Top n) ordenados com base nos valores de confiança de cada uma das medidas de classificação descritas na seção anterior.

A Figura 3.6 demonstra a ontologia de entrada, na ferramenta Protègè, antes do processo de povoamento, e na Figura 3.7 apresenta-se a classe País, da ontologia preenchida com as instâncias selecionadas, considerando os 10 primeiros candidatos a instância(Top10) da classe em questão.

3.2 VALIDAÇÃO DO MÉTODO PARA POVOAMENTO DE ONTOLOGIAS

Com o intuito de verificar a eficácia do Método para Povoamento de Ontologias, um experimento modular com as 4 etapas do método foi executado. Para a sua validação foi desenvolvido um sistema computacional que objetivou efetuar a coleta, extração, classificação e povoamento da ontologia como descrito neste capítulo e nas próximas seções.

3.2.1 Etapa I - Coleta

A coleta foi executada automaticamente e objetivou formar o corpus de trabalho, para isso, 15 classes foram selecionadas em uma ontologia de topo customizada: *Cidade*, *País*, *Pássaro*, *Peixe*, *Sintoma*, *Esporte*, *Inseto*, *Mamífero*, *Doença*, *Universidade*, *Ator*, *Atriz*, *Filme*, *Rio* e *Hotel*.

Nesta etapa foram coletados 1000 documentos/dia entre o período de 19/09/2014 a 20/11/2014 e foram recuperados 62.409 documentos relevantes, como pode-se observar na Tabela 3.5.

Observa-se ainda que essas classes foram escolhidas por representarem conceitos de diferentes domínios, e alguns desses terem sidos utilizados nos trabalhos de McDowell e

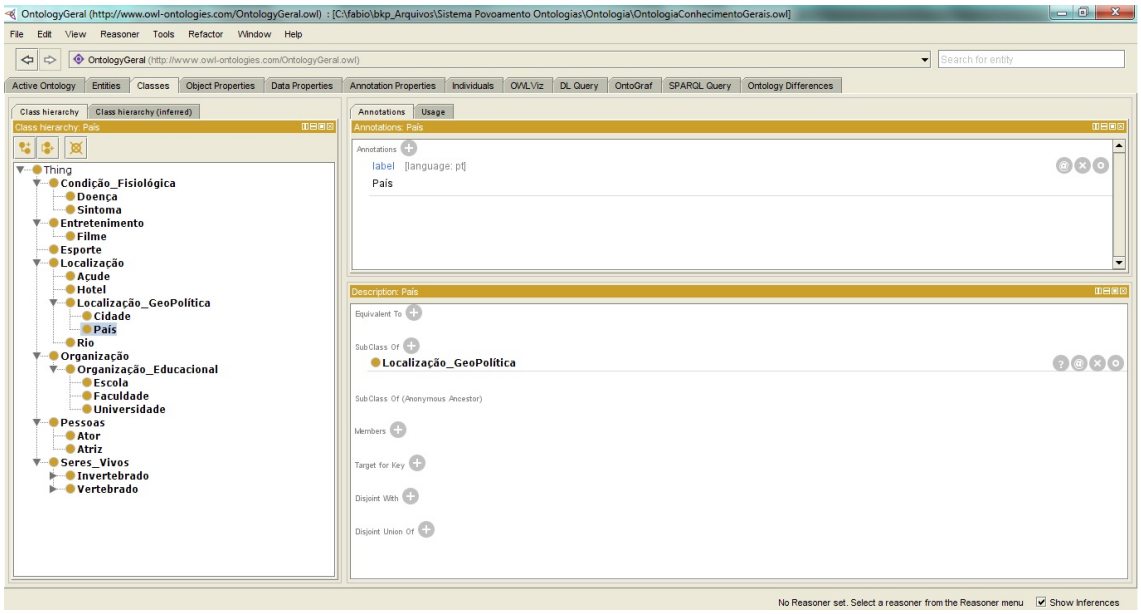

Figura 3.6 Demonstração da ontologia no Protégé antes do povoamento

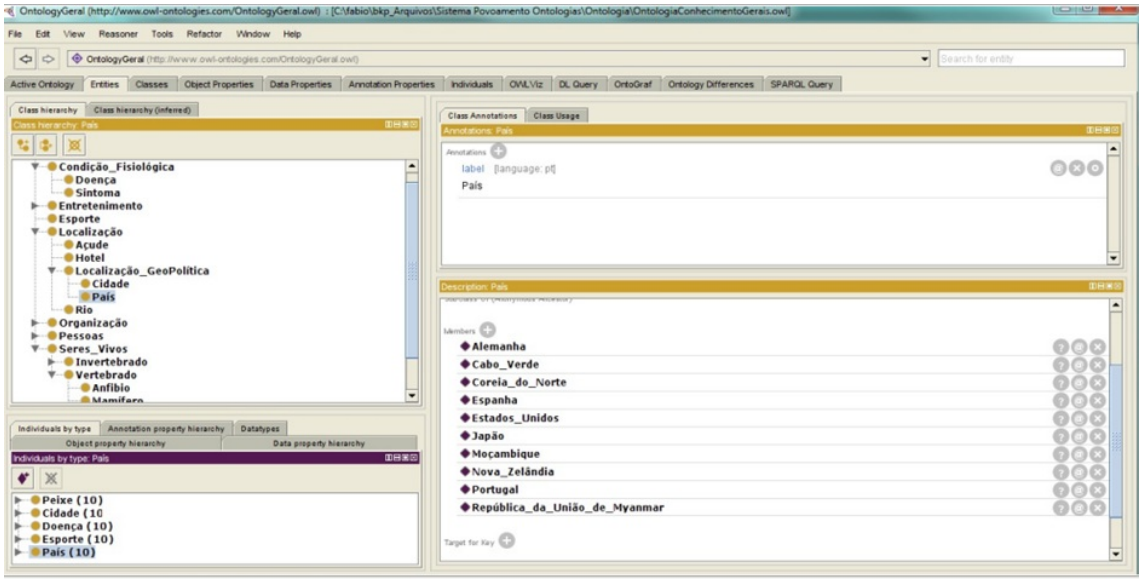

Figura 3.7 Demonstração da ontologia no Protégé após a fase de povoamento

Cafarella (2008), Oliveira (2013), Etzioni et al. (2004).

Classe	CLASSE(S) como CANDIDATOS	CLASSE(S) tais como CANDIDATOS	tais CLASSE(S) como CANDIDATOS	CANDIDATOS ou outro(s) CLASSE(S)	CANDIDATOS e outro(S) CLASSE(S)	CLASSE(S) incluindo CANDIDATOS	CLASSE(S) especialmente CANDIDATOS
Cidade	899	275	68	333	748	401	349
País	1384	339	46	277	1001	3	1
Pássaro	358	149	12	200	270	179	246
Peixe	562	217	6	225	494	247	232

Classe	CLASSE(S) como CANDIDATOS	CLASSE(S) tais como CANDIDATOS	tais CLASSE(S) como CANDIDATOS	CANDIDATOS ou outro(s) CLASSE(S)	CANDIDATOS e outro(S) CLASSE(S)	CLASSE(S) incluindo CANDIDATOS	CLASSE(S) especialmente CANDIDATOS
Sintoma	815	331	145	337	321	270	284
Esporte	669	320	41	312	559	285	378
Inseto	499	249	17	225	492	198	245
Mamífero	420	162	1	185	320	339	340
Doença	1233	346	29	424	862	528	461
Universidade	816	187	6	168	276	220	272
Ator	469	212	34	181	408	234	202
Atriz	504	59	0	12	233	139	114
Filme	264	386	139	191	295	485	368
Rio	872	184	3	115	216	187	316
Hotel	190	173	0	88	147	194	181
Réptil	296	131	0	120	188	250	131
Anfíbio	315	114	5	16	175	134	267
Média	621	226	32	201	412	253	258
Total							

Tabela 3.4: Quantidade de snippets coletados por classe - parte I

Classe	CLASSE(S) principalmente CANDIDATOS	CLASSE(S) particularmente CANDIDATOS	CLASSE(S) em especial CANDIDATOS	CLASSE(S) em particular CANDIDATOS	CLASSE(S) sobretudo CANDIDATOS	CANDIDATO é ART CLASSE	Total snippets
Cidade	427	163	209	190	303	1221	5586
País	0	0	210	143	0	1189	4593
Pássaro	237	90	171	83	149	676	2820
Peixe	335	144	182	170	412	787	4013
Sintoma	327	195	165	133	155	585	4063
Esporte	563	133	369	155	224	1118	5126
Inseto	317	129	192	134	156	645	3498
Mamífero	265	170	97	135	140	797	3371
Doença	523	0	271	153	503	1609	6942
Universidade	320	197	167	160	198	338	3325
Ator	230	107	188	128	154	514	3061
Atriz	151	25	104	8	53	713	2115
Filme	543	132	213	150	195	1107	4468
Rio	351	164	181	91	168	594	3442
Hotel	247	60	27	48	134	663	2152
Réptil	219	41	64	13	70	446	1969
Anfíbio	164	134	47	44	51	399	1865
Média	307	111	168	114	180	788	
Total							62409

Tabela 3.5: Quantidade de snippets coletados por classe - parte II

Nas Tabelas 3.4 e 3.5 pode-se observar que alguns padrões não obtiveram bons resultados, sendo o caso do padrão [tais CLASSE(S) como CANDIDATOS] que resultou uma média de 32 snippets por classe. Os padrões [CLASSE(S) particularmente CANDIDATOS; CLASSE(S) em especial CANDIDATOS e CLASSE(S) em particular CANDIDATOS] também não tiveram bons resultados, retornando menos de 200 snippets em média por classe. Em contrapartida os padrões [CLASSE(S) como CANDIDATOS e CANDIDATO é ART CLASSE] obtiveram respectivamente uma média de 621 e 788 snippets por classe, as maiores médias da coleta.

Observa-se ainda na Tabela 3.5 que as classes que obtiveram mais de 5000 snippets coletados foram as classes: Doença, Cidade e Esporte, devido à relevância desses assuntos. Também existem classes com pouca abrangência, com menos de 2000 snippets, sendo o

caso das classes Réptil e Anfíbio, as quais foram excluídas da fase de povoamento, por não atingirem uma quantidade expressiva de instâncias válidas.

3.2.2 Etapa II - Extração

Inicialmente a execução dessa etapa gerou 15.739 instâncias candidatas, das quais, cerca de 73% foram descartadas por não contemplarem o padrão de busca e conterem informações repetidas. Na Tabela 3.6 observa-se, na segunda coluna, a quantidade de documentos recuperados e na terceira a porcentagem de candidatos removidos pela subetapa(extração de sentenças). Ainda é possível observar na quinta coluna a porcentagem de candidatos eliminados pelo processo de filtragem que foi capaz de remover em média 40% dos candidatos inválidos. Na última coluna, observa-se o total de instâncias candidatas selecionadas, as quais foram processadas pela Etapa de Classificação, discutida na próxima seção.

Classe	Total documentos recuperados	Percentagem de eliminação	Total de candidatos gerados	Percentagem de eliminação (filtragem)	Total de instâncias selecionadas
Cidade	5586	62%	2136	19%	1736
País	4593	77%	1063	34%	702
Pássaro	2820	84%	440	50%	222
Peixe	4013	82%	709	41%	418
Sintoma	4063	91%	363	42%	212
Esporte	5126	84%	821	47%	437
Inseto	3498	87%	460	53%	218
Mamífero	3371	88%	394	47%	210
Doença	6942	87%	871	38%	542
Universidade	3325	64%	1207	32%	825
Ator	3061	30%	2157	27%	1584
Atriz	2115	36%	1353	27%	986
Filme	4468	67%	1490	16%	1247
Rio	3442	77%	807	36%	519
Hotel	2152	62%	812	33%	544
Réptil	1969	86%	274	64%	100 **
Anfíbio	1865	80%	382	72%	107 **
Média		73%		40%	
Total	62409		15739		10402 [8]

Tabela 3.6: Percentagem documentos eliminados na Etapa I e II

[8]Itens sinalizados com ** não contabilizado na soma por apresentarem valores abaixo do Limiar 200

3.2.3 Etapa III - Classificação

Para esse método foi elencado como medida de classificação a medida da PMI combinada(C9), discutida no experimento III. Nela considerou-se o limiar Top 10 que obteve uma média de 70% de precisão e conseguiu classificar uma maior quantidade de instâncias válidas para a maioria das classes estudas.

