

**UNIVERSIDADE FEDERAL DE SÃO CARLOS**  
CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA  
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

**LEITURA DA WEB EM PORTUGUÊS EM  
AMBIENTE DE APRENDIZADO SEM-FIM**

**MAÍSA CRISTINA DUARTE**

**ORIENTADOR: PROF. DR. ESTEVAM R. HRUSCHKA JR.**

São Carlos – SP  
Janeiro/2016

**UNIVERSIDADE FEDERAL DE SÃO CARLOS**

CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

**LEITURA DA WEB EM PORTUGUÊS EM  
AMBIENTE DE APRENDIZADO SEM-FIM**

**MAÍSA CRISTINA DUARTE**

Tese apresentada ao Programa de Pós-Graduação em  
Ciência da Computação da Universidade Federal de  
São Carlos, como parte dos requisitos para a ob-  
tenção do título de Doutora em Ciência da Compu-  
tação, área de concentração: Aprendizado de Má-  
quina/Inteligência Artificial

Orientador: Prof. Dr. Estevam R. Hruschka Jr.

São Carlos – SP

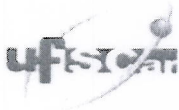
Janeiro/2016

Ficha catalográfica elaborada pelo DePT da Biblioteca Comunitária UFSCar  
Processamento Técnico  
com os dados fornecidos pelo(a) autor(a)

D812L Duarte, Maísa Cristina  
Leitura da web em português em ambiente de  
aprendizado sem-fim / Maísa Cristina Duarte. -- São  
Carlos : UFSCar, 2016.  
78 p.

Tese (Doutorado) -- Universidade Federal de São  
Carlos, 2016.

1. Aprendizado de máquina semissupervisionado. 2.  
Aprendizado sem-fim. 3. NELL. 4. Acoplamento. 5.  
Resolução de correferência. I. Título.



UNIVERSIDADE FEDERAL DE SAO CARLOS

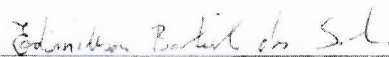
Centro de Ciências Exatas e de Tecnologia  
Programa de Pós-Graduação em Ciência da Computação


Folha de Aprovação


Assinaturas dos membros da comissão examinadora que avaliou e aprovou a Defesa de Tese de Doutorado da candidata  
Maísa Cristina Duarte realizada em 04/01/2016

  
Prof. Dr. Estevam Rafael Hruschka Junior  
UFSCar

  
Prof. Dr. Daniel Lucrécio  
UFSCar

  
Prof. Dr. Edmilson Batista dos Santos  
UFSJ

  
Prof. Dr. João Gama  
UP

  
Prof. Dr. Nelson Francisco Favilla Ebecken  
UFRJ

*Aos meus pais*

## AGRADECIMENTOS

Resumir em somente uma página meus agradecimentos é muito injusto, entretanto mesmo se houvesse um espaço maior também seria impossível agradecer a todos como merecem. Várias pessoas me apoiaram desde o início deste projeto, algumas de bem pertinho, outras de um pouco mais distante, mas todas participaram de forma ativa e importante.

Dentre tudo o que esse período proporcionou, as amizades que surgiram possuem o maior valor, sobretudo de meu orientador e amigo Estevam Hruschka. Obrigada por me orientar desde o mestrado, por me apoiar em todas as dificuldades, por ser meu modelo pessoal e profissional, por me ensinar pelo exemplo, por ser sempre prestativo, e, principalmente, por me ensinar a ver um pouco do mundo com seus olhos. Você me fez descobrir minha vocação.

Obrigada a outra grande amizade conquistada, Maria do Carmo, que me recebeu de braços abertos em São Carlos e sempre me apoiou. Você foi o motivo da minha escolha ser São Carlos e UFSCar, agradeço-lhe imensamente por me direcionar a um caminho que me faz tão feliz!

Obrigada a toda minha família, namorado e amigos pelo apoio, atenção, orações e compreensão. Sem vocês nada seria possível, nem mesmo respirar! Vocês são a razão do meu empenho.

Obrigada a todos membros do MaLL (Machine Learning Laboratory) pelo companheirismo e compartilhamento de conhecimento.

Obrigada à toda equipe da NELL na CMU por todo suporte. Em especial agradeço a Bryan Kisiel pelo suporte técnico e paciência durante todo o projeto, desde o mestrado! Apesar de não nos conhecermos pessoalmente, compartilho com você grande parte de tudo que atingimos!

Agradeço também a todos funcionários (e amigos!) do Departamento de Computação da UFSCar (DC), em especial à Cristina Trevelin, Augusto César e Ivan Rogério. Obrigada pela paciência e por sempre estarem prontos a resolver qualquer problema!

Todos vocês estão para sempre em meu coração! Obrigada imensamente a cada um de vocês!

*Se “ aprender é a única coisa de que a mente nunca se cansa, nunca tem medo e nunca se arrepende. ” (Leonardo da Vinci), imagine uma máquina aprendendo!*

## RESUMO

A NELL é um sistema de computador que possui o objetivo de executar 24 horas por dia, 7 dias por semana, sem parar. A versão atual da NELL foi iniciada em 12 de Janeiro de 2010 e continua ativa. Seu objetivo é aprender cada vez mais fatos da web para popular sua base de conhecimento (Knowlegde Base - KB). Além de aprender cada vez mais, a NELL também objetiva alcançar alta confiança no aprendizado para garantir a continuidade do aprendizado.

