

UNIVERSIDADE FEDERAL DE SÃO CARLOS

CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

**APRENDIZADO SEMISSUPERVISIONADO ATRAVÉS DE
TÉCNICAS DE ACOPLAMENTO**

MAÍSA CRISTINA DUARTE

ORIENTADOR: PROF. DR. ESTEVAM RAFAEL HRUSCHKA JUNIOR

São Carlos - SP
Fevereiro/2011

UNIVERSIDADE FEDERAL DE SÃO CARLOS

CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

**APRENDIZADO SEMISSUPERVISIONADO ATRAVÉS
DE TÉCNICAS DE ACOPLAMENTO**

MAÍSA CRISTINA DUARTE

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: Prof. Dr. Estevam Rafael Hruschka Junior.

São Carlos - SP
Fevereiro/2011

**Ficha catalográfica elaborada pelo DePT da
Biblioteca Comunitária da UFSCar**

D812as

Duarte, Maisa Cristina.

Aprendizado semissupervisionado através de técnicas de acoplamento / Maisa Cristina Duarte. -- São Carlos : UFSCar, 2011.

102 f.

Dissertação (Mestrado) -- Universidade Federal de São Carlos, 2011.

1. Aprendizado do computador. 2. Auto-supervisão. 3. Entidades nomeadas. I. Título.

CDD: 006.31 (20^a)

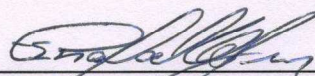
Universidade Federal de São Carlos
Centro de Ciências Exatas e de Tecnologia
Programa de Pós-Graduação em Ciência da Computação

“Aprendizado Semissupervisionado através de
Técnicas de Acoplamento”

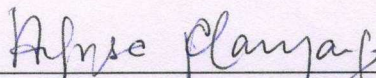
MAÍSA CRISTINA DUARTE

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

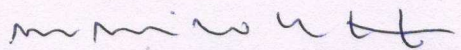
Membros da Banca:



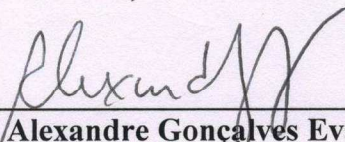
Prof. Dr. Estevam Rafael Hruschka Júnior
(Orientador - DC/UFSCar)



Profa. Dra. Heloisa de Arruda Camargo
(DC/UFSCar)



Profa. Dra. Maria do Carmo Nicoletti
(PPG-CC/UFSCar)



Prof. Dr. Alexandre Gonçalves Evsukoff
(COPPE/UFRJ)

São Carlos
Fevereiro/2011

Aos meus pais que sempre foram fundamentais para a realização de todos os meus sonhos, sem os quais não seria possível sequer pensar na realização de algum deles.

AGRADECIMENTO

A Deus pela oportunidade e por iluminar todo o andamento do trabalho.

Ao orientador deste trabalho, Estevam Hruschka Jr., o qual é um dos principais responsáveis pela motivação e entusiasmo para a realização deste, foi sensível a entraves que surgiram e em nenhum momento foi menos confiante, pelo contrário, sempre demonstrou apoio e segurança.

A Maria do Carmo Nicoletti que se manteve ao meu lado sempre e me apoiou desde o início nas minhas mais diferentes necessidades. Em todos os momentos se fez muito presente, cuidadosa e compreensiva. Sua dedicação e empenho me impulsionaram e impulsionam a sentir mais confiança e motivação.

A minha família que eu amo, principalmente aos meus pais, Antônio e Lourdes, que não mediram esforços e apoio. E também, de forma especial, meu irmão Alexandre que é meu porto seguro em todos os momentos que preciso.

A todos meus amigos que caminharam comigo me ajudando das mais diversas formas, seja com palavras, motivação, auxílio e principalmente presença, dentre eles, de forma especial Flávio Montoro, Eduardo Cirilo, Edimilson B. dos Santos, Wesley Willy, Juciara Nepomuceno, Michelle Zattoni, Rafael Durelli, Valéria Lauande e Kamila Rios.

Ao MaLL, aos colegas do PPGCC, ao Departamento de Computação, todos os professores e funcionários.

A todos que de alguma forma estiveram no meu caminho e acreditaram nos meus sonhos.

*É melhor tentar e falhar,
que preocupar-se e ver a vida passar;
é melhor tentar, ainda que em vão,
que sentar-se fazendo nada até o final.
Eu prefiro na chuva caminhar,
que em dias tristes em casa me esconder.
Prefiro ser feliz, embora louco,
que em conformidade viver ..."*
Martin Luther King

RESUMO

O Aprendizado de Máquina (AM) pode ser visto como uma área de pesquisa dentro da Inteligência Artificial (IA) que busca o desenvolvimento de programas de computador que possam evoluir à medida que vão sendo expostos a novas experiências. O principal objetivo de AM é a busca por métodos e técnicas que permitem a concepção de sistemas computacionais capazes de melhorar seu desempenho, de maneira autônoma, usando informações obtidas ao longo de seu uso; tal característica pode, de certa forma, ser considerada como um dos mecanismos fundamentais que regem os processos de aprendizado automático. O principal objetivo da pesquisa descrita neste documento foi investigar, propor e implementar métodos e algoritmos que permitissem a construção de um sistema computacional de aprendizado contínuo capaz de realizar a extração de conhecimento a partir da Web em português, por meio da criação de uma base de conhecimento atualizada constantemente à medida que novos conhecimentos vão sendo extraídos.

Palavras-chave: Aprendizado de Máquina, Auto-Supervisão, Acoplamento, Entidades Nomeadas

ABSTRACT

Machine Learning (ML) can be seen as research area within the Artificial Intelligence (AI) that aims to develop computer programs that can evolve with new experiences. The main ML purpose is the search for methods and techniques that enable the computer system improve its performance autonomously using information learned through its use. This feature can be considered the fundamental mechanisms of the processes of automatic learning. The main goal in this research project was to investigate, propose and implement methods and algorithms to allow the construction of a continuous learning system capable of extracting knowledge from the Web in Portuguese, throughout the creation of a knowledge base which can be constantly updated as new knowledge is extracted.

Keywords: Machine Learning, Self Supervised, Coupling, Named Entities

LISTA DE FIGURAS

Figura 3.1 Base de Conhecimento	44
Figura 3.2 Base de Conhecimento com Relações Semânticas	45
Figura 3.3 Extração de ENs e PTs	48
Figura 3.4 Extração de ENs e PTs – Exemplo	49
Figura 3.5 Aprendizado de ENs e PTs para categorias.....	52
Figura 3.6 Extração de Pares de ENs e PTs de Relações Semânticas - Exemplo	54
Figura 3.7 Extração de Pares de ENs e PTs de Relações Semânticas	55
Figura 3.8 Aprendizado de Pares ENs e PTs de Relações Semânticas	57
Figura 3.9 Aprendizado de Pares de ENs e PTs de Relações Semânticas com Tipagem	61
Figura 3.10 RTWP com todos acoplamentos	63
Figura 3.11 Visão Futura do NELL	64
Figura 3.12 Visão Atual do NELL	65
Figura 4.1 Resultados - Experimento Não Acoplado e Não Cumulativo 10....	71
Figura 4.2 Resultados - Experimento Não Acoplado e Não Cumulativo 3.....	72
Figura 4.3 Resultados - Experimento Acoplado e Não Cumulativo 10	73
Figura 4.4 Resultados - Experimento Acoplado e Não Cumulativo 3	74
Figura 4.5 Aprendizado Acoplado e Não Acoplado Cumulativo 10 para Cidade	75
Figura 4.6 Aprendizado Acoplado e Não Acoplado Cumulativo 10 para Ator.	75
Figura 4.7 Aprendizado Acoplado e Não Acoplado Cumulativo 10 para Atleta	75
Figura 4.8 Aprendizado Acoplado e Não Acoplado Cumulativo 10 para Jogo	75
Figura 4.9 Figura	75
4.10 Aprendizado Acoplado e Não Acoplado Cumulativo 10 para Treinador .	75
4.11 Aprendizado Acoplado e Não Acoplado Cumulativo 10 para Companhia	75

Figura 4.12 Aprendizado Acoplado e Não Acoplado Cumulativo 10 para País	76
Figura 4.13 Aprendizado Acoplado e Não Acoplado Cumulativo 10 para Setor Econômico.....	76
Figura 4.14 Aprendizado Acoplado e Não Acoplado Cumulativo 10 para Hobby.....	76
Figura 4.15 Aprendizado Acoplado e Não Acoplado Cumulativo 10 para Pessoa	76
Figura 4.16 Aprendizado Acoplado e Não Acoplado Cumulativo 10 para Político	76
Figura 4.17 Aprendizado Acoplado e Não Acoplado Cumulativo 10 para Produto.....	76
Figura 4.18 Aprendizado Acoplado e Não Acoplado Cumulativo 10 para Tipo de Produto.....	77
Figura 4.19 Aprendizado Acoplado e Não Acoplado Cumulativo 10 para Cientista	77
Figura 4.20 Aprendizado Acoplado e Não Acoplado Cumulativo 10 para Esporte.....	77
Figura 4.21 Aprendizado Acoplado e Não Acoplado Cumulativo 10 para Equipe Esportiva	77
Figura 4.22 Percentual de erros e acertos em todas as categorias no Aprendizado Acoplado Cumulativo 10	77
Figura 4.23 Percentual de erros e acertos em todas as categorias no Aprendizado Não Acoplado Cumulativo 10.....	77
Figura 4.24 Erros e Acertos no Aprendizado Acoplado e Não Acoplado Cumulativo 10	78
Figura 4.25 Aprendizado Acoplado e Não Acoplado Cumulativo 3 para Cidade	81
Figura 4.26 Aprendizado Acoplado e Não Acoplado Cumulativo 3 para Ator.....	81
Figura 4.27 Aprendizado Acoplado e Não Acoplado Cumulativo 3 para Atleta	81
Figura 4.28 Aprendizado Acoplado e Não Acoplado Cumulativo 3 para Jogo.....	81

Figura 4.29 Aprendizado Acoplado e Não Acoplado Cumulativo 3 para Treinador.....	81
Figura 4.30 Aprendizado Acoplado e Não Acoplado Cumulativo 3 para Companhia.....	81
Figura 4.31 Aprendizado Acoplado e Não Acoplado Cumulativo 3 para País	82
Figura 4.32 Aprendizado Acoplado e Não Acoplado Cumulativo 3 para Setor Econômico.....	82
Figura 4.33 Aprendizado Acoplado e Não Acoplado Cumulativo 3 para Hobby	82
Figura 4.34 Aprendizado Acoplado e Não Acoplado Cumulativo 3 para Pessoa	82
Figura 4.35 Aprendizado Acoplado e Não Acoplado Cumulativo 3 para Político	82
Figura 4.36 Aprendizado Acoplado e Não Acoplado Cumulativo 3 para Produto.....	82
Figura 4.37 Aprendizado Acoplado e Não Acoplado Cumulativo 3 para Tipo de Produto.....	83
Figura 4.38 Aprendizado Acoplado e Não Acoplado Cumulativo 3 para Cientista	83
Figura 4.39 Aprendizado Acoplado e Não Acoplado Cumulativo 3 para Esporte.....	83
Figura 4.40 Aprendizado Acoplado e Não Acoplado Cumulativo 3 para Equipe Esportiva	83
Figura 4.41 Percentual de erros e acertos em todas as categorias no Aprendizado Acoplado e Não Acoplado Cumulativo 3.....	83
Figura 4.42 Percentual de erros e acertos em todas as categorias no Aprendizado Acoplado e Não Acoplado Cumulativo 3.....	83
4.43 Erros e Acertos no Aprendizado Acoplado e Não Acoplado Cumulativo 3	84
Figura 4.44 Aprendizado de Acoplado Cumulativo de Pessoa.....	85
Figura 4.45 Erros e Acertos no Aprendizado de Pares de ENs com Relações Semânticas com seis iterações.....	89

Figura 4.46 Experimento de Acoplamento de Relações Semânticas e Tipagem com Filtro de 1/3 dos Pares de ENs	91
Figura 4.47 Experimento de Acoplamento de Relações Semânticas e Tipagem sem Filtro antes da Promoção.....	92
Figura 4.48 Aprendizado de ENs com todos os acoplamentos	93
Figura 5.1 Base de Conhecimento - Nova Proposta	99

LISTA DE TABELAS

Tabela 3.1 Padrões Positivos e Negativos	50
Tabela 3.2 Padrões Negativos de Cidade	51
Tabela 3.3 Padrões Negativos de Pessoa.....	51
Tabela 3.4 PTs para categoria	59
Tabela 3.5 Par de EN	59
Tabela 3.6 Exemplo de Tipagem.....	59
Tabela 4.1 Experimentos para o Aprendizado de categorias	68
Tabela 4.2 Experimentos para o Aprendizado de categorias e Relações Semânticas.....	69
Tabela 5.1 Detalhamento de Nova Proposta da Base de Conhecimento.....	100

LISTA DE ABREVIATURAS E SIGLAS

EN – *Entidade Nomeada*

PT – *Padrão Textual*

NELL – *Never Ending Language Learning*

RTW – *Read The Web*

RTWP – *Read The Web in Portuguese*

SASF – *Sistema de Aprendizado Sem Fim*

SUMÁRIO

CAPÍTULO 1 - INTRODUÇÃO.....	16
1.1. Contextualização.....	16
1.2. Objetivos da Pesquisa	20
1.3. Metodologia de Trabalho	21
1.4. Organização do Trabalho.....	23
CAPÍTULO 2 - FUNDAMENTAÇÃO TEÓRICA E TRABALHOS RELACIONADOS	25
2.1 Conceituação	25
2.2 Aprendizado Semissupervisionado	26
2.3 Aprendizado Semissupervisionado no Leitura da Web em Português.....	27
2.4 Co-Training.....	28
2.5 Entidade Nomeada – EN.....	29
2.6 Aprendizado Semissupervisionado usando Co-Training e <i>Bootstrap</i>	31
2.7 <i>Bootstrap</i> no Aprendizado Semissupervisionado de Entidades Nomeadas e categorias.....	32
2.8 Trabalhos relacionados à leitura da Web	34
CAPÍTULO 3 - LEITURA DA WEB EM PORTUGUÊS.....	38
3.1 RTWP - Leitura da Web em Português	38
3.2 Extração de Conhecimento a partir da Web.....	40
3.3 Base de Conhecimento	43
3.4 Padrões Negativos	46
3.5 Identificação e Extração de ENs e PTs para categorias	47
3.5.1 Padrões Negativos na Promoção de ENs e PTs.....	50
3.5.2 Promoção de ENs e PTs para categorias	51
3.5.3 Processo de Aprendizado	52
3.6 Identificação e Extração de Pares de ENs e PTs de Relações Semânticas	54

3.6.1 Padrões Negativos para Promoção Pares ENs e PTs de Relações Semânticas	56
3.6.2 Promoção de Pares de ENs e PTs de Relações Semânticas	56
3.6.3 Processo de Aprendizado	57
3.7 Identificação e Extração de Pares de ENs com Tipagem e PTs de Relações Semânticas.....	58
3.7.1 Promoção de Pares de ENs e PTs de Relações Semânticas com Tipagem....	60
3.7.2 Processo de Aprendizado	61
3.8 Funcionamento de Todos os Acoplamentos: Aprendizado de ENs, Pares de ENs, PTs de Categorias e PTs de Relações Semânticas com Tipagem	63
3.9 Visão Geral do Projeto	64
CAPÍTULO 4 - EXPERIMENTOS E ANÁLISE	67
4.1 Experimentos	67
4.2 Experimentos I e II – Não Acoplados e Não Cumulativos 10 e 3	71
4.3 Experimentos III e IV – Acoplados e Não Cumulativos 10 e 3	72
4.4 Experimentos V e VII – Acoplados e Não Acoplados Cumulativos 10	74
4.5 Experimentos VI e VIII – Acoplados e Não Acoplados Cumulativos 3	80
4.6 Considerações dos Experimentos VII e VIII	88
4.7 Experimento IX – Relações Semânticas Acoplado com Padrões Negativos	89
4.8 Experimento X – Relações Semânticas Acoplado com Tipagem e Padrões Negativos (Com filtro antes da promoção)	90
4.9 Experimento XI - Relações Semânticas Acoplado com Tipagem e Padrões Negativos (Com filtro antes da promoção)	91
4.10 Considerações dos Experimentos IX, X e XI – Aprendizado de ENs	93
4.11 Análise de Todos os Experimentos	94
CAPÍTULO 5 - CONCLUSÃO	96
5.1 Objetivos Alcançados	96
5.2 Contribuições e Limitações	98
5.3 Trabalhos Futuros	101
REFERÊNCIAS BIBLIOGRÁFICAS.....	103

Capítulo 1

INTRODUÇÃO

Um dos objetivos principais do Aprendizado de Máquina é buscar métodos que aprendam e sejam capazes de melhorar seu desempenho com base em experiências anteriores. Neste Capítulo serão introduzidos o problema investigado, a motivação para este trabalho, as hipóteses de pesquisa e as justificativas. O capítulo está estruturado da seguinte maneira: Contexto, Objetivos, Metodologia e Organização do Trabalho.

1.1. Contextualização

O Aprendizado de Máquina (AM) pode ser visto como uma área de pesquisa dentro da Inteligência Artificial (IA) que tem como um de seus principais objetivos o desenvolvimento de programas de computador que possam evoluir à medida que vão sendo expostos a novas experiências [30]. Assim, há grande investimento na busca por métodos e técnicas que permitam a concepção de sistemas computacionais capazes de melhorar seu desempenho, de maneira autônoma, usando informações obtidas ao longo de seu uso; tal característica pode de certa forma, ser considerada como um dos mecanismos fundamentais que regem os processos de aprendizado automático [31].

Existem basicamente três abordagens de aprendizado de máquina, a saber, Aprendizado Supervisionado [30], Semissupervisionado [46] e Não Supervisionado [30]. A terceira abordagem, chamada Aprendizado Semissupervisionado, a qual foi usada nesse trabalho, pode ser ilustrada por meio de uma situação em que há um número pequeno de clientes de L já rotulados, mas esse número não é suficiente para guiar de maneira adequada a indução de um modelo usando técnicas de Aprendizado Supervisionado. Assim, os poucos clientes já rotulados são utilizados na categorização de novos clientes, os quais serão utilizados para ampliar a quantidade de dados rotulados a serem utilizados na supervisão do aprendizado. Pode-se notar então que um processo de supervisão é utilizado na transformação de instâncias não-rotuladas em rotuladas.

Mesmo com todo o avanço e conquistas na área de AM, entretanto, ainda se busca um sistema computacional que aprenda de maneira incremental e que seja capaz de usar o conhecimento aprendido previamente para, continuamente, ir refinando sua capacidade de aprendizado ao longo do tempo. Essa característica, que pode ser chamada de “Aprendizado Sem Fim” (ASF), permitiria que um sistema computacional de aprendizado evoluísse constantemente.

Um processo de ASF deve ir além da definição de um modelo fixo e estático a ser utilizado para inferência e, também, além do simples processo incremental de acúmulo de conhecimento. O aprendizado deve acontecer continuamente de forma que o conhecimento adquirido (além de ser usado para incrementar uma base de conhecimento) sirva também para, dinamicamente, fazer com que o desempenho do sistema melhore continuamente. Assim, caso a mesma fonte de informação, já

utilizada para o aprendizado anterior, seja novamente fornecida, o sistema deverá ser capaz de aprender mais e melhor do que havia aprendido anteriormente.

Observando as características desejáveis em um Sistema de Aprendizado Sem Fim (SASF), é possível notar que são, de certa forma, similares às de sistemas que implementam o Aprendizado Semissupervisionado; em ambas as abordagens de aprendizado, a partir de uma pequena quantidade de conhecimento inicial, busca-se aprender mais utilizando informações previamente não rotuladas. Esta similaridade levou parte da comunidade de AM à investigação e ao desenvolvimento de métodos de Aprendizado Semissupervisionado que incorporassem as características específicas de aprendizado sem fim. O uso de técnicas de Aprendizado Semissupervisionado na geração de um SASF, entretanto, provoca um problema conhecido como “desvio do conceito” (concept drift)¹, como será visto com mais detalhes no Capítulo 2. Esse desvio ocorre no Aprendizado Semissupervisionado, por exemplo, quando o modelo rotula de maneira incorreta um novo exemplo E1 (anteriormente não-rotulado) e, a partir daí, passa a utilizá-lo para rotular novos exemplos (E2, E3, ..., En) incorretamente. Essa característica faz com que, com o passar do tempo, o sistema aprenda incorretamente mais e mais conceitos o que, de certa forma, inviabiliza um SASF [17].

Além das dificuldades intrinsecamente ligadas à definição de um modelo de AM adequado a um SASF, existem ainda dificuldades relacionadas à forma de representação e recuperação do conhecimento. É fundamental e determinante para a criação de um SASF a definição de uma estrutura que viabilize a contínua

¹ O desvio de conceito não significa necessariamente um erro. Alguns conceitos podem mudar com passar do tempo e a identificação destas mudanças pode, neste caso, ser muito importante para manter a correteude no aprendizado.

evolução e crescimento da base de conhecimento do sistema, bem como permita a manipulação de uma enorme quantidade de informações (no estudo de caso proposto neste projeto, a Web em português).

Assim, nota-se que a busca por um SASF constitui-se ainda em um grande desafio para a comunidade de AM e IA em geral. Essa é uma das principais motivações e justificativas deste trabalho de pesquisa ser direcionado à investigação do problema de aprendizado sem fim.

A hipótese de pesquisa é que o problema de desvio de conceito pode ser minimizado nos métodos de aprendizado semissupervisionado. Para tanto, deve-se acoplar tarefas de aprendizado de maneira a permitir que resultados de algoritmos de aprendizado semissupervisionado possam ser (automaticamente) utilizados pelo SASF para minimizar o desvio de conceito.

Buscando evidências empíricas para dar suporte para a hipótese de pesquisa, acima apresentada, após o estudo e investigação de diferentes métodos e técnicas relacionadas à construção de um SASF, foi desenvolvido um SASF para a extração de conhecimento com base na ideia de acoplamento de tarefas de aprendizado semissupervisionado. Este SASF chama-se RTWP (iniciais para “Leitura da Web em Português” em inglês - “Read The Web in Portuguese”). O RTWP pode ser visto como a maior contribuição deste trabalho e permitiu a verificação empírica de que técnicas de acoplamento de algoritmos de aprendizado semissupervisionado podem trazer ganho para a extração de conhecimento a partir de textos em português, assim como já haviam trazido em tarefas de extração de conhecimento a partir de textos em inglês (como pode ser visto nos resultados do projeto RTW (<http://rtw.ml.cmu.edu/>) disponíveis em:

<http://rtw.ml.cmu.edu/readtheWeb.html> que mostram a viabilidade do trabalho.

[6][7][13][29]

1.2. Objetivos da Pesquisa

O principal objetivo da pesquisa descrita neste documento foi investigar, propor e implementar métodos e algoritmos de aprendizado semissupervisionado que permitam a minimização do problema de desvio de conceito. Além disso, para mostrar empiricamente os resultados obtidos, foi construído um sistema computacional capaz de realizar a extração de conhecimento a partir da Web em português, por meio da criação de uma base de conhecimento consistente, atualizada constantemente à medida que novos conhecimentos vão sendo extraídos através do aprendizado semissupervisionado.

Mais especificamente, o sistema computacional proposto e implementado é capaz de, a partir de uma base de conhecimento inicial (inspirada na estrutura de ontologias), ser executado continuamente por um período de tempo com dois objetivos específicos, a saber:

Extração: extrair mais “conhecimento” a partir da Web em português visando à expansão da base de conhecimento inicial (ontologia inicial);

Aprendizado: aprender a extrair melhor e com mais precisão que no ‘dia anterior’.

A ontologia inicial a ser fornecida ao sistema deve especificar um conjunto de categorias (que são tipos de informações ou instâncias presentes na ontologia, por exemplo, Cidade, País, Estado, Ator, Político, Empresa, Universidade, etc.) e

Relações Semânticas (são as relações entre as categorias, por exemplo, Prefeito(Político, Cidade), CapitalDoPaís(Cidade, País), CapitalDoEstado(Cidade, Estado), etc.) usadas como guia na extração do conhecimento e que são populadas com instâncias encontradas na Web. Assim, pode-se ter, por exemplo, São Carlos, Araraquara, Campinas, etc. como instâncias da categoria “Cidade”. E como instâncias para popular a relação “Prefeito(X,Y)” podem ser encontrados os Pares “Oswaldo Barba; São Carlos”, “Eduardo Paes; Rio de Janeiro”, etc. Um exemplo de ontologia inicial pode ser visto em <http://rtw.ml.cmu.edu/readtheWeb.html>. A ontologia deve ser constantemente atualizada pelo sistema, formando assim uma base conhecimento dinâmica.

Os algoritmos desenvolvidos e implementados neste projeto foram avaliados com relação ao desempenho bem como à capacidade de continuar aprendendo com o passar do tempo. Assim, a avaliação foi realizada com base nas duas tarefas descritas anteriormente, a saber, “**extração**” e “**aprendizado**”. A partir de uma avaliação manual é possível identificar se o sistema conseguiu melhorar seu desempenho com o uso dos métodos de acoplamento desenvolvidos.

1.3. Metodologia de Trabalho

A ideia básica utilizada para definir a metodologia de trabalho é que com a integração de vários processos de aprendizado semissupervisionado [46] é possível minimizar a divergência do aprendizado que comumente ocorre com o problema do desvio de conceito [17]. Mesmo tendo alguns pontos em comum, a metodologia aqui utilizada não está vinculada às ideias de extração de informação aberta (como

descrito em Banko e Etzioni, 2007 [5]) e nem tampouco ao uso de aprendizado incremental, mas sim aos princípios de aprendizado sem fim (never-ending learning) definidos em [6].

Resultados mostrados em [7] sugerem que o aprendizado integrado de tarefas de Aprendizado Semissupervisionado pode viabilizar o desenvolvimento de um SASF.

Neste trabalho também foram obtidos resultados que, empiricamente, mostram a viabilidade do uso de tarefas/métodos integrados para o desenvolvimento de um sistema computacional que incorpore características do ASF para a Web em português. Assim, o trabalho mostrou que o aprendizado semissupervisionado pode ser realizado com muito mais êxito por meio da integração contínua de tarefas de aprendizado do que por meio de tarefas de aprendizado isoladas. Como consequência, o problema de extração de conhecimento a partir da Web pode obter resultados mais satisfatórios. As tarefas de aprendizado Semissupervisionado integradas foram:

- Identificação e extração a partir de páginas Web de “Entidades Nomeadas” (**ENs**) a partir de Padrões Textuais (**PTs**);
- Identificação e extração a partir de páginas Web de “Padrões Textuais” (**PTs**) a partir de **ENs**;
- Identificação e extração a partir de páginas Web de Pares de **ENs** a partir de **PTs** de Relações Semânticas;
- Identificação e extração a partir de páginas Web de **PTs de Relações Semânticas** a partir de Pares de **ENs**;

ENs, neste projeto, são considerados todos os substantivos. Pode-se ler mais sobre o assunto em [42]. Os **PTs** foram definidos aqui como as palavras que estão antes ou depois de uma EN, ou ainda entre Pares de ENs. **Relações Semânticas** são as relações entre ENs.

A integração das tarefas acima discriminadas seguiu a metodologia definida em [7]. Assim, a identificação de ENs com base em PTs foi integrada (através de um mecanismo de *Bootstrapping*) formando um método acoplado de identificação de ENs e PTs (Acoplamento 1). Da mesma forma, a identificação e extração de Pares de ENs através de PTs de Relações Semânticas foi acoplada à identificação e extração de PTs de Relações Semânticas a partir de Pares de ENs (Acoplamento 2). Na sequência, o Acoplamento 1 e o Acoplamento 2 foram acoplados formando o Acoplamento 3. A ideia é que quanto mais acoplamentos o processo possa ter (desde que sejam acoplamentos adequados), maior será o ganho no desempenho. Mais detalhes podem ser vistos no Capítulo 3.

1.4. Organização do Trabalho

O Capítulo 2 é referente à fundamentação teórica e aos trabalhos correlatos, em que são apresentados métodos usados ou que podem ser usados futuramente no RTWP, ou ainda, abordagens de problemas relacionados ao objetivo buscado. No Capítulo 3 são abordadas como foram realizadas e usadas as formas de acoplamento, as Stop Words, a estrutura de ontologia (base de conhecimento) e o modo de extração e identificação de ENs, PTs, Pares de ENs e PTs de Relações Semânticas a partir da Web. No Capítulo 4 são mostrados os resultados obtidos e suas respectivas análises. Cada experimento é analisado quanto à precisão, cobertura e ao uso da estrutura ontológica escolhida. O Capítulo 5 apresenta os

objetivos alcançados, contribuições e limitações, e os projetos futuros juntamente com possíveis melhoras na solução do problema que foi abordado neste trabalho. No final estão apresentados os anexos mais detalhados sobre os experimentos e as Stop Words.

Capítulo 2

FUNDAMENTAÇÃO TEÓRICA E TRABALHOS RELACIONADOS

O aprendizado de máquina estuda algoritmos capazes de melhorar com a experiência. Neste capítulo serão apresentadas as formas básicas de aprendizado de máquina necessárias para o entendimento deste projeto. As seções são as seguintes: Fundamentação Teórica, Aprendizado Semissupervisionado no Leitura da Web, Aprendizado Semissupervisionado usando Bootstrap, Bootstrap no Aprendizado de Entidades Nomeadas e categorias. Aprendizado Semissupervisionado com Modelos Exponenciais e Co-Regularização e Visões Múltiplas.

2.1 Conceituação

“Desde a invenção dos computadores questiona-se se eles poderiam aprender. Se cientistas pudessem compreender como programá-los para aprender a melhorar automaticamente com a experiência o impacto seria dramático. Imagine computadores aprendendo a partir de registros médicos quais são os tratamentos mais eficazes para novas doenças, sistemas de gerência habitacionais aprendendo a partir da experiência como otimizar custos de energia baseando-se nos padrões de uso particular dos seus ocupantes, ou assistentes de softwares aprendendo a

evoluir a partir dos interesses de seus usuários, a fim de destacar notícias novas e interessantes durante a manhã” [30].

Um sistema de Aprendizado de Máquina “aprende” com suas experiências para resolver problemas que surgirem posteriormente [41]. Para ser construído precisam ser definidos três componentes: A tarefa a ser realizada (classificação de texto, aprendizado de entidade, etc.), a fonte de conhecimento (arquivos texto, Web, banco de dados, etc.) e a medida de desempenho a ser otimizada.

2.2 Aprendizado Semissupervisionado

Tradicionalmente, o Aprendizado de Máquina pode ser dividido em Aprendizado Supervisionado e Aprendizado Não Supervisionado, exemplos de aplicações, algoritmos, etc. destes 2 tipos de aprendizado podem ser vistos em [30]. Mais recentemente passou-se também a investigar o processo de aprendizado chamado Aprendizado Semissupervisionado [47].

O Aprendizado Supervisionado é a forma de aprendizagem em que o processo é guiado por alguma forma de supervisão. Tal supervisão pode, por exemplo, estar vinculada a exemplos previamente rotulados que são utilizados para se identificar padrões que permitam a classificação/agrupamento de novos exemplos sem rótulos.

No Aprendizado Não Supervisionado, ao contrário do Supervisionado, não há supervisão no processo de aprendizado. Não há, por exemplo, exemplos previamente rotulados, mas, somente exemplos não rotulados. Ao se executar o

algoritmo de aprendizagem, são formados grupos dos exemplos de acordo com a similaridade (ou dissimilaridade) entre eles.

Já o Aprendizado Semissupervisionado pode ser visto como a mistura entre o Supervisionado e Não Supervisionado, pois há “um pouco” de supervisão no processo, mas há também etapas em que a supervisão não está presente. Para o exemplo em que a supervisão se dá com base em exemplos rotulados, no aprendizado semissupervisionado são utilizados exemplos previamente rotulados e a partir deles são classificados/agrupados novos exemplos sem rótulo que passarão a servir de base para a fase “supervisionada” do processo. Este aprendizado é bastante comum quando se deseja, por exemplo, classificar/agrupar grandes amostras de dados não rotulados a partir de uma pequena amostra de dados rotulados [47].

2.3 Aprendizado Semissupervisionado no Leitura da Web em Português

Neste trabalho o Aprendizado Semissupervisionado é utilizado. Como pode ser visto em [47], o Aprendizado de Máquina Semissupervisionado possui dentre suas características principais a possibilidade de rotular uma grande amostra de dados não rotulados a partir de uma pequena amostra rotulada. Considere, por exemplo, a existência de um conjunto de dados $D1$ formado por instâncias rotuladas ($IR1$) e por instâncias não rotuladas ($INR1$). Assim, $D1 = IR1 \cup INR1$. A semissupervisão, neste caso, ocorre pelo fato de se usar inicialmente o conjunto $IR1$ para iniciar o processo de aprendizagem (etapa de treinamento) e depois, inserir

rótulos em instâncias do conjunto $INR1$ (etapa de predição), gerando assim um novo conjunto de instâncias rotuladas ($IRDA1$) durante o processo de aprendizado. Uma vez gerado o conjunto $IRDA1$, as instâncias deste conjunto passam a ser utilizadas em conjunto com as instâncias do conjunto $IR1$, ou seja, o aprendizado segue utilizando na etapa de treinamento as instâncias do conjunto $IR1 \cup IRDA1$ para identificar padrões. Assim, o conjunto de instâncias utilizados no treinamento é expandindo a cada nova iteração.

2.4 Co-Training

A rotulagem de dados não rotulados precisam de cuidado na definição dos padrões prévios, e dependendo da aplicação, esse é um trabalho difícil de ser feito. Suponha a rotulação de páginas na web, humanamente não seria possível devido ao grande volume de dados, às atualizações ocorrentes a todo instante, etc.

Tendo o objetivo de rotular páginas web, entre outros problemas com abrangência parecida, em [8] foi proposto por Blum e Mitchell o Co-Training. Foram dois os problemas principais abordados: o treinamento e rotulagem automática de páginas web, e a avaliação dos dados de treinamento. Tais problemas foram tratados investindo-se em usar diferentes tipos de informações para se chegar ao mesmo objetivo. Por exemplo, imagine que seja extraída uma página de uma universidade. A abordagem tratou as palavras que ocorreram na página, por exemplo: cursos, professores, graduação, pós-graduação, vestibular, etc. E tratou também os links, por exemplo: “meu orientador”, “minha universidade”, etc. levam

normalmente a sites de uma instituição de ensino. A ocorrência de ambos (palavras e links) indicam maior chance da página ser de uma universidade.

Em [8] o Co-training foi usado no aprendizado semissupervisionado, assim foi dada uma pequena amostra de dados rotulados para que fosse rotulada uma amostra maior. Os dados que foram rotulados eram usados para rotular outras páginas que ainda não eram rotuladas.

Neste trabalho foi usado o Co-training com dois focos principais para o aprendizado: as ENs e os PTs. Com as ENs foram aprendidos novos PTs e com os PTs novas ENs.

2.5 Entidade Nomeada – EN

Diferentes definições sobre Entidades Nomeadas (ENs). Em alguns casos ENs são definidas como palavras da classe de substantivos próprios, como: pessoas, locais, organizações, obras, etc. Além disso, a mesma definição é encontrada às vezes como Entidades Mencionadas (EMs). Em alguns casos encontra-se ENs definidas como substantivos próprios, enquanto EMs possuem a mesma definição, porém dependem do contexto.

Devido as diferentes definições existentes na literatura, a abordagem usada neste trabalho para ENs é voltada ao problema estudado. A definição de ENs para este trabalho é apresentada na Definição 1.

Definição 1 **EN:** são palavras da classe de substantivos, próprios ou não. Ex: Pessoa, esporte, hobby, lugar, organização, marca, produto, etc.

A abordagem para essa definição foi de acordo com o uso realizado em [13], em que são apresentados os resultados iniciais do NELL.

O foco de extração e aprendizado neste projeto são as ENs, para melhor entendimento são dadas algumas definições abaixo.

Definição 2 **PTs:** são padrões textuais que acompanham uma EN, após, ou antes. Ex: “X **é uma bela cidade**”, em que X é uma EN e o restante é um PT.

Definição 3 **Pares de ENs:** são os Pares de ENs que são aprendidos a partir de PTs de Relações Semânticas. Ex: **São Carlos** é uma cidade do **Brasil**, em que **São Carlos** e **Brasil** é um Par de EN.

Definição 4 **PTs de Relações Semânticas:** são os PTs entre os Pares de ENs. Ex: São Carlos **é uma cidade do** Brasil, em que “**é uma cidade do**” é um PT de Relação Semântica.

Definição 5 **Cobertura:** foi calculada de acordo com o total de aprendizado promovido (ENs).

Definição 6 **Precisão:** foi calculada de acordo com a porcentagem do aprendizado promovido (ENs) corretamente.

Definição 7 **Categoria:** são diferentes tipos de informações que o RTWP objetivou aprender, conforme foi definido no NELL em [13].

Definição 8 **Padrões:** exemplos de ENs, PTs, Pares de ENs e PTs de Relações Semânticas que são promovidos e podem ser usados como padrões para novas extrações/aprendizado. Quando são chamados de padrões podem ser entendidos também como dados rotulados (que se conhece a categoria).

Definição 9 Extração: é a tarefa de extrair possíveis padrões vindos de páginas web.

Definição 10 Promoção: a promoção é realizada após a extração, em que são promovidos os padrões com evidências suficientes para que o RTWP possa “acreditar” que são corretos.

Definição 11 Padrões fracos: são padrões que não possuem evidências suficientes para serem promovidos, ou que em futuras extrações podem trazer novos padrões errados.

Definição 12 Padrões fortes: são padrões que possuem evidências suficientes para serem promovidos.

No decorrer do projeto serão dadas outras definições quando se fizerem necessárias.

A extração de padrões foi realizada de acordo com as necessidades para a língua portuguesa vistas no decorrer do desenvolvimento do projeto.

2.6 Aprendizado Semissupervisionado usando Co-Training e Bootstrap

A técnica de *Bootstrap* [44] usa diferentes informações para alcançar um mesmo objetivo. Neste trabalho o objetivo é aprender ENs e as diferentes visões são o aprendizado de ENs e de PTs, em que os PTs ajudam a popular ENs e vice-versa. Tal abordagem tem apresentado resultados promissores em aplicações como, por exemplo, na eliminação de ambiguidade das palavras [28][44], na classificação de

páginas Web, na classificação de entidades nomeadas [16], *parsing* para desambiguação [28] e na tradução automática [39].

Uma das primeiras abordagens do *Bootstrap* foi dada em [44], este trabalho teve foco no problema de eliminação de ambigüidade das palavras. Para eliminação da ambigüidade eram contadas as ocorrências em cada palavra em cada categoria de informação (Ex: Manga da blusa ou manga fruta?). Em [8], para classificar páginas Web, foi utilizado o Co-Training [8] juntamente com *Bootstrap*. O primeiro classificador utiliza como atributos as palavras presentes no texto das páginas Web (numa abordagem chamada “bag of words”). O segundo classificador utiliza como atributos palavras usadas em *hiperlinks* presentes na página. As predições mais confiáveis de cada classificador em cada iteração são então utilizadas para rotular mais documentos, e os classificadores retreinados. Os autores justificam que essa abordagem trata da maximização das previsões de classificação de texto sobre os dados rotulados. Em [1] argumenta-se que o algoritmo *Co-Training* não faz realmente a maximização, devido a isso é proposto o algoritmo *Greedy Agreement Algorithm (GAA)* [44] com os dados de entrada usados em [16], em que foi realizado um emparelhamento de regras com dados não rotulados, obtendo-se resultados similares ao *Co-training*.

2.7 *Bootstrap* no Aprendizado Semissupervisionado de Entidades Nomeadas e categorias

A partir do momento em que o algoritmo *Bootstrapping* é executado muitas vezes, há uma tendência em gerar erros de sentido semântico, ou seja, de

ambiguidade, então o aprendizado pode passar a trazer cada vez mais dados incorretos (desvio de conceito) [17]. Esse problema foi abordado neste trabalho, em que se focou em minimizar o desvio de conceito.

Um dos primeiros trabalhos no domínio classificação de entidades nomeadas é descrito em [16], em que um dos algoritmos usados foi baseado no trabalho anterior de Yarowsky [44] e Blum e Mitchell [8]. Dado uma amostra pequena de regras, foram aprendidas de acordo com a definição do Bootstrap.

Em [35], foi descrito um método de aprendizado categorias de entidades nomeadas através de um *parser* superficial, chamado “shallow parser”. Os padrões rotulados de cada categoria de interesse foram usadas para aprender novos padrões. Esses padrões são usados para aprender outros novos padrões

O “*Mutual Bootstrapping*” é o método também adotado no trabalho descrito neste documento no Capítulo 3. Ainda nesse aprendizado é introduzida a ideia de *meta-Bootstrapping*, em que são executadas muitas iterações do algoritmo, e é realizado um *ranking* dos padrões e das instâncias que mais ocorreram, então os melhores colocados no *ranking* são promovidos, ou seja, passam a fazer parte dos exemplos. Foram obtidos bons resultados nesta abordagem, os quais se mantiveram perto de 80% de precisão.

Para extração de Relações Semânticas, um dos primeiros trabalhos foi o método chamado DIPRE (*Dual Iterative Pattern Relation Expansion*) descrito em [9], que possui um pequeno número de padrões de relações rotulados inicialmente (Ex: AuthorOf(Author,Book)) e através do *bootstrapping* descobre padrões de extração da Web (Ex: Isaac Asimov, The Robots of Dawn).

Em [2] foi construído o *SnowBall*, uma melhoria do *DIPRE* utilizando padrões mais flexíveis, métodos adicionais de marcação e Relações Semânticas padrões potencialmente melhores de instância, além de uma nova metodologia de avaliação.

Em [34] o objetivo foi extrair Relações Semânticas de documentos Web (Ex: BornIn(person, date)), usando-se de generalização de padrões baseados em classes Semânticas de palavras (Ex: is, was, has, does, could).

Em [35] e [43] foi apresentada mais uma forma de acoplamento com padrões positivos e negativos. Em que os padrões negativos de uma categoria eram os PTs positivos de outras categorias.

2.8 Trabalhos relacionados à leitura da Web

Sarawagi realizou em [36] uma pesquisa de técnicas de extração de informações a partir da Web. Para tal trabalho foi abordado o reconhecimento de entidades nomeadas (ENs) de [32]. Um dos primeiros trabalhos desenvolvidos que usa extração de padrões (para a extração de ENs) é o de Hearst [21]. Essa abordagem propôs o uso de hiponímias, as quais podem ser usadas para extrair candidatos. Exemplo: IsA(X,Y) (protozoa,paramécium).

Etzioni et al.[19] também apresentaram métodos para aumentar a cobertura do sistema “KnowItAll” enquanto buscava-se a alta precisão. O KnowItAll busca descobrir categorias específicas, as quais são usadas tanto para extrair novas instâncias candidatas quanto para evoluir instâncias candidatas. Por exemplo, sabe-se que “biologist” é um tipo de cientista, então é possível usar o padrão “biologists such as X” para extrair mais cientistas. Os métodos usados incluem: descoberta de

padrões, extração de subclasses e extração de listas. A extração de lista usa técnicas de *indução wrapper* para descobrir listas de instâncias da Web, e então extrai novas instâncias candidatas da categoria descoberta na lista. Todos os métodos obtiveram melhor precisão.

Em [19] foi apresentado um modelo probabilístico formal chamado URNS, este modelo funciona como urnas, ou seja, são pontuados votos. Cada ocorrência de uma instância extraída por padrão é representada como um voto na urna. A questão chave para se responder com o modelo é: Se uma instância ocorre K vezes de N retiradas da urna, qual a probabilidade que isto seja uma extração correta?

Em [23] foram usados dois itens importantes para melhorar a precisão, em: “cities, such Pittsburgh and X”; Pittsburgh sabe-se que é uma cidade e “cities, such” é um padrão para cidades, ou seja, há validação dos PTs e das ENs antes de se promover tais informações. Essa abordagem foi usada nesse projeto a chamada *Tipagem* como pode ser visto no Capítulo 3.

Kushmerick [25] abordou de forma geral métodos de indução e em [15] seu trabalho foi expandido e passou a explorar estruturas de páginas HTML, e não somente prefixo e sufixo do que foi extraído. Ex: métodos *wrappers* para extrair registros que contenham autor, título, preço de um livro, em que a página descreve a venda desse livro (site específico de venda de livros).

Wang e Cohen em [40] usaram padrões de sufixos e prefixos em wrappers que podem ser usados para extrair a partir de uma variedade de linguagens de marcação. Foram obtidos bons resultados utilizando esse modelo.

Hipótese de distribuição indica que uma palavra pode ser caracterizada por uma outra palavra que co-ocorre no texto (sinônimos) [20]. Essa ideia é aplicada em [22] em um corpus de 6 milhões de palavras, em que se conseguiu determinar que

palavras diferentes possuem o mesmo valor semântico (Ex: *Decision* e *Ruling*). Em [33] também foi utilizada esta abordagem juntamente com pareamento, em que melhores resultados foram obtidos. Concluiu-se que o tamanho do corpus teve efeito positivo nos resultados, por conta da maior repetição de termos;

O Open IE é um projeto que extrai informações da Web, a diferença entre ele e este projeto apresentado nesta dissertação, é que o Open IE busca informações sem um domínio específico, enquanto o RTWP extrai a partir de uma ontologia já definida. A extração de domínio aberto de texto da Web também ocorre em [18], em que são produzidos uma coleção de conjuntos de exemplos rotulados.

Shinyama e Sekine [2006] descobriram relações entre Pares de entidades através de agrupamento (*clustering*). A partir disso foi realizado o alinhamento relacional em uma tabela (ex: Uma coluna numa tabela seria “furacão” e em outra coluna é o local danificado). Por utilizar pareamento e agrupamento de documentos, essa abordagem torna-se cara computacionalmente.

Ainda sobre o Open IE, foi apresentado por Banko et al. em [4] o sistema *TextRunner*, que usa o *Naïve Bayes* para classificar Pares de entidades nomeadas. Novamente Banko e Etzioni em [5] apresentam um novo método, o CRF (*Conditional Random Field*) que substituiu o *Naïve Bayes* no *TextRunner*. Com esta troca de algoritmos (NB pelo CRF) foi possível obter melhor precisão. Em [45] foi criado o *Resolver* inspirado em *URNS*, para avaliar a probabilidade de duas palavras serem sinônimas, e tratar triplas extraídas pelo *TextRunner*. O *Resolver* faz agrupamento por sinônimos (Ex: Marte e Planeta vermelho) e as Relações Semânticas entre eles (gira em torno de órbitas).

Outro trabalho de destaque, voltado ao Open IE, é descrito por Cafarella et. al [12], em que são agregadas saídas do *TextRunner*, *WebTables* e um sistema de

busca na *Deep Web*. O *TextRunner* extrai triplas de strings textuais que denotam uma relação entre um Par de entidades. *WebTables* [10][11] extrai tuplas de elementos de tabelas HTML. E o terceiro método citado apresenta consultas a motores de busca na Web [26].

As técnicas principais que foram usadas no trabalho apresentado nesta dissertação de mestrado foram o “*Mutual Bootstrapping*” [44] com base no *Co-Training* [8], *Tipagem* [23] com o Aprendizado Semissupervisionado [46]. Toda a abordagem e proposta teve por base o projeto NELL [6], [7], [13], e [29].

Capítulo 3

LEITURA DA WEB EM PORTUGUÊS

Leitura da Web em Português, "Read The Web in Portuguese" (RTWP) ou ainda a busca por um Sistema de Aprendizado Sem Fim (SASF), é a aplicação apresentada neste projeto de mestrado a qual foi inspirada no NELL (<http://rtw.ml.cmu.edu>). Este Capítulo está organizado da seguinte forma: Primeiramente são apresentadas as formas que conceitos usados foram implementados no RTWP e em seguida os resultados obtidos e suas respectivas considerações.

3.1 RTWP - Leitura da Web em Português

O RTWP foi inspirado e se inclui no NELL (Never-Ending Language Learning) [6], [7], [29] e [13], em desenvolvimento pela Carnegie Mellon University, porém desenvolvido para a língua portuguesa. Toda a concepção foi investigada e proposta para as características específicas da língua portuguesa. Da mesma forma, a implementação foi desenvolvida do início (não foram traduzidos e nem utilizados métodos prontos já definidos para o NELL). Um outro ponto que diferencia o RTWP do componente de extração de ENs a partir de texto implementado no NELL, é o fato do RTWP extrair conhecimento diretamente da Web, enquanto no NELL a extração é feita a partir de um cópulus de páginas Web previamente armazenado em disco.

O NELL está sendo desenvolvido com a intenção de definir formalmente e comprovar que a técnica de “aprendizado sem fim” é eficiente e viável na teoria e também em aplicações reais. Para tanto, pretende-se mostrar que um computador pode aprender de forma autônoma e utilizar seus conhecimentos para evoluir seu próprio processo de aprendizado e ampliar, cada vez mais, sua base de conhecimento. Pretende-se mostrar que um computador pode continuamente adquirir conhecimento e ter autonomia suficiente para revisar e ampliar seu conhecimento a partir de novas descobertas sem o uso de supervisão e com confiabilidade. Existem técnicas que fazem isso, porém a confiabilidade tende a cada vez cair mais a cada iteração.

Este tipo de aprendizado contínuo é inspirado na forma seres humanos aprendem. Um ser humano inicia a construção de sua base de conhecimento a partir de pequenas dicas (de seus pais, por exemplo) e da exploração do ambiente, quanto mais aprende, mais independente e capaz se torna o ser humano para continuar aprendendo, e assim também é o NELL. A ideia da "Leitura da Web" é utilizar uma aplicação real para mostrar a viabilidade do NELL. Assim, o NELL recebeu (no início) algumas informações de temas sobre os quais ele deveria aprender (localidades, empresas, livros, pessoas, etc.) e a partir daí começou a "ler" a Web para extrair conhecimento acerca dos temas definidos.

A base de conhecimento do NELL (bem como informações mais técnicas sobre os algoritmos computacionais utilizados) podem ser encontradas no endereço eletrônico <http://rtw.ml.cmu.edu>.

O RWTP seguiu os passos do NELL e recebeu informações sobre temas (categorias) diversos, com isso extrai páginas da Web e adquire conhecimento para extrair outras páginas e assim por diante.

3.2 Extração de Conhecimento a partir da Web

Neste trabalho, entende-se por **extração de conhecimento a partir da Web** a capacidade de realizar a leitura de páginas da Web e identificar estruturas que possam ser Entidades Nomeadas (ENs)² ou Padrões Textuais (PTs)³.

Para que o aprendizado sem fim seja viável é necessário definir uma forma na qual o conhecimento adquirido possa ser reutilizado na sequência do aprendizado. Imagine que o RTWP deva aprender quais são as cidades do mundo. Assim, inicialmente fornecem-se a ele algumas dicas de leitura (que são os chamados ENs e PTs aqui) que o auxiliarão na identificação de cidades em textos disponíveis na Web em português. Sempre (ou quase sempre) que é encontrada a sentença "**X é uma cidade localizada...**" o termo X refere-se a uma cidade. Assim, o sistema passa ler a Web em busca destes padrões. Após alguma leitura, suponha que as cidades "**São Paulo**", "**São Carlos**" e "**Curitiba**" podem ter sido identificadas. Tais padrões extraídos são candidatos à promoção, em que candidato(s) são:

Definição 13 Candidatos: são todos os padrões extraídos.