1. Método de Classificação

   Para cada classe selecionada na ontologia de entrada aplica-se a medida combinada(C9) que é responsável por atribuir um grau de confiança a cada candidato a instância que foi selecionado na etapa II. Na Tabela 3.7, terceira coluna, observa-se a taxa de precisão da medida combinada(C9), a qual obteve uma média de 70% de precisão para 11 das 15 classes estudadas. Na última coluna, observa-se o total de instâncias selecionadas para povoar a ontologia de entrada.

Classe	Total de candidatos selecionados	Precisão da medida combinada(C9)	Total de instâncias selecionadas para povoamento
Cidade	1736	70%	1215
País	702	90%	632
Pássaro	222	90%	200
Peixe	418	90%	377
Sintoma	212	30%	64
Esporte	437	70%	306
Inseto	218	60%	131
Mamífero	210	80%	168
Doença	542	90%	488
Universidade	825	60%	495
Ator	1584	70%	1109
Atriz	986	80%	789
Filme	1247	50%	623
Rio	519	90%	467
Hotel	544	30%	163
Média Total	 10402	70%	 7225

Tabela 3.7: Total de instâncias selecionadas para povoamento

3.2.4 Etapa IV - Povoamento

Com objetivo de transferir para a ontologia de entrada os candidatos selecionados na fase de classificação, o processo se deu de forma guiada pela ontologia de entrada e pela lista de candidatos oriundos da etapa III. Na Tabela 3.7, na última coluna, observa-se o total de candidatos validados e utilizados para povoar a ontologia de entrada.

3.2.5 Resultados e Discussões

O método proposto comportou-se de maneira similar diante dos diferentes domínios e não necessitou de alterações durante sua execução. Seus resultados são promissores e em todos os experimentos realizados encontrou-se uma grande quantidade de instâncias corretas para a maioria das classes analisadas, mesmo variando de acordo com domínio de cada classe selecionada e considerando a complexidade do domínio analisado. Entretanto, a fase de classificação do método extraiu uma grande quantidade de falsos candidatos a instâncias, justamente, pelo fato dessas informações serem escritas por pessoas que, em geral, não são especialistas do domínio abordado e em virtude do alto valor que muitos candidatos obtiveram nas medidas da PMI.

A medida combinada(C9), escolhida para a fase de classificação do método para povoamento, obteve uma média de precisão em torno de 70% para a maioria das classes analisadas. Entretanto, as classes Hotel e Sintoma, tiveram uma baixa precisão, cerca de 30%, como apresentada na Tabela 3.7, terceira coluna. Tal fato justifica-se pela ausência de estruturação dos documentos coletados e pela falta de conhecimento prévio dos validadores humanos que, na maioria das vezes, tiveram que recorrer à outras fontes de dados e a própria web para buscar informações da instância em questão.

Em todas as variações da PMI analisadas foram identificadas dificuldades que influenciaram negativamente os resultados, mesmo assim, a medida de combinada(C9) apresentou uma melhor precisão do que as outras variações da PMI estudadas. Entre as principais dificuldades encontradas destacam-se:

- Falsos candidatos a instâncias

  São candidatos que possuem um alto valor de coocorrência apenas para poucos padrões linguísticos e valor zero nos demais. Por exemplo, na classe Cidade, o candidato a instância [Belo Horizonte] presente no Top50 obteve um alto valor de coocorrência em seis padrões linguísticos e valor zero nos outros sete. Muitos falsos candidatos a instâncias obtiveram alto valor de PMI por causa desse problema.

- A presença de candidatos a instâncias incompletos

  Por exemplo, na classe Cidade foi identificada a extração do candidato a instância [Conquista], na qual possivelmente o correto seria [Vitória da Conquista]. Possivelmente o valor da coocorrência hits(c, ci, p) do real candidato [Vitoria da Conquista] está contido no valor do candidato [Conquista], que também está contido em outros padrões e que resultou na atribuição de um alto valor na medida da PMI para o candidato a instância [Conquista]. Essa mesma situação foi observada em outros candidatos, como por exemplo [Cruz | Alta], no qual, possivelmente, o correto seria

[Cruz Alta]. Esse tipo de problema foi gerado devido aos erros na ferramenta de identificação de sintagmas nominais ou por causa de fragmentos de textos incompletos.

- Estrutura é identificada como um sintagma nominal único

  Quando uma estrutura é identificada como um sintagma nominal único gera, na maioria dos casos, um falso candidato a instância formada por duas instâncias corretas, por exemplo, [Campina Grande e Patos], essas instâncias deveriam estar separadas[Campina Grande] e [Patos]. Tal problema provocou um impacto negativo devido ao alto valor obtido na medida da PMI.

- Grafias diferentes e mesmo significado

  Todas as classes analisadas apresentaram esse problema, por exemplo, na classe Cidade foram encontradas as instâncias [Camcun] e [Camcum], ambas instâncias corretas.

- Grafias erradas e com validades na Web

  Algumas instâncias foram duplicadas por terem uma grafia correta e uma incorreta, porém, ambas conhecidas pela maioria dos usuários e com alto índice de referência, o que levaram candidatos falsos a serem considerados instâncias reais, como Ornitorrinco e onitorrinco.

- Siglas

  A falta de uma ferramenta que identificasse as siglas e apelidos gerou um impacto bastante negativo na validação das classes, principalmente, nas classes Universidade e Cidade. Para a classe Universidade tem-se muitas instâncias duplicadas e sem relações entre elas, é o caso da sigla UFBA, ela refere-se a Universidade Federal da Bahia ou Universidade Federal Baiana. Além disso, para a classe cidade tem-se as siglas SSA, VCA, GBI, BH e Foz, Floripa, Sampa como apelidos.

- Formação radical

  Instâncias que contêm a mesma formação radical acabaram gerando falsos candidatos, como Maragogi, Maragogipe e maragogipinho, candidatos com o mesmo radical e com pós-fixo diferentes.

- Termos pertencentes ao domínio

  A presença de termos pertencentes ao domínio, mas, que não são instâncias para a classe selecionada, como por exemplo, para a classe Doença os candidatos a instâncias [Coração] e [HIV - Vírus da Imunodeficiência Humana] representam termos relevantes para o domínio de doenças, porém não são instâncias dessa classe. Contudo, por serem termos importantes para o domínio, eles obtiveram um alto valor nas medidas da PMI. Esse problema levou o método na etapa de classificação, a promover falsos candidatos a instâncias, principalmente, quando se utilizou os limiares mais abrangentes (Top100 e Top200).

3.3 EXPERIMENTOS

Objetivando analisar qual melhor medida para classificar instâncias de classes ontológicas para o domínio de textos extraídos da web e escrito em português. Neste capítulo apresentam-se os experimentos realizados que visam, principalmente, avaliar as medidas PMI apresentadas.

3.3.1 Visão geral

Cada experimento objetivou analisar e comparar a precisão das medidas de classificação estudadas e definir qual a medida poderia ser utilizada na fase de classificação do método para Povoamento de Ontologias, sendo eles:

No primeiro experimento (seção 3.3.2) avaliou-se comparando-as entre si, as variações das PMI Strenght, PMI Str I-Norm, PMI Str IC-Norm e a heurística NPE. Desta forma esperava-se identificar qual a variação apresenta os melhores resultados em termos de precisão.

O segundo experimento (seção 3.3.3) visou avaliar a precisão das medidas propostas: PMI Str I-Norm-Z; PMI Str IC-Norm-Z; PMI Str I-Norm-Hits0-Z, em detrimento das medidas PMI Str I-Norm, PMI Str IC-Norm, PMI Str I-Norm-Hits0.

No terceiro experimento (seção 3.3.4) executou-se uma combinação linear entre a medida NPE, melhor medida do experimento I e a PMI Str I-Norm-Z, melhor medida do experimento II. Com isso, esperava-se identificar se a combinação entre as medidas geraria uma métrica mais eficaz do que as medidas analisadas.

O quarto experimento (secão 3.3.5) efetuou-se uma comparação entre a melhor medida do experimento I, II e III. Desta forma buscou-se identificar a medida mais adequada para utilizar no Método para Povoamento de Ontologias.

1. Mecanismo de avaliação

   No capítulo dois foram discutidas algumas medidas utilizadas para avaliar sistemas de extração de informação. Neste trabalho foi utilizada a medida que avalia a corretude de um SEIBO, a Precisão, a qual pode ser definida pela razão entre a quantidade de instâncias corretas extraídas pelo total de instâncias recuperadas.

   Para o cálculo dessa medida faz-se necessário analisar cada instância extraída e verificar se a mesma é uma real instância para a classe ao qual foi atribuída.

   A avaliação da precisão dos experimentos foi realiza variando diferentes limiares. O limiar define o valor dos N candidatos a instâncias (Top N) ordenados com base na medida de classificação utilizada. Nestes experimentos foram utilizados quatro limiares: Top 10, Top 50, Top 100 e Top 200. Os limiares Top 10 e Top 50 são mais restritivos, ou seja, poucas instâncias são extraídas. Já os limiares Top 100 e Top 200 visam promover um número maior de instâncias por iteração.

   Para o cálculo da Precisão foi necessário analisar cada termo candidato extraído e verificar se ele era realmente uma instância para a classe ao qual foi atribuída. Para isso, 27 humanos foram responsáveis pelo processo de validação de cada um dos *Top N* melhores candidatos a instâncias classificados.

A Equação 3.9 é usada para calcular a precisão nos experimentos descritos neste capítulo, na qual, N é a quantidade de instâncias candidatas promovidas à instância com base na medida escolhida.

$$Precisao(TopN) = \frac{total\_de\_intancias\_corretas}{N} \tag{3.9}$$

2. Critérios adotados para validação de instâncias

   Os critérios de validações de instâncias definem as regras para aceite ou recusa de cada candidato a instância selecionado e é utilizado pelo especialista de domínio na fase de classificação.

   Para a fase de validação foram selecionadas apenas classes que obtiveram mais de 200 instâncias classificadas, o que resultou na exclusão das classes Anfíbio e Réptil e na consideração apenas das classes listadas na Tabela 3.8.

Número	Classe		Classe		Classe		Classe
1	Cidade	2	País	3	Pássaro	4	Peixe
5	Sintoma	6	Esporte	7	Inseto	8	Mamífero
9	Doença	10	Universidade	11	Ator	12	Atriz
13	Filme	14	Rio	15	Hotel		

Tabela 3.8: Classes utilizadas na etapa de Classificação

Os candidatos a instâncias classificados nesta fase foram analisados e validados por especialistas da área observando alguns critérios que variaram de acordo com o domínio da classe selecionada, sendo eles:

- Filme

  Nesta classe foram consideradas como instâncias corretas somente as instâncias que continham o nome completo do filme, por exemplo: considerou-se Batman Begin ao invés de Batman;

- Ator e Atriz

  As classes Ator e Atriz seguiram a mesma abordagem da classe Filme, na qual, os candidatos a instância Ator e Atriz eram computados somente quando o nome de uma provável instância fosse mais completo possível, por exemplo, computava-se Ana Paula Arósio ao invés de Ana Paula;

- Esporte

  Devido essa classe possuir muitas especificidades e diferentes formas corretas de grafias, foram consideradas instâncias corretas, ambas as formas, por exemplo: Futsal, Futsal Feminino; Futebol, Futebol Americano; Handebol, Handball;

- Cidade e País

  As classes Cidade e País mesmo possuindo instâncias de fácil identificação, apresentaram dificuldades durante a avaliação. Na classe Cidade, por exemplo,

houve instâncias que foram classificadas de forma separada, sendo o caso de [Cruz e Alta]. Na classe País houve alguns casos de instâncias que obtiveram alto índice de relevância e foram selecionadas, sendo o caso de instâncias como Braziu e Brasiu;

- Rio

  A classe Rio teve uma complexidade pequena em sua identificação, muitas instâncias incorretas foram apresentadas para validação devido ao uso de metáforas contidas nos documentos analisados, como por exemplo, foram encontrados candidatos a instâncias [tudo; vida] que originavam de sintagmas nominais como "nessa vida tudo é um rio", " A vida é um rio de palavras". Nessa classe também foram consideradas como corretas as instâncias que continham o nome da classe em sua formação, por exemplo, [Rio Tietê | Tietê]; outro caso foi a atribuição de apelidos ao mesmo candidato a instância,como por exemplo, instâncias como Velho chico e São Francisco foram aceitas. Também pode-se identificar instâncias com diferentes grafias, é o caso da instância [Ganges | Ganjes], ambas também consideradas como corretas;

- Peixe

  A classe Peixe apresentou um pouco de complexidade, devido a grande diversidade do domínio, a qual necessitava de pesquisas constantes a outras fontes de informações, como é o caso de candidatos a instâncias [Peixe espada | Peixe espadarte] serem peixes diferentes, entretanto, com a mesma construção do radical. Nessa classe, também não considerou-se como corretos os nomes próprios, por exemplo, [Nemo, Aneci] por outro lado, foi considerado correto, as especialidades da classe como: Peixe [Pacu | Pacu Borracha | Pacu Vermelho];

- Inseto, Mamífero e Pássaro

  As classes Inseto, Mamífero e Pássaro, por serem classes complexas, foram consideradas como instâncias corretas, tanto o gênero como também a espécie, por exemplo: a instância aranha é um gênero e aranha Marrom uma espécie para a classe Inseto; para a classe Pássaro tem-se Agapórnis como gênero e tem-se periquito como espécie. Também não foram considerados nomes próprios como, por exemplo [Chita | Monga] para a classe Mamífero;

- Doença e Sintoma

  As classes Doença e Sintoma também foram de difícil identificação em virtude de suas características serem bastante específicas. Para essas classes conseguirem determinar termos, como: Câncer para a classe Doença e Raiva para a classe Sintoma, necessitava-se, além de um bom conhecimento do domínio, de pesquisas em outras fontes textuais;

- Universidade

  Esta classe contém a peculiaridade da instância possuir tanto o nome completo quanto uma sigla. Neste caso foi contabilizado como instâncias corretas ambas as formas, por exemplo [UFBA | Universidade Federal da Bahia]. Outra

peculiaridade, mas apenas contabilizado uma vez, foi o caso de diversos nomes da mesma classe, ou seja, [Instituto Tecnológico de Massachusetts | Instituto de Tecnologia de Massachusetts] ambos os nomes corretos. Essa mesma definição foi aplicada a nomes de universidade que poderiam ter outra interpretação, por exemplo: São Paulo pode ser uma cidade ou estado e somente foi considerado a instância que continha o nome completo, ou seja, Universidade de São Paulo;

- Hotel
  Para validação dessa classe todas as instâncias foram validadas fazendo uma pesquisa na WEB a fim de confirmar a instância em questão. Desta forma considerou-se como candidatos a instâncias corretos: Plaza Atené e Plaza Camboriú, ambos hotéis Plaza mas com diferente nomes fantasias.

3.3.2 Experimento I

Uma análise foi conduzida para avaliar as variações PMI Strength, PMI Str-INorm e a PMI Str-ICNorm isoladamente e comparadas com a heurística NPE proposta por Tomaz et al. (2012). Desta forma, espera-se identificar qual dessas medidas obtêm melhor acurácia na classificação dos candidatos a instâncias extraídos.

O experimento iniciou com a execução da etapa 2 (Extração) e foi responsável por extrair um conjunto de candidatos a instâncias para cada uma das classes apresentadas na Tabela 3.8 selecionadas para este experimento.

Após a extração dos candidatos a instâncias, foram realizadas as execuções da etapa denominada Classificação (seção 3.1.3). Cada execução realizou o processo de classificação utilizando uma variação da medida de PMI adotada como medida de classificação de forma isolada e da medida heurística NPE.

Na próxima subseção são apresentados os resultados e as discussões deste experimento.

3.3.2.1 Resultados e Discussões do Experimento I

Os gráficos apresentados nas Figuras 3.8, 3.9, 3.10 e 3.11 demonstram os resultados obtidos para as 15 classes selecionadas na ontologia de entrada. Em todos os gráficos são realizadas as comparações entre as medidas de PMI Strength, PMI Str-INorm-Thresh, PMI Str-ICNorm-Thresh e a heurística NPE.

Analisando os resultados apresentados nas Figuras 3.8, 3.9, 3.10 e 3.11 é possível observar uma grande quantidade de instâncias corretas para a maioria das classes selecionadas. Nos limiares mais restritivos, Top 10 e Top 50, maiores valores de precisão foram alcançados, enquanto que nos limiares mais abrangentes, Top 100 e Top 200, houve uma tendência natural de perda de precisão. Observa-se também que existe uma variação de precisão entre as classes selecionadas, isso ocorreu devido aos fatores como complexidade do domínio e da coocorrência de instâncias das classe selecionadas com os padrões linguísticos utilizados. Por exemplo, nas classes Cidade e País que são de domínio geral, as medidas avaliadas obtiveram uma alta taxa de precisão em relação as outras classes de domínio mais específico, como Inseto, Pássaro e Sintoma.

Considerando o limiar Top 10, como apresentado na Figura 3.8, a heurística de NPE apresentou uma maior taxa de precisão em 9 das 15 classes analisadas, obtendo uma

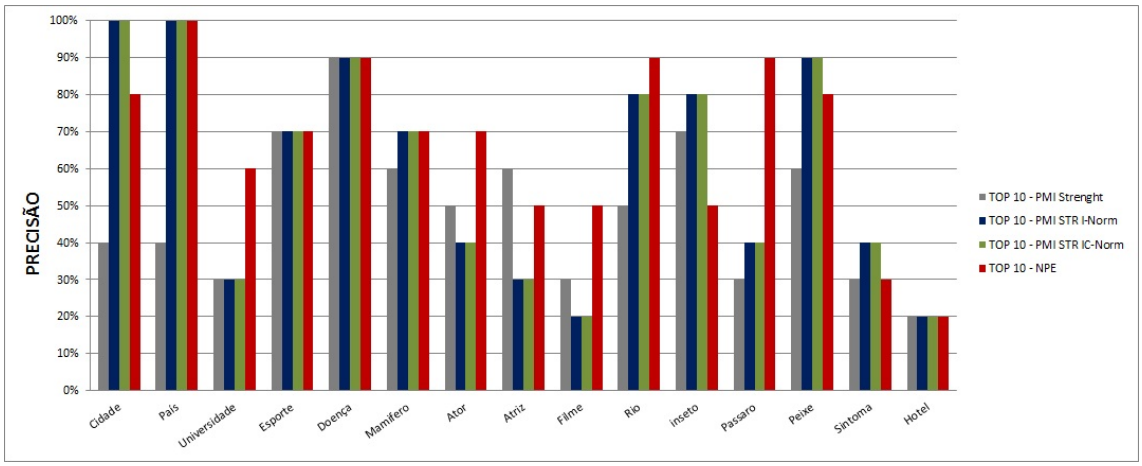

Figura 3.8 Resultados do Experimento I no limiar Top 10.

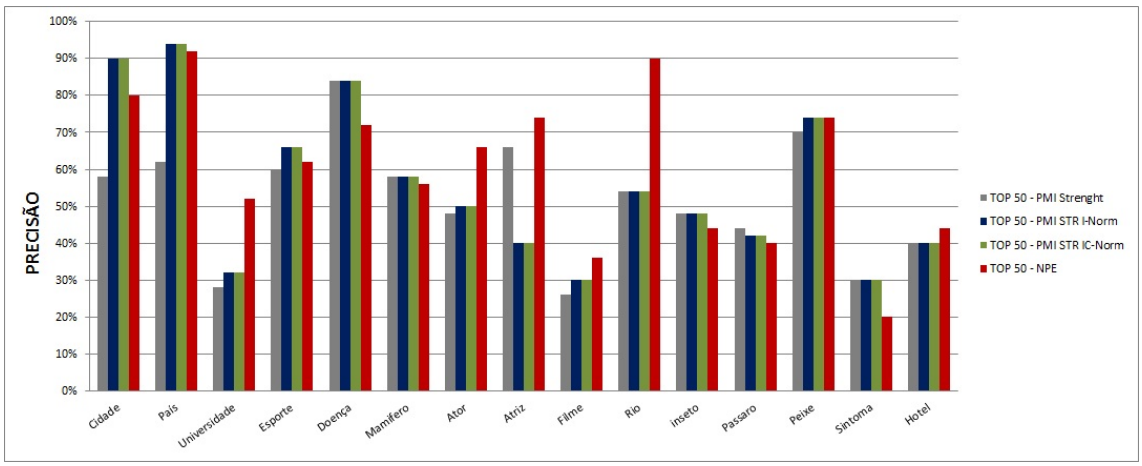

Figura 3.9 Resultados do Experimento I no limiar Top 50.

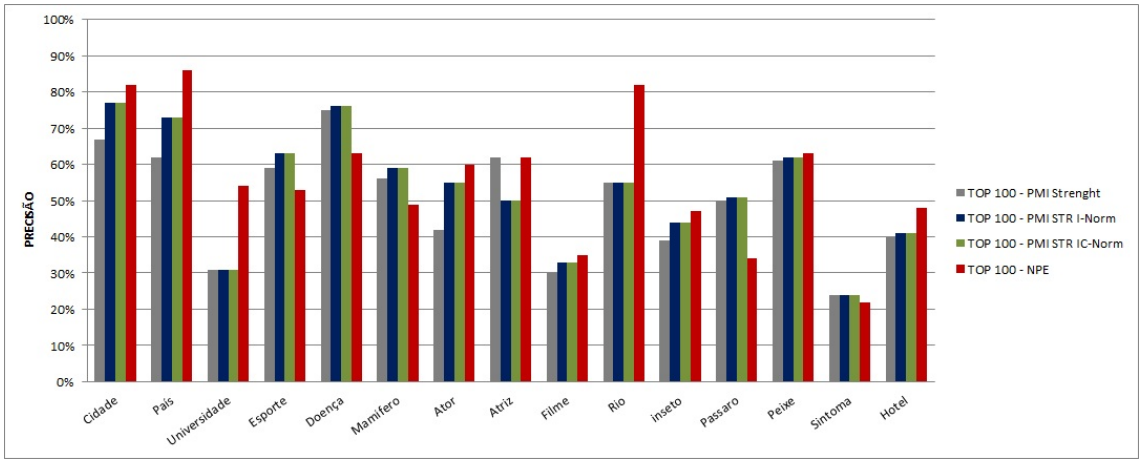

Figura 3.10 Resultados do Experimento I no limiar Top 100.

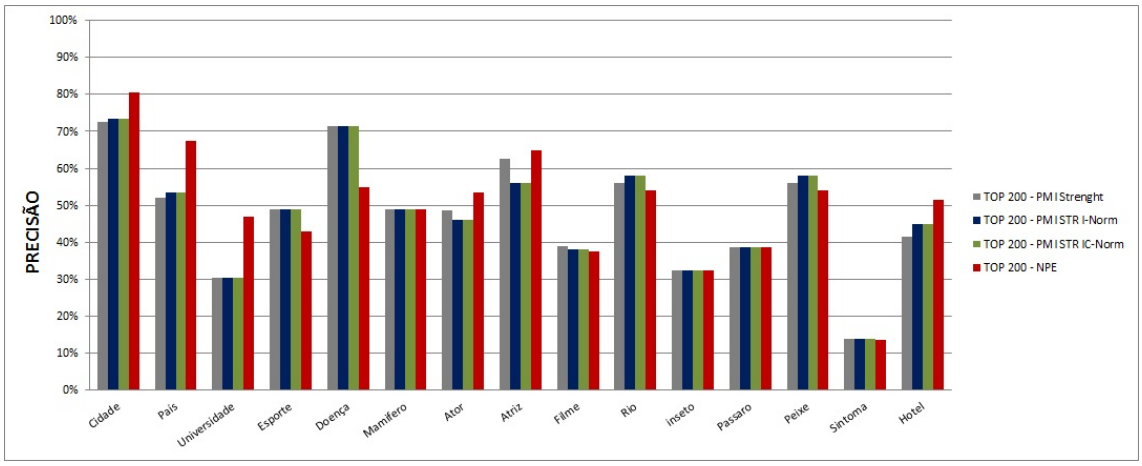

Figura 3.11 Resultados do Experimento I no limiar Top 200.

média geral de 67% de precisão. As variações PMI Str-INorm-Thresh e PMI Str-ICNorm-Thresh apresentaram maiores taxas de precisão em 8 das 15 classes, ficando ambas com uma média de 60% de precisão. Por fim, a variação PMI Strength apresentou melhores taxas de precisão apenas em 3 classes, ficando com uma média de precisão geral de 49%.

Analisando os limiares Top 50 e Top 100, como apresentado nas Figuras 3.9 e 3.10, o mesmo comportamento foi observado. A heurística de NPE continuou com melhor desempenho apresentando uma média geral de precisão de 60% no Top 50 e 48% no Top 100. Em segundo lugar, as variações PMI Str-INorm-Thresh e PMI Str-ICNorm-Thresh continuaram empatadas com 55% de precisão média no Top 50 e 41% no Top 100. Por último, a variação de PMI Strength obteve 52% de precisão média no Top 50 e 40% no Top 100.

