Um Método Não Supervisionado para o Povoamento de Ontologias a partir de Fontes Textuais na Web

Alternative Title: An Unsupervised Method for Ontology Population from Textual Sources on the Web

Fábio Lima
Universidade Federal da Bahia - UFBA
Rua Ademar de Barros - Ondina
Salvador, Bahia
fabio.lima@dcc.ufba.br

Hilário Oliveira
Universidade Federal de Pernambuco - UFPE
Av. Professor Morais Rego, 1235 - Cidade Universitária
Recife, Pernambuco
htao@cin.ufpe.br

Laís Salvador
Universidade Federal da Bahia - UFBA
Fraunhofer Project Center for Software and Systems Engineering
Salvador, Bahia
laisns@ufba.br

RESUMO

O crescimento na produção e disponibilização de informações não estruturadas na Web aumenta diariamente. Essa abundância de informações desestruturadas representa um grande desafio para a aquisição de conhecimento que seja processado por seres humanos e também por máquinas. Nesse sentido, ao longo dos anos 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 em português. Diante disso, o objetivo deste trabalho é propor e avaliar um método não supervisionado para o povoamento de ontologias utilizando a Web como grande fonte de informações, no contexto da língua Portuguesa. Os resultados obtidos com os experimentos realizados foram encorajadores e demonstraram que a abordagem proposta obteve uma taxa de precisão média de 67% na extração de instâncias de classes ontológicas.

Palavras-Chave

Ontologias, Povoamento de Ontologias, Extração de Informações

ABSTRACT

The increasing in the production and availability of unstructured information on the Web grows daily. This abundance of unstructured information is a great challenge for acquisition of structured knowledge. Many approaches have been proposed for extracting information from texts written in natural language. However, only a few studies have investigated the extraction of information from texts written in Portuguese. Thus, this work aims to propose and evaluate an unsupervised method for ontology population using the Web as a big source of information in the context of the Portuguese language. The results of the experiments are encouraging and demonstrated that the proposed approach reached a precision rate of 67% in the instances of ontological classes extraction.

SBSI 2015, May 26th-29th, 2015, Goiânia, Goiás, Brazil

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—Information filtering; I.2.4 [**Artificial Intelligence**]: Knowledge Representation Formalisms and Methods—Representation languages

General Terms

Measures, Method

Keywords

Ontologies, Ontology Population, Information Extraction

1. INTRODUÇÃO

Com o crescimento na produção de informações em formato digital disponibilizadas na Web 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 como política, esporte, economia, dentre outras.

Todas essas informações não são exploradas em todo o seu potencial devido à capacidade humana de processamento manual ser limitada. Surge assim, a necessidade da criação de sistemas computacionais que sejam capazes de analisar

automaticamente o enorme volume de informações disponíveis na Web. Contudo, a maioria das informações na Web está representada em formato textual, escrita em linguagem natural, sendo destinada à consulta, análise e interpretação realizadas por pessoas.

O armazenamento em formato textual não é a forma mais apropriada para o processamento computacional, uma vez que não é estruturado 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 essas informações sejam armazenadas em um formato estruturado permitindo que sejam interpretadas de maneira não ambígua. Com isso, é possível que tanto pessoas, quanto agentes computacionais possam analisar e extrair conhecimento útil. A transformação de construções sintáticas em semânticas é o principal objetivo da construção de bases de conhecimento.

A tarefa de construir Bases de Conhecimento 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 desestruturados (textos) em um formato estruturado. Essa tarefa apresenta diversos desafios e tem demandado interesse da comunidade de Inteligência Artificial ao longo dos anos, tendo sido propostos diversos trabalhos como o PANKOW [4], KnowItAll [6], Never-Ending Language Learning (NELL) [3], UMO-POW [14], OntoLP [17], Poronto [18], PSAPO [1], entre outros.

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 Ontologia(s). A base de conhecimento então instancia os elementos presentes na ontologia para um determinado domínio.

O termo *Ontologia* surgiu originalmente na Filosofia, como uma área que trata da natureza e organização dos seres [8]. No contexto da Ciência da Computação, ontologias podem ser definidas como especificações explícitas e formais de uma conceitualização compartilhada [13]. Uma ontologia representa um domínio através de seus conceitos (classes), propriedades, relações, axiomas, hierarquia de conceitos (taxonomia de conceitos) e hierarquia de relações (taxonomia de relações).

