UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
CURSO DE CIÊNCIA DA COMPUTAÇÃO

EDUARDO DELAZERI FERREIRA

Criação de Ontologias Linguísticas Automáticas a partir de Texto

Monografia apresentada como requisito parcial para a obtenção do grau de Bacharel em Ciência da Computação

Orientador: Prof. Dr. Aline Villavicencio
Co-orientador: Dr. Rodrigo Souza Wilkens

Porto Alegre
2016

RESUMO

Ontologias linguísticas são recursos importantes na área de Processamento de Linguagem Natural, sendo a WordNet um exemplo de construção manual dessas ontologias. Infelizmente a criação e expansão dessas ontologias é difícil devido a necessidade de alta supervisão de especialistas. Para simplificar o processo de manutenção de ontologias foram criadas ontologias automáticas através de extração de padrões de texto. O método comumente usado para a criação automática de ontologias foi realizado em cima de corpus, buscando pares de palavras nestes textos para então descobrir padrões de ocorrência destas palavras. Esse é aplicável para aquisição de ontologia para qualquer domínio e quaisquer relações, dependendo apenas do domínio do corpus e das relações dos pares de entrada. Neste trabalho evidenciamos o comportamento de padrões em textos livres, avaliando um dos trabalho originários da área, os padrões de Hearst e o método criado nesse trabalho. Também avaliamos os padrões, os pares de palavras e seus comportamentos em textos livres. Os resultados apresentados demonstram que os padrões não se comportam, em textos livres, como esperado. Os resultados indicam que os padrões, mesmo os de Hearst, conhecidos por terem uma alta precisão, não apresentam uma unicidade de relações (são encontrados em mais de uma relação).

Palavras-chave: Ontologias Linguísticas. Extração de Informações. Aquisição de padrões. Processamento de Linguagem Natural. Extração automática de ontologia.

Automatic Ontology Creation from text

ABSTRACT

Linguistic ontologies are important resources to the field of Natural Language Processing, WordNet is an example of such an ontology which was manually built. Unfortunately, the expansion of these ontologies is hard due to the need of a high level of specialized supervision. In order to simplify the process of maintenance of these ontologies, we developed an automatic process of building them through the extraction of text patterns.
The proposed method for the automatic generation of ontologies was built using corpus, through the search of word pairs in texts in order to uncover patterns and their occurrences. Our method is suitable to the acquisition of ontologies for all domains and relations, and depends only upon the domain of the corpus and the input pair relationships.
Our work shows how the above mentioned patterns happen for free texts, through an evaluation of both a previous work present in the literature - the Hearst patterns - and our own proposed method. Our work also evaluates the patterns, word pairs and their behavior in free texts.
Our results show that resulting patterns do not behave in free texts as one would expect, in that even the Hearst patterns, known for their high level of precision, do not exhibit unicity of meaning.

Palavras-chave: Automatic Ontology Extraction. Pattern Extraction. Natural Language Processing.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

PLN	Processamento de Linguagem Natural
PMI	*Pointwise Mutual Information*
IMP	Informação Mútua Pontual

SUMÁRIO

1 INTRODUÇÃO

Uma ontologia é uma especificação formal explicita e compartilhada que representa um domínio de interesse (BUITELAAR; CIMIANO; MAGNINI, 2005), como por exemplo a WordNet[1](MILLER, 1995). O caráter formal de uma ontologia se deve ao fato dela ser *machine readable*, ou seja, possuir regras restritas para a sua criação, não podendo ser um texto livre como a linguagem natural. Esta, assim como a linguagem natural, é compartilhada, pois deve ser aceita por um grupo ou comunidade (BUITELAAR; CIMIANO; MAGNINI, 2005).

As ontologias são uma estruturação importante do conhecimento, a qual permite que o entendimento de um determinado domínio seja formalizado, facilitando a análise e o entendimento sobre o mesmo. Antes de ontologias serem empregadas na computação, o conceito de ontologias já havia sido definido, tendo sua origem da filosofia, onde uma ontologia tinha como objetivo definir os termos usados para descrever e representar uma área do conhecimento.

Na computação, o uso de ontologias ocorre em diversos tipos de aplicações em diferentes áreas. Na área de Inteligência Artificial, ontologias podem ser usadas por exemplo para bases de conhecimento para sistemas especialistas (MATKAR; PARAB, 2011). Na área de representação de conhecimento são utilizadas em sistemas de conhecimento de senso comum (ERNEST, 1990)[2]. Na área de Processamento de Linguagem Natural ontologias, como por exemplo, a WordNet (MILLER, 1995), são utilizadas para representar o domínio linguístico e representar os léxicos da linguagem.

As abordagens para criação de ontologias podem ser dividas quanto a necessidade de intervenção humana. Assim, o estado da arte pode ser divido em três abordagens:

Manual necessita especialistas do domínio desejado, os quais manualmente identificam as classes e as relações da ontologia. A grande vantagem dessa abordagem é a sua qualidade da ontologia, como por exemplo, a WordNet (MILLER, 1995) que é ontologia lexical manualmente construída amplamente utilizada por sua qualidade e cobertura reconhecidos na comunidade de pesquisa. A grande desvantagem da abordagem manual é o tempo de criação, o alto custo para a sua geração e a dificuldade na sua manutenção.

Automática utiliza apenas métodos computacionais sem interferência humana. É reali-

[1]https://wordnet.princeton.edu/
[2]https://www.commonsensemedia.org/

zada normalmente através de extração de informação com base em grandes conjuntos de textos. É estabelecida uma metodologia que pode ser aplicada para qualquer domínio de texto, desa forma podendo produzir ontologias de domínios distintos. A grande vantagem desta técnica é a velocidade para se obter um recurso para diversas áreas diferentes através de uma mesma metodologia. Alguns exemplos desta metodologia automática são os trabalhos propostos por Brin (1998) e Caraballo (1999). A maior desvantagem é que os resultados não são tão precisos quanto os de ontologias manualmente criadas. Um exemplo desta técnica é apresentado na Seção 2.3

Semi-automática ocorre quando existe qualquer forma de combinação entre as duas abordagens citadas anteriormente. Essa abordagem combina a dinamicidade das técnicas automáticas com a qualidade das técnicas manuais. Um exemplo desta combinação de abordagens pode ser encontrado na Seção 2.3, onde os padrões são criados manualmente e a população da ontologia através destes padrões é feita automaticamente. O resultado destas abordagens é mais preciso que os automáticos, mas sua construção acaba sendo mais demorada.

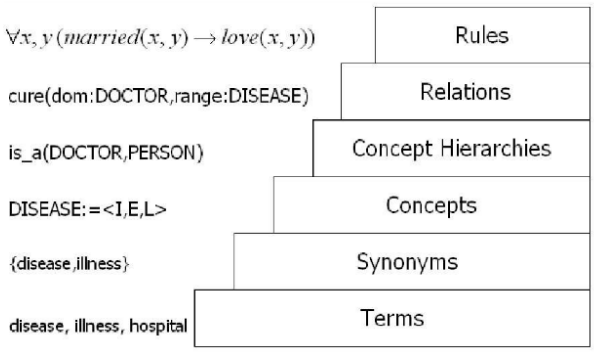
Neste trabalho damos enfase para as ontologias linguísticas, que tem por objetivo representar a linguagem e suas relações[3]. A abordagem de criação de ontologias deste trabalho é automática, visando reduzir a quantidade de trabalho dos métodos manuais e reduzir o tempo de espera na criação de ontologias (AUGER; BARRIÈRE, 2008).

A metodologia básica para a criação de ontologias linguísticas consiste dos seguintes passos (AUGER; BARRIÈRE, 2008):

1. estabelecer as relações a serem adquiridas. As relações se dão entre palavras e estabelecem um tipo específico de relação linguística. Por exemplo, a relação de sinônimo indica que uma palavra possui o mesmo significado que outra, enquanto que a relação de antônimo indica que duas palavras possuem significados contrários.
2. Extrair pares que possuam a relação desejada. Por exemplo, *alto* e *baixo* para o caso de antônimos.
3. Procurar as ocorrências dos pares em textos do domínio desejado
4. Extrair os padrões destas ocorrências. Esses padrões são construções léxico-sintáticas, nas quais ocorre a relação desejada. Por exemplo, o padrão *"...pode ser* $palavra_1$

[3] Ontologias linguísticas são introduzidas em profundidade na Seção 2.1

Figura 1.1: Divisão de criação de ontologia através da complexidade da tarefa



Fonte: (BUITELAAR; CIMIANO; MAGNINI, 2005)

ou $palavra_2$*"* pode indicar a existência de antônimos, como em *"O preço pode ser alto ou baixo"*. Esses padrões devem ser recorrentes da linguagem e apresentar uma única relação entre seus componentes.

5. Procurar os padrões adquiridos nos textos do domínio
6. Popular a ontologia com os pares encontrados.

Existe uma grande variedade de métodos criados e com objetivos diferentes para o aprendizado de ontologias a partir de texto. Com o objetivo de permitir uma comparação entre estes métodos, Buitelaar, Cimiano e Magnini (2005) estabeleceram uma classificação composta de camadas que definem diferenças de complexidade. A representação de aumento da complexidade de representação é ilustrada na Figura 1.1, através do exemplo *disease*, na qual são considerados 6 níveis:

Termos São compreensões linguísticas de conceitos de domínios específicos;

Sinônimos e variantes linguísticas Foco em aquisição de variantes de termos semânticos em e entre línguas;

Conceitos Extração do aspecto intensional, instâncias do seu aspecto extensional e um conjunto de termos para este contexto;

Taxonomias Criação de taxonomias a partir de 3 paradigmas: utilização de padrões léxico-sintáticos, clusterização hierárquica a partir da hipótese distribucional de Harris e a última baseada em recuperação de informações para definir a hierarquia de termos;

Relações Definir novos relacionamentos entre conceitos conhecidos, indo além das relações de sinonímia antes citadas; e

Regras Extrair regras para construção de ontologias a partir de regras de vinculação lexical.

Neste trabalho apresentamos a construção automática de ontologias linguísticas, com foco no nível de *taxonomias* de Buitelaar, Cimiano e Magnini (2005). As *taxonomias*, neste trabalho, são baseadas em relações hierárquicas entre palavras. Assim, este trabalho visa gerar ontologias linguísticas de forma dinâmica para diversas aplicações na área de Processamento de Linguagem Natural (PLN). Essas ontologias servem como recursos para serem explorados na resolução de diversas tarefas de PLN, como análise de sentimentos (BALDONI et al., 2012), mineração de textos (SPASIC et al., 2005) e sistema de perguntas e respostas (LOPEZ et al., 2007).

Para a realização do objetivo deste trabalho avaliamos o processo automático de criação de ontologias. Para tanto investigamos as seguintes hipóteses do processo, visto que os algoritmos são baseados em ideias intuitivas, mas com poucos esforços de validação:

H1 poucas sementes são capazes de gerar padrões representativos da linguagem

H2 padrões da linguagem representam uma relação linguística específica

Este trabalho está organizado nas seguintes seção:

Seção 2 apresenta uma revisão dos trabalhos relacionados.

Seção 3 apresenta os materiais que foram utilizados para a criação da metodologia e dos experimentos apresentados. Também é apresentado uma explicação mais detalhada da metodologia criada.

Seção 4 apresenta as avaliações realizadas para analisar as hipóteses levantadas e o resultado final do trabalho

Seção 5 apresenta a conclusão do trabalho realizado, fazendo uma revisão dos pontos levantados e uma avaliação geral do trabalho e de suas contribuições

2 TRABALHOS RELACIONADOS

Este trabalho se baseia fortemente em dois conceitos de criação de ontologias: *relações entre conceitos* e *padrões linguísticos*. As relações deste trabalho se restringem a relações semânticas entre palavras e as seguintes relações são de interesse neste trabalho: hiperônimos, merônimos, sinônimos e antônimos, que serão mais detalhados na Seção 2.1 Os padrões deste trabalho se referem a padrões léxico-sintáticos recorrentes da linguagem que representem as relações acima citadas. Esses padrões não são fixos. O que definem esses padrões são que a existência deles da uma intuição geral a respeito da sentença onde se encontram, mesmo que os elementos principais da sentença não sejam conhecidos. Por exemplo, na sentença *"theremim é um tipo de instrumento"*, é possível deduzir que *theremim* é um dos muitos tipos de instrumentos, mesmo sem saber o que é um theremim. Dessa forma, o padrão *"*$Palavra_1$ *é um tipo de* $Palavra_2$*"* pode ser classificado como um padrão que indica a presença da relação de hiperônimos.

Neste capítulo contextualizamos diferentes ontologias linguísticas (Seção 2.1) a fim de identificar os diferentes tipos de relações representadas. Com foco em criação automática de ontologias, apresentamos abordagens para extração de relações (Seção 2.2) e padrões (Seção 2.4).

2.1 Ontologias linguísticas

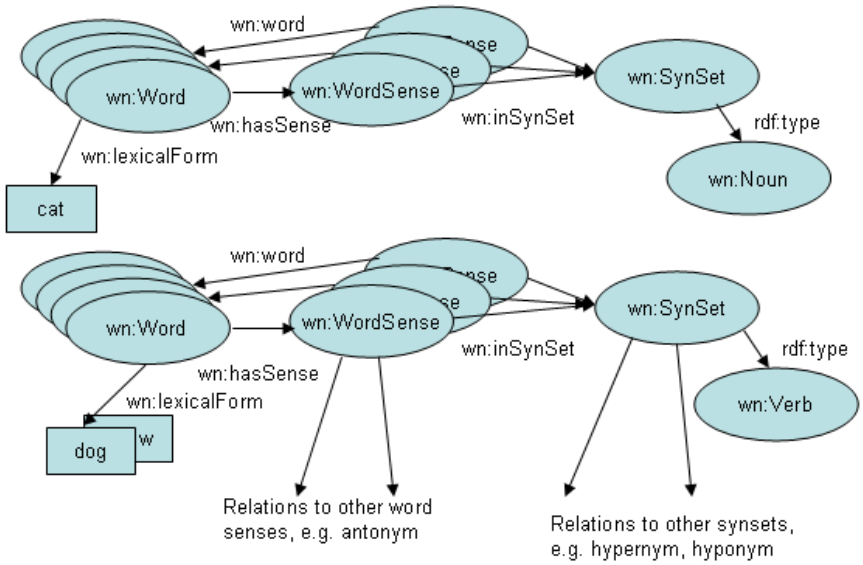
No contexto deste trabalho consideramos ontologias linguísticas como ontologias que representam a linguagem e sua estruturação. Nessas ontologias linguísticas os termos de interesse são os léxicos da língua e as relações da ontologia se dão sobre léxicos. Dessa forma, a ontologia apresenta uma série de relações entre os léxicos, essas relações podendo ser, por exemplo, antonímia, sinonímia, hiperonímia e meronímia.

Existem diversas ontologias linguísticas, as quais, em sua maioria, seguem o padrão estabelecido pela WordNet (MILLER, 1995), e posteriormente definido pela W3C[1] mostrado na Figura 2.1. Nesse padrão cada palavra possui uma lista de sentidos e cada um desses sentidos está associado com um *synset*. Portando os *synsets* representam um sentido global, e cada palavra se relaciona com os *synsets* através dos diferentes sentidos que ela possue. Essa abrangência de sentidos para uma única palavra é conhecida como polissemia. Por exemplo, *manga* pode ser ou uma fruta ou uma parte da roupa, e

[1] https://www.w3.org

cabo pode ser uma parte da vassoura ou uma patente militar. Portanto, um *synset* representa um conceito da língua e esse conceito contêm os sentidos das palavras usadas para expressá-lo, e uma palavra pode estar presente em diversos *synsets* através de seus diversos sentidos. Por exemplo, *banco* possui 3 sentidos facilmente identificáveis: (1) assento, (2) instituição financeira e (3) formação geológica de rios. Cada um desses diferentes sentidos colocará a palavra *banco* no respectivo *synset*.

Figura 2.1: Padrão W3C para ontologias linguísticas



https://www.w3.org/2001/sw/BestPractices/WNET/wordnet-sw-20040713.html

As ontologias linguísticas apresentam diversas relações entre conceitos, apresentadas por Miller et al. (1990). Entre elas as mais utilizadas são:

Sinonímia possui duas definições, a primeira, mais restrita, diz respeito a sinônimos verdadeiros, onde duas palavras são consideradas sinônimos caso a substituição de uma pela outra mantenha o sentido original da sentença. Por esta definição ser bastante restrita, palavras com sinônimos verdadeiros são bastante raras. A outra definição se da em relação ao contexto, ou seja, duas palavras são consideradas sinônimos se uma pode ser trocada pela outra sem perder o sentido em um contexto definido. Por exemplo, *subir* e *alçar* são sinônimos em certos contextos. A segunda definição é a utilizada pela WordNet para definir os seus *synsets* e também é a utilizada neste trabalho.

Antonímia é uma relação que descreve quando palavras cujos sentidos são opostos. Por exemplo, subir e descer.

Hiperonímia é uma relação que descreve quando duas palavras estão no mesmo campo semântico, mas possuem uma hierarquia, ou seja, uma delas tem um sentido mais abrangente. Por exemplo, *planta* é um hiperônimo de *árvore* e *árvore* é um hiperônimo de *carvalho*.

Hiponímia é uma relação simétrica a hiperonímia, portanto hipônimos são palavras mais específicas de um conceito mais abrangente. Por exemplo, se *árvore* é um hiperônimo de *carvalho* (como mostrado na relação acima), *carvalho* é um hipônimo de *árvore*.

Meronímia é uma relação entre duas palavras, quando estas possuem uma relação de parte de, onde um substantivo é uma parte que compõe o outro substantivo. Por exemplo, o *galho* é uma parte da *árvore*.