As cidades são promovidas⁴ ou não de acordo com a ocorrência nas páginas web. A partir da promoção⁴, o RTWP possui agora condições de definir (autonomamente) novas formas de identificação de cidades. Estas novas formas podem ser, por exemplo, através da constatação de que a sentença "**a prefeitura**

² Definição na página 47

³ Definição na página 48

⁴ Definição na página 49

municipal de X" ocorre com todas as cidades já aprendidas (ou a maioria delas). Assim, este novo PT de identificação de cidades passa a ser usado para que novas cidades possam ser identificadas. Um ponto muito importante no aprendizado sem fim é que haja uma forma de validação interna que evite o aprendizado e a propagação de erros. Este ponto um dos mais críticos da metodologia. Em outras palavras, considere que o RTWP tenha detectado que o padrão textual "***moro em X***" seja muito adequado para se identificar cidades. Assim, ao encontrar uma frase do tipo "moro em Portugal" pode ser uma evidência de que Portugal é uma cidade. Para evitar tais equívocos, o RTWP possui um conjunto de componentes de aprendizado e um conjunto de temas a aprender. Desta forma, um dado conhecimento só será considerado verdadeiro (ou aprendido) caso haja evidência suficiente da veracidade de tal conhecimento e pouca (ou nenhuma) evidência contrária a este conhecimento.

O ponto crítico citado é chamado desvio de conceito, o qual é o foco principal deste trabalho. O comportamento do desvio de conceito foi visto com o uso da cobertura⁵ e precisão⁶, conforme definidos neste trabalho.

Como já mencionado anteriormente, o RTWP realiza um processo inspirado no "*Mutual Bootstrapping*" [44] com base no Co-Training [1] em que é dada uma base de dados inicial e, a partir disso, novos dados são adquiridos e intercalados (como definido no Bootstrapping). A extração é realizada para ENs e para Pares de ENs (representando Relações Semânticas). O processo busca ENs e PTs de forma intercalada, ou seja, primeiramente são aprendidas ENs em seguida são aprendidos

⁵ Definição na página 48

⁶ Definição na página 48

PTs, depois são aprendidas novas ENs, e depois novamente novos PTs, e assim por diante.

Foi usado o *Yahoo Boss* para a recuperação de páginas Web em que havia a ocorrência de sementes. Suponha que um dos PTs utilizados para a extração de ENs para a categoria Cidade seja “prefeitura municipal de X”. Então o *Yahoo Boss* utiliza este PT como semente para extração de possíveis ENs de Cidade, ou seja, é feita uma busca na Web fornecendo-se a query: “prefeitura municipal de” e as páginas que contém esta frase são armazenadas para serem processadas pelo RTWP. O *Yahoo Boss* é uma API gratuita acessível a qualquer usuário; basta acessar o site <http://developer.yahoo.com/search/boss/>, criar uma conta e solicitar o ID, que é liberado automaticamente e no mesmo instante. Possui um número de acesso máximo por IP e por ID, além da quantidade máxima de 500 páginas por query.

A partir da página em XML extraída do *Yahoo Boss* são feitas filtragens (Stop Words, por exemplo), e depois, são extraídas as possíveis ENs e PTs.

É usado padrão 5-gramas para identificar ENs no XML, isso significa que as ENs terão no máximo 5 palavras, já os PTs não possuem tamanho definido, porém é bom lembrar que também passam por filtros, como por exemplo: um PT não pode conter ponto final.

No aprendizado de Pares de ENs (em Relações Semânticas Semânticas) não há um número de gramas máximo, somente é executado um filtro de pontuação, como por exemplo um contexto de relação não define um relacionamento se possui um ponto final.

Os filtros citados são usados para marcar os pontos de parada na contagem de gramas. Estes filtros são de *Stop Words*, que incluem além de palavras

irrelevantes, pontuação, números e alguns caracteres. *Stop Words* ou *Stop List*, neste trabalho é definido como é uma lista de palavras e caracteres considerados irrelevantes que foi usada para o aprendizado apresentado neste projeto. Irrelevantes são Strings que não são ENs ou PTs e ocorrem muito em uma página web. A lista inicial foi fornecida pelo NILC (<http://www.nilc.icmc.usp.br>), porém passou por uma adaptação que gerou a adição de palavras e caracteres.

Depois dos filtros serem executados, são procuradas as ENs/PTs, que são contadas de uma a cinco palavras antes e depois de cada marcador (contador de final ou ponto final de uma provável EN ou PT), e são formadas todas as combinações possíveis sem se alterar a sequência das palavras.

Para o aprendizado de Relações Semânticas somente é usado o filtro de pontuação.

Existe também um filtro para ENs de categorias⁷ que são nomes próprios, países por exemplo. Este filtro trabalha da seguinte forma: A possível EN é extraída e em seguida os filtros analisam a String da direita para a esquerda e marcam com \$ até que seja reconhecida uma palavra que a primeira letra seja maiúscula, depois a tarefa é refeita da esquerda para a direita.

3.3 Base de Conhecimento

A base de conhecimento usada para este projeto foi construída inspirada em ontologias, e a sua estrutura foi definida como em [7], porém não com a mesma organização de exclusividade mútua. Em [7] todas as categorias escolhidas e que se

⁷ Definição na página 48

referem à pessoas (ex: Pessoa, Ator, Atleta, Cientista, etc.) não são mutuamente exclusivas. Já no RTWP, tais categorias foram dadas como mutuamente exclusivas, exceto à categoria Pessoa.

Duas categorias (A e B) são ditas mutuamente exclusivas quando as instâncias de A são padrões negativos da categoria B. Por exemplo, como uma Equipe esportiva não é um Esporte, as categorias “Equipe Esportiva” e “Esporte” são mutuamente exclusivas.

A Figura 3.1 ilustra a base de conhecimento usada para o aprendizado de categorias, que será visto na Seção 3.5.

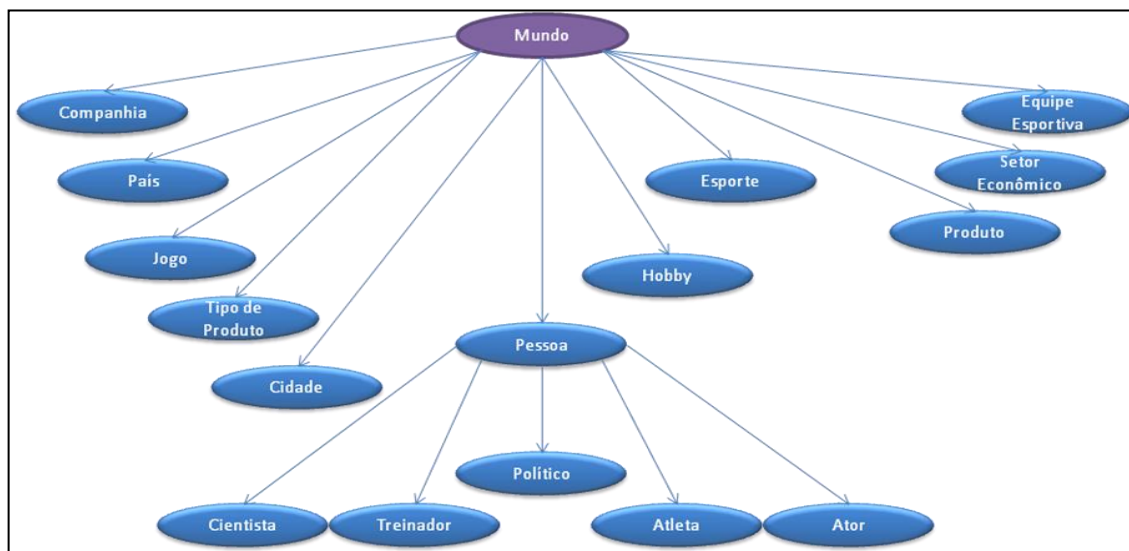


Figura 3.1 Base de Conhecimento

Como pode ser visto na Figura 3.1, neste projeto foram usadas 16 categorias, sendo elas: Cidade, Ator, Atleta, Jogo (tabuleiro), Treinador, Companhia, País, Setor Econômico, Hobby, Pessoa, Político, Produto, Tipos de Produto, Cientista, Esporte e Equipe Esportista. A estrutura da ontologia permite explorar a ideia da relação de exclusividade mútua. Além disso, a estrutura hierárquica da ontologia permite ainda, a representação de subcategorias. Se uma categoria C é uma subcategoria de D,

então as instâncias de C são um subconjunto das instâncias de D. Esta ideia de subcategorias refere-se à categorias que estão inclusas em uma “super categoria”. As categorias Cientista, Treinador, Político, Atleta e Ator são subcategorias de Pessoa, pois a última engloba todas as anteriores.

As categorias presentes na ontologia inicial, Cientista, Treinador, Político, Atleta e Ator, só não são mutuamente exclusivas somente com a categoria Pessoa; o que faz gerar um problema que será citado no Capítulo 4. Categorias Esporte e Hobby não são mutuamente exclusivas entre elas. As demais são mutuamente exclusivas entre si.

A estrutura ontológica que define Relações Semânticas é ilustrada na Figura 3.2.

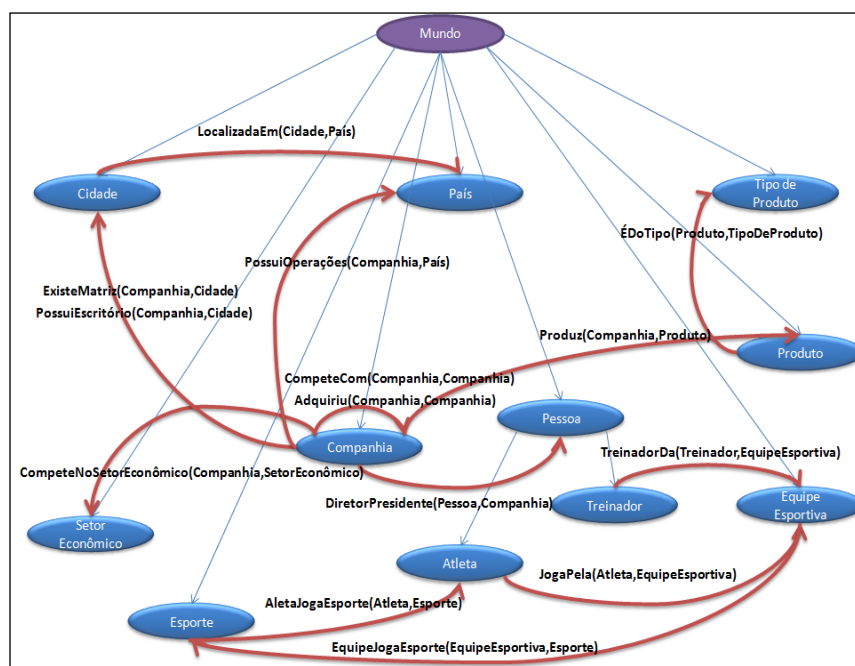


Figura 3.2 Base de Conhecimento com Relações Semânticas

No aprendizado de relações semânticas são utilizadas 14 relações como pode ser visto na Figura 3.2: `CompeteCom(Companhia, Companhia)`, `Adquiriu(Companhia, Companhia)`, `DiretorPresidente(Pessoa, Companhia)`,

TreinadorDa(Treinador, EquipeEsportiva), CompeteNoSetorEconômico(Companhia, SetorEconômico), PossuiEscritório(Companhia, Cidade), PossuiOperações(Companhia, País), ExisteMatriz(Companhia, Cidade), LocalizadoEm(Cidade, País), JogaPela(A atleta, EquipeEsportiva), AtletaJogaEsporte(A atleta, Esporte), EquipeJogaEsporte(EquipeEsportiva, Esporte) e Produz(Companhia, Produto), ÉDoTipo(Produto, TipoDeProduto). Com estas relações semânticas as mutuamente exclusivas são aquelas que não possuem os mesmos primeiro e segundo parâmetro, como por exemplo: Produz(Companhia, Produto) com TreinadorDa(Treinador, EquipeEsportiva). Já aquelas que não são mutuamente exclusivas os mesmos parâmetros, como: CompeteCom(Companhia, Companhia) e Adquiriu(Companhia, Companhia).

3.4 Padrões Negativos

Os padrões negativos [35] [43], foram usados para ampliar a capacidade do sistema de identificar erros no processo de aprendizado de ENs e PTs. Com isso melhorar a aprendizagem em cobertura e precisão. Os resultados do aprendizado utilizando padrões positivos e padrões negativos são acoplados para aumentar a precisão do sistema.

A definição usada neste trabalho pode ser vista abaixo:

Definição 14 Padrões Negativos: são todos os padrões de categorias que são mutuamente exclusivas à categoria que está sendo aprendida no momento.

Definição 15 Padrões Positivos: são os padrões promovidos de categorias que não são mutuamente exclusivas

Se, por exemplo, a categoria Objeto é mutuamente exclusiva com a categoria Cidade, logo, padrões promovidos de Cidade são usados como padrões negativos para Objeto. Ex: “Lápis de cor é uma cidade”. Quanto maior a ocorrência negativa, menor a confiabilidade do conhecimento extraído, a não ser em casos que a ocorrência positiva e negativa são altas, como por exemplo: “São Paulo é um time” e “São Paulo é uma cidade”.

3.5 Identificação e Extração de ENs e PTs para categorias

Neste projeto, “*Extração de ENs e PTs para categorias*” é entendida a extração de ENs e PTs de forma simples, sem relação entre ENs. Suponha as cidades: Miami, Brasília, São Carlos; e os países: Brasil, Colômbia, USA; Suponha também que não se sabe que as cidades citadas do Brasil e USA. Assim, a extração é realizada com base em predicados unários: Cidade(Miami), País(Brasil), etc.

Quando as ENs são nomes próprios, são filtradas por métodos que procuram características de nomes (Letra maiúscula no início); Já as ENs que não são nomes próprios (instâncias de *Hobby*, por exemplo) e os PTs, são analisados quanto a maior ocorrência e maior quantidade de strings.

De forma geral, a extração é executada como na Figura 3.3, em que na primeira iteração são extraídas ENs e PTs somente a partir dos padrões previamente definidos.

Definição 16 Padrões Previamente Definidos: são todos os padrões inseridos previamente de forma manual no sistema para o início da aprendizagem

Já na segunda iteração e nas seguintes são utilizados os dados que foram aprendidos ao longo das iterações anteriores. Em outras palavras, a partir da segunda iteração, os padrões aprendidos anteriormente são somados aos aprendidos na iteração corrente, como mostrado na Figura 3.3.

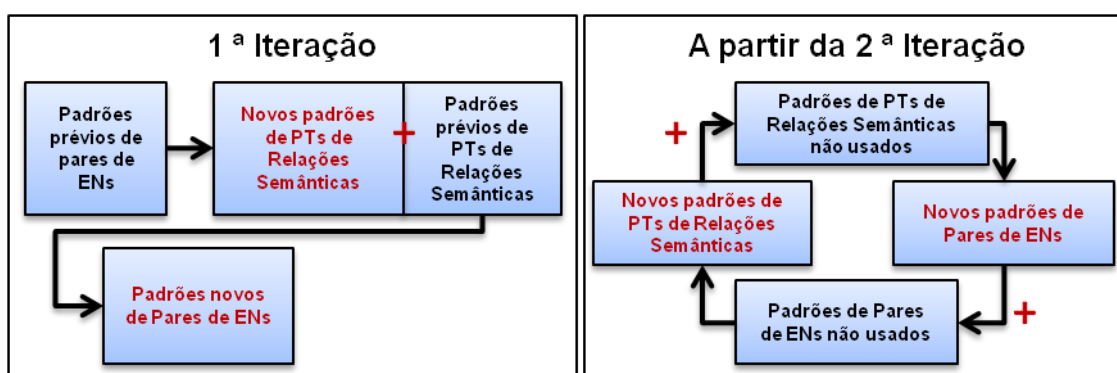


Figura 3.3 Extração de ENs e PTs

A Figura 3.3. mostra que na primeira iteração somente são usados padrões previamente⁸ fornecidos manualmente ao RTWP; A partir da segunda iteração, o resultado do aprendizado passa a ser utilizado para realimentar o processo (conhecimento passa a se acumular). Os padrões não usados são as ENs ou PTs que foram promovidos porém ainda não foram usados para extrair conhecimento.

Um exemplo de funcionamento da extração pode ser vista de forma exemplificada na Figura 3.4.

⁸ Definição na página 65

Em vermelho estão os padrões previamente inseridos, em preto o que foi aprendido corretamente e em rosa o que também foi aprendido, porém são incorretos.



Figura 3.4 Extração de ENs e PTs – Exemplo

A Figura 3.4 ilustra o funcionamento do aprendizado da seguinte forma:

- 1º Passo: São fornecidos manualmente alguns padrões de cidade e de PTs (em vermelho).
- 2º Passo: A partir dos padrões prévios de ENs são identificados novos padrões para PTs, e com os padrões prévios de PTs são identificados novos padrões para ENs,
- 3º Passo: A partir dos padrões de ENs que foram aprendidos e promovidos, novos padrões de PTs são extraídos.
- 4º Passo: A partir dos padrões de PTs que foram aprendidos e promovidos, novos padrões de ENs são extraídos.

5º Os passos 3 e 4 são repetidos até se atingir o critério de parada escolhido.

3.5.1 Padrões Negativos na Promoção de ENs e PTs

Para se construir as queries (que serão utilizadas para a recuperação de páginas Web) são unidas ENs promovidas, de categorias mutuamente exclusivas, aos PTs candidatos⁹ (Ex: “Google é uma cidade perto da”). Da mesma forma, PTs promovidos, de categorias mutuamente exclusivas, são unidos à ENs candidatas (Ex: “Piracicaba é uma companhia”). A Tabela 3.1 mostra um exemplo, em que são apresentadas as categorias cidade(X) e pessoa(Y), com isso os padrões negativos para cidade serão os padrões positivos de pessoa e vice-versa.

Considere a Tabela 3.1, em que na parte superior estão apresentados os padrões positivos e negativos¹⁰ para cada categoria (Cidade e Pessoa).

Tabela 3.1 Padrões Positivos e Negativos

Padrões Positivos		Padrões Negativos	
Cidade (X)	Pessoa(Y)	Cidade (X)	Pessoa(Y)
EN's		EN's	
São Paulo	Carolina Ferraz	Carolina Ferraz	São Paulo
São Carlos	Jô Soares	Jô Soares	São Carlos
Nova York	Raul Gil	Raul Gil	Nova York
PT's		PT's	
X é uma cidade	Y é uma atriz famosa	Y é uma atriz famosa	X é uma cidade
X é a cidade do desenvolvimento	Y é uma pessoa conhecida	Y é uma pessoa conhecida	X é a cidade do desenvolvimento
X é uma famosa cidade	Y é um apresentador de tv	Y é um apresentador de tv	X é uma famosa cidade

⁹ Definição na página 58

¹⁰ Definições na página 64

A formação das queries é realizada como na Tabela 3.2 para cidade e na Tabela 3.2, “**Y é uma atriz famosa**” é padrão negativo para cidade, o que gera: “**São Paulo é uma atriz famosa**”, “**São Carlos é uma atriz famosa**”, etc.

Tabela 3.2 Padrões Negativos de Cidade

São Paulo é uma atriz famosa	São Carlos é uma atriz famosa	Nova York é uma atriz famosa
São Paulo é uma pessoa conhecida	São Paulo é uma pessoa conhecida	Nova York é uma pessoa conhecida
São Paulo é um apresentador de tv	São Paulo é um apresentador de tv	Nova York é um apresentador de tv

Já na Tabela 3.3, “**X é uma cidade**” é dado como padrão negativo para pessoa, ficando então: “**Carolina Ferrraz é uma cidade**”, “**Jô Soares é uma cidade**”, etc.

Tabela 3.3 Padrões Negativos de Pessoa

Carolina Ferrraz é uma cidade	Jô Soares é uma cidade	Raul Gil é uma cidade
Carolina Ferrraz é a cidade do desenvolvimento	Jô Soares é a cidade do desenvolvimento	Raul Gil é a cidade do desenvolvimento
Carolina Ferrraz é uma famosa cidade	Jô Soares é uma famosa cidade	Raul Gil é uma famosa cidade

Depois da extração a partir das queries criadas são contadas as ocorrências e co-ocorrências para cada EN ou Contexto candidato a ser promovido. Com isso é calculado o *Score Negativo* dado na equação (1).

3.5.2 Promoção de ENs e PTs para categorias

A promoção de ENs e PTs é realizada de acordo com o peso, o qual é calculado a partir do *score* positivo e negativo, que considera o número de ocorrências positivas e negativas e co-ocorrências positivas e negativas de sementes.

Os cálculos citados anteriormente estão definidos nas expressões (1), (2) e (3).

$$\text{Score}_{\text{Negativo}} = \log_{10}(\text{NroOcorrênciaNegativa}^{\text{NroCoOcorrênciaNegativa}}) \quad (1)$$

$$\text{Score}_{\text{Positivo}} = \log_{10}(\text{NroOcorrênciaPositiva}^{\text{NroCoOcorrênciaPositiva}}) \quad (2)$$

$$\text{Weight}_{\text{EN/PT}} = \text{Score}_{\text{Positivo}} - \text{Score}_{\text{Negativo}} \quad (3)$$

O número de ocorrências positivas é simplesmente a soma de vezes que uma EN ou um PT ocorre. Por exemplo, se São Paulo é extraído de “X é uma cidade” com número de ocorrência igual a 10 e “X é uma bela cidade” igual a 5, o número de ocorrência da EN (São Paulo) é igual a 15 e o número de co-ocorrência com diferentes PTs é igual a 2 (2 PTs diferentes extraíram a EN). O mesmo vale para padrões negativos.

3.5.3 Processo de Aprendizado

O aprendizado é organizado em duas partes principais: aprendizado de ENs e aprendizado de PTs. O funcionamento ocorre de acordo com ideia do “*Mutual Bootstrapping*”[44], ou seja, eles acontecem de forma intercalada. O processo de aprendizado de categorias segue o esquema da Figura 3.5.

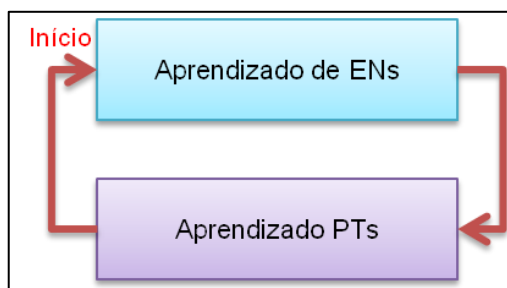


Figura 3.5 Aprendizado de ENs e PTs para categorias

De acordo com a Figura 3.5, são aprendidas ENs a partir de PTs e com estes são aprendidas novas ENs, e isso segue. De uma forma mais detalhada, primeiramente são selecionados os PTs promovidos e ainda não usados. A partir destes PTs são extraídas novas ENs. Para cada EN extraída é calculado o Score a partir da quantidade de ocorrência e co-ocorrência com PTs para cada EN, com isso é gerada uma lista com até 1000 ENs com maiores chances de serem promovidas (por categoria). Esta lista é submetida ao acoplamento de padrões negativos. Neste acoplamento executada a contagem de ocorrência e co-ocorrência nos padrões negativos, os quais são padrões positivos de categorias mutuamente exclusivas para a categoria que é analisada.

Durante a execução do acoplamento de padrões negativos são executados os cálculos utilizando-se as equações (1), (2) e (3). Em seguida é reordenada a lista de ENs com maiores scores e somente 1/3 delas (por categoria) é promovido.

Para o aprendizado de PTs de Relações Semânticas o ciclo é o mesmo, em que são selecionados para extração os Pares de ENs promovidos ainda não usados, e aprendidos novos padrões de PTs de Relações Semânticas. Durante o aprendizado já é calculado o Score de cada contexto e depois é gerada a lista de até 1000 com os maiores scores. Em seguida é executado o acoplamento de padrões negativos e também os cálculos com base em (1), (2) e (3). Depois dos cálculos é reordenada a lista e são promovidos 3 ou 10 dos PTs (dependendo da especificação do usuário).

O número máximo selecionado de ENs e PTs candidatos para a promoção é de 1000 para o aprendizado de categorias e de 500 para o de PTs de Relações

Semânticas. Possuir esse número máximo na lista de candidatos funciona como um corte no ranking de candidatos.

No método de promoção a quantidade de ENs, PTs e PTs de Relações Semânticas promovidos também varia. Para ENs em todos os métodos implementados é promovido 1/3 do ranking. Já PTs variam em 3 ou 10 o número de promoções, e para PTs de Relações Semânticas somente é promovido 1/3.

Depois das tarefas anteriores serem executadas elas são re-executadas até se alcançar o critério de parada.

3.6 Identificação e Extração de Pares de ENs e PTs de Relações Semânticas

Como visto anteriormente, a extração de ENs e PTs para categorias é realizada com base em predicados unários: Cidade(Miami), País(Brasil), etc. em que Miami é uma cidade e Brasil é um país. Já na extração para Relações Semânticas (descrita nesta seção) tem-se com base predicados binários, como por exemplo: LocalizadaEm(Brasília, Brasil). Um exemplo pode ser visto na Figura 3.6.



Figura 3.6 Extração de Pares de ENs e PTs de Relações Semânticas - Exemplo

Com os predicados binários extrai-se um ou mais Pares de ENs para cada PT de Relação Semântica promovido. Os PTs de Relações Semânticas por sua vez, relacionam as duas ENs envolvidas no processo de extração.

A extração de Pares de ENs e PTs de Relações Semânticas é realizada da mesma forma descrita na seção anterior (3.5), porém os padrões agora contam com a relação, assim como pode ser visto na Figura 3.6.