Uma das principais tarefas para a construção e manutenção de Bases de Conhecimento é o *Povoamento de Ontologias* [9]. Povoamento de Ontologias é o processo de inserção de novas instâncias de classes, propriedades e/ou relações em uma ontologia existente[12]. Além disso, essa tarefa permite relacionar o conhecimento descrito em linguagem natural com ontologias, auxiliando o processo de geração de conteúdo semântico [16]. Por fim, a ontologia povoada pode ser usada em diversas aplicações, como gerenciamento de conteúdo, recuperação de informação, raciocínio automático, dentre outras.

Atualmente, ainda existem poucos trabalhos que investigam a criação de abordagens automáticas capazes de realizarem a tarefa de povoamento de ontologias, a partir de textos escritos em português. Em sua grande maioria os trabalhos adotam a língua inglesa, principalmente devido à quantidade de informações disponíveis nesse idioma e pela quantidade de ferramentas que podem auxiliar no processo de povoamento.

Tentando suprir tal problema, o objetivo deste trabalho é propor e avaliar um método não supervisionado para o povoamento de ontologias utilizando a Web como grande corpus de trabalho. A abordagem proposta é capaz de extrair instâncias de classes ontológicas a partir de fontes textuais escritos em linguagem natural disponíveis na Web para o idioma português. O método proposto é guiado por uma ontologia de entrada que define quais conceitos devem ser povoados, e por um conjunto de padrões linguísticos usados para extrair e classificar um conjunto de termos candidatos a instâncias.

O restante deste artigo está organizado da seguinte forma: Na Seção 2, são apresentados os principais trabalhos relacionados com a tarefa de Povoamento de Ontologias, focando principalmente nos trabalhos que utilizam textos escritos em português como fonte de informações. Na Seção 3, são expostas as principais tecnologias e o método envolvido no seu desenvolvimento. Na Seção 4, os experimentos, resultados e discussões são ressaltados. Por fim, na Seção 5 são delineadas as conclusões e trabalhos futuros.

2. 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.

Em [11] um processo para extrair informações estruturadas a partir de páginas da web é proposto. 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 [17] apresentam um estudo sobre a extração de uma estrutura ontológica contendo relações de hiponímia (é uma) e localização a partir da Wikipédia em língua portuguesa. A abordagem visa capturar a estrutura de categorias de enciclopedias, 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 [5] 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 e também valida seu estudo no domínio de Turismo.

Baségio [2] 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. O autor, para a validação de sua proposta, conduz experimentos sobre o domínio de Turismo.

Tomaz et. al. [14] propõem um método não supervisionado para o povoamento de ontologias a partir de textos

escritos em inglês disponíveis na Web. O método extrai termos candidatos a instância utilizando a Web como corpus e posteriormente combina diferentes medidas estatísticas e semânticas para classificar os termos extraídos. Os experimentos realizados obtiveram bons resultados utilizando uma ontologia de topo com classes de diferentes domínios.

A abordagem proposta neste trabalho baseia-se no trabalho de Tomaz et. al. [14] que foi desenvolvido originalmente para o idioma inglês. No melhor do conhecimento dos autores nenhum trabalho anterior investigou a aplicação de um processo de povoamento de ontologias similiar ao proposto por Tomaz et. al.[14] para textos escritos em português. Essa lacuna motivou o desenvolvimento deste trabalho, no qual o método proposto se diferencia de outras propostas que lidam com textos em português por: (1) Utilizar a Web como grande corpus de trabalho para extração de termos candidatos a instâncias e (2) Avaliar diferentes medidas não supervisionadas para classificação de candidatos a instâncias.

3. ABORDAGEM PARA O POVOAMENTO DE ONTOLOGIAS PROPOSTA

O objetivo principal da abordagem proposta é extrair instâncias de classes ontológicas a partir de textos escritos em linguagem natural, em português, encontrados na Web. O processo de extração é guiado por uma ontologia de entrada que define os conceitos que devem ser povoados, e por um conjunto de padrões linguísticos independente de domínio adaptados de Hearst [7]. Esses padrões possuem dois papéis fundamentais: (1) Extratores, guiando o processo de extração de candidatos a instâncias; e (2) Discriminadores, sendo utilizados durante a classificação dos termos candidatos a instâncias. Os padrões de Hearst [7] foram originalmente utilizados para documentos em inglês. Neste trabalho, esses padrões foram traduzidos para o português.

A abordagem proposta é composta por 4 etapas principais: Coleta, Extração, Classificação e Povoamento. Uma visão geral das etapas e do fluxo de execução são apresentados na Figura 1 e detalhados nas seções a seguir.