Existem diversas ontologias linguísticas criadas. Neste trabalho nós destacamos a WordNet (MILLER, 1995), por ser uma base importante para o surgimento de diversas outras ontologias baseadas na sua metodologia, o BabelNet (NAVIGLI; PONZETTO, 2010), por ter gerado uma grande ontologia para diversas línguas, e OntoPT (OLIVEIRA; GOMES, 2014), por ser uma importante ontologia focada para a língua portuguesa.

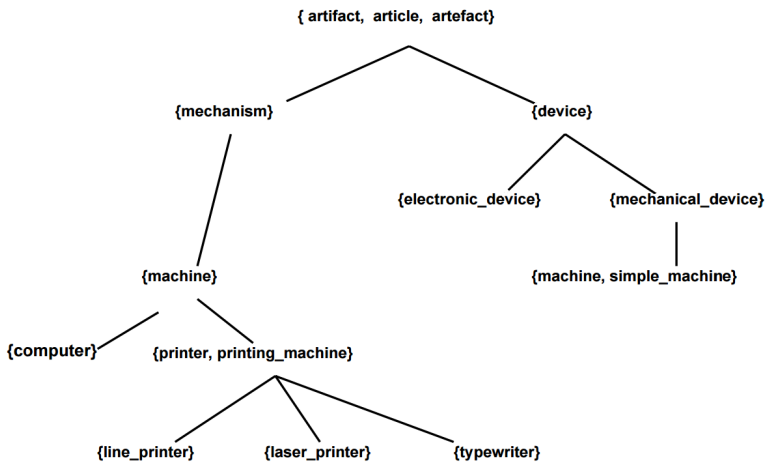
A WordNet (MILLER, 1995) é uma ontologia lexical, como explicado no início deste capítulo, que foi manualmente criada. Na WordNet as palavras são agrupadas através de *synsets*, onde cada *synset* representa um conceito da língua. Uma exemplificação para melhor compreender como os *synsets* interagem com os hiperônimos, formando uma hierarquia pode ser vista na Figura 2.2, onde os níveis de *synsets* na árvore demonstram o quão genérico é o conceito que este representa. A relação de hiperonímia é dada de cima pra baixo, de forma que o nó pai na árvore é um hiperônimo de seus nós filhos.

Na Tabela 2.1 mostramos algumas medidas do tamanho da WordNet.

BabelNet (NAVIGLI; PONZETTO, 2010) é uma ontologia linguística criada para mais de 200 idiomas, incluindo o português e o inglês. Ela foi criada conectando a Wikipedia com a WordNet através de mapeamento automático entre línguas e, em caso de línguas com poucos recursos, tradução automática de textos. O resultado desta abordagem é uma ontologia que provê conceitos e relações entre os léxicos de diversas línguas. Na Tabela 2.1 mostramos algumas medidas do tamanho da BabelNet.

Outro recurso voltado para criação de ontologias lexicais é a Onto.PT (OLIVEIRA; GOMES, 2014). Esse é um recurso baseado na WordNet para a criação de ontologias em português, que utiliza diversos recursos para a sua criação, como thesauros, dicionários digitais, corpora e enciclopédias. Na Tabela 2.1 mostramos algumas medidas do tamanho

Figura 2.2: Hierarquia dos synsets presentes na WordNet



Fone: (HEARST, 1992)

do Onto.PT.

Na Tabela 2.1 mostramos a quantidade total de palavras presentes nestas ontologias, o total de *synsets* criados por estas ontologias, cada *synset* representando um conceito da linguagem e as relações apresentadas por estas ontologias.

Tabela 2.1: Tamanho das ontologias

	Palavras	***synsets***	**Relações**
WordNet3.0	155.000	117.000	285.000
BabelNet3.7	$764mi$	$14mi$	$380mi$
Onto.PT	160.000	109.000	175.000

2.2 Padrões para Extração de Relações

Padrões, como definidos no inicio deste capítulo, são estruturas linguísticas recorrentes em textos e que indicam alguma relação específica na sentença em que ocorrem. Dessa forma, um padrão famoso para a identificação de um hiperônimo é *"cachorro é um animal"*, onde o padrão *é um* indica que o substantivo anterior ao padrão é um hipônimo do substantivo posterior ao padrão, ou o substantivo posterior ao padrão é um hiperônimo do substantivo anterior ao padrão. Portanto, tais construções léxicas seriam construções recorrentes na língua.

Segundo Auger e Barrière (2008) a utilização de padrões para a extração de relações entre palavras foi iniciada por Hearst (1992), que deu o primeiro passo para a

aplicação de padrões para extração automática de relações em texto. Hearst observou que dois sintagmas nominais (*noun phrases* - NP) conectados por construções do tipo NP_x *such as* NP_y frequentemente resultam em NP_x sendo um hipônimo de NP_y. Desta forma, neste capítulo, apresentamos a abordagem original da utilização de padrões para aquisição de relações na Seção 2.3 e alternativas para tornar esta construção de padrões automática na Seção 2.4

2.3 Padrões manualmente construídos

Existem algumas abordagens que definem manualmente padrões para a extração de relações em sentenças. Dentre essas, destacamos a criada por Hearst (1992).

Os padrões de Hearst são padrões léxico-sintáticos criados manualmente que tem por objetivo detectar relações de hiperonímia entre palavras. Estes padrões atingem um alto nível de precisão, mas possuem uma baixa revocação. A grande ideia por trás deste trabalho está em levantar a hipótese de que padrões léxico sintáticos podem indicar algum tipo específico de relação entre as palavras em que estes ocorrem. Desta forma, presume-se que um padrão descrito por $SN\ SN$, onde SN representa um sintagma nominal, identificará em sentenças como *"Bebidas, tais como Juripenho, tem sua origem..."* que *Juripenho* é um sub-tipo de bebida, portanto *hiperônimo(juripenho, bebida)*. Esses padrões manualmente criados devem encontrar poucos casamentos em textos, mas destes casamentos muitos devem possuir relações verdadeiras.

Esses padrões são genéricos o suficiente para serem aplicados em todos tipos de gêneros literários e quaisquer domínios. Os padrões de Hearst foram criados pensando em três preceitos básicos:

1. Ocorrem frequentemente e em diversos gêneros literários;
2. Quase sempre indicam a relação de interesse; e
3. Podem ser reconhecidos com pouco ou nenhum conhecimento pré-codificado.

Os Padrões são descritos pelas seguintes regras, onde NP representa um sintagma nominal, as palavras entre chaves são opcionais e os parênteses listam um conjunto de opções válidas:

1. NP such as {NP, NP ..., (and | or)} NP
   Authors such as Herrick, Goldsmith and Shakespeare.

2. such NP as {NP, }*{(or|and)} NP
   Such authors as Herrick, Goldsmith and Shakespeare.
3. NP{, NP}*{,} or other NP
   Herrick, Goldsmith, Shakespeare or other authors.
4. NP{, NP}*{,} and other NP
   Herrick, Goldsmith, Shakespeare and other authors.
5. NP{,} including {NP, }*{(or|and)} NP
   Authors, including Herrick, Goldsmith and Shakespeare.
6. NP{,} especially {NP, }*{(or|and)} NP
   Authors especially Herrick, Goldsmith and Shakespeare

2.4 Aquisição Automática de Padrões

A aquisição automática de padrões tem por objetivo obter padrões a partir de técnicas de extração de informações a partir de conjuntos de texto (corpus ou da web). Tais padrões objetivam representar uma construção da linguagem que represente uma relação única, ou seja, encontrar padrões semelhantes aos de Hearst para diversas relações. Uma das abordagens iniciais para isso utilizou pares de palavras com relação conhecida como base para descobrir os padrões, como mostrado na Seção 2.4.1, para encontrar os padrões desejados. A partir desta abordagem outras ideias foram utilizadas, por exemplo, um aprimoramento da filtragem de padrões (Seção 2.4.2) e utilização de notações sintáticas para criação destes padrões (Seção 2.4.3).

2.4.1 DIPRE

O algoritmo de Expansão de Relação de Padrão Iterativo Dual (*Dual Iterative Pattern Relation Expansion* – DIPRE) (BRIN, 1998) propôs um método de extração de padrões a partir da web e aquisição de relações em cima destes padrões extraídos. Esse método estabeleceu um algoritmo bastante popular e ainda recorrente, que funciona através dos seguintes passos:

1. Extrair um pequeno conjunto de sementes da relação alvo;
2. Encontrar todas as ocorrências das sementes no corpus;

3. Gerar os padrões (com baixa taxa de erro) baseado no conjunto de ocorrências adquiridas;
4. Encontrar no corpus quaisquer ocorrências dos padrões aprendidos; e
5. Voltar para a etapa 2, caso o conjunto de sementes não seja grande o suficiente.

A maior contribuição do trabalho de Brin (1998) foi a dupla-iteratividade, chamada de *bootstrapping*, onde as relações encontradas pelo sistema servem para re-alimentar o processo, dessa forma um pequeno conjunto inicial de sementes resultaria em uma grande quantidade de pares ao final do processo.

Outro ponto bastante importante levantado é sobre a qualidade das sementes e dos padrões aprendidos. Nesse assume-se que boas sementes geram bons padrões, e bons padrões ajudam a gerar boas sementes. Dessa forma o algoritmo irá a partir de boas sementes encontrar bons padrões, que irão retornar boas sementes novamente. Isso é importante, pois num método de *bootstrapping* o erro se espalha muito facilmente, como o algoritmo itera sobre sobre as próprias instâncias coletadas, caso essas instâncias não sejam boas o suficiente e acabem por extrair padrões ruins, esses vão encontrar muitas relações ruins, tendo um alto impacto na performance do algoritmo.

Brin (1998) utilizou o algoritmo para extração de relação entre autores e livros, tendo como entrada um conjunto de 5 pares de autores e livros. Para a avaliação dos resultados foram escolhidos de forma aleatória 20 pares encontrados e buscados na web. Destes 20 pares, 19 estavam corretos.

2.4.2 WWW2REL

O método de aquisição apresentado por Halskov e Barrière (2008), conhecido como WWW2REL é inspirado no DIPRE, tendo a base do algoritmo de dupla-iteratividade bastante similar. Esse trabalho ignora os possíveis padrões encontrados a direita e a esquerda das palavras alvo, baseando-se na ideia de que o contexto principal a ser olhado para encontrar padrões que identifiquem relações entre palavras é o contexto entre as duas palavras alvo.

Uma das questões levantas por este trabalho são os desafios para encontrar padrões de qualidade, através da seguinte análise a respeito destes padrões:

Imprevisíveis: por fazer parte da linguagem natural, os padrões ocorrem sem regras bem definidas e não possuem um conjunto total fixo, pois vão sendo criados novos

padrões conforme a linguagem vai se alterando

Polissêmicos: os padrões, assim como muitas palavras podem se referir a conceitos completamente diferentes, também podem ter conceitos distintos

Referências anafóricas: referências anafóricas são referências ao contexto em que o padrão esta inserido

Dependentes de domínio: podem ou não ser considerados dependendo do domínio em que se inserem

São apresentadas as seguintes métricas para a filtragem de padrões, tendo por objetivo remover padrões que são muito genéricos, se associando com muitos tipos de pares de palavras.

- Remover padrões sem verbos;
- Avaliar a precisão dos padrões, utilizando de pares negativos da relação. Dessa forma, pretende-se avaliar se o padrão possui uma relação maior com os pares da relação desejada ou com pares aleatórios que não pertencem a relação desejada; e
- Avaliar a quantidade de pares com os quais o padrão ocorre.

Para o ordenamento dos pares de palavras são apresentadas as seguintes estratégias:

- Frequência dos pares junto dos padrões;
- Quantidade de diferentes padrões com os quais os pares se associam; e
- Pointwise Mutual Information (PMI) das contagens dos pares em relação aos padrões.

No trabalho de Halskov e Barrière (2008) sao utilizadas consultas ao Google para encontrar instâncias das sementes, utilizando um máximo de 40 pares de palavras para cada relação desejada. O fato do corpus utilizado ser a web em si, e não um extrato dela, faz com que mesmo pares de baixa frequência possam ser encontrados e avaliados. A avaliação deste trabalho foi realizada para relações do domínio médico. No total foram 2.000 relações avaliadas, utilizando 4 juízes para analisar as relações encontradas.

2.4.3 Utilização de árvores de dependência

A metodologia criada por Snow, Jurafsky e Ng (2004) é um classificador automático de hipônimos e hiperônimos. Esse tem por objetivo automatizar a criação de padrões

apresentada na Seção 2.4.1, através de técnicas de aprendizado de máquina. Esta técnica utiliza um pipeline inspirado no apresentado na Seção 2.4.1.

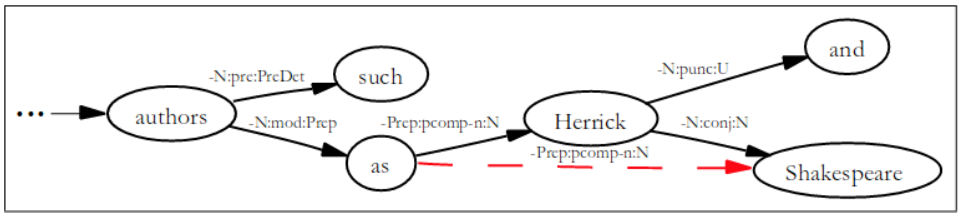
Para o treinamento deste classificador são encontrados pares de hiperônimos, através de ontologias existentes, em um corpus. Essa forma de aquisição dos pares sementes do algoritmo não necessita de interferência humana na escolha, mas necessita que uma ontologia do domínio desejado já tenha sido criada. Essas sementes não são limitadas por um número pequeno, podendo ser automaticamente extraídas diversas sementes do corpus. Para identificar os padrões léxico-sintáticos, a abordagem realizada para o treinamento segue abaixo:

1. Encontrar pares de substantivos em sentenças de corpora, que possuam a relação de hiperonímia ou hiponímia na WordNet;
2. Para cada, par encontrar todas suas ocorrências em um corpus, extraindo as sentenças em que estes ocorrem;
3. Extrair a árvore de dependência das sentenças e, a partir desta árvore, extrair os padrões que conectam o par desejado; e
4. Treinar um classificador de hiperonímia baseado nesses padrões extraídos.

A utilização de um parser para a geração de um árvore de dependência de uma sentença, tem por objetivo encontrar os padrões léxico-sintático presentes nas sentenças extraídas. O padrão existente é o menor caminho na árvore sintática que conecte as duas palavras alvo, desconsiderando as palavras alvo com o intuito de generalizar o padrão para qualquer par de palavra. Aos padrões também são adicionados elementos opcionais, o primeiro deles é chamado de satélites, que são palavras que se conectam com os substantivos, mas que não entrariam no menor caminho da árvore sintática. O segundo é o aproveitamento de conjunções da língua.

É possível analisar um exemplo da árvore de dependência na Figura 2.3. Nesse exemplo, a frase analisada é *"... authors such as Herrick and Shakespeare"*. Nesse exemplo, o padrão que desejá-se extrair é *such as*, tendo *hiperônimo(Herrick, authors)* e *hiperônimo(Shakespeare, authors)*. Neste exemplo podemos ver a importância da utilização de satélites na aquisição de padrões, pois o satélite *such* faz com que este padrão ganhe muito em qualidade, caso contrário o padrão seria apenas *as*. Outro ponto é o aproveitamento das conjunções, como dito acima, esta sentença aproveita a conjunção *and* para adquirir *Herrick* e *Shakespeare*.

Figura 2.3: Exemplo da árvore de dependência no MINIPAR



Fone: (SNOW; JURAFSKY; NG, 2004)

3 MATERIAIS E MÉTODOS

Neste trabalho temos como objetivo detalhar o método utilizado, como apresentado no Capítulo 1, para produzir ontologias linguísticas através de padrões léxico-sintáticos automaticamente extraídos de textos. Com o intuíto de realizar esse objetivo, nessa seção, apresentamos os recursos e ferramentas utilizados (Seção 3.1) e a metodologia empregada para a criação automática de ontologia (Seção 3.2).

3.1 Materiais

Para alcançar o objetivo deste trabalho são necessários recursos computacionais que sejam representativos o uso da língua e que permitam uma representação computacional dessa. Neste trabalho o uso da língua é representado por corpora e dicionários que expressão relações ontológicas entre palavras. A representação computacional é realizada através de análise sintática, a qual é utilizada para obtenção de lemas e árvore de dependências. Assim, utilizamos o corpus anotado com árvore de dependências ukWaC (BARONI et al., 2009) para representar o uso da língua e os dicionários WordNet (MILLER, 1995), BLESS (BARONI; LENCI, 2011) e HyperLex (VULIĆ et al., 2016) para representar as relações entre as palavras.

O corpus ukWaC (BARONI et al., 2009) é um corpus criado a partir da web, extraindo sentenças a partir de um método de *web crawling*[1]. Esse corpus possui uma grande variedade de tópicos, dentre eles textos técnicos, blogs e transcrições de linguagem falada (FERRARESI et al., 2008). O método de *web crawling* foi realizado através de uma escolha de pares de palavras sementes. Esses pares foram extraídos do corpus BNC (LEECH, 1992) e de uma lista de palavras em inglês para estrangeiros, totalizando um total de 2.000 pares. Esses pares de palavras foram buscados no *Google*, extraindo então 10 URLs dos resultados das buscas. Essas URLs são então unificadas e mantido apenas uma URL para cada domínio. Estas URLs serviram de entrada para o *crawler Heritrix*[2]. O corpus ukWaC, anotado com o analisador sintático MaltParser (NIVRE; HALL; NILSSON, 2006), apresenta quatro informações sobre cada palavra: forma de superfície, forma lematizada, classe gramatical e informações sobre as relações de dependência criadas pelo

[1] *Web crawling* é um método que navega automaticamente pela web para adquirir textos que serão processados posteriormente (KILGARRIFF; GREFENSTETTE, 2003)

[2] crawler.archieve.org

analisador sintática. O corpus possui um total de 1.914.150.197 tokens e 3.798.106 types.

BLESS é um conjunto de palavras criado para a avaliação de modelos de semântica distribucional, que possue para cada palavra diversas palavras divididas em 5 relações extraídas da WordNet. Estas relações são:

COORD: palavras que possuem hiperônimos em comum, ou seja, são co-hipônimas. Por exemplo, *COORD(alligator, lizard)*.

HYPER: palavras que possuem relação de hiperonímia. Por exemplo, *HYPER(alligator, animal)*.

MERO: palavras com relação de meronímia. Por exemplo, *MERO(alligator, mouth)*.

ATTRI: adjetivos que expressem um atributo do conceito. Por exemplo, *ATTRI(alligator, aquatic)*.

EVENT: verbos que representam uma ação com a qual o conceito está relacionado. Por exemplo, *EVENT(alligator, swim)*.

São um total de 200 entradas no conjunto. As relações com essas entradas são apresentadas na Tabela 3.1.

Tabela 3.1: Tabela sobre as relações do BLESS

Relações	**Quantidade**
total	26.554
hiperônimos	1.343
atributos	2.736
merônimos	2.944
co-hipônimos	3.570
eventos	3.384
aleatórias	12.577

Esses pares foram utilizados para gerar sementes para o algoritmo de extração de padrões.

A fim de expandir o conjunto de relações do BLESS para sinônimos, utilizamos os substantivos do BLESS presentes nas relações de HYPER e MERO e os substantivos comuns da língua inglesa. Para a identificação de substantivos comuns foi utilizado o *Oxford 3000*[3]. Esse possui um total de 3.352 palavras selecionadas pelo critério das palavras de maior frequência na língua. Essas palavras foram utilizadas, juntamente com todas as palavras do BLESS das relações de merônimos e sinônimos, para gerar uma lista de palavras que seriam buscadas na WordNet para a aquisição de sinônimos. Através deste conjunto de palavras foi buscado na WordNet as relações de sinônimos, resultando numa

[3] www.oxfordlearnersdictionaries.com

lista de 2.606 pares de sinônimos e um total de 3.127 palavras distintas.

HyperLex é um conjunto de dados de palavras baseado no BLESS, mas que estabelece uma pontuação dos pares referente a relação apresentada por eles[4]. Um exemplo das pontuações pode ser encontrado na Tabela 3.2. Esse conjunto de dados se baseia na premissa de que palavras possuem graus continuo de relação e não binária. Para avaliar os diferentes graus de relação este recurso elaborou uma avaliação humana com 10 juízes para cada par de relação, não necessariamente os juízes devem ser diferentes para cada par. Para avaliar o quão relacionado são os pares, foi feito um questionário, no qual foi perguntado *"Em que grau X é um tipo de Y"*, podendo haver uma resposta de 0 a 10. O conjunto total apresenta um total de 2.616 pares.

Tabela 3.2: Tabela de pares e pontuações

Par	**Pontuação**
girl / person	9.85
citizen / person	8.63
person / citizen	5.17
idol / person	4.28

3.2 Método

A extração automática de ontologias apresentada neste trabalho é fortemente baseada nos trabalhos de Brin (1998), que apresenta uma baseline seguido por muitos trabaho da área, e Snow, Jurafsky e Ng (2004), que apresenta uma forma de extração de padrões baseada em árvores de dependência, apresentados na Seção 2.4. Esses trabalhos foram escolhidos por apresentarem uma proposta simples para identificação de relações entre palavras. O trabalho realizado por Halskov e Barrière (2008) também foi importante para criar uma intuição geral sobre filtragem de relações e pares de palavras adquiridos, para tentar reduzir o ruído do algoritmo. Assim, os passos do algoritmo apresentado neste trabalho, ilustrados na Figura 3.1, são:

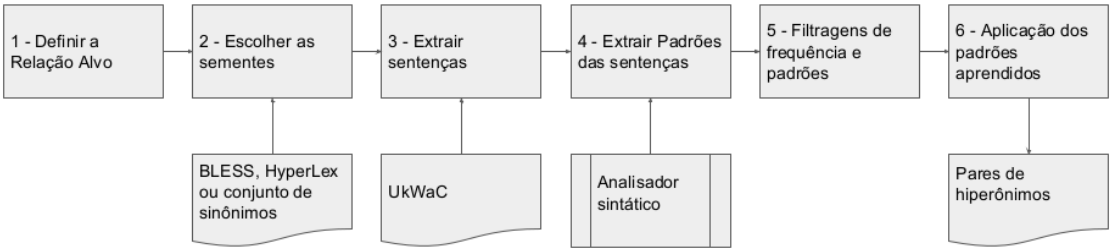
1. Definição da nossa relação de interesse;
2. Definição do conjunto de pares de palavras como sementes;
3. Extração de sentenças;
4. Extração do padrão que une o par semente através da árvore sintática da sentença;
5. Filtragem:

[4]http://people.ds.cam.ac.uk/iv250/hyperlex.html

- Filtragem de frequência;
- Filtragem de padrões; e

6. Utilização dos padrões encontrados para encontrar pares da relação alvo.

Figura 3.1: Pipeline do algoritmo para extração de ontologias



Neste trabalho a relação de interesse explorada foi a relação linguística de hiperonímia. Essa relação foi escolhida por ser a principal relação definicional em uma ontologia, sendo utilizada em muitors trabalhos, como por exemplo Hearst (1992), Snow, Jurafsky e Ng (2004), Caraballo (1999) e Ravichandran e Hovy (2002) A partir da escolha da relação alvo foram buscados pares que possuam essa relação. Esses pares foram selecionados através do BLESS e o do HyperLex, que são conjuntos de palavras com categorias de relação específicas. Os pares do HyperLex possuem uma pontuação que indica a relação de uma forma mais contínua, portanto foram selecionados pares do HyperLex com uma pontuação maior de $9,7$, de forma a indicar pares com relações fortes. No BLESS foram selecionados todos os pares com a relação de hiperônimos, sem distinção. A partir destes dois conjuntos foram criadas duas ontologias discutidas no capítulo 4.

A partir das sementes se da a extração das sentenças em corpus. O corpus escolhido foi o ukWaC (apresentado na seção 3.1). As sentenças encontradas mantêm a ordem dos pares, de forma que a primeira palavra encontrada é mais específica que a segunda e ocorrem na forma apresentada pela Tabela 3.3. Essa tabela apresenta exemplos de sentenças para os pares de palavras *(cachorro, animal)* e *(manga, fruta)*.

Tabela 3.3: Sentenças e palavras alvo extraídas do corpus

Esquerda	**Palavra 1**	**Centro**	**Palavra 2**	**Direita**
... sabemos que	cachorro	é um tipo conhecido de	animal	doméstico ...
...	manga	e abacaxi são	frutas	tropicais
São	cachorros		animais	?
A	manga	é uma	fruta	deliciosa ....

A partir dessas sentenças extraídas, é necessário descobrir o padrão presente nelas. O algoritmo para extração dos padrões foi baseado no método apresentado por Snow,

Jurafsky e Ng (2004), no qual inicialmente é extraído o caminho mais curto da árvore sintática que une as duas palavras alvo. Posteriormente são removidos os padrões que possuam substantivos no caminho entre as duas palavras alvo. Este filtro é feito porque se um substantivo ocorre no meio do padrão, é mais provável que o padrão esteja referindo-se ao substantivo mais próximo, e não aos dois substantivos que nos interessam. Esses padrões são ainda generalizados, substituindo o substantivo original por um *wildcard*, chamado de NP. Padrões longos também são removidos, consideramos padrões com tamanho maior que de 3 palavras desnecessários para compor os padrões desejados, restringindo o tamanho para padrões com até 3 palavras. Após a filtragem os padrões são limpos de categorias sintáticas consideradas desnecessárias e as categorias alvo são adjetivos, pontuações e conjunções coordenadas. Desta forma, os padrões desejadas devem apresentar a estrutura apresentada na Tabela 3.4

Tabela 3.4: Padrões criados a partir das sentenças extraídas

Esquerda	**Palavra 1**	**Centro**	**Palavra 2**	**Direita**
	NP	é tipo de	NP	
	NP	é	NP	
São	NP		NP	
	NP	é	NP	

Para filtrar a quantidade de padrões, dois tipos de filtros foram utilizados. O primeiro, frequência, remove padrões gerados apenas por um dos pares sementes e pares que adquiriram um só padrão. Esse filtro foi aplicado iterativamente até estabilizar o número de padrões. O segundo filtro utilizado, remove padrões que podem indicar outras relações. Para tanto, foram extraídos os padrões de sinônimos e merônimos e comparados com os padrões de hiperônimos, permanecendo apenas os padrões que ocorriam somente na relação de hiperonímia. A análise referente a esse filtro é apresentada na Seção 4.2.

A avaliação deste trabalho é realizada de forma intrínseca e extrínseca.

Intrínseca visa avaliar a qualidade dos padrões gerados em relação as sementes e a exclusividade que um padrão pode representar. Segundo Brin (1998), um pequeno conjunto de sementes é capaz de gerar os padrões desejados para o algoritmo de *bootstrapping*[5] e esses padrões podem gerar pares de qualidade suficiente para realimentar o algoritmo. Portanto, desejamos avaliar se estes pares gerados possuem qualidade suficiente para realimentar o algoritmo, tendo em vista que pares ruins podem arruinar a performance do algoritmo. Acreditamos que isso é importante,

[5] o algoritmo de *bootstrapping* utiliza os resultados gerados para realimentar a sua entrada 2.4.1

pois esses passos todos são dados como intuitivos, mas segundo Auger e Barrière (2008) não chegam a ser realizadas avaliações intrínsecas sobre as abordagens, sendo avaliado apenas o resultado obtido (avaliação extrínseca).

Extrínseca busca avaliar a qualidade geral do método estabelecido. Isso é feito através de uma análise da ontologia criada. Esta análise pode ser de diversas formas, uma avaliação manual ou uma comparação com um *Gold Standard* previamente estabelecido. Nesse trabalho, a avaliação extrínseca se dará em comparação com um *Gold Standard* e sua performance comparada com o algoritmo proposto por Hearst (1992).

4 ANÁLISES

Neste capítulo, temos por objetivo apresentar as análises intrínsecas realizadas para avaliar os pares (Seção 4.1) e padrões (Seção 4.2 e 4.3) extraídos. Também é apresentado as análises extrínsecas realizadas, através de uma comparação do resultado final do pipeline proposto com o resultado da aplicação dos padrões de Hearst (Seção 4.4).

4.1 Avaliação das sementes

Primeiramente procuramos avaliar a premissa de que um conjunto pequeno de pares é capaz de representar a linguagem e as suas relações (BRIN, 1998). Através dessa premissa é possível dizer que o número de padrões converge em relação aos pares de palavra. Dessa forma, existiria um momento em que a adição de novos pares nas sementes do algoritmo se tornaria redundante e seria então possível estabelecer um número ideal de pares para a aquisição dos padrões. Para analisar essa hipótese, estudamos o crescimento do número de padrões em relação ao tamanho do conjunto inicial de sementes, e sua distribuição para diferentes tamanhos de padrões.

O crescimento se deu através da seleção aleatória de *X* pares de palavras do conjunto total de sementes, onde *X* inicia em 6 e vai até o número total de pares. A progressão dos pares ocorre da seguinte forma: $25 \times i$ pares, onde $i = \{1, 2, 3....\}$. Para cada *X* pares selecionados, foram avaliados 10 amostras desses pares, para termos uma melhor análise do comportamento geral e reduzir possível viés de ruído amostral.

Os resultados do número médio de padrões obtidos em relação ao número de pares utilizados podem ser vistos na Figura 4.1[1]. A Tabela 4.1 apresenta as regressões linear e logarítmicas feitas para Figura 4.1. Este ajuste de curva serve para que possamos avaliar melhor o comportamento da taxa de aprendizado dos padrões em relação à quantidade de pares. A figura também apresenta os desvios padrões relacionados a cada média extraída, representando o quanto a quantidade de padrões para N pares é próxima da média destes N pares.

Com base nas medidas de erro (R^2)[2] apresentadas na Tabela 4.1, observamos que o ajuste linear é mais adequado que o ajuste logarítmico. Isso indica que a progressão de padrões em relação aos pares sementes não corresponde a nenhuma convergência pros

[1] Na Figura 4.1 as barras de erro indicam o desvio padrão.

[2] o erro avaliado é o coeficiente de determinação, em que quanto maior o R^2 mais explicativo é o modelo

Tabela 4.1: Ajustes linear e logarítmico sobre as curvas da Figura 4.1

(a) Ajustes linear sobre as curvas da Figura 4.1

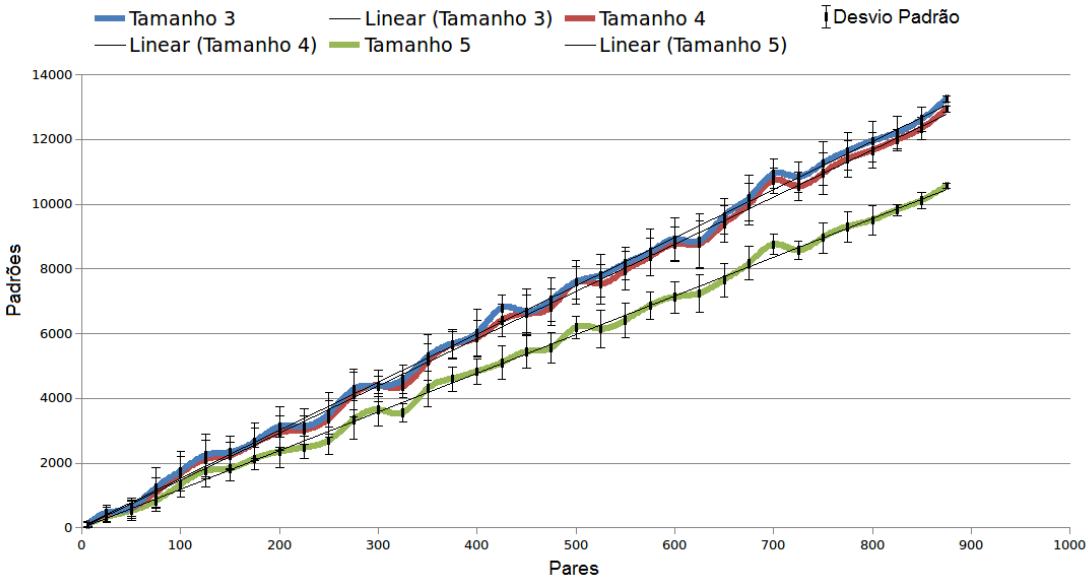
Tamanho	f(x)	R^2
1	14,8785x - 49,2246	0,997711
2	14,6210x - 2,0151	0,997868
3	11,9538x - 6,8460	0,997887

(b) Ajustes logarítmico sobre as curvas da Figura 4.1

Tamanho	f(x)	R^2
1	2492,66ln(x) - 9079,40	0,740
2	3050,34ln(x) - 11103,48	0,741
3	3108,47ln(x) - 11277,15	0,743

parâmetros utilizados. Portanto a adição de pares continua sendo significativa para a aquisição de padrões. Analisando as funções lineares que descrevem o crescimento do número de padrões, observamos um crescimento maior e similar para as relações de tamanho 1 e 2 que para as relações de tamanho 3. Isso possivelmente é devido à maior facilidade em se descobrir padrões menos específicos.

Figura 4.1: Progressão de padrões por tamanho das sementes



Uma vez que os pares de palavras utilizados para buscar os padrões mudam em cada iteração, analisamos o desvio padrão do número de padrões encontrados. Isso pode indicar uma diferença acentuada em padrões adquiridos entre os pares utilizados. Como ilustrado na Figura 4.1, o desvio padrão é similar para todos os números de pares de pa-

lavras utilizados, indicando que a quantidade de padrões adquiridos por par de palavras varia consistentemente. Isso indica que sementes possuem uma alta variação na quantidade de pares que elas adquirem, podendo-se dizer, então, que elas possuem diferente qualidade.

Os resultados apresentado nessa seção, ilustrados na Figura 4.1, indicam que a aquisição de padrões não tende a estabilizar[3]. Indicando assim que a premissa de poucos pares de palavras serem representativos da linguagem não parece válida, pois não há redução do número de padrões encontrados. Sendo que uma representação linear explica melhor o crescimento do número de padrões encontrados em função do número de pares que uma representação logarítmica. Contudo, destacamos que muitos dos padrões aprendidos podem ser utilizados para representar outras relações além de hiperonímia.

4.2 Avaliação da exclusividade dos padrões

Neste trabalho também avaliamos se um padrão aprendido representa somente um tipo relação, ou se talvez um par de palavras gere relações que poderiam indicar outras relações. Como apresentado na Seção 2, autores, tais como, Snow, Jurafsky e Ng (2004), Hearst (1992) e Brin (1998) consideram que os padrões adquirem um tipo de relação. Para avaliar isto consideramos também as relações de merônimo e sinônimo. Os merônimos foram obtidos a partir das relação do BLESS, enquanto que os sinônimos foram obtidos utilizando a relação de sinonímia indicada pela WordNet, a partir do conjunto de palavras adquiridas pelo Oxford3000 em conjunto das palavras já adquiridas de sinônimos e merônimos. Dessa forma obtivemos um total de 6.886 pares de palavras dividas em três relações (Tabela 4.2 apresenta o número de pares por relação).

Tabela 4.2: Total de pares por relação

Relações	**Total de Pares**
Hiperônimos	1.337
Sinônimos	2.606
Merônimos	2.943

Dessas relações, foram extraídos todos os padrões das sentenças encontradas no ukWaC, que potencialmente representassem a relação em questão. Posteriormente foram analisados as intersecções entre os conjuntos de padrões aprendidos, visando identificar

[3]Neste trabalho exploramos até 897 pares de entrada, por possuírem ocorrências no corpus, dos 1.343 existentes no BLESS.

quantos padrões únicos foram extraídos.

Também foi criado um conjunto de padrões a partir da intersecção entre hiperônimos e a união dos sinônimos com os merônimos (Hiperônimos - (Sinônimos + Merônimos)). Com isso pretendemos identificar se as relações aprendidas pelos pares de hiperônimos foram únicas, podendo gerar, dessa forma, apenas hiperônimos, ou se eram ambíguas, podendo extrair pares de palavras com uma relação diferente da desejada. O resultado dessa análise pode ser vista na Tabela 4.3, com base na qual pode-se observar que a intersecção entre as relações de fato acontece.

Como é possível observar na Tabela 4.3, os padrões possuem uma intersecção alta entre os diferentes tipos de relações. Dessas, o percentual de padrões de hiperonímia que ocorrem também na relação de sinonímia é de 28,64% e de 25,20% quando comparado com a relação de meronimia, sendo considerado o total de padrões de hiperonímia para o cálculo da percentagem.

Tabela 4.3: Ocorrência dos padrões nos diferentes tipos de relação

Relações	**Total de Padrões**
Hiperônimos	15.651
Sinônimos	84.547
Merônimos	49.803
Hiperônimos e sinônimos	4.484
Hiperônimos e merônimos	3.944
Merônimos e sinônimos	9.337
Apenas hiperônimos	10.016

Com base na ocorrência de padrões nos diferentes tipos de relações, é possível afirmar que os padrões extraídos de pares que possuem uma relação específica, não necessariamente irão ter uma exclusividade de significado. O que nos leva a crer que a utilização de padrões para extrair um tipo específico de relação é mais complexa do que é atualmente reconhecido (como nos trabalhos apresentados por Brin (1998) e Snow, Jurafsky e Ng (2004)).

Esse resultado é importante para avaliarmos que os padrões aprendidos podem representar relações indesejadas, o que, segundo Brin (1998) é algo que interferiria no algoritmo de *bootstrapping*, pois pares de hiperônimos que acabam por gerar padrões incorretos, ou seja, que possam adquirir pares de palavras diferentes, são bastante prejudiciais para toda a ideia de auto-realimentação do algoritmo.

4.3 Avaliação das sementes reconhecidas apenas como Hiperônimos

Com base nas avaliações anteriores, analisamos se ao utilizarmos somente padrões cuja relação seja de hiperônimos exclusivos se observaria um saturamento no crescimento de padrões, ou seja, a partir de um ponto a adição de novos pares sementes não seria mais significativa para a aquisição dos padrões desejados. O objetivo dessa análise é avaliarmos se, ao adquirir apenas hiperônimos, seria possível que a quantidade de padrões aprendidos convergisse.

A Figura 4.2 mostra o número de padrões extraídos para os padrões considerados hiperônimos e para os padrões considerados apenas hiperônimos (não contendo merônimos e sinônimos), da Tabela 4.3. Nessa figura, é possível comparar a diferença na progressão das duas relações com o tamanho dos seus padrões. As três linhas mais acima são referentes aos padrões representando hiperônimos e as três linhas mais abaixo, somente hiperônimos. É possível perceber que todos os segmentos caíram, o que já era esperado, tendo em vista que os padrões somente hiperônimos são um subconjunto dos hiperônimos originais. Nesse gráfico é visível que os novos padrões de tamanho 1 e 2 continuam tendo um crescimento linear, mas o padrão de tamanho 3 (linha roxa) a identificação da progressão não é tão facilmente observável.

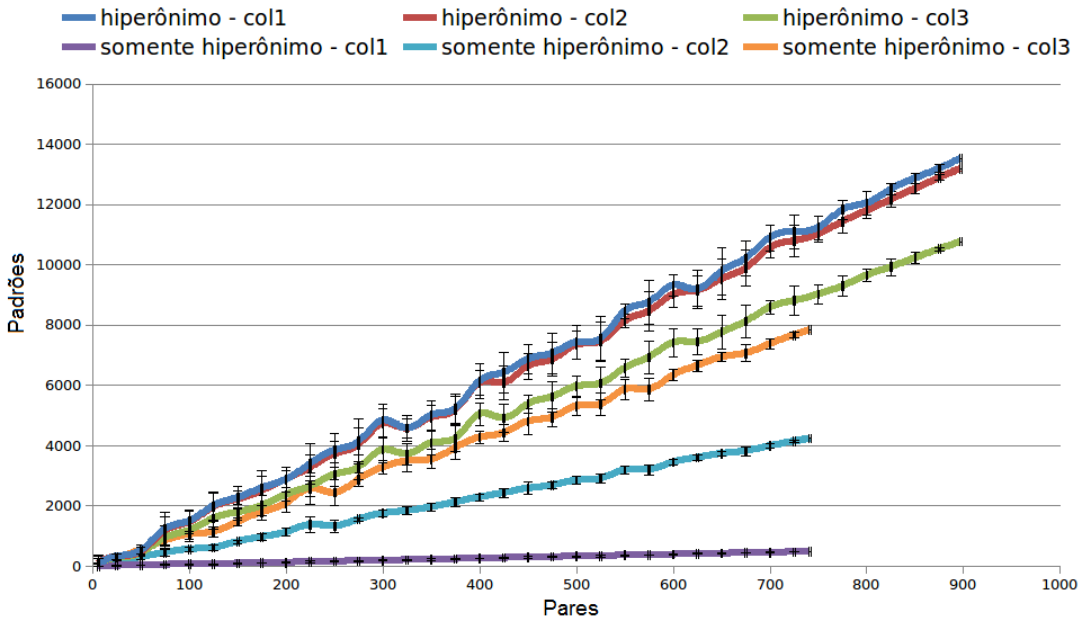
Para verificar se esta linha é logarítmica ou linear, realizamos uma regressão linear e polinomial. Na regressão linear o coeficiente de determinação (R^2) é de $0,9989\%$ e na regressão polinomial de $0,7429\%$. Portanto, os resultados da Seção 4.1 se mantem e a adição de novos pares continua sendo relevante para a aquisição de padrões.

Nessa figura, a diferença no eixo X dos conjuntos de dados se deve somente a quantidade total de padrões resultantes ao serem removidos os padrões que se encaixavam em outras relações, por isto a diferença de 897 pares de hiperônimos totais para 741 pares somente de hiperônimos.

4.4 Avaliação da ontologia gerada por sementes restritas

Avaliando a ontologia adquirida de forma extrínseca, ou seja, avaliando os resultados por ela criados. Esta avaliação é importante para poder comparar os nossos resultados com os resultados de outros trabalhos da área. Nesta avaliação é feita uma comparação entre os resultados do nosso trabalho com o trabalho de Hearst (1992). A comparação dos resultados adquiridos pelos dois trabalhos é feita contra um *gold standard*.

Figura 4.2: Progressão de padrões por tamanho das sementes para hiperônimos e apenas hiperônimos



Para esta avaliação, executamos o método utilizando apenas sementes de boa qualidade, extraídas do HyperLex e filtrando pares com pontuação maior que 9,7, resultando em um conjunto de sementes de 128 pares de hiperônimos. Esta escolha teve base em um dos fundamentos apresentados por Brin (1998), que diz que boas sementes geram bons padrões. O que esperamos desta decisão é que ao restringir a qualidade das sementes, teremos padrões mais sólidos, representando melhor a relação desejada.

Também executamos o método apresentado por Hearst (1992). Esse método apresenta padrões criados manualmente, e tem por objetivo uma precisão maior. Dessa forma, é esperado que os padrões apresentados por Hearst (1992) retornem apenas palavras com a relação de hiperonímia.

Ambos os experimentos foram feitos utilizando um extrato do ukWaC, contendo 6.000.000 de sentenças, buscando apenas relações de hiperonímia. Os resultados, encontrados na Tabela 4.4, mostram a revocação dos algoritmos para o conjunto de hiperônimos, e também os pares de relações incorretas (sinônimos e merônimos).

A Tabela 4.4 apresenta uma comparação entre duas abordagens para criação de ontologia, ambas através de padrões textuais, mas uma tendo estes padrões construídos de forma manual (Padrões de Hearst) e outra de forma automática (Algoritmo Proposto). As ontologias geradas foram analisadas em relação à sua revocação em frente a 3 *gold*

Tabela 4.4: Comparação da Ontologia linguística gerada com pares de hiperônimos do BLESS e de sinônimos e o total de pares apresentados por relação no BLESS

	Total de pares	**Algoritmo Proposto**	**Padrões de Hearst**
hiperônimos	1.337	83	21
sinônimos	2.606	258	14
merônimos	2.943	131	4

standards de relações (hiperônimos, sinônimos e merônimos). Buscando avaliar a revocação do algoritmo proposta com um *gold standard* de hiperônimos e também analisar os pares extraídos com uma relação indesejada, comparando com os *gold standards* de sinônimos e merônimos. O algoritmo proposto adquiriu mais pares de hiperonímia, mas também adquiriu muitos pares indesejados, mostrando que ele possui uma baixa acurácia em seus padrões. Os padrões de Hearst possuem resultados um pouco mais precisos, adquirindo mais hiperônimos que relações indesejadas. Tendo em vista que os padrões de Hearst deveriam ter uma precisão bastante alta, verificamos que ele também não atinge seu objetivo, visto que quase metade das relações adquiridas são indesejadas.

Os resultados da abordagem proposta não conseguem diferenciar sinônimos, merônimos e hiperônimos. Provavelmente isso se deve ao fato de os padrões aprendidos não conseguirem reproduzir padrões específicos de hiperônimos, adquirindo desta forma apenas palavras relacionadas de alguma forma, mas sem uma acurácia de qual relação. Os padrões de Hearst deveriam deveriam ser bastante precisos. Apesar de ter usado um pequeno conjunto de padrões de alta qualidade, a precisão dos pares adquiridos deixou a desejar, pois foram retornados bastante pares de sinônimos e merônimos, além dos pares desejados de hiperônimos.

Esses resultados nos levam a crer que, no corpora utilizado, os padrões não são utilizados de forma tão restrita. Mesmo padrões que deveriam indicar uma relação de hiperonímia, podem acontecer em uma relação de sinonímia, até pela proximidade destas duas relações. Devido a isso, é possível que a informação de relações entre as palavras não esteja tão facilmente disponível, o que justificaria a intersecção de padrões apresentadas na Seção 4.2

5 CONCLUSÕES

Ontologias linguísticas são um recurso muito importante em PLN, servindo como base para aplicações de análise de sentimentos (BALDONI et al., 2012), mineração de textos (SPASIC et al., 2005) e sistema de perguntas e respostas (LOPEZ et al., 2007). A criação e manutenção de ontologias manuais é muito cara e demorada, não sendo possível fazer com que elas se mantenham atualizadas na mesma velocidade que as informações da web. Tendo isso em vista, é importante que seja possível uma forma de manter uma ontologia que represente o estado atual da informação.

Com o objetivo de ter uma ontologia que possa ser mantida atualizada com maior facilidade, este trabalho apresentou um método para criação de ontologias de forma automática a partir de conjunto de textos do domínio desejado. Esse método é genérico e pode ser utilizado para encontrar qualquer tipo de relação desejada.

Para realizar nosso objetivo de estabelecer um método para criação de ontologias, foram analisados fatores para confirmar algumas hipóteses bastante intuitivas de trabalhos importantes na área. Essas hipóteses estabelecidas são: H1: uma pequena quantidade de pares como sementes do algoritmo são capazes de representar a linguagem e H2: os padrões exprimem uma relação específica entre seus pares.

Com relação a hipótese H1, os resultados obtidos indicam que uma baixa quantidade de sementes não é suficiente para uma estabilização dos pares adquiridos, não confirmando a hipótese inicial. Isto mostra que uma análise mais profunda sobre a aquisição de padrões por pares é necessária para que se possa restringir os padrões apreendidos, em busca de uma estabilização com o aumento do número de pares. Também é necessário uma melhor análise das diferentes formas como um padrão ocorre em textos da web, buscando unificar estas ocorrências.

Com relação a hipótese H2, os resultados da abordagem proposta por Hearst (1992) para a aquisição de pares relacionados indicam que os padrões criados não se encaixam em apenas uma relação, adquirindo pares de relações indesejadas e não apenas de hiperonímia. Estes resultados mostram que existe uma falta de unicidade nos padrões criados. Estes resultados podem ser levados para o trabalho proposto e também o de Brin (1998), onde os padrões são criados de forma automática. Os padrões criados apresentam a mesma falta de unicidade de relação encontrado nos padrões de Hearst, indicando que a pura aplicação dos padrões não é suficiente para identificar uma relação nos textos explorados e que novas abordagens devem ser utilizadas para filtrar os pares encontrados, como

por exemplo a utilização de padrões de diferentes relações para remover pares ambíguos.

Como trabalhos futuros pretendemos analisar mais a fundo essa dualidade de significados nos padrões, buscando identificar se a causa está atrelada a linguagem comum utilizada ou se esta dualidade faz parte da linguagem em si. Deste modo seria possível avaliar se em textos de domínio específicos os padrões são menos ambíguos. A partir deste estudo poderemos entender como utilizar padrões mais específicos na criação de ontologias automáticas.

Pretende-se também realizar o segundo passo do *bootstrapping*, realimentando o algoritmo através dos resultados por ele adquirido, isso quando os resultados apresentarem um maior grau de precisão. Além do método estatístico para aquisição de padrões, também sera feita uma análise utilizando aprendizado de máquina.

REFERÊNCIAS

AUGER, A.; BARRIÈRE, C. Pattern-based approaches to semantic relation extraction: A state-of-the-art. **Terminology**, v. 14, n. 1, p. 1–19, 2008.

BALDONI, M. et al. From tags to emotions: Ontology-driven sentiment analysis in the social semantic web. **Intelligenza Artificiale**, IOS Press, v. 6, n. 1, p. 41–54, 2012.

BARONI, M. et al. The wacky wide web: a collection of very large linguistically processed web-crawled corpora. **Language resources and evaluation**, Springer, v. 43, n. 3, p. 209–226, 2009.

BARONI, M.; LENCI, A. How we blessed distributional semantic evaluation. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. **Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics**. [S.l.], 2011. p. 1–10.

BRIN, S. Extracting patterns and relations from the world wide web. In: SPRINGER. **International Workshop on The World Wide Web and Databases**. [S.l.], 1998. p. 172–183.

BUITELAAR, P.; CIMIANO, P.; MAGNINI, B. Ontology learning from text: An overview. **Ontology learning from text: Methods, evaluation and applications**, IOS Press, Amsterdam, v. 123, p. 3–12, 2005.

CARABALLO, S. A. Automatic construction of a hypernym-labeled noun hierarchy from text. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. **Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics**. [S.l.], 1999. p. 120–126.

ERNEST, D. **Representations of Commonsense Knowledge**. [S.l.]: Morgan Kaufmann, San Mateo, 1990.

FERRARESI, A. et al. Introducing and evaluating ukwac, a very large web-derived corpus of english. In: **Proceedings of the 4th Web as Corpus Workshop (WAC-4) Can we beat Google**. [S.l.: s.n.], 2008. p. 47–54.

HALSKOV, J.; BARRIÈRE, C. Web-based extraction of semantic relation instances for terminology work. **Terminology**, v. 14, n. 1, p. 20–44, 2008.

HEARST, M. A. Automatic acquisition of hyponyms from large text corpora. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. **Proceedings of the 14th conference on Computational linguistics-Volume 2**. [S.l.], 1992. p. 539–545.

KILGARRIFF, A.; GREFENSTETTE, G. Introduction to the special issue on the web as corpus. **Computational linguistics**, MIT press, v. 29, n. 3, p. 333–347, 2003.

LEECH, G. 100 million words of english: the british national corpus (bnc). **Language Research**, v. 28, n. 1, p. 1–13, 1992.

LOPEZ, V. et al. Aqualog: An ontology-driven question answering system for organizational semantic intranets. **Web Semantics: Science, Services and Agents on the World Wide Web**, Elsevier, v. 5, n. 2, p. 72–105, 2007.

MATKAR, R.; PARAB, A. Ontology based expert systems–replication of human learning. In: **Thinkquest˜ 2010**. [S.l.]: Springer, 2011. p. 43–47.

MILLER, G. A. Wordnet: a lexical database for english. **Communications of the ACM**, ACM, v. 38, n. 11, p. 39–41, 1995.

MILLER, G. A. et al. Introduction to wordnet: An on-line lexical database. **International journal of lexicography**, Oxford Univ Press, v. 3, n. 4, p. 235–244, 1990.

NAVIGLI, R.; PONZETTO, S. P. Babelnet: Building a very large multilingual semantic network. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. **Proceedings of the 48th annual meeting of the association for computational linguistics**. [S.l.], 2010. p. 216–225.

NIVRE, J.; HALL, J.; NILSSON, J. Maltparser: A data-driven parser-generator for dependency parsing. In: **Proceedings of LREC**. [S.l.: s.n.], 2006. v. 6, p. 2216–2219.

OLIVEIRA, H. G.; GOMES, P. ECO and Onto.PT: A flexible approach for creating a Portuguese wordnet automatically. **Language Resources and Evaluation Journal**, Springer, v. 48, n. 2, p. 373–393, 2014. Available from Internet: <http://dx.doi.org/10.1007/s10579-013-9249-9>.

RAVICHANDRAN, D.; HOVY, E. Learning surface text patterns for a question answering system. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. **Proceedings of the 40th annual meeting on association for computational linguistics**. [S.l.], 2002. p. 41–47.

SNOW, R.; JURAFSKY, D.; NG, A. Y. Learning syntactic patterns for automatic hypernym discovery. **Advances in Neural Information Processing Systems 17**, 2004.

SPASIC, I. et al. Text mining and ontologies in biomedicine: making sense of raw text. **Briefings in bioinformatics**, Oxford Univ Press, v. 6, n. 3, p. 239–251, 2005.

VULIĆ, I. et al. Hyperlex: A large-scale evaluation of graded lexical entailment. **arXiv preprint arXiv:1608.02117**, 2016.