No limiar Top 200, apresentado na Figura 3.11, as 4 medidas ficaram com médias de precisão muito próximas, com uma diferença de apenas 1% para a heurística de NPE que obteve precisão média de 49%. Enquanto isso, as 3 variações de PMI ficaram empatadas com uma média de 48% de precisão.

3.3.2.2 Conclusões do Experimento I

Com base nos experimentos executados, conclui-se que a heurística de NPE obteve melhores resultados do que as 3 variações da medida PMI em todos os limiares. Contudo, a diferença entre elas foi diminuindo à medida que o limiar foi aumentando. As variações PMI Str-INorm-Thresh e PMI Str-ICNorm-Thresh obtiveram melhores resultados do que a variação PMI Strength. Tais resultados indicam que os fatores de normalização usados pelo PMI Str-INorm-Thresh e PMI Str-ICNorm-Thresh apresentaram um impacto positivo nos resultados. Esse fato encorajou a execução do Experimento II que teve como finalidade identificar se a alteração do fator de normalização melhora a acurácia das medidas PMI estudadas.

3.3.3 Experimento II

Neste experimento são analisadas e validadas as medidas PMI Str-INorm-Z, PMI Str-ICNorm-Z, PMI Str-I-Norm-Hits0-Z, propostas na seção 3.1.3. Os experimentos foram conduzidos comparando com as medidas PMI Str-INorm-Thresh e PMI Str-ICNorm-Thresh, avaliadas por McDowell e Cafarella (2008), e com a medida PMI Str-I-Norm-Hits0 avaliada por (TOMAZ et al., 2012).

Neste experimento espera-se verificar quais das medidas PMI, propostas neste trabalho ou as apresentadas por Tomaz et al. (2012) e McDowell e Cafarella (2008), obtém a melhor acurácia na classificação de candidatos a instância.

Com finalidade de proporcionar as medidas testadas as mesmas chances, todos os testes foram executados com o mesmo conjunto de candidatos a instâncias e iniciaram-se na etapa de Classificação.

Na próxima subseção são apresentados os resultados e as discussões deste experimento.

3.3.3.1 Resultados e Discussões do Experimento II

Os gráficos apresentados nas Figuras 3.12, 3.13, 3.14, 3.15 demonstram os resultados obtidos para as classes selecionadas neste experimento. Em todos os gráficos foram feitas comparações em termos da medida de precisão para cada variação de PMI testada.

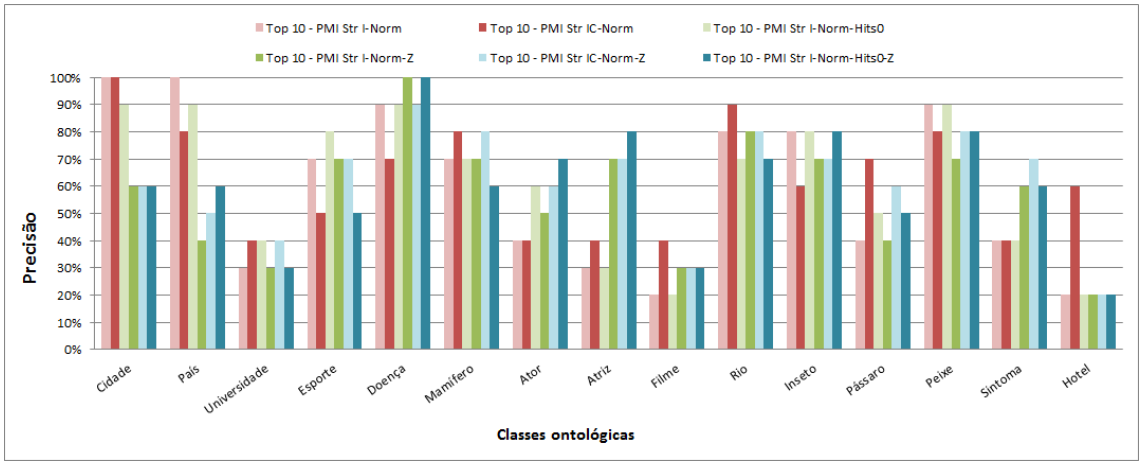

Figura 3.12 Resultados do Experimento II no limiar Top 10.

Pelos resultados apresentados nas Figuras 3.12, 3.13, 3.14 e 3.15 é possível identificar que as medidas propostas obtiveram uma média de precisão maior que as médias das outras medidas demonstradas nesses experimentos e em todos os limiares. Contudo, o teste estatístico **T-Student** demonstrou que não há diferença significativamente relevante entre elas, adotando 95% de confiança. Entretanto, é possível observar que a medida proposta *PMI Str IC-Norm-Z* obteve uma média de **52**% de precisão contra **49**% da *PMI Str IC-Norm*. Já as medidas *PMI Str I-Norm e PMI Str I-Norm Hits0* diferenciaram de suas variações em apenas (um) ponto em média.

Nos limiares Top 10 e Top 50 identificam-se maiores valores de precisão do que os limiares mais abrangentes Top 100 e Top 200 que tiveram uma tendência natural de perda

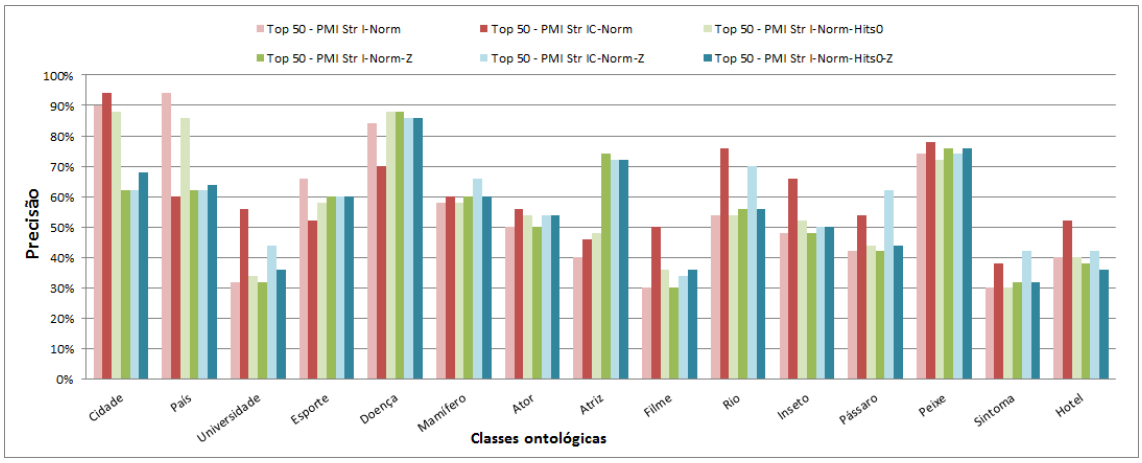

Figura 3.13 Resultados do Experimento II no limiar Top 50.

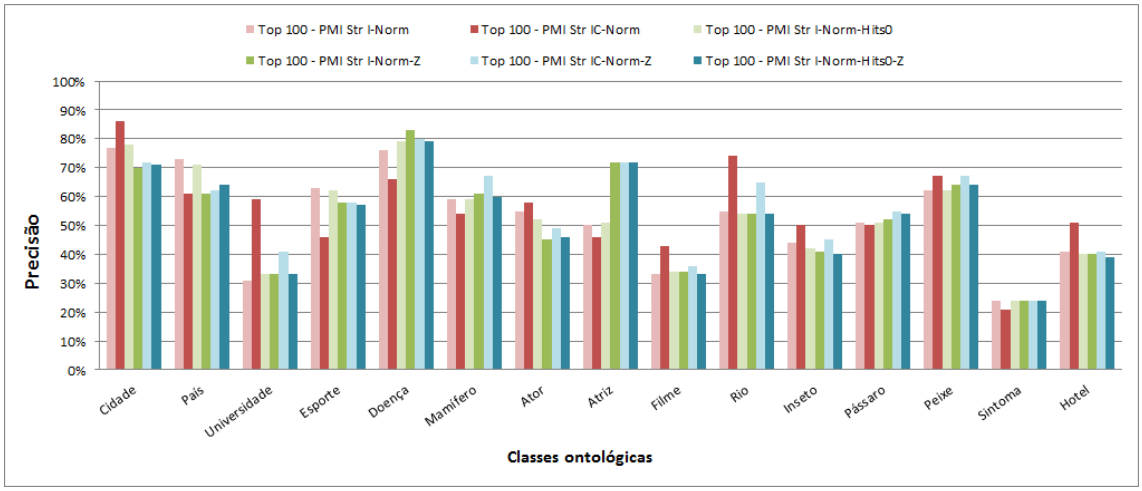

Figura 3.14 Resultados do Experimento II no limiar Top 100.

de precisão. Da mesma forma que o Experimento I, também se observou uma variação de precisão entre as classes selecionadas, ainda devido aos fatores como complexidade do domínio e a coocorrência de instâncias da classe selecionada com os padrões linguísticos utilizados.

3.3.3.2 Conclusões do Experimento II

Após a análise dos resultados foi possível concluir que, estatisticamente, as métricas propostas não obtiveram uma diferença significativamente relevante em comparação com as outras medidas estudadas, entretanto, as medidas propostas obtiveram uma grande quantidade de instâncias corretas para a maioria das quinze classes selecionadas. Esses resultados são encorajadores para que as medidas propostas possam ser evoluídas para atingirem melhores resultados.

Diante desse fato, surgiu a ideia de combinar a heurística de NPE, melhor medida

Figura 3.15 Resultados do Experimento II no limiar Top 200.

do experimento I, com a melhor medida PMI obtida nesse experimento, ou seja, PMI Str IC-Norm-Z. Sendo isto a motivação para o experimento III apresentado na próxima seção.

3.3.4 Experimento III

Neste experimento, as análises foram conduzidas efetuando uma combinação entre a medida heurística NPE com a medida PMI Str IC-Norm-Z. Esperava-se com isso identificar se a combinação entre as medidas geraria uma medida melhor do que as outras medidas discutidas e já abordadas no Experimento I e II.

Com a finalidade de proporcionar as medidas testadas as mesmas chances, todos os testes foram executados com o mesmo conjunto de candidatos a instâncias e ambas as medidas foram combinadas com valores variando de 0.1 a 0.9, como demonstra a Tabela 3.9.

Combinação	NPE		PMI Str IC-Norm-Z
C1 =	NPE * 0.1	+	PMI Str IC-Norm-Z * 0.9
C2 =	NPE * 0.2	+	PMI Str IC-Norm-Z * 0.8
C3 =	NPE * 0.3	+	PMI Str IC-Norm-Z * 0.7
C4 =	NPE * 0.4	+	PMI Str IC-Norm-Z * 0.6
C5 =	NPE * 0.5	+	PMI Str IC-Norm-Z * 0.5
C6 =	NPE * 0.6	+	PMI Str IC-Norm-Z * 0.4

Combinação	NPE		PMI Str IC-Norm-Z
C7 =	NPE * 0.7	+	PMI Str IC-Norm-Z * 0.3
C8 =	NPE * 0.8	+	PMI Str IC-Norm-Z * 0.2
C9 =	NPE * 0.9	+	PMI Str IC-Norm-Z * 0.1

Tabela 3.9: Combinação linear utilizada no experimento III

Supondo que a instância Salvador tem NPE = 3 e considerando a PMI Str IC-Norm-Z = 0,4, a combinação C1 conforme apresentada na Tabela 3.9 resulta na seguinte expressão: C1 = 3 * 0.1 + 0.4 * 0.9, ou seja, C1 = 0.66.

Na próxima seção são apresentadas as discussões e os resultados obtidos com este experimento.

3.3.4.1 Resultados e Discussões do Experimento III

Os gráficos apresentados nas Figuras 3.16, 3.17, 3.18, 3.19 demonstram os resultados obtidos para as classes selecionadas neste experimento. Em todos os gráficos foram feitas comparações em termos da medida de precisão para cada variação da PMI testada.

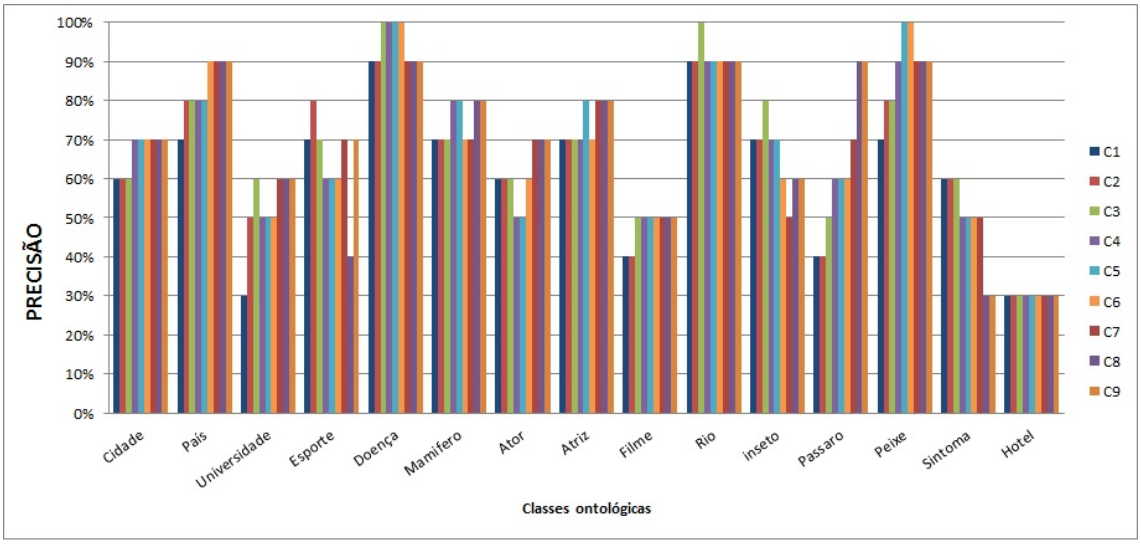

Figura 3.16 Resultados do Experimento III no limiar Top 10.

Analisando os resultados apresentados nas Figuras 3.16, 3.17, 3.18 e 3.19 foi possível observar que existem diferenças entre as medidas combinadas alcançando no limiar Top 10, uma diferença de 14% entre a menor combinação(C1) e a maior combinação(C9). Já nos limiares Top50, Top100 e Top200 essas diferenças foram diminuindo obtendo respectivamente 6%, 4% e 2% de precisão. Também é possível observar que a combinação(C9) obteve a maior média dentre as combinações analisadas, atingindo nos limiares Top10 70%, Top100 56% e Top200 48% de precisão.

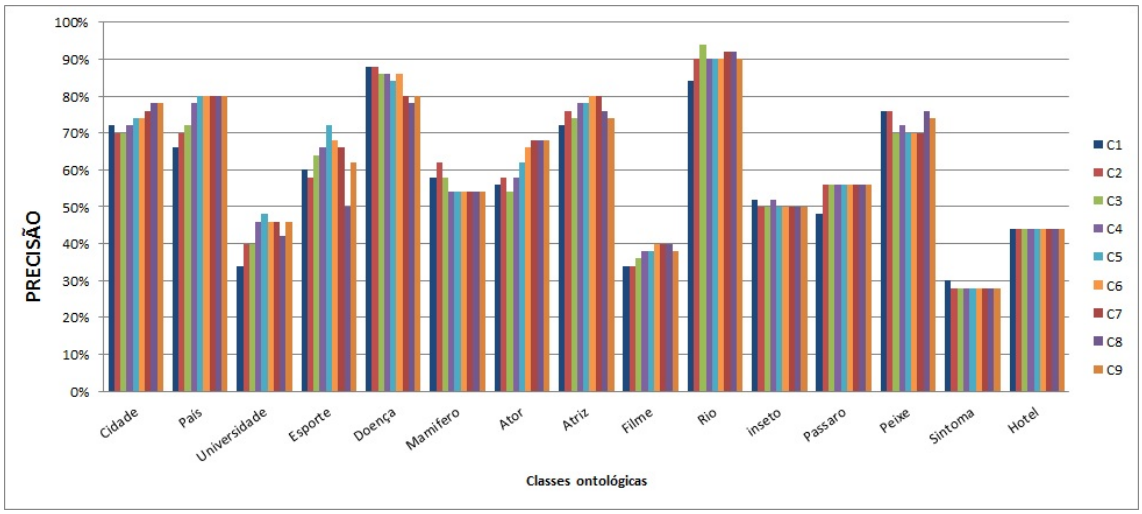

Figura 3.17 Resultados do Experimento III no limiar Top 50.

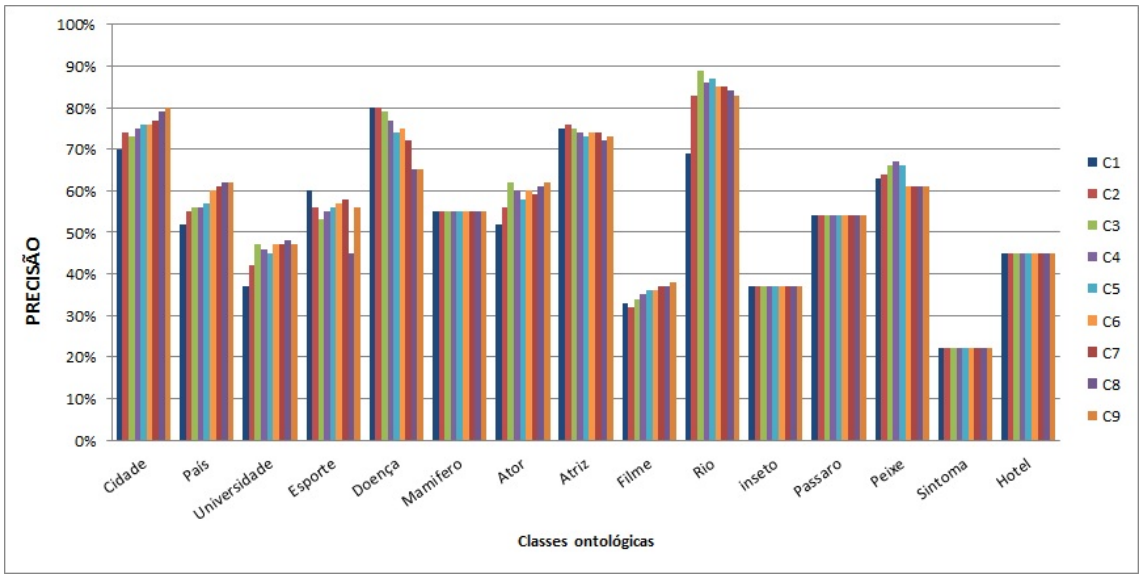

Figura 3.18 Resultados do Experimento III no limiar Top 100.

3.3.4.2 Conclusões do Experimento III

Após a análise dos resultados foi possível concluir, considerando os limiares (Top 10, 100 e 200), que a combinação(C9) foi capaz de extrair uma grande quantidade de instâncias corretas para a maioria das quinze classes selecionadas, especificamente no limiar Top 10 que obteve uma media superior a 70% de precisão para 10 das 15 classes analisadas.

Diante desses fatos e com finalidade de dirimir as dúvidas entre quais das medidas analisadas nos experimentos I, II e III devem ser utilizadas no método para Povoamento de Ontologias, foi apresentada na próxima seção uma análise comparativa entre as medidas NPE, a medida Combinada(C9) e a medida PMI Str-IC-Norm-Z.

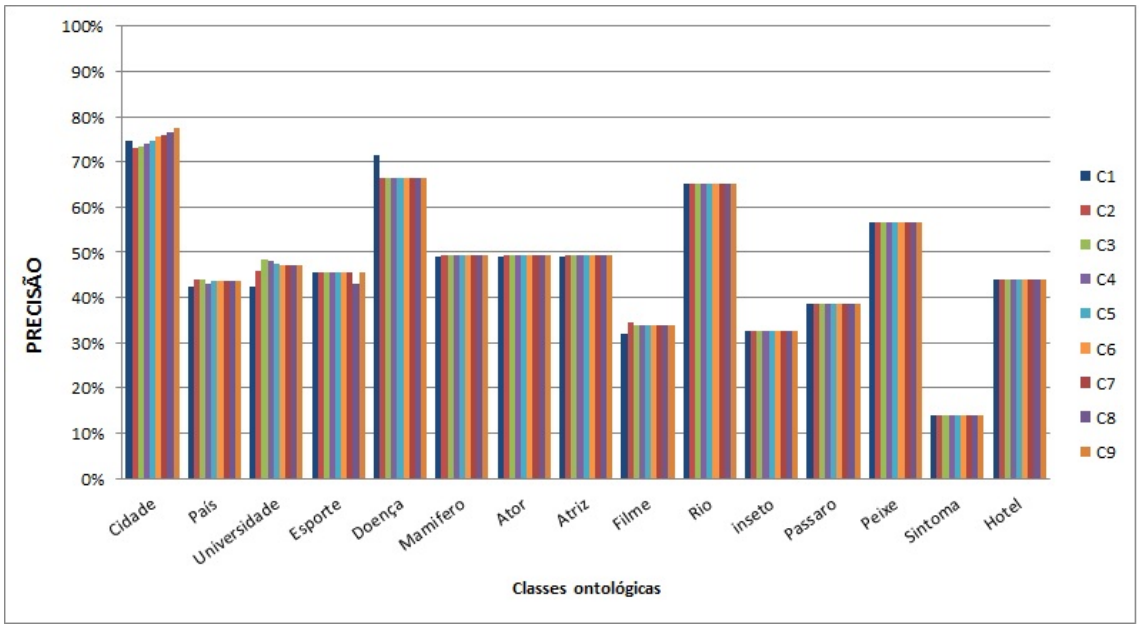

Figura 3.19 Resultados do Experimento III no limiar Top 200.

3.3.5 Comparativo entre os experimentos I, II e III

Esta seção demonstra um comparativo entre as melhores medidas destacadas dos experimentos I, II e III. Diante disso, considerando o Método para Povoamento de Ontologias proposto, é necessário identificar qual dessas medidas obtem uma melhor precisão para o corpus analisado. Para isso, foram comparados os resultados obtidos de cada uma das medidas analisadas: NPE, PMI Str IC-Norm-Z e a combinação(C9), como apresentadas nas Figuras 3.20, 3.21, 3.22 e 3.23.

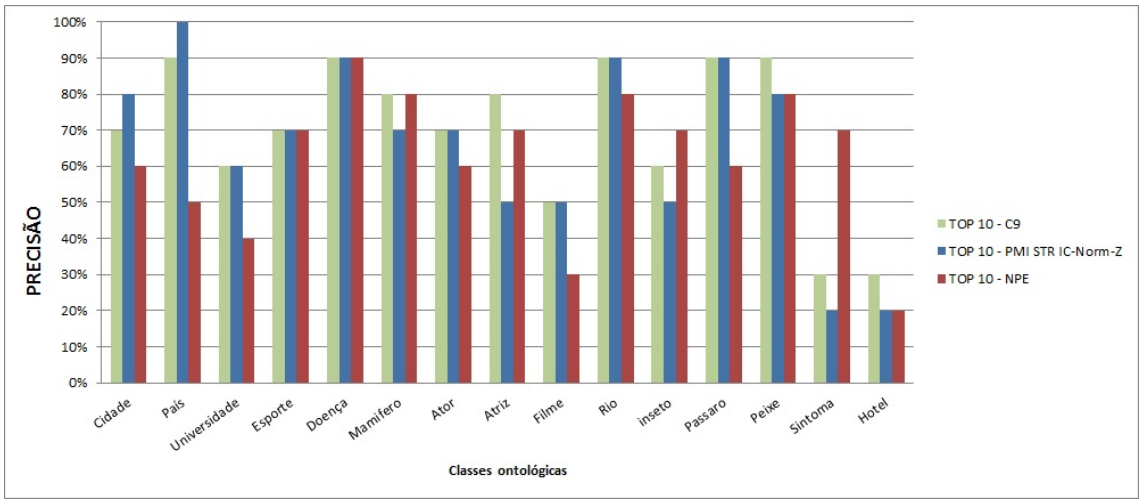

Figura 3.20 Resultados Comparativos no limiar Top 10.

Observou-se na Figura 3.20 que a medida combinada**(C9)** obteve uma média de 70%

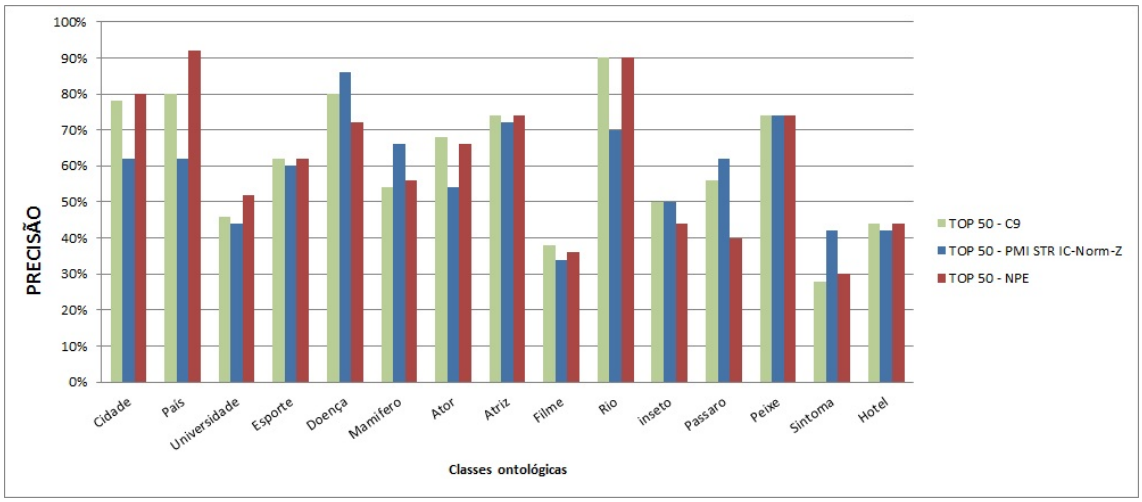

Figura 3.21 Resultados Comparativos no limiar Top 50.

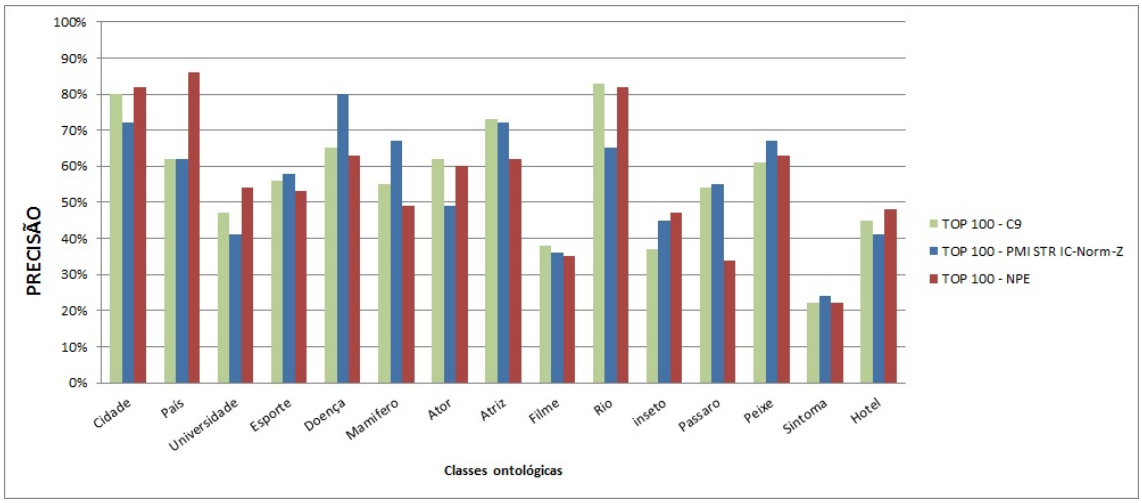

Figura 3.22 Resultados Comparativos no limiar Top 100.

de precisão, enquanto as medidas **PMI Str-IC-norm-Z** e **NPE** obtiveram 63%. Já no limiar Top 50, a medida **PMI Str-IC-Norm-Z** obteve uma precisão de 59%, em contra partida, a NPE e a medida combinada**(C9)** ficaram empatadas, ambas com 61% de precisão. Este mesmo comportamento foi identificado no limiar Top 100, no qual, todas as medidas obtiveram 56% de precisão. No Top 200, diferentemente dos outros limiares, a medida **PMI Str-IC-Norm-Z** obteve a maior media de precisão, 52%, diferenciando-se da NPE em 1 ponto em média e da combinação(C9) em 4 pontos em média.

Também é possível observar que a combinação(C9) obteve as maiores precisões considerando os limiares Top 10, Top 50 e Top 100, na qual, conseguiu classificar uma maior quantidade de instâncias válidas para a maioria das classes analisadas. Mesmo diante das diferenças obtidas, observou-se em alguns casos que a diferença entre as medidas combinadas chegaram a 100%, como é o caso das classes Universidade, Esporte, Pássaro

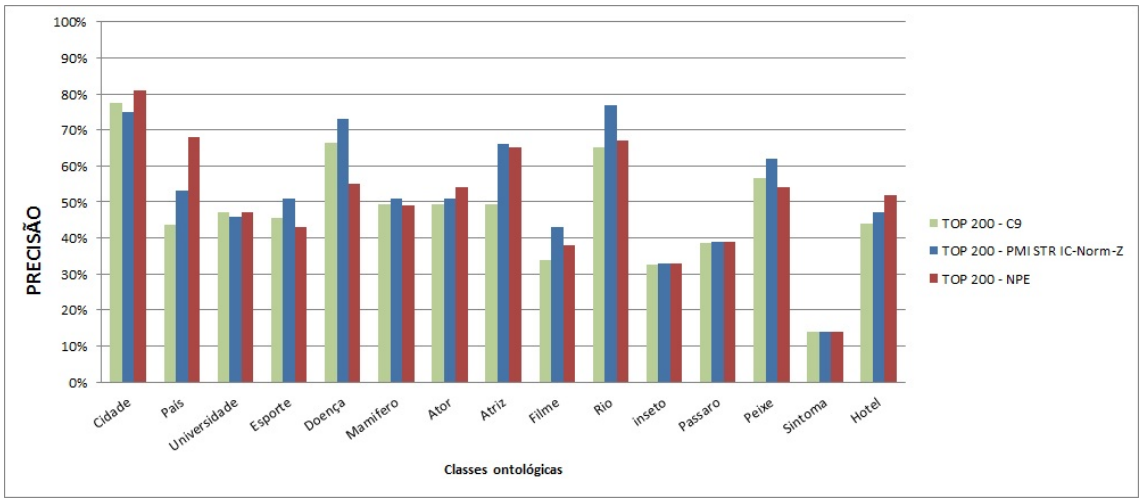

Figura 3.23 Resultados Comparativos no limiar Top 200.

e Sintoma.

Para avaliar se houve realmente uma diferenciação significativa entre os resultados obtidos, foi aplicado o teste de significância estatística T-Student. Esse teste estatístico foi realizado considerando os quatro limiares analisados (Top10, Top50, Top100 e Top200) e cada uma das variações de PMI analisadas. Diante desses fatos e da comparação estatística (Apêndice A), elencou-se a medida combinada(C9) para utilização na fase de Classificação do Método para Povoamento de Ontologias proposta.

Capítulo

4

CONCLUSÕES E TRABALHOS FUTUROS

Na conclusão deste trabalho, apresentam-se as considerações finais, as percepções acerca de seus resultados, limitações e contribuições científicas, bem como propostas de trabalhos futuros.

4.1 CONCLUSÕES

Este trabalho apresentou um método não supervisionado para o povoamento de ontologias a partir de textos escritos em linguagem natural em português, utilizando a Web como grande fonte de informações. Além disso, executou-se experimentações das medidas estatísticas de classificação de instâncias de classes ontológicas com textos extraídos da *Web* e no idioma português. Nesse mesmo sentido, foi analisada a proposta da alteração do normalizador das medidas *PMI Str I-Norm, PMI Str IC-Norm* e *PMI Str I-NormHits0* para remoção do Percent25 e a inclusão do Desvio Padrão.

A relevância do problema abordado decorre da quantidade de informações desestruturadas presentes na *Web* e da necessidade da aquisição e gerenciamento dessas informações para a criação de bases de conhecimento. Além disso, observou-se a ausência de experimentos de medidas de classificação de instâncias de classes ontológicas direcionados aos textos no idioma português extraídos da *Web*.

O Método para Povoamento de Ontologias apresentado difere-se de outros trabalhos nos seguintes quesitos: (a) utiliza técnicas e ferramentas de Processamento de Linguagem Natural voltadas para o idioma português; (b) utiliza uma ontologia como guia do processo de extração, como também para armazenamento; (c) utiliza os padrões de Hearst (traduzidos para o português) tanto na coleta quanto na etapa de extração; (d) linguagem de formação do corpus (textos na língua portuguesa); (e) domínio das classes ontológicas; (f) alteração do normalizador das medidas clássicas da PMI para inclusão do Desvio Padrão e (g) execução de experimentações com corpus coletado da web no idioma português com ranqueamento utilizando a medida PMI.

Após a análise dos resultados foi possível concluir que estatisticamente as medidas PMI propostas não obtiveram uma diferença significativamente relevante em comparação

com as outras medidas estudadas. Entretanto, uma das medidas propostas, a PMI combinada(C9) obteve uma precisão média de 70% e conseguiu elencar uma grande quantidade de instâncias corretas para a maioria das classes selecionadas, mesmo adotando limiares mais abrangentes, o que implicou diretamente na eficácia do Método para Povoamento de Ontologias e na ratificação da utilização da Web como fonte de informações. Tal fato demostrou ser uma boa opção para a classificação dos candidatos a instâncias, utilizando a medida PMI, como medida de relevância para explorar a alta redundância do conteúdo disponível na web e como corpus para a extração de candidatos a instâncias.

Em todos os experimentos realizados, a execução do método foi executado sem nenhuma interação manual. Os resultados produzidos demonstraram que as medidas propostas e as analisadas apresentaram resultados variados de acordo com domínio da classe selecionada, nesse sentido, optar em realizar customizações de forma a escolher filtros específicos para o domínio da classe selecionada poderá melhorar a acurácia dos resultados obtidos.

Os resultados do trabalho como um todo são satisfatórios, tanto pela integração das diferentes ferramentas e tarefas empregadas quanto pela qualidade dos recursos produzidos.

Ressalta-se ainda o escopo das áreas investigadas durante a realização deste trabalho que circundam a área de Ontologias, PLN, Extração de Informação e SEIBOs. Considera-se que as contribuições resultantes deste trabalho são um acréscimo significativo para a área de SEIBOs no contexto da língua portuguesa e a integração entre as áreas analisadas, permitindo também efetuar importantes considerações sobre as aplicações e as diferentes técnicas existentes e disponíveis para o idioma português.

4.2 CONTRIBUIÇÕES

- Contribuições principais

  (1) Método não supervisionado, automático e independente de domínio para o povoamento de ontologias com termos extraídos da Web no idioma português que permite a extração de instâncias de classes a partir de uma ontologia de entrada;

  (2) Sistema para validação do método proposto com validação de todas suas fases;

  (3) Medida de confiança, para classificação de instâncias de classes ontológicas, validada com termos extraídos da Web no idioma português;

  (4) Avaliação das medidas da PMI disponíveis na literatura em detrimento das medidas propostas. Essa análise permitiu identificar quais as limitações de cada medida PMI estudada.

- Recursos

  (1) Sistema para povoamento de ontologias, validado para diferentes domínios, focado no idioma português, classificado com base na medida de Precisão e que utiliza uma medida de classificação de instâncias de classes ontológicas baseada na PMI.

  (2) Mapeamento de novos padrões de busca: [CANDIDATO] É UM TIPO DE [CLASSE]; [CANDIDATOS] SÃO [CLASSE]; [CLASSE] COMO POR EXEMPLO

[CANDIDATO]; [CANDIDATO] È UM EXEMPLO DE [CLASSE]; que podem integrar os padrões de Hearts traduzidos para português por Baségio.

- Artigos Publicados

  (1) LIMA, F.; OLIVEIRA, H.; SALVADOR, L. An unsupervised method for ontology population from textual sources on the web. In: Proceedings of the Annual Conference on Brazilian Symposium on Information Systems: Information Systems: A Computer Socio-Technical Perspective - Volume 1. Porto Alegre, Brazil, Brazil: Brazilian Computer Society, 2015. (SBSI 2015), p. 23:16323:170. Disponível em: http://dl.acm.org/ citation.cfm?id=2814058.2814086.

  (2) LIMA, F. d. S.; SALVADOR, L. d. N. Toward a scoring schema to rank candidate instances of ontological classes: Extracting brazilian portuguese texts from the web. In: Proceedings of the 21st Brazilian Symposium on Multimedia and the Web. New York,NY, USA: ACM, 2015. (WebMedia '15), p. 8184. ISBN 978-1-4503-3959-9. Disponível em: http://doi.acm.org/10.1145/2820426.2820465.

4.3 LIMITAÇÕES DO MÉTODO

(1) Extração de instâncias de classes ontológicas. Diante do alcance da proposta, este trabalho teve como foco apenas a extração de instâncias de classes ontológicas;

(2) Abrangência da extração dos candidatos a instância. O método proposto seleciona apenas palavras que são classificadas como sintagmas nominais. Apesar da grande quantidade de instâncias que podem ser classificadas, diversos candidatos que não estiverem nesse padrão nunca serão encontrados;

(3) Contexto da Extração. O método proposto seleciona conceitos baseados apenas nas classes, não considerando o conjunto de conceitos e relações sobre um domínio especifico, o que implica que cada conceito definido possui algum significado para o domínio representado, ou seja, um determinado conceito pode ter várias interpretações, um significado X para um domínio e um Y para outro;

(4) Diferenciação entre subclasses e instâncias. No método proposto não são executadas essas diferenciações. Como afirmam McDowell e Cafarella (2008), dependendo do candidato avaliado, essa diferenciação é complexa até mesmo para um ontologista. Nos experimentos realizados todos os candidatos foram considerados corretos, seguindo os mesmos critérios adotados nos trabalhos de (Etzioni et al., 2004a; McDowell e caffarella, 2008; Tomaz et. a, 2012);

4.4 TRABALHOS FUTUROS

Ao decorrer deste trabalho foram identificadas algumas limitações que definiram as ideias para expandir e contemplar outros aspectos do método proposto. São elas:

(1) Estender o método proposto para contemplar o povoamento de relações e também de propriedades pertencentes aos conceitos;

(2) Integrar um módulo para identificação de sinônimos para a filtragem de candidatos a instâncias que possuem o mesmo valor semântico;

(3) Explorar recursos semânticos disponíveis para o idioma português, por exemplo, o OpenWordnet-PT [1];

(4) Realizar outros experimentos simulando novas iterações do método proposto, buscando identificar o comportamento do processo de extração a cada iteração executada;

(5) Explorar axiomas como filtros, funcionando como mais uma medida para avaliar os candidatos a instâncias extraídos;

(6) Desenvolvimento de um método que realize a diferenciação entre o que deve ser extraído como instância e o que deve ser subclasse;

(7) Construção de um corpus anotado que permitirá a realização de experimentos utilizando as tradicionais medidas de Precisão, Cobertura e Medida-F. Esse corpus possibilitará que metodologias propostas por outros autores sejam comparadas com o método demonstrado neste trabalho;

(8) Execução do método proposto com os novos padrões identificados manualmente neste trabalho.

[1] http://logics.emap.fgv.br/wn/

Apêndice

A

TESTES ESTATÍSTICOS

Para avaliar as Medidas de Classificação de Candidatos a Instâncias foi executado um teste estatístico visando identificar o comportamento da média de precisão de cada medida estudada.

1) Definição das Hipóteses

Com o objetivo de analisar a existência de uma diferença na precisão entre as medidas de classificação dos candidatos a instâncias apresentadas no capitulo 3, e, em caso positivo, qual a melhor e em quais limiares. As medidas ou suas variações foram testadas em pares e para cada par de medidas, foram testadas duas hipóteses, nesta ordem: (1) médias iguais ou diferentes e (2) médias iguais ou a média do primeiro ser maior que a do segundo. Caso no primeiro teste a hipótese nula seja rejeitada (p -valor ¡= a), realizados o segundo teste. Caso o segundo teste não rejeite a hipótese nula (p - valor ¿ a), conclui-se que a média de precisão da primeira medida (M1) é menor.

Os testes de hipóteses utilizados foram o Teste T-student pareado ou Teste Wilcoxon pareado, ambos, com nível de significância de 5% (a = 0,05). Para decidir qual teste utilizar, foi aplicado um teste de aderência conforme descrito a seguir.

2) Teste de Aderência

Para decidir qual teste de hipótese utilizar, paramétrico ou não paramétrico, primeiro foi necessário aplicar um teste de aderência para verificar a normalidade dos dados. Para tal, o teste de Shapiro foi aplicado. Esse teste foi escolhido por ser um dos mais utilizados na literatura e por ter apresentado os melhores resultados na comparação entre testes de normalidade, conforme demonstrado em (Leotti et al., 2005).

São apresentados na Tabela A.1 os p-valores obtidos com a execução do teste de Shapiro.

PMIs	TOP200	TOP100	TOP50	TOP10
PMI Streght	0.9515	0.4585	0.5756	0.4071
PMI I-norm	0.9147	0.7708	0.1535	0.0643
PMI IC-norm	0.8494	0.7571	0.4720	0.0739

PMI I-norm-Hits0	0.9829	0.8580	0.1211	0.0562
PMI NPE	0.9483	0.6586	0.6615	0.1876
PMI I-norm-Z	0.6518	0.9843	0.5877	0.4736
PMI IC-norm-Z	0.5722	0.7709	0.8954	0.2138
PMI I-norm-Hits0-Z	0.8336	0.8111	0.6246	0.6176

Tabela A.1: P-valores do teste de Shapiro

Analisando a Tabela A.1, pode-se observar que todos os p-values foram superiores ao nível de significância adotado (5%), logo, a hipótese nula não é rejeitada e, portanto as amostras se aproximam de uma distribuição normal. Diante disso, será utilizado para as comparações dessas medidas o Teste T-Student pareado, apresentados na Tabela A.2.

PMIs	TOP200	TOP100	TOP50	TOP10
PMI Streght x PMI I-norm	0.6487	0.3775	0.4304	0.2272
PMI Streght x PMI IC-norm	0.5379	0.1746	0.0735	0.0622
PMI I-norm x PMI IC-norm	0.8898	0.6612	0.456	0.7763
PMI IC-norm x NPE	0.8092	0.9312	0.9663	0.6889
PMI IC-norm x PMI I-norm-Hits0	0.9637	0.6454	0.4851	0.8777
NPE x PMI I-norm-Hits0	0.7851	0.6172	0.517	0.6175
NPE x PMI I-norm-Z	0.8696	0.6188	0.3232	0.3149
NPE x PMI IC-norm-Z	0.8113	0.9494	0.7358	0.6257
PMI I-norm-Z x PMI IC-norm-Z	0.6905	0.6385	0.429	0.5443

PMIs	TOP200	TOP100	TOP50	TOP10
PMI IC-norm-Z x PMI I-norm-Hits0-Z	0.7248	0.6207	0.5571	0.7936

Tabela A.2: P-valores do Teste-T-student

É possível observar na Tabela A.2, que em todos os limiares adotados não houve variação de precisão signifiicativamente relevante, o que foi comprovado usando o teste de significância estatístico T-Student com um nível de confiança de 95% (a = 0,05). Diante disso, para identificar a medida mais adequada para a abordagem proposta, utilizou-se o total de intâncias validadas que cada medida consegue identificar, considerando todos os limiares.

REFERÊNCIAS BIBLIOGRÁFICAS

ALVES, C. G. d. F. Um Processo Independente de Domínio para o Povoamento Automático de Ontologias a partir de Fontes Textuais. 2013.

AMANN, B.; FUNDULAKI, I. Integrating ontologies and thesauri to build rdf schemas. In: *In ECDL-99: Research and Advanced Technologies for Digital Libraries, Lecture Notes in Computer Science*. [S.l.]: Springer-Verlag, 1999. p. 234–253.

APPELT, D.; ISRAEL, D. *Introduction to Information Extraction Technology*. Stockholm, Schweden: Joint Conference of Artificial Intelligence, Stockholm, Schweden, 1999.

BASéGIO, T. L. Uma Abordagem Semi-automática para Identificação de Estruturas Ontológicas a partir de Textos na Língua Portuguesa do Brasil. p. 1–124, 2007.

BEN-DOV, M.; FELDMAN, R. *Text Mining and information extraction. In The Data Mining and Knowledge Discovery Handbook*. Tel-Aviv, Israel: Springer Science, 2010. ISBN 978-0-387-09822-7.

BERNERSLEE, T.; HENDLER, J.; LASSILA, O. The semantic web - a new form of web content that is meaningful to computers will unleash a revolution of new possibilities. *Scientific American*, 2001.

BONTCHEVA, K. et al. *Semantic Annotation and Human Language Technology*. John Wiley & Sons, Ltd, 2006. 29–50 p. ISBN 9780470030332. Disponível em: ⟨http://dx.doi.org/10.1002/047003033X.ch3⟩.

BRILL, E. Processing natural language without natural language processing. In: GELBUKH, A. (Ed.). *Computational Linguistics and Intelligent Text Processing*. Springer Berlin Heidelberg, 2003, (Lecture Notes in Computer Science, v. 2588). p. 360–369. ISBN 978-3-540-00532-2. Disponível em: ⟨http://dx.doi.org/10.1007/3-540-36456-0_37⟩.

BRUCKSCHEN, M. *Reconhecimento de Entidades Nomeadas e Relações no Domínio de Privacidade e Responsabilização* . Dissertação (Dissertação de Mestrado) — PUCRS, 2010.

BUITELAAR, P. et al. Ontology-based information extraction and integration from heterogeneous data sources. *Int. J. Hum.-Comput. Stud.*, Academic Press, Inc., Duluth, MN, USA, v. 66, n. 11, p. 759–788, nov. 2008. ISSN 1071-5819. Disponível em: ⟨http://dx.doi.org/10.1016/j.ijhcs.2008.07.007⟩.

BUITELAAR, P.; SIEGEL, M. Ontology-based information extraction with soba. In: *15th International Conference on Language Resources and Evaluation*. [S.l.: s.n.], 2006.

CARLOS, L. R. J. et al. Ontolp : Engenharia de ontologias em língua portuguesa. 2008.

CARLSON, A. et al. Toward an architecture for never-ending language learning. In: *In AAAI*. [S.l.: s.n.], 2010.

CHAVES, A.; RINO, L. The mitkov algorithm for anaphora resolution in portuguese. In: TEIXEIRA, A. et al. (Ed.). *Computational Processing of the Portuguese Language*. Springer Berlin Heidelberg, 2008, (Lecture Notes in Computer Science, v. 5190). p. 51–60. ISBN 978-3-540-85979-6. Disponível em: ⟨http://dx.doi.org/10.1007/978-3-540-85980-2_6⟩.

CIMIANO, P. *Ontology Learning and Population from Text: Algorithms, Evaluation and Applications*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006. ISBN 0387306323.

CIMIANO, P.; HANDSCHUH, S.; STAAB, S. Towards the self-annotating web. *Proceedings of the 13th conference on World Wide Web - WWW '04*, ACM Press, New York, New York, USA, p. 462, 2004. Disponível em: ⟨http://portal.acm.org/citation.cfm?doid=988672.988735⟩.

CORCHO, O.; LOPEZ, M. F.; PEREZ, A. G. Ontologies for software engineering and software technology. Springer-Verlagl. 2006.

CORREIA, J. D. S. S. *Um processo para a aquisição de relações taxonômicas de uma ontologia*. Dissertação (Dissertação de Mestrado) — Universidade Federal do Maranhão, 2011.

CUNNINGHAM, H. GATE, a General Architecture for Text Engineering. *In Computers and the Humanities*, v. 36, p. 223–254, 2002.

CUNNINGHAM, H. *Information Extraction, Automatic*. [S.l.]: Elsevier, 2006. 665–677 p.

DRUMOND, L.; GIRARDI, R. Extracting ontology concept hierarchies from text using markov logic. In: *Proceedings of the 2010 ACM Symposium on Applied Computing*. New York, NY, USA: ACM, 2010. (SAC '10), p. 1354–1358. ISBN 978-1-60558-639-7. Disponível em: ⟨http://doi.acm.org/10.1145/1774088.1774379⟩.

EMBLEY, D. Toward semantic understanding: an approach based on information extraction ontologies. *In Proceedings of the 15th Australasian Database Conference*, Darlinghurst, Australia, 2004.

EMBLEY, D. W. et al. Ontology-based extraction and structuring of information from data-rich unstructured documents. In: . [S.l.: s.n.], 1998. p. 52–59.

ETZIONI, O. et al. Web-Scale Information Extraction in KnowItAll (Preliminary Results). p. 100–110, 2004.

FENSEL, D. et al. Product data integration in b2b e-commerce. *IEEE Intelligent Systems*, IEEE Educational Activities Department, Piscataway, NJ, USA, v. 16, n. 4, p. 54–59, jul. 2001. ISSN 1541-1672. Disponível em: ⟨http://dx.doi.org/10.1109/5254.941358⟩.

FENSEL, D. et al. Oil in a nutshell. In: *Proceedings of the 12th European Workshop on Knowledge Acquisition, Modeling and Management*. London, UK, UK: Springer-Verlag, 2000. (EKAW '00), p. 1–16. ISBN 3-540-41119-4. Disponível em: ⟨http://dl.acm.org/citation.cfm?id=645361.650830⟩.

FERREIRA, E.; BALSA, J.; BRANCO, A. Combining rule-based and statistical methods for named entity recognition in portuguese. In: *TIL- V Workshop de Tecnologia da Informação e da Linguagem Humana*. Lisbon,PT: SBC, 2007. p. 1615–1624. Disponível em: ⟨www.di.fc.ul.pt/~ahb/FerreiraBalsaBranco2007.pdf⟩.

GóMEZ-PÉREZ, A.; FERNáNDEZ-LóPEZ, M.; CORCHO, O. *Ontological Engineering: With Examples from the Areas of Knowledge Management, e-Commerce and the Semantic Web. (Advanced Information and Knowledge Processing)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2007. ISBN 1846283965.

GRISHMAN, R. Information extraction: Techniques and challenges. In: *International Summer School on Information Extraction: A Multidisciplinary Approach to an Emerging Information Technology*. London, UK, UK: Springer-Verlag, 1997. (SCIE '97), p. 10–27. ISBN 3-540-63438-X. Disponível em: ⟨http://dl.acm.org/citation.cfm?id=645856.669801⟩.

GRUBER, T. R. A translation approach to portable ontology specifications. *Knowl. Acquis.*, Academic Press Ltd., London, UK, UK, v. 5, n. 2, p. 199–220, jun. 1993. ISSN 1042-8143. Disponível em: ⟨http://dx.doi.org/10.1006/knac.1993.1008⟩.

GUARINO, N. Formal ontologies and information systemsl. First International Conference. 1998.

HEARST, M. A. Automatic Acquisition of Hyponyms from Large Text Corpora. p. 23–28, 1992.

JACOBS, P. S.; RAU, L. F. Innovations in text interpretation. *Artif. Intell.*, Elsevier Science Publishers Ltd., Essex, UK, v. 63, n. 1-2, p. 143–191, out. 1993. ISSN 0004-3702. Disponível em: ⟨http://dx.doi.org/10.1016/0004-3702(93)90016-5⟩.

KIM, H. M. *Developing Ontologies to Enable Knowledge Management: Integrating Business Process and Data Driven Approaches*. Keele Street Toronto, Canada, 2000.

LAVELLI, A. et al. Evaluation of machine learning-based information extraction algorithms: criticisms and recommendations. *Language Resources and Evaluation*, Springer Netherlands, v. 42, n. 4, p. 361–393, 2008. ISSN 1574-020X. Disponível em: ⟨http://dx.doi.org/10.1007/s10579-008-9079-3⟩.

LI, Y.; BONTCHEVA, K. Hierarchical, perceptron-like learning for ontology-based information extraction. In: *Proceedings of the 16th International Conference on World Wide Web*. New York, NY, USA: ACM, 2007. (WWW '07), p. 777–786. ISBN 978-1-59593-654-7. Disponível em: ⟨http://doi.acm.org/10.1145/1242572.1242677⟩.

LIMA, F.; OLIVEIRA, H.; SALVADOR, L. An unsupervised method for ontology population from textual sources on the web. In: *Proceedings of the Annual Conference on Brazilian Symposium on Information Systems: Information Systems: A Computer Socio-Technical Perspective - Volume 1*. Porto Alegre, Brazil, Brazil: Brazilian Computer Society, 2015. (SBSI 2015), p. 23:163–23:170. Disponível em: ⟨http://dl.acm.org/citation.cfm?id=2814058.2814086⟩.

LIMA, F. d. S.; SALVADOR, L. d. N. Toward a scoring schema to rank candidate instances of ontological classes: Extracting brazilian portuguese texts from the web. In: *Proceedings of the 21st Brazilian Symposium on Multimedia and the Web*. New York, NY, USA: ACM, 2015. (WebMedia '15), p. 81–84. ISBN 978-1-4503-3959-9. Disponível em: ⟨http://doi.acm.org/10.1145/2820426.2820465⟩.

MAEDCHE, A.; NEUMANN, G.; STAAB, S. Bootstrapping an ontology-based information extraction system. In: SZCZEPANIAK, P. et al. (Ed.). *Intelligent Exploration of the Web*. Physica-Verlag HD, 2003, (Studies in Fuzziness and Soft Computing, v. 111). p. 345–359. ISBN 978-3-7908-2519-0. Disponível em: ⟨http://dx.doi.org/10.1007/978-3-7908-1772-0_21⟩.

MAEDCHE, A.; STAAB, S. Ontology learning for the semantic web. *Intelligent Systems, IEEE*, v. 16, n. 2, p. 72–79, Mar 2001. ISSN 1541-1672.

MAEDCHE, A.; STAAB, S. Ontology learning. In: STAAB, S.; STUDER, R. (Ed.). *Handbook on Ontologies*. Springer Berlin Heidelberg, 2004, (International Handbooks on Information Systems). p. 173–190. ISBN 978-3-662-11957-0. Disponível em: ⟨http://dx.doi.org/10.1007/978-3-540-24750-0_9⟩.

MAHESH, K. *Ontology Development for Machine Translation: Ideology and Methodology*. 1996.

MAYNARD, D.; LI, Y.; PETERS, W. Nlp techniques for term extraction and ontology population. In: *Proceedings of the 2008 Conference on Ontology Learning and Population: Bridging the Gap Between Text and Knowledge*. Amsterdam, The Netherlands, The Netherlands: IOS Press, 2008. p. 107–127. ISBN 978-1-58603-818-2. Disponível em: ⟨http://dl.acm.org/citation.cfm?id=1563823.1563834⟩.

MCDOWELL, L. K.; CAFARELLA, M. Ontology-driven information extraction with ontosyphon. In: *Proceedings of the 5th International Conference on The Semantic Web*. [S.l.: s.n.], 2006. (ISWC'06), p. 428–444. ISBN 3-540-49029-9, 978-3-540-49029-6.

MCDOWELL, L. K.; CAFARELLA, M. Ontology-driven, unsupervised instance population. *Web Semant.*, Elsevier Science Publishers B. V., Amsterdam, The Netherlands,

The Netherlands, v. 6, n. 3, p. 218–236, set. 2008. ISSN 1570-8268. Disponível em: ⟨http://dx.doi.org/10.1016/j.websem.2008.04.002⟩.

MONLLAÓ, C. V. *Ontology-based Information Extraction*. Tese (Doutorado) — Citeseer, 2011.

MORAIS, E. A. M.; AMBRóSIO, A. P. L. *Ontologias: conceitos, usos, tipos, metodologias ferramentas e linguagens*. [S.l.], 2007.

MOTTA, E.; ANDREATTA, A.; SIQUEIRA, S. Populating a domain ontology from web historical dictionaries and encyclopedias. *Proceedings of the 2008 Euro American Conference on Telematics and Information Systems - EATIS '08*, ACM Press, New York, New York, USA, p. 1–8, 2008. Disponível em: ⟨http://portal.acm.org/citation.cfm?doid=1621087.1621108⟩.

MOTTA, E. N. *Preenchimento Semi-automático de Ontologis de Domínio a Partir de Textos em Língua Portuguesa*. Dissertação (Dissertação de Mestrado) — Universidade Federal do Estado do Rio de Janeiro, 2009.

MUSLEA, I.; MINTON, S.; KNOBLOCK, C. A hierarchical approach to wrapper induction. In: . [S.l.]: ACM Press, 1999. p. 190–197.

NOY, N. F.; MCGUINNESS, D. L. et al. *Ontology development 101: A guide to creating your first ontology*. [S.l.]: Stanford knowledge systems laboratory technical report KSL-01-05 and Stanford medical informatics technical report SMI-2001-0880, 2001.

OLIVEIRA, H. T. A. de. *Um Método Não Supervisionado para Povoamento de Ontologias na Web*. Dissertação (Dissertação (mestrado)) — Universidade Federal de Pernambuco, 2013.

PETASIS, G. et al. Ontology population and enrichment: State of the art. In: PALIOURAS, G.; SPYROPOULOS, C.; TSATSARONIS, G. (Ed.). *Knowledge-Driven Multimedia Information Extraction and Ontology Evolution*. Springer Berlin Heidelberg, 2011, (Lecture Notes in Computer Science, v. 6050). p. 134–166. ISBN 978-3-642-20794-5. Disponível em: ⟨http://dx.doi.org/10.1007/978-3-642-20795-2_6⟩.

SAGGION, H. et al. Ontology-based information extraction for business intelligence. In: *6th International and 2nd Asian Semantic Web Conference*. [S.l.: s.n.], 2007.

SILVA, T. D. M. S. et al. Extraçao de informaçao para busca semântica na web baseada em ontologias. Florianópolis, SC, 2003.

SMITH, M. K.; WELTY, C.; MCGUINNESS, D. L. *OWLWeb ontology language guide*. 2004. Disponível em: ⟨http://www.w3.org/TR/owl-guide/⟩.

SMITH, M. K.; WELTY, C.; MCGUINNESS, D. L. *OWL 2 Web ontology language Documente Overview(Second Edition)*. 2012. Disponível em: ⟨http://www.w3.org/TR/owl2-overview/⟩.

SáNCHEZ, L. *Atribuição de Papéis Semânticos a Argumentos de Nominalizações: Um Método Semi-automático*. Dissertação (Dissertação de Mestrado) — Instituto Militar de Engenharia, Rio de Janeiro, RJ, 2007.

SODERLAND, S. et al. Crystal inducing a conceptual dictionary. In: *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1995. (IJCAI'95), p. 1314–1319. ISBN 1-55860-363-8. Disponível em: ⟨http://dl.acm.org/citation.cfm?id=1643031.1643069⟩.

SOWA, J. F. *Knowledge Representation: Logical, Philosophical and Computational Foundations*. Pacific Grove, CA, USA: Brooks/Cole Publishing Co., 2000. ISBN 0-534-94965-7.

STUDER, S. et al. Knowledge processes and ontologies. *IEEE Intelligent Systems*, v. 1, n. 16, p. 26–34, 2001.

TOMAZ, H. et al. An unsupervised method for ontology population from the web. In: PAVóN, J.; DUQUE-MéNDEZ, N.; FUENTES-FERNáNDEZ, R. (Ed.). *Advances in Artificial Intelligence – IBERAMIA 2012*. Springer Berlin Heidelberg, 2012, (Lecture Notes in Computer Science, v. 7637). p. 41–50. ISBN 978-3-642-34653-8. Disponível em: ⟨http://dx.doi.org/10.1007/978-3-642-34654-5_5⟩.

TURNEY, P. Mining the web for synonyms: Pmi-ir versus lsa on toefl. 2001.

USCHOLD, M. et al. Ontologies: Principles, methods and applications. *Knowledge Engineering Review*, v. 11, p. 93–136, 1996.

W3C. *Resource Description Framework (RDF)*. 2015. Disponível em: ⟨http://www.w3.org/RDF/⟩.

WIMALASURIYA, D. C. Ontology-based information extraction: An introduction and a survey of current approaches. *Journal of Information Science*, v. 36, n. 3, p. 306–323, mar. 2010. ISSN 0165-5515. Disponível em: ⟨http://jis.sagepub.com/cgi/doi/10.1177/0165551509360123⟩.

WIVES, L. K.; LOH, S. Tecnologias de descoberta de conhecimento em informações textuais; ênfase em agrupamento de informações. In: *OFICINA DE INTELIGÊNCIA ARTIFICIAL (OIA) III*. Pelotas , RS: EDUCAT, 1999. p. 28–48.

WU, F.; WELD, D. S. Autonomously semantifying wikipedia. In: *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*. New York, NY, USA: ACM, 2007. (CIKM '07), p. 41–50. ISBN 978-1-59593-803-9. Disponível em: ⟨http://doi.acm.org/10.1145/1321440.1321449⟩.

XAVIER, C. C. a.; LIMA, V. L. S. A Semi-Automatic Method for Domain Ontology Extraction from Portuguese Language Wikipedia's Categories. 2008.

YILDIZ, B.; MIKSCH, S. ontox - a method for ontology-driven information extraction. In: GERVASI, O.; GAVRILOVA, M. (Ed.). *Computational Science and Its Applications – ICCSA 2007*. Springer Berlin Heidelberg, 2007, (Lecture Notes in Computer Science, v. 4707). p. 660–673. ISBN 978-3-540-74482-5. Disponível em: ⟨http://dx.doi.org/10.1007/978-3-540-74484-9_57⟩.

ZAHRA, F. M.; CARVALHO, D. R.; MALUCELLI, A. Poronto : ferramenta para construção semiautomática de ontologias em português Poronto : herramienta para construcción semiautomática de ontologías en portugués. *journal of Health Informatics*, v. 5, n. 2, p. 52–59, 2013.