3.1 Etapa 1 - Coleta dos Documentos

A primeira tarefa da abordagem proposta é selecionar uma classe c na ontologia de entrada. Após isso, utilizando um conjunto de padrões linguísticos independentes de domínio adaptados de Hearst [7], consultas são formuladas e aplicadas a um mecanismo de busca na Web para recuperação de um conjunto de documentos relevantes. São apresentados na Tabela 1, na 1ª coluna, os sete padrões linguísticos utilizados; e na 2ª coluna, exemplos de consultas formuladas para a classe Cidade.

Os padrões linguísticos listados na Tabela 1, em geral, são precedidos ou seguidos de exemplos de instâncias para a classe selecionada. O processo de formulação das consultas apresentadas na 2ª coluna é realizado da seguinte maneira: (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, indicando que a busca deve ser exata, ou seja, os documentos só devem ser recupados se possuírem exatamente a consulta usada.

Tabela 1: Padrões linguísticos com a respectiva consulta formulada

	Padrões Linguísticos	Consultas
1	*Classe(s)* tais como *Candidatos*	”cidades tais como”
2	tais *Classe(s)*) como *Candidatos*	”tais cidades como”
3	*Candidatos* ou outro(a) *Classe(s)*	”ou outras cidades”
4	*Candidatos* e outro(a) *Classe(s)*	e ”outras cidades”
5	*Classe(s)* incluindo *Candidatos*	”cidades incluindo”
6	*Classe(s)* especialmente *Candidatos*	”cidades especialmente”
7	*Candidato* é ART *Classe*	”é uma cidade”

Após a formulação das consultas, essas são aplicadas a um mecanismo de buscas na Web para recuperação de n documentos relevantes para cada consulta apresentada na Tabela 1. Neste trabalho, o Microsoft Bing [1] foi utilizado como mecanismo de busca.

Ao final desta etapa, um conjunto de documentos relevantes para a classe selecionada na ontologia de entrada está disponível para realização do processo de extração de candidatos a instâncias. Este processo é descrito na próxima seção.

3.2 Etapa 2 - Extração dos Candidatos a Instâncias

Após a etapa de coleta, cada um dos n documentos recuperados é processado. Para isso, a ferramenta CoGrOO [2], foi utilizada para realização das tarefas de tokenização, divisão de sentenças, etiquetagem das classes gramaticais, stemming e identificação de sintagmas nominais.

O objetivo é extrair sentenças relevantes, ou seja, são extraídas apenas sentenças que possuem o padrão linguístico que originou a consulta, no qual *Candidatos* representa um conjunto de sintagmas nominais extraídos como candidatos a instâncias. Por exemplo, 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. Para o caso do padrão *Candidato é ART Classe*, o elemento Candidato denota um único sintagma nominal. Cada candidato a instância extraído, mantém a lista de padrões linguísticos responsáveis por sua extração sem repetição. Essa informação será usada posteriormente na fase de classificação dos candidatos a instâncias.

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. [6]; McDowell e Cafarella [10]; Tomaz et al.[14] também utilizaram com sucesso sintagmas nominais como candidatos a instâncias na tarefa de Povoamento de Ontologias.

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

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

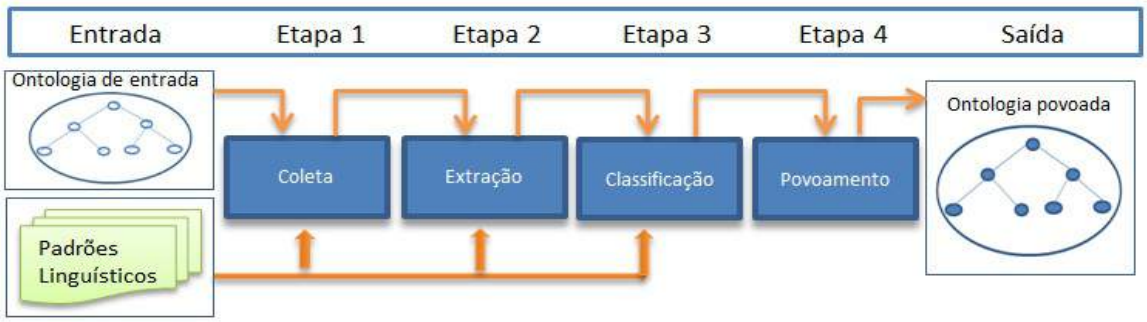

Figura 1: Visão geral da abordagem proposta

Com o objetivo de eliminar candidatos a instâncias inválidos ou repetidos, foram utilizadas alguns filtros:

a) Filtragem Candidatos a Instâncias repetidos: essa filtragem tem o objetivo de eliminar candidatos a instâncias que representam informações já existentes na ontologia de entrada. Por exemplo, dada a classe *Cidade* e um possível candidato a instância *Cidade*, ou variações, por exemplo, *cidades*, *a cidade*, *uma cidade*, *na cidade*, entre outros. Esses candidatos são eliminados, pois, representam a própria classe selecionada. Candidatos a instâncias que sejam instâncias já presentes na classe em povoamento ou que sejam instâncias de classes disjuntas também são eliminados. Para aumentar a cobertura desse filtro, o algoritmo de *stemming* é utilizado para encontrar o radical das palavras.

b) Filtragem de Candidatos a Instância sem valor semântico: o objetivo deste filtro é eliminar candidatos a instâncias que não possuem valor semântico. No método proposto, candidatos que não possuem substantivo são removidos. Por exemplo, candidatos a instâncias como: *aqueles*, *a seguir*, *o quê* e *embora*, entre outros, são eliminados 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 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, aplica-se o algoritmo de *stemming* para evitar variações do singular e do plural. 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 a coocorrência, a medida Pontuação de Informação Mútua (PMI) apresentada na Equação 1 e os padrões linguísticos listados na Tabela 1 são utilizados. Maiores detalhes sobre a medida de PMI são apresentados na próxima seção. O candidato c com maior valor de coocorrência é inserido na lista de candidatos a instâncias, além disso, a lista de padrões linguísticos dos candidatos removidos são adicionadas na lista de padrões linguístico de c.

Ao final desta etapa, após o processamento de todos os documentos coletados, uma lista formada pelos candidatos a instâncias extraídos para a classe selecionada e os padrões linguísticos que o extraíram, é produzida. É ilustrado na Tabela 2 exemplos de candidatos a instâncias extraídos para a classe *Cidade* e os respectivos padrões que o extraíram.

Tabela 2: Lista de candidatos a instância com padrões linguísticos que o extraíram

Candidato a instância	Padrões linguísticos que o extraíram
Alexandria	*Candidatos* ou outro(s) *Classe(s)* *Candidato* é ART *Classe* *Classe(s)* incluindo *Candidatos*
Teresópolis	*Classe(s)* tais como *Candidatos* *Classe(s)* incluindo *Candidatos*
Viçosa	*Candidato* é ART *Classe*
Belo Horizonte	*Candidatos* ou outro(s) *Classe(s)* *Classe(s)* tais como *Candidatos* *Candidato* é ART *Classe Candidatos* e outro(s) *Classe(s)*

3.3 Etapa 3 - Classificação dos Candidatos a Instâncias

Após a extração dos candidatos a instâncias, é necessário avaliar a confiabilidade de cada candidato classificado. Esta etapa tem por objetivo avaliar cada candidato identificado pela etapa Extração dos Candidatos a Instâncias e atribuir um grau de confiança a cada candidato a instância extraído.

Para isso, neste trabalho foram avaliadas três variações tradicionais da medida de Pontuação de Informação Mútua, do inglês *Pointwise Mutual Information* (PMI). O PMI é uma medida estatística cujo objetivo é mensurar o grau de coocorrência entre dois termos. Em geral, as medidas estatísticas sofrem com o problema de esparsidade dos dados [4], ou seja, dependendo da fonte de informação utilizada, os dados disponíveis nem sempre apresentam um indicativo de sua relevância, refletindo assim em uma baixa performance, principalmente quando utiliza-se palavras relativamente raras.

Para resolver tal problema, diversos autores [4], [6], [10], [14] demonstraram que a utilização de medidas estatísticas explorando a grande quantidade de dados disponíveis na Web apresenta-se como uma solução viável.

Diante disso, neste trabalho, a medida de PMI objetiva 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 são formuladas utilizando c, ci e cada padrão linguístico $p \in \mathcal{P}$.

Tabela 3: Lista de candidatos a instância com padrões linguísticos que o extraíram

(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

Posteriormente essas consultas são aplicadas a algum mecanismo de busca na Web com o objetivo de obter a quantidade de ocorrências de cada consulta executada. São ilustradas na Tabela 3 as consultas formuladas para o cálculo do PMI utilizando a classe Cidade, o candidato a instância Salvador e os padrões linguísticos apresentados na Tabela 1.

O processo de formulação das consultas apresentadas na Tabela 3 é realizado de maneira semelhante à etapa de Extração de Candidatos. A diferença está no uso do elemento Candidato(s) que é substituído pelo candidato a instância em avaliação.

Diversos trabalhos utilizam diferentes variações de fórmulas aplicadas no cálculo do PMI [15], [4], [6], [10], [14]. Em particular, McDowell e Cafarella [10] e Tomaz et al. [14], apresentam um estudo realizado para comparar diferentes variações para o cálculo do PMI aplicados no Povoamento de Ontologias na Web para textos escritos em inglês. Seguindo a mesma ideia, um experimento foi realizado com o objetivo de avaliar as 3 variações do PMI previamente investigadas por [10] e [14].

Nas equações 1, 2 e 3 são apresentadas as três variações da medida de PMI analisadas neste trabalho. Nestas equações, *hits(ci, c, p)* representa o número de ocorrências retornadas pelo mecanismo de busca na Web para a consulta formada por um candidato a instância *ci*, uma classe *c*, usando o padrão linguístico *p*. Enquanto isso, *hits(ci)* e *hits(c)* representam o total de ocorrência do candidato a instância *ci* e da classe *c* isoladamente.

PMI Strength: Essa variação do PMI é calculada pelo somatório de todas as coocorrência retornadas pela consulta *hits(c, ci, p)*. Para isso, consultas são formuladas para o par *(c, ci)* e cada padrão linguístico $p \in \mathcal{P}$ listados na Tabela 1. Na Equação 1 é apresentado como é realizado o cálculo desta variação de PMI.

$$PMI_{Strength} = \sum_{p \in P} hits(c, ci, p) \quad (1)$$

PMI Str-INorm-Thresh: Segundo McDowell e Cafarella [10], a variação *PMI Strength* pode ser tendenciosa para instâncias muito frequentes. Outras instâncias para a classe selecionada podem não ser tão frequentes. Tentando solucionar esse problema, algumas variações utilizam como fator de normalização o total de ocorrências retornado pelo *hits(ci)*. Contudo, mesmo usando esse fator de normalização em casos onde o candidato à instância é raro, o problema da tendenciosidade persiste. Diante disso, o fator de normalização precisa ter um valor mínimo definido. Nesta versão, o fator de normalização utilizado é resultado do maior valor entre *hits(ci)* do candidato a instância em classificação e o 25^{o} percentil da distribuição de *hits(ci)* de todos os candidatos a instância para a classe *c* selecionada. Na Equação 2 é apresentado como é realizado o cálculo desta variação de PMI.

$$PMI_{Str-INorm} = \frac{\sum_{p \in P} hits(c, ci, p)}{max(hits(ci), Percen_{25})} \quad (2)$$

PMI Str-ICNorm-Thresh: Seguindo a mesma ideia de normalização do *PMI Str-INorm-Thresh*, esta variação utiliza, como fator de normalização, ambos, o candidato a instância *ci* e a classe *c*. Na Equação 3 é apresentada como é calculada esta variação.

$$PMI_{Str-ICNorm} = \frac{\sum_{p \in P} hits(c, ci, p)}{max(hits(ci), Percen_{25}) * hits(c)} \quad (3)$$

Além das 3 variações de PMI apresentadas, neste trabalho, a heurística de Número de Padrões que Extraíram (NPE) proposta por [14] foi analisada. A heurística de NPE 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. Baseado 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 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 = 1.

Ao final desta etapa todos os candidatos a instâncias possuem um valor de confiança atribuído por cada uma das 3 variações de PMI e também pela heurística de NPE. Na seção 4 são apresentados e discutidos os experimentos realizados para investigar qual das 4 medidas de classificação apresenta a melhor performance na tarefa de classificar os candidatos a instâncias.

3.4 Etapa 4 - Povoamento

Nesta última etapa, o objetivo é decidir quais desses candidatos a instâncias serão promovidas a instâncias da classe selecionada, sendo assim utilizadas para povoar a ontologia de entrada. Diante disso, a escolha de um limiar é muito importante principalmente levando em consideração que essa escolha impacta diretamente na taxa de acerto e na 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. [14], a medida de 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.[14] 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.

4. EXPERIMENTO E DISCUSSÕES

Para demonstrar a eficácia da abordagem proposta, um experimento comparando as 3 variações da medida de PMI e também a heurística de NPE foi realizado. Esse experimento vislumbra avaliar cada uma das etapas do método proposto, focando principalmente na análise comparativa entre as 4 medidas apresentadas aplicadas na tarefa de classificação dos candidatos a instâncias das classes ontológicas.

4.1 Configurações do Experimento

O corpus utilizado neste experimento é formado por um conjunto de *snippets* coletados utilizando o mecanismo de busca Bing. *Snippets* são textos simples que em geral possuem as palavras chave que formam a consulta aplicada, apresentando uma prévia da informação contida nos documentos recuperados. Essas prévias, mesmo possuindo um tamanho reduzido, são informativos o suficiente para extrair conhecimento relacionado com a consulta aplicada sem a necessidade de processar o documento inteiro [14].

Para este experimento 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*.

A coleta de documentos foi efetuada de forma automática e buscava 1000 documentos/dia. No período de 19/09/2014 a 20/11/2014 foram recuperados para as 15 classes selecionadas um total de 62.409 documentos relevantes.

A medida utilizada para avaliar os experimentos realizados foi a medida de *Precisão*. No contexto deste trabalho, a precisão pode ser definida pela razão entre a quantidade de instâncias corretas extraídas pelo total de instâncias recuperadas. A Equação 4 foi usada para calcular a precisão dos experimentos.

$$Precisao(Top_N) = \frac{total\_de\_instancias\_corretas}{N} \quad (4)$$

Nesta equação, N é a quantidade de termos candidatos promovidos a instância.

A avaliação da precisão neste experimento foi realizada variando diferentes limiares. O limiar definiu o valor dos N candidatos a instâncias (Top N) ordenados com base na medida de classificação utilizada. Neste experimento 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 é necessário analisar cada termo candidato extraído e verificar se ela é 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.

Para promover uma comparação justa entre as medidas de classificação avaliadas, as etapas de coleta e extração de candidatos a instâncias foram executadas uma única vez. Em seguida, os conjuntos de candidatos a instâncias extraídos para cada uma das 15 classes selecionadas foram passados como parâmetros de entrada para a etapa de classificação que utilizou as 3 medidas de PMI e a heurística de NPE individualmente para classificar os candidatos a instâncias.

4.2 Resultados e Discussões

Os gráficos apresentados nas Figuras 2, 3, 4 e 5 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 de NPE.

Analisando os resultados apresentados nas Figuras 2, 3, 4 e 5 é possível observar que a abordagem proposta é capaz de extrair 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 perca de precisão. Observa-se também que existe uma variação de precisão entre as classes selecionadas, isso ocorre devido a fatores como complexidade do domínio e coocorrência de instâncias da classe selecionada 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 a outras classes de domínio mais específico como Inseto, Passáro e Sintoma.

Um fator importante a ser ressaltado é a falta de confiabilidade das informações presentes na Web. A maioria dessas informações é escrita por pessoas que em geral não são especialistas do domínio abordado. Quanto mais complexo o domínio analisado, maior a probabilidade de encontrar informações erradas.

Considerando o limiar Top 10, como apresentado na Figura 2, a heurística de NPE apresentou uma maior taxa de precisão em 9 das 15 classes analisadas obtendo uma 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 e 4, o mesmo comportamento foi observado. A heurística de NPE continuou melhor com 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 5, 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.

Com base nos experimentos executados, conclui-se que a heurística de NPE obteve melhores resultados do que as 3 variações da medida de 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. Tal resultado indica que os fatores de normalização usados pelo PMI Str-INorm-Thresh e PMI

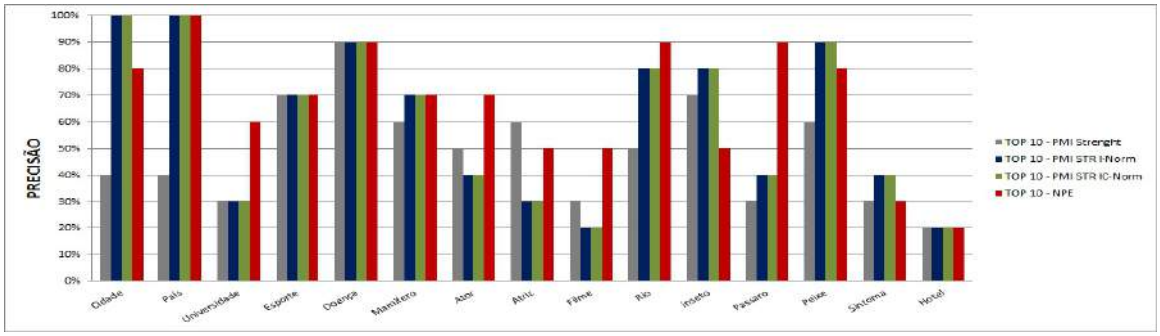

Figura 2: Resultados do Experimento no limiar Top 10.

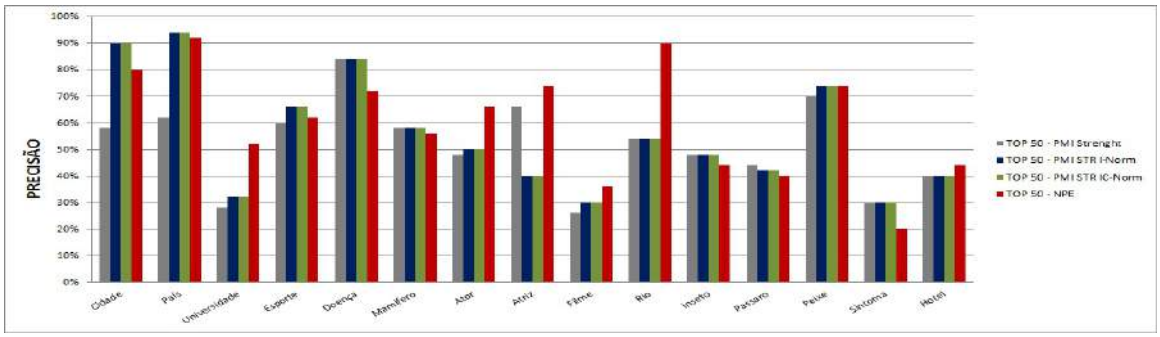

Figura 3: Resultados do Experimento no limiar Top 50.

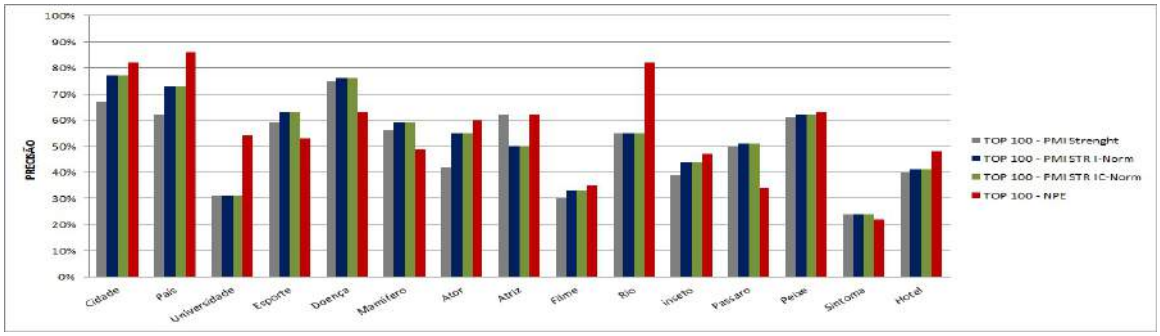

Figura 4: Resultados do Experimento no limiar Top 100.

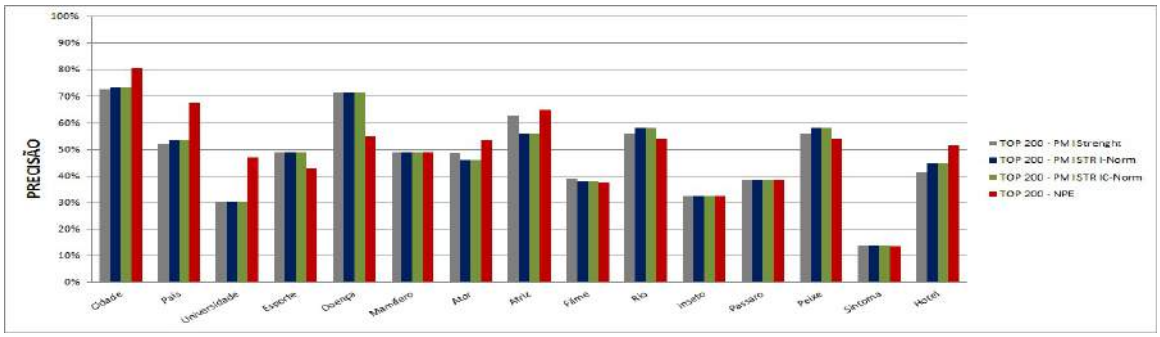

Figura 5: Resultados do Experimento no limiar Top 200.

Str-ICNorm-Thresh apresentaram um impacto positivo nos resultados.

5. CONCLUSÕES

Este trabalho apresentou uma abordagem não supervisionada para o povoamento de ontologias a partir de textos escritos em linguagem natural em português utilizando a Web como grande fonte de informações. A relevância desse problema é decorrente da abundância de informações desestruturadas presentes na Web e da aquisição de conhecimento estruturado a partir de textos desestruturados para criação de bases de conhecimento. Além disso, acrescenta-se os poucos estudos que investigam a exploração de textos escritos em português como fonte de informação.

Após a análise dos resultados é possível concluir que a abordagem para o povoamento de ontologias na Web foi capaz de extrair uma grande quantidade de instâncias corretas para a maioria das quinze classes selecionadas. Esses resultados são encorajadores para que a abordagem proposta possa ser evoluída para atingir melhores resultados e também expandida para contemplar outros aspectos na ontologia de entrada como relações entre classes e propriedades.

Como trabalhos futuros sugere-se: (1) Propor uma medida combinada que integre diferentes medidas e heurísticas para a classificação dos candidatos a instâncias extraídos; (2) Integrar um módulo para identificação de sinônimos para a filtragem de candidatos a instâncias que possuem o mesmo valor semântico; e (3) Explorar recursos semânticos disponíveis para o idioma português, por exemplo, o OpenWordnet-PT [3].

6. REFERÊNCIAS

[1] C. G. d. F. Alves. Um Processo Independente de Domínio para o Povoamento Automático de Ontologias a partir de Fontes Textuais. 2013.

[2] T. L. Baségio. Uma Abordagem Semi-automática para Identificação de Estruturas Ontológicas a partir de Textos na Língua Portuguesa do Brasil. pages 1–124, 2007.

[3] A. Carlson, J. Betteridge, B. Kisiel, B. Settles, E. R. Hruschka, and T. M. Mitchell. Toward an architecture for never-ending language learning. In *In AAAI*, 2010.

[4] P. Cimiano, S. Handschuh, and S. Staab. Towards the self-annotating web. *Proceedings of the 13th conference on World Wide Web - WWW '04*, page 462, 2004.

[5] L. Drumond and R. Girardi. Extracting ontology concept hierarchies from text using markov logic. In *Proceedings of the 2010 ACM Symposium on Applied Computing*, SAC '10, pages 1354–1358, New York, NY, USA, 2010. ACM.

[6] O. Etzioni, S. Kok, S. Soderland, M. Cafarella, A. m. Popescu, D. S. Weld, D. Downey, T. Shaked, and A. Yates. Web-Scale Information Extraction in KnowItAll (Preliminary Results). pages 100–110, 2004.

[7] M. A. Hearst. Automatic Acquisition of Hyponyms from Large Text Corpora. pages 23–28, 1992.

[8] A. Maedche and S. Staab. Ontology learning for the semantic web. *Intelligent Systems, IEEE*, 16(2):72–79, Mar 2001.

[9] D. Maynard, Y. Li, and W. Peters. 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*, pages 107–127, Amsterdam, The Netherlands, The Netherlands, 2008. IOS Press.

[10] L. K. McDowell and M. Cafarella. Ontology-driven, unsupervised instance population. *Web Semant.*, 6(3):218–236, Sept. 2008.

[11] E. N. Motta. Preenchimento Semi-automático de Ontologis de Domínio a Partir de Textos em Língua Portuguesa. Master's thesis.

[12] G. Petasis, V. Karkaletsis, G. Paliouras, A. Krithara, and E. Zavitsanos. Ontology population and enrichment: State of the art. In G. Paliouras, C. Spyropoulos, and G. Tsatsaronis, editors, *Knowledge-Driven Multimedia Information Extraction and Ontology Evolution*, volume 6050 of *Lecture Notes in Computer Science*, pages 134–166. Springer Berlin Heidelberg, 2011.

[13] S. Studer, S. Rudi, H.-P. Schnurr, and Y. SURE. Knowledge processes and ontologies. *IEEE Intelligent Systems*, 1(16):26–34, 2001.

[14] H. Tomaz, R. Lima, J. Emanoel, and F. Freitas. An unsupervised method for ontology population from the web. In J. Pavón, N. Duque-Méndez, and R. Fuentes-Fernández, editors, *Advances in Artificial Intelligence – IBERAMIA 2012*, volume 7637 of *Lecture Notes in Computer Science*, pages 41–50. Springer Berlin Heidelberg, 2012.

[15] P. Turney. Mining the web for synonyms: Pmi-ir versus lsa on toefl. 2001.

[16] D. C. Wimalasuriya. Ontology-based information extraction: An introduction and a survey of current approaches. *Journal of Information Science*, 36(3):306–323, Mar. 2010.

[17] C. C. a. Xavier and V. L. S. d. Lima. A Semi-Automatic Method for Domain Ontology Extraction from Portuguese Language Wikipedia's Categories.

[18] F. M. Zahra, D. R. Carvalho, and A. Malucelli. Poronto : ferramenta para construção semiautomática de ontologias em português Poronto : herramienta para construcción semiautomática de ontologías en portugués. *journal of Health Informatics*, 5(2):52–59, 2013.

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