A NELL foi desenvolvida e atua no contexto da macroleitura, no qual é necessária uma grande quantidade e redundância de dados. Para que o sistema possa aprender, o primeiro passo é criar uma base preprocessada (*all-pairs-data*) a partir do uso de técnicas linguísticas. O *all-pairs-data* deve possuir todas as estatísticas suficientes para a execução da NELL e também deve ser de um tamanho suficientemente grande para que o aprendizado possa ocorrer.

Neste projeto, foi proposta a criação de uma nova instância da NELL em português. Inicialmente foi proposta a criação de um *all-pairs-data* e, em seguida, a criação de uma abordagem híbrida para a resolução de correferências independente de língua por base em características semânticas e morfológicas. A proposta híbrida objetivou aperfeiçoar o processo atual de tratamento de correferências na NELL, melhorando assim a confiabilidade no aprendizado.

Todas as propostas foram desenvolvidas e a NELL em português obteve bons resultados. Tais resultados evidenciam que a leitura da web em português poderá se tornar um sistema de aprendizado sem-fim. Para que isso ocorra são também apresentadas as futuras abordagens e propostas.

Além disso, este projeto apresenta a metodologia de criação da instância da NELL em português, uma proposta de resolução de correferência que explora atributos linguísticos, bem como a ontologia da NELL, além de apontar trabalhos futuros, nos quais inclui-se processos de adição de outras línguas na NELL, pricipalmente para aquelas que possuem poucas páginas web disponíveis para o aprendizado.

**Palavras-chave:** Aprendizado de Máquina Semisupervisionado, Aprendizado Sem-Fim, NELL, Acomodamento, Resolução de Correferência, Leitura da Web em Português



## ABSTRACT

NELL is a computer system that has the goal of learn to learn 24 hours per day, continuously and learn more an better than the last day, to perform the knowledge base (KB). NELL is running since January 12 of 2010. Furthermore, NELL goals is have hight precision to be able to continue the learning.

NELL is developed in macro-reading context, because this NELL needs very much redundancy to run. The first step to run NELL is to have an big (*all-pairs-data*). An *all-pairs-data* is a preprocessed base using Natural Language Processing (NLP), that base has all sufficient statistics about a *corpus* of web pages.

The proposal of this project was to create a instance of NELL (currently in English) in Portuguese. For this, the first goal was the developing an *all-pairs-data* in Portuguese. The second step was to create a new version of Portuguese NELL. And finally, the third goal was to develop a coreference resolution hybrid method focused in features semantics and morphologics. This method is not dependent of a specific language, it is can be applied for another languages with the same alphabet of Portuguese language.

The NELL in Portuguese was developed, but the *all-pairs-data* is not big enough. Because it Portuguese NELL is not running for ever, like the English version. Even so, this project present the steps about how to develop a NELL in other language and some ideas about how to improve the *all-pairs-data*. By the way, this project present a coreference resolution hybrid method with good results to NELL.

**Keywords:** Semi-Supervised Learning, Never-Ending Learning, NELL, Coupling, Correference Resolution, Read The Web in Portuguese

## LISTA DE FIGURAS

1.1	Exemplo de um subconjunto da ontologia da NELL. . . . .	13
2.1	Arquitetura básica NELL. Figura inspirada em (CARLSON et al., 2010a). . . . .	20
2.2	Aprendizado de Categorias - Categoria cidade . . . . .	23
2.3	Aprendizado de Relações - Relação localizadaEm . . . . .	24
3.1	Instâncias aprendidas corretamente com e sem o uso de sementes de PTs. . . . .	43
A.1	Fluxograma de processo na extração de categorias . . . . .	71
A.2	Fluxograma de processo na extração de relações . . . . .	74
A.3	Implementação em Java do <i>Map</i> (Hadoop) para categorias . . . . .	75
A.4	Implementação em Java do <i>Reduce</i> ( <i>Hadoop</i> ) para categorias . . . . .	76
A.5	Implementação em Java do Map (Hadoop) para relações semânticas . . . . .	77

## LISTA DE TABELAS

2.1	Arquitetura básica NELL. Figura adaptada e inspirada em (VIEIRA, 2015). . . . .	22
2.2	Instâncias de relações usadas como características semânticas para resolução de correferências. . . . .	25
2.3	Exemplo de <i>slots</i> vazios que podem ocorrer com instâncias de relações usadas como características semânticas para a resolução de correferência. . . . .	26
2.4	Número de páginas do ClueWeb09. Fonte: <a href="http://boston.lti.cs.cmu.edu/clueweb09">http://boston.lti.cs.cmu.edu/clueweb09</a>	29
3.1	Resultados do primeiro experimento - Sem adição de novas sementes de PT. CI (Correct Instances): Número de Instâncias Corretas; LI (Learned Instances): Número de Instâncias Aprendidas . . . . .	41
3.2	Resultados do primeiro experimento - Com adição de novas sementes de PTs. CI (Correct Instances): Número de Instâncias Corretas; LI (Learned Instances): Número de Instâncias Aprendidas . . . . .	42
3.3	Leitura da Web em Português com Supervisão Humana para categorias - Resultados Cumulativos . . . . .	46
3.4	Leitura da Web em Português com Supervisão Humana para relações semânticas - Resultados Cumulativos . . . . .	47
3.5	Instâncias de relações usadas como características semânticas para a resolução de correferência. . . . .	50
3.6	Sumário dos resultados empíricos para as 3 configurações de experimentos. Os resultados são as médias obtidas utilizando o 10-fold cross-validation. Os melhores resultados estão em negrito . . . . .	53