De forma geral, RTWP executa o algoritmo de aprendizado para Pares de ENs e PTs de Relações Semânticas como na Figura 3.6, em que, na primeira iteração, são usados os Pares de ENs manualmente fornecidos para extrair novos PTs de Relações Semânticas. Os novos PTs são somados aos já existentes, com isso novos Pares de ENs são extraídos. Já a partir da segunda iteração são usados os padrões tanto de PTs de Relações Semânticas quanto de Pares de ENs que ainda não foram usados antes, para serem extraídos novos padrões de ambos.

O algoritmo executa de forma intercalada seguindo a ideia do “*Mutual Bootstrapping*”, assim como a extração de ENs e PTs para categoria (seção 3.5), em que para aprender ENs são usados PTs e vice-versa.

Diferentemente do aprendizado de ENs e PTs para categorias, nesta extração a primeira iteração não se difere das seguintes, como pode ser visto na Figura 3.7.

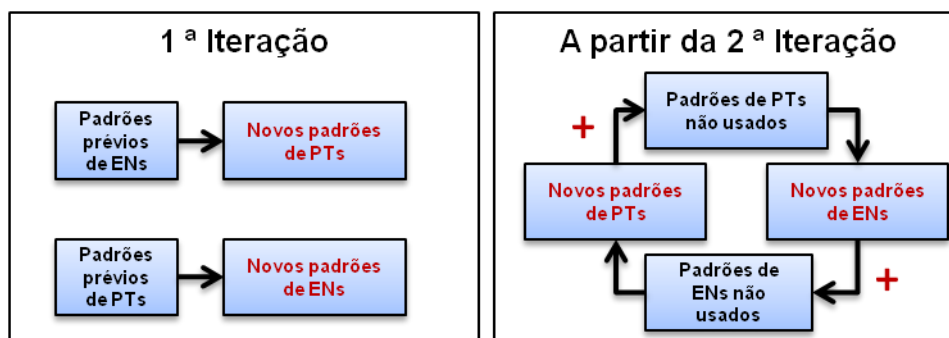


Figura 3.7 Extração de Pares de ENs e PTs de Relações Semânticas

Isso foi assim definido (Figura 3.7) porque as Relações Semânticas são mais difíceis de serem extraídas, pois a ocorrência em páginas Web é mais baixa. Na primeira iteração, os PTs de Relações Semânticas previamente definidos já foram somados aos aprendidos pelos Pares de ENs para que houvesse mais padrões disponíveis para a próxima extração.

3.6.1 Padrões Negativos para Promoção Pares ENs e PTs de Relações Semânticas

Os padrões negativos aqui também seguem a mesma forma citada na Seção 3.5.1; a única diferença é a utilização de um Par de ENs ao invés de uma única ENs por instância. Por exemplo, para a promoção de um Par de EN, seja ele: **“São Carlos & Brasil”**, em que se tem a relação de localização, sendo ela **“LocalizadoEm(X,W)”**, em que X é uma cidade e W é um país. Com este Par (X e W) poderia ser dado como padrão negativo a relação **“FabricaDe(A,B)”**, em que A é uma fábrica e B é um produto. Com isso, o padrão para esta relação, *“é uma fábrica de”* é usada de padrão negativo como *“São Carlos é uma fábrica de Brasil”*.

3.6.2 Promoção de Pares de ENs e PTs de Relações Semânticas

É executada da mesma forma que a usada para as categorias na seção 3.5.2 com a diferença que os cálculos são executados para Pares de ENs e PTs de Relações Semânticas.

3.6.3 Processo de Aprendizado

O aprendizado tratado nesta seção funciona da mesma forma que o já apresentado na seção 3.6, a diferença é que a extração é a partir de Pares de ENs e PTs que estão entre as ENs.

A Figura 3.7 apresenta o funcionamento do aprendizado de Pares de ENs e PTs de Relações Semânticas:

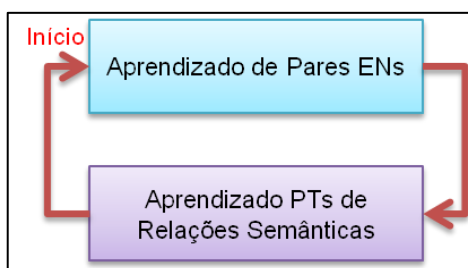


Figura 3.8 Aprendizado de Pares ENs e PTs de Relações Semânticas

O funcionamento dado na Figura 3.8 ocorre da seguinte forma: são aprendidos Pares de ENs a partir de PTs de Relações Semânticas e com estes são aprendidos novos Pares de ENs, e isso segue. De uma forma mais detalhada, primeiramente executa-se o aprendizado de Pares de ENs e em seguida o aprendizado de PTs de Relações Semânticas. Para o aprendizado de Pares de ENs são selecionados os PTs de Relações Semânticas ainda não usados. A partir destes são extraídos Pares de ENs, e nesta mesma tarefa já é calculado o Score para cada Par. Em seguida é gerada a lista com até 500 candidatos à promoção, e esta é submetida ao acoplamento com padrões negativos. Nesta mesma tarefa é calculado o peso (equações (1), (2) e (3)) para cada item. Após os cálculos a lista é reordenada a partir do peso e, é executada a promoção de até 1/3 dos candidatos.

Para o aprendizado de PTs de Relações Semânticas são selecionados os Pares de ENs promovidos (ou os padrões previamente definidos caso esteja na 1ª

iteração), com estes são extraídos novos PTs de Relações Semânticas. Na tarefa de extração já é executado o cálculo do Score para cada PT. Em seguida é gerada uma lista com os 500 PTs candidatos, por categoria. Os componentes desta lista são submetidos ao acoplamento de padrões negativos, e em tempo de execução também é calculado o peso (equações (1), (2) e (3)) de cada um. Após a tarefa anterior a lista é reordenada e ocorre a promoção do primeiro 1/3.

Depois das tarefas anteriores serem executadas elas são re-executadas até se alcançar o critério de parada.

3.7 Identificação e Extração de Pares de ENs com Tipagem e PTs de Relações Semânticas

Esta versão do RTWP é um acoplamento do aprendizado de Pares de ENs e PTs de Relações Semânticas com o aprendizado de ENs de categorias (*tipagem [23]*), que é executada antes do método de promoção.

Tipagem é a confirmação de cada um dos argumentos do predicado binário antes que seja realizada a promoção dos Pares de ENs. Na tipagem verifica-se se ambos os argumentos são válidos para o predicado. Por exemplo, considere o seguinte predicado *LocalizadaEm*("Cidade","País"), em que "Cidade" e "País" definem o "tipo" do primeiro e do segundo argumento respectivamente, ou seja, definem à qual categoria os argumentos devem pertencer. Considere também que se tenha o Par de ENs "São Carlos" e "Brasil" (*LocalizadaEm*(São Carlos, Brazil)) como candidato a ser promovido. Neste exemplo, antes de se promover o Par de ENs, é primeiro analisado se *São Carlos* realmente é uma **cidade** e se *Brasil*

realmente é um **País**. Esta confirmação é executada utilizando os métodos de aprendizado ENs para categorias (referente à seção 3.5). Um exemplo é dado com as tabelas Figura 3.4, Figura 3.5 e Figura 3.7.

Tabela 3.4 PTs para categoria

PTs para País	PTs de Cidade
é um país	é uma cidade
é um país subdesenvolvido	é uma pequena cidade
é um belo país	é uma bela cidade

Tabela 3.5 Par de EN

Par de EN	
País	Cidade
Brasil	São Carlos

Tabela 3.6 Exemplo de Tipagem

PTs com Tipagem para País	PTs com Tipagem de Cidade
Brasil é um país	São Carlos é uma cidade
Brasil é um país subdesenvolvido	São Carlos é uma pequena cidade
Brasil é um belo país	São Carlos é uma bela cidade

A Tabela 3.4 possui os PTs utilizados para a extração de ENs (para categoria), na Tabela 3.5 tem-se o Par de EN a ser confirmado. Para que aconteça este teste são formadas queries como as linhas da Tabela 3.6. Em seguida tais linhas são buscadas pelo Yahoo Boss. As páginas extraídas passam pelo mesmo processo já descrito anteriormente, ou seja, são contadas quantas vezes ocorreram “**Brasil** é um país”, “**Brasil** é um país subdesenvolvido”, “**São Carlos** é uma cidade”, etc. Esta contagem é realizada também separa número de co-ocorrência e ocorrência.

Ocorrência Positiva: número de vezes que **São Carlos** ocorreu como **Cidade** em todas as queries;

Co-ocorrência Positiva: Número de PTs que confirmam que **São Carlos** é uma cidade:

Ocorrência Negativa: número de vezes que **São Carlos** ocorreu como qualquer outra classe mutuamente exclusiva em todas as queries.

Co-ocorrência Negativa: Número de PTs que dizem que **São Carlos** é de qualquer outra classe mutuamente exclusiva (objeto, por exemplo).

As ENs de um Par não são promovidas separadamente, ou seja, o peso é calculado pelo Par.

3.7.1 Promoção de Pares de ENs e PTs de Relações Semânticas com Tipagem

Segue a mesma forma já citada anteriormente, porém com um score adicionado, o chamado *ScoreTyping*, o qual é calculado com as expressões (4), (5) e (6).

$$\begin{aligned} & \textit{ScoreTyping}_{\textit{Negativo}} \\ & = \log_{10}(\textit{NroOcorrênciaNegativaTyping}^{\textit{NroCoOcorrênciaNegativaTyping}}) \end{aligned} \quad (4)$$

$$\begin{aligned} & \textit{ScoreTyping}_{\textit{Positivo}} \\ & = \log_{10}(\textit{NroOcorrênciaPositivaTyping}^{\textit{NroCoOcorrênciaPositivaTyping}}) \end{aligned} \quad (5)$$

$$\textit{ScoreTyping} = \textit{ScoreTyping}_{\textit{Positivo}} - \textit{ScoreTyping}_{\textit{Negativo}} \quad (6)$$

O *ScoreTyping* é usado no cálculo do peso da seguinte forma:

Se (*ScoreTyping* > 0) então está apto a ser promovido, senão não está apto a ser promovido. Se está apto a ser promovido, verifique o peso.

O ScoreTyping somente diz se pode ser promovido ou não, porém a lista de promoção, ou seja a ordem de promoção, depende do peso (definido na expressão (3).

3.7.2 Processo de Aprendizado

O processo de aprendizado com tipagem é executado da mesma forma já citada nas sessões 3.5 e 3.6, a diferença é a adição de mais uma forma de acoplamento, a tipagem. A tipagem, como já falado no início desta seção, é mais uma forma de confirmação do conhecimento, ou seja, é um valor a mais para que um Par de EN seja avaliado. A tipagem só é usada para o aprendizado de Pares de ENs, não para o aprendizado de PTs de Relações Semânticas, devido a sua particularidade de testar as duas ENs do Par .

A Figura 3.9 ilustra como é executado o aprendizado de Pares de ENs e PTs de Relações Semânticas com tipagem.

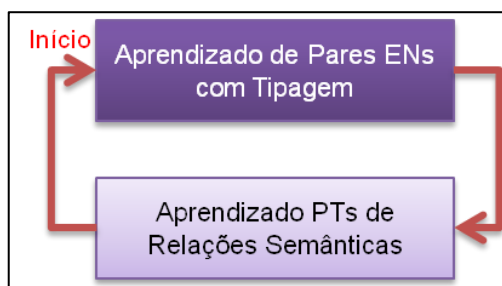


Figura 3.9 Aprendizado de Pares de ENs e PTs de Relações Semânticas com Tipagem

De acordo com a Figura 3.9, de forma geral é executado o aprendizado de Pares de ENs a partir de PTs de Relações Semânticas aprendidas, antes da promoção os Pares passam pelo método de tipagem. Depois da promoção é executado o aprendizado de PTs de Relações Semânticas a partir dos Pares de ENs, e assim segue. Mais detalhadamente, primeiramente são selecionados os PTs

de Relações Semânticas já promovidos (se está na primeira iteração, estes são os previamente definidos), em seguida são extraídos, ou seja, novos Pares de ENs, os quais já, em tempo de execução, têm seu Score calculado. Depois é gerada uma lista com até 500 candidatos a partir do Score e esta lista é submetida ao acoplamento de padrões negativos, no qual em tempo de execução é calculado o peso (expressões (1), (2) e (3)). Após a finalização e antes da promoção é reordenada a lista e submetida ao acoplamento de tipagem, no qual é calculado o peso de tipagem (expressões (4), (5) e (6)). A lista é reordenada e é executada a promoção. Neste caso, foram executados dois testes, em um há um filtro de 1/3 dos melhores Scores antes de ser realizada a tipagem, e no outro todos que possuem valores de Scores são submetidos à tipagem. Em ambos casos, após a tipagem é executada a promoção de Pares de ENs.

Após a primeira parte da tarefa ser concluída inicia-se o aprendizado de PTs de Relações Semânticas, em que são selecionados todos os Pares de ENs promovidos e não usados anteriormente e é executada a extração, na qual já é realizada o cálculo de Score. Depois é gerada uma lista com até 500 PTs de Relações Semânticas a partir do Score calculado. A lista é submetida ao acoplamento de padrões negativos e em tempo de execução são já realizados os cálculos de weight (expressões (1), (2) e (3)). Em seguida é realizada a promoção em que 1/3 dos candidatos são promovidos.

Após finalizar o ciclo, ele é reiniciado e somente para quando chega no critério de parada.

3.8 Funcionamento de Todos os Acoplamentos: Aprendizado de ENs, Pares de ENs, PTs de Categorias e PTs de Relações Semânticas com Tipagem

Pode-se definir três formas principais de acoplamentos neste projeto: Aprendizado para categorias (versão da seção 3.5) que acopla aprendizado de ENs e PTs, 3.6 aprendizado para Relações Semânticas (versão da seção 3.7), que acopla aprendizado de Pares de ENs e PTs de predicados binário, e finalmente a tipagem (versão em seção da 3.6) que acopla aprendizado de Relações Semânticas e de categorias. Finalmente, há também um acoplamento que faz com que os Pares de ENs aprendidos para as Relações Semânticas sejam promovidos como novas instâncias de categorias.

Para o funcionamento na prática de todos os acoplamentos propostos neste projeto, será executada a primeira iteração completa do aprendizado para categorias, depois que encerrada, é iniciada a primeira iteração do aprendizado para Relações Semânticas com tipagem (estas duas iterações poderiam ser executadas em paralelo). Esta última ao encerrar faz com que inicie-se a segunda iteração para categorias, que ao terminar causa o início da segunda iteração para Relações Semânticas e assim por diante até se alcançar o critério de parada.

A Figura 3.10 ilustra o funcionamento completo do RTWP.

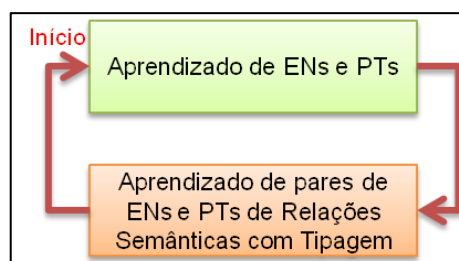


Figura 3.10 RTWP com todos acoplamentos

Primeiramente é executada a versão conforme a seção 3.5 e em seguida é executada a versão da seção 3.7 de forma intercalada.

A Figura 3.10 também mostra que ambos aprendizados trabalham de forma intercalada, em que o controle de iteração é compartilhado e a base de dados também. Ao encerrar uma iteração do aprendizado de categorias as ENs e PTs promovidos são salvos na mesma base que os Pares de ENs e PTs de Relações Semânticas. Com este acoplamento de todas as versões, espera-se que a base de conhecimento seja mais robusta.

3.9 Visão Geral do Projeto

O RTWP é um software que implementa a ideia de aprendizado apresentada no NELL, não de forma sem fim, porém como o RTW, conclui que é possível melhorar o aprendizado com acoplamentos de conhecimento. No caso do RTWP ele é voltado à Web em português e o RTW para a em inglês. Assim, mais um acoplamento é esperado com a integração do RTW e do RTWP, como pode ser visto na Figura 3.11.

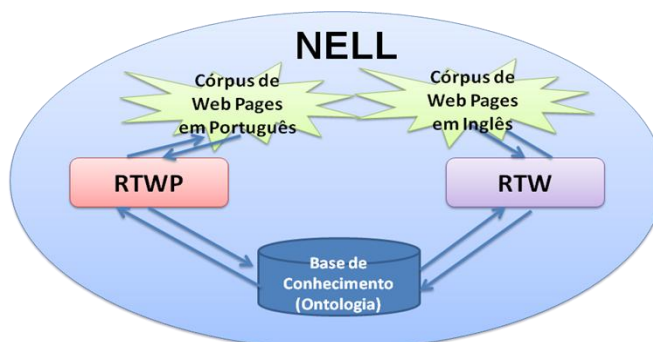


Figura 3.11 Visão Futura do NELL

Tal integração será realizada futuramente. De qualquer forma, para que se tenha uma ideia de como tal acoplamento irá acontecer, a Figura 3.11 mostra a organização do NELL e como as aplicações RTW e RTWP irão interagir.

Como na Figura 3.11, o RTWP e o RTWP fazem parte do projeto NELL e tais aplicações posteriormente serão vinculadas, porém atualmente são como descritas na Figura 3.12.

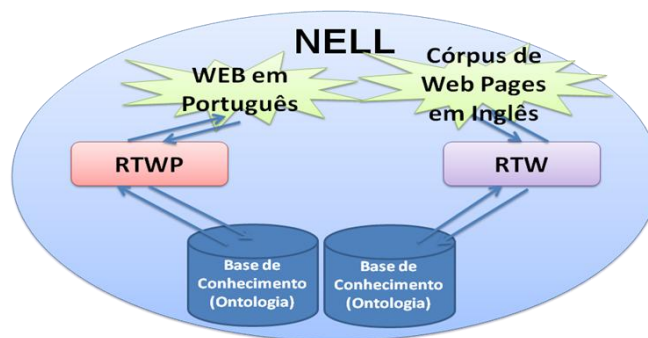


Figura 3.12 Visão Atual do NELL

As diferenças principais entre a Figura 3.11 e a Figura 3.12 são:

- RTWP extrai conhecimento diretamente da Web, usando a API *Boss da Yahoo*.
- RTW extrai conhecimento de um cópus da Web em inglês (com isso, pode-se ter mais resultados na fase de extração de candidatos do que no caso do RTWP).

Futuramente a base de conhecimento do RTWP será vinculada à do RTW. O RTWP, assim como o RTW extrairá conhecimento a partir de um cópus previamente extraído da Web (além de manter seus métodos de extração direta).

Existem características que diferem na implementação do RTW e do RTWP. Para a construção do RTWP, não se teve apoio de buscadores/API como no RTW, o que traria maior número de páginas. Não se contou também com um computador tão potente quanto o usado no RTW (M45 da Yahoo) para preprocessar o cópus de

páginas Web. No RWTP foram usados computadores comuns (vários computadores para execução de mais versões ao mesmo tempo, uma versão em cada computador).

Capítulo 4

EXPERIMENTOS E ANÁLISE

A partir das formas de acoplamentos apresentadas no Capítulo 3: Aprendizado de ENs e PTs com e sem padrões negativos, aprendizado de Pares de ENs e PTs de Relações Semânticas com e sem tipagem, neste Capítulo são mostradas discutidos os resultados de execuções de combinações de diferentes tipos de acoplamentos.

4.1 Experimentos

Os experimentos realizados foram criados de acordo com o Capítulo 3 - , em que foram apresentadas diferentes formas e combinações de acoplamentos. As formas de acoplamentos podem ser resumidas em aprendizagem de categorias, de Relações Semânticas e Relações Semânticas com tipagem.

A aprendizagem de categorias contou com oito formas testadas com diferentes parâmetros que se referiram a: ser ou não cumulativos, acoplamento ou não de padrões negativos e o número máximo de contexto a serem promovidos (3 ou 10). Estes experimentos estão organizados na Tabela 4.1.

As tabelas Tabela 4.1 e Tabela 4.2 possuem células em verde para identificar experimentos com e sem acoplamento de padrões negativos, cumulação e quantidade de PTs promovidos.

Tabela 4.1 Experimentos para o Aprendizado de categorias

Número de Experimento	Acoplamento de Padrões Negativos	Cumulativo	Número de PTs Promovidos
I.			10
II.			3
III.			10
IV.			3
V.			10
VI.			3
VII.			10
VIII.			3

A Tabela 4.1 é descreve características dos experimentos de I a VIII. Nela, a coluna Acoplamento de padrões negativos é referente ao que foi visto na Seção 3.4 e em cada subseção referentes aos diversos tipos de acoplamentos apresentados. Cumulativo refere-se ao acúmulo de conhecimento a cada iteração; caso o experimento não seja cumulativo, os dados aprendidos em uma iteração só poderão ser usados na mesma não sendo transferidos para as próximas. O Número de PTs promovidos refere-se ao parâmetro usado para a promoção de PTs, o qual permite somente a promoção de ou 3 ou 10.

A motivação para se ter experimentos cumulativos e não cumulativos no decorrer das iterações, é avaliar o quanto o resultado do processo de aprendizado pode auxiliar na sequência do processo. Em outras palavras, busca-se verificar se o conhecimento adquirido hoje pode auxiliar a melhorar o aprendizado de amanhã.

Os experimentos acoplados e não acoplados de padrões negativos se referem à execução, ou não, da tarefa de busca de padrões negativos e investigam empiricamente o quanto a inserção de restrições pode ajudar o aprendizado a minimizar o desvio de conceito.

O número de PTs a serem promovidos refere-se ao ajuste de precisão; quanto maior esse parâmetro provavelmente menor a precisão e maior a cobertura. Isso acontece porque a promoção de PTs é mais difícil de ser avaliada. Quanto menos PTs forem promovidos maior será a precisão (considerando uma lista de ranking sem empates, por exemplo), pois serão selecionados os PTs que estão no topo.

A Tabela 4.2 mostra os experimentos de aprendizado de Relações Semânticas, os quais também contam com o acoplamento de categorias (exceto no experimento IX, que é referente somente a Relações Semânticas). Os resultados foram obtidos com a versão de aprendizado de categorias já concluído. A base de conhecimento do aprendizado de categorias já havia sido preenchida e não teve adição de dados durante a execução. Com isso, não foi possível verificar qual o ganho em relação ao aumento do número de ENs aprendidas com ou sem a integração destes dois componentes de aprendizado.

Tabela 4.2 Experimentos para o Aprendizado de categorias e Relações Semânticas

Número de Experimento	Acoplamento de categorias	Acoplamento de Relações Semânticas	Acoplamento de Tipagem	Seleção de 1/3 de ENs Candidatas antes da Tipagem
IX.				
X.				
XI.				

A Tabela 4.2 possui as colunas: acoplamento de categorias, acoplamento de Relações Semânticas, acoplamento de tipagem e seleção de 1/3 de ENs candidatas antes da tipagem. O Acoplamento de categorias engloba todos os parâmetros mencionados na Tabela 4.1, todos são ativos e o número de PTs promovidos é 3,

devido aos resultados previamente obtidos com as categorias e que serão mostrados neste capítulo.

A validação para a geração dos gráficos foi realizada de forma manual e bastante rígida. Foram dadas como erradas ENs incompletas ou com palavras a mais, como por exemplo, para a categoria Produto: nova soja. Para as categorias Cientista, Ator e Atleta não foram dados como certos nomes sem sobrenome (exceto casos que não deixam dúvida como, por exemplo: Neymar).

Os gráficos apresentados no decorrer desta seção possuem o cálculo de cobertura de forma simples, pois somente considera-se o número de ENs promovidas.

Todos os gráficos de resultados por categoria a seguir, ilustram o aprendizado obtido. O eixo Y indica a precisão e o X indica a cobertura em uma iteração. Tanto cobertura quanto precisão são mostradas de forma cumulativa por iteração. O número de iterações inicia em 1 e vai até 5.

Em alguns casos com baixo aprendizado, a definição de uma estrutura ontológica que possua diferentes assuntos e ao mesmo tempo seja específica, pode possibilitar melhores resultados. Quanto a ser específica significa que ela deve possuir categorias mais vinculadas ou parecidas, por exemplo: Município, Cidade, Bairro, Vila, Praia, Estado, País, Ilha e Continente.

Outro fator importante na causa para poucos resultados foi à API usada para a extração de páginas da Web, pois o *Yahoo Boss* possui limite diário de acessos e além disso, são retornadas somente 500 páginas por consulta. Portanto, pelo número de páginas que são permitidas, a cobertura realmente é pequena, pois a base de fonte de dados se torna mais restrita, mesmo ela sendo a internet.

4.2 Experimentos I e II – Não Acoplados e Não Cumulativos 10 e 3

Os experimentos I e II são não acoplados e não cumulativos, pois não possuem nenhum tipo de acoplamento e não acumulam o conhecimento com o passar das iterações. O experimento I é não acoplado e não cumulativo com parâmetro 10 de promoção de PTs, enquanto o II se diferencia somente este parâmetro, sendo então 3.

Não acumular conhecimento impõe que o sistema use apenas conhecimento adquirido na iteração corrente.

Estes dois experimentos, I e II, apresentados nas figuras Figura 4.1 e Figura 4.2, obtiveram resultados ruins devido a não serem cumulativos. Foi obtido pouco conhecimento e os experimentos não avançaram para a próxima iteração, se mantiveram somente na primeira. Isto não é surpresa, pois estes experimentos apenas mostram empiricamente como se comportam de métodos de aprendizado com estas características.

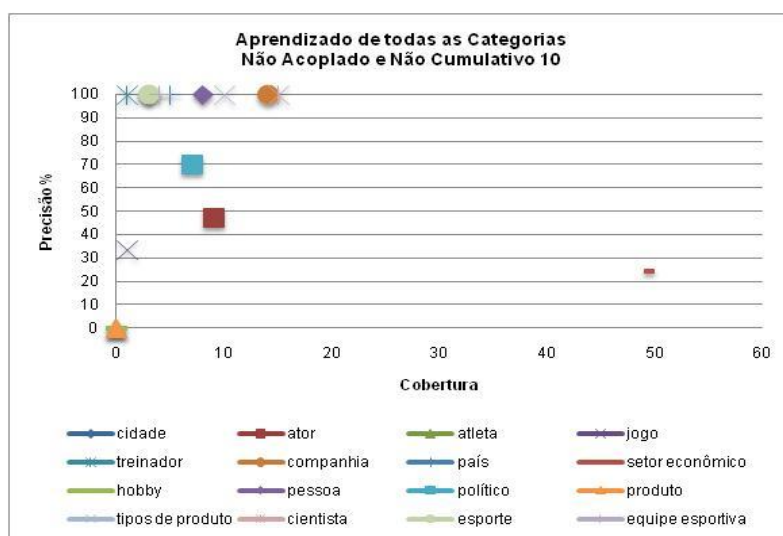


Figura 4.1 Resultados - Experimento Não Acoplado e Não Cumulativo 10

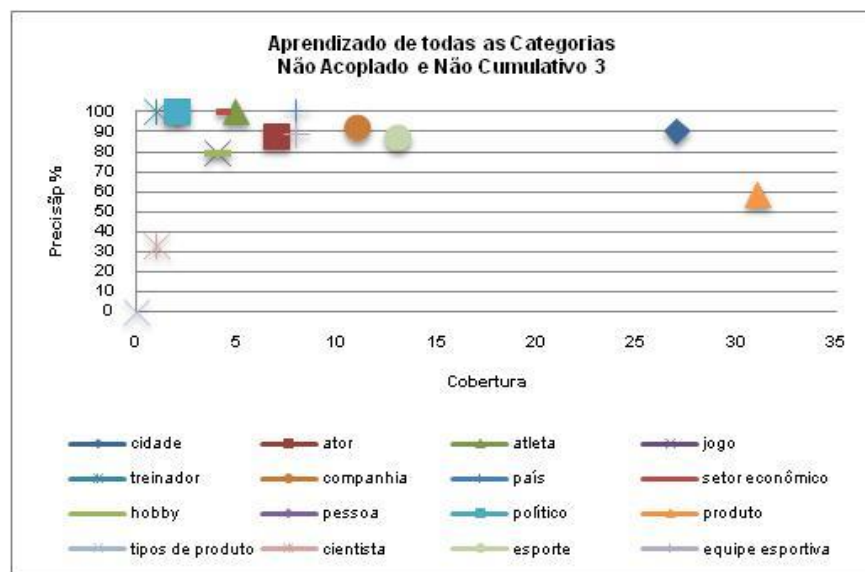


Figura 4.2 Resultados - Experimento Não Acoplado e Não Cumulativo 3

A partir dos resultados obtidos nestes experimentos e dos próximos pode-se observar que tanto no experimento I quanto no II, o uso do aprendizado de forma não cumulativa interferiu negativamente sobre o resultado. Isso deve-se ao fato de não se ter permitido que os dados obtidos na primeira iteração pudessem ser adicionados ao que se aprendeu na segunda. Tais resultados podem ser observados comparando-se estes experimentos e os próximos que são cumulativos, os quais tiveram mais iterações.

Ambos os resultados (Figura 4.2 e Figura 4.1) somente mostram a promoção de conhecimento para a primeira iteração, já na segunda não há dados plausíveis a serem promovidos e com isso o sistema é interrompido.

4.3 Experimentos III e IV – Acoplados e Não Cumulativos 10 e 3

Estes experimentos também não são cumulativos, portanto terão poucos resultados, porém possuem acoplamento de padrões negativos.

Os resultados podem ser vistos na Figura 4.3 e na Figura 4.4, em que os gráficos mostram o conhecimento adquirido em todas as categorias e em todas

iterações. No experimento III foram realizadas 2 iterações enquanto o IV alcançou 5 iterações.

Comparando-se com os resultados obtidos nos experimentos I e II (Subseção 4.2), o acoplamento ajudou no funcionamento do sistema, pois os experimentos acoplados executaram por mais iterações de modo geral e ainda obtiveram mais resultados. Quanto à precisão, o acoplamento obteve o comportamento esperado. Através dos padrões negativos foi possível minimizar o desvio de conceito. Isso pode ser notado analisando a Figura 4.3, em que de modo geral, em categorias que houveram quedas conseguiu se manter por mais tempo maior precisão. As quedas tiveram “passos” curtos, ou seja, não caíram drasticamente.

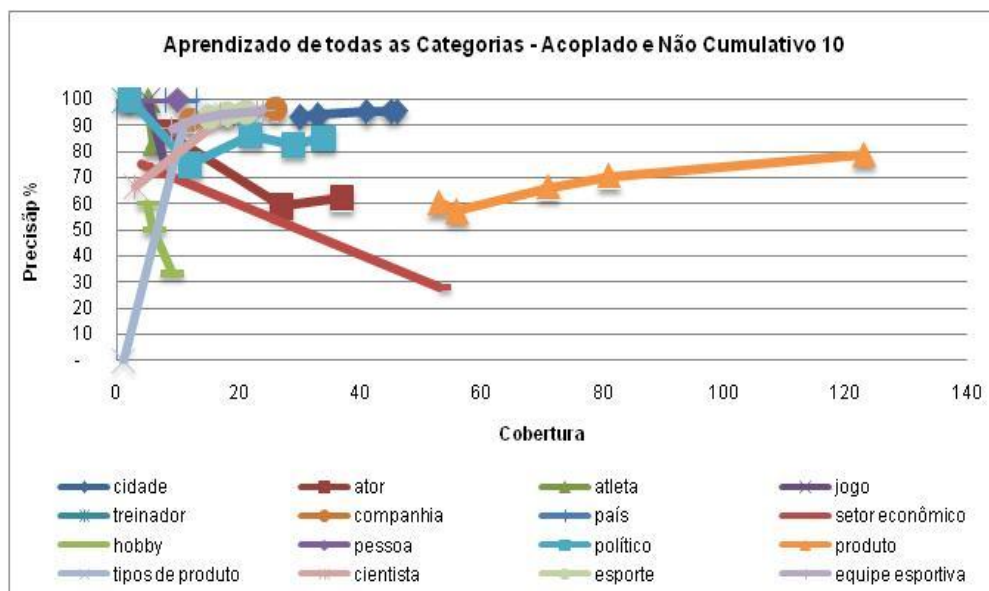


Figura 4.3 Resultados - Experimento Acoplado e Não Cumulativo 10

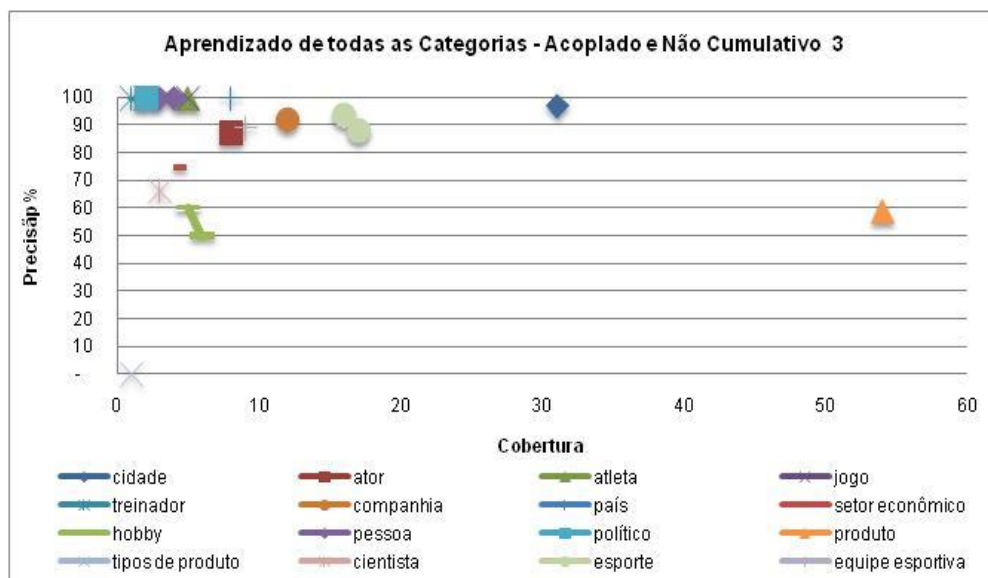


Figura 4.4 Resultados - Experimento Acoplado e Não Cumulativo 3

O maior número de resultados foi possível porque com o acoplamento de padrões negativos foi realizada a filtragem de melhores ENs e PTs, o que possibilitou maior número de iterações devido ao aumento de dados confiáveis.

Comparando o experimento I (Figura 4.1) e III (Figura 4.3) é possível afirmar que as diferenças são maiores que as apontadas para II e IV devido ao número de até 10 PTs poderem ser promovidos. Contudo também reafirma que o acoplamento ajuda o sistema a prosseguir por mais iterações, além de garantir maior número de acertos.

4.4 Experimentos V e VII – Acoplados e Não Acoplados Cumulativos 10

Estes experimentos comparam o uso do acoplamento, com o parâmetro de promoção de 10 PTs por categoria/iteração, com o aprendizado sem o uso do acoplamento, mas com as demais características idênticas. Os resultados podem ser vistos nas figuras de 3.5 a 3.21, em que cada figura mostra a cobertura e precisão de uma categoria.

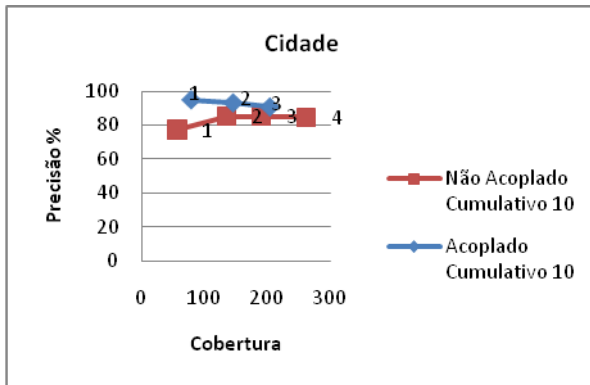


Figura 4.5 Aprendizado Acoplado e Não Acoplado Cumulativo 10 para Cidade

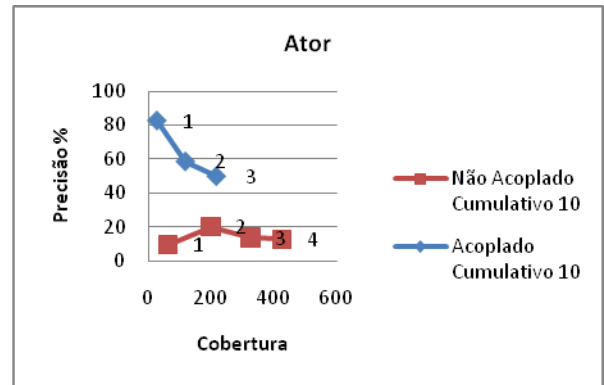


Figura 4.6 Aprendizado Acoplado e Não Acoplado Cumulativo 10 para Ator

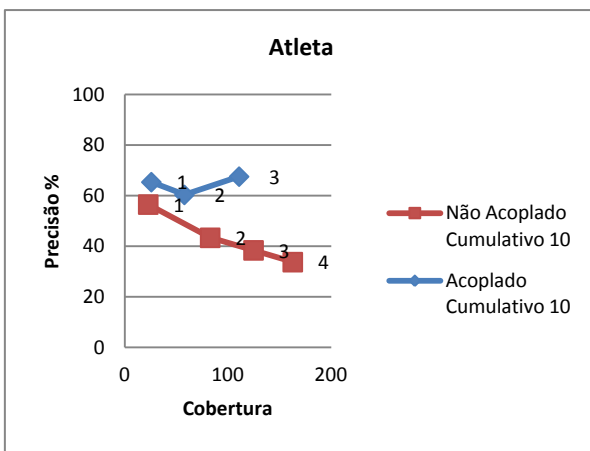


Figura 4.7 Aprendizado Acoplado e Não Acoplado Cumulativo 10 para Atleta

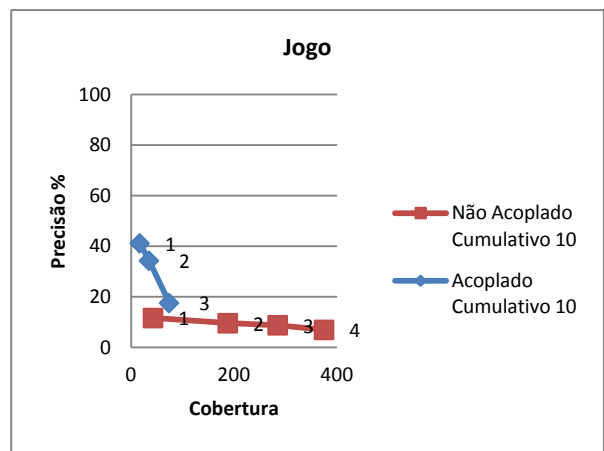


Figura 4.8 Aprendizado Acoplado e Não Acoplado Cumulativo 10 para Jogo

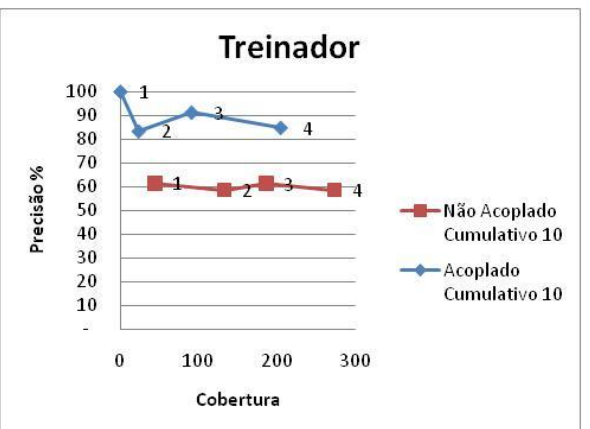
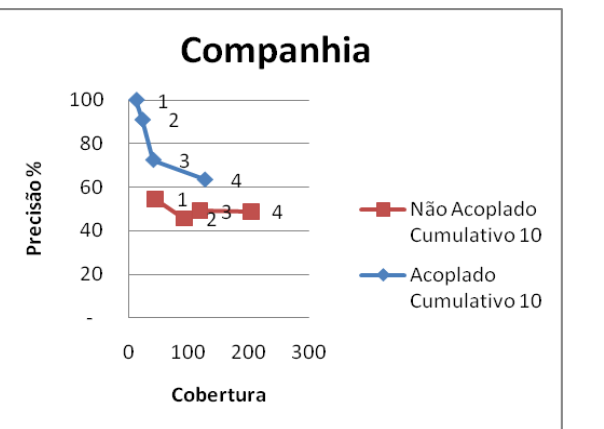


Figura 4.9 Figura

4.10 Aprendizado Acoplado e Não Acoplado Cumulativo 10 para Treinador



4.11 Aprendizado Acoplado e Não Acoplado Cumulativo 10 para Companhia

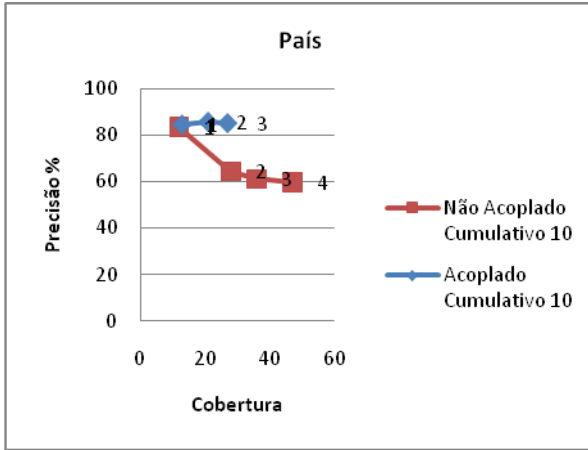


Figura 4.12 Aprendizado Acoplado e Não Acoplado Cumulativo 10 para País

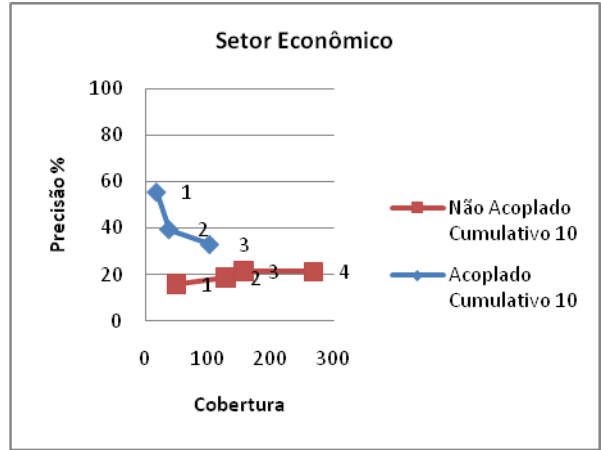


Figura 4.13 Aprendizado Acoplado e Não Acoplado Cumulativo 10 para Setor Econômico

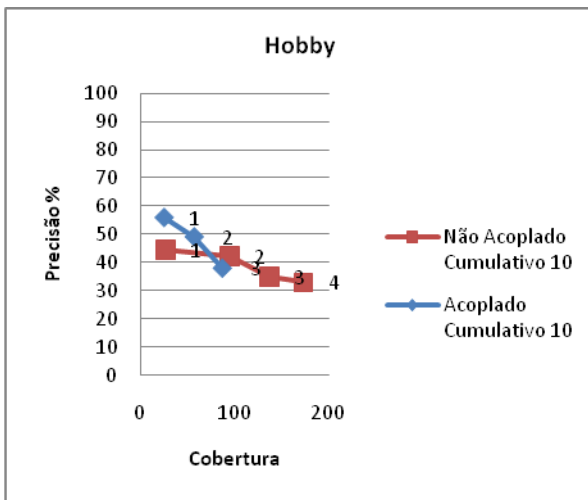


Figura 4.14 Aprendizado Acoplado e Não Acoplado Cumulativo 10 para Hobby

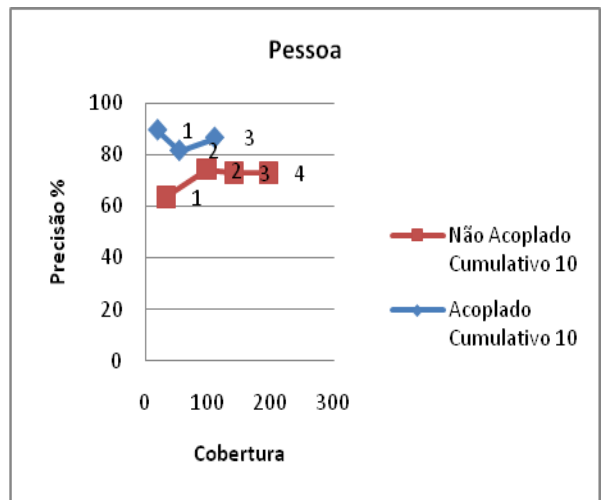


Figura 4.15 Aprendizado Acoplado e Não Acoplado Cumulativo 10 para Pessoa

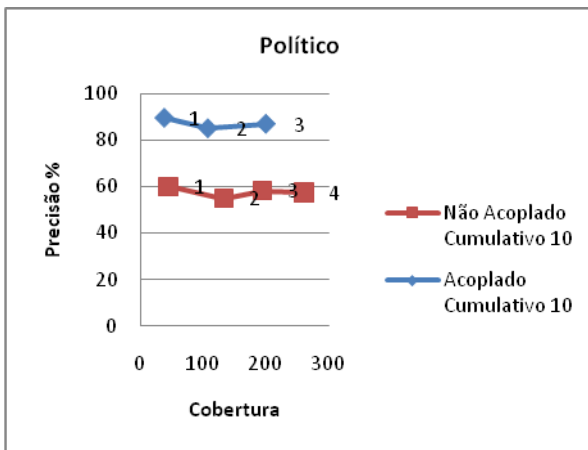


Figura 4.16 Aprendizado Acoplado e Não Acoplado Cumulativo 10 para Político

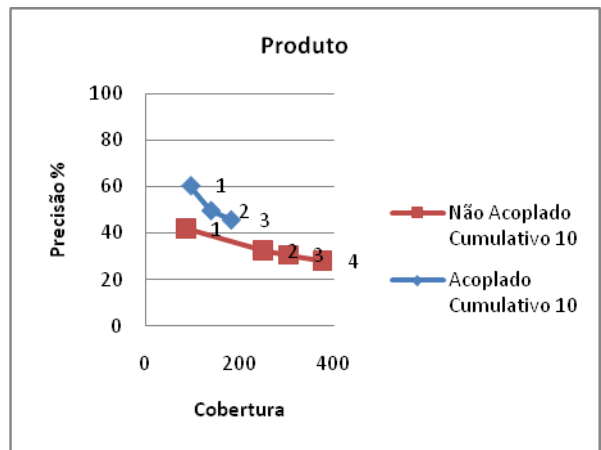


Figura 4.17 Aprendizado Acoplado e Não Acoplado Cumulativo 10 para Produto

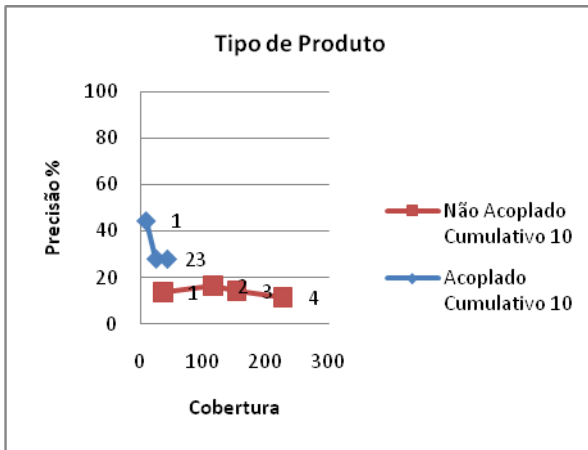


Figura 4.18 Aprendizado Acoplado e Não Acoplado Cumulativo 10 para Tipo de Produto

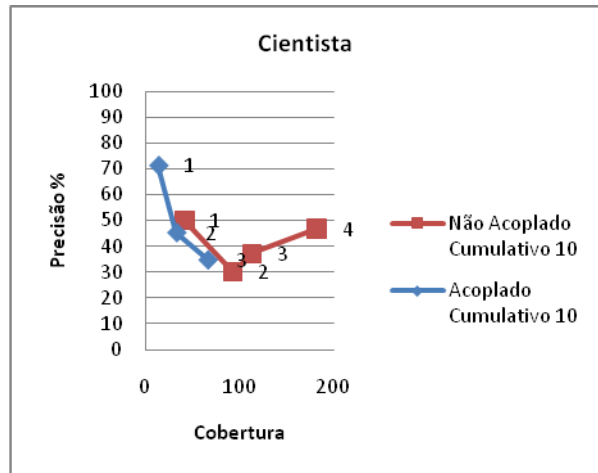


Figura 4.19 Aprendizado Acoplado e Não Acoplado Cumulativo 10 para Cientista

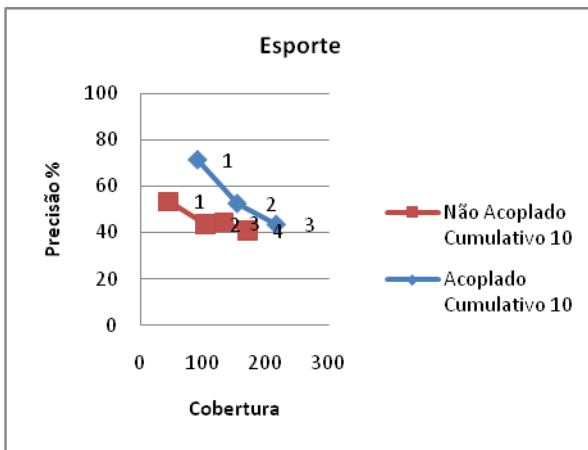


Figura 4.20 Aprendizado Acoplado e Não Acoplado Cumulativo 10 para Esporte

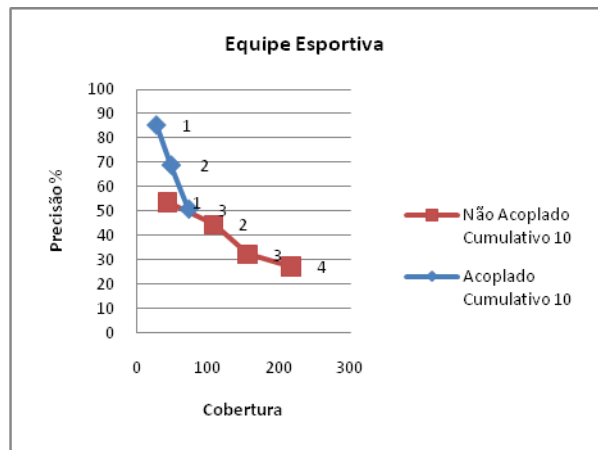


Figura 4.21 Aprendizado Acoplado e Não Acoplado Cumulativo 10 para Equipe Esportiva

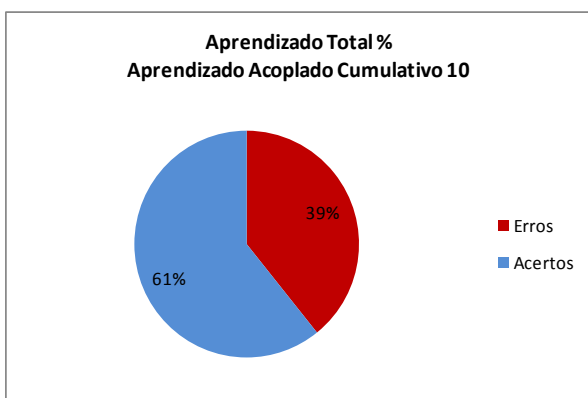


Figura 4.22 Percentual de erros e acertos em todas as categorias no Aprendizado Acoplado Cumulativo 10

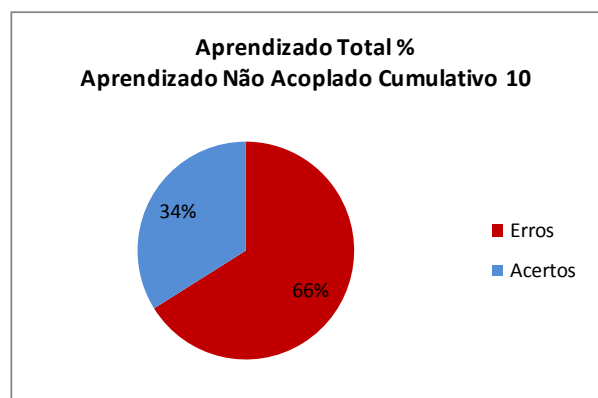


Figura 4.23 Percentual de erros e acertos em todas as categorias no Aprendizado Não Acoplado Cumulativo 10

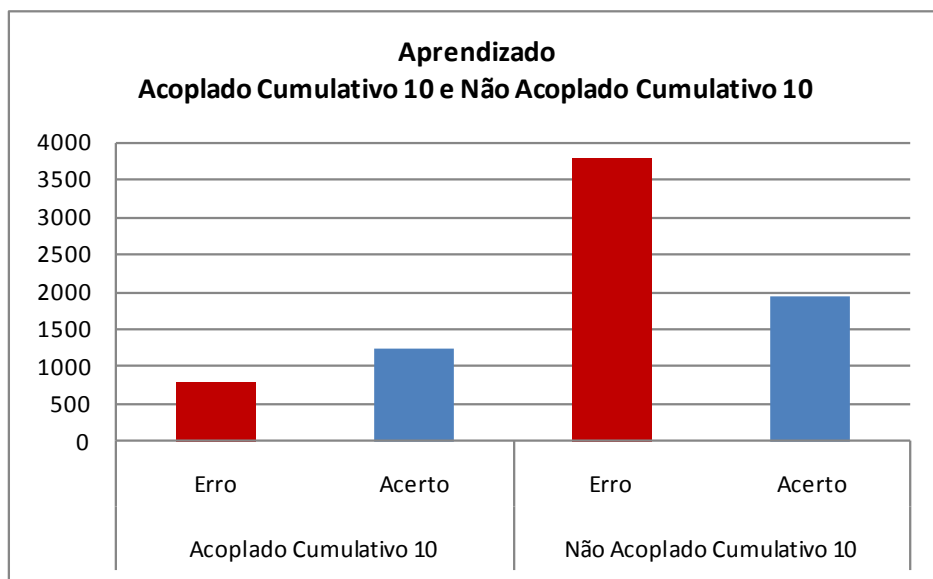


Figura 4.24 Erros e Acertos no Aprendizado Acoplado e Não Acoplado Cumulativo 10

De forma geral, em quase todos os gráficos apresentados nessa seção, é possível observar que a precisão é maior com o uso do acoplamento. Em quase todos experimentos a cobertura de conhecimento no experimento V (não acoplado com padrões negativos) é mais alta que para o VII (acoplado com padrões negativos). Isso significa que embora tenha ocorrido um número menor de cobertura com o uso de acoplamento, a taxa de acerto comparado à quantidade de conhecimento promovido foi maior que o não uso do acoplamento, como pode ser visto na Figura 4.22 e na Figura 4.23. Esses resultados mostram empiricamente que o acoplamento de padrões negativos auxilia na minimização do problema de desvio de conceito. Na Figura 4.24 pode-se perceber que proporcionalmente o acoplamento apresenta resultados mais precisos (menos desvio de conceito).

Algumas categorias tiveram resultados que podem ser melhorados nos dois experimentos, principalmente no VII, com as categorias: Ator, Atleta, Treinador, Pessoa, Político e Cientista. Tais categorias tiveram seus resultados com a cobertura baixa devido à base ontológica usada, na qual Ator, Atleta, Treinador,

Político e Cientista são mutuamente exclusivos. O acoplamento através de padrões negativos é importante, mas precisa ser definido de maneira muito cuidadosa. Isso também aconteceu com as categorias Jogo e Esporte, as quais também foram dadas como mutuamente exclusivas. O uso do acoplamento nestes casos piorou os resultados justamente por utilizar como padrões negativos os padrões de categorias que não deveriam ser mutuamente exclusivas. Além disso, os resultados para a categoria Pessoa não contam com os resultados de todas suas subcategorias (Atleta, Ator, Político, etc.), ou seja, a ideia de subcategorias foi explorada nesta implementação.

Na categoria Hobby foram atingidos poucos resultados em ambos os experimentos, principalmente no acoplado, também devido à configuração da ontologia, pois são extraídos muitos verbos (sair, dançar, correr), o provoca confusão em iterações seguintes.

Nas categorias Produto, País, Setor Econômico e Tipo de Produto os resultados, tanto em precisão quanto em cobertura caíram; isso ocorreu devido a dificuldade de extração e de padrões fracos ¹¹usados.

Em Produto e Tipo de Produto também houve confusão de resultados entre estas categorias, isso devido provavelmente à ontologia e os exemplos usados. Já em Equipe Esportiva, os resultados obtidos mostraram confusão entre equipes profissionais (ex: Equipe Google) e atletas, o que significa que a ontologia precisa de uma definição mais específica neste caso.

Apesar dos detalhes mencionados anteriormente que podem ser melhorados, em Político, País, Atleta, Cidade, Esporte e Ator foram obtidos bons resultados com

¹¹ Definição na página 49

o acoplamento ao se considerar a cobertura total atingida para cada uma das categorias e a configuração usada na ontologia. Bons resultados, pois o acoplamento conseguiu manter por mais tempo melhor precisão que o não acoplado, além da abrangência de ambos ficarem próximas.

O experimento VII deve ser ajustado, pois com a configuração atual da ontologia e o número de PTs a serem promovidos ele se tornou computacionalmente inviável quanto ao tempo de execução. Por isso, para este experimento são apresentados resultados somente até a quarta iteração. O experimento foi parado manualmente depois de uma grande demora (em torno de 1 mês para executar uma iteração). Essa demora aconteceu pelo elevado número de padrões negativos a serem executados. Já o experimento V obteve maior número de resultados, porém a precisão das informações caiu.

Realizando a comparação entre os experimentos, o VII pode ser considerado melhor, apesar dos ajustes. Um dos ajustes pode ser visto na próxima subseção, cujo número de PTs é somente 3 e não 10.

4.5 Experimentos VI e VIII – Acoplados e Não Acoplados Cumulativos 3

Estes experimentos comparam o uso e o não uso do acoplamento, assim como a Subseção 4.4 com os experimentos V e VII, porém com o parâmetro de promoção máxima de 3 PTs por categoria/iteração.

Os resultados podem ser vistos nas figuras de 4.25 a 4.43, em que cada figura mostra a cobertura e precisão de uma das categorias. A mudança de 10 para 3 no número de PTs fez com que o sistema escolhesse os pesos melhores na promoção de PTs trazendo melhores extrações.

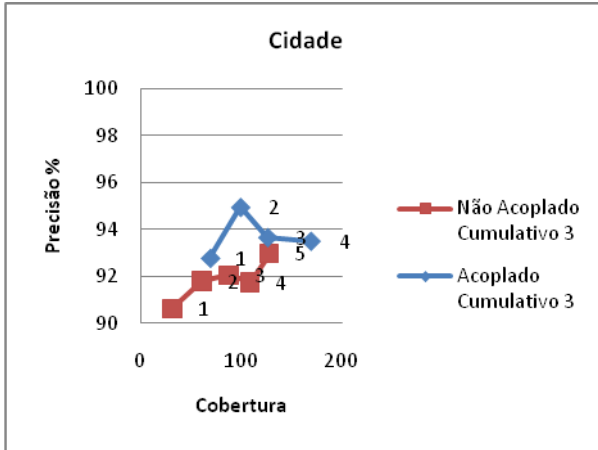


Figura 4.25 Aprendizado Acoplado e Não Acoplado Cumulativo 3 para Cidade

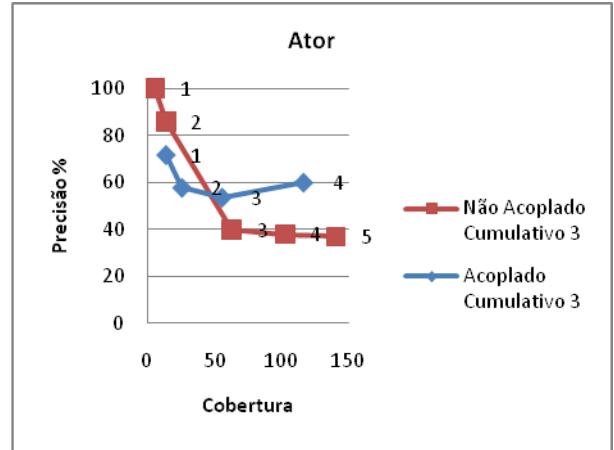


Figura 4.26 Aprendizado Acoplado e Não Acoplado Cumulativo 3 para Ator

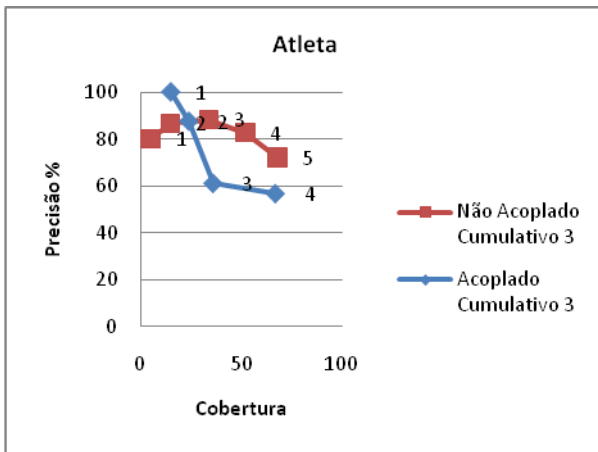


Figura 4.27 Aprendizado Acoplado e Não Acoplado Cumulativo 3 para Atleta

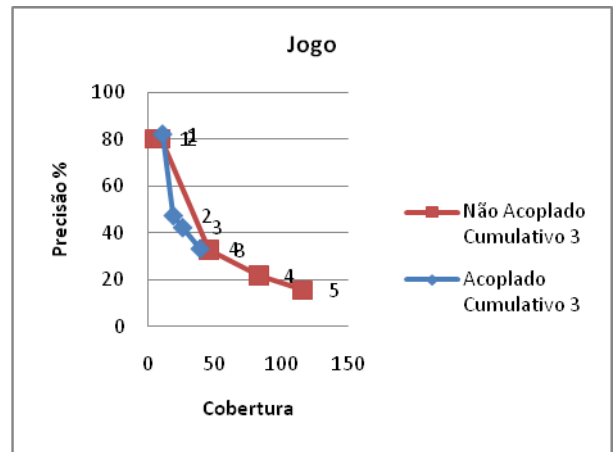


Figura 4.28 Aprendizado Acoplado e Não Acoplado Cumulativo 3 para Jogo

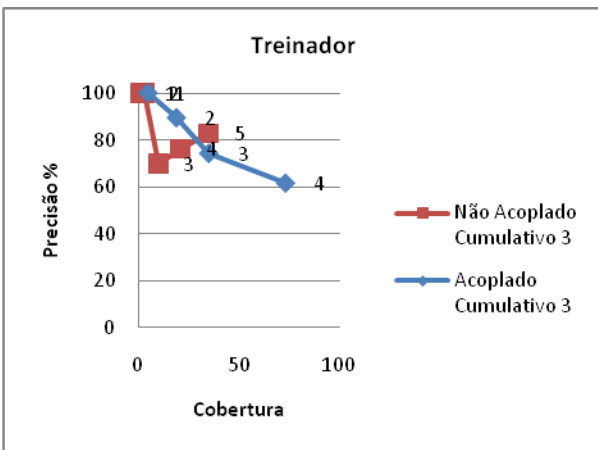


Figura 4.29 Aprendizado Acoplado e Não Acoplado Cumulativo 3 para Treinador

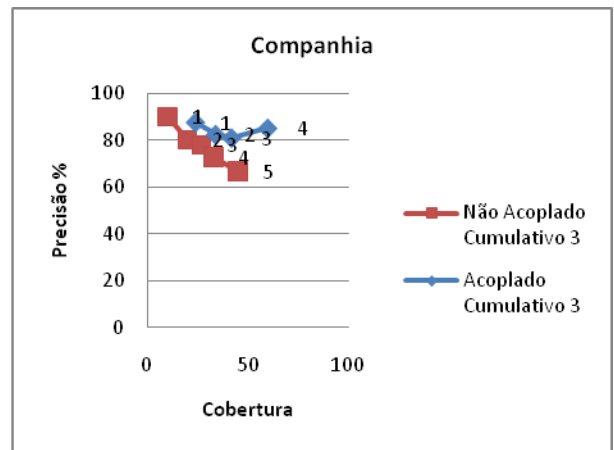


Figura 4.30 Aprendizado Acoplado e Não Acoplado Cumulativo 3 para Companhia

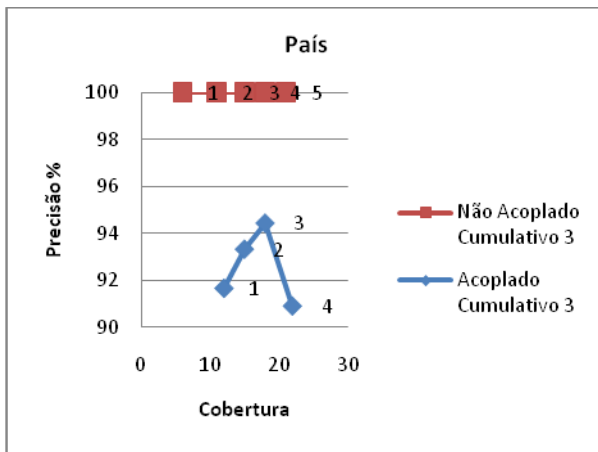


Figura 4.31 Aprendizado Acoplado e Não Acoplado Cumulativo 3 para País

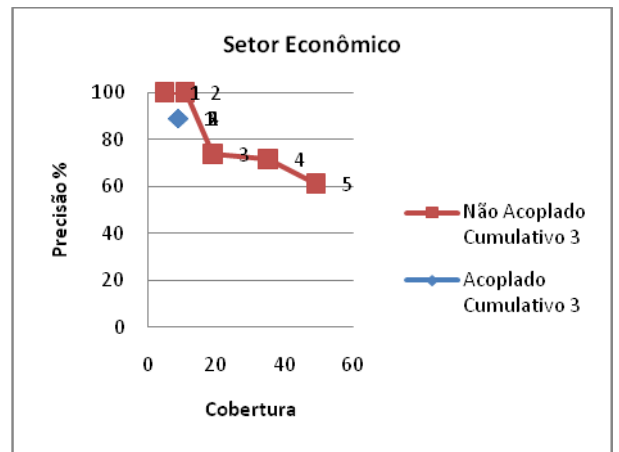


Figura 4.32 Aprendizado Acoplado e Não Acoplado Cumulativo 3 para Setor Econômico

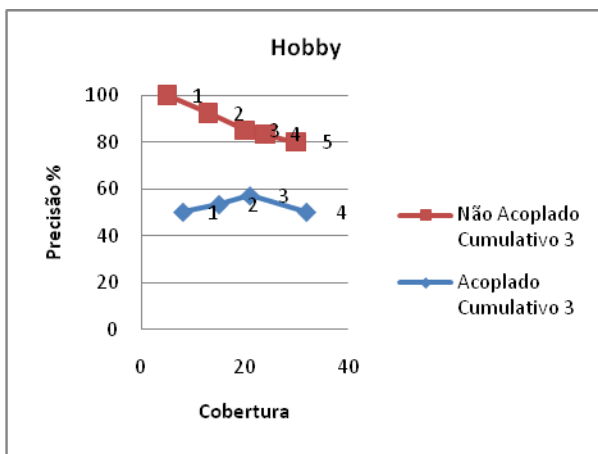


Figura 4.33 Aprendizado Acoplado e Não Acoplado Cumulativo 3 para Hobby

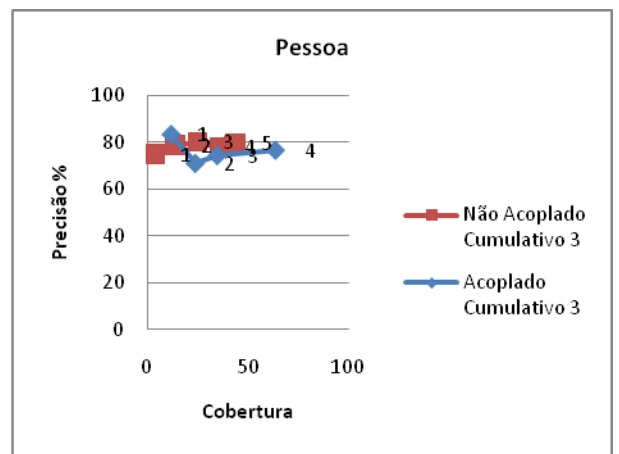


Figura 4.34 Aprendizado Acoplado e Não Acoplado Cumulativo 3 para Pessoa

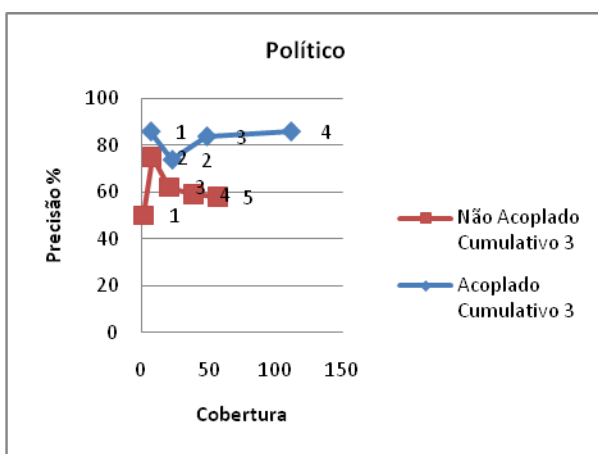


Figura 4.35 Aprendizado Acoplado e Não Acoplado Cumulativo 3 para Político

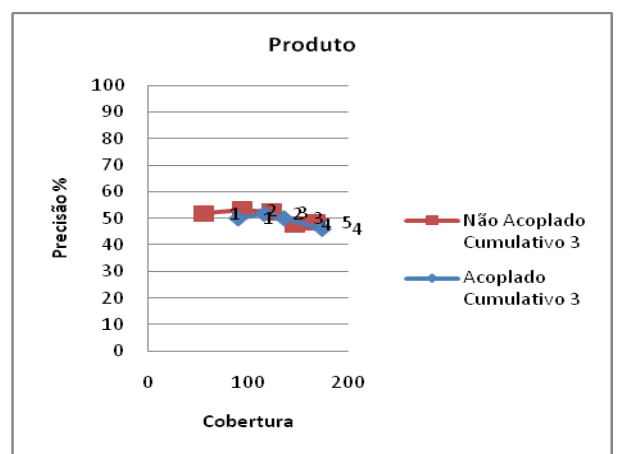


Figura 4.36 Aprendizado Acoplado e Não Acoplado Cumulativo 3 para Produto

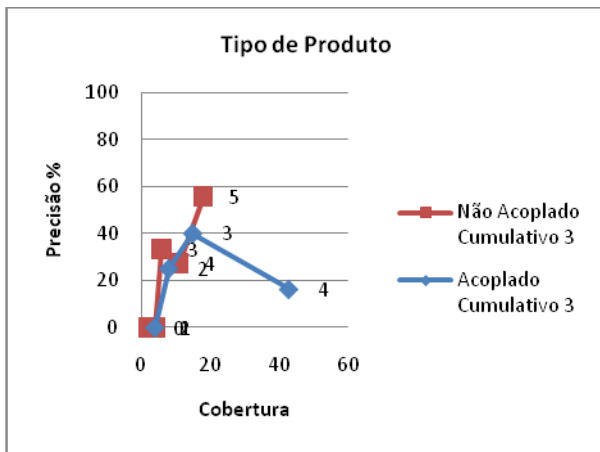


Figura 4.37 Aprendizado Acoplado e Não Acoplado Cumulativo 3 para Tipo de Produto

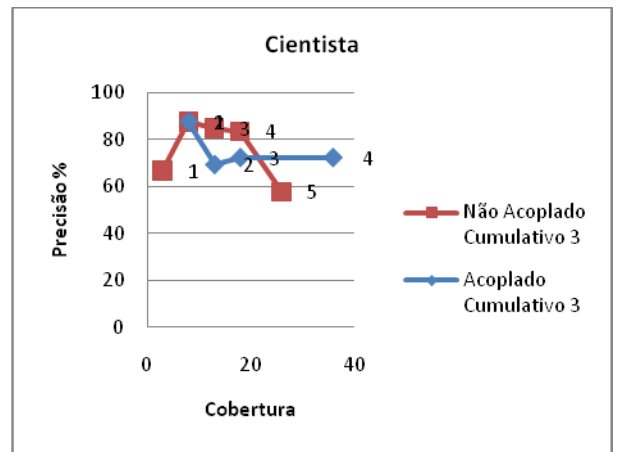


Figura 4.38 Aprendizado Acoplado e Não Acoplado Cumulativo 3 para Cientista

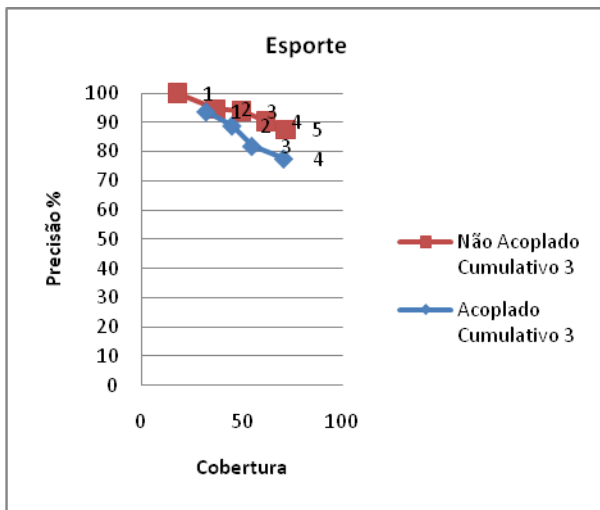


Figura 4.39 Aprendizado Acoplado e Não Acoplado Cumulativo 3 para Esporte

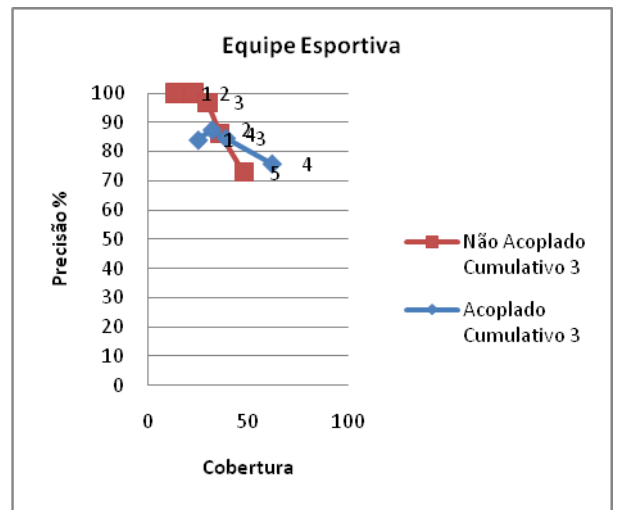


Figura 4.40 Aprendizado Acoplado e Não Acoplado Cumulativo 3 para Equipe Esportiva

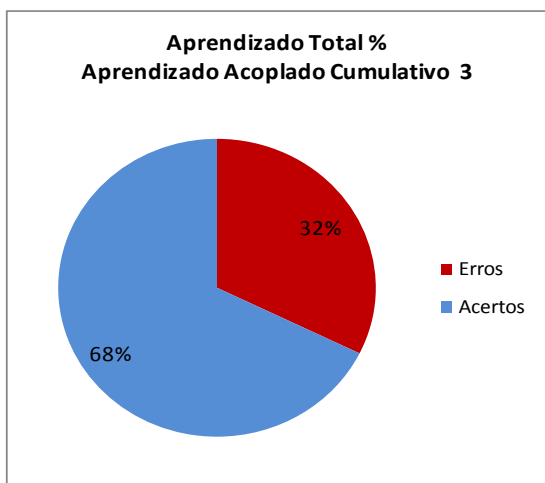


Figura 4.41 Percentual de erros e acertos em todas as categorias no Aprendizado Acoplado e Não Acoplado Cumulativo 3

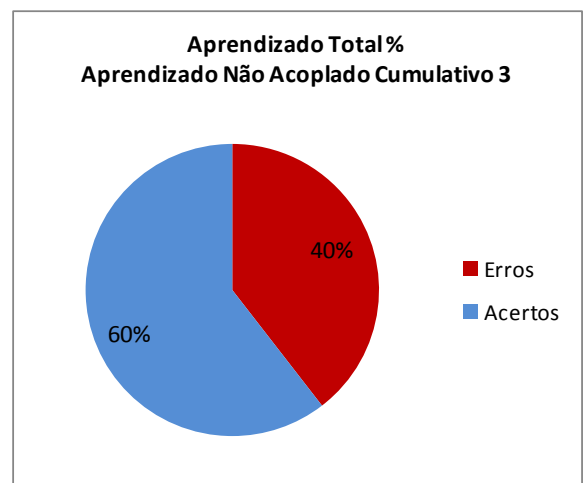
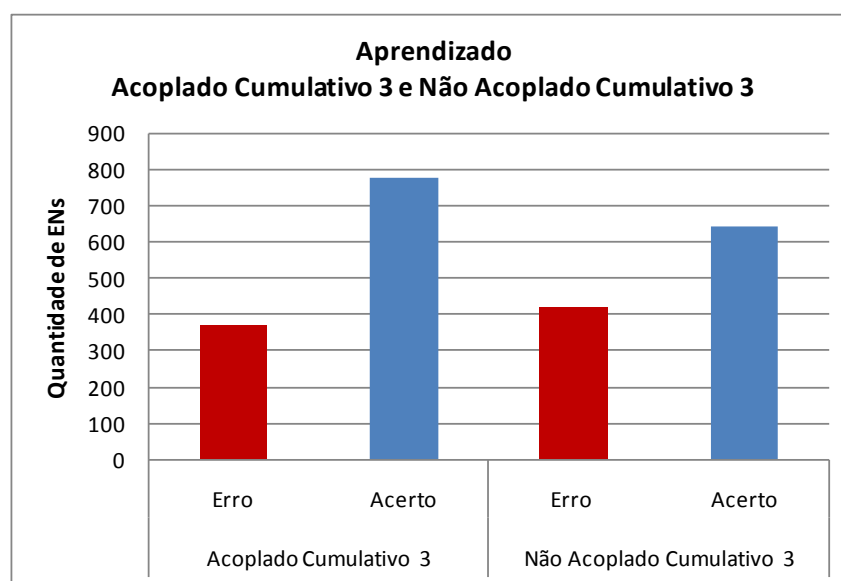


Figura 4.42 Percentual de erros e acertos em todas as categorias no Aprendizado Acoplado e Não Acoplado Cumulativo 3



4.43 Erros e Acertos no Aprendizado Acoplado e Não Acoplado Cumulativo 3

De forma geral, em quase todos os gráficos de resultados de aprendizado das categorias, é possível observar que a precisão é maior com o uso do acoplamento, assim como visto na comparação dos experimentos V e VII na Subseção 4.4. Por outro lado, diferentemente dos resultados dos experimentos V e VII, a cobertura também se mostra maior em quase todos os gráficos com o uso do acoplamento. Assim também como na subseção 4.4, no experimento acoplado a taxa de acerto em relação a quantidade de conhecimento foi maior que no sem acoplamento, como pode ser visto na Figura 4.41 e na Figura 4.42. Na Figura 4.24, é mostrado que foi alcançada maior quantidade de acertos com o uso do acoplamento, enquanto no experimento sem acoplamento esta a quantidade foi menor e a de erro maior (maior desvio de conceito).

As categorias referentes a pessoas, tais como Ator, Atleta, Treinador, Pessoa, Político e Cientista tiveram o mesmo problema apresentado na Subseção 4.4 referente à configuração da ontologia. Desse modo todas as categorias são mutuamente exclusivas às demais, com exceção de Pessoa. Além disso, a categoria

Pessoa não é o conjunto de todas as outras categorias referentes a pessoas, devido a não usar a característica de subcategorias. O mesmo ocorreu com Jogo e Esporte, os quais também foram dados como mutuamente exclusivos. Devido a essas configurações o acoplamento não conseguiu melhores resultados em alguns casos e/ou não obteve uma quantidade melhor de resultados que poderia ser possível com outra modelagem da ontologia. Por outro lado, mesmo assim, o acoplamento se manteve com melhores precisões durante mais iterações nas categorias Esporte, Ator, Atleta, Pessoa, Político e Cientista.

Com relação à categoria Pessoa, seguindo outra estrutura ontológica na qual se considera incluso em Pessoa todas as suas subcategorias, o resultado obtido está ilustrado na Figura 4.44.

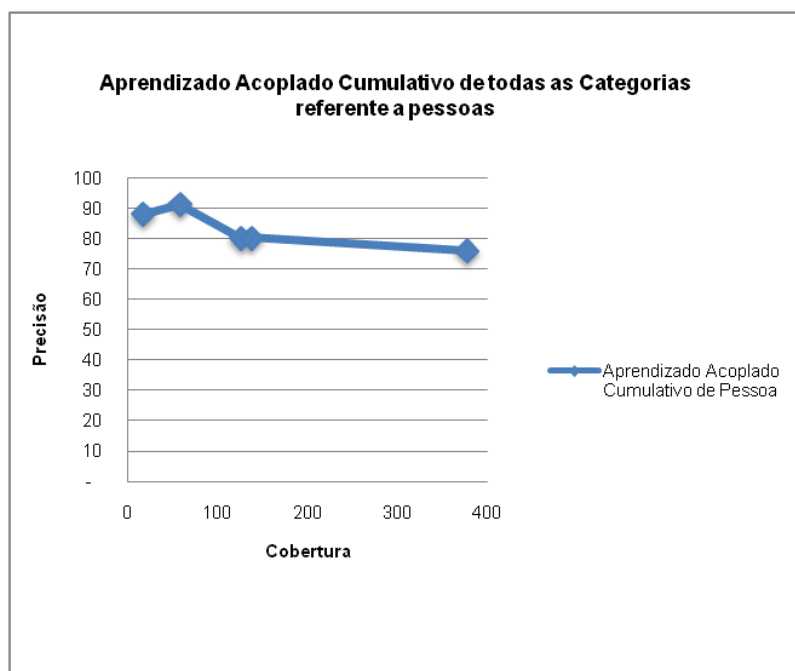


Figura 4.44 Aprendizado de Acoplado Cumulativo de Pessoa

De acordo com a Figura 4.44, os resultados obtidos mostram que a cobertura e precisão foram melhores do que na abordagem usada neste experimento e isto

evidencia de que a ideia de subcategorias (como um novo tipo de acoplamento) pode trazer ganhos ao processo de aprendizado.

Da mesma forma que na subseção 4.4 nas categorias Setor Econômico e Tipo de Produto os resultados, tanto em precisão quanto em cobertura, caíram nos dois experimentos. Este fato ocorreu devido à dificuldade de extração e uso de padrões fracos usados.

Em Produto e Tipo de Produto também houve confusão de resultados devido aos padrões prévios fracos e a ontologia que foi definida, por isso acoplamento também piorou a cobertura. A categoria Equipe Esportiva também teve resultados que mostraram confusão entre equipes profissionais, pois a ontologia precisa de uma definição mais específica neste caso, assim como já dito na Subseção 4.4.

Ao analisar os resultados de acordo com a configuração usada na ontologia, na categoria Cidade a precisão e a cobertura do acoplamento tenderam a melhorar, assim como ocorreu também para as categorias: Companhia, Político, Cientista e Equipe Esportiva.

Já na categoria Produto, Esporte e Equipe Esportiva, foram obtidos poucos resultados, e ambos os experimentos se comportaram de forma parecida. Isso deve-se ao fato dos padrões iniciais fornecidos e a falta de padrões negativos voltados a estas categorias.

A categoria Jogo foi prejudicada pelo acoplamento, devido às categorias mutuamente exclusivas definidas de maneira inadequada e a dificuldade de extração de jogos de tabuleiro. Foi observado na validação de conhecimento que na categoria jogos está ocorrendo confusão com, jogos que não são de tabuleiro, como por exemplo, jogos online. Por outro lado, mesmo assim o acoplamento manteve a precisão melhor por mais iterações nesta categoria.

Um caso diferente nestes experimentos aconteceu com a categoria País, provavelmente o acoplamento atingiu menor taxa de acerto devido a diferentes datas de execução do sistema. Além disso, foram atingidos poucos resultados para ambos os experimentos. Também foram encontrados dados confusos, por exemplo, no que se refere a estados e grandes cidades. O que leva a concluir que uma ontologia mais robusta, com ramificações mais abrangentes e vinculadas pode melhorar a extração.

No caso da categoria Hobby foram obtidos poucos resultados e o acoplamento não ajudou. Isso aconteceu devido aos padrões negativos (por conta da ontologia usada, sem o uso de subgrupos de categorias ex: Jogo não consta como Hobby) e a grande presença de verbos (sair, dançar, correr, etc.) na categoria Hobby. A ocorrência de verbos na categoria Hobby trouxe confusão em iterações seguintes, o que deixou ruim os padrões positivos. Para esta categoria também é necessária uma melhor definição da base ontológica.

No caso da categoria Tipo de Produto, ambos os experimentos caminharam de forma similar até a quarta iteração, porém por conta das confusões nos padrões, no experimento com acoplamento a cobertura e precisão caíram drasticamente.

Os experimentos VI e VIII mostraram que o acoplamento traz bons resultados tanto em cobertura quanto em precisão, além de reafirmarem que a estrutura da ontologia deve ser bem estudada e definida com cuidado.

4.6 Considerações dos Experimentos VII e VIII

O experimento VII, como já dito anteriormente não é viável de ser executado devido ao tempo gasto, já o VIII, tendeu a ser mais preciso e conseguiu executar até a 5ª iteração sem precisar ser interrompido, ou seja, com isso ele é viável para os próximos experimentos a serem mostrados neste projeto. Em outras palavras, o VIII mostrou que o parâmetro 3 para a promoção de PTs é melhor que 10, tanto por ser viável quanto pela precisão. Com o VIII reafirmou-se que alguns resultados ruins são decorrentes da forma da estrutura da ontologia. Além disso, o VIII atingiu melhores precisões nas categorias Cidade, Ator, Jogo, Companhia, país, Setor Econômico, Hobby, Produto, Tipo de Produto, Cientista, Esporte e Equipe Esportiva.

Já nas categorias restantes (Ator, Treinador e Político) o acoplamento com parâmetro 10 obteve maior precisão. Isso ocorreu porque apesar da estrutura atual da ontologia tratar tais categorias como mutuamente exclusivas, o parâmetro 10 usado minimiza um pouco desta característica.

Nas categorias apontadas como melhores em precisão no experimento VIII é importante observar que a categoria Setor Econômico parou de adquirir conhecimento a partir da iteração 3, e as demais não conseguiram alcançar a mesma cobertura do experimento VII. A taxa de acerto e erro de ambos os experimentos pode ser observada na Figura 4.22 e na Figura 4.41.

Foi escolhida abordagem do VIII para ser usado nos próximos experimentos a seguir, por possuir maior taxa de acerto, embora a cobertura seja menor, e pela inviabilidade de realizar o VII com a configuração atual.

4.7 Experimento IX – Relações Semânticas Acoplado com Padrões Negativos

Negativos

Este experimento é o primeiro que trata de acoplamento de Relações Semânticas, que inclui também acoplamento de padrões negativos. A sua parte conceitual foi abordada na subseção 3.6.

Os experimentos voltados a Relações Semânticas não obtiveram muitos resultados. A cobertura foi pequena como pode ser visto na Figura 4.45, mas foi obtida metade com 100% de precisão, apesar de atingir a sexta iteração. Somente seis Relações Semânticas tiveram aprendido.

Os experimentos com Relações Semânticas não envolvem todo o sistema, como apresentado na Subseção 3.8, pois este seria o último experimento a ser mostrado, porém não foi finalizado (não alcançou ainda a iteração final 5).

O experimento IX tem seus resultados mostrados na Figura 4.45.

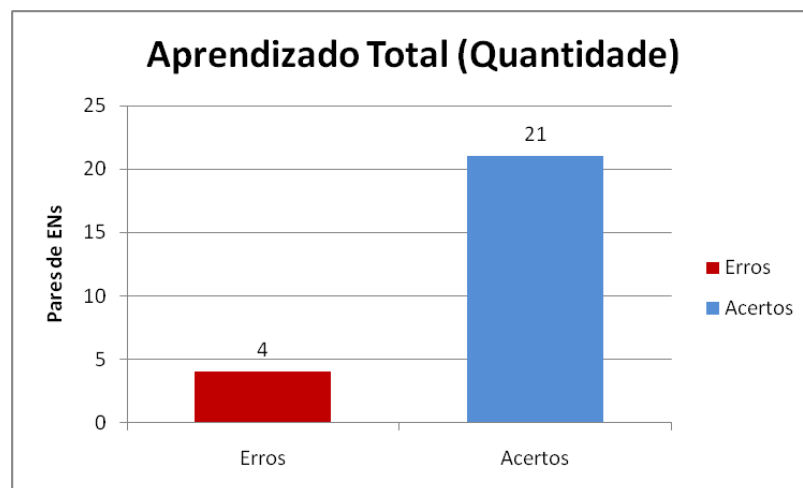


Figura 4.45 Erros e Acertos no Aprendizado de Pares de ENs com Relações Semânticas com seis iterações

Como já mencionado, a quantidade conhecimento alcançada foi muito baixa em todas as categorias, algumas Relações Semânticas não tiveram nenhuma promoção (ex: EquipeJogaEsporte(EquipeEsportiva, Esporte), ExisteMatriz(Companhia, Cidade), etc.). Isso ocorreu devido a dificuldade em se extrair Relações Semânticas, mas também tiveram influência dos padrões previamente definidos, que não ajudaram o quanto era esperado.

4.8 Experimento X – Relações Semânticas Acoplado com Tipagem e Padrões Negativos (Com filtro antes da promoção)

Este experimento foi executado da mesma forma que o IX, porém agora conta também com o acoplamento de categorias (VIII), tipagem e padrões negativos.

Estes acoplamentos funcionam da seguinte forma: Se um Par de ENs é candidato à promoção, primeiramente é verificado se ambas ENs existem na base do experimento VIII, se a resposta for positiva o Par é promovido, caso contrário é realizada a tipagem, conforme foi abordada na Subseção 3.7.

Os resultados obtidos também foram poucos, como podem ser vistos na Figura 4.46, em que somente houve resultados na primeira iteração. O uso da tipagem neste experimento fez com que menos Pares de ENs fossem promovidos, porém atingiu 100% de acerto em 7 categorias (as únicas que tiveram aprendizado), e o número de Relações Semânticas atingido foi muito baixo, 17 no total.

Lembrando que o aprendizado de relações possui maior dificuldade que os outros tipos citados. Isso pode ser visto tanto na implementação deste trabalho

quanto na do NELL. Será necessário um estudo mais detalhado e cuidadoso para que se melhore este aprendizado.

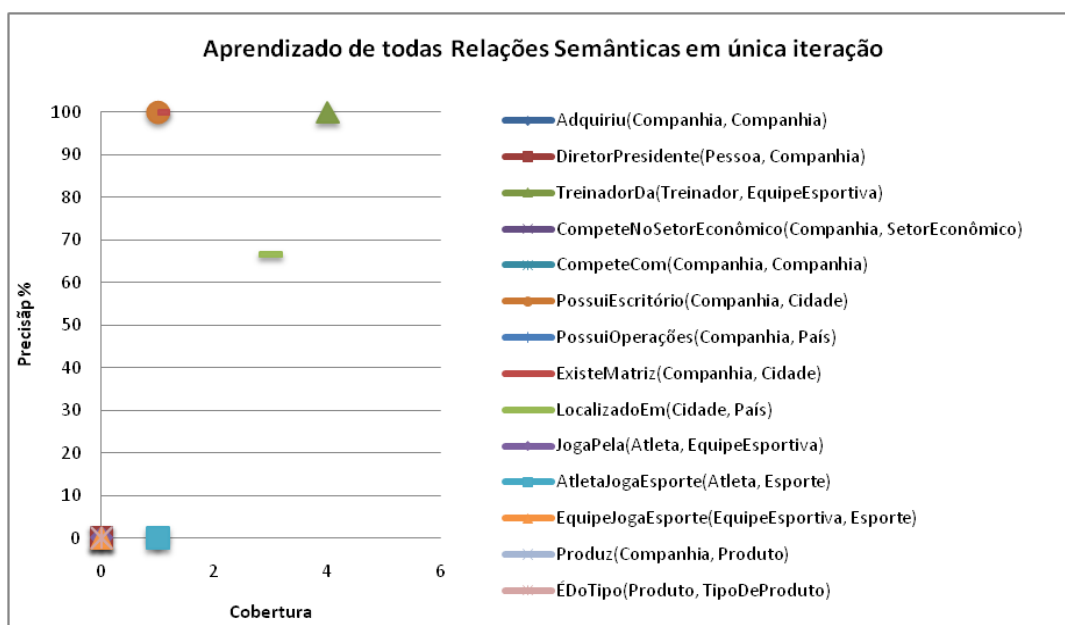


Figura 4.46 Experimento de Acoplamento de Relações Semânticas e Tipagem com Filtro de 1/3 dos Pares de ENs

Assim como no experimento IX, obtido por conta da dificuldade em se extrair Relações Semânticas e pelos fracos padrões definidos previamente.

4.9 Experimento XI - Relações Semânticas Acoplado com Tipagem e Padrões Negativos (Com filtro antes da promoção)

Neste experimento foram realizados os mesmos acoplamentos feitos em X, a única diferença é o filtro antes da tipagem, pois neste não é usado. O intuito disso foi obter maior número de resultados para Relações Semânticas usando tipagem.

Embora a quantidade de novas ENs tenha sido baixa neste experimento, a base de conhecimento aumentou, porém essa forma de aprendizado deve ser melhorada para ampliar a cobertura. Em precisão foram obtidos bons resultados, como pode ser visto na Figura 4.47, em que a maioria das categorias obteve 100% de acerto. Devido à baixa cobertura e a dificuldade de extração, que ocorre por conta dos padrões prévios e da baixa ocorrência nas páginas web, somente foi possível a execução de uma iteração.

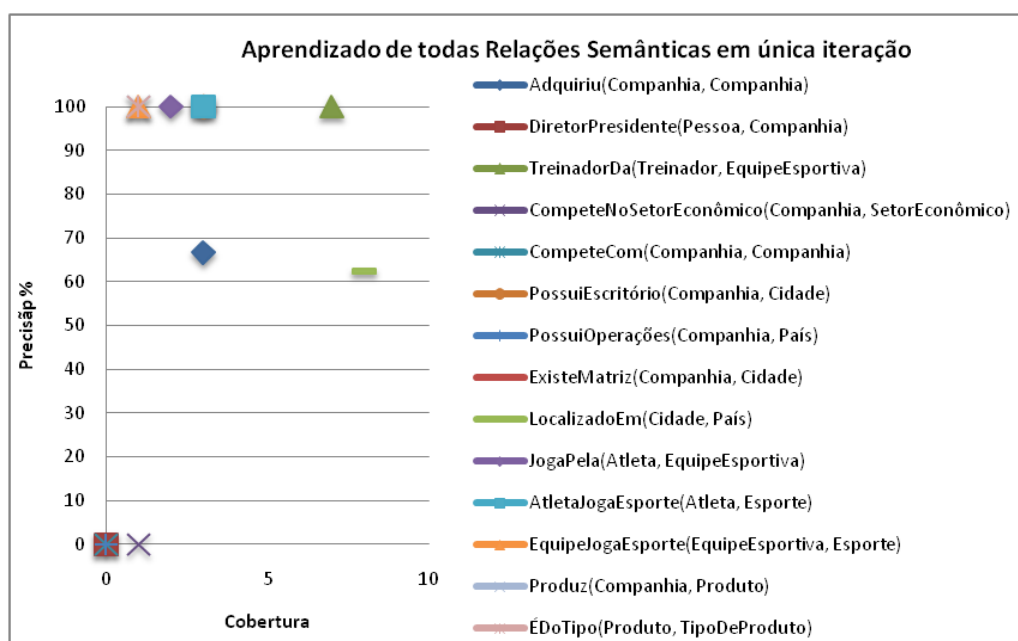


Figura 4.47 Experimento de Acoplamento de Relações Semânticas e Tipagem sem Filtro antes da Promoção

De maneira geral, como pode ser visto no gráfico da Figura 4.47, três categorias atingiram 100% de acerto, uma alcançou 67% e mais uma 63%. Tais resultados indicam a mesma tendência dos experimentos para Relações Semânticas quanto ao número baixo de resultados, apesar de haver maior dificuldade na extração de Relações Semânticas: os padrões prévios devem ser melhorados bem como a estrutura da base de conhecimento usada.

4.10 Considerações dos Experimentos IX, X e XI – Aprendizado de ENs

Nesta Subseção constam considerações dos experimentos IX, X e XI de acordo com acerto e erro na promoção de ENs, não Pares de ENs, por isso os valores são diferentes dos resultados de relações semânticas.

Os experimentos IX e XI tiveram maior taxa de cobertura e menor taxa de acerto comparados ao X, que se comportou da forma contrária. O comportamento aconteceu devido ao filtro de 1/3 de candidatos antes da tipagem, o que levou à maior precisão e menor cobertura se usado, caso contrário aumenta a cobertura e diminui a precisão. Tais resultados podem ser vistos na Figura 4.48.

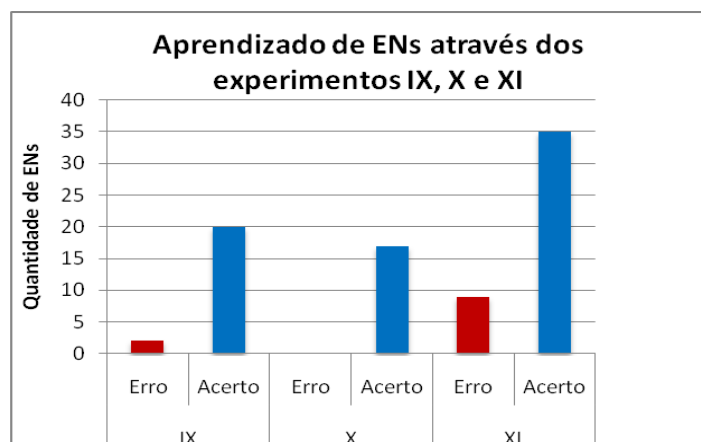


Figura 4.48 Aprendizado de ENs com todos os acoplamentos

Mesmo com os resultados exibindo uma alta taxa de acerto, é bom salientar que estes experimentos precisam ter a ontologia e os padrões prévios melhor definidos para a obtenção de maior cobertura.

Apesar da baixa coberta, estes experimentos mostraram a tendência a melhor precisão do que sem o uso do acoplamento de relações semânticas.

4.11 Análise de Todos os Experimentos

Os experimentos não cumulativos (I,II,III e IV) obtiveram poucos resultados e não passaram da segunda iteração. Isso aconteceu porque não houve conhecimento promovido suficiente para que houvessem mais promoções

Já os experimentos cumulativos (V, VI, VII e VIII) tiveram maior número de iterações e de resultados, isso foi devido ao uso de conhecimento adquirido anteriormente para aprender mais posteriormente. A cada iteração pode ser usada toda a base de conhecimento para obtenção de mais aprendizado, mas isso não quer dizer que mesmos PTs e ENs podem ser usados mais de uma vez, pois somente são usados uma única vez. Significa que o conhecimento extraído em iterações anteriores, mas não promovido, pode ser novamente candidato a promoção, caso haja aprendizagem nas iterações seguintes. Por outro lado, nestes mesmos experimentos o uso do acoplamento de padrões negativos fez com que houvesse tendência a melhor cobertura e precisão, com exceção dos casos que tiveram configurações restritivas na base de conhecimento (Ex: categoria Pessoa), os quais foram justificados anteriormente neste capítulo. O uso do acoplamento de padrões negativos trouxe maior aprendizado porque fortaleceu a base de conhecimento com padrões mais precisos, que conseqüentemente tendem a extrair melhores ENs e PTs.

Os experimentos VII e VIII obtiveram os melhores resultados de maneira geral. O VII possui maior cobertura, mas além de ser inviável computacionalmente, o parâmetro de promoção de PTs igual a 10 faz com a precisão caia rapidamente com o passar das iterações. Já o VIII foi viável devido a alteração do parâmetro para 3, mas a cobertura diminuiu, mas mesmo assim teve bons resultados, pois conseguiu fazer com que a precisão caísse mais suavemente, e com isso alcançou o objetivo deste projeto: mostrou empiricamente que a integração de diferentes tarefas (diversos acoplamentos) tendem a minimizar a divergência do aprendizado que comumente ocorre com o problema do desvio de conceito [17]. Os resultados obtidos, entretanto, não solucionaram o problema por completo. Assim, mais acoplamentos podem ser necessários como sugerem resultados obtidos em trabalhos do RTW.

Os experimentos de aprendizado de Relações Semânticas (IX, X e XI) obtiveram poucos resultados (cobertura muito baixa). Entretanto, mesmo com baixa cobertura, os resultados destes experimentos sugerem que este acoplamento é um bom aliado no aprendizado de ENs. O experimento IX obteve precisão 100% em três de cinco Relações Semânticas que tiveram promoções. Nos experimentos X e XI notou-se que a precisão continuou alta, na maioria dos casos a precisão se manteve em 100%. Portanto mais uma vez os resultados mostram a tendência de que o acoplamento de diferentes tarefas pode melhorar o aprendizado. No experimento X, o uso da tipagem ajudou a apurar os padrões mais fortes¹², que são os primeiros do ranking, porém só executou uma iteração.

¹² Definição na página 49

Capítulo 5

CONCLUSÃO

5.1 Objetivos Alcançados

O objetivo inicial proposto neste projeto de pesquisa foi investigar, propor e implementar métodos e algoritmos que permitissem a construção de um sistema computacional capaz de realizar a extração de conhecimento a partir da Web em português, por meio da criação de uma base de conhecimento consistente, atualizada constantemente à medida que novos conhecimentos são extraídos. Foi proposto que diferentes tarefas integradas minimizariam erros no Aprendizado Semissupervisionado na extração de conhecimento a partir da leitura da Web em português.

Foram usados vários métodos acoplados, estes com base no co-training com bootstrapping. O foco principal foi o acoplamento com os métodos combinados de aprendizado de: ENs, PTs, Pares de ENs, PTs de Relações Semânticas, padrões negativos e tipagem. Tais métodos, os quais já haviam sido propostos no NELL, foram reimplementados de acordo com as necessidades do aprendizado a partir da

leitura da web em português. Os algoritmos foram implementados desde o início, pois além de necessária adaptação à língua portuguesa, não havia corpus da web em português disponível, como no NELL para a língua inglesa.

Foram obtidos bons resultados que mostram empiricamente a tendência à minimização do desvio de conceito com o uso dos acoplamentos propostos, que são a integração das diferentes tarefas sugeridas a partir do NELL (apresentadas no Capítulo 3).

Embora apontada tendência à minimização do desvio de conceito a partir do uso de acoplamentos, houve limitações em alguns casos, as quais foram detalhadas no Capítulo 4. No aprendizado de categorias houve limitações por conta da estrutura da base de conhecimento, o que pode ser melhorado com a criação de novas categorias, melhora dos padrões prévios e a reestruturação das Relações Semânticas. Isso levou também à conclusão de que é necessário um estudo bem cuidadosa na criação da estrutura da base de conhecimento, pois se bem criada ajuda, caso contrário pode trazer maiores problemas de desvio de conceito.

No aprendizado de Relações Semânticas foram alcançados poucos resultados e por isso será necessária melhora nos padrões prévios, além desta tarefa também poder ser melhorada a partir de uma melhoria na estrutura da base de conhecimento.

5.2 Contribuições e Limitações

De uma maneira geral, as principais contribuições científicas e tecnológicas deste projeto estão vinculadas aos desafios apresentados e discutidos na seção de Introdução. Assim, a principal contribuição deste projeto está na investigação, proposta e implementação de um sistema com características vinculadas aos conceitos do aprendizado sem fim apresentados no NELL, e capaz de extrair conhecimento a partir da Web em português.

Como principal contribuição adjacente pode ser destacada a futura integração de dois sistemas: o RTWP subsidiado pela Web em português, um dos principais objetivos deste projeto e o RTW subsidiado pela Web em inglês [7]. A integração dos dois sistemas permitirá a construção de um sistema único que utiliza conhecimentos extraídos nas duas línguas e, assim, possa trazer ganhos aos sistemas individuais.

Houve algumas limitações durante o desenvolvimento deste projeto, como a quantidade de páginas que puderam ser acessadas, pois o *Yahoo Boss* possui um número de acessos e páginas permitidos diariamente. Além disso, por conta do alto número de acessos a partir da UFSCar, várias vezes ocorreu bloqueio e/ou bloqueio por ID(número de identificação fornecido pela Yahoo). Um número menor de páginas puderam ser extraídas diariamente. Não se teve apoio de buscadores ou APIs que pudessem possibilitar maior acesso, somente foi usado o *Yahoo Boss*, exatamente porque é uma API gratuita.

Por outro lado, houve limitação de resultados também devido a estrutura da base de conhecimento, que pode ser melhorada como a nova proposta dada na Figura 5.1, em que várias categorias foram mais detalhadas e/ou generalizadas.

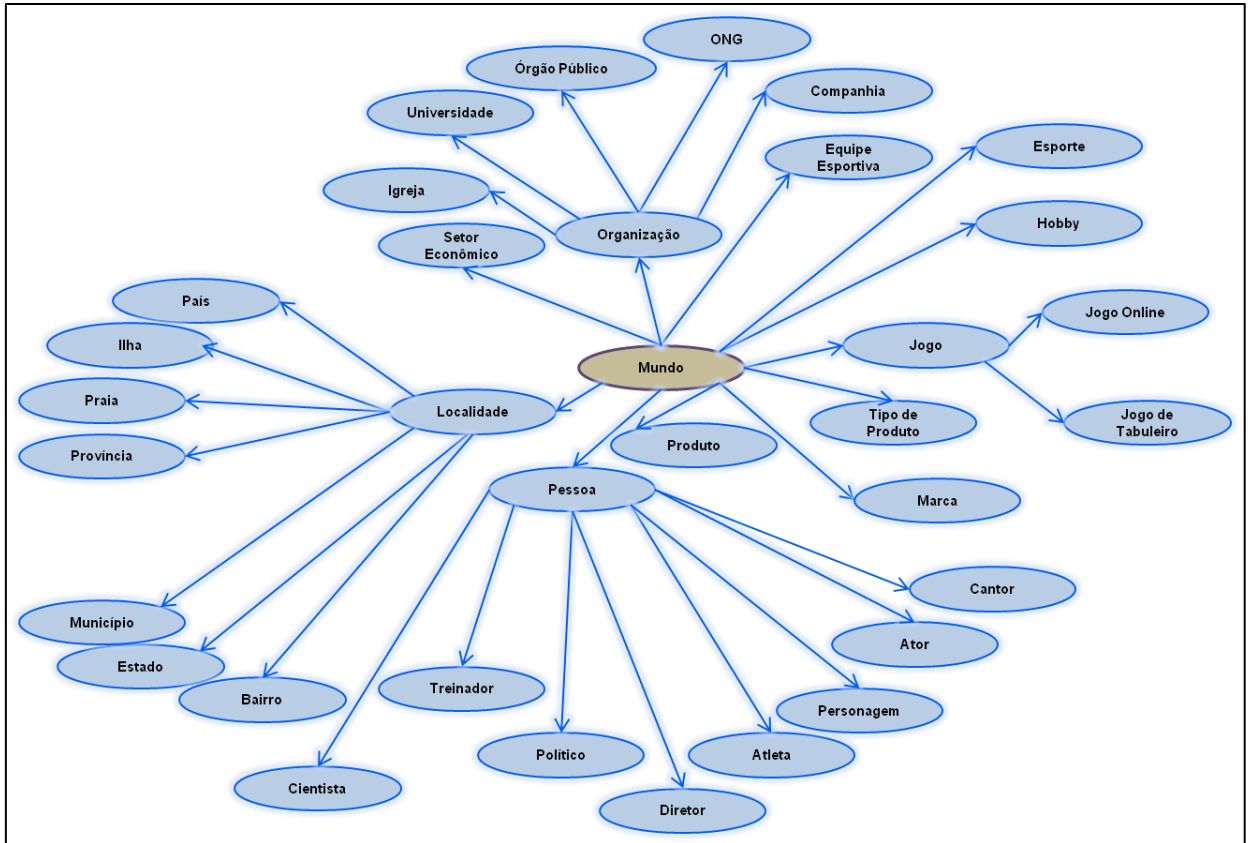


Figura 5.1 Base de Conhecimento - Nova Proposta

Como pode ser visto na Figura 5.1, existem mais categorias que estão ligadas à Companhia, e estas estão organizadas como: Organização, Igreja, Universidade, Governamental, ONG e Companhia. Igreja é mutuamente exclusiva com Universidade, Governamental e ONG, etc. Da mesma forma, a categoria Pessoa também conta agora com mais subcategorias derivadas, permitindo maior aprendizado. O mesmo aconteceu com Localidade (que antes nem existia) e Jogo.

Foi também adicionada a categoria Marca, para que haja melhor aprendizagem de produtos.

A nova organização proposta está detalhada na Tabela 5.1.

Tabela 5.1 Detalhamento de Nova Proposta da Base de Conhecimento

Grupo	Subgrupo	Subgrupo	Não é Mutuamente Exclusivo com
Organização	Igreja Universidade Órgão Público ONG		ONG Órgão Público, Companhia Companhia Pública Igreja
	Companhia	Companhia Privada Companhia Pública	Companhia Privada, Universidade Companhia Pública, Universidade
Localidade	País		Ilha
	Ilha		País
	Praia		
	Província		
	Cidade		
	Bairro Estado		
Setor Econômico			Tipo de Produto
Equipe Esportiva			
Esporte			Jogo de Tabuleiro, Hobby
Hobby			Esporte, Jogo de Tabuleiro, Jogo Online
Jogo	Jogo Online		
	Jogo de Tabuleiro		
Produto			Marca
Tipo de Produto			
Marca			Produto
Pessoa	Cientista		Treinador, Político, Diretor, Atleta, Personagem, Ator, Cantor
	Treinador		Cientista, Político, Diretor, Atleta, Personagem, Ator, Cantor
	Político		Cientista, Treinador, Diretor, Atleta, Personagem, Ator, Cantor
	Diretor		Cientista, Treinador, Político, Atleta, Personagem, Ator, Cantor
	Atleta		Cientista, Treinador, Político, Diretor, Personagem, Ator, Cantor
	Personagem		Cientista, Treinador, Político, Diretor, Atleta, Ator, Cantor
	Ator		Cientista, Treinador, Político, Diretor, Atleta, Personagem, Cantor
	Cantor		Cientista, Treinador, Político, Diretor, Atleta, Personagem, Ator

Na Tabela 5.1 a coluna *Grupo* significa o grupo maior, que engloba outras categorias que são *Subgrupos*. Ex: Jogo Online e Jogo de Tabuleiro são subgrupos de Jogo. A coluna “*Não é Mutuamente Exclusivo com*” mostra quais são as categorias que não são usadas para os padrões negativos. A organização de grupos é a seguinte: Subgrupo 1 é *pai* do Subgrupo 2 e ambos são filhos de Grupo.

Conforme apresentado na Tabela 5.1, os padrões negativos são todas as categorias que estão em grupos que são mutuamente exclusivos. Ex: O grupo de categorias Pessoa é mutuamente exclusivo ao grupo Jogo.

Com a implementação deste trabalho conclui-se também que a construção bem estruturada da base de conhecimento é muito importante para que o aprendizado obtenha bons resultados, esta precisa ser estudada e bem projetada de acordo com o problema proposto.

5.3 Trabalhos Futuros

Os trabalhos futuros propostos são os seguintes:

- Realizar análise de co-referência dos dados na base de conhecimento em crescimento.
- Melhorar os padrões prévios fracos (que extraíram poucos resultados).
- Buscar formas que melhorem o aprendizado de Relações Semânticas obtenha maior cobertura com melhor precisão.
- Propor, investigar e implementar métodos que formem PTs dinamicamente de acordo com particularidades da língua portuguesa que podem fazer com que a base cresça mais, por exemplo a flexibilidade de masculino e feminino em: “X ÉDonaDaEmpresa Y”, “X ÉDonoDaEmpresa Y”, em que X é proprietário/proprietária e Y é empresa.
- Realizar ajustes para que o RTWP possa estar ativo continuamente.

- A partir deste projeto apresentado planeja-se acoplá-lo com o aprendizado a partir de padrões HTML, que possui as seguintes tarefas:
 - Identificação e extração de ENs a partir de padrões HTML da Web;
 - Identificação e extração de Relações Semânticas entre ENs a partir de padrões HTML da Web.
- Investigar novas medidas para promoção de ENs e PTs que melhorem a cobertura e precisão do aprendizado. A nova ontologia proposta também será usada nos experimentos e será avaliada.
- Os resultados obtidos vieram diretamente da Web, porém isso demanda muito tempo para o pré-processamento do texto, por isso será criado um cópulus para a Web em português, que será a nova fonte de dados para este sistema.

Futuramente serão integrados o RTWP ao NELL, em que as bases de ambos os sistemas estarão vinculadas.

REFERÊNCIAS BIBLIOGRÁFICAS

- [1] ABNEY, S. Bootstrapping. In: ANNUAL MEETING ON ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 40., 2002, Philadelphia, Pennsylvania. **Proceedings** of the 51st Annual Meeting On Association For Computational Linguistics. Philadelphia: Association for Computational Linguistics, 2002. p. 360-367.
- [2] AGICHTEIN, E.; GRAVANO, L. Snowball: extracting relations from large plain-text collections. In: ACM CONFERENCE ON DIGITAL LIBRARIES, 15., 2000, San Antonio, Texas. **Proceedings** of the 5th ACM Conference On Digital Libraries. San Antonio: ACM, 2000. p. 85-94.
- [3] BALCAN, M.-F.; BLUM, A. A PAC-style Model for Learning from Labeled and Unlabeled Data. In: ANNUAL CONFERENCE ON COMPUTATIONAL LEARNING THEORY (COLT), 18., 2005, Bertinoro, Italie. **Proceedings** of the 18th Annual Conference On Computational Learning Theory (COLT). Berlin: Springer, 2005. p. 111-126.
- [4] BANKO, M.; CAFARELLA, M. J.; SODERLAND, S.; BROADHEAD, M.; ETZIONI, O. Open information extraction from the Web. In: INTERNATIONAL JOINT CONFERENCE ON ARTIFICIAL INTELLIGENCE, 20., 2007, Hyderabad, India. **Proceedings** of the 12th International Joint Conference On Artificial Intelligence. California: Morgan Kaufmann Publishers Inc., 2007. p. 2670-2676.
- [5] BANKO, M.; ETZIONI, O. The tradeoffs between open and traditional relation extraction. In: ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 46., 2008, Philadelphia. **Proceedings** of the 8th Annual Meeting Of The Association For Computational Linguistics. Philadelphia: Association for Computational Linguistics, 2008. p. 28-36.
- [6] BETTERIDGE, J.; CARLSON, A.; HONG, S. A. ; ESTEVAM R. HRUSCHKA, J.; LAW, EDITH L. M. ; MITCHELL, T.; WANG, S. Toward never ending language learning. In: AAAI 2009 SPRING SYMPOSIUM ON LEARNING BY READING AND LEARNING TO READ, 24., 2009 Palo Alto, Canada. **Proceedings** of the AAAI 2009 Spring Symposium On Learning By Reading And Learning To Read. Palo Alto: Association for the Advancement of Artificial Intelligence, 2009.
- [7] BETTERIDGE, J.; CARLSON, A.; ESTEVAM R. HRUSCHKA, J.; MITCHELL, T. M. Coupling semi-supervised learning of categories and relations. In: THE NAACL HLT 2009 WORKSHOP ON SEMI-SUPERVISED LEARNING FOR NATURAL LANGUAGE PROCESSING, 2009, Boulder, Colorado, USA. **Proceedings** of the The Naacl Hlt 2009 Workshop On Semi-Supervised

Learning For Natural Language Processing. Colorado: Association for Computational Linguistics, 2009. p. 1-9.

- [8] BLUM, A.; MITCHELL, T. Combining labeled and unlabeled data with co-training. In: ANNUAL CONFERENCE ON COMPUTATIONAL LEARNING THEORY (COLT), 11., 1998, Madison, Wisconsin, USA. **Proceedings** of the Annual Conference On Computational Learning Theory (Colt). Madison: ACM, 1998. p. 92-100.
- [9] 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, 26., 1999, Valencia, Spain. **Proceedings** of the Selected Papers From The International Workshop On The World Wide Web And Databases. London: Springer-Verlag, 1999. p. 172-183.
- [10] CAFARELLA, M. J.; DOWNEY, D.; SODERLAND, S.; ETZIONI, O. KnowItNow: fast, scalable information extraction from the Web. In: CONFERENCE ON HUMAN LANGUAGE TECHNOLOGY AND EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING, 2005, Vancouver, British Columbia, Canada. **Proceedings** of the Conference On Human Language Technology And Empirical Methods In Natural Language Processing. Stroudsburg, Pennsylvania, USA: Association for Computational Linguistics, 2005. p. 563-570.
- [11] CAFARELLA, M. J.; HALEVY, A.; WANG, D. Z.; WU, E.; ZHANG, Y. WebTables: exploring the power of tables on the Web. **Proceedings of the Very Large Data Base Endow.**, v. 1, n. 1, p. 538-549, 2008.
- [12] CAFARELLA, M. J.; MADHAVAN, J.; HALEVY, A. Web-scale extraction of structured data. **Special Interest Group on Management Of Data Rec.**, v. 37, n. 4, p. 55-61, 2008.
- [13] CARLSON, A. **Coupled Semi-Supervised Learning**. 159 f. PhD. Thesis – School of Computer Science, Carnegie Mellon University, Pittsburgh, USA, 2010.
- [14] CHANG, M.-W.; RATINOV, L.; ROTH, D. Guiding Semi-Supervision with Constraint-Driven Learning. In: ANNUAL MEETING OF THE ASSOCIATION OF COMPUTATIONAL LINGUISTICS, 45., 2007, Prague, Czech Republic. **Proceedings** of the Annual Meeting Of The Association Of Computational Linguistics. Prague: Association for Computational Linguistics C1, 2007. p. 280-287.
- [15] COHEN, W. W.; HURST, M.; JENSEN, L. S. A flexible learning system for wrapping tables and lists in HTML documents. In: INTERNATIONAL CONFERENCE ON WORLD WIDE WEB, 11., 2002, Honolulu, Hawaii, USA. **Proceedings** of the International Conference On World Wide Web. Honolulu: ACM, 2002. p. 232-241.

- [16] COLLINS, M.; SINGER, Y. Unsupervised models for named entity classification. In: JOINT SIGDAT CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING AND VERY LARGE CORPORA, 1999, Singapore, **Proceedings...** Singapore: 1999. p. 100-110.
- [17] CURRAN, J. R.; MURPHY, T.; SCHOLZ, B. Minimising semantic drift with Mutual Exclusion Bootstrapping. In: PACIFIC ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 10., 2007, Melbourne, Australia. **Proceedings** of the Pacific Association For Computational Linguistics. Melbourne: 2007. p. 172–180.
- [18] DURME, B. V.; PASCA, M. Finding cars, goddesses and enzymes: parametrizable acquisition of labeled instances for open-domain information extraction. In: NATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE, 23., 2008, Chicago, Illinois, USA. **Proceedings** of the National Conference On Artificial Intelligence. Chicago: AAAI Press, 2008. p. 1243-1248.
- [19] ETZIONI, O.; CAFARELLA, M.; DOWNEY, D.; POPESCU, A.-M.; SHAKED, T.; SODERLAND, S.; WELD, D. S.; YATES, A. Unsupervised named-entity extraction from the Web: An experimental study. **Artificial Intelligence**, v. 165, n. 1, p. 91-134, 2005.
- [20] HARRIS, Z. Distributional structure. **Word**, v. 10, n. 23, p. 146-162, 1954.
- [21] HEARST, M. A. Automatic acquisition of hyponyms from large text corpora. In: CONFERENCE ON COMPUTATIONAL LINGUISTICS, 14., 1992, Nantes, France. **Proceedings** of the Conference On Computational Linguistics. Nantes: Association for Computational Linguistics, 1992. p. 539-545.
- [22] HINDLE, D. Noun classification from predicate-argument structures. In: ANNUAL MEETING ON ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 28., 1990, Singapore. **Proceedings** of the Annual Meeting On Association For Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 1990. p. 268-275.
- [23] HOVY, E.; KOZAREVA, Z.; RILOFF, E. Toward completeness in concept extraction and classification. In: CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING, 2009, Singapore. **Proceedings** of the Conference On Empirical Methods In Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2009. p. 948-957.
- [24] JØRGENSEN, B. Exponential dispersion models (with discussion). **Journal of the Royal Statistical Society, Series B**, v. 49, n. 2, p. 127-162, 1987.
- [25] KUSHMERICK, N. **Wrapper induction for information extraction**, 246 f. PhD Thesis - Department of Computer Science & Engineering University of Washington, Washington, USA, 1997.

- [26] MADHAVAN, J.; KO, D.; KOT, Ł.; GANAPATHY, V.; RASMUSSEN, A.; HALEVY, A. Google's Deep Web crawl. **Proceedings VLDB Endow.**, v. 1, n. 2, p. 1241-1252, 2008.
- [27] MANN, G. S.; MCCALLUM, A. Simple, robust, scalable semi-supervised learning via expectation regularization. In: INTERNATIONAL CONFERENCE ON MACHINE LEARNING, 24., 2007, Corvallis, Oregon, USA. **Proceedings** of the International Conference On Machine Learning. Corvallis: Association for Computing Machinery, 2007. p. 593-600.
- [28] MCCLOSKEY, D.; CHARNIAK, E.; JOHNSON, M. Effective self-training for parsing. In: HUMAN LANGUAGE TECHNOLOGY CONFERENCE OF THE NAACL, MAIN CONFERENCE, 2006, New York, USA. **Proceedings** of the Human Language Technology Conference Of The Naacl, Main Conference. Stroudsburg: Association for Computational Linguistics, 2006. p. 152-159.
- [29] MITCHELL, T. M.; BETTERIDGE, J.; CARLSON, A.; HRUSCHKA, E.; WANG, R. Populating the semantic web by macro-reading internet text. In: INTERNATIONAL SEMANTIC WEB CONFERENCE, 8., 2009, Chantilly, Virginia, USA. **Proceedings** of the International Semantic Web Conference. Chantilly: Springer-Verlag, 2009. p. 998-1002.
- [30] MITCHELL, T. M. **Machine Learning**. New York, USA: McGraw-Hill, 1997. 432 p.
- [31] MITCHELL, T. M. **The discipline of machine learning**, 2006. Disponível em: <<http://www.cs.cmu.edu/~tom/pubs/MachineLearning.pdf>>. Acesso em: 24/01/2010.
- [32] NADEAU, D.; SEKINE, S. A survey of named entity recognition and classification. **Journal of Linguisticae Investigationes**, v. 30, n. 1, p. 1-20, 2007.
- [33] PANTEL, P.; CRESTAN, E.; BORKOVSKY, A.; POPESCU, A.-M.; VYAS, V. Web-scale distributional similarity and entity set expansion. In: CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING, 2009, Singapore. **Proceedings** of the Conference On Empirical Methods In Natural Language Processing. Singapore: Association for Computational Linguistics, 2009. p. 938-947.
- [34] PASCA, M.; LIN, D.; BIGHAM, J.; LIFCHITS, A.; JAIN, A. Organizing and searching the world wide web of facts - step one: the one-million fact extraction challenge. In: NATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE, 21., 2006, Boston, Massachusetts, USA. **Proceedings** of the National Conference On Artificial Intelligence. Boston: Association for the Advancement of Artificial Intelligence Press, 2006. p. 1400-1405.
- [35] RILOFF, E.; JONES, R. Learning dictionaries for information extraction by multi-level bootstrapping. In: NATIONAL CONFERENCE ON ARTIFICIAL

- INTELLIGENCE AND THE ELEVENTH INNOVATIVE APPLICATIONS OF ARTIFICIAL INTELLIGENCE, 16., 1999, Orlando, Florida, USA. **Proceedings** of the National Conference On Artificial Intelligence And The Eleventh Innovative Applications Of Artificial Intelligence. Orlando: American Association for Artificial Intelligence, 1999. p. 474-479.
- [36] SARAWAGI, S. Information Extraction. **Found. Trends databases**, v. 1, n. 3, p. 261-377, 2008.
- [37] SINDHWANI, V.; NIYOGI, P.; BELKIN, M. A co-regularization approach to semi-supervised learning with multiple views. In: WORKSHOP ON LEARNING WITH MULTIPLE VIEWS, 22., 2005, Bonn, Germany. **Proceedings** of the Workshop On Learning With Multiple Views. Bonn: International Conference on Machine Learning, 2005. p. 824-831.
- [38] SMITH, D. A.; EISNER, J. Bootstrapping feature-rich dependency parsers with entropic priors. In: JOINT CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING AND COMPUTATIONAL NATURAL LANGUAGE LEARNING (EMNLP-CoNLL), 2007, Prague, Czech Republic. **Proceedings** of the JOINT CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING AND COMPUTATIONAL NATURAL LANGUAGE LEARNING (EMNLP-Conll). Prague: Association for Computational Linguistics, 2007. p. 667-677.
- [39] UEFFING, N. Self-training for machine translation. In: NIPS WORKSHOP ON MACHINE LEARNING FOR MULTILINGUAL INFORMATION ACCESS, 2006, Whistler, British Columbia, Canada. **Proceedings** of the Nips Workshop On Machine Learning For Multilingual Information Access. Whistler: Machine Learning For Multilingual Information Access, 2006.
- [40] WANG, R. C.; COHEN, W. W. Language-independent set expansion of named entities using the web. In: IEEE INTERNATIONAL CONFERENCE ON DATA MINING, 7., 2007, Omaha, Nebraska, USA. **Proceedings** of the IEEE International Conference On Data Mining. Nebraska: IEEE Computer Society, 2007. p. 342-350.
- [41] WEISS, S. M.; KULIKOWSKI, C. A. **Computer systems that learn: classification and prediction methods from statistics, neural nets, machine learning, and expert systems**. San Fransisco, California, USA: Morgan Kaufmann Publishers, 1991. 223 p.
- [42] WHITELAW, C.; KEHLENBECK, A.; PETROVIC, N.; UNGAR, L. **Web-scale named entity recognition**. Napa Valley, California, USA: ACM, 2008. p. 123-132.
- [43] YANGARBER, R. Counter-training in discovery of semantic patterns. In: ANNUAL MEETING ON ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 41., 2003, Morristown, NJ, USA. **Proceedings** of the Annual

Meeting On Association For Computational Linguistics. Morristown: Association for Computational Linguistics, 2003. p. 343-350.

- [44] YAROWSKY, D. Unsupervised word sense disambiguation rivaling supervised methods. In: ANNUAL MEETING ON ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 33., 1995, Cambridge, Massachusetts, USA. **Proceedings** of the Annual Meeting on Association For Computational Linguistics. Cambridge: Association for Computational Linguistics, 1995. p. 189-196.

- [45] YATES, A.; ETZIONI, O. Unsupervised resolution of objects and relations on the web. In: ANNUAL CONFERENCE OF THE NORTH AMERICAN CHAPTER OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 2007, Rochester, New York, USA. **Proceedings** of the Annual Conference of The North American Chapter Of The Association For Computational Linguistics. Rochester: Association for Computational Linguistics, 2007. p. 121-130.

- [46] ZHU, X.; GOLDBERG, A. B.; BRACHMAN, R.; DIETTERICH, T. **Introduction to Semi-Supervised Learning**. San Rafael, California, USA: Morgan and Claypool Publishers, 2009. 130 p.

- [47] ZHU, X.; GHAHRAMANI, Z.; LAFFERTY, J. Semi-supervised learning using Gaussian fields and harmonic functions. In: MACHINE LEARNING INTERNATIONAL WORKSHOP, 20., 2003, Washington, District of Columbia, USA. **Proceedings** of the Machine Learning International Workshop. Washington: International Conference on Machine Learning, 2003. p. 912-919.