Leonardo Sameshima Taba

Extração automática de relações semânticas a partir de textos escritos em português do Brasil

São Carlos – SP, Brasil

Junho de 2013

Leonardo Sameshima Taba

Extração automática de relações semânticas a partir de textos escritos em português do Brasil

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de São Carlos, como parte dos requisitos para a obtenção do título de Mestre em Ciência da Computação, área de concentração: Inteligência artificial

Orientador:
Helena de Medeiros Caseli

DEPARTAMENTO DE COMPUTAÇÃO
CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA
UNIVERSIDADE FEDERAL DE SÃO CARLOS

São Carlos – SP, Brasil

Junho de 2013

Ficha catalográfica elaborada pelo DePT da Biblioteca Comunitária da UFSCar

T112ea

Taba, Leonardo Sameshima.
Extração automática de relações semânticas a partir de textos escritos em português do Brasil / Leonardo Sameshima Taba. -- São Carlos : UFSCar, 2013.
85 f.

Dissertação (Mestrado) -- Universidade Federal de São Carlos, 2013.

1. Inteligência artificial. 2. Processamento de linguagem natural (Computação). 3. Extração de informação. 4. Extração de relações semânticas. I. Título.

CDD: 006.3 (20ª)

Universidade Federal de São Carlos

Centro de Ciências Exatas e de Tecnologia

Programa de Pós-Graduação em Ciência da Computação

"Extração Automática de Relações Semânticas a partir de Textos escritos em Português do Brasil"

Leonardo Sameshima Taba

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de São Carlos, como parte dos requisitos para a obtenção do título de Mestre em Ciência da Computação

Membros da Banca:

Helena de M. Caseli

Profa. Dra. Helena de Medeiros Caseli
(Orientadora - DC/UFSCar)

Profa. Dra. Gladis Maria de Barcellos Almeida
(DL/UFSCar)

Profa. Dra. Renata Vieira
(PUC/RS)

São Carlos
Julho/2013

Agradecimentos

Primeiramente, agradeço à minha família pelo apoio constante e por sempre acreditarem em meu potencial. Mãe, pai, Lê, Renato e Gui, agora o caçula, muito obrigado por tudo.

Meu segundo agradecimento vai para a Renata, minha namorada. Obrigado por não me dizer o que eu queria ouvir mas sim o que precisava, e por todos os momentos, alegrias e lições. Este mestrado foi muito mais feliz com você ao meu lado.

Um agradecimento especial à minha orientadora, professora Helena. Obrigado por toda a orientação, paciência, compreensão e amizade. Esta dissertação só existe graças a você.

Agradeço aos amigos que fiz nessa jornada, Thiago, Augusto e Ricardo; à turma do café pelas conversas e risadas na copa às quatro da tarde; ao Clube de Go: Eric, Flávio, Pinguim, Luiz, Jú e Alfredo, sentirei saudades das tardes e noites de jogo e diversão e das golodices. E não poderia deixar de agradecer ao Seinenkai da Associação Nipo-Brasileira de São Carlos: meu "primo" Paulo, Massa, João, Vô, Cris, Amy, Hiro3, Hiro4, Jonatan, Shinji, Hiroki, Toca, Osmar, Grace, Eiki, Panda, Tonhão, Raj, Minoru, Mário, Wendel, Dóris, Eiji, Murilo, Caique, Jairo, Tai, e todos os demais. Levo cada um de vocês aqui comigo, junto com as boas lembranças de todos os Integras, Matsuris e Undoukais.

Aos meus amigos que seguiram nessa trilha comigo e também estão a defender: Carlos, Diego, Dias, Paulo, César, Douglas; aos meus amigos que não seguiram nessa trilha: Mikio, Andressa, Gui, Jú, Daniel, Victor, Laís, Felps, Chohfi, Aramizu, Cuppi, Borgo, e toda a turma 07; aos meus velhos amigos de Bauru: Bruno "Mutt", Bruno e Felipe.

Aos docentes e funcionários da universidade por toda a presteza e trabalho durante todos esses anos em que estive por aqui; aos professores Marcondes e Delano, em especial, pelo entusiasmo com a Maratona de Programação.

Ao CNPq e à FAPESP pelo apoio financeiro.

A todos que fizeram a diferença em minha vida.

Obrigado!

Resumo

A extração de informação (EI) é uma das muitas aplicações do Processamento de Língua Natural (PLN); seu foco é o processamento de textos com o objetivo de recuperar informações específicas sobre uma determinada entidade ou conceito. Uma de suas subtarefas é a extração automática de relações semânticas entre termos, que é muito útil na construção e melhoramento de recursos linguísticos como ontologias e bases lexicais. A esse contexto soma-se o fato de que há uma demanda crescente por conhecimento semântico, visto que diversos sistemas computacionais de PLN necessitam dessas informações em seu processamento. Aplicações como recuperação de informação em documentos web e tradução automática para outros idiomas podem se beneficiar desse tipo de conhecimento. No entanto, não há recursos humanos suficientes para produzir esse conhecimento na mesma velocidade que sua demanda. Com o objetivo de remediar essa escassez de dados semânticos, esta dissertação apresenta a investigação da extração auomática de relações semânticas binárias a partir de textos escritos no português do Brasil. Tais relações se baseiam na teoria de Minsky (1986) e são usadas para representar conhecimento de senso comum no projeto *Open Mind Common Sense* no Brasil (OMCS-Br) desenvolvido no LIA (Laboratório de Interação Avançada), laboratório parceiro do LaLiC (Laboratório de Linguística Computacional) no qual esta pesquisa se desenvolveu, ambos da Universidade Federal de São Carlos (UFSCar). As primeiras estratégias para essa tarefa se basearam na busca de padrões textuais em textos, onde uma determinada expressão textual indica que há uma relação específica entre dois termos em uma sentença. Essa abordagem tem alta precisão mas baixa cobertura, o que levou ao estudo de métodos que utilizam aprendizado de máquina como modelo principal, englobando o uso de técnicas como classificadores probabilísticos e estatísticos, além de métodos de *kernel*, que atualmente figuram no estado da arte. Esta dissertação apresenta a investigação, implementação e avaliação de algumas dessas técnicas com o objetivo de determinar como e em que medida elas podem ser aplicadas para a extração automática de relações semânticas binárias em textos escritos em português. Desse modo, este trabalho é um importante passo no avanço do estado da arte em extração de informação com foco no idioma português, que ainda carece de recursos na área semântica, além de um avanço no cenário de PLN do português como um todo.

Abstract

Information extraction (IE) is one of the many applications in Natural Language Processing (NLP); it focuses on processing texts in order to retrieve specific information about a certain entity or concept. One of its subtasks is the automatic extraction of semantic relations between terms, which is very useful in the construction and improvement of linguistic resources such as ontologies and lexical bases. Moreover, there's a rising demand for semantic knowledge, as many computational NLP systems need that information in their processing. Applications such as information retrieval from web documents and automatic translation to other languages could benefit from that kind of knowledge. However, there aren't sufficient human resources to produce that knowledge at the same rate of its demand. Aiming to solve that semantic data scarcity problem, this work investigates how binary semantic relations can be automatically extracted from Brazilian Portuguese texts. These relations are based on Minsky's (1986) theory and are used to represent common sense knowledge in the *Open Mind Common Sense no Brasil* (OMCS-Br) project developed at LIA (*Laboratório de Interação Avançada*), partner of LaLiC (*Laboratório de Linguística Computacional*), where this research was conducted, both in *Universidade Federal de São Carlos* (UFSCar). The first strategies for this task were based on searching textual patterns in texts, where a certain textual expression indicates that there is a specific relation between two terms in a sentence. This approach has high precision but low recall, which led to the research of methods that use machine learning as their main model, encompassing techniques such as probabilistic and statistical classifiers and also kernel methods, which currently figure among the state of the art. Therefore, this work investigates, implements and evaluates some of these techniques in order to determine how and to which extent they can be applied to the automatic extraction of binary semantic relations in Portuguese texts. In that way, this work is an important step in the advancement of the state of the art in information extraction for the Portuguese language, which still lacks resources in the semantic area, and also advances the Portuguese language NLP scenario as a whole.

Lista de Figuras

Lista de Tabelas

Sumário

1 Introdução

A extração de informação (EI) é uma das áreas do Processamento de Língua Natural (PLN) na qual se busca obter informações estruturadas a partir de textos escritos em língua natural (COWIE; LEHNERT, 1996). O advento da internet e a consequente explosão na geração de conteúdo incentivam novos estudos nessa área, visando obter informação útil e computacionalmente utilizável a partir da enorme quantidade de novos dados que são gerados a cada instante.

Uma das subtarefas da EI é a extração de relações semânticas de textos, que pode ser definida como a extração de relações de significado, normalmente binárias, explícitas ou implícitas, entre entidades mencionadas em um texto. Por exemplo, a partir da sentença "A doença afetou animais como cachorros, gatos e ratos", podem ser extraídas três relações de hiponímia (relação is-a, ou é-um), que são is-a(cachorros, animais), is-a(gatos, animais) e is-a(ratos, animais). Da mesma forma, a partir da sentença "Os dentes do mais antigo orangotango", pode ser extraída a relação de meronímia (part-of, ou parte-de) part-of(orangotango, dentes), denotando que os dentes são uma parte do orangotango. Essa tarefa tem aplicações principalmente na recuperação de informação e na construção e melhoramento de ontologias e bases léxicas, que são recursos custosos e trabalhosos de se construir manualmente.

Existem basicamente duas estratégias para a extração de relações semânticas: a abordagem de padrões textuais, proposta inicialmente por (HEARST, 1992), e a abordagem de aprendizado de máquina, mais recente. A primeira é mais simples e se baseia na definição de padrões textuais que indicam que existe uma relação binária entre dois termos em um texto. No entanto, essa técnica tem alta especificidade devido ao uso de *templates* fixos, resultando em uma boa precisão na extração, mas baixa cobertura. Procurando superar essas limitações, pesquisadores passaram a utilizar abordagens de aprendizado de máquina, como classificadores *naive* Bayes (SNOW et al., 2005), regressão logística (MINTZ et al., 2009) e *Support Vector Machines* (ZELENKO et al., 2003; ZHANG et al., 2006). Nas duas abordagens podem ser utilizados recursos linguísticos tais como tesauros e *WordNets* para melhorar o desempenho da tarefa.

Neste contexto, esta dissertação apresenta os resultados da pesquisa realizada com vistas

à extração automática de relações semânticas binárias a partir de textos escritos em português. Nesse processo foram investigadas tanto técnicas baseadas em padrões textuais como métodos de aprendizado de máquina, inovando no uso de uma base de informações de Senso Comum para apoiar a extração de relações.

O Senso Comum pode ser definido como um conjunto de fatos e crenças compartilhados por um grupo de pessoas em um dado tempo e espaço (SINGH et al., 2002). Sentenças como "O céu é azul" e "Bolas são usadas para brincar" podem ser consideradas como senso comum, pois representam conhecimento compartilhado por muitas pessoas no mundo. O projeto *Open Mind Common Sense*[1] (OMCS), do *Massachusetts Institute of Technology* (MIT), busca coletar informações de Senso Comum pela internet, usando informações fornecidas pelo preenchimento de *templates* por voluntários – um exemplo de *template* é "Um navio de grande porte é usado para X", onde X é a informação de Senso Comum fornecida pelo usuário. Neste *template*, a informação sublinhada é variável e pode ser realimentada por dados já inseridos na base; o restante do *template* é fixo e criado especificamente para a extração de um tipo de relação, neste caso a relação used-for (usado-para). Dessa forma, seria extraída uma instância de relação binária representada como used-for(navio de grande porte, X).

Um dos braços do projeto OMCS ao redor do mundo é o *Opem Mind Common Sense* no Brasil[2] (OMCS-Br) mantido pelo Laboratório de Interação Avançada[3] (LIA), parceiro do Laboratório de Linguística Computacional[4] (LaLiC) no qual esta pesquisa se desenvolveu, ambos do Departamento de Computação da Universidade Federal de São Carlos. A base de senso comum do OMCS-Br foi utilizada em uma das estratégias investigadas neste trabalho, fornecendo sementes das 7 relações semânticas consideradas (descritas na próxima seção) para realizar a extração de novas instâncias de relações. A combinação da informação de Senso Comum com o PLN ainda é um tema pouco explorado e que poderia trazer benefícios para diversas aplicações. Algumas possibilidades são utilizar Senso Comum para customizar textos a fim de torná-los mais compreensíveis para um determinado grupo social ou traduzir gírias e regionalismos (SUGIYAMA et al., 2011).

Este trabalho, portanto, apresenta os resultados da investigação de duas abordagens para a extração de relações semânticas binárias: a de padrões textuais e a de aprendizado de máquina, usando como apoio a base de dados de Senso Comum do projeto OMCS-Br. Deste modo, este trabalho é um importante passo no avanço das pesquisas sobre extração automática de relações

[1] `http://openmind.media.mit.edu/`
[2] `http://www.sensocomum.ufscar.br/`
[3] `http://lia.dc.ufscar.br/`
[4] `http://www.lalic.dc.ufscar.br`

semânticas, área ainda em aberto e pouco explorada sobretudo para a língua portuguesa.

1.1 Relações semânticas

As relações semânticas extraídas automaticamente variam muito entre os trabalhos pesquisados, mas algumas das mais frequentemente consideradas são a hiponímia/hiperonímia (is-a) (HEARST, 1992, 1998; SNOW et al., 2005), meronímia/holonímia (part-of) (BERLAND; CHARNIAK, 1999; GIRJU, 2003), sinonímia (LIN et al., 2003) e causa/efeito (GIRJU; MOLDOVAN, 2002; GIRJU et al., 2003).

Neste trabalho, especificamente, visa-se a extração de relações semânticas derivadas da teoria de Minsky (1986), utilizadas pelo projeto *Open Mind Common Sense* no Brasil para expressar conhecimento de Senso Comum. De um total de 20 relações semânticas, as 7 apresentadas na Tabela 1.1 foram as selecionadas para investigação neste trabalho. As outras 13 relações utilizadas no projeto OMCS-Br, que não serão consideradas neste trabalho, são apresentadas na Tabela 1.2. Essas relações foram desconsideradas devido ao alto grau de abstração de seus significados, diferentemente das outras relações que envolvem conceitos mais concretos.

Tabela 1.1: Relações semânticas investigadas neste trabalho

	Relação semântica	**Sentença exemplo**	**Relação extraída**
1	location-of(algo/alguém, local)	Uma secretária pode ser encontrada em um escritório	location-of(secretária, escritório)
2	is-a(subclasse, superclasse)	Maçã é uma fruta	is-a(maçã, fruta)
3	property-of(algo/alguém, característica)	O prédio é alto	property-of(prédio, alto)
4	part-of(todo, parte)	Parafuso é uma parte de uma máquina	part-of(máquina, parafuso)
5	made-of(produto, substância)	Cacau é utilizado para fazer chocolate	made-of(chocolate, cacau)
6	effect-of(ação/estado, consequência)	Gripe causa febre	effect-of(gripe, febre)
7	used-for(entidade, função)	Pás são usadas para cavar	used-for(pás, cavar)

1.2 Motivação

Atualmente há uma demanda crescente por conhecimento semântico, não apenas em aplicações específicas de PLN mas também em outras como web semântica, que procura atribuir significado às páginas da internet, e busca na web[5]. Aplicações como extração de informação em documentos

[5] `http://www.google.com/`

Tabela 1.2: Relações semânticas não consideradas

	Relação semântica	**Sentença exemplo**	**Descrição**
1	defined-as(algo/alguém, sinônimo)	Uma fração é definida como a divisão de dois números inteiros	defined-as(fração, divisão de dois números inteiros)
2	capable-of(algo/alguém, habilidade)	Cangurus são capazes de saltar longe	capable-of(cangurus, saltar longe)
3	desirous-effect-of(ação, consequência)	Para ser aprovado deve-se estudar	desirous-effect-of(estudar, ser aprovado)
4	desire-of(alguém, desejo)	Toda pessoa quer ser feliz	desire-of(pessoa, ser feliz)
5	motivation-of(ação, motivação)	O que motiva fazer exercícios é a sensação de bem estar	motivation-of(fazer exercícios, sensação de bem estar)
6	first-subevent-of(evento, subevento)	A primeira coisa a ser feita para tomar banho é abrir a torneira	Um subevento é a primeira coisa a ocorrer na execução de um evento
7	subevent-of(evento, subevento)	Para respirar é necessário aspirar o ar para dentro dos pulmões	Um subevento ocorre no decorrer de um evento
8	last-subevent-of(evento, subevento)	Ao sair de casa, deve-se trancar a porta	Um subevento é a última coisa a ocorrer na execução de um evento
9	prerequisite-event-of(evento, pre-requisito)	Para entrar na universidade, é preciso ter o ensino médio concluído	Um pré-requisito é necessário para outro evento ocorrer
10	capable-of-receiving-action(algo/alguém, ação)	Uma bola pode ser arremessada	Algo ou alguém pode sofrer uma determinada ação
11	conceptually-related-to(X, Y)	Mau hálito é aliviado com uma bala de hortelã	Existe uma relação entre dois termos, mas ela não é conhecida (no exemplo, “mau hálito” tem alguma relação com “hortelã”)
12	thematic-k-line(X, Y)	–	Dois termos compartilham o mesmo tema, p. ex. “champanhe” e “fim de ano”, “brinquedos” e “infância”
13	super-thematic-k-line(X, Y)	–	Um termo é um supertema de outro, p. ex. “comprar” é um supertema para “comprar comida”

web (CARLSON et al., 2010), sistemas de perguntas e respostas e tradução automática para outros idiomas[6] também podem se beneficiar do conhecimento semântico contido nas relações abordadas por este trabalho. Porém, não há recursos humanos suficientes para produzir esse conhecimento na mesma velocidade de sua demanda.

Nesse cenário, este trabalho de mestrado surge como alternativa ao processo custoso de identificação manual de informações semânticas, pois se propõe a desenvolver sistemas que realizem a extração de relações semânticas automaticamente a partir de textos escritos em língua natural. Dessa forma, busca-se superar o gargalo existente devido à grande demanda por dados semânticos e a escassez de tal conhecimento e de mão de obra qualificada e disponível para gerá-lo em tempo hábil.

[6] `http://translate.google.com`

Além da utilidade do conhecimento gerado nesta pesquisa, outra motivação reside na escolha do idioma português que, comparado com outras línguas como o inglês, possui ainda poucos recursos semânticos. Assim, este trabalho é motivado pela necessidade crescente de geração de recursos semânticos aliada à falta dos mesmos para o idioma português. Além disso, busca-se, com este trabalho, avançar o estado da arte em extração de relações semânticas a partir de textos livres, área cujos sistemas ainda não têm alto nível de performance (GIRJU et al., 2010). Por fim, vale salientar que o uso de conhecimento de Senso Comum no PLN ainda é um tema pouco explorado e que poderia trazer um impacto positivo em várias aplicações.

1.3 Objetivos

O objetivo deste trabalho de mestrado é verificar como e em que medida métodos automáticos podem ser aplicados para a extração de relações semânticas binárias em textos escritos em português.

Esse objetivo foi buscado por meio da investigação de técnicas e métodos propostos na literatura para realizar a tarefa em questão, avaliando-os, comparando-os e adaptando-os para o português. Diferentes abordagens linguísticas e de aprendizado de máquina foram investigadas e comparadas quantitativa e qualitativamente, buscando avançar o estado da arte em extração automática de relações semânticas a partir de textos, além de contribuir para o cenário de PLN do português de uma forma geral.

1.4 Organização do texto

O restante deste documento está estruturado como se segue.

O Capítulo 2 apresenta a revisão bibliográfica sobre o tema de extração automática de relações semânticas a partir de textos. São explorados os principais trabalhos da área envolvendo as abordagens linguística e de aprendizado de máquina. São apresentadas pesquisas de estado da arte, suas técnicas de extração, *corpora* de treinamento, métodos de avaliação e resultados encontrados. Também é feito um levantamento sobre os trabalhos que têm a língua portuguesa como foco principal.

O Capítulo 3 trata da metodologia de desenvolvimento do trabalho e dos experimentos desenvolvidos no decorrer do projeto de mestrado. São delineados os métodos, ferramentas e recursos utilizados para a produção de sistemas que realizam automaticamente a tarefa de extração de relações semânticas a partir de textos. Com base na revisão bibliográfica, foram implementa-

das estratégias que utilizam padrões textuais e aprendizado de máquina, com diferentes algoritmos de aprendizagem. Essas implementações, juntamente com suas avaliações, possibilitaram a comparação entre as técnicas investigadas e permitem a verificação da melhor abordagem para a extração automática de relações semânticas para textos escritos em português.

O Capítulo 4 apresenta os resultados obtidos com a aplicação dos métodos citados no capítulo anterior e traz uma discussão sobre os mesmos, comparando as estratégias utilizadas e o desempenho de cada uma.

Por fim, o Capítulo 5 apresenta as considerações finais deste trabalho, enfatizando os resultados obtidos e a contribuição desta pesquisa para o cenário de PLN do português e internacional, além de sugestões e apontamentos para trabahos futuros.

2 Revisão Bibliográfica

Este capítulo tem o propósito de contextualizar o leitor sobre os principais conceitos, métodos e estratégias encontrados na literatura científica a respeito da extração automática de relações semânticas a partir de textos. Serão apresentadas as estratégias/abordagens mais utilizadas para realizar a tarefa (Seção 2.1) e trabalhos que têm foco na língua portuguesa (Seção 2.2).

2.1 Estratégias de extração automática de relações semânticas

Não há uma categorização clara das abordagens para extração automática de relações semânticas, mas, considerando-se os métodos apresentados na literatura, elas podem ser divididas grosso modo em dois grupos não mutuamente exclusivos: as que usam padrões textuais e as que usam técnicas de aprendizado de máquina, ambas detalhadas nas subseções a seguir. Todos os trabalhos apresentados nesta seção têm como foco a língua inglesa; trabalhos desenvolvidos especificamente para a língua portuguesa estão descritos na Seção 2.2.

Os trabalhos descritos a seguir são:

- Padrões textuais – Na Seção 2.1.1, (HEARST, 1992, 1998), (BERLAND; CHARNIAK, 1999), (BRIN, 1999), (AGICHTEIN; GRAVANO, 2000) e (GIRJU; MOLDOVAN, 2002) para a língua inglesa. Na Seção 2.2, (FREITAS; QUENTAL, 2007), (OLIVEIRA et al., 2010), (CARDOSO, 2008), (CHAVES, 2008) e (BRUCKSCHEN et al., 2008) para a língua portuguesa.
- Aprendizado de máquina – Na Seção 2.1.2, (CARABALLO, 1999), (GIRJU, 2003; GIRJU et al., 2003), (ZELENKO et al., 2003), (SNOW et al., 2005), (ZHANG et al., 2006), (YAP; BALDWIN, 2009), (GIRJU et al., 2010) e (CARLSON et al., 2010), todos para a língua inglesa.

2.1.1 Abordagem de padrões textuais

As propostas mais antigas e mais simples são as que realizam a extração de relações semânticas com base em padrões textuais. Em outras palavras, esses estudos fazem uso de padrões textuais como indicativos (pistas) de que determinada construção textual denota uma relação entre duas entidades.

Hearst (1992, 1998)

Hearst (1992) foi uma das primeiras a explorar a abordagem de padrões textuais. Um dos padrões definidos em seu trabalho é apresentado na Figura 2.1, na forma de expressão regular[1].

Figura 2.1: Um dos padrões definidos por Hearst (1992)

$$NP_1\{,\}\ especially\ \{NP_2, NP_3\ ...\}\ \{or \mid and\}\ NP_n$$

Esse padrão indica uma relação de hiponímia, também conhecida como relação is-a (é-um), ou seja, uma relação onde uma entidade é um subtipo de outra. Nesse caso, a aplicação do padrão a uma sentença permite inferir uma relação is-a(NP_i, NP_1) (com $i = 2, ..., n$). Por exemplo, a aplicação desse padrão à frase "*most countries, especially France, England and Spain*" ("a maioria dos países, especialmente França, Inglaterra e Espanha"), permite a inferência de três instâncias da relação is-a: is-a(France, country), is-a(England, country), e is-a(Spain, country). Outros exemplos de padrões definidos em (HEARST, 1992) podem ser vistos na Figura 2.2.

Esta abordagem necessita como entrada apenas de textos processados com *shallow parsing*, ou processamento sintático raso, já que os padrões se atêm às formas superficiais das sentenças (as palavras da sentença como aparecem no texto) e requerem apenas a identificação de sintagmas nominais.

Em (HEARST, 1992, 1998; BERLAND; CHARNIAK, 1999; GIRJU; MOLDOVAN, 2002) esses padrões são construídos manualmente ou semiautomaticamente por meio da observação do *corpus*[2] utilizado. O procedimento descrito em Hearst (1992, 1998) para a extração semiautomática de relações consiste basicamente em 5 passos apresentados na Figura 2.3.

Assim, a posterior aplicação dos novos padrões encontrados sobre o *corpus* permite a descoberta de outros novos padrões, sendo um procedimento iterativo.

[1] No padrão apresentado como exemplo, NP denota um sintagma nominal (*noun phrase*), { e } representam a repetição de 0 ou mais vezes do padrão entre as chaves e | indica uma opção de escolha entre valores.

[2] Um *corpus* é um conjunto de textos escritos em língua natural. Seu plural é *corpora*.

Figura 2.2: Outros padrões definidos por Hearst (1992)

1. NP_0 such as $\{ NP_1, NP_2, ..., (\text{and} \mid \text{or}) \} NP_n$
   Frase exemplo: *"The bow lute, such as the Bambara ndang, is ..."*
   Relação extraída: is-a("Bambara ndang", "bow lute")

2. *such NP as* $\{NP ,\}^* \{(or \mid and)\}$ *NP*
   Frase exemplo: *"... works by such authors as Herrick, Goldsmith and Shakespeare"*
   Relações extraídas: is-a("Herrick", "author"), is-a("Goldsmith", "author")

3. *NP* $\{, NP\}^*$ $\{,\}$ *or other NP*
   Frase exemplo: *"Bruises, wounds, broken bones or other injuries ..."*
   Relações extraídas: is-a("bruise", "injury"), is-a("wound", "injury"), is-a("broken bone", "injury")

4. *NP* $\{, NP\}^*$ $\{,\}$ *and other NP*
   Frase exemplo: *"... temples, treasuries, and other important civic buildings"*
   Relações extraídas: is-a("temple", "civic building"), is-a("treasury", "civic building")

5. *NP* $\{,\}$ *including* $\{NP ,\}^* \{or \mid and\}$ *NP*
   Frase exemplo: *"All common-law countries, including Canada and England"*
   Relações extraídas: is-a("Canada", "common-law country"), is-a("England", "common-law country")

Figura 2.3: Algoritmo de Hearst (1992)

1. Primeiramente, uma relação semântica de interesse é escolhida (p. ex. hiponímia, meronímia, etc.);

2. Constrói-se uma lista de termos para os quais se sabe que a relação é válida (p. ex. "Brasil-país" e "cachorro-animal", para a relação de hiponímia). Essa lista pode ser obtida por meio da aplicação de padrões feitos manualmente ou a partir de alguma base léxica ou de conhecimento pré-existente;

3. Então, procura-se no *corpus* por sentenças em que esses termos ocorrem sintaticamente próximos e armazena-se o contexto (palavras ao redor ou sentença) em que eles aparecem (p. ex. "O Brasil é um país em desenvolvimento.");

4. A seguir, procura-se por similaridades entre esses contextos e hipotetiza-se que contextos comuns indicam a relação de interesse;

5. Quando um padrão é identificado positivamente, ele é utilizado para encontrar mais instâncias da relação alvo e volta-se ao passo 2.

Para avaliar sua proposta, Hearst aplicou os seus padrões textuais sobre um *corpus* jor-

nalístico contendo seis meses de texto do *New York Times*. Os resultados obtidos foram validados comparando as relações extraídas com os dados presentes na WordNet (versão 1.5) (FELLBAUM, 1998). Nessa avaliação, se a relação e os termos da relação já estivessem presentes na WordNet, a relação era considerada correta. Se a relação ou algum dos termos não estivesse presente na base de conhecimento, era feita uma avaliação subjetiva da qualidade da relação, considerando-a como uma candidata "forte", "boa", ou "ruim" para inclusão na base léxica – essa classificação foi feita pela própria autora comparando as relações candidatas com as relações já existentes na WordNet. Utilizando o padrão *"or other"* ("*NP* {, *NP*}* {,} *or other NP*"), a autora reporta que, de uma amostra de 200 sentenças que continham os termos *"or other"*, foram extraídas 166 relações válidas (sem erros sintáticos e não eram repetições de outras relações), e dessas, 104 (cerca de 63% das relações válidas) já estavam presentes ou eram fortes candidatas à inclusão na WordNet.

A principal diferença entre (HEARST, 1992) e (HEARST, 1998) consiste no *corpus* utilizado na avaliação experimental: o trabalho de 1992 utiliza cerca de 8,6 milhões de palavras de textos da *Grolier's American Academic Encyclopedia*, enquanto o de 1998 utiliza um conjunto de 6 meses de textos do jornal *New York Times*.

Berland e Charniak (1999)

Berland e Charniak (1999) utilizam o algoritmo de Hearst mas procuram pela relação de meronímia (part-of ou parte-de) em um *corpus* do domínio jornalístico com cerca de 100 milhões de palavras etiquetadas com etiquetas *part-of-speech* (POS). Dois padrões que indicam essa relação foram considerados: "$N's\ N$" e "$N\ (of|the)\ N$" (onde N representa um substantivo). Embora esses padrões sejam ambíguos (HEARST, 1992; GIRJU, 2003), podendo indicar outras relações semânticas como posse (p. ex. "*The girl's car*" – "O carro da menina"), o uso de métricas estatísticas e o tamanho do *corpus* possibilitaram resultados satisfatórios. As métricas estatísticas consistem em ordenar as relações extraídas de acordo com a probabilidade de realmente serem um par que indica meronímia, com base na contagem de coocorrência das palavras no *corpus* e no número de correspondências com os padrões. A avaliação dos 50 merônimos mais prováveis encontrados para seis palavras ("*book*", "*building*", "*car*", "*hospital*", "*plant*", "*school*") feita por cinco humanos mostrou que, em média, 55% deles estavam corretos.

DIPRE (BRIN, 1999)

Brin (1999) propõe o DIPRE (*Dual Iterative Pattern Relation Extraction*), um método para extrair conhecimento de milhares de documentos da web. O algoritmo é similar ao de Hearst, iniciando

com um pequeno conjunto de instâncias da relação que se deseja extrair, procurando trechos em que os termos das instâncias apareçam próximos, guardando os contextos, gerando novos padrões e recomeçando o processamento. A relação exemplificada no trabalho é a autor-livro, ou seja, um certo autor é o escritor de um certo livro, mas a abordagem pode ser utilizada para outras relações, bastando fornecer um conjunto de relações semente adequado. O *corpus* total disponível neste estudo consiste em 24 milhões de páginas da web, totalizando 147 gigabytes de dados, embora apenas um subconjunto desses dados tenha sido utilizado nos experimentos relatados.

A avaliação experimental feita por Brin (1999) consistiu em procurar a relação livro-autor, fornecendo ao sistema uma lista criada manualmente com apenas 5 relações semente. Após 5 iterações do sistema, usando um subconjunto de cerca de 5 milhões de páginas do *corpus*, chegou-se a uma lista de mais de 15000 relações. Alguns dos padrões encontrados pelo sistema foram "`<LI><B>`*`title`*`</B> by `*`author`*", "`<i>`*`title`*`</i> by `*`author`*" e "*`author || title ||`*" (os padrões são aplicados sobre o código-fonte das páginas web, por isso a presença das marcações HTML – elementos envoltos por "<" e ">"). Para avaliar a qualidade das relações extraídas, foram selecionadas 20 tuplas aleatórias e verificou-se na internet se eram realmente livros verdadeiros. O resultado foi que, dos 20 pares livro-autor, 19 estavam corretos, sendo que o incorreto na verdade era um artigo de revista.

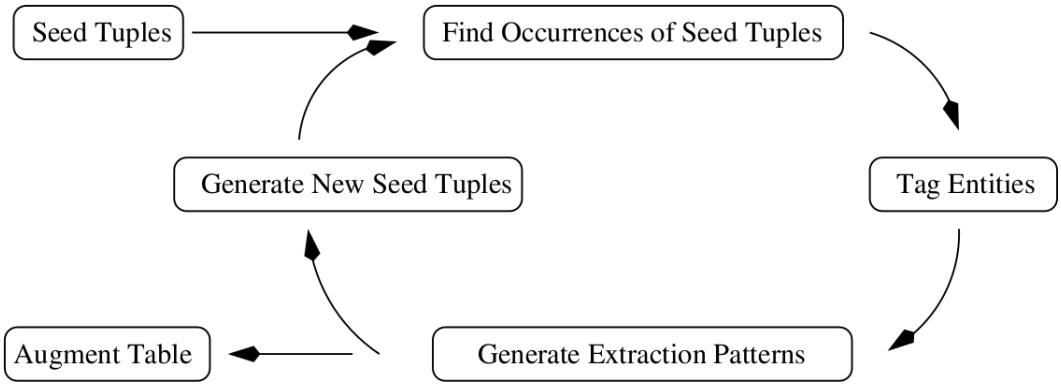
Snowball (AGICHTEIN; GRAVANO, 2000)

Agichtein e Gravano (2000) descrevem o *Snowball*, um método que expande o DIPRE citado anteriormente. Uma das melhorias sobre o DIPRE é a forma de gerar padrões: o *Snowball* faz uma etapa de reconhecimento de entidades nomeadas (*named entity recognition* – NER) de forma que os padrões gerados tenham restrições semânticas quanto à classe dos termos. Por exemplo, para a relação organização-local, usada como exemplo no artigo, um padrão que pode ser gerado é "<LOCATION>-based <ORGANIZATION>", ou seja, apenas termos etiquetados como localidade e como organização serão extraídos. Outros exemplos de padrões possíveis são "<ORGANIZATION>'s headquarters in <LOCATION>" e "<ORGANIZATION>, <LOCATION>".

A principal melhoria sobre o DIPRE é a etapa de avaliação dos padrões e das relações extraídas: de forma simplificada, um padrão é considerado confiável se as relações que encontra são consistentes com as relações já constantes na tabela de relações semente; de forma análoga, uma relação é considerada confiável se vários padrões distintos a encontram no *corpus*. Essa medida de confiança impede que padrões e relações não confiáveis sejam passados para iterações seguintes do algoritmo, gerando cada vez mais erros. A Figura 2.4 mostra uma visão geral do

sistema.

Figura 2.4: Visão geral do sistema *Snowball* (AGICHTEIN; GRAVANO, 2000, p. 87)



A avaliação do Snowball foi feita utilizando um *corpus* composto por cerca de 320 mil artigos jornalísticos, divididos em 178 mil documentos para treinamento e 142 mil documentos para teste. A avaliação dos resultados foi feita comparando-se as relações extraídas com uma tabela de organizações e suas respectivas localidades. Inicialmente o sistema contava com apenas 5 relações semente feitas manualmente; o número total de relações extraídas não foi publicado, mas uma avaliação manual de uma amostra de 100 relações mostrou que 52% das relações estavam corretas. Se um dos parâmetros do sistema, o limiar de confiabilidade das relações extraídas, era colocado em 0,8, a precisão subia para 93%, à custa de uma redução acentuada na cobertura.

Girju e Moldovan (2002)

Girju e Moldovan (2002) estudam a extração da relação de causalidade (causa-efeito, descrita na Tabela 1.1 como effect-of) e utilizam o mesmo processo de descoberta de padrões de Hearst (1998), porém, utilizando a WordNet versão 1.7 (FELLBAUM, 1998) para a obtenção da lista de termos relacionados (passo 2) e para desambiguação de verbos polissêmicos. O padrão utilizado para a procura de relações tem a forma "*NP verbo/expressao verbal NP*".

A avaliação dessa proposta foi feita utilizando-se um *corpus* de 3GB de textos jornalísticos (provenientes do *Wall Street Journal*, *Financial Times*, entre outros) e uma lista de 60 verbos que exprimem causalidade. Assim, foram selecionadas 50 sentenças do *corpus* que continham cada verbo da lista, formando um novo *corpus* contendo 3000 sentenças que foram analisadas sintaticamente e etiquetadas com etiquetas de *part-of-speech*. De 1321 relacionamentos obtidos, foi selecionada uma amostra de 300 para serem avaliados manualmente por dois anotadores humanos. A precisão média obtida foi de 65,6%, um pouco melhor que a de Hearst. No entanto, a

comparação deve ser cuidadosa, pois os dois trabalhos têm foco em relações distintas (causalidade e hiponímia), e Girju et al. (2010) sugerem que relações distintas têm dificuldades diferentes para serem extraídas, além dos métodos de avaliação serem distintos.

Deficiências da abordagem baseada em padrões

Uma das deficiências da abordagem baseada em padrões textuais é que, devido à sua especificidade, eles têm alta precisão, mas baixa cobertura (YAP; BALDWIN, 2009). Além disso, alguns padrões, como por exemplo "*NP*, *such as NP*" (para sentenças em inglês), definem com alta confiabilidade uma relação semântica, neste caso a hiponímia (is-a); no entanto, outros padrões, como por exemplo "$NP\ of\ NP$", um dos mais encontrados para expressar relações de meronímia (part-of), são ambíguos e podem ser associados a mais de uma relação (GIRJU et al., 2010). Para superar essas limitações, estudos mais recentes passaram a utilizar informações linguísticas mais profundas e variadas em conjunto com métodos de aprendizado de máquina.

2.1.2 Abordagem de aprendizado de máquina

Diferente das técnicas baseadas em padrões textuais, que utilizam pouco conhecimento linguístico, as técnicas baseadas em aprendizado de máquina fazem uso de atributos (*features*[3]) variados, que podem ser tanto léxicos, quanto sintáticos ou semânticos. Essas anotações podem ser produzidas por ferramentas linguísticas como etiquetadores de *part-of-speech* (POS) e analisadores sintáticos (*parsers*).

Essas *features* são fornecidas a classificadores que, utilizando métodos probabilísticos ou estatísticos, são capazes de predizer se uma relação ocorre entre termos de uma sentença. Esses classificadores podem ser não supervisionados (*clustering*) (CARABALLO, 1999) ou supervisionados como árvores de decisão (GIRJU et al., 2006), *naive* Bayes e regressão logística (SNOW et al., 2005), etc.

Alguns dos classificadores muito utilizados e que figuram entre o estado da arte são as *Support Vector Machines* (SVMs) (VAPNIK, 1995). As SVMs são classificadores binários que usam propriedades geométricas para encontrar um hiperplano que melhor separe amostras de treinamento em duas classes. Uma das características mais úteis das SVMs é que elas podem utilizar funções de *kernel*, que são basicamente funções que computam a similaridade entre duas instâncias

[3] *Features* são atributos que descrevem uma instância; elas são utilizadas por métodos de aprendizado de máquina como forma de generalizar e discriminar instâncias. Por exemplo, algumas *features* de uma maçã são sua cor, tamanho e massa.

quaisquer. Diferentes funções de *kernel* permitem que o hiperplano tenha diferentes formatos, não apenas lineares – dessa forma, o classificador gerado não precisa ser linear, sendo essa uma técnica bastante versátil. Existem outros classificadores que podem utilizar funções de *kernel*, como por exemplo os *Voted Perceptrons* (FREUND; SCHAPIRE, 1999); a ideia por trás destes é bastante parecida com a das SVMs, procurando encontrar a maior margem que separe dois conjuntos de dados. Uma das vantagens deste método em relação às SVMs é sua maior simplicidade para implementação e maior eficiência computacional, embora à custa de uma piora na performance na classificação.

Os métodos de *kernel* foram utilizados em estudos sobre a extração automática de relações semânticas (ZELENKO et al., 2003; ZHANG et al., 2006). Uma característica muito útil das funções de *kernel* é que elas podem explorar *features* estruturadas em suas representações originais (ZHANG et al., 2006). Isso significa que estruturas como árvores sintáticas não precisam ser modificadas – transformadas em um vetor de palavras, por exemplo – e dessa forma informações estruturais implícitas possivelmente importantes são mantidas.

Caraballo (1999)

Caraballo (1999) utiliza a técnica de aprendizado de máquina não supervisionado de agrupamento (*clustering*) sobre *corpora* jornalísticos do *Wall Street Journal* analisados sintaticamente, com o objetivo de encontrar relações de hiponímia. Inicialmente, todos os substantivos do *corpus* são considerados nodos distintos. Através do cálculo de similaridade distribucional entre os substantivos, os dois substantivos/nodos mais similares são agrupados sob um novo nodo pai comum a ambos, formando um grupo (*cluster*). Esse processo se repete até que todos os nodos estejam presentes na hierarquia. Essa hierarquia não é etiquetada, no sentido em que os substantivos estão agrupados, mas não se sabe claramente qual a razão para estarem agrupados dessa forma (esse é um dos vieses dos métodos de aprendizado não supervisionado). Desse modo, o próximo passo é etiquetar cada *cluster* com o hiperônimo que seria mais provável, de acordo com os substantivos que ele engloba. Essa verificação decorre da aplicação dos padrões "*NP* {, *NP*}* {,} *or other NP*" e "*NP* {, *NP*}* {,} *and other NP*" (HEARST, 1992) sobre o *corpus* e verificando qual hiperônimo ocorre com mais frequência para os substantivos do *cluster*.

A avaliação desse sistema foi feita selecionando-se uma amostra aleatória de 10 *clusters* da hierarquia, cada um englobando pelo menos 20 substantivos. Foi pedido a três juízes humanos que avaliassem se os *clusters* estavam etiquetados adequadamente com os hiperônimos correspondentes aos substantivos que constavam do *cluster*. De acordo com o critério de avaliação mais rígido, onde a maioria dos juízes deveria concordar, a precisão encontrada foi de 33%.

Girju et al. (2003), Girju (2003)

Girju et al. (2003) e Girju (2003) investigam, respectivamente, as relações de meronímia (part-of) e causalidade. Apesar de serem trabalhos separados, ambos são semelhantes em sua metodologia, utilizando a WordNet como recurso de conhecimento semântico e árvores de decisão C4.5 como algoritmo de aprendizado. Os *corpora* utilizados também são os mesmos, cobrindo textos do jornal *Los Angeles Times* e parte do SemCor (MILLER et al., 1993) – o SemCor é uma pequena parte do *English Brown Corpus* (KUCERA; FRANCIS, 1967) anotado semanticamente com os sentidos da WordNet (FELLBAUM, 1998).

Essencialmente, ambos os trabalhos executam o algoritmo de Hearst, primeiro coletando pares semente das respectivas relações de interesse (meronímia e causalidade) a partir da WordNet; em seguida, são procurados nos *corpora* sentenças em que esses termos apareçam próximos. Foram encontrados os padrões "*NP of PP*" (onde PP significa *prepositional phrase* – sintagma preposicional), "*NP's PP*" e "*NP verb NP*" para a relação de meronímia e "*NP verb NP*" para causalidade. Esses são padrões bastante ambíguos, no sentido em que podem expressar mais de uma relação semântica (como o padrão "*NP verb NP*" presente nos dois estudos). Para tratar essa ambiguidade, a ideia proposta nos artigos é definir restrições semânticas sobre os argumentos das relações, ou seja, definir pares de classes semânticas válidas para os participantes da relação de meronímia e causalidade. Essa etapa é feita da seguinte forma:

- Um conjunto de treinamento é construído manualmente contendo tuplas representando uma possível relação de meronímia ou causalidade e sua classificação em correta ou incorreta: p. ex. "<aria, opera, yes>", indicando que ária é parte de uma ópera;
- Esses exemplos de treinamento são generalizados substituindo-se os termos participantes pelos seus sentidos mais gerais da Wordnet: p. ex. <aria, opera, yes> se torna <entity#1, abstraction#6, yes>, sendo entity#1 o sentido mais geral de "aria" e abstraction#6 o sentido mais geral de "opera".

Os algoritmos de árvore de decisão são então treinados sobre os exemplos de treinamento generalizados. Na etapa de testes o sistema faz o mesmo processo de generalização sobre as relações candidatas extraídas para poder avaliá-las.

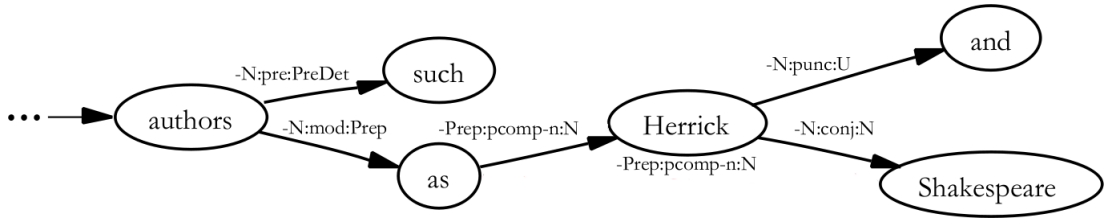
Girju et al. (2003) construíram, para a relação de meronímia, um conjunto de 81580 relações de treinamento, sendo 34609 exemplos positivos e 46971 negativos. O sistema foi avaliado sobre 10000 sentenças do *corpus* do *Los Angeles Times*, que continham 162 relações de

meronímia. O sistema recuperou 140 relações das quais 117 eram verdadeiramente meronímia, resultando em uma precisão de 83% e cobertura de 72%. Já em (GIRJU, 2003), que explora a relação de causalidade, o conjunto de treinamento continha 6523 relações, das quais 2101 tinham etiqueta positiva e 4422 negativa. O conjunto de teste continha 153 instâncias. O sistema recuperou 138 relações, sendo 102 identificando corretamente a relação procurada, resultando em uma precisão de 73,91% e cobertura de 66,6%.

Snow et al. (2005)

Snow et al. (2005) investigam a extração de relações de hiperonímia. Sua proposta tem basicamente a mesma forma que o procedimento desenvolvido por Hearst (1992). O *parser* de dependência utilizado foi o MINIPAR (LIN, 1998), que gera árvores de dependência, as quais indicam as dependências sintáticas existentes entre cada termo da sentença. Um trecho de árvore de dependência gerada pelo MINIPAR pode ser visto na Figura 2.5. As arestas da árvore são representadas no seguinte formato: (word1, CATEGORY1 : RELATION : CATEGORY2, word2), onde word1 e word2 são as formas lematizadas/base das palavras relacionadas, e correspondem a nodos da árvore de dependência; CATEGORY 1 e 2 são as etiquetas de *part-of-speech* de cada palavra (p. ex. N para substantivo, PREP para preposição); e RELATION é a relação sintática existente entre os termos (p. ex. OBJ para objeto, MOD para modificador ou CONJ para conjunção). Seguindo o exemplo da Figura 2.5, a relação entre "*authors*" e "*as*" pode ser representada como "author, -N : mod : Prep, as", indicando que "*author*" é um substantivo (N), "*as*" é uma preposição (Prep) e a relação existente entre eles é de modificador (mod). Um caminho de dependência é definido como uma sequência de arestas separadas por vírgula. Os padrões textuais, usados nas abordagens da seção anterior, podem ser representados como caminhos de dependência, como pode ser observado na Tabela 2.1.

Figura 2.5: Trecho de árvore de dependência gerada pelo MINIPAR (LIN, 1998) – Figura modificada de Snow et al. (2005, p. 1298)



O *corpus* utilizado neste estudo é composto por mais de 6 milhões de sentenças de artigos jornalísticos. Para a construção do conjunto de treinamento, primeiramente todas as sentenças

Tabela 2.1: Alguns caminhos de dependência utilizados em Snow et al. (2005)

Padrão	Caminho de dependência
NP like NP	N:PCOMP-N:PREP,like,like,PREP:MOD:N
NP called NP	N:DESC:V,call,call,V:VREL:N
NP is a NP	N:S:VBE,be,be,-VBE:PRED:N
NP, a NP (appositive)	N:APPO:N

do *corpus* foram processadas pelo MINIPAR e todos os pares de substantivos de cada sentença foram extraídos. De todo o conjunto de pares extraídos, 752311 foram selecionados para serem etiquetados como exemplos de pares hiperônimos ou não hiperônimos, de acordo com a WordNet. Desse total, 14387 foram classificados como hiperônimos e o restante como não hiperônimos.

As *features* utilizadas nesse trabalho são 69592 caminhos de dependência, que correspondem a todos os caminhos de dependência que ocorrem entre pelo menos 5 diferentes pares de substantivos no *corpus*. Então, para cada par de substantivos no *corpus* é construído um vetor de *features* com 69592 dimensões, cada uma correspondendo ao número de vezes que um determinado caminho de dependência foi o menor caminho a ligar o par de substantivos em questão no *corpus*. Assim, a tarefa em questão é definida como a classificação binária de um par de substantivos em hipônimo ou não hipônimo baseada no seu vetor de caminhos de dependência.

Foram testados classificadores *naive* Bayes e regressão logística. A etapa de testes foi feita através de *10-fold cross-validation* – os dados etiquetados foram divididos aleatoriamente em 10 partes, sendo 9 utilizadas para o treinamento dos classificadores e 1 para teste, e o procedimento é repetido 10 vezes alternando as partições. Os padrões de Hearst (1992) foram codificados na forma de caminhos de dependência (Tabela 2.2) e testados por um classificador simples, que apenas checa se o caminho está presente ou não entre o par de substantivos avaliado, para possibilitar uma comparação entre os resultados.

Tabela 2.2: Padrões de Hearst (1992) codificados como caminhos de dependência

Padrão	Caminho de dependência
NPX and other NPY	(and,U:PUNC:N),-N:CONJ:N, (other,A:MOD:N)
NPX or other NPY	(or,U:PUNC:N),-N:CONJ:N, (other,A:MOD:N)
NPY such as NPX	N:PCOMP-N:PREP,such as,such as,PREP:MOD:N
Such NPY as NPX	N:PCOMP-N:PREP,as,as,PREP:MOD:N,(such,PREDET:PRE:N)
NPY including NPX	N:OBJ:V,include,include,V:I:C,dummy node,dummy node,C:REL:N
NPY , especially NPX	-N:APPO:N,(especially,A:APPO-MOD:N)

O melhor resultado foi obtido utilizando o classificador de regressão logística, que obteve uma medida-F[4]. de 34,8%, enquanto o classificador *naive* Bayes obteve 31,7%. Em comparação, os padrões de Hearst obtiveram uma medida-F de 15%.

Zelenko et al. (2003)

Zelenko et al. (2003) é um dos trabalhos da literatura cujo algoritmo de aprendizado se baseia em *kernel*. Os autores utilizam um *parse tree kernel* com os classificadores SVM e *Voted Perceptron* para encontrar relações semânticas do tipo pessoa-filiação (uma dada pessoa está empregada em uma certa empresa) e organização-local (uma empresa tem sede em determinado local). O *corpus* utilizado é formado por cerca de 200 artigos jornalísticos processados por *shallow parsing*. Foram gerados *shallow parses* para treinamento e teste, que foram etiquetados manualmente como sendo exemplos positivos ou negativos para as duas relações. Ao total foram etiquetados 5439 *shallow parses*, sendo 3524 para a relação pessoa-filiação (1262 positivos e 262 negativos) e 1915 para a relação organização-local (506 positivos e 1409 negativos).

Conforme citado anteriormente, uma das vantagens dos métodos que usam *kernel* é permitir o uso de *features* estruturadas sem a necessidade de ter que transformá-las em *features* numéricas; isso é feito com a definição da função de *kernel*, que é uma função da forma $K : A \times A \rightarrow \mathbb{R}^+$. Essa função recebe dois objetos $x, y \in A$, sendo A o conjunto de objetos possíveis, gerando um número real $K(x, y)$ que representa a similaridade entre os dois objetos. A definição dessa função é genérica; no caso do *parse tree kernel* utilizado nesta abordagem, a função de *kernel* é definida recursivamente de forma que, considerando x e y duas árvores de *shallow parse*, todos os nodos e sequências de filhos de a são comparados com todos os nodos e sequências de filhos de b. A função de *kernel* irá então, simplificadamente, retornar o quão similares as duas árvores são. Desse modo, ao invés de usar certas características dos dados de treinamento como *features* – como feito por Snow et al. (2005) que usam como *features* os possíveis caminhos de dependência entre dois substantivos – é possível usar a própria estrutura de árvore dos dados como *feature*.

O classificador SVM com *parse tree kernel* foi comparado com outros métodos de aprendizado de máquina como *naive* Bayes e o classificador *Voted Perceptron* com o *parse tree kernel*. A avaliação foi feita treinando os classificadores sobre os exemplos etiquetados divididos aleatoriamente em um conjunto de treinamento (60% dos exemplos) e um conjunto de teste (40% dos exemplos). Os resultados finais são a média do resultado obtido pelos classificadores sobre 10 partições aleatórias dos dados. O classificador SVM com *parse tree kernel* foi o que obteve melhor desempenho para ambas as relações de pessoa-filiação e organização-local, com medidas-F

[4]A medida-F é a média harmônica entre precisão e cobertura

de 86,8% e 83,3%, respectivamente. O classificador *Voted Perceptron* com *parse tree kernel* teve medidas-F de 85,61% e 80,05% e o *naive* Bayes resultou em 82,93% e 80,04%.

Zhang et al. (2006)

Zhang et al. (2006) é outro trabalho que se baseia em *kernels*. Os autores utilizam um classificador SVM com um *kernel* composto, formado por um *kernel* de entidade e um *parse tree kernel*. Segundo os autores desse estudo, o *kernel* composto é bastante eficaz para a extração de relações e pode ser facilmente adaptado para cobrir mais conhecimento através da adição de mais *kernels*.

Os *corpora* utilizados nesse estudo foram o *corpus* do evento ACE de 2003, composto por um conjunto de treinamento de 674 documentos e 9683 instâncias de relações e um conjunto de testes de 97 documentos e 1386 instâncias de relações, e o *corpus* do evento ACE de 2004, composto por 451 documentos e 5702 instâncias de relações. São considerados 7 tipos de relações semânticas gerais, que são (1) relações físicas (localização, proximidade, parte-todo), (2) relações pessoais/sociais (parentesco), (3) emprego/pertinência/subsidiária (contexto empresarial/organizacional), (4) agente-artefato, (5) afiliação entre pessoa e organização (etnicidade, ideologia), (6) afiliação com local (residência em local, organização baseada em algum local) e (7) discurso (relações anafóricas). Ambos os *corpora* foram analisados sintaticamente.

O *kernel* de entidades foi construído para tirar proveito de *features* relativas às entidades nomeadas nos *corpora* do ACE 2003 e 2004 e é definido como um *kernel* simples, que retorna o número de *features* em comum entre duas entidades. O *parse tree kernel*, embora diferente do definido por Zelenko et al. (2003), tem um funcionamento similar, retornando o número de subárvores em comum entre duas árvores de *parsing* como medida de similaridade entre elas.

Os melhores resultados obtidos pelo classificador SVM com o *kernel* composto em termos de precisão, cobertura e medida-F, considerando a média dos resultados obtidos para cada tipo de relação, são 76,1%, 68,4% e 72,1%, respectivamente, sobre o *corpus* do ACE 2004. Esses valores são melhores que os reportados por estudos anteriores sobre os mesmos dados, mostrando que o *kernel* composto utilizado nessa abordagem é bastante eficaz para a extração de relações semânticas.

Mintz et al. (2009)

Mintz et al. (2009) propõem uma abordagem de aprendizado de máquina chamada *distant supervision*, que une algumas vantagens dos três tipos de aprendizado – supervisionado (ZELENKO et al., 2003; ZHANG et al., 2006), não supervisionado (CARABALLO, 1999) e semissupervisionado

(BRIN, 1999; AGICHTEIN; GRAVANO, 2000). O método proposto usa o Freebase[5], uma grande base semântica livre e online, para supervisionar a tarefa de extração de relações. Enquanto uma abordagem supervisionada é limitada pelo número de dados etiquetados disponíveis para treinamento, esta abordagem pode usar quantidades muito maiores de dados: são usados 1,2 milhões de artigos da Wikipedia[6] em inglês como *corpus* e 1,8 milhões de instâncias de 102 relações semânticas conectando 940 mil entidades distintas da Freebase como base de conhecimento.

A premissa da *distant supervision* é que, se uma sentença contém um par de entidades que participam em uma relação conhecida da base semântica, essa sentença possivelmente expressa essa relação de alguma forma. Como podem existir muitas sentenças que contenham um dado par de entidades – expressando ou não a relação semântica – pode ser extraído um grande número de *features*, possivelmente ruidosas, dessas sentenças. Essas *features* são combinadas em um classificador de regressão logística multiclasse que aprende pesos para cada *feature*. Por exemplo, considerando-se a relação de localidade e que na Freebase houvesse as relações (Virginia, Redmond) e (France, Nantes); ao encontrar sentenças como "*Richmond, the capital of Virginia*" ("Redmond, a capital da Virginia") e "*Henry's Edict of Nantes helped the Protestants of France*" ("O Edito de Nantes de Henrique ajudou os protestantes da França"), seriam extraídas *features* de ambas. Algumas das *features*, como as da primeira sentença, seriam mais úteis do que as da segunda. Na fase de testes, se fosse encontrada a sentença "*Vienna, the capital of Austria*" ("Viena, a capital da Áustria"), uma ou mais *features* dessa sentença iriam corresponder a *features* da sentença de Richmond, fornecendo evidências de que (Vienna, Austria) é uma instância da relação de localidade.

As *features* utilizadas são léxicas e sintáticas. As léxicas são referentes às entidades e às palavras que as circundam nas sentenças em que aparecem; algumas delas são a sequência de palavras entre as duas entidades, as etiquetas *part-of-speech* dessas palavras, e janelas de termos à esquerda e à direita das entidades. A *feature* sintática engloba o caminho de dependência entre as duas entidades, similar a Snow et al. (2005). Também é utilizado um reconhecedor de entidades nomeadas, fornecendo mais uma *feature* que são etiquetas de entidade nomeadas para cada palavra. Essa etiqueta pode ser pessoa, local, organização, miscelâneo ou nenhum.

A avaliação do sistema foi feita dividindo o conjunto de artigos da Wikipedia em 800 mil para treinamento e 400 mil para teste. São necessários exemplos negativos para treinar o classificador, então foi construído um vetor de *features* para uma relação "unrelated" ("não relacionado"), selecionando pares de entidades aleatórias que não apareciam em nenhuma relação da Freebase

[5] `http://www.freebase.com/`

[6] `http://en.wikipedia.org/` – a Wikipedia é uma enciclopédia online livre e colaborativa

e extraindo as *features* para as sentenças em que eles apareciam. A precisão obtida para as 100 instâncias mais confiáveis das 10 relações que ocorreram mais frequentemente nos dados de teste foi de 69%, segundo a avaliação feita por humanos através do serviço Mechanical Turk[7] da Amazon[8]. As 10 relações mais frequentes encontradas foram diretor-filme, roteirista-filme, países de bacias hidrográfias-rios, países-divisões administrativas, localidade, condado dos Estados Unidos-sede do condado, artista-origem, falecido-data de morte, pessoa-nacionalidade e pessoa-data de nascimento.

Yap e Baldwin (2009)

Entre os trabalhos que figuram como estado da arte na extração automática de relações semânticas, podemos citar (YAP; BALDWIN, 2009) e (GIRJU et al., 2010). Yap e Baldwin (2009) expandem o trabalho de Snow et al. (2005), extraindo relações de hiponímia e averiguando se as relações de sinonímia e antonímia também podem ser extraídas utilizando o mesmo método de Snow et al., além de explorar o impacto da escolha do *corpus* e o tamanho dos dados de treinamento. Foram utilizados dois conjuntos de textos nesse estudo, (1) o *English GigaWord corpus*, composto por textos jornalísticos, contendo cerca de 84 milhões de sentenças, e (2) o *dump* XML de julho de 2008 da Wikipedia em inglês, contendo cerca de 38 milhões de sentenças. O classificador utilizado nesse trabalho é do tipo *Support Vector Machine*, diferente de Snow et al. que utilizou *naive* Bayes e regressão logística.

Os resultados deste estudo mostram que, para a relação de hiponímia, Yap e Baldwin supera Snow et al., com medidas-F de 65,4% (*corpus* Gigaword) e 44,5% (*corpus* Wikipedia) contra 34,8% de Snow et al. (2005). Para as relações de sinonímia e antonímia, as medidas-F são 89,2% e 82% para o *corpus* Gigaword, e 89,5% e 86,1% para o *corpus* Wikipedia. No entanto, ao comparar o resultado para a relação de hiponímia com Snow et al., deve-se levar em conta que a quantidade de dados de treinamento utilizada por Yap e Baldwin é muito maior que a utilizada por Snow et al. (84 milhões e 38 milhões de sentenças contra 6 milhões), além dos métodos de aprendizado de máquina serem diferentes (SVMs contra *naive* Bayes e regressão logística).

Girju et al. (2010)

Girju et al. (2010) apresenta outro sistema de estado da arte para a tarefa de extração de relações semânticas. O sistema apresentado obteve o melhor desempenho em sua categoria na tarefa *Classification of Semantic Relations between Nominals* do SemEval 2007 (GIRJU et al., 2007). Esse

[7] `https://www.mturk.com/`
[8] `http://www.amazon.com/`

sistema utiliza conhecimento intensivo (*knowledge-intensive*) na forma de diversas ferramentas e recursos linguísticos, como um *tokenizer*, lematizador, etiquetador POS, analisador sintático, analisador de papéis semânticos, bases de conhecimento como a WordNet e dicionários construídos manualmente.

O *corpus* utilizado para treinamento não contém um número uniforme de instâncias de exemplo para cada uma das sete relações consideradas, sendo que elas se distribuem dessa forma: 1460 instâncias para causa-efeito, 140 para instrumento-agente, 973 para produto-produtor, 307 para origem-entidade, 231 para tema-ferramenta, 1143 para parte-todo e 140 para conteúdo-continente.

Uma das hipóteses defendidas por Girju et al. é que diferentes relações semânticas são mais bem definidas por determinadas *features*, portanto, é empregado um grande conjunto de atributos variados dos níveis léxico, sintático e semântico, como: a posição dos termos na relação; se um dos termos codifica informações de tempo ou espaço; o papel sintático dos termos (sujeito, objeto direto ou nenhum); a lista de etiquetas POS entre os termos da relação; entre outras. Essas *features* são empregadas no treinamento de sete classificadores binários SVM, um para cada relação considerada no SemEval.

Os resultados mostram valores de precisão, cobertura e medida-F, fazendo a média entre os resultados para as sete relações consideradas, de 79,7%, 69,8% e 72,4%. Embora os valores médios apresentados sejam menores que os de (YAP; BALDWIN, 2009), é importante ressaltar que as relações consideradas (hiponímia, sinonímia e antonímia para Yap e Baldwin contra sete relações distintas neste trabalho) e os *corpora* utilizados para treinamento são distintos (os *corpora* utilizados por Yap e Baldwin são consideravelmente maiores – dezenas de milhões contra alguns milhares de sentenças neste estudo), portanto comparações devem ser feitas criteriosamente.

NELL: Never-Ending Language Learning (CARLSON et al., 2010)

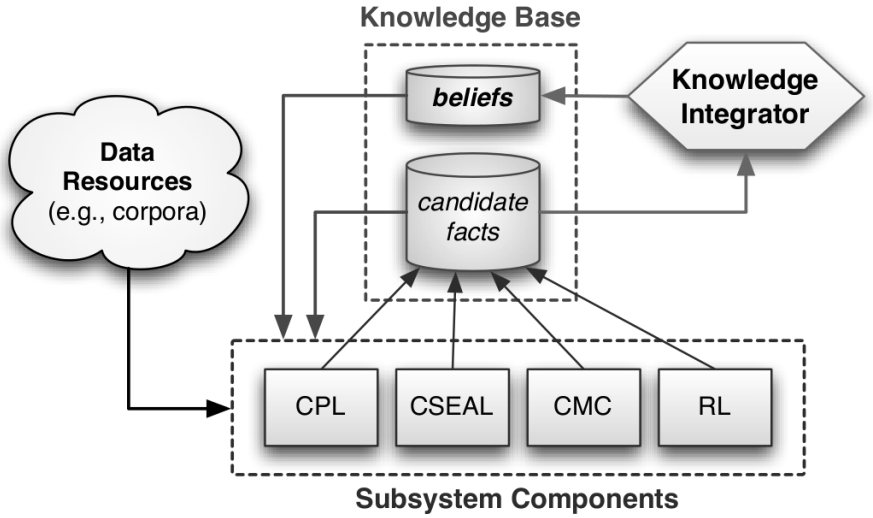
O projeto NELL[9] (*Never-Ending Language Learning*) visa o aprendizado contínuo e incremental a partir de dados da web. Para tanto, está baseado na execução de duas tarefas que são a *leitura* e o *aprendizado*. Enquanto a primeira visa a extração automática de informação de textos não estruturados da web para alimentar uma base estruturada de fatos, a segunda foca o aprendizado evolutivo, ou seja, o sistema deve aprender com a experiência, realizando a extração de informações de forma cada vez melhor (CARLSON et al., 2010). Para realizar as tarefas citadas são usados métodos de aprendizado semissupervisionado e uma variedade de métodos de extração de conhecimento.

[9] `http://rtw.ml.cmu.edu/rtw/`

O NELL coleta dois tipos de conhecimento: (i) a *categoria semântica* de sintagmas nominais (se são cidades, companhias ou times esportivos, por exemplo) e (ii) quais pares de sintagmas nominais fazem parte de quais *relações semânticas*, como hasOfficesIn(organization, location) (háEscritóriosEm(organização, local)). Esses dois tipos de informação – instâncias de categorias e instâncias de relações – são armazenados na peça central da proposta, que é a *knowledge base* (KB – base de conhecimento). A base de conhecimento segue a estrutura de uma ontologia, definindo as categorias semânticas que serão consideradas e relações válidas entre elas. O objetivo do NELL é expandir continuamente e automaticamente a KB, populando a ontologia através da leitura da web e do aprendizado contínuo para permitir uma melhor extração de conhecimento.

A base de conhecimento é alimentada e utilizada por um conjunto de módulos que são os responsáveis pela extração das instâncias de categorias e relações de páginas web. As instâncias candidatas extraídas por esses módulos vão para o conjunto de *candidate facts* (fatos candidatos) da KB, juntamente com uma medida de confiança do(s) módulo(s) que as extraiu (mais de um módulo pode ter extraído a mesma relação). Então, o *knowledge integrator* (KI – integrador de conhecimento) analisa esse conjunto de *candidate facts* e promove as instâncias que têm maior medida de confiança para o conjunto de *beliefs* (fatos consolidados) da base de conhecimento. Uma visão geral da arquitetura do NELL pode ser vista na Figura 2.6.

Figura 2.6: Arquitetura do NELL (CARLSON et al., 2010, p. 1307)



Dessa forma, o sistema funciona iterativamente: em cada iteração, todos os módulos são treinados e executados utilizando a KB atual, e em seguida o KI promove as relações mais confiáveis. Assim, a KB cresce em cada etapa, recebendo mais fatos consolidados que são utilizados para retreinar os módulos, que extraem mais fatos candidatos e assim sucessivamente. Um

dos problemas que podem surgir nesse tipo de ciclo é a propagação de erros quando uma instância incorreta é classificada como correta, gerando uma sucessão de conhecimento espúrio. Para amenizar esse risco, o sistema é supervisionado por humanos durante 10 a 15 minutos por dia, embora no experimento relatado essa supervisão tenha sido limitada.

A implementação do protótipo do NELL descrita em (CARLSON et al., 2010) foi composta por 4 módulos distintos de aprendizado e extração de informações de textos. A ideia em se utilizar diversos sistemas de aprendizado distintos juntos é combinar várias abordagens de forma que elas se complementem, possibilitando chegar a melhores resultados do que cada método conseguiria sozinho. Os 4 módulos utilizados foram:

- *Coupled Pattern Learner* (CPL) – Extrator de instâncias de categorias e relações a partir de padrões contextuais como os de Hearst (1992): por exemplo "*mayor of X*" ("prefeito de X") e "*X plays for Y*" ("X joga para Y"). O CPL usa medidas de coocorrência entre sintagmas nominais e contextos (ambos definidos como sequências de etiquetas POS) para definir novos padrões de extração.

- *Coupled SEAL* (CSEAL) – Extrai informações de fontes semiestruturadas de páginas da internet, como listas e tabelas. O CSEAL faz pesquisas na internet usando instâncias de categorias e relações da base de conhecimento para encontrar novas instâncias dessas categorias e relações.

- *Coupled Morphological Classifier* (CMC) – Conjunto de classificadores de regressão logística, um para cada categoria semântica da ontologia da KB, que classificam os sintagmas nominais de acordo com várias *features* morfológicas, como por exemplo a forma superficial das palavras do sintagma, presença de letras maiúsculas, afixos, POS, entre outras. Os fatos constantes na base de conhecimento são usados como treinamento para os classificadores, mas se restringindo às categorias e relações que têm pelo menos 100 instâncias no conjunto *beliefs*. O CMC também examina fatos candidatos propostos pelos outros módulos.

- *Rule Learner* (RL) – Algoritmo de aprendizado relacional baseado em lógica de primeira ordem, aprende cláusulas de Horn probabilísticas. Essas regras são usadas para inferir novas instâncias de relações a partir das relações constantes na base de conhecimento.

O experimento que avaliou esta primeira implementação do NELL teve a seguinte configuração: foi fornecida ao sistema uma ontologia inicial contendo 123 categorias semânticas, cada uma com 10 a 15 instâncias semente, e 5 padrões iniciais de extração para o módulo CPL.

Algumas das categorias semânticas da ontologia incluíam locais (p. ex. montanhas, lagos, cidades), pessoas (p. ex. cientistas, escritores, políticos), animais (p. ex. répteis, pássaros, mamíferos), organizações (p. ex. empresas, universidades), entre outras. A ontologia também definia 55 relações semânticas que também acompanhavam de 10 a 15 instâncias semente, mais 5 instâncias negativas. As relações englobavam relacionamentos entre as categorias da ontologia, como livro-escritor, companhia-produz-produto e time-joga-esporte.

Os módulos CPL, CSEAL e CMC foram executados uma vez por iteração, enquanto o RL foi executado uma vez a cada 10 iterações. As regras de inferência para encontrar novas relações propostas pelo RL foram filtradas manualmente por um humano.

Após 67 dias de execução, o sistema completou 66 iterações. Verificou-se que 242453 fatos haviam sido consolidados na base de conhecimento, sendo que desses, 95% (cerca de 230 mil) eram instâncias de categorias e 5% (cerca de 12 mil) instâncias de relações. A avaliação da precisão dos fatos foi feita selecionando-se uma amostra aleatória de 100 instâncias promovidas ao conjunto *beliefs* em três momentos diferentes da execução do NELL: das iterações 1 a 22, 23 a 44 e 45 a 66. Os 3 conjuntos de 100 instâncias foram avaliados por juízes humanos e chegou-se a uma precisão estimada média de 74%[10]. Alguns exemplos de relações extraídas são *cityInState(troy, Michigan)*, *musicArtistGenre(Nirvana, Grunge)* e *sportUsesEquip(soccer, balls)*.

2.1.3 Trabalhos relacionados

As duas abordagens citadas acima (baseada em padrões e em aprendizado de máquina) não são mutuamente exclusivas. Snow et al. (2005) e Yap e Baldwin (2009), por exemplo, realizam o aprendizado de padrões textuais por meio do uso de classificadores SVM.

Ambas as abordagens também podem utilizar conhecimento especializado proveniente de ontologias de domínio e bases léxicas. Na língua inglesa, um dos recursos mais utilizados nesse sentido é a WordNet (FELLBAUM, 1998), atualmente em sua terceira versão. A WordNet é uma base léxico-semântica extensa que contém palavras e seus sinônimos, organizadas em estruturas chamadas *synsets*, e suas relações de hiponímia e hiperonímia com outras palavras, definindo dessa forma uma estrutura hierárquica. Vários dos trabalhos levantados (HEARST, 1992, 1998; SNOW et al., 2005; YAP; BALDWIN, 2009; GIRJU; MOLDOVAN, 2002; GIRJU, 2003; GIRJU et al., 2010) utilizam a WordNet como um de seus principais componentes. Infelizmente, para o português do Brasil – idioma de interesse neste trabalho – ainda não existe um recurso dessa mag-

[10]Na avaliação relatada em (CARLSON et al., 2010), a precisão dada é geral, ou seja, a precisão não foi calculada separadamente para instâncias de categorias e instâncias de relações.

nitude e de cobertura tão ampla, dificultando a condução de trabalhos que utilizam conhecimento intensivo (*knowledge-rich approaches*). Alguns recursos próximos existentes para o português são o TeP (Thesaurus Eletrônico para o Português do Brasil) e o PAPEL (Palavras Associadas Porto Editora – Linguateca), que serão descritos no Capítulo 3 e na Seção 2.2, respectivamente.

Existe uma intersecção considerável entre o tema de extração automática de relações semânticas e trabalhos da área de bioinformática – o uso da computação para auxiliar em tarefas do domínio da biologia e medicina – como pode ser visto em trabalhos como (OAKES, 2005), (ROBERTS et al., 2008), (AHMED et al., 2010) e (SHANG et al., 2011). Algumas tarefas recorrentes são a extração de interações entre proteínas, genes e substâncias químicas e a identificação da relação entre determinados tratamentos e seus efeitos. Uma diferença importante desses trabalhos é o fato de que normalmente são utilizados dicionários de domínio para realizar a identificação dos termos que serão relacionados; ou seja, os termos que compõem as relações semânticas que se deseja extrair normalmente já são conhecidos de antemão, diferentemente do trabalho descrito nesta dissertação, que pode extrair relações entre termos desconhecidos.

2.2 Trabalhos com a língua portuguesa como foco

Todos os trabalhos descritos na seção anterior têm como língua-alvo o inglês. Em comparação com a língua inglesa existem poucos estudos sobre extração de relações semânticas com o português em primeiro plano. Alguns desses trabalhos serão descritos nesta seção.

Freitas e Quental (2007)

Um dos trabalhos com o português em foco é (FREITAS; QUENTAL, 2007), que adapta os padrões de Hearst (1992, 1998) para a língua portuguesa e propõe alguns novos, produzindo expressões regulares para extração de relações de hiperonímia entre termos.

Os padrões foram avaliados aplicando-os sobre um *corpus* de 11MB (1846502 palavras) de textos sobre saúde pública coletados da internet, etiquetados pelo *parser* PALAVRAS (BICK, 2000) na opção "*morphological tagging*" e então processados por um identificador de sintagmas nominais (SANTOS; OLIVEIRA, 2005). Os padrões construídos nesse estudo foram:

- Padrão "tais como"

  Esse padrão é análogo ao "*such as*" de Hearst (1992). "*such as*" pode ser traduzido como "tais como", embora apenas "como" seja frequentemente utilizado em português. Um pro-

blema da palavra "como" é que ela pode pertencer a diferentes classes gramaticais; o "como" que interessa a esse padrão é a palavra denotativa, equivalente a "por exemplo". Portanto, o "como" utilizado nos padrões é o como_PDEN (PDEN é uma etiqueta para palavra denotativa).

Um outro complicador é a ambiguidade existente em sentenças que contêm sintagmas preposicionais. Em estruturas como "[Infecções por bactérias] como [a Salmonella] e [a Shighella]" (sintagmas nominais representados entre colchetes) são extraídas as relações is-a(Salmonella, Infecções por bactérias) e is-a(Shighella, Infecções por bactérias), sendo que na verdade o hiperônimo é apenas "bactéria". A solução para esse problema foi criar, além do SN Hiper (sintagma nominal hiperônimo), o SN HHiper, que considera como SN hiperônimo apenas o primeiro substantivo à esquerda de "como/tais como". Uma análise do *corpus* mostrou que, se houver vírgula antecedendo a expressão "como/tais como", o hiperônimo pode ser considerado o SN Hiper tradicional, ou seja, o SN completo. Assim, o padrão "tais como" é dividido em duas regras:

(1) *SN HHiper* (*tais como* | *como_PDEN*) SN_1 { , SN_2 ... ,} (*e*|*ou*) SN_i

(2) *SN Hiper* , (*tais como* | *como*_PDEN) SN_1 { , SN_2 ... ,} (*e*|*ou*) SN_i

- Padrão "e outros"

  Equivalente ao padrão "(*and*|*or*) *other*" de Hearst. Assim como o padrão "tais como", esse também sofre do problema de ambiguidade do sintagma preposicional, como ilustrado no exemplo:

  ... [a experiência subjetiva com [o LSD-25]] e outros [alucinógenos]

  Onde o hipônimo corresponde apenas a "o LSD-25". Neste caso, a dificuldade de segmentação é nos SNs hipônimos. A solução encontrada foi criar o SN HHipo, que funciona da mesma forma que o SN HHiper: é considerado o SN hipônimo apenas o primeiro substantivo à esquerda de "*e*|*ou outros*".

  SN HHipo { ,*SN* $Hipo_i$ } * { , } *e*|*ou outros SN Hiper*

- Padrão "tipos de"

  Esse padrão foi derivado a partir da análise do *corpus*, observando que a expressão "tipos de" expressa relação de hiponímia.

  tipos de *SN Hiper*: SN_1 { , SN_2 ... ,} (*e*|*ou*) SN_i

- Padrão "chamado"

Esse padrão também foi derivado a partir da análise do *corpus*. Ele também sofre do problema da ambiguidade do sintagma preposicional, e novamente foi utilizada a estrutura SN HHiper.

Sentença exemplo: ... [a alta frequência da doença mental] chamada [esquizofrenia].

(IV) $SN\ HHiper\ chamado/s/a/as\ \{de\}\ SN\ Hipo$

A avaliação do trabalho foi feita da seguinte maneira: dos resultados obtidos da aplicação dos padrões léxicos sobre o *corpus*, primeiramente foram retiradas as relações com erros sintáticos. As relações restantes foram então validadas por 3 avaliadores de diferentes formações (biologia, educação física e direito), que deveriam pontuar uma amostra das relações de acordo com as seguintes categorias:

- 3: a relação está correta da forma como foi extraída;
- 2: a relação está parcialmente correta, isto é, o substantivo núcleo está correto, mas preposições, adjetivos, etc. que o acompanham deixam a relação estranha;
- 1: a relação está correta em termos gerais, mas é muito geral ou muito específica para ser útil;
- 0: a relação está errada.

Foram obtidas 2244 relações sintaticamente corretas, e delas 436 (cerca de 1/3) foram selecionadas aleatoriamente como amostra para avaliação manual, onde avaliadores pontuaram cada relação de acordo com os critérios mostrados acima. Essa avaliação apontou um índice de 73,4% de relações corretas (primeira categoria), resultado melhor que o obtido por (HEARST, 1998). Apesar disso, devido à natureza subjetiva dessa avaliação, não é possível fazer comparações diretas com outros trabalhos. Os resultados para as demais categorias foram 3,5% de relações parcialmente corretas, 16% de relações muito gerais e 7,1% de relações incorretas.

PAPEL (OLIVEIRA et al., 2010)

Oliveira et al. (2010) apresentam o PAPEL[11] (Palavras Associadas Porto Editora – Linguateca), um recurso lexical para o português similar à WordNet (FELLBAUM, 1998), constituído por um conjunto de relações semânticas entre termos e construído semiautomaticamente. Sua construção

[11] `http://www.linguateca.pt/PAPEL/`

Tabela 2.3: Número de relações semânticas do PAPEL (OLIVEIRA et al., 2010) por categoria

Relação	Número
Sinonímia	82137
Hiperonímia	49945
Meronímia	12616
Antonímia	388
Causa	7758
Produtor	1311
Finalidade	8561
Estado	394
Qualidade	1628
Local	834
Maneira	1262
Material	335
Referente	24005
Total	191174

foi feita utilizando a estratégia de padrões textuais, mas a partir de definições de um dicionário e não de textos livres. Essa escolha por uma fonte textual mais estruturada facilita a extração de relações semânticas, já que dicionários têm formato e vocabulário mais previsíveis e simples. É importante mencionar que o trabalho foi desenvolvido sobre a variante lusitana do idioma português, que apresenta distinções em relação à variante brasileira. O dicionário utilizado para a extração das relações foi o *Dicionário PRO da Língua Portuguesa da Porto Editora*.

As relações semânticas presentes no PAPEL somam cerca de 190 mil, divididas entre várias categorias: sinonímia e hiperonímia (as mais frequentes), meronímia (dividida entre as relações parte-de, membro-de e contido-em), antonímia, causalidade, produto-produtor, finalidade, estado (p. ex. "limpeza" tem o estado "asseio"), qualidade (p. ex. "célebre" tem a qualidade "celebridade"), localidade, maneira (a maneira como algo é feito, p. ex. "simplesmente" vem de "simplicidade"), material (p. ex. "vaso" é feito de "barro") e referente (p. ex. "renal" se refere a "rim"). A Tabela 2.3 mostra a quantidade de cada uma dessas relações na base.

O processo de construção do PAPEL envolveu primeiramente a construção de gramáticas semânticas para cada tipo de relação que se pretendia extrair (Tabela 2.4) e a criação de descrições para cada relação, de acordo com seu tipo e especificando a categoria gramatical de seus argumentos (Figura 2.7). Então, utilizando um *parser*, é feita a análise superficial de definições de verbetes do dicionário (Figura 2.8), da qual são extraídas automaticamente relações utilizando os padrões das gramáticas descritas anteriormente. Os resultados obtidos são avaliados manualmente.

Tabela 2.4: Exemplos de padrões usados nas gramáticas de Oliveira et al. (2010)

Padrão	Relação associada
tipo|gênero|classe|forma de	Hiperonímia
parte|membro de	Meronímia
que causa|provoca|origina	Causa
usado|utilizado para	Objetivo
uma palavra ou lista de palavras	Sinonímia

Figura 2.7: Exemplo de descrição de grupo de relações relativas à meronímia (OLIVEIRA et al., 2010)

```
PARTE {
    nome:nome * PARTE_DE:INCLUI;
    nome:adj * PARTE_DE_ALGO_COM_PROPRIEDADE:PROPRIEDADE_DE_ALGO_QUE_INCLUI;
    adj:nome * PROPRIEDADE_DE_ALGO_PARTE_DE:INCLUI_ALGO_COM_PROPRIEDADE;
}
```

As gramáticas não realizam a análise sintática das definições, o que pode levar à construção de relações entre palavras de categorias gramaticais incompatíveis, situação que é verificada automaticamente, primeiro verificando-se o próprio verbete do dicionário, e em seguida aplicando-se um analisador morfológico. Se for constatado que as categorias de uma relação estão incorretas mas isso puder ser corrigido trocando a relação por outra do mesmo grupo, essa substituição é feita, caso contrário a relação é descartada. Por exemplo, a relação "loucura ACÇÃO_QUE_CAUSA desvario", que espera um verbo como primeiro argumento, é transformada automaticamente em "loucura CAUSADOR_DE desvario", pois ambos os argumentos são substantivos.

O PAPEL (em sua versão 2.0) foi avaliado em duas instâncias, uma para a relação de sinonímia e outra para as relações de hiperonímia, meronímia, causalidade, produtor e finalidade. A avaliação da sinonímia foi realizada comparando-se os dados do PAPEL com o TeP 2.0 – Thesaurus Eletrônico para o Português do Brasil (descrito no Capítulo 3): primeiramente, foi feita uma análise da cobertura de um recurso sobre o outro, procurando quantos termos do TeP se encontravam no PAPEL e vice-versa. O resultado foi que existiam 28971 termos em comum entre os dois recursos, o que corresponde a 30% dos termos do PAPEL e 68,2% dos termos do TeP. A avaliação da relação de sinonímia do PAPEL foi feita sobre esse subconjunto de termos comuns para evitar enviesamentos. Verificou-se que 50,1% das relações de sinonímia do PAPEL estavam presentes no TeP, e 21,3% das relações do TeP estavam presentes no PAPEL. Isso mostra que, apesar de os dois recursos terem propósitos similares, são complementares.

Figura 2.8: Definição de dicionário para “cometa”, resultado da análise sintática do verbete e relações extraídas (OLIVEIRA et al., 2010)

cometa, s. m. - astro geralmente constituído por núcleo, cabeleira e cauda

```
[RAIZ]
    [QUALQUERCOISA]
    > [astro]
    [QUALQUERCOISA]
    > [geralmente]
    [PADRAO\_CONSTITUIDO]
        [VERBO\_PARTE\_PP]
        > [constituído]
        [PREP]
        > [por]
        [ENUM\_PARTE]
            [PARTE\_DE]
            > [núcleo]
            [VIRG]
            > [,]
            [ENUM\_PARTE]
                [PARTE\_DE]
                > [cabeleira]
                [CONJ]
                > [e]
                [PARTE\_DE]
                > [cauda]
```

núcleo PARTE DE cometa
cabeleira PARTE DE cometa
cauda PARTE DE cometa

A avaliação das relações de hiperonímia, meronímia, causalidade, produtor e finalidade foi feita buscando-se as relações manualmente no *corpus* CETEMPúblico (ROCHA; SANTOS, 2000). Primeiramente, foram desconsideradas as relações que tivessem argumentos que não constassem no *corpus*, resultando nas seguintes quantidades de instâncias a serem procuradas: 40079 para hiperonímia, 3746 para meronímia, 557 para causalidade, 414 para produtor e 1718 para finalidade. Porém, por restrições de tempo foram selecionadas amostras aleatórias de 8% das relações de hiperonímia (3145 instâncias) e 63% das relações de meronímia (2343 instâncias) para serem averiguadas. Uma média de 20% das amostras de relações de hiperonímia e meronímia foi encontrada no *corpus*; para as relações de causalidade, produtor e finalidade, uma média de 5,6% das relações foi verificada. Esses valores são explicados pelo pequeno tamanho do *corpus* e pelos padrões utilizados na construção do PAPEL serem bastante simples (textos livres têm muito mais possibilidades de definição de relações que verbetes de dicionários).

Tarefa ReRelEM do segundo HAREM (MOTA; SANTOS, 2008)

O HAREM[12] (**H**AREM é uma **A**valiação de **R**econhecedores de **E**ntidades **M**encionadas) foi um evento organizado pela Linguateca[13] em 2006 e 2008, com o propósito de avaliar sistemas de reconhecimento de entidades nomeadas (entidades mencionadas, em português de Portugal) para o português principalmente lusitano (mas não limitado a este). O HAREM de 2008 contou com a tarefa ReRelEM, que significa "Reconhecimento de Relações entre Entidades Mencionadas", o primeiro evento de avaliação conjunta de sistemas que fazem a identificação de relações semânticas entre entidades para a língua portuguesa. Uma iniciativa semelhante a eventos similares para a língua inglesa, como as *Message Understanding Conferences* (MUC) (GRISHMAN; SUNDHEIM, 1996), *Automatic Content Extraction* (ACE) (DODDINGTON et al., 2004), e as *Semantic Evaluation* de 2007 (GIRJU et al., 2007) e de 2010 (HENDRICKX et al., 2009).

As relações semânticas em foco no ReRelEM foram "identidade" (duas entidades se referem ao mesmo referente – engloba sinonímia, definições, acrônimos e relações anafóricas), "inclusão" (uma entidade inclui outra entidade – engloba relações de hiponímia e meronímia), "localização" (um evento ocorre em um local) e "outras", que engloba relações relevantes mas que não se encaixam nas categorias anteriores. O *corpus* de referência construído para a tarefa é composto por 1416 relações do tipo identidade, 1650 de inclusão, 1232 de localização e 2179 na categoria "outras". Uma das características do ReRelEM é que, por ser uma tarefa do HAREM, todas as relações semânticas anotadas ocorrem entre entidades nomeadas de um texto. Assim, os sistemas avaliados com esse *corpus* de referência devem realizar primeiramente o reconhecimento de entidades nomeadas, para então procurarem as relações entre elas.

Foram submetidos três sistemas para o ReRelEM: REMBRANDT (CARDOSO, 2008), SEI-Geo (CHAVES, 2008) e SeRELeP (BRUCKSCHEN et al., 2008).

- O REMBRANDT utiliza a porção em português da Wikipedia[14] como base de conhecimento externo juntamente com regras gramaticais simples na etapa de reconhecimento de entidades nomeadas. A etapa de identificação de relações semânticas utiliza heurísticas simples, como a comparação de similaridade entre as formas superficiais das entidades para encontrar relações "identidade" e a proximidade entre a posição das entidades "acontecimento" e "local" para definir relações de localidade.
- O SEI-Geo é parte de um sistema maior de extração e integração de conhecimento ge-

[12] `http://www.linguateca.pt/HAREM/`
[13] `http://www.linguateca.pt/`
[14] `http://pt.wikipedia.org/`

ográfico, portanto, procura apenas pela relação "inclusão" entre entidades do tipo "local". Para reconhecer as entidades nomeadas referentes a locais, é utilizada uma ontologia de dados geográficos aliada a padrões textuais como os de Hearst (1992), do tipo "[conceito geográfico] tal(is) como [local]". A identificação da relação de inclusão é feita mapeando as entidades nomeadas relativas a locais em ontologias geográficas e procurando relações dentro dessas ontologias.

- O SeRELeP utiliza algumas das etiquetas produzidas pelo *parser* PALAVRAS na etapa de reconhecimento de entidades nomeadas (p. ex. a etiqueta "prop", que indica nome próprio, e algumas etiquetas semânticas) e heurísticas simples para encontrar as relações entre as entidades, como a comparação direta entre a similaridade das formas superficiais das entidades para procurar pela relação "identidade" e a ligação de uma entidade classificada como "local" à entidade "organização" ou "acontecimento" mais próxima no texto. O SeRELeP não procura pela relação do tipo "outras".

Como cada um dos três sistemas considera um subconjunto distinto das relações propostas no ReRelEM, a avaliação dos resultados individuais foi feita sobre o subconjunto de relações adequado do *corpus* de referência. Apenas o REMBRANDT busca todas as quatro relações; o SeRELeP não considera a relação "outras" e o SEI-Geo só busca a relação "inclusão".

O sistema que obteve melhor desempenho foi o REMBRANDT, que apresentou valores para precisão, cobertura e medida-F de 58,2%, 36,7% e 45%, respectivamente. Os melhores resultados para o SEI-Geo foram 91,7%, 16,2% e 27,5% – uma alta precisão, mas baixa cobertura – e para o SeRELeP foram 58,3%, 26,7% e 36,6% – a precisão é ligeiramente maior que a do REMBRANDT, mas a cobertura é menor. Como se vê, comparados com os trabalhos realizados para a língua inglesa apresentados anteriormente neste capítulo, esses são valores ainda bem abaixo do estado da arte.

Trabalhos relacionados

Entre outros trabalhos sobre extração automática de relações semânticas com o português em foco podem ser citados (GASPERIN; LIMA, 2002, 2003), que pesquisam a extração de palavras semanticamente relacionadas. Essa extração é feita através do cálculo da similaridade distribucional e da similaridade do contexto sintático entre as palavras do *corpus*. Gasperin e Lima expandiram o trabalho de Grefenstette (1993) para o português brasileiro, usando o *parser* PALAVRAS e o *corpus* FolhaNOT, um subconjunto de cerca de 1,4 milhão de palavras do *corpus* do NILC[15].

[15] `http://www.nilc.icmc.usp.br/nilc/`

Os resultados dos métodos de Gasperin e Lima são listas de palavras que são possivelmente relacionadas semanticamente a alguma outra palavra. A avaliação dessas listas é subjetiva, avaliando a homogeneidade das palavras nelas contidas; Gasperin e Lima definem que uma lista é homogênea se as palavras que a compõem são semanticamente relacionadas. Analisando uma amostra das listas encontradas, verificou-se que elas eram bastante homogêneas. Os trabalhos de Gasperin e Lima têm uma granularidade mais geral que o presente trabalho de mestrado; enquanto esta dissertação propõe encontrar certas relações semânticas específicas entre termos (no caso, relações de Minsky), Gasperin e Lima procuram palavras semanticamente relacionadas de qualquer natureza.

Percebe-se que trabalhos sobre extração automática de relações semânticas que têm o português como língua-alvo utilizam apenas a abordagem de padrões textuais, sendo a de aprendizado de máquina ainda pouco ou nada explorada. Um dos motivos para isso pode ser os poucos recursos e ferramentas disponíveis para o processamento computacional da língua portuguesa, dificultando o uso da abordagem de aprendizado de máquina, que se baseia em diversas *features* de diferentes níveis linguísticos e que necessita de sistemas e bases de conhecimento sofisticadas para anotação.

A ferramenta mais utilizada em pesquisas com o português é o *parser* PALAVRAS (BICK, 2000), usado em (GASPERIN; LIMA, 2003; FREITAS; QUENTAL, 2007; BRUCKSCHEN et al., 2008). Algumas das bases de conhecimento externo utilizadas nas pesquisas levantadas são dicionários (OLIVEIRA et al., 2010) e a Wikipedia (CARDOSO, 2008), e as relações semânticas mais frequentemente consideradas foram hiponímia (FREITAS; QUENTAL, 2007; MOTA; SANTOS, 2008; OLIVEIRA et al., 2010), meronímia e localidade (MOTA; SANTOS, 2008; OLIVEIRA et al., 2010).

Considerando-se que as abordagens investigadas e os resultados obtidos até o momento para o português ainda estão distantes daqueles para outros idiomas, como o inglês, esse trabalho investigou como e em que medida métodos automáticos podem ser aplicados para a extração de relações semânticas binárias em textos escritos no português do Brasil, com o auxílio de informações de senso comum. Para tanto, este trabalho desenvolveu pesquisas em ambas as abordagens – tanto a de padrões textuais quanto a de aprendizado de máquina – e explorou diferentes relações semânticas e um recurso ainda pouco utilizado no PLN: uma base de dados com informações de Senso Comum.

2.3 Tabela comparativa dos trabalhos apresentados

Concluindo este capítulo, as Tabelas 2.5 e 2.6 a seguir resumem os principais trabalhos relacionados e permitem uma visualização mais direta das características de cada um. A avaliação dos trabalhos mencionados não consta das tabelas, pois (i) as relações extraídas são muito variadas e (ii) os métodos e recursos usados nas avaliações de cada um são distintos, impedindo uma comparação objetiva entre as abordagens. A descrição mais detalhada dos métodos pode ser revista no texto apresentado anteriormente nesta seção.

Tabela 2.5: Resumo dos principais trabalhos relacionados descritos neste capítulo

Referência	**Língua-alvo**	**Relações extraídas**	***Corpus***	**Abordagem**
(HEARST, 1992, 1998)	Inglês	Hiponímia	6 meses de texto do New York Times	Padrões textuais
(BRIN, 1999)	Inglês	Autor-livro	Subconjunto de 24 milhões de páginas da web	Padrões textuais
(AGICHTEIN; GRAVANO, 2000)	Inglês	Organização-local	Mais de 300 mil artigos jornalísticos	Padrões textuais
(GIRJU; MOLDOVAN, 2002)	Inglês	Causalidade	3GB de textos jornalísticos	Padrões textuais
(CARABALLO, 1999)	Inglês	Hiponímia	Seleção de textos jornalísticos (tamanho não especificado)	Aprendizado de máquina
(GIRJU et al., 2003) e (GIRJU, 2003)	Inglês	Meronímia e causalidade	Cerca de 150 mil sentenças do *corpus* SemCor 1.7 e de artigos do *Los Angeles Times* do *corpus* da conferência TREC-9	Aprendizado de máquina
(SNOW et al., 2005)	Inglês	Hiponímia	Mais de 6 milhões de sentenças de textos jornalísticos	Aprendizado de máquina
(ZELENKO et al., 2003)	Inglês	Pessoa-filiação e organização-local	Cerca de 200 artigos jornalísticos	Aprendizado de máquina
(ZHANG et al., 2006)	Inglês	Relações físicas, pessoais, emprego, agente-artefato, afiliação com organização, afiliação com local e relações anafóricas	15385 instâncias de relações dos ACEs de 2003 e 2004	Aprendizado de máquina
(MINTZ et al., 2009)	Inglês	Diretor-filme, localidade, pessoa-nacionalidade, pessoa-data de nascimento e outras 98	1,2 milhões de artigos da Wikipedia e 1,8 milhões de instâncias de 102 tipos de relações semânticas da Freebase	Aprendizado de máquina
(YAP; BALDWIN, 2009)	Inglês	Hiponímia, sinonímia e antonímia	*English Gigaword Corpus* (84 milhões de sentenças de textos jornalísticos) e *dump* da Wikipedia (38 milhões de sentenças)	Aprendizado de máquina
(GIRJU et al., 2010)	Inglês	Causa-efeito, instrumento-agente, produto-produtor, origem-entidade, tema-ferramenta, parte-todo, conteúdo-continente	4394 instâncias distribuídas entre 7 tipos de relações semânticas	Aprendizado de máquina
(CARLSON et al., 2010)	Inglês	Empresa-produz-produto, empresa-localizada-em, livro-autor, time-joga-para, entre outras	Web	Aprendizado de máquina

Tabela 2.6: Resumo dos principais trabalhos relacionados descritos neste capítulo – continuação

Referência	**Língua-alvo**	**Relações extraídas**	**Corpus**	**Abordagem**
(FREITAS; QUENTAL, 2007)	Português do Brasil	Hiponímia	11MB de textos sobre saúde pública	Padrões textuais
(OLIVEIRA et al., 2010)	Português de Portugal	Hiponímia, sinonímia, antonímia, meronímia, causa-efeito, produto-produtor, finalidade, estado, qualidade, local, maneira, material, referente	Dicionário PRO da Língua Portuguesa da Porto Editora (tamanho não especificado)	Padrões textuais
(CHAVES, 2008), (CARDOSO, 2008) e (BRUCKSCHEN et al., 2008) – ReRelEM	Português de Portugal	Identidade, inclusão, localização e "outras"	*Corpus* anotado com 6477 relações distribuídas entre as 4 consideradas	Padrões textuais

3 *Metodologia*

A partir dos métodos e estratégias descritos no capítulo anterior para a extração automática de relações semânticas a partir de textos, decidiu-se investigar as duas principais abordagens enfocadas na literatura:

1. Abordagem de padrões textuais;
2. Abordagem de aprendizado de máquina com diferentes classificadores (árvores de decisão C4.5 e *Support Vector Machines*)

Essas abordagens são apresentadas em ordem crescente de complexidade e quantidade de recursos utilizados e, como tal, também foram investigadas segundo essa ordenação. A primeira abordagem, descrita na Seção 3.3, se baseia nos trabalhos de Hearst (1992) e Freitas e Quental (2007), tendo sido utilizada para encontrar apenas instâncias da relação de hiponímia (is-a). A segunda abordagem, derivada dos trabalhos baseados em aprendizado de máquina averiguados e descrita na Seção 3.4, utiliza *corpora* anotados e algoritmos de aprendizado de máquina para encontrar instâncias de todas as sete relações de interesse descritas na Tabela 1.1.

Na Seção 3.1 a seguir, serão apresentadas algumas definições que serão utilizadas no decorrer do capítulo. A Seção 3.2 detalha os *corpora* utilizados neste trabalho, recurso imprescindível para ambas as abordagens. Nesta seção também serão descritos outros recursos e ferramentas que foram utilizados no desenvolvimento desta pesquisa. Conforme já citado, a Seção 3.3 trata dos experimentos com a extração de relações semânticas usando a estratégia de padrões textuais e a Seção 3.4 trata dos experimentos que utilizam algoritmos de aprendizado de máquina. Os resultados dos experimentos realizados serão descritos no próximo capítulo.

3.1 Definições

Antes de iniciar a descrição do desenvolvimento desta pesquisa, é importante estabelecer algumas definições que serão utilizadas nos textos que se seguem.

- ***Token*** – uma sequência de quaisquer caracteres com exceção do espaço;
- **Termo** – uma sequência de *tokens* que representa uma entidade ou tem algum significado específico em uma sentença;
- **Relação semântica** – nesta pesquisa em específico, é uma tripla *<relação, termo1, termo2>*, onde *relação* é a denominação de uma relação semântica (p. ex. is-a, made-of, part-of, como as descritas na Tabela 1.1) e *termo1* e *termo2* são dois termos distintos em uma sentença. Relações semânticas podem ser denotadas simplificadamente com a notação *relação(termo1, termo2)* que será usada no decorrer do trabalho.

3.2 Recursos e ferramentas

Nesta seção são descritos os recursos (seção 3.2.1) e as ferramentas (seção 3.2.2) utilizados nesta pesquisa.

3.2.1 Recursos

A tarefa de extração automática de relações semânticas por meio de métodos de aprendizado de máquina necessita de um *corpus* de treinamento anotado manualmente com as relações semânticas existentes entre os termos de cada sentença. As anotações das relações precisam ser feitas manualmente devido à inexistência de um sistema que faça essas anotações automaticamente para o idioma português. Além dessas anotações que representam os exemplos de treinamento, são interessantes outras que podem servir como *features* para o aprendizado, como por exemplo anotações morfossintáticas na forma de etiquetas *part-of-speech* (POS), árvores sintáticas e, possivelmente, etiquetas semânticas. Para algumas dessas anotações, estão disponíveis ferramentas para o português do Brasil como o *parser* PALAVRAS (BICK, 2000) e os etiquetadores morfossintáticos do projeto ReTraTos (CASELI, 2007), detalhados na subseção 3.2.2.

Dois *corpora* foram utilizados neste estudo. O primeiro *corpus* é composto por 646 artigos científicos da revista Pesquisa FAPESP[1], escritos em português do Brasil. Esse *corpus* foi originalmente utilizado e enriquecido com informações linguísticas no ReTraTos (CASELI, 2007). No entanto, para o presente estudo, o *corpus* foi processado morfossintaticamente pelo *parser* PALAVRAS (BICK, 2000). Esse *corpus* está composto por 17397 sentenças (cerca de 870 mil palavras) enriquecidas com informações complementares para cada forma superficial (a palavra da maneira

[1] `http://revistapesquisa.fapesp.br/`

como ocorre no texto, ex: meninos) das palavras em português: lema (a forma base da palavra, ex: menino), *part-of-speech* (categoria gramatical, ex: substantivo), traços morfológicos (valores dos atributos das POS da palavra, ex: masculino plural) e atributo de dependência sintática, indicando qual o papel do *token* na árvore de dependências da sentença. Embora o PALAVRAS também produza algumas etiquetas semânticas para algumas poucas palavras na sua análise final, as mesmas não satisfazem o objetivo deste trabalho (extração de relações como as apresentadas na Seção 1.1).

O segundo *corpus* utilizado nos demais experimentos desta pesquisa é o CETENFolha[2], composto por cerca de 24 milhões de palavras provenientes de artigos de diversos cadernos do jornal *Folha de São Paulo* do ano de 1994. Esse *corpus* foi também enriquecido com informações morfossintáticas pelo *parser* PALAVRAS (BICK, 2000), trazendo portanto informações morfológicas e o papel sintático na árvore de constituintes da sentença para cada *token*.

Como mencionado anteriormente, as relações semânticas de interesse precisam ser manualmente anotadas nos *corpora* de treinamento. A anotação de relações semânticas envolve na verdade dois passos: (1) a anotação dos termos de interesse e (2) a anotação das relações em si. Por exemplo, considerando a sentença:

```
Os dentes do mais antigo orangotango
```

Temos três termos de interesse: *dentes*, *mais antigo* e *orangotango*. Usualmente todos os substantivos, tanto comuns quanto próprios, e todos os adjetivos e seus modificadores são considerados como termos. Dependendo da situação, verbos também podem ser marcados se indicam alguma relação effect-of ou used-for. Entre esses três termos da sentença, podemos inferir a relação semântica de meronímia (part-of), ou seja, *dentes* são uma parte de *orangotango*, e a relação semântica property-of entre *mais antigo* e *orangotango*. É papel do anotador verificar quais relações semânticas existem entre todos os pares de termos da sentença. Pode-se representar essas relações como:

```
Os [dentes]1 do [mais antigo]2 [orangotango]3 -- property-of(3, 2) e
                                                 part-of(3, 1)
```

Onde os termos de interesse são identificados entre colchetes e pelos números 1, 2 e 3 e as relações semânticas part-of e property-of indicam a relação de parte e todo e de propriedade existentes na sentença. A Figura 3.1 mostra outros exemplos de sentenças anotadas dessa forma.

[2] `http://www.linguateca.pt/cetenfolha/`

Figura 3.1: Exemplos de sentenças do *corpus* FAPESP anotadas para indicar as relações semânticas entre os termos

1. `uma equipe da [universidade estadual de campinas]1 ( [unicamp]2 ) conseguiu isolar e caracterizar pela primeira vez o [vírus respiratório sincicial bovino]3 ( [brsv]4 ) no brasil , que causa graves problemas respiratórios sobretudo em bezerros . [is-a(1,2), is-a(3,4)]`
2. `uma nova [espécie de hominídeo]1 encontrado na [tailândia]2 , com estimados 12 milhões de anos , tornou - se o parente mais remoto dos atuais [orangotangos]3 ( [pongo pygmaeus]4 ) . [location-of(1,2), is-a(3,4)]`
3. `mas só agora começam a reunir condições de responder à pergunta : quando é que virá o próximo el niño ? []`

Essa notação *inline* poderia ser utilizada para a anotação das relações semânticas, mas considerando que os *corpora* trazem outras informações além da forma superficial de cada *token* (p. ex. traços morfológicos, lema) e estes poderiam poluir visualmente e dificultar a marcação das relações, preferiu-se a utilização de uma estrutura própria de anotação baseada no formato JSON[3] (*JavaScript Object Notation*). Nessa estrutura, *tokens*, termos e relações são tratados separadamente, cada qual como um objeto autocontido com seus próprios atributos. O formato completo está detalhado no Apêndice A. A sentença exemplo apresentada anteriormente representada nesse formato pode ser vista a seguir:

```
{
    "texto": "Os dentes de o mais antigo orangotango .",
    "tokens": [
        { "t": "Os", "l": "o", "pos": "<artd> DET M P", "sin": "@>N" },
        { "t": "dentes", "l": "dente", "pos": "N M P", "sin": "@NPHR" },
        { "t": "de", "l": "de", "pos": "PRP",  "sin": "@N<" },
        { "t": "o", "l": "o", "pos": "<-sam> <artd> DET M S", "sin": "@>N" },
        { "t": "mais", "l": "mais", "pos": "<quant> <KOMP> ADV", "sin": "@>A" },
        { "t": "antigo", "l": "antigo", "pos": "ADJ M S", "sin": "@>N" },
        { "t": "orangotango", "l": "orangotango", "pos": "N M S",  "sin": "@P<" },
        { "t": ".", "l": ".", "pos": null, "sin": null }
    ],
    "termos": [
```

[3] http://www.json.org/

```
        { "de": 1, "ate": 1 },
        { "de": 7, "ate": 7 }
    ],
    "relações": [
        { "r": "property-of", "t1": 2, "t2": 1 }
    ]
}
```

Esse formato foi escolhido sobre outros formatos tradicionais de codificação de *corpus* baseados em XML como o Tiger XML (MENGEL; LEZIUS, 2000) e o XCES (IDE et al., 2000) por ser mais sucinto mas igualmente portável e simples de se manipular computacionalmente. Para auxiliar a tarefa de anotação manual das sentenças nesse formato, foi desenvolvida uma ferramenta que será detalhada na subseção a seguir.

Devido a seu maior tamanho e abrangência de assuntos, o *corpus* CETENFolha foi escolhido para anotação manual. Visto que a anotação de todas as suas sentenças com termos e relações semânticas de interesse seria uma tarefa inviável dado o tamanho do *corpus*, foi selecionado um subconjunto de suas sentenças. Essa amostra foi selecionada de acordo com a frequência de ocorrência de sintagmas nominais (SNs) no *corpus*. Primeiramente, todos os SNs de todas as sentenças foram identificados usando o identificador de SNs descrito em (SANTOS; OLIVEIRA, 2005) e então foram escolhidas as sentenças que continham sintagmas que ocorriam de 18 a 51 vezes no *corpus*. Essa faixa de frequência foi delimitada considerando-se as frequências ordenadas e acumuladas de todos os SNs. Desse modo, a amostra representa cerca de 15 mil sintagmas que correspondem a uma faixa de 10% do total de SNs do *corpus* cuja frequência está entre 40% e 50% da distribuição acumulada; não são os mais frequentes, localizados no início da curva de distribuição, e nem os menos frequentes, localizados no fim da curva. O intervalo entre 40% e 50% foi definido pois resulta em uma boa variedade de sintagmas distintos (15 mil) com um número significativo de ocorrências (18 a 51).

As sentenças em que esses 15 mil SNs apareciam foram selecionadas, resultando em uma amostra de aproximadamente 230 mil sentenças a serem anotadas. Essas sentenças foram transformadas do formato de saída do PALAVRAS para o formato JSON citado anteriormente e divididas em pacotes com cerca de 1000 sentenças.

De forma a obter maior qualidade e coerência nas anotações manuais das relações semânticas, visto que essa é uma tarefa subjetiva, decidiu-se que o processo de anotação seria realizado por dois anotadores em paralelo, seguindo as regras especificadas em (TABA; CASELI, 2013). A fim de calcular a concordância entre as anotações de ambos, cada par de conjuntos de

1000 sentenças tem por volta de 100 sentenças em comum; a concordância é calculada sobre essas sentenças em comum como o número de relações anotadas idênticas pelos dois anotadores dividido pelo número total de relações distintas anotadas por ambos. É pertinente ressaltar que, como a anotação de relações semânticas compreende duas etapas – anotação de termos e das relações entre eles – e os termos podem ser marcados de formas distintas pelos anotadores, não é possível utilizar o coeficiente kappa (CARLETTA, 1996) para cálculo da concordância. Por exemplo, se um anotador marcou a instância de relação property-of(orangotango, antigo) e o outro anotador marcou a instância property-of(orangotango, mais antigo), elas são consideradas distintas. Uma instância de relação é considerada idêntica apenas se a relação e os termos participantes são iguais.

Até o momento da escrita desta dissertação, 4 pacotes haviam sido anotados, 2 por cada anotador, com uma taxa de concordância de 69,15%. Esses 4 pacotes compreendem cerca de 3800 sentenças distintas e 11757 instâncias de relações semânticas diferentes, ou seja, anotadas por pelo menos um anotador. O número de instâncias de cada relação anotadas por cada anotador pode ser visto na Tabela 3.1. A tabela contém também o número de instâncias negativas (pares de termos entre os quais não existe relação) – elas são usadas no treinamento dos métodos de aprendizado de máquina, onde é importante que os classificadores saibam reconhecer os pares de termos que não são relacionados. As instâncias negativas são geradas automaticamente a partir de todos os pares de termos marcados em sentenças, mas entre os quais não existe nenhuma relação.

Tabela 3.1: Número de instâncias de cada relação marcadas por cada anotador e, na última coluna, o número total de instâncias distintas, mais as instâncias negativas geradas automaticamente

Relação	**Anotador A**	**Anotador B**	$\mathbf{A \cup B - A \cap B}$
property-of	2843	2271	4909
is-a	1333	1617	2816
part-of	902	1203	1976
location-of	775	967	1683
effect-of	58	111	163
used-for	21	117	136
made-of	23	59	74
Subtotal	5955	6345	11757
Classe negativa	–	–	104135
Total	–	–	115892

O *corpus* FAPESP também teve cerca de 500 sentenças anotadas pelo anotador B para realização de testes com diferentes *corpora* na estratégia de aprendizado de máquina. Essa porção do *corpus* foi anotada por apenas um anotador já que, após a marcação dos 4 pacotes do CETENFolha, ambos os anotadores obtiveram uma taxa de concordância aceitável, não sendo mais

necessária a anotação em paralelo. O número total de instâncias anotadas para este *corpus* foi 1986, sendo o número de instâncias de cada relação apresentado na Tabela 3.2.

Tabela 3.2: Número de instâncias de cada relação marcadas no *corpus* FAPESP mais as instâncias negativas geradas automaticamente

Relação	**Instâncias**
property-of	853
is-a	367
part-of	306
location-of	248
effect-of	84
used-for	68
made-of	60
Subtotal	1986
Classe negativa	23985
Total	25971

Após a etapa de preparação dos *corpora* com a anotação de relações semânticas, eles podem ser utilizados tanto na etapa de treinamento quanto na etapa de testes, onde as relações extraídas pelos sistemas implementados são comparadas com as relações anotadas manualmente. Detalhes de como essas instâncias de relações anotadas manualmente foram utilizadas em cada experimento são apresentados juntamente com a descrição dos mesmos.

Outro recurso linguístico utilizado nesta pesquisa foi a base de dados de Senso Comum do projeto *Open Mind Common Sense* do Brasil, composta por cerca de 115 mil instâncias das relações de Minsky binárias englobando as 7 relações descritas na Tabela 1.1[4]. A quantidade de instâncias de cada relação na base pode ser vista na Tabela 3.3. Essas instâncias foram utilizadas como semente nos experimentos com padrões textuais, como descrito em detalhes na Seção 3.3.

3.2.2 Ferramentas

Além dos recursos linguísticos citados anteriormente, este trabalho também utilizou ferramentas computacionais como o analisador sintático PALAVRAS (BICK, 2000), já citado anteriormente, o etiquetador morfossintático derivado no projeto ReTraTos (CASELI, 2007) e o identificador de sintagmas nominais descrito em (SANTOS; OLIVEIRA, 2005). Para os experimentos com aprendizado de máquina, foram utilizados o pacote de algoritmos WEKA (HALL et al., 2009), muito usado devido à sua simplicidade e versatilidade, e o software SVM Light (JOACHIMS,

[4]Dados extraídos da base em outubro de 2011.

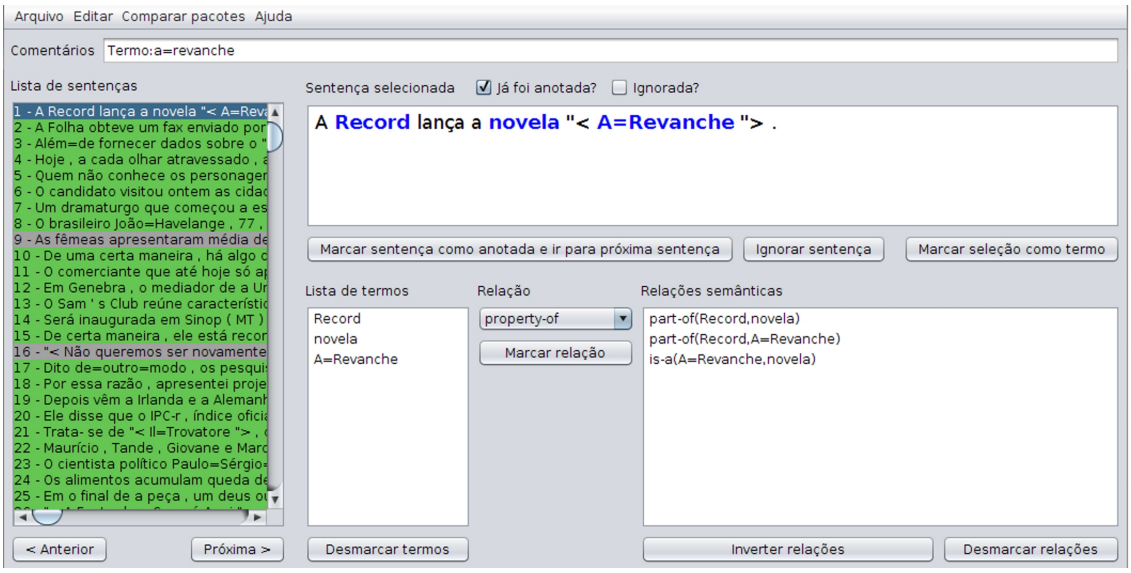
Tabela 3.3: Número de instâncias de cada relação presentes na base de dados do projeto OMCS-Br

Relação	Instâncias
property-of	15589
is-a	12436
part-of	6615
location-of	40509
effect-of	7057
used-for	27347
made-of	5248
Total	114801

1998), utilizado em alguns dos trabalhos citados na revisão bibliográfica (ZELENKO et al., 2003; ZHANG et al., 2006) para realizar experimentos com *Support Vector Machines*.

Foi também desenvolvida uma ferramenta específica para auxiliar a tarefa de anotação manual de relações semânticas em *corpora*: a ARS (Anotador de Relações Semânticas). Essa ferramenta, desenvolvida em Java, manipula sentenças codificadas no formato JSON citado anteriormente, permitindo a marcação visual de termos e de relações entre eles. Sua interface principal pode ser vista na Figura 3.2. A ARS também traz outras funcionalidades como o cálculo da concordância entre pacotes de dois anotadores distintos, marcação de comentários em cada sentença, entre outros.

Figura 3.2: Interface principal da ferramenta ARS usada para auxiliar a anotação de relações semânticas. Os termos aparecem em azul na sentença sendo anotada e como itens de uma lista de termos ao lado das relações já marcadas.



Preferiu-se o desenvolvimento de uma ferramenta própria e alinhada com a tarefa em questão pois as plataformas de anotação de *corpora* já existentes, como p. ex. a GATE (CUNNINGHAM et al., 2011) e a SALTO (BURCHARDT et al., 2006), traziam diversas funcionalidades que não seriam aproveitadas em nosso estudo. Além disso, a ARS inclui algumas funções não presentes nas demais ferramentas como a manipulação do formato JSON e o cálculo da concordância entre a anotação de dois pacotes distintos, além de poder ser alterada conforme a necessidade. Informações detalhadas a respeito do desenvolvimento da ferramenta podem ser obtidas em (TABA; CASELI, 2012b).

3.3 Experimentos com padrões textuais

A estratégia que se baseia na busca por padrões textuais para identificar relações semânticas, por ser a mais simples, foi a primeira investigada neste estudo. Particularmente, os trabalhos de Hearst (1992) e Freitas e Quental (2007) foram as principais referências nesta etapa do trabalho. Como já foi dito no Capítulo 2, Hearst (1992) foi uma das pioneiras a utilizar padrões para encontrar relações semânticas, no caso a hiponímia. Além disso, a pesquisadora também definiu um algoritmo iterativo para a descoberta de novos padrões e relações. Freitas e Quental (2007), baseando-se no trabalho de Hearst, traduziram seus padrões textuais para o português do Brasil.

Assim, foi realizado o **experimento 1** que consistiu na aplicação dos 4 padrões de Freitas e Quental (2007), resumidos na Figura 3.3, sobre o *corpus* FAPESP para encontrar relações de hiponímia. Após a aplicação dos padrões, o algoritmo de Hearst (veja Figura 2.3) foi implementado de forma semiautomática, usando como sementes as 12436 (veja Tabela 3.3) instâncias da relação de hiponímia da base do projeto OMCS-Br e as relações encontradas pela aplicação dos padrões de Freitas e Quental. Após a aplicação do algoritmo, uma análise manual dos contextos recuperados foi realizada resultando na definição de 3 novos padrões (Figura 3.4), que novamente foram utilizados para encontrar novas instâncias de relações de hiponímia. A avaliação das instâncias de relações encontradas neste experimento foi feita manualmente, já que a parte do *corpus* FAPESP usada neste experimento não está anotada com as relações semânticas de interesse. Os resultados serão descritos no próximo capítulo. Mais informações sobre este experimento podem ser obtidas em (TABA; CASELI, 2012a).

O **experimento 2** foi realizado com o intuito de definir padrões textuais para as outras 6 relações semânticas, algo até então não encontrado na literatura para o português do Brasil. Para tanto, esses padrões foram definidos a partir da análise manual e semiautomática do *corpus* CETENFolha. Inicialmente, por meio da observação do *corpus* foram definidos manualmente os 13

Figura 3.3: Padrões de Freitas e Quental (2007)

is-a 1: SN_Hiper (tais como|como) SN {, SN}* (e|ou) SN
is-a 2: SN {, SN}* ,? (e|ou) outros SN_Hiper
is-a 3: tipos de SN_Hiper: SN {, SN}* (e|ou) SN
is-a 4: SN_Hiper chamad(o|a|os|as) de? SN

Figura 3.4: Novos padrões identificados com a aplicação do algoritmo de Hearst (1992)

is-a 5: SN {, SN}* ,? (e|ou) (qualquer|quaisquer) outro{s}? SN_Hiper
is-a 6: SN é (o|a|um|uma) SN_Hiper
is-a 7: SN são SN_Hiper

padrões apresentados na Figura 3.4[5]. Depois, o algoritmo de Hearst (1992) (veja Figura 2.3) foi aplicado (1 iteração apenas) usando a base de dados do projeto OMCS-Br como semente (veja Tabela 3.3). Da aplicação do algoritmo surgiram novos contextos que foram analisados manualmente resultando em outros 4 padrões textuais apresentados na Tabela 3.5.

Tabela 3.4: Padrões definidos manualmente para as 6 demais relações semânticas

Relação	**#**	**Padrão**
property-of	1	T1_N T2_ADJ
	2	T2_ADJ T1_N
	3	T1_N " T2_ADJ "
part-of	1	T1 com T2
	2	T1 {verbo fazer} parte de T2
	3	T1 {verbo ser} parte de T2
made-of	1	T1_N de T2_N
	2	T1 (é|são)? feit(o|a|os|as) de T2
location-of	1	T1 entrou em T2
	2	T1 ,? localizad(a|o) em T2
effect-of	1	T2_V .* devido=a T1
	2	T2_V por=causa=de (a|o|as|os)? T1
used-for	1	T1 (que podem ser)? usadas? para T2_V

Os 17 novos padrões (Tabelas 3.4 e 3.5) e os 7 padrões do experimento 1 (Figuras 3.3 e 3.4) foram, então, aplicados sobre as mesmas sentenças do *corpus* CETENFolha anotadas manualmente. As instâncias de relações encontradas pela aplicação dos 24 padrões textuais a essa parte do *corpus* CETENFolha foram comparadas com as instâncias marcadas manualmente, possibilitando

[5]As notações "_N", "_ADJ" e "_V" indicam que um termo deve ser um substantivo, um adjetivo ou um verbo, respectivamente.

Tabela 3.5: Padrões definidos a partir de uma iteração do algoritmo de Hearst (1992) e análise manual dos contextos recuperados para as 6 demais relações semânticas

Relação	**#**	**Padrão**
property-of	4	de T1_ADJ T2_N
part-of		–
made-of		–
location-of	3	T1 chega a o T2
	4	T1 em (o|a|os|as) T2
effect-of		–
used-for	2	T1 para (o|a|os|as) T2_V

o cálculo automático dos valores de precisão, cobertura e medida-F (veja subseção 3.5). Os resultados mostram que os padrões textuais tiveram boa precisão mas baixa cobertura, corroborando o que já foi dito na Seção 2.1.1. Esses resultados serão detalhados no capítulo a seguir.

A baixa cobertura na recuperação automática de instâncias das relações obtida com a estratégia baseada em padrões textuais, aliada ao alto custo da análise manual (seja do *corpus* para extração direta de padrões, seja dos contextos após a aplicação do algoritmo de Hearst), desmotivou a continuidade da investigação desta estratégia.

3.4 Experimentos com aprendizado de máquina

O aprendizado de máquina (AM) é uma subárea da inteligência artificial que trata de algoritmos que, quanto mais "experiência" (exemplos) tem, melhor "aprendem" a realizar uma tarefa (classificação ou regressão). Os métodos de AM se tornaram cada vez mais populares devido ao barateamento do poder computacional, à grande disponibilidade de dados e ao seu bom desempenho. Atualmente todos os estudos de estado da arte sobre extração automática de relações semânticas usam algum algoritmo de AM, como pode ser visto na subseção 2.1.2 do capítulo anterior.

A abordagem para extração de relações semânticas usando AM é mais complexa que a abordagem baseada em padrões descrita na seção anterior, pois faz uso de recursos e ferramentas mais específicos. Enquanto o primeiro experimento realizado com padrões textuais utilizava um *corpus* com apenas seus sintagmas nominais identificados, os experimentos com AM necessitam de *corpora* enriquecidos com mais informações (etiquetas *part-of-speech*, atributos sintáticos, semânticos, etc.) e com seus termos de interesse e as respectivas relações identificados para realização do treinamento e teste dos modelos computacionais.

Antes de prosseguir, é importante formalizar a tarefa de extração automática de relações semânticas como um problema tratável por modelos de AM. Dessa forma, a tarefa é definida como: dado um par de termos distintos <*termo1, termo2*> em uma sentença, um classificador deve determinar se entre eles existe uma das 7 relações de interesse (Tabela 1.1) ou se não há nenhuma relação (classe negativa). Todos os pares de termos distintos de uma sentença são verificados pelos classificadores. A demarcação dos termos não é realizada pelos classificadores, sendo essa uma tarefa do anotador (Seção 3.2.1) ou da ferramenta ARS (TABA; CASELI, 2012b). Assim, a extração automática de relações semânticas é definida como um problema de classificação com múltiplas classes.

Os experimentos desenvolvidos neste trabalho utilizam modelos de AM supervisionado. De forma simplificada, um modelo supervisionado é treinado sobre um conjunto de dados etiquetados (entradas) com suas classificações corretas (saídas desejadas) – essas etiquetas normalmente são fornecidas por especialistas na tarefa em questão. A partir desse treinamento o modelo pode ser usado para classificar novas instâncias de dados. A abordagem supervisionada difere da não supervisionada e da semissupervisionada em que a não supervisionada não utiliza etiquetas para dados e a semissupervisionada utiliza tanto dados etiquetados quanto não etiquetados no treinamento dos modelos.

Uma etapa importante para a realização de experimentos com AM é a *featurização*, ou decomposição dos dados em atributos. Essa etapa será descrita a seguir. Em seguida serão descritos os experimentos com os dois algoritmos escolhidos para realizar os testes com AM: as árvores de decisão C4.5 (QUINLAN, 1993) e *Support Vector Machines* (VAPNIK, 1995). Concluindo a seção, será mostrada a forma de avaliação dos resultados obtidos.

3.4.1 *Features*

Features (ou *atributos*) são características que descrevem dados e são utilizadas pelos algoritmos de AM como forma de discriminar e classificar instâncias. A seleção de *features* é uma das etapas mais importantes na definição de experimentos com AM já que *features* ruins – que são pouco relevantes para a tarefa em questão ou não auxiliam a separação de instâncias – podem trazer resultados ruins, enquanto *features* adequadas – que sejam relevantes e auxiliem a discriminiação de instâncias – podem trazer bons resultados, mesmo utilizando o mesmo conjunto de treinamento em ambos os casos.

Com base em alguns dos trabalhos vistos na revisão bibliográfica (MINTZ et al., 2009; GIRJU et al., 2010) e na observação dos dados anotados, foram definidas 288 *features* de di-

ferentes níveis linguísticos para realizar a *featurização* dos dados de treinamento, apresentadas nas Tabelas 3.6 a 3.8 a seguir. É importante relembrar que, conforme as definições mostradas na Seção 3.1, uma instância é considerada um par de termos. Portanto, as *features* apresentadas sempre se referem a um par de termos de uma sentença.

Tabela 3.6: Lista de *features* superficiais utilizadas pelos algoritmos de AM

Feature	**Descrição**	**Tipo**
Distância entre termos	A distância (em número de *tokens*) entre os termos	Numérico
Vírgulas entre termos	O número de vírgulas entre os termos	Numérico
Palavras capitalizadas entre termos	O número de palavras capitalizadas entre os termos	Numérico
Termos entre termos	O número de termos que existem entre os dois termos atuais	Numérico
Padrões textuais entre termos	Verdadeiro se algum padrão textual de hiponímia (Figuras 3.3 e 3.4) existe entre os termos, falso caso contrário	Booleano
Ordem dos termos	Verdadeiro se o primeiro termo vem antes do segundo, falso caso contrário	Booleano
Tamanho de cada termo (em *tokens*)	O tamanho de um termo em número de *tokens*	Numérico
Tamanho de cada termo (em caracteres)	O tamanho de um termo em número de caracteres	Numérico
Palavras capitalizadas em cada termo	O número de palavras capitalizadas que compõem um termo	Numérico
Termo completamente capitalizado	Verdadeiro se um termo é escrito todo em maiúsculas, falso caso contrário	Booleano
Parêntese antes de cada termo	Verdadeiro se um parêntese ocorre antes de um termo, falso caso contrário	Booleano
Parêntese depois de cada termo	Verdadeiro se um parêntese ocorre depois de um termo, falso caso contrário	Booleano
Aspas antes de cada termo	Verdadeiro se aspas ocorrem antes de um termo, falso caso contrário	Booleano
Aspas depois de cada termo	Verdadeiro se aspas ocorrem depois de um termo, falso caso contrário	Booleano
Vírgula antes de cada termo	Verdadeiro se uma vírgula ocorre antes de um termo, falso caso contrário	Booleano
Vírgula depois de cada termo	Verdadeiro se uma vírgula ocorre depois de um termo, falso caso contrário	Booleano
Termos contêm números	Verdadeiro se um termo contém números, falso caso contrário	Booleano
Termos começam com “de”	Verdadeiro se um termo começa com a preposição “de”, falso caso contrário	Booleano

As primeiras seis features da Tabela 3.6 representam o contexto entre os dois termos, enquanto as demais descrevem características superficiais de cada termo que podem ajudar no pro-

cesso de aprendizado. As últimas doze *features* são duplicadas (uma para cada termo) e a quinta *feature* ("Padrões textuais entre termos") corresponde, na verdade, a sete *features* (uma para cada padrão das Figuras 3.3 e 3.4). Note que o uso dos padrões textuais como uma *feature* permite que o conhecimento derivado do experimento 1, específico para detecção de instâncias da relação is-a, seja utilizado também pelos métodos de AM. Totalizando as *features*, a Tabela 3.6 engloba 36 *features* superficiais.

Tabela 3.7: Lista de *features* morfológicas utilizadas pelos algoritmos de AM

Feature	**Descrição**	**Tipo**
Etiquetas *part-of-speech* ao redor de cada termo	As etiquetas POS dos *tokens* ao redor dos termos, considerando uma janela de tamanho 3 (3 *tokens* antes e 3 *tokens* depois)	Nominal
Verbo "ser" entre termos	Verdadeiro se o verbo "ser" ocorre entre os termos, falso caso contrário	Booleano
Preposição entre termos	Verdadeiro se uma preposição ocorre entre os termos, falso caso contrário	Booleano
Etiquetas *part-of-speech* de cada termo	Verdadeiro se uma determinada etiqueta POS está presente no termo, falso caso contrário	Booleano
Artigo antes de cada termo	Verdadeiro se um artigo ocorre antes de um termo, falso caso contrário	Booleano

Como na última tabela, as três primeiras *features* da Tabela 3.7 representam o contexto ao redor dos dois termos. Neste caso elas englobam o contexto morfológico ao redor e entre os dois termos. A primeira *feature* na verdade é composta por doze *features*: as etiquetas POS dos três *tokens* antes e depois de cada termo (pode haver sobreposição entre esses *tokens*). Como um termo pode conter vários *tokens*, cada um podendo ter diferentes etiquetas POS, a penúltima *feature*, "Etiquetas *part-of-speech* de cada termo", é dividida em 10 *features* binárias – cada uma delas representa se uma de 10 classes gramaticais ocorre dentro de um termo. Como as duas últimas *features* são duplicadas (uma para cada termo), no total há 36 *features* morfológicas.

Tabela 3.8: Lista de *features* sintáticas utilizadas pelos algoritmos de AM

Feature	**Descrição**	**Tipo**
Etiquetas sintáticas ao redor dos termos	Verdadeiro se uma determinada etiqueta sintática está presente na janela (de tamanho 1) ao redor do termo, falso caso contrário	Booleano
Etiquetas sintáticas de cada termo	Verdadeiro se uma determinada etiqueta sintática está presente no termo, falso caso contrário	Booleano

Finalmente, a Tabela 3.8 traz uma *feature* para representar o contexto sintático e outra

para descrever as características sintáticas de cada termo. A informação utilizada nessas *features* é o papel sintático de cada *token* na árvore de dependências da sentença como etiquetado pelo PALAVRAS (BICK, 2000). O PALAVRAS define um conjunto de 36 etiquetas sintáticas possíveis, e mais de uma etiqueta pode ser atribuída a um *token*. Assim, a primeira *feature* é na verdade composta por 4×36 *features* binárias, ou 36 *features* para cada *token* antes e depois de cada termo. Da mesma forma, a segunda *feature* é composta por 2×36 *features* binárias, 36 para cada termo. No total há 216 *features* sintáticas.

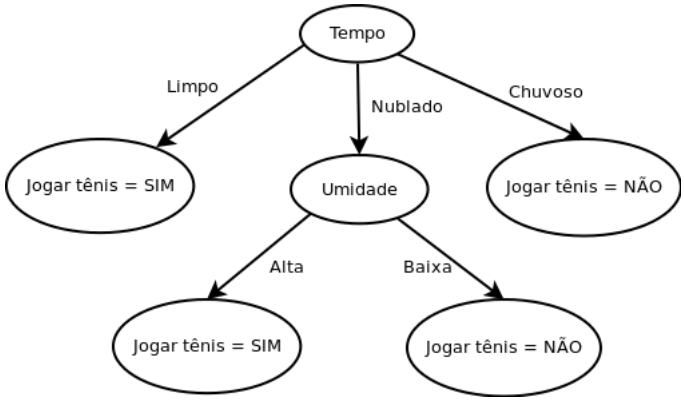
Todas as *features* derivadas somam, portanto, 288 *features*.

3.4.2 Experimento com árvores de decisão

As árvores de decisão são um dos algoritmos mais utilizados em estudos com AM devido à sua simplicidade e bons resultados. Seu funcionamento se baseia na criação de árvores onde cada nodo representa uma tomada de decisão sobre uma variável da representação do problema, com as folhas da árvore indicando a classificação de uma dada instância. Essa simplicidade estrutural torna fácil para um leitor humano observar uma árvore de decisão e compreender seu funcionamento (como pode ser visto na Figura 3.5), diferentemente de outros modelos de AM que são menos interpretáveis.

Outra vantagem das árvores de decisão é que elas tratam naturalmente o problema de classificação com múltiplas classes, que é o caso deste trabalho. Outros classificadores, como as *Support Vector Machines*, que serão detalhadas na próxima subseção, são estritamente binários, sendo necessário o uso de adaptações para tratar o problema de classificação multiclasse.

Figura 3.5: Exemplo de árvore de decisão para a pergunta "É um bom dia para jogar tênis?" baseada em duas variáveis (tempo e umidade)



O algoritmo de árvore de decisão utilizado neste trabalho foi o C4.5 (QUINLAN, 1993),

que baseia a construção das árvores sobre a noção de entropia da teoria da informação (SHANNON, 1948). Simplificadamente, a árvore é construída no sentido *top-down*, ou seja, de cima para baixo, e os nodos mais próximos da raiz tomam decisões sobre as variáveis que trazem maior ganho de informação (maior poder discriminatório). A escolha de colocar as variáveis que têm maior ganho de informação próximas da base torna a árvore mais curta, já que quanto mais a fundo se percorre a árvore, menos decisões precisam ser tomadas.

A implementação do algoritmo C4.5 escolhida para realizar os experimentos foi o J48 do pacote WEKA (HALL et al., 2009) com o parâmetro[6] $C = 0{,}25$. O **experimento 3** consistiu, então, em fornecer as instâncias de relações marcadas no *corpus* CETENFolha (descritos na Seção 3.2) como dados de treinamento ao algoritmo, que foi avaliado usando o método *10-fold cross-validation*. Os resultados serão mostrados e discutidos no próximo capítulo.

3.4.3 Experimento com *Support Vector Machines*

O segundo algoritmo de AM escolhido foram as *Support Vector Machines* (SVMs) (VAPNIK, 1995), classificadores binários que figuram no estado da arte entre os algoritmos de AM devido ao seu bom desempenho. Basicamente, as SVMs utilizam propriedades geométricas para encontrar um hiperplano que melhor separe dois conjuntos de dados; esse hiperplano é encontrado através da resolução de um problema de otimização quadrático.

Dado que as SVMs são classificadores binários mas nossa tarefa envolve o discernimento entre 7 diferentes classes mais uma classe negativa, foi adotada a estratégia *one-vs-all*, onde para cada classe é treinado um classificador que discerne entre essa classe e todas as demais. A classificação de uma instância é feita processando-a por todos os 8 classificadores e atribuindo-lhe a classe do classificador que tenha relatado maior confiança em sua predição.

A implementação de SVM escolhida foi a do software SVM Light[7] (JOACHIMS, 1998), utilizada em (ZELENKO et al., 2003; ZHANG et al., 2006), com *kernel* polinomial de grau 3 e parâmetro[8] $C = 0{,}01$. Não existe uma forma única de determinar os melhores parâmetros a serem utilizados, já que esses dependem dos dados disponíveis. Portanto, eles foram selecionados empiricamente após testes feitos com todas as combinações de graus de 1 a 4 e C de 10^2 a 10^{-4} (em intervalos de potências de 10). Preferiu-se o uso do SVM Light ao invés do Weka pois os testes com o último tomavam cerca de 5 vezes mais tempo que com o primeiro.

[6] Fator de confiança na poda da árvore. Quanto menor, mais podas serão feitas.

[7] `http://svmlight.joachims.org/`

[8] Fator de compensação entre os erros no treinamento e a margem dos vetores de suporte.

Assim, para a realização do **experimento 4**, as instâncias de relações marcadas no *corpus* CETENFolha foram fornecidas como treinamento para 8 classificadores – um para cada uma das 7 relações mais a classe negativa (quando não há relação entre os termos). Da mesma forma que com as árvores de decisão, foi utilizado o método de avaliação *10-fold cross-validation*. Os resultados obtidos serão mostrados e discutidos no próximo capítulo.

3.4.4 Experimento com diferentes subconjuntos de *features*

Com o propósito de verificar como cada subconjunto de *features* (veja seção 3.4.1) influencia nos resultados obtidos pelos classificadores, foram realizados experimentos com os mesmos dados que os experimentos anteriores (cerca de 100 mil instâncias marcadas do CETENFolha) mas *featurizados* de diferentes maneiras.

Dessa forma, o **experimento 5** consistiu na realização de 7 novos testes com os classificadores de árvore de decisão e SVM. Cada teste foi feito com um determiando subconjunto de *features*, conforme apresentado na Tabela 3.9. O subconjunto 7, cobrindo todas as *features*, equivale aos experimentos mencionados anteriormente (experimento 3 para árvores de decisão e experimento 4 para SVM). Foi realizada a avaliação usando *10-fold cross-validation* para cada um dos subconjuntos descritos e para ambos os classificadores, utilizando os mesmos parâmetros descritos anteriormente para cada classificador. Os resultados serão descritos no capítulo seguinte.

Tabela 3.9: Subconjuntos de features utilizadas

	Features		
Subconjunto	Superficiais	Morfológicas	Sintáticas
1	X		
2		X	
3			X
4	X	X	
5	X		X
6		X	X
7	X	X	X

3.4.5 Experimento com classificadores testados sobre *corpus* FAPESP

Nos experimentos citados até então, ambos os algoritmos de AM – árvores de decisão e SVM – foram treinados e testados no mesmo *corpus*: usando as instâncias anotadas do *corpus* CETENFolha e aplicando o método *10-fold cross-validation*. Embora essa seja uma boa forma de obter uma

estimativa do desempenho de um classificador quando confrontado com dados ainda não vistos[9], idealmente devem ser realizados testes com dados distintos dos utilizados no treinamento para averiguar de fato a capacidade de generalização dos classificadores treinados.

Com essa finalidade, o **experimento 6** consistiu em treinar um classificador de árvore de decisão e um classificador SVM sobre todos os dados anotados do CETENFolha, com todas as *features* (ou seja, os classificadores resultantes dos experimentos 3 e 4, respectivamente), e testá-los sobre as cerca de 500 sentenças anotadas do *corpus* FAPESP. Além de avaliar a capacidade de generalização dos classificadores treinados, este experimento busca verificar se tais classificadores podem ser aplicados a *corpora* de gêneros distintos, já que o *corpus* de treinamento (CETENFolha) é jornalístico e o *corpus* de teste (FAPESP) é de divulgação científica.

3.4.6 Resumo dos experimentos

Conforme apresentado, foram realizados 6 experimentos no decorrer deste mestrado, resumidos na Tabela 3.10.

Tabela 3.10: Resumo dos experimentos realizados

Experimento	Abordagem	Descrição
1	Padrões textuais	Extração da relação de hiponímia do *corpus* FAPESP com os padrões de Freitas e Quental (2007) e novos padrões encontrados com o algoritmo de Hearst (1992)
2	Padrões textuais	Extração das 7 relações semânticas de interesse do *corpus* CETENFolha com padrões definidos manualmente e novos padrões encontrados com o algoritmo de Hearst (1992)
3	AM	Treinamento e teste de classificador de árvore de decisão sobre o *corpus* CETENFolha avaliado por *10-fold cross-validation*
4	AM	Treinamento e teste de classificador SVM sobre o *corpus* CETENFolha avaliado por *10-fold cross-validation*
5	AM	Treinamento e teste de ambos os classificadores sobre o *corpus* CETENFolha com diferentes subconjuntos de *features* avaliados por *10-fold cross-validation*
6	AM	Treinamento de ambos os classificadores sobre o *corpus* CETENFolha e testados sobre o *corpus* FAPESP

[9]Lembrando que no *10-fold cross-validation* o *corpus* de treinamento é dividido em 10 partições, sendo 9 usadas para treinar o classificador e a última para testá-lo. Esse processo é repetido 10 vezes, sempre variando a partição deixada para teste, e o resultado final é dado pela média dos resultados para as 10 iterações.

3.5 Avaliação

A avaliação dos métodos de extração automática de relações semânticas pode ser realizada automaticamente comparando-se as instâncias de relações identificadas pelos métodos automáticos com aquelas pertencentes a um conjunto de referência (com instâncias corretamente anotadas). Essa estratégia foi usada a partir do experimento 2 e, principalmente, na avaliação dos métodos que usam AM. Apenas para o primeiro experimento não foi possível realizar a avaliação automática porque não havia um conjunto de referência para o *corpus* utilizado, sendo necessário, portanto, realizar a avaliação manual das instâncias extraídas no experimento 1 com o cálculo apenas da precisão.

A partir das instâncias extraídas automaticamente (A) e das instâncias de referência (R), as métricas de precisão, cobertura e medida-F são calculadas como mostrado nas equações a seguir:

$$\text{Precisão} = \frac{\text{A} \cap \text{R}}{\text{A}} \tag{3.1}$$

$$\text{Cobertura} = \frac{\text{A} \cap \text{R}}{\text{R}} \tag{3.2}$$

$$\text{Medida-F} = 2 \times \frac{\text{Precisão} \times \text{Cobertura}}{\text{Precisão + Cobertura}} \tag{3.3}$$

Assim, a precisão representa a capacidade de um sistema de obter apenas resultados corretos, enquanto a cobertura indica a sua capacidade de obter a maior quantidade possível de instâncias corretas do conjunto de dados. A medida-F, por sua vez, é a média harmônica das duas medidas anteriores. Essas métricas são utilizadas na grande maioria dos trabalhos investigados na revisão bibliográfica. Portanto, o seu uso permite a comparação das abordagens implementadas no âmbito deste trabalho com os resultados do estado da arte encontrados na literatura. A faixa possível de valores para essas medidas vai de 0 a 1, sendo que quanto mais próximos de 1, melhor o desempenho do sistema.

4 Resultados e discussão

Neste capítulo são apresentados os resultados obtidos pelos experimentos descritos no capítulo anterior, além de uma discussão sobre os mesmos.

4.1 Experimento 1

O experimento 1 consistiu na identificação de instâncias da relação de hiponímia (is-a) através da aplicação dos 4 padrões de Freitas e Quental (2007) e da aplicação do algoritmo de Hearst (1992) sobre o *corpus* FAPESP, que auxiliou na definição de 3 novos padrões. Os resultados obtidos com a aplicação dos 4 padrões de Freitas e Quental (2007) foram próximos com os obtidos pelas pesquisadoras: de 1309 relações encontradas, 816 (62,34%) foram consideradas corretas por avaliação manual.

Vale mencionar que outras 317 relações encontradas (24,22%) estavam parcialmente corretas, com o erro localizado na identificação dos sintagmas nominais devido a sintagmas preposicionados – alguns sintagmas englobavam mais ou menos palavras do que deveriam. Por exemplo, na sentença "(...) [centros urbanos] em [crescimento contínuo] como [Uberaba] (...)", o SN[1] correto para a relação deveria englobar os dois primeiros SNs, começando em "centros" e terminando em "contínuo". Essa identificação incorreta de SNs resulta na relação parcialmente correta *is-a(Uberaba, crescimento contínuo)*, quando deveria ser *is-a(Uberaba, centros urbanos em crescimento contínuo)*.

O algoritmo iterativo de Hearst (1992), cuja finalidade é auxiliar na descoberta de novos padrões que indiquem a relação de interesse, foi aplicado sobre o *corpus*. As instâncias de relações de hiponímia constantes na base de dados do projeto OMCS-Br foram utilizadas como semente para o segundo passo do algoritmo. Após a primeira iteração, foram encontrados os três novos padrões descritos anteriormente na Tabela 3.4. Apenas uma iteração do algoritmo foi executada devido à limitações de tempo.

[1] Os sintagmas nominais aparecem delimitados por colchetes

Os três novos padrões foram então procurados sobre o *corpus*, resultando em 854 novas instâncias encontradas, das quais 605 (70,84%) foram consideradas corretas e 163 (19,09%) estavam parcialmente corretas.

A Tabela 4.1 sumariza os resultados obtidos e a Tabela 4.2 traz alguns exemplos de relações corretamente identificadas.

Tabela 4.1: Experimento 1 – Número de relações corretas, parcialmente corretas e incorretas identificadas usando os padrões das Figuras 3.3 e 3.4

Padrões aplicados	**Corretas**	**Parcialmente corretas**	**Incorretas**	**Total**
Freitas e Quental (Figura 3.3)	816 (62,34%)	317 (24,22%)	176 (13,44%)	1309 (100%)
Novos padrões (Figura 3.4)	605 (70,84%)	163 (19,09%)	86 (10,07%)	854 (100%)
Total	1421 (65,70%)	480 (22,19%)	262 (12,11%)	2163 (100%)

Tabela 4.2: Experimento 1 – Exemplos de instâncias de relações identificadas corretamente, os contextos em que ocorreram e os padrões que as encontraram

Relação	**Contexto original**	**Padrão**
is-a(Rio Branco, capital)	...aeroportos de capitais como Rio Branco...	1
is-a(formigas, insetos)	...formigas e outros insetos acabam grudados...	2
is-a(macrófagos, células)	...ativa células chamadas macrófagos, que...	4
is-a(cerveja, bebida)	...tomar cerveja, vinho ou qualquer outra bebida...	5
is-a(dióxido de estanho, semicondutor)	...o dióxido de estanho é um semicondutor...	6

A partir da avaliação manual das instâncias da relação is-a extraídas automaticamente no experimento 1, usando padrões textuais, pode-se concluir que a precisão foi de 62,34% quando só os padrões de Freitas e Quental (2007) foram usados, subiu para 70,84% quando só os padrões derivados deste trabalho foram aplicados e ficou em 65,70% quando ambos os padrões foram empregados.

4.2 Experimento 2

O experimento 2 consistiu na definição de 17 padrões textuais (ver Tabelas 3.4 e 3.5) para extrair as outras 6 relações semânticas enfocadas neste trabalho. Esses padrões, juntamente com os padrões de Freitas e Quental (2007) e os 3 padrões encontrados anteriormente para a relação de hiponímia (Tabela 3.4), foram aplicados sobre as mesmas sentenças anotadas do *corpus* CETENFolha (conjunto de referência). Desse modo foi possível calcular não apenas a precisão, mas também a cobertura e a medida-F do método de extração de relações semânticas usando padrões textuais. Os valores de precisão, cobertura e medida-F, obtidos para cada padrão, podem ser consultados na Tabela 4.3[2].

A precisão obtida na extração de instâncias da relação is-a no *corpus* CETENFolha, 61,1%, foi bem próxima à obtida no experimento 1 usando os mesmos 7 padrões textuais aplicados ao *corpus* FAPESP (65,7%). Resumindo os valores de precisão apresentados na Tabela 4.3, as relações com padrões mais precisos foram, em ordem decrescente de precisão: effect-of (69,2%), is-a (61,1%), property-of (57,5%), used-for (42,3%), part-of (12,3%), location-of (9%) e made-of (1,3%). A precisão média da estratégia de padrões textuais aplicada ao *corpus* CETENFolha foi, portanto, de 36,5%.

A cobertura média, 18,2%, foi obtida dividindo o número total de instâncias corretas encontradas (2140) pelo número total de instâncias de relações marcadas manualmente no *corpus* (11757). Como esperado, os valores de cobertura no geral foram bastante baixos, corroborando o que já foi levantado na revisão bibliográfica (YAP; BALDWIN, 2009).

Buscando melhorar a cobertura dos resultados, métodos de AM foram investigados e os resultados dos experimentos com esses métodos são mostrados nas próximas seções.

4.3 Experimento 3

O experimento 3 consistiu no treinamento do algoritmo de árvore de decisão C4.5 (QUINLAN, 1993) sobre as cerca de 100 mil instâncias de treinamento do *corpus* CETENFolha (Tabela 3.1) usando todas as *features* descritas na Seção 3.4.1. O resultado da aplicação do *10-fold cross-validation* com árvores de decisão é mostrado na Tabela 4.4.

A partir dos resultados do experimento 3, é possível notar um aumento significativo na cobertura média quando árvores de decisão são usadas em comparação ao uso de padrões textuais.

[2]É importante mencionar que o total de instâncias encontradas para uma relação pode ser diferente da soma do número de instâncias encontradas por cada padrão, já que dois padrões distintos podem encontrar a mesma instância.

Tabela 4.3: Experimento 2 – Resultados da aplicação de todos os 24 padrões textuais no *corpus* CETENFolha em termos de precisão, cobertura e medida-F

Relação	**#**	**Instâncias encontradas**	**Instâncias corretas**	**Precisão**	**Cobertura**	**Medida-F**
is-a	1	0	0	0,0%	0,0%	0,0%
	2	7	3	42,8%	0,1%	0,2%
	3	0	0	0,0%	0,0%	0,0%
	4	5	4	80,0%	0,1%	0,2%
	5	0	0	0,0%	0,0%	0,0%
	6	32	19	59,4%	0,7%	1,4%
	7	10	7	70,0%	0,2%	0,4%
Total		54	33	61,1%	1,2%	2,3%
property-of	1	1048	524	50,0%	10,7%	17,6%
	2	2287	1417	61,9%	28,9%	39,4%
	3	10	10	100,0%	0,2%	0,4%
	4	59	30	50,8%	0,6%	1,2%
Total		3397	1954	57,5%	39,8%	47,0%
part-of	1	48	2	4,2%	0,1%	0,2%
	2	5	3	60,0%	0,1%	0,2%
	3	2	2	100,0%	0,1%	0,2%
Total		57	7	12,3%	0,3%	0,6%
location-of	1	6	5	83,3%	0,3%	0,6%
	2	2	2	100,0%	0,1%	0,2%
	3	4	4	100,0%	0,2%	0,4%
	4	585	43	7,3%	2,5%	3,7%
Total		597	54	9,0%	3,2%	4,7%
effect-of	1	7	5	71,4%	3,1%	5,9%
	2	6	4	66,7%	2,4%	4,6%
Total		13	9	69,2%	5,5%	10,2%
made-of	1	1660	21	1,3%	28,4%	2,5%
	2	1	1	100,0%	1,3%	2,6%
Total		1661	22	1,3%	29,7%	2,5%
used-for	1	1	1	100,0%	0,7%	1,4%
	2	84	35	42,7%	25,7%	32,1%
Total		85	36	42,3%	26,5%	32,6%
Média		5864	2140	36,5%	18,2%	24,3%

Esse aumento impactou diretamente no aumento na média harmônica de precisão e cobertura, a medida-F. Assim, ordenando as relações decrescentemente pela medida-F, tem-se: property-of (84,9%), is-a (65,0%), part-of (47,7%), location-of (31,8%), effect-of (16,0%), used-for (8,0%) e made-of (5,2%).

No entanto, é importante observar que as relações made-of e used-for individualmente

Tabela 4.4: Experimento 3 – Resultados do *10-fold cross-validation* para o algoritmo de árvore de decisão em termos de precisão, cobertura e medida-F usando todas as *features* da Seção 3.4.1

Relação	**Precisão**	**Cobertura**	**Medida-F**
property-of	90,5%	80,0%	84,9%
is-a	76,9%	56,3%	65,0%
part-of	66,4%	37,2%	47,7%
location-of	62,2%	21,4%	31,8%
effect-of	34,0%	10,4%	16,0%
made-of	33,3%	2,8%	5,2%
used-for	17,5%	5,1%	8,0%
Média	54,4%	30,4%	39,0%

obtiveram maior cobertura utilizando padrões (29,7% e 26,5%, respectivamente) do que árvores de decisão (2,8% e 5,1%). Ademais, a relação used-for obteve uma medida-F maior com padrões: 32,6% contra 8,0%, o que mostra que pelo menos para essa relação, a abordagem de padrões textuais traz melhores resultados.

Em resumo, tanto a precisão média, 54,4%, quanto a cobertura média, 30,4%, obtidas com árvores de decisão foram maiores do que as obtidas com padrões textuais, 36,5% e 18,2%, respectivamente. Esse ganho de desempenho se reflete na medida-F média obtida com o uso de árvores de decisão, 39%, que também foi bem maior do que a obtida com padrões textuais, 24,3%. Esses resultados demonstram a efetividade do uso de árvores de decisão na extração automática de relações semânticas, com a exceção da relação used-for, que obteve melhores resultados com a aplicação de padrões textuais.

4.4 Experimento 4

O experimento 4 consistiu no treinamento do algoritmo *Support Vector Machine* (VAPNIK, 1995) sobre as cerca de 100 mil instâncias de treinamento do *corpus* CETENFolha (Tabela 3.1) usando todas as *features* descritas na Seção 3.4.1. Os resultados da avaliação com *10-fold cross-validation* estão descritos na Tabela 4.5 a seguir.

Assim como verificado no experimento 3, o experimento 4 apresentou um aumento significativo na cobertura quando SVMs são usadas em comparação ao uso de padrões textuais, o que também impactou diretamente nos valores de medida-F. A ordem decrescente das relações, pela medida-F, manteve-se quase a mesma obtida com árvores de decisão, com melhora significativa de desempenho nas relações made-of, used-for e effect-of: property-of (85,4%), is-a (71,0%), part-of

Tabela 4.5: Experimento 4 – Resultados do *10-fold cross-validation* para o algoritmo SVM em termos de precisão, cobertura e medida-F usando todas as *features* da Seção 3.4.1

Relação	**Precisão**	**Cobertura**	**Medida-F**
property-of	89,6%	81,6%	85,4%
is-a	78,2%	65,1%	71,0%
part-of	56,9%	41,4%	47,9%
location-of	51,8%	27,8%	36,2%
effect-of	45,5%	16,9%	24,6%
made-of	58,2%	24,3%	34,3%
used-for	50,8%	17,7%	26,2%
Média	61,6%	39,2%	47,9%

(47,9%), location-of (36,2%), made-of (34,3%), used-for (26,2%) e effect-of (24,6%).

Contudo, as relações made-of e used-for ainda têm maior cobertura com padrões textuais – 29,7% e 26,5% contra 24,3% e 17,7% – e a relação used-for utilizando padrões ainda apresenta maior medida-F (32,6% contra 26,2%). As SVMs trazem um grande ganho de cobertura para algumas relações em comparação com as árvores de decisão, mas esse aumento não é suficiente para superar a medida-F da relação used-for com padrões.

A precisão, cobertura e medida-F médias obtidas com SVMs foram 61,6%, 39,2% e 47,9%, respectivamente, valores superiores aos obtidos com árvores de decisão (54,4%, 30,4% e 39,0%, respectivamente). Consequentemente, os valores médios obtidos com SVMs foram também maiores do que os obtidos com padrões textuais. Esses resultados também demonstram a efetividade do uso de SVMs na extração automática de relações semânticas, novamente com a ressalva da relação used-for, que ainda apresenta melhores resultados com a aplicação de padrões textuais.

4.5 Experimento 5

O experimento 5 consistiu no treinamento dos classificadores de árvore de decisão e SVM com todos os dados anotados do CETENFolha mas com diferentes subconjuntos de *features* (conforme a Tabela 3.9). Esse experimento foi realizado com o intuito de se verificar o impacto de cada conjunto de *features* no AM.

O desempenho dos classificadores treinados com cada subconjunto de features, analisado através de *10-fold cross-validation*, é mostrado nas Tabelas 4.6 a 4.12 a seguir. A partir dos valores apresentados nestas tabelas é possível notar que os métodos de AM obtiveram melhor desempe-

nho quando todas as *features* são utilizadas no treinamento dos classificadores (Tabela 4.12, que reproduz os valores dos experimentos 3 e 4).

Tabela 4.6: Experimento 5 – Resultados do *10-fold cross-validation* para os classificadores treinados com as *features* do subconjunto 1

	Árvore de decisão			**SVM**		
Relação	**Precisão**	**Cobertura**	**Medida-F**	**Precisão**	**Cobertura**	**Medida-F**
property-of	77,3%	64,6%	70,4%	0,0%	0,0%	0,0%
is-a	74,7%	49,7%	59,7%	0,0%	0,0%	0,0%
part-of	69,0%	29,0%	40,8%	0,0%	0,0%	0,0%
location-of	66,2%	15,3%	24,8%	0,0%	0,0%	0,0%
effect-of	25,0%	3,7%	6,4%	0,0%	0,0%	0,0%
made-of	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%
used-for	14,3%	0,7%	1,4%	0,0%	0,0%	0,0%
Média	46,6%	23,3%	31,7%	0,0%	0,0%	0,0%

Tabela 4.7: Experimento 5 – Resultados do *10-fold cross-validation* para os classificadores treinados com as *features* do subconjunto 2

	Árvore de decisão			**SVM**		
Relação	**Precisão**	**Cobertura**	**Medida-F**	**Precisão**	**Cobertura**	**Medida-F**
property-of	88,8%	72,9%	80,1%	53,4%	85,4%	67,8%
is-a	73,1%	36,9%	49,1%	44,8%	67,5%	53,8%
part-of	30,0%	0,2%	0,3%	33,3%	44,0%	37,9%
location-of	50,0%	0,7%	1,3%	33,9%	39,5%	36,5%
effect-of	20,0%	1,8%	3,4%	25,2%	28,1%	26,6%
made-of	0,0%	0,0%	0,0%	20,3%	25,7%	22,7%
used-for	0,0%	0,0%	0,0%	26,7%	29,2%	27,9%
Média	37,4%	16,1%	22,5%	34,4%	45,6%	39,2%

Tabela 4.8: Experimento 5 – Resultados do *10-fold cross-validation* para os classificadores treinados com as *features* do subconjunto 3

	Árvore de decisão			**SVM**		
Relação	**Precisão**	**Cobertura**	**Medida-F**	**Precisão**	**Cobertura**	**Medida-F**
property-of	66,2%	27,8%	39,2%	0,0%	0,0%	0,0%
is-a	55,0%	7,3%	12,9%	0,0%	0,0%	0,0%
part-of	25,0%	0,1%	0,2%	0,0%	0,0%	0,0%
location-of	66,7%	0,1%	0,2%	1,4%	100,0%	2,8%
effect-of	41,4%	7,4%	12,5%	0,0%	0,0%	0,0%
made-of	50,0%	4,2%	7,8%	0,0%	0,0%	0,0%
used-for	45,5%	3,7%	6,8%	0,0%	0,0%	0,0%
Média	50,0%	7,2%	12,6%	0,2%	14,3%	0,4%

Tabela 4.9: Experimento 5 – Resultados do *10-fold cross-validation* para os classificadores treinados com as *features* do subconjunto 4

	Árvore de decisão			**SVM**		
Relação	**Precisão**	**Cobertura**	**Medida-F**	**Precisão**	**Cobertura**	**Medida-F**
property-of	92,0%	79,5%	85,3%	90,1%	80,8%	85,2%
is-a	78,6%	55,5%	65,0%	78,3%	61,5%	68,9%
part-of	69,8%	34,3%	46,0%	63,1%	36,3%	46,1%
location-of	67,5%	18,8%	29,4%	56,7%	19,9%	29,4%
effect-of	20,7%	3,7%	6,3%	40,3%	13,7%	20,5%
made-of	50,0%	1,4%	2,7%	67,0%	21,4%	32,5%
used-for	15,0%	2,2%	3,8%	57,3%	15,4%	24,3%
Média	56,2%	27,9%	37,3%	64,7%	35,6%	45,9%

Tabela 4.10: Experimento 5 – Resultados do *10-fold cross-validation* para os classificadores treinados com as *features* do subconjunto 5

	Árvore de decisão			**SVM**		
Relação	**Precisão**	**Cobertura**	**Medida-F**	**Precisão**	**Cobertura**	**Medida-F**
property-of	90,0%	79,1%	84,2%	80,4%	66,5%	72,8%
is-a	76,1%	54,8%	63,7%	84,1%	13,6%	23,4%
part-of	66,1%	35,4%	46,1%	75,0%	0,4%	0,9%
location-of	58,3%	19,7%	29,4%	30,0%	0,2%	0,3%
effect-of	22,9%	6,7%	10,4%	66,7%	15,0%	24,5%
made-of	37,5%	4,2%	7,6%	0,0%	0,0%	0,0%
used-for	13,3%	2,9%	4,8%	0,0%	0,0%	0,0%
Média	52,0%	29,0%	37,2%	48,0%	13,7%	21,3%

Tabela 4.11: Experimento 5 – Resultados do *10-fold cross-validation* para os classificadores treinados com as *features* do subconjunto 6

	Árvore de decisão			**SVM**		
Relação	**Precisão**	**Cobertura**	**Medida-F**	**Precisão**	**Cobertura**	**Medida-F**
property-of	88,2%	74,8%	80,9%	88,4%	79,0%	83,5%
is-a	71,4%	42,2%	53,0%	75,8%	58,2%	65,8%
part-of	41,5%	1,1%	2,2%	53,1%	29,0%	37,5%
location-of	46,8%	3,0%	5,7%	47,3%	19,8%	27,9%
effect-of	31,7%	8,0%	12,7%	41,4%	17,5%	24,6%
made-of	50,0%	2,8%	5,3%	60,8%	24,3%	34,7%
used-for	0,0%	0,0%	0,0%	54,4%	16,9%	25,8%
Média	47,1%	18,8%	26,9%	60,2%	35,0%	44,2%

Tabela 4.12: Experimento 5 – Resultados do *10-fold cross-validation* para os classificadores treinados com as *features* do subconjunto 7

	Árvore de decisão			**SVM**		
Relação	**Precisão**	**Cobertura**	**Medida-F**	**Precisão**	**Cobertura**	**Medida-F**
property-of	90,5%	80,0%	84,9%	89,6%	81,6%	85,4%
is-a	76,9%	56,3%	65,0%	78,2%	65,1%	71,0%
part-of	66,4%	37,2%	47,7%	56,9%	41,4%	47,9%
location-of	62,2%	21,4%	31,8%	51,8%	27,8%	36,2%
effect-of	34,0%	10,4%	16,0%	45,5%	16,9%	24,6%
made-of	33,3%	2,8%	5,2%	58,2%	24,3%	34,3%
used-for	17,5%	5,1%	8,0%	50,8%	17,7%	26,2%
Média	54,4%	30,4%	39,0%	61,6%	39,2%	47,9%

4.6 Experimento 6

O último experimento realizado neste trabalho, o experimento 6, consistiu no treinamento de um classificador de árvore de decisão e um SVM sobre todas as instâncias de treinamento do *corpus* CETENFolha, seguido do teste sobre as sentenças anotadas do *corpus* FAPESP. O intuito deste experimento foi, então, verificar o desempenho dos classificadores sobre dados inéditos e a adequação do *corpus* de treinamento, que é jornalístico, sobre um *corpus* de gênero distinto (divulgação científica). Os resultados do experimento 6 são apresentados nas Tabelas 4.13 e 4.14 a seguir.

Tabela 4.13: Experimento 6 – Resultados do classificador de árvore de decisão testado sobre o *corpus* FAPESP, em termos de precisão, cobertura e medida-F

Relação	**Precisão**	**Cobertura**	**Medida-F**
property-of	89,1%	83,3%	86,1%
is-a	60,0%	26,2%	36,4%
part-of	59,5%	35,0%	44,1%
location-of	39,1%	10,9%	17,0%
effect-of	40,0%	7,1%	12,1%
made-of	0,0%	0,0%	0,0%
used-for	38,5%	7,4%	12,3%
Média	23,3%	22,8%	23,0%

Tabela 4.14: Experimento 6 – Resultados do classificador SVM testado sobre o *corpus* FAPESP, em termos de precisão, cobertura e medida-F

Relação	**Precisão**	**Cobertura**	**Medida-F**
property-of	91,0%	84,3%	87,5%
is-a	57,6%	32,1%	41,2%
part-of	52,1%	37,1%	43,3%
location-of	39,6%	17,7%	24,5%
effect-of	50,0%	8,3%	14,3%
made-of	36,4%	7,3%	12,1%
used-for	40,9%	13,2%	20,0%
Média	52,5%	28,6%	37,0%

4.7 Discussão

4.7.1 Estratégia de padrões textuais

Os experimentos com padrões textuais confirmaram o que já havia sido discutido no levantamento bibliográfico para determinadas relações: a extração de relações semânticas com padrões textuais tem boa precisão mas baixa cobertura. Isso pode ser observado especialmente na relação is-a, que obteve 61% de precisão e apenas 1% de cobertura. No entanto, as relações part-of, location-of e made-of tiveram precisão baixa; como discutido em (GIRJU et al., 2010), um dos motivos para esse resultado pode ser o fato de essas relações terem padrões muito ambíguos em nível superificial. De fato, as relações part-of e location-of muitas vezes ocorrem juntas (p. ex. "o cômodo de a casa" implica as relações part-of(casa, cômodo) e location-of(cômodo, casa)), dificultando a distinção entre as duas relações. O padrão mais comum para a relação made-of ("T1_N de T2_N") também

é muito ambíguo, podendo indicar outras relações como part-of e location-of. Essa ambiguidade é refletida em uma menor precisão no resultado dessas relações.

O bom desempenho da relação property-of em termos de cobertura pode ser explicado pela sua simplicidade: a grande maioria dos substantivos seguidos de adjetivos e vice-versa configuram uma instância dessa relação. Assim, ela pode ser encontrada facilmente pelos padrões definidos, especialmente "T1_N T2_ADJ" e "T2_ADJ T1_N". As relações made-of e used-for, no experimento 2, obtiveram também uma cobertura um pouco maior que a média pois têm menos exemplos de treinamento. Assim, mesmo algumas poucas dezenas de instâncias corretas já têm impacto significativo.

O experimento 1 mostrou também que o identificador de sintagmas nominais tem influência considerável nos resultados obtidos, visto que 480 (22,19%) do total de 2163 instâncias da relação de hiponímia encontradas estavam parcialmente corretas devido a problemas na identificação dos SNs.

Os resultados dos experimentos com padrões textuais mostram que determinadas relações são mais simples de serem extraídas, como as relações is-a e property-of, enquanto outras, como part-of e made-of, são mais difíceis devido à ambiguidade e plasticidade inerentes à língua. A abordagem de AM, que engloba conhecimento linguístico mais profundo (codificado nas diversas *features* utilizadas), atenua esse problema e traz maior cobertura na extração de relações semânticas.

4.7.2 Estratégia de aprendizado de máquina

Como pode ser observado comparando-se os resultados do experimento 2 (Tabela 4.3) com os valores das Tabelas 4.4 e 4.5, os métodos de AM superam quase todos os resultados obtidos pela aplicação de padrões textuais para a tarefa de extração de relações semânticas, com exceção da relação used-for. Pode-se dizer que as *features* definidas (Seção 3.4.1) conseguem codificar boa parte do conhecimento implícito nos padrões textuais utilizados. De fato, uma das *features* utilizadas é a existência ou não de um dos sete padrões textuais para a relação is-a observados entre termos.

As relações que obtiveram melhores resultados, tanto com as árvores de decisão quanto com SVMs – property-of, is-a, part-of e location-of – são as que tinham maior número de exemplos de treinamento. Esse fato pode ser observado mais facilmente na Tabela 4.15. Isso já era esperado devido à natureza dos métodos de aprendizado supervisionado, que necessitam de grande quantidade de dados de treinamento para obterem bom desempenho. A anotação de mais exemplos para as classes com menores amostras é essencial para melhorar o desempenho dos classificadores para

essas classes.

Tabela 4.15: Resumo dos resultados (em termos de medida-F) da aplicação dos métodos de AM com o número de instâncias de treinamento para cada relação

Relação	**Instâncias de treinamento**	**Árvores de decisão**	**SVM**
property-of	4909	84,9%	85,4%
is-a	2816	65,0%	71,0%
part-of	1976	47,7%	47,9%
location-of	1683	31,8%	36,2%
effect-of	163	16,0%	24,6%
made-of	74	5,2%	34,3%
used-for	136	8,0%	26,2%

Entre os resultados dos algoritmos de árvore de decisão (Tabela 4.4) e SVM (Tabela 4.5), percebe-se que, no geral, o último obteve melhores resultados que o primeiro, com uma diferença de 8,9% na medida-F média. Isso já era esperado, já que SVMs estão entre os classificadores de estado da arte atualmente. No entanto, quando se observa o desempenho individual para cada relação, observa-se que as relações property-of e part-of extraídas por árvores obtiveram resultados bastante próximos das SVMs: 84,9% contra 85,4% e 47,7% contra 47,9%, respectivamente. Para todas as demais relações, as SVMs obtiveram resultados com diferença expressiva, em especial as relações made-of e used-for, que tiveram uma melhora de 29,1 e 18,2 pontos percentuais, respectivamente, em relação às árvores.

Também é pertinente ressaltar que, no geral, as SVMs apresentam melhor precisão mas principalmente uma melhor cobertura que as árvores. Isso pode indicar que as SVMs têm uma maior capacidade de generalização, encontrando mais instâncias de relações. Essa capacidade pode explicar o fato de as SVMs terem obtido resultados melhores com as relações menos representadas no conjunto de dados de treinamento, como made-of e used-for.

O experimento 5, que utiliza diferentes subconjuntos de *features*, mostrou que o classificador SVM usando apenas as *features* superficiais (Tabelas 4.6) não consegue distinguir entre as diferentes relações e atribuiu todas as instâncias de teste à classe negativa, resultando em uma medida-F de 0%. Um resultado similar ocorre nos testes com apenas as *features* sintáticas (Tabela 4.8), mas nesse caso todas as instâncias são atribuídas à classe location-of. No entanto, em ambos os casos as árvores de decisão realizaram a classificação normalmente.

Um possível motivo para esse resultado pode ser que as *features* superficiais ou sintáticas em separado não são adequadas o suficiente para realizar a distinção entre as diferentes classes, ao menos utilizando o classificador SVM. Outra possível explicação é o problema do desbalan-

ceamento de classes, já que a classe negativa é cerca de dez vezes maior que a soma das demais classes (Tabela 3.1). Esse desbalanceamento favorece a classificação de instâncias como negativas, muitas vezes prejudicando o resultado final das classificações. O tratamento desse problema pode trazer benefícios para os experimentos tanto com árvores quanto SVMs.

O experimento 5 também mostrou que as *features* sintáticas tiveram um impacto positivo mas pequeno nos classificadores. Isso pode ter acontecido pelo fato das informações sintáticas contidas no *corpus* CETENFolha não se distinguirem muito dos traços morfológicos, já que tanto as informações sintáticas quanto morfológicas têm a forma de etiquetas que se associam a *tokens*. É possível que o uso da sintaxe em uma forma mais sofisticada – p. ex. como caminhos de dependência entre termos, como analisado em (SNOW et al., 2005) e (ZELENKO et al., 2003) – possa trazer benefícios. A sintaxe nessa forma não foi utilizada pois o *corpus* CETENFolha não contém esse tipo de marcação, e a sua reetiquetação para obter essas informações demandaria muito tempo.

A aplicação dos classificadores treinados com o CETENFolha e testados sobre o *corpus* FAPESP, que compõe o experimento 6, mostra que os algoritmos utilizados no treinamento tiveram boa capacidade de generalização quando confrontrados com um *corpus* de gênero distinto: o CETENFolha é jornalístico, enquanto o FAPESP é de divulgação científica. Os valores obtidos (Tabelas 4.13 e 4.14) ficaram próximos dos resultados obtidos através de *10-fold cross-validation* (Tabelas 4.4 e 4.5), mostrando a aplicabilidade dos classificadores treinados em novos textos. Apesar disso, é importante ressaltar que o tipo de linguagem utilizado no *corpus* FAPESP não difere muito do utilizado no CETENFolha; provavelmente a aplicação dos classificadores em textos de gêneros textuais mais distintos, como o literário, traria resultados mais fracos.

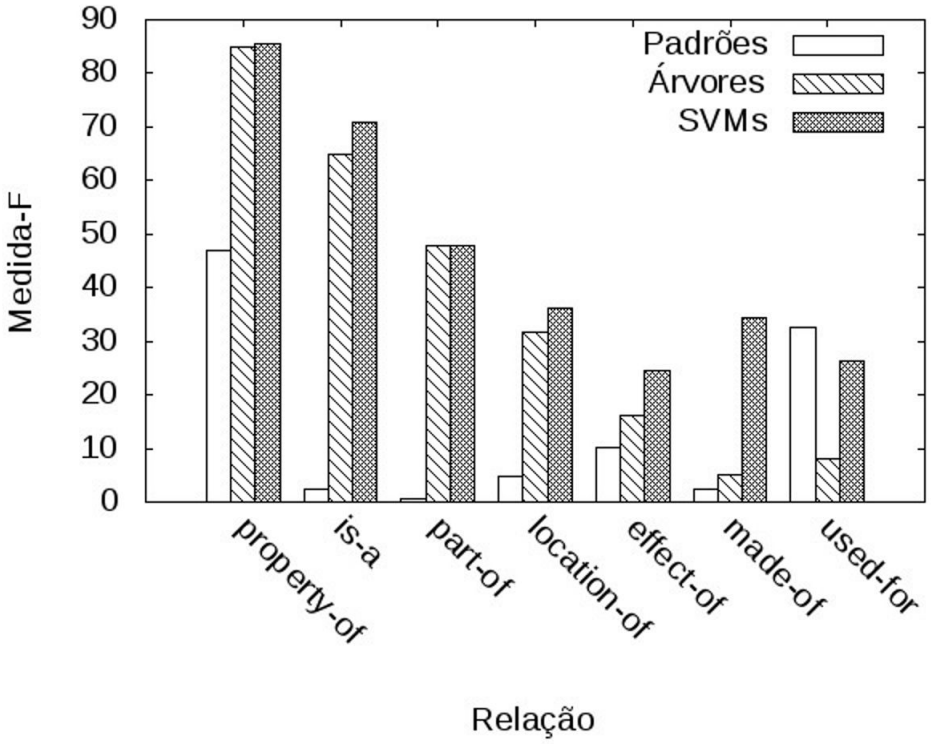
4.7.3 Comparação entre as duas estratégias

O gráfico da Figura 4.1 mostra os resultados obtidos pelos padrões textuais, os classificadores de árvore de decisão e SVM para cada uma das 7 relações semânticas, permitindo uma comparação visual entre o desempenho de cada um.

Um dos problemas que dificultam a abordagem de padrões textuais é a ambiguidade inerente à língua natural. Como já discutido, certas relações como a part-of e location-of compartilham alguns padrões como “T1 de T2”, tornando a distinção entre essas relações difícil utilizando apenas padrões em nível superficial e morfológico. A abordagem de AM consegue lidar melhor com esse tipo de ambiguidade por envolver *features* de diferentes níveis linguísticos.

Outro problema da abordagem de padrões é o alto custo para a definição de novos padrões

Figura 4.1: Desempenho em medida-F obtido pelos padrões textuais (experimento 2), árvores de decisão (experimento 3) e SVMs (experimento 4) para cada relação semântica



textuais, já que estes dependem da análise do *corpus* e do seu estilo e gênero. Por exemplo, o padrão "T1_N T2_ADJ", indicando a relação property-of, é comum em textos jornalísticos, mas pode não ser em outros gêneros de escrita.

Essas dificuldades são atenuadas na abordagem de AM, embora nesta também seja necessária uma etapa custosa de etiquetação de *corpora*. De fato, por terem sido utilizados classificadores supervisionados, a anotação de mais dados de treinamento pode trazer melhores resultados, especialmente com relação às relações com menos exemplos de treinamento (made-of, used-for, effect-of).

A relação used-for foi a única que obteve melhores resultados com a aplicação de padrões textuais do que com métodos de AM, apresentando uma medida-F de 32,6%. Uma possível explicação para esse resultado pode ser que as *features* utilizadas pelos classificadores de AM não conseguem capturar informações relevantes para a extração dessa relação, que seria então mais bem expressada pelos padrões textuais. O baixo número de exemplos de treinamento anotados para essa relação também agrava o baixo desempenho dos classificadores.

Um fator que pode influenciar nos resultados de ambas as abordagens é o *parser* que foi

utilizado para processar os *corpora*, no caso o PALAVRAS (BICK, 2000). Em algumas situações o *parser* marca certos *tokens* com etiquetas incorretas, confundindo adjetivos com substantivos, entre outros problemas. Experimentos com outros *parsers* podem elucidar a influência do processador morfossintático nos resultados encontrados. Um exemplo de erro de etiquetação é o nome da cidade "Belo Horizonte" ser anotado como "Belo_ADJ Horizonte_N", quando ambos deveriam ser substantivos. Essa marcação errônea levou à extração de várias instâncias property-of(Horizonte, Belo) incorretas utilizando a estratégia de padrões textuais.

Concluindo, os experimentos mostraram que a abordagem de AM (particularmente usando classificadores SVM) obteve melhores resultados que a baseada em padrões textuais para todas as relações exceto a used-for, embora seja mais complexa e envolva mais recursos. No entanto, a etapa de preparação e anotação do *corpus* pode ser bastante custosa; se for necessário obter resultados rápidos e não houver textos etiquetados disponíveis, a abordagem de padrões textuais pode ser preferível. Contudo, se há tempo e pessoal disponível para anotar *corpora* de treinamento e treinar classificadores de AM, a adoção dessa estratégia pode trazer melhores resultados.

Nesse sentido, vale mencionar que a anotação de novas instâncias de relações pode ser realizada a partir de uma marcação inicial feita pelos métodos de AM treinados neste trabalho. Na verdade, esta estratégia de anotação em duas etapas – primeiro aplicam-se os métodos treinados neste trabalho e, em seguida, as instâncias de relações anotadas automaticamente são corrigidas e novas instâncias são inseridas usando a ferramenta ARS – já foi adotada na anotação do *corpus* FAPESP usado no experimento 6.

A Tabela 4.16 mostra alguns exemplos de instâncias de relações identificadas corretamente tanto pela estratégia de padrões textuais quanto de AM.

4.7.4 Comparação com trabalhos na literatura

Uma comparação dos valores obtidos neste trabalho com alguns dos principais resultados relatados na literatura é apresentada a seguir.

Considerando a relação is-a, pode-se comparar os trabalhos de Hearst (1992) (para o inglês) e Freitas e Quental (2007) (para o português), que buscam extrair essa relação utilizando padrões textuais. Hearst obteve uma precisão de 63% e Freitas e Quental obtiveram 73,4%, enquanto o experimento 1 (Tabela 4.1) resultou em uma precisão média de 65,7%. Quando os 7 padrões de hiponímia foram aplicados sobre o *corpus* CETENFolha (experimento 2), a precisão manteve-se quase a mesma, atingindo 61,1%. A aplicação dos classificadores de árvore de decisão (expeirmento 3) e SVM (experimento 4) resultou nas precisões de 76,9% e 78,2%, respectivamente, para

Tabela 4.16: Exemplos de instâncias de relações identificadas corretamente. os contextos em que ocorreram e os métodos que as encontraram

Relação	**Contexto original**	**Método**
is-a(Itália, país)	...ir para Itália ou qualquer outro país para...	Padrão is-a 5
property-of(mito, velho)	...enterra em=parte o velho mito explicitado por...	Padrão property-of 1
part-of(painéis, campanha)	...que os painéis fizessem parte de a campanha de o...	Padrão part-of 2
made-of(medalhas, ouro)	...Vegard=Ulvang , que ganhou três medalhas de ouro em o...	Padrão made-of 1
effect-of(úlcera, morreu)	...Jandira morreu devido=a a úlcera perfurada .	Padrão effect-of 1
location-of(carro de FHC, Colégio Alberto=Levy)	O carro de FHC chega a o Colégio Alberto=Levy , em...	Padrão location-of 3
used-for(recurso, alterar a foto)	...sobre o recurso usado para alterar a foto	Padrão used-for 1
is-a(Jovem=Pesquisador, programa)	...participou de o programa Jovem=Pesquisador com...	Árvore de decisão
property-of(explicação, esquemática)	Essa é uma explicação esquemática	Árvore de decisão
part-of(USP, Equipe)	Equipe de a USP detalha os mecanismos...	Árvore de decisão
made-of(shorts, algodão com lycra)	...e shorts em algodão e algodão com lycra...	Árvore de decisão
location-of(Centre=Georges=Pompidou, Paris)	...situada em o Centre=Georges=Pompidou , em Paris...	Árvore de decisão
effect-of(começou a chover e a ventar forte, desistir)	...a volta começou a chover e a ventar forte , perto=de Ribeirão=Preto , e tivemos de desistir...	Árvore de decisão
used-for(terapia celular, tratar)	...usam terapia celular para tratar experimentalmente...	Árvore de decisão
is-a(P10, sigla)	...conhecido por a sigla P10 .	SVM
property-of(ação, predatória)	...a ação predatória de o homem é...	SVM
part-of(desenhos, palmeiras)	...de as palmeiras de os desenhos de...	SVM
made-of(jatos, gases)	...evolução de jatos de gases com...	SVM
effect-of(faz surgir um caroço central, dar origem a a estrela)	...faz surgir um caroço central , que vai dar origem a a estrela...	SVM
used-for(natação, conter)	...que a natação ajuda a conter a...	SVM
location-of(sangue, corpo)	...o sangue que corre por o corpo contém...	SVM

a relação is-a. Percebe-se que os valores obtidos pelos experimentos de 1 a 4 se situam na média de 70%, com destaque para a precisão de 78,2% obtida pelo classificador SVM, superando tanto os valores de Hearst quanto de Freitas e Quental. A diferença desses resultados para os valores obtidos pelos trabalhos da literatura (63% e 73,4%) pode ser devido aos diferentes métodos e *corpora* utilizados em cada trabalho. A cobertura não pôde ser comparada já que não foi calculada em (HEARST, 1992) e (FREITAS; QUENTAL, 2007).

Uma comparação direta com os demais trabalhos que enfocam a língua portuguesa – PAPEL (OLIVEIRA et al., 2010) e HAREM (MOTA; SANTOS, 2008) – não é possível pois a

definição das relações semânticas e sua avaliação nesses trabalhos é distinta da realizada nesta pesquisa.

Em relação aos melhores resultados obtidos para as outras relações, pode-se citar (GIRJU et al., 2003) e (GIRJU, 2003), que extraem a relação part-of e effect-of, respectivamente, ambos utilizando árvores de decisão C4.5. O primeiro obteve 83% de precisão e 72% de cobertura e o segundo 74% de precisão e 67% de cobertura, resultando em medidas-F de 77,1% e 70,3%. Esses valores de medida-F são consideravelmente maiores que os obtidos no experimento 3 (47,7% para part-of e 16,0% para effect-of) e no experimento 4 (47,9% para part-of e 24,6% para effect-of). Um dos possíveis motivos para essa discrepância é a quantidade de exemplos de treinamento utilizados nesses trabalhos: 81580 instâncias para a relação part-of e 6523 instâncias para a effect-of, contra 1976 e 163 instâncias de cada relação usadas nos experimentos 3 e 4. Como já mostrado na Tabela 4.15, a quantidade de exemplos de treinamento de cada relação influi diretamente no desempenho dos algoritmos de AM.

Zelenko et al. (2003) enfocam a relação location-of, e utiliza um classificador SVM com um *parse tree kernel* e 1915 exemplos de treinamento. O valor de medida-F obtido foi de 83,3%, superando facilmente os resultados obtidos nos experimentos 3 (31,8%) e 4 (36,2%), apesar de o número de instâncias de treinamento ser próximo (1915 contra 1683 neste trabalho). É possível que o uso do *kernel* específico tenha bastante influência no desempenho do algoritmo. Já em (SNOW et al., 2005) foram utilizados classificadores *naive* Bayes e regressão logística para extrair a relação is-a, com cerca de 750 mil instâncias de treinamento. O melhor resultado obtido foi uma medida-F de 34,8% usando regressão logística, valor menor que o apresentado pelos experimentos 3 (65,0%) e 4 (71,0%).

Finalmente, Girju et al. (2010) procuram por 7 relações semânticas, entre elas 2 que também são enfocadas neste trabalho: effect-of e part-of. Foram utilizados classificadores SVM com 1460 instâncias de treinamento para a relação effect-of e 1143 para a part-of. Os valores de medida-F obtidos para cada relação foram 82,0% (effect-of) e 68% (part-of), ambos maiores que os obtidos nos experimentos 3 e 4 deste trabalho. Deve-se considerar que (GIRJU et al., 2010) é um dos trabalhos levantados que utilizam mais recursos e ferramentas linguísticas em seu método.

Concluindo, embora os resultados apresentados neste trabalho não superem todos os relatados na literatura, é importante observar que cada estudo investigado utiliza diferentes *corpora* e métodos de extração de relações semânticas, além de enfocar tipos distintos de relações, fatores que devem ser levados em conta ao fazer comparações. Vale ressaltar também que este é o primeiro estudo que investiga técnicas de AM para extrair relações semânticas com a língua portuguesa em foco e que, comparado com alguns dos trabalhos levantados na literatura para o inglês (SNOW et

al., 2005; GIRJU et al., 2010), utiliza poucas ferramentas e recursos linguísticos. Além disso, as relações property-of, made-of e used-for ainda não haviam sido alvo de estudos voltados para a extração automática de relações semânticas. A Tabela 4.17 resume a comparação com os trabalhos levantados na literatura em termos de medida-F (com exceção da primeira linha, apresentada com valores de precisão).

Tabela 4.17: Resumo da comparação com trabalhos da literatura em termos de medida-F, exceto quando especificado de outra maneira

Relação	**Melhor resultado obtido neste trabalho**	**Melhor resultado relatado na literatura**
is-a	78,2% (precisão)	73,4% (precisão) (FREITAS; QUENTAL, 2007)
property-of	85,4%	–
part-of	47,9%	77,1% (GIRJU, 2003)
location-of	36,2%	83,3% (ZELENKO et al., 2003)
effect-of	24,6%	82,0% (GIRJU et al., 2010)
made-of	34,3%	–
used-for	26,2%	–

5 Conclusões e trabalhos futuros

A extração automática de relações semânticas ainda é uma tarefa para a qual os sistemas existentes não têm alta performance (GIRJU et al., 2010), sobretudo devido a sua complexidade. Considerando-se a língua portuguesa, em especial a variante brasileira, a situação é ainda mais precária, com poucos trabalhos e pesquisas feitas sobre esse tema. Agravando esse cenário, a língua portuguesa carece de recursos linguísticos e bases de conhecimento de grande cobertura, como a WordNet para a língua inglesa, e ferramentas como analisadores sintáticos e etiquetadores de papéis semânticos de alta precisão. Pesquisas nessa área colaborariam no aumento da quantidade e qualidade de recursos linguísticos para o idioma português, além de beneficiar o cenário de PLN do português como um todo.

Sendo assim, este trabalho de mestrado visou o estudo e a comparação das duas principais abordagens para extração automática de relações semânticas, uma área subexplorada considerando-se a língua portuguesa. Os resultados obtidos mostram que a abordagem baseada em aprendizado de máquina traz melhores resultados que a baseada em padrões textuais, indicando que abordagens baseadas em AM são uma direção promissora para estudos sobre o tema com a língua portugesa em foco.

O método, as *features* e os resultados apresentados podem ser utilizados e trazer avanços em aplicações como extração e recuperação de informação, tradução automática e sistemas de respostas e perguntas, e como apoio para a construção e melhoramento de recursos léxicos como ontologias, terminologias e dicionários. Sobretudo, estudos sobre extração de relações semânticas com o idioma português ainda são escassos, reforçando a importância deste trabalho para o avanço da pesquisa nessa área tão ampla e complexa.

A inexistência de certos recursos e ferramentas linguísticas de alta qualidade para a língua portugesa limita os resultados alcançados nas pesquisas atuais. Entre eles, uma Wordnet de grande cobertura tal qual a de Princeton (FELLBAUM, 1998) poderia trazer melhoras significativas em trabalhos sobre semântica com o português, visto que alguns dos trabalhos de estado da arte para o inglês, como (GIRJU et al., 2010), utilizam a WordNet como um de seus componentes. Ferra-

mentas como um identificador de entidades nomeadas e um *parser* aberto e de alto desempenho também trariam grandes benefícios para o processamento semântico do português.

A fim de beneficiar e contribuir com a comunidade e as pesquisas em PLN, os recursos (*corpus* FAPESP e o *corpus* CETENFolha anotados manualmente com relações semânticas), bem como a ferramenta ARS incrementada com a funcionalidade de identificação automática de relações semânticas por meio das árvores de decisão e SVMs treinadas conforme descrito neste documento serão disponibilizados de forma aberta (*open-source*) no Portal de Tradução Automática[1] (PorTAl) do Laboratório de Linguística Computacional (LaLiC) da UFSCar. Para ter acesso a estes recursos e ferramenta basta se cadastrar no PorTAl e acessar a área de download.

É pertinente ressaltar que o projeto foi desenvolvido no LaLiC, com o apoio da FAPESP (processo #2011/04482-4) e a colaboração do Núcleo Interinstitucional de Linguística Computacional[2] (NILC), grupo do qual a orientadora e o aluno fazem parte. O NILC é um grupo de pesquisa interdisciplinar de linguística e computação composto por professores e alunos de várias universidades brasileiras, dentre as quais participam a UFSCar (campus de São Carlos) com integrantes de seus departamentos de Computação (LaLiC) e de Letras, e a USP com alunos e pesquisadores do ICMC (Instituto de Ciências Matemáticas e de Computação). O NILC conta com o suporte de cientistas da computação e linguistas especializados, além de diversos recursos linguístico-computacionais de grande utilidade para o projeto aqui apresentado.

5.1 Trabalhos futuros

Entre as continuações e futuras melhorias possíveis para este trabalho, pode-se citar a definição de novas *features* que auxiliem os métodos de AM na classificação de relações. Outra possibilidade é a utilização de diferentes etiquetadores e identificadores de sintagmas nominais para o processamento do *corpus* de treinamento. A anotação de mais exemplos de treinamento também pode trazer melhores resultados para os classificadores. É interessante verificar se a anotação de mais exemplos da relação used-for irá fazer com que os métodos de AM superem os resultados da aplicação de padrões textuais para essa relação.

Os dois *corpora* utilizados neste estudo, o FAPESP e o CETENFolha, são dos gêneros de divulgação científica e jornalístico, respectivamente. A utilização de *corpora* de outros gêneros e domínios poderia elucidar o impacto do *corpus* de treinamento nesse tipo de estudo, avaliando se os métodos são genéricos o suficiente para poderem ser aplicados a quaisquer tipos de textos.

[1] `http://www.lalic.dc.ufscar.br/portal/`
[2] `http://www.nilc.icmc.usp.br`

Certas relações têm restrições quanto às classes morfológicas de seus termos participantes: por exemplo, a relação property-of ocorre geralmente entre um substantivo e um adjetivo. Para melhorar a qualidade das relações extraídas poderiam ser implementados filtros que descartem instâncias de relações que ocorram entre termos de classes morfossintáticas incompatíveis. Outra possibilidade é realizar *cotraining* (CARLSON et al., 2010), combinando as visões de diferentes classificadores buscando obter melhores resultados.

Outra sugestão para trabalhos futuros é a adaptação da proposta de estudos como (ZELENKO et al., 2003) e (ZHANG et al., 2006) que utilizam *kernels* que comparam árvores sintáticas ao invés de vetores de *features*, como é feito neste trabalho. A análise do impacto de cada *feature* também pode ser feita, averiguando quais têm maior peso na identificação de uma determinada relação semântica. Novas *features* também podem ser criadas.

Referências Bibliográficas

AGICHTEIN, E.; GRAVANO, L. Snowball: extracting relations from large plain-text collections. In: *Proceedings of the fifth ACM conference on Digital libraries*. New York, NY, USA: ACM, 2000. (DL '00), p. 85–94.

AHMED, S. T.; NAIR, R.; PATEL, C.; KANWAR, S. P.; HAKENBERG, J.; DAVULCU, H. Semantic classification and dependency parsing enabled automated bio-molecular event extraction from text. In: *Proceedings of the First ACM International Conference on Bioinformatics and Computational Biology*. New York, NY, USA: ACM, 2010. (BCB '10), p. 370–373.

BERLAND, M.; CHARNIAK, E. Finding parts in very large corpora. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*. College park, MD, 1999. p. 57–64.

BICK, E. *The Parsing System"Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Denmark: Aarhus University Press, 2000.

BRIN, S. Extracting patterns and relations from the world wide web. In: *Selected papers from the International Workshop on The World Wide Web and Databases*. London, UK: Springer-Verlag, 1999. p. 172–183.

BRUCKSCHEN, M.; SOUZA, J. G. C. de; VIEIRA, R.; RIGO, S. Sistema serelep para o reconhecimento de relações entre entidades mencionadas. In: ____. Portugal: Linguateca, 2008. cap. 14, p. 247–260.

BURCHARDT, A.; ERK, K.; FRANK, A.; KOWALSKI, A.; PADO, S.; PINKAL, M. Salto – a versatile multi-level annotation tool. In: *Proceedings of LREC 2006*. Genoa, Italy: [s.n.], 2006.

CARABALLO, S. A. Automatic construction of a hypernym-labeled noun hierarchy from text. In: *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*. Stroudsburg, PA, USA: Association for Computational Linguistics, 1999. (ACL '99), p. 120–126.

CARDOSO, N. Rembrandt - reconhecimento de entidades mencionadas baseado em relações e análise detalhada do texto. In: ____. Portugal: Linguateca, 2008. cap. 11, p. 195–211.

CARLETTA, J. Squibs and discussions assessing agreement on classification tasks: The kappa statistic. *Computational linguistics*, v. 22, n. 2, p. 249–254, 1996.

CARLSON, A.; BETTERIDGE, J.; KISIEL, B.; SETTLES, B.; JR., E. R. H.; MITCHELL, T. M. Toward an architecture for never-ending language learning. In: *Proceedings of the Twenty-Fourth Conference on Artificial Intelligence (AAAI 2010)*. [S.l.]: AAAI Press, 2010. p. 1306–1313.

CASELI, H. M. *Indução de léxicos bilíngües e regras para a tradução automática*. Tese (Doutorado) — Instituto de Ciências Matemáticas e de Computação - ICMC-USP, Junho 2007.

CHAVES, M. S. Geo-ontologias e padrões para reconhecimento de locais e de suas relações em textos: o sei-geo no segundo harem. In: ____. Portugal: Linguateca, 2008. cap. 13, p. 231–245.

COWIE, J.; LEHNERT, W. Information extraction. *Communications of the ACM*, ACM, New York, NY, USA, v. 39, p. 80–91, January 1996.

CUNNINGHAM, H.; MAYNARD, D.; BONTCHEVA, K.; TABLAN, V.; ASWANI, N.; ROBERTS, I.; GORRELL, G.; FUNK, A.; ROBERTS, A.; DAMLJANOVIC, D.; HEITZ, T.; GREENWOOD, M. A.; SAGGION, H.; PETRAK, J.; LI, Y.; PETERS, W. *Text Processing with GATE (Version 6)*. Gateway Press CA, 2011. ISBN 978-0956599315. Disponível em: <http://tinyurl.com/gatebook>.

DODDINGTON, G.; MITCHELL, A.; PRZYBOCKI, M.; RAMSHAW, L.; STRASSEL, S.; WEISCHEDEL, R. The automatic content extraction (ace) program–tasks, data, and evaluation. In: CITESEER. *Proceedings of LREC*. [S.l.], 2004. v. 4, p. 837–840.

FELLBAUM, C. *WordNet: An electronic lexical database*. Cambridge, MA: The MIT press, 1998.

FREITAS, M. C. de; QUENTAL, V. Subsídios para a elaboração automática de taxonomias. In: *Anais do XXVII Congresso da SBC*. Rio de Janeiro, Rio de Janeiro: [s.n.], 2007. (V Workshop em Tecnologia da Informação e da Linguagem Humana TIL), p. 1585–1594.

FREUND, Y.; SCHAPIRE, R. E. Large margin classification using the perceptron algorithm. *Mach. Learn.*, Kluwer Academic Publishers, Hingham, MA, USA, v. 37, p. 277–296, December 1999.

GASPERIN, C. V.; LIMA, V. L. S. de. Semantic similarity from syntactic relations. In: *Workshop Multilingual Information Acces & Natural Language Processing (VIII Iberoamerican Conference on Artificial Intelligence - IBERAMIA 2002)*. Sevilla: [s.n.], 2002. v. 1, p. 55–62.

GASPERIN, C. V.; LIMA, V. L. S. de. Experiments on extracting semantic relations from syntactic relations. In: *Proceedings of the 4th international conference on Computational linguistics and intelligent text processing*. Berlin, Heidelberg: Springer-Verlag, 2003. (CICLing'03), p. 314–324.

GIRJU, R. Automatic detection of causal relations for question answering. In: *Proceedings of the ACL 2003 workshop on Multilingual summarization and question answering - Volume 12*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2003. (MultiSumQA '03), p. 76–83.

GIRJU, R.; BADULESCU, A.; MOLDOVAN, D. Learning semantic constraints for the automatic discovery of part-whole relations. In: *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2003. (NAACL '03), p. 1–8.

GIRJU, R.; BADULESCU, A.; MOLDOVAN, D. Automatic discovery of part-whole relations. *Computational Linguistics*, MIT Press, Cambridge, MA, USA, v. 32, p. 83–135, March 2006.

GIRJU, R.; BEAMER, B.; ROZOVSKAYA, A.; FISTER, A.; BHAT, S. A knowledge-rich approach to identifying semantic relations between nominals. *Information Processing and Management*, v. 46, n. 5, p. 589–610, 2010.

GIRJU, R.; MOLDOVAN, D. Text mining for causal relations. In: AAAI PRESS. *Proceedings of the International Florida Artificial Intelligence Research Society (FLAIRS 2002)*. Pensacola, Florida, 2002. p. 360–364.

GIRJU, R.; NAKOV, P.; NASTASE, V.; SZPAKOWICZ, S.; TURNEY, P.; YURET, D. Semeval-2007 task 04: Classification of semantic relations between nominals. In: *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*. Prague, Czech Republic: Association for Computational Linguistics, 2007. p. 13–18.

GREFENSTETTE, G. Evaluation techniques for automatic semantic extraction: comparing syntactic and window based approaches. In: ____. Cambridge, MA, USA: MIT Press, 1993. p. 143–153.

GRISHMAN, R.; SUNDHEIM, B. Message understanding conference-6: A brief history. In: *Proceedings of COLING*. [S.l.: s.n.], 1996. v. 96, p. 466–471.

HALL, M.; FRANK, E.; HOLMES, G.; PFAHRINGER, B.; REUTEMANN, P.; WITTEN, I. H. The weka data mining software: An update. *SIGKDD Explorations*, v. 11, n. 1, p. 10–18, June 2009.

HEARST, M. A. Automatic acquisition of hyponyms from large text corpora. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the 14th International Conference on Computational linguistics - Volume 2*. Nantes, France, 1992. p. 539–545.

HEARST, M. A. Automated discovery of wordnet relations. In: ____. [S.l.]: The MIT press, 1998. WordNet: An electronic lexical database, cap. 5, p. 131–151.

HENDRICKX, I.; KIM, S. N.; KOZAREVA, Z.; NAKOV, P.; SéAGHDHA, D. O.; PADó, S.; PENNACCHIOTTI, M.; ROMANO, L.; SZPAKOWICZ, S. Semeval-2010 task 8: multi-way classification of semantic relations between pairs of nominals. In: *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2009. (DEW '09), p. 94–99.

IDE, N.; BONHOMME, P.; ROMARY, L. Xces: An xml-based encoding standard for linguistic corpora. In: *Proceedings of LREC 2000*. Atenas, Grécia: [s.n.], 2000. p. 825–830.

JOACHIMS, T. *Making Large-Scale SVM Learning Practical*. [S.l.], 1998.

KUCERA, H.; FRANCIS, W. N. *Computational analysis of present-day American English*. Providence, RI: Brown University Press, 1967.

LIN, D. Dependency-based evaluation of minipar. In: SPRINGER. *Workshop on the Evaluation of Parsing systems*. [S.l.], 1998. p. 317–330.

LIN, D.; ZHAO, S.; QIN, L.; ZHOU, M. Identifying synonyms among distributionally similar words. In: *Proceedings of the 18th international joint conference on Artificial intelligence*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2003. p. 1492–1493.

MENGEL, A.; LEZIUS, W. An xml-based encoding format for syntactically annotated corpora. In: *Proceedings of LREC 2000*. Atenas, Grécia: [s.n.], 2000. p. 121–126.

MILLER, G. A.; LEACOCK, C.; TENGI, R.; BUNKER, R. T. A semantic concordance. In: *Proceedings of the workshop on Human Language Technology*. Stroudsburg, PA, USA: Association for Computational Linguistics, 1993. (HLT '93), p. 303–308.

MINSKY, M. *The Society of Mind*. [S.l.]: Simon and Schuster, 1986.

MINTZ, M.; BILLS, S.; SNOW, R.; JURAFSKY, D. Distant supervision for relation extraction without labeled data. In: *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - Volume 2*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2009. (ACL '09), p. 1003–1011.

MOTA, C.; SANTOS, D. (Ed.). *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*. Portugal: Linguateca, 2008.

OAKES, M. P. Using hearst's rules for the automatic acquisition of hyponyms for mining a pharmaceutical corpus. In: *RANLP Text Mining Workshop*. [S.l.: s.n.], 2005. p. 63–67.

OLIVEIRA, H. G.; SANTOS, D.; GOMES, P. Extracção de relações semânticas entre palavras a partir de um dicionário: o papel e sua avaliação. *Linguamática*, v. 2, n. 1, p. 77–94, 2010.

QUINLAN, J. R. *C4.5: programs for machine learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1993. ISBN 1-55860-238-0.

ROBERTS, A.; GAIZAUSKAS, R.; HEPPLE, M.; GUO, Y. Mining clinical relationships from patient narratives. *BMC Bioinformatics*, v. 9, n. Suppl 11, p. S3, 2008.

ROCHA, P. A.; SANTOS, D. Cetempúblico: Um corpus de grandes dimensões de linguagem jornalística portuguesa. In: NUNES, M. das G. V. (Ed.). *V Encontro para o processamento computacional da língua portuguesa escrita e falada (PROPOR 2000)*. São Paulo: ICMC/USP, 2000. p. 131–140. Disponível em: <http://www.linguateca.pt/documentos/RochaSantosPROPOR2000.pdf>.

SANTOS, C. N. dos; OLIVEIRA, C. Aplicação de aprendizado baseado em transformações na identificação de sintagmas nominais. In: *Anais do XXV Congresso da Sociedade Brasileira de Computação*. Brasil: [s.n.], 2005. (III TIL), p. 2138–2147.

SHANG, Y.; LI, Y.; LIN, H.; YANG, Z. Enhancing biomedical text summarization using semantic relation extraction. *PLoS ONE*, Public Library of Science, v. 6, n. 8, p. e23862, 08 2011.

SHANNON, C. E. A mathematical theory of communication. *Bell system technical journal*, v. 27, p. 379–423, 1948.

SINGH, P.; LIN, T.; MUELLER, E. T.; LIM, G.; PERKINS, T.; ZHU, W. L. Open mind common sense: Knowledge acquisition from the general public. In: *On the Move to Meaningful Internet Systems, 2002 - DOA/CoopIS/ODBASE 2002 Confederated International Conferences DOA, CoopIS and ODBASE 2002*. [S.l.]: Springer-Verlag, 2002. p. 1223–1237.

SNOW, R.; JURAFSKY, D.; NG, A. Y. Learning syntactic patterns for automatic hypernym discovery. In: *Advances in Neural Information Processing Systems 17*. [S.l.]: MIT Press, 2005. p. 1297–1304.

SUGIYAMA, B. A.; ANACLETO, J. C.; CASELI, H. de M. Assisting users in a cross-cultural communication by providing culturally contextualized translations. In: *Proceedings of the 29th ACM international conference on Design of communication*. Pisa, Italy: ACM, 2011. (SIGDOC'11), p. 189–194.

TABA, L. S.; CASELI, H. de M. Automatic hyponymy identification from brazilian portuguese texts. In: *Proceedings of the International Conference on Computational Processing of the Portuguese Language (PROPOR 2012)*. [S.l.: s.n.], 2012. (in press).

TABA, L. S.; CASELI, H. de M. Uma ferramenta para anotação de relações semânticas entre termos. In: *Anais do XI Encontro de Linguística de Corpus - ELC 2012*. Instituto de Ciências Matemáticas e de Computação da USP, São Carlos/SP: [s.n.], 2012. Disponível em: <http://143.107.232.109/elc-ebralc2012/anais/completos/104005.pdf>.

TABA, L. S.; CASELI, H. M. *ARS – Ferramenta de anotação de relações semânticas em textos escritos em português do Brasil. Série de relatórios do NILC (NILC-TR-13-03)*. São Carlos, 2013.

VAPNIK, V. *The nature of statistical learning theory*. [S.l.]: Springer Verlag, 1995.

YAP, W.; BALDWIN, T. Experiments on pattern-based relation learning. In: *Proceeding of the 18th ACM conference on Information and knowledge management*. New York, NY, USA: ACM, 2009. (CIKM '09), p. 1657–1660.

ZELENKO, D.; AONE, C.; RICHARDELLA, A. Kernel methods for relation extraction. *The Journal of Machine Learning Research*, v. 3, p. 1083–1106, March 2003.

ZHANG, M.; ZHANG, J.; SU, J.; ZHOU, G. A composite kernel to extract relations between entities with both flat and structured features. In: *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2006. (ACL-44), p. 825–832.

APÊNDICE A

Estrutura JSON para codificação de sentenças

Este apêndice apresenta a estrutura JSON para codificação de sentenças utilizada pela ferramenta de auxílio à anotação de relações semânticas, a ARS.

O formato JSON se baseia em duas estruturas básicas, objetos e vetores. Objetos (definidos entre "{" e "}") são conjuntos de pares chave:valor, similares a vetores associativos, onde a chave é uma *string*; e vetores (definidos entre "[" e "]") são sequências ordenadas de valores. Valores podem ser *strings*, números, objetos, vetores, `true`/`false` (valores booleanos) ou `null` (valor inexistente). Baseado nessas estruturas, uma sentença é um objeto que contém outros objetos (*tokens*, termos, relações) e alguns campos (id, texto, entre outros). As estruturas definidas para a codificação de sentenças, *tokens*, termos e relações são mostradas nas Tabelas A.1 a A.4 a seguir.

Tabela A.1: Campos do objeto JSON que representam uma sentença

Campo	**Tipo**	**Descrição**
id	Numérico	Número que identifica a sentença
texto	String	Texto da sentença
comentarios	String	Comentários adicionados pelos anotadores
anotada	Booleano	Se a sentença já foi anotada por algum anotador
ignorada	Booleano	Se a sentença foi marcada como ignorada (porque contém algum erro) pelos anotadores
tokens	Vetor	Vetor de objetos que representam cada *token* da sentença
termos	Vetor	Vetor de objetos que representam os termos marcados na sentença
relacoes	Vetor	Vetor de objetos que representam as relações semânticas marcadas na sentença
anotadores	Vetor	Vetor de strings que armazenam o nome dos anotadores que trabalharam na sentença

Tabela A.2: Campos do objeto JSON que representam um *token*

Campo	Tipo	Descrição
t	String	Forma superficial do *token*
l	String	Lema (forma base) do *token*
pos	String	Etiquetas *part-of-speech* do *token* (cf. anotado pelo PALAVRAS (BICK, 2000)
sin	String	Papel sintático do *token* na árvore de dependências da sentença (cf. anotado pelo PALAVRAS (BICK, 2000))

Tabela A.3: Campos do objeto JSON que representam um termo

Campo	Tipo	Descrição
de	Numérico	Índice do *token* (no vetor de *tokens* da sentença) onde o termo se inicia
para	Numérico	Índice do *token* (no vetor de *tokens* da sentença) onde o termo termina

A seguir é mostrado um exemplo completo de uma sentença codificada nesse formato, retirada dos pacotes anotados do *corpus* CETENFolha.

```
{
    "id": 271463,
    "texto": "Em a virada para o real , os valores foram convertidos para a nova moeda ,
              mas sem considerar a desvalorização de o dólar .",
    "comentarios": "Termo:desvalorização=de=o=dólar",
    "anotada": true,
    "ignorada": false,
    "tokens": [
        { "t": "Em", "l": "em", "pos": "<sam-> PRP", "sin": "@ADVL>" },
        { "t": "a", "l": "o", "pos": "<-sam> <artd> DET F S", "sin": "@>N" },
        { "t": "virada", "l": "virada", "pos": "N F S", "sin": "@P<" },
        { "t": "para", "l": "para", "pos": "PRP", "sin": "@N<" },
```

Tabela A.4: Campos do objeto JSON que representam uma relação

Campo	Tipo	Descrição
r	String	Qual relação semântica está marcada (is-a, property-of, etc.)
t1	Numérico	Índice do termo que é o primeiro participante da relação
t2	Numérico	Índice do termo que é o segundo participante da relação

```
        { "t": "o", "l": "o", "pos": "<artd> DET M S", "sin": "@>N" },
        { "t": "real", "l": "real", "pos": "N M S", "sin": "@P<" },
        { "t": ",", "l": ",", "pos": null, "sin": null },
        { "t": "os", "l": "o", "pos": "<artd> DET M P", "sin": "@>N" },
        { "t": "valores", "l": "valor", "pos": "N M P", "sin": "@SUBJ>" },
        { "t": "foram", "l": "ser", "pos": "<fmc> V PS/MQP 3P IND VFIN", "sin": "@FAUX" },
        { "t": "convertidos", "l": "converter", "pos": "V PCP M P", "sin": "@IMV @#ICL-AUX<" },
        { "t": "para", "l": "para", "pos": "PRP", "sin": "@<ADVL" },
        { "t": "a", "l": "o", "pos": "<artd> DET F S", "sin": "@>N" },
        { "t": "nova", "l": "novo", "pos": "ADJ F S", "sin": "@>N" },
        { "t": "moeda", "l": "moeda", "pos": "N F S", "sin": "@P<" },
        { "t": ",", "l": ",", "pos": null, "sin": null },
        { "t": "mas", "l": "mas", "pos": "<co-advl> KC", "sin": "@CO" },
        { "t": "sem", "l": "sem", "pos": "PRP", "sin": "@<ADVL" },
        { "t": "considerar", "l": "considerar", "pos": "V INF", "sin": "@IMV @#ICL-P<" },
        { "t": "a", "l": "o", "pos": "<artd> DET F S", "sin": "@>N" },
        { "t": "desvalorização", "l": "desvalorização", "pos": "N F S", "sin": "@<ACC" },
        { "t": "de", "l": "de", "pos": "<sam-> PRP", "sin": "@N<" },
        { "t": "o", "l": "o", "pos": "<-sam> <artd> DET M S", "sin": "@>N" },
        { "t": "dólar", "l": "dólar", "pos": "N M S", "sin": "@P<" },
        { "t": ".", "l": ".", "pos": null, "sin": null }
    ],
    "termos": [
        { "de": 13, "ate": 13 },
        { "de": 14, "ate": 14 }
    ],
    "relacoes": [
        { "r": "property-of", "t1": 1, "t2": 0 }
    ],
    "anotadores": [
        "Leo"
    ]
}
```


Essa sentença tem dois termos marcados (“nova” e “moeda”) e a relação property-of entre eles.