# SUMÁRIO

<b>CAPÍTULO 1 – INTRODUÇÃO</b>	<b>11</b>
1.1 Motivação e Justificativa . . . . .	14
1.2 Objetivos da Pesquisa . . . . .	14
1.3 Organização do Trabalho . . . . .	16
<b>CAPÍTULO 2 – FUNDAMENTAÇÃO TEÓRICA E TRABALHOS CORRELATOS</b>	<b>17</b>
2.1 Tipos de Aprendizado . . . . .	17
2.1.1 Aprendizado Supervisionado . . . . .	17
2.1.2 Aprendizado Não Supervisionado . . . . .	18
2.1.3 Aprendizado Semissupervisionado . . . . .	18
2.2 Never-Ending Language Learning - NELL . . . . .	19
2.3 Características Linguísticas . . . . .	26
2.4 Corpora Importantes . . . . .	28
2.5 Início da Leitura da Web em Português . . . . .	30
2.6 Trabalhos Correlatos de Resolução de Correferência . . . . .	32
<b>CAPÍTULO 3 – A LEITURA DA WEB EM PORTUGUÊS E A RESOLUÇÃO DE CORREFERÊNCIA INDEPENDENTE DE LÍNGUA</b>	<b>35</b>
3.1 RWTP & NELL . . . . .	35
3.1.1 All-Pairs-Data . . . . .	36
3.1.2 Criação de Instância do Sistema NELL para o Português e sua Ontologia	38

3.1.3	Experimentos e Análise . . . . .	40
3.2	Leitura da Web em Português: A NELL em Português . . . . .	43
3.2.1	Experimentos e Análise . . . . .	45
3.3	Resolução de Correferência independente de língua na NELL . . . . .	48
3.3.1	Experimentos e Análise . . . . .	52
<b>CAPÍTULO 4 – CONCLUSÕES</b>		<b>54</b>
4.1	Objetivos Alcançados . . . . .	55
4.2	Contribuições e Limitações . . . . .	55
4.3	Trabalhos Futuros . . . . .	57
<b>REFERÊNCIAS</b>		<b>58</b>
<b>GLOSSÁRIO</b>		<b>65</b>
<b>CAPÍTULO A –PRÉ-PROCESSAMENTO E CRIAÇÃO DE ALL-PAIRS-DATA A PARTIR DO CLUEWEB</b>		<b>68</b>
A.1	Contextualização . . . . .	68
A.2	<i>Corpus</i> ClueWeb - Coleta de Páginas Web . . . . .	68
A.3	Desenvolvimento do <i>All-Pairs-Data</i> do <i>corpus</i> ClueWeb . . . . .	70
A.4	Fluxogramas de processo de execução do pré-processamento do ClueWeb . . .	70
A.4.1	Fluxograma de processo para as categorias . . . . .	71
A.4.2	Fluxograma de Relações Semânticas . . . . .	73
A.5	Uso do Hadoop . . . . .	75

# Capítulo 1

## INTRODUÇÃO

---

---

Aprendizado de Máquina (AM) é uma subárea de Inteligência Artificial (IA) com foco em algoritmos, métodos e abordagens, os quais visam permitir que sistemas de computador aprendam com a experiência.

Nos últimos anos, AM tem sido usado em muitas aplicações de diferentes domínios, resultando em várias experiências bem sucedidas. A maioria das histórias de sucesso são, no entanto, restritas a abordagens em que uma única função alvo é definida, sendo o objetivo principal aproximar um modelo à esta função.

Tais abordagens podem ser exemplificadas por métodos supervisionados modelados para detectar *spam*, transações financeiras fraudulentas, realizar predições de diagnósticos médicos e previsão do tempo. Em todos esses exemplos, normalmente uma simples função objetivo é definida, e o modelo é construído para se aproximar desta função.

Existem, entretanto, muitos domínios complexos em que essa tradicional abordagem de aproximação de uma função objetivo tende a não obter bons resultados. Um exemplo é o domínio chamado *Machine Reading* (MR).

MR pode ser descrita como uma área de pesquisa com foco em Compreensão da Linguagem Natural (CLN), a qual vai além de Processamento de Linguagem Natural (PLN). Seguindo a ideia apresentada em (ETZIONI; BANKO; CAFARELLA, 2007), o objetivo principal de MR é "*o entendimento autônomo do texto*".

Considerando que um dos mais importantes métodos pelos quais o ser humano aprende é pela leitura, muitas iniciativas, bem como projetos de pesquisa, têm dedicado suas investigações a fim de construir máquinas capazes de aprender a partir de leitura (CLARK et al., 2007).

Para alcançar bons resultados em sistemas de MR, uma alternativa ao tradicional método de

aproximação da função objetivo é o never-ending learning (NEL), também chamado de aprendizado de máquina sem-fim (AMSF), i.e., um paradigma de aprendizado no qual o *aprendiz*, autonomamente, consegue evoluir de maneira contínua e incremental ao longo do tempo.

Mais importante que apenas continuar a evoluir, nesse novo paradigma, um conhecimento adquirido pode, de forma dinâmica, ser usado para expandir o escopo e melhorar o desempenho da tarefa de aprendizado como um todo. Em outras palavras, em uma abordagem de aprendizado de máquina sem-fim o *aprendiz* deve, constantemente e autonomamente, aprender a aprender mais e melhor a cada nova leitura.

Considerando suas capacidades de aprendizado, o NEL pode ser visto como um candidato a ser usado em domínios complexos, como em MR. Não coincidentemente, MR é o domínio de aplicação do primeiro sistema de aprendizado de máquina sem-fim, apresentado em (CARLSON et al., 2010a) e chamado de NELL (Never-Ending Language Learning).

A NELL é um sistema de computador que possui o objetivo de executar 24 horas por dia, 7 dias por semana, sem parar. A versão atual da NELL foi iniciada em 12 de Janeiro de 2010 e continua ativa, objetivando coletar mais e mais fatos a partir da web para aumentar e popular a sua base de conhecimento (*Knowledge Base - KB*).

Resumidamente, a KB inicial da NELL é organizada na forma de uma ontologia, na qual foram inseridas centenas de categorias e relações. As categorias são os tipos de conhecimento e.g.: pessoa, equipeEsportiva, fruta, emoção, etc. Já as relações são os relacionamentos entre as categorias, e.g., atletaJogaParaEquipeEsportiva(atleta, equipeEsportiva), musicoTocaInstrumento(músico, instrumento), etc.

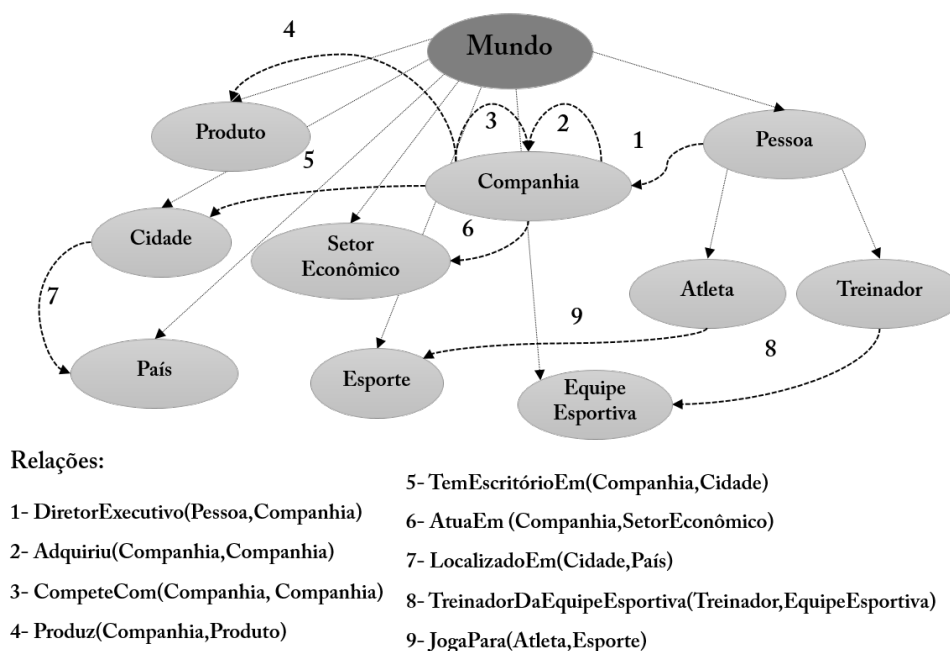
Na Figura 1.1 é apresentado um pequeno subconjunto da ontologia atual da NELL. Os círculos referem-se às categorias enquanto os números às relações entre as categorias.

Para que a NELL inicie o aprendizado, a ontologia também deve possuir exemplos de instâncias (sementes) para cada uma das categorias, e.g., pessoa(Angelina Jolie), fruta(morango), etc. E da mesma forma, deve possuir exemplos de cada uma das relações, e.g., atletaJogaParaEquipeEsportiva(Kobe Bryant, LA Lakers), musicoTocaInstrumento(Eric Clapton, guitarra), etc.

Trabalhos relatados<sup>1</sup> até o momento revelam que bons resultados têm sido obtidos quando é realizada a Leitura da Web em Inglês. Quando realizada a tentativa de execução da mesma tarefa, mas aplicada a páginas escritas em português, os resultados relatados em (DUARTE; NICOLETTE; HRUSCHKA, 2011) e (HRUSCHKA; DUARTE; NICOLETTI, 2013) não mantiveram o de-

---

<sup>1</sup><http://rtw.ml.cmu.edu/rtw/publications>



**Figura 1.1: Exemplo de um subconjunto da ontologia da NELL.**

sempenho tão bom quanto o aprendido em inglês.

Um dos componentes principais da NELL, ilustrado na Figura 2.1, é o *ConceptResolver* (KRISHNAMURTHY; MITCHELL, 2011), projetado para atuar na resolução de correferência a partir de características semânticas, as quais são extraídas das relações aprendidas na KB.

Neste projeto, resolução de correferência refere-se a duas ou mais instâncias que possuem o mesmo significado. Por exemplo Florianópolis e Floripa; São Carlos e Sanca; Nova York, NY e New York, etc.

O *ConceptResolver* lida com polissemia<sup>2</sup> e com resolução de sinonímia<sup>3</sup> a fim de identificar conceitos latentes aos quais um *noun phrase* (sintagma nominal) se refere. Basicamente, o componente executa um processo de *word sense disambiguation* (busca identificar palavras que possuam o mesmo significado) e aplica a resolução de sinonímia sobre as relações extraídas do texto. Para isso, o *ConceptResolver* usa a ontologia da NELL e uma pequena quantidade de dados rotulados.

Um problema de difícil tratamento para o *ConceptResolver* é a ocorrência de várias relações vazias, ou seja, não populadas, dentre o grupo selecionado para a análise. Em outras palavras, o *ConceptResolver* é fortemente vinculado às características semânticas, as quais dependem unicamente do aprendizado de relações. Por isso, quando poucas relações de um conceito são

<sup>2</sup>Instâncias iguais ou muito parecidas, mas com significados diferentes. Exemplo: apple (fruta) e Apple (empresa)

<sup>3</sup>Instâncias sinônima. Exemplo: Sampa e São Paulo



aprendidas, mesmo que haja a correferência, o componente não conseguirá identificá-la.

## 1.1 Motivação e Justificativa

A principal motivação deste trabalho foi desenvolver por completo a leitura da web em português em ambiente de aprendizado de máquina sem-fim. A insistência em melhorar trabalhos anteriores é justificada devido à importância da incorporação de outras línguas ao projeto NELL (o português é a primeira língua, após o inglês), visando, assim, a melhoria do aprendizado de linguagens a partir da web.

Outra motivação foi a de melhorar a acurácia do aprendizado da NELL de forma independente de língua. Para isso, foi proposto o tratamento de correferência a partir de categorias. A justificativa é que o atual componente, o *ConceptResolver*, trabalha diretamente vinculado somente a relações e, por isso, existe a dificuldade na resolução de correferência quando relações vazias (relações não aprendidas/não populadas) são analisadas em um grupo de clusterização. E no caso da NELL, relações vazias são comumente encontradas.

## 1.2 Objetivos da Pesquisa

O objetivo principal da pesquisa descrita neste documento é investigar, propor e implementar métodos e algoritmos que possibilitassem a execução da NELL em português, e que possibilitassem também contribuir para o processo de resolução de correferência originalmente presente no sistema de aprendizado sem-fim.

Como primeiro objetivo específico, tem-se a criação de um conjunto de dados pré-processado, chamado *all-pairs-data*. Em seguida, tem-se como segundo objetivo específico, a criação de uma nova instância do sistema NELL para o português.

O terceiro objetivo específico é propor um método de resolução de correferência de forma independente de língua e que tome como entrada resultados gerados pelo *ConceptResolver*. Além da melhoria da acurácia do sistema, tal método visa diminuir uma dificuldade do componente *ConceptResolver* da NELL: a resolução de correferência quando existem poucas relações aprendidas.

No terceiro objetivo específico, a independência de língua é abordada como a independência entre línguas que possuam alfabeto similar ao português.

A proposta para o terceiro objetivo foi criar um método que pudesse relacionar caracterís-

ticas semânticas, já abordadas pelo *ConceptResolver*, e características morfológicas, que ainda não são abordadas. O objetivo, aqui, foi o de melhorar o desempenho da resolução de correferência em conjunto com a abordagem do *ConceptResolver*. Nesse caso, desempenho significa melhorar a confiabilidade e a cobertura da KB da NELL.

O objeto de unir a abordagem de tratamento de correferência baseado em relações semânticas ao tratamento de correferência baseado em categorias foi abordado devido à dificuldade em se aprender relações semânticas e à facilidade em se aprender categorias pela NELL. As categorias possuem menos chances de estarem vazias, porém são menos assertivas, diferentemente das relações semânticas, que são mais assertivas, mas podem estar vazias mais frequentemente. Sendo assim, a união de ambas as abordagens visou unir o ganho de cada uma dessas visões independentes. A assertividade refere-se, neste projeto, à capacidade de aprender corretamente. Quanto maior a possibilidade do aprendizado ser correto, mais assertivo ele é.

Ao se ter atingido o primeiro e o segundo objetivos específicos foi possível a execução da NELL em português, enquanto que o sucesso na busca pelo terceiro objetivo específico permitiu a melhoria da acurácia dos fatos aprendidos de forma independente de língua.

Resumidamente, a NELL é capaz de, a partir de uma KB inicial, ser executada continuamente com dois objetivos específicos:

- **Extração:** extrair mais conhecimento a partir da Web em português, visando a expansão da KB inicial;
- **Aprendizado:** aprender a extrair melhor e com mais precisão que no "*dia anterior*".

A instância da NELL criada para o português foi avaliada 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, "*extração*" e "*aprendizado*".

A partir de uma avaliação manual, foi possível validar empiricamente a execução da leitura da web em português no contexto do aprendizado de máquina sem-fim.

A avaliação foi realizada de forma manual e empírica. A partir dos resultados obtidos, foi possível concluir que o uso do método proposto para o tratamento de correferência com base em categorias possibilita a melhoria de acurácia na NELL.

Além disso, o método proposto para o tratamento de correferência foi desenvolvido de forma independentemente de língua, podendo assim ser usado para outras línguas.

## 1.3 Organização do Trabalho

Este documento é organizado de forma a apresentar inicialmente os objetivos, as motivações e as justificativas e, em seguida, a trajetória realizada para alcançar cada ponto ressaltado na proposta.

Para isso, no *Capítulo 2* é abordada a fundamentação teórica, base para entendimento do projeto, juntamente com os trabalhos correlatos mais importantes vinculados aos objetivos abordados. Inicialmente, são abordados os tipos de aprendizado de máquina e a NELL, em seguida, algumas características linguísticas e alguns *corpora* importantes para a definição da trajetória da NELL em português. Por fim, são apontados os trabalhos correlatos desenvolvidos no início do desenvolvimento da NELL em português, o tratamento de correferências atual da NELL e outros trabalhos encontrados na literatura voltados ao tratamento de correferências.

No *Capítulo 3* é relatada toda a trajetória da NELL em português e a resolução de correferências independente de língua. Para ambos os assuntos são apresentados os experimentos realizados juntamente com análises, discussões e apontamentos importantes levantados durante todo o desenvolvimento do projeto.

As conclusões, contribuições, publicações, limitações e trabalhos futuros são apresentadas no *Capítulo 4*.

# Capítulo 2

## FUNDAMENTAÇÃO TEÓRICA E TRABALHOS CORRELATOS

---

---

O Aprendizado de Máquina (AM) é uma subárea de Inteligência Artificial (IA). O AM busca o desenvolvimento de programas de computador que possam evoluir à medida que são expostos a novas experiências (MITCHELL, 1997).

O principal objetivo do AM é a busca por métodos e técnicas que permitam a concepção de sistemas computacionais capazes de melhorar seu desempenho, de maneira autônoma e a partir de informações obtidas ao longo de seu uso; característica considerada um dos mecanismos fundamentais que regem os processos de aprendizado automático (MITCHELL, 2006).

### 2.1 Tipos de Aprendizado

Tradicionalmente, o aprendizado de máquina pode ser dividido em *aprendizado supervisionado* e *aprendizado não supervisionado*. Mais recentemente, passou-se também a investigar o processo de aprendizado chamado *aprendizado semissupervisionado* (ZHU et al., 2003).

#### 2.1.1 Aprendizado Supervisionado

O aprendizado supervisionado é um processo guiado por um modelo gerado através da tarefa de treinamento. Caracteriza-se por possuir duas etapas:

1. Criação do modelo: São duas as entradas - um conjunto de dados conhecidos e os possíveis rótulos. A saída é o modelo que será usado na próxima etapa para rotular as respostas da base de dados.

2. Predição: São duas as entradas - o modelo obtido na etapa anterior e a base de dados. A saída são os rótulos previstos pelo modelo.

O aprendizado supervisionado pode ser usado para a classificação de textos, a identificação de tipos de clientes a partir de classes definidas, a identificação de alfabeto através de imagem, a identificação de alfabeto através de som, etc.

### 2.1.2 Aprendizado Não Supervisionado

O aprendizado não supervisionado distingue-se do aprendizado supervisionado já que nenhum dado de treinamento é fornecido. Os algoritmos de aprendizado não supervisionado são caracterizados por identificar padrões de similaridade/dissimilaridade sem o conhecimento das classes (rótulos).

O uso do aprendizado não supervisionado é importante em bases de dados em que há dificuldade de identificação das classes existentes. O objetivo de algoritmos de aprendizado não supervisionado é identificar grupos de dados que possuam maior similaridade.

A entrada do aprendizado são dados desconhecidos e a saída são grupos sem rótulos, os quais devem ser rotulados por um especialista.

O aprendizado não supervisionado pode ser usado para a clusterização de tipos de textos, a identificação de grupos de clientes, a identificação de grupos de imagens parecidas, a identificação de grupos de sons parecidos, etc.

### 2.1.3 Aprendizado Semissupervisionado

O aprendizado semissupervisionado, contexto em que este trabalho se insere, utiliza a abordagem supervisionada e não supervisionada de forma conjunta.

A característica principal do aprendizado semissupervisionado é que, a partir de uma pequena amostra de dados rotulados (aprendizado supervisionado), é possível rotular uma grande amostra de dados não rotulados (não supervisionado). Em outras palavras, o aprendizado semissupervisionado usa uma pequena amostra de dados rotulados combinada com uma grande amostra de dados não rotulados, visando obter melhores predições.

A amostra de dados rotulados é incrementada automaticamente a cada iteração, tornando possível a rotulagem de uma grande amostra de dados sem a necessidade de um especialista adicionar novos rótulos a cada iteração.

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 semisupervisão, nesse caso, ocorre devido ao conjunto  $IR1$  ser usado para iniciar o processo de aprendizagem (etapa de treinamento) e, em seguida, rotular as instâncias do conjunto  $INR1$  (etapa de predição), o que gera um novo conjunto de instâncias rotuladas aprendidas ( $IRDA1$ ) durante o processo.

Uma vez gerado o conjunto  $IRDA1$ , as suas instâncias passam a ser utilizadas juntamente 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 utilizadas no treinamento é expandido a cada nova iteração.

O aprendizado semissupervisionado pode ser usado para a classificação de textos ou palavras, a identificação de padrões de clientes em uma base muito grande ou em crescimento, a identificação de perfis de terroristas, etc.

## 2.2 Never-Ending Language Learning - NELL

A primeira definição formal da NELL (Never-Ending Language Learning) foi apresentada em (BETTERIDGE et al., 2009). A evolução e os refinamentos da NELL foram apresentados nas publicações subsequentes (CARLSON et al., 2009), (CARLSON et al., 2010a), entre outras.

A entrada da NELL é uma ontologia inicial  $O$  e, para cada categoria pré-definida  $C$  e para cada relação  $R$ , em  $O$ , a especificação é:

- alguns *noun phrases* (NP) (ou sintagmas nominais ou ainda Entidades Nomeadas - EN), que são instâncias confiáveis de  $C$ ,
- alguns pares de entidades nomeadas (pares de EN), que são instâncias confiáveis de  $R$ ,
- alguns padrões textuais (PT ou ainda Contextos), os quais são conhecidos pela alta precisão em extrações (HEARST, 1992),
- uma lista de categorias mutualmente exclusivas  $C$ ,
- uma lista de categorias na qual  $C$  é um subconjunto de cada uma delas,
- uma lista de relações mutualmente exclusivas  $R$ ,
- uma lista de relações na qual  $R$  é um subconjunto de cada uma delas.

O objetivo principal da NELL é aumentar, o máximo possível, suas listas de instâncias para cada categoria e para cada relação, mantendo boa precisão.

Na tentativa de atingir seu objetivo, a NELL também amplia o conjunto de padrões textuais confiáveis associados a cada categoria e a cada relação, como também cria novas relações para expandir a ontologia inicial  $O$ , como descrito em (MOHAMED et al., 2011) e (OLIVERIO; JR, 2012). Também são aplicados diferentes algoritmos que ajudam o sistema a popular automaticamente sua própria KB (como descrito em (GARDNER et al., 2014) e (GARDNER et al., 2013).

Na Figura 2.1 é apresentada a estrutura da arquitetura original do sistema NELL. Todos os componentes são acoplados de forma a cooperar entre si e, continuamente, aumentar a KB em número de instâncias (número de fatos).

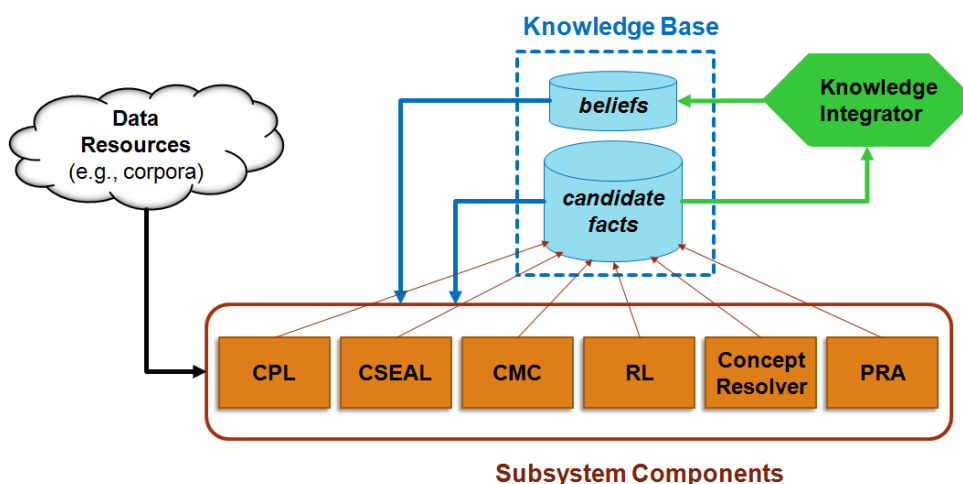


Figura 2.1: Arquitetura básica NELL. Figura inspirada em (CARLSON et al., 2010a).

A estrutura da arquitetura original da NELL possui cinco subsistemas principais de extração de conhecimento (de 1 a 6), um componente para avaliação dos resultados dos outros componentes (7) e a base de conhecimento (8). Tais componentes são:

1. *Coupled Pattern Learning* - CPL realiza o aprendizado a partir da extração de ENs e PTs de páginas web.
2. *Coupled SEAL* - CSEAL realiza *queries* na web e extrai conhecimento a partir de padrões HTML.
3. *Coupled Morphological Classifier* - CMC é um classificador morfológico que examina os resultados obtidos através do CPL e do CSEAL, visando encontrar padrões morfológicos. Por exemplo, "pólis" pode ser apontado como um padrão comumente encontrado em nomes de cidades: Pratápolis, Jardinópolis, Pradópolis, Fernandópolis, etc.

4. *Rule Learner - RL* infere cláusulas de Horn a partir do conhecimento que já faz parte da KB da NELL. Para isso, o RL usa o ILP, um sistema similar ao FOIL (QUINLAN; CAMERON-JONES, 1993). O RL usa uma implementação adaptada do algoritmo FOIL para induzir regras probabilísticas.
5. *ConceptResolver* procura por instâncias correferentes a partir de clusterização baseada em relações semânticas da KB da NELL.
6. *Path Ranking Algorithm - PRA* infere novas crenças a partir de análise de caminhos percorridos pelo fato aprendido (GARDNER et al., 2013). Esse caminho é referente aos vínculos entre as relações que extraíram um fato.
7. *Knowledge Integrator - KI* funciona como um árbitro para avaliar os resultados dos componentes e é executado ao final da etapa de promoção.
8. *Knowledge Base - KB* é uma reimplementação do THEO (MITCHELL et al., 1991), um *framework* para sistemas de auto-aperfeiçoamento, o qual pode lidar com milhões de entradas em uma única máquina.

Na Tabela 2.1 são apresentados todos os componentes atuais que podem ser executados na NELL, os quais podem ser vistos com detalhes em (MITCHELL et al., 2015). Resumidamente, as seguintes publicações são referentes especificamente a cada componente: CPL (CARLSON et al., 2009) e (CARLSON et al., 2010b), CSEAL (CARLSON et al., 2010b) e (WANG; COHEN, 2009), ConceptResolver (KRISHNAMURTHY; MITCHELL, 2011), PRA (GARDNER et al., 2013), OntExt (MOHAMED et al., 2011), OpenEval (SAMADI; VELOSO; BLUM, 2013), Prophet (APPEL; HRUSCHKA, 2011) e NEIL (CHEN; SHRIVASTAVA; GUPTA, 2013).

Neste documento somente são abordados em detalhes os componentes mais relevantes para o trabalho desenvolvido.

Os dois subsistemas mais importantes para este trabalho são o CPL e o ConceptResolver. O CPL é um extrator de novas instâncias de classes e de padrões textuais a partir da leitura de páginas da web. O CPL aprende e usa padrões contextuais como "*prefeito de X*" e "*X joga para Y*" para extrair instâncias de categorias e de relações, respectivamente.

O CPL usa estatísticas de co-ocorrência entre EN e PT (ambos definidos usando *tags* de um *part-of-speech*) a fim de aprender padrões de extração para cada predicado de interesse e, com isso, encontrar instâncias adicionais de cada predicado. Relacionamentos entre predicados são usados para filtrar padrões de saída, os quais são muito genéricos.

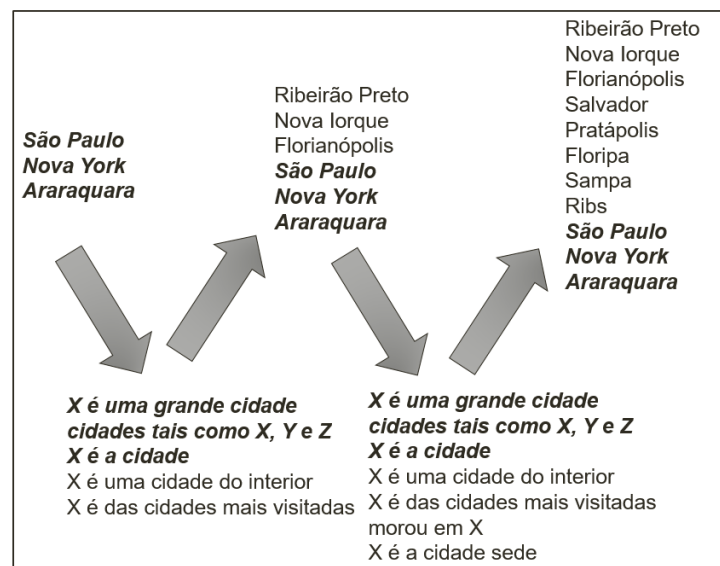


Tabela 2.1: Arquitetura básica NELL. Figura adaptada e inspirada em (VIEIRA, 2015).

Componentes	CPL	CSEAL	CMC	ConceptResolver	PRA	OntExt	OpenEval	Prophet	NEIL
Entrada	<i>all-pairs-data</i>	HTML de páginas web	KB	KB	KB	<i>all-pairs-data</i>	KB	KB	imagens
Aprendizado de instâncias de categorias	Sim	Sim	Sim	Não	Não	Não	Sim	Não	Sim
Aprendizado de instâncias de relações	Sim	Sim	Sim	Não	Sim	Sim	Sim	Sim	Sim
Inferência de novas relações	Não	Não	Não	Não	Não	Sim	Não	Não	Não
Modelo	Padrões textuais	Wrappers de documentos etiquetados	Características morfológicas	Características semânticas (relações)	Instâncias de relações	Padrões extraídos anteriormente	Instâncias e palavras-chave baseadas em contexto	Instâncias de relações	Relações entre objetos, cenas e atributos
Aprendizado online (Direto da web)	Não	NELL Inglês - Não NELL Português - Sim	Não	Não	Não	Não	Não	Não	Não
Escopo	Global	Local	Global	Global	Global	Global	Local	Global	Global
Foco do aprendizado/Vantagens	Aprendizado de padrões textuais	Pode aprender diretamente da web ou de uma base local; Independente de Língua.	Aumenta a precisão do aprendizado	Indução de sentido de palavra e resolução de coreferência	Aumenta a precisão da KB	Gera novas relações	Avalia instâncias	Identifica as relações corretas e incorretas	Extraí conhecimento a partir de imagens

Na Figura 2.2 é apresentado um exemplo simples do funcionamento do CPL no aprendizado de categorias. Como já citato, a NELL é baseada no aprendizado semissupervisionado então inicialmente, são dadas algumas sementes de ENs e PTs (em negrito e itálico).

As setas indicam metade de uma iteração. Uma iteração completa do aprendizado de categorias é o seguinte ciclo: com as sementes de ENs de cidade são aprendidos novos PTs, aumentando assim a amostra de PTs, e com toda a amostra de PTs (aprendidos e dados como sementes) novas instâncias de ENs são aprendidas, aumentando assim a amostra de ENs.



**Figura 2.2: Aprendizado de Categorias - Categoria cidade**

Exemplificando uma iteração do aprendizado de categorias para cidade, de acordo com Figura 2.2, com as sementes de EN "São Paulo", "Nova York" e "Araraquara" foram aprendidos os PTs "\_ é uma cidade do interior" e "\_ é uma das cidades mais visitadas".

Em seguida, com os PTs aprendidos e com as sementes inseridas previamente ("\_ é uma grande cidade", "cidades tais como \_" e "\_ é a cidade"), foram aprendidas novas ENs: "Ribeirão Preto", "Nova Iorque" e "Florianópolis". Após esse ciclo é iniciada outra iteração.

Na Figura 2.3 é apresentado um exemplo simples do funcionamento do CPL no aprendizado de relações, em o funcionamento é basicamente o mesmo do aprendizado de categorias. A partir de sementes de pares de ENs de cidade foram aprendidos novos PTs de relações, aumentando assim a amostra de PTs de relações, e com toda a amostra de PTs de relações (aprendidos e dados como sementes) novos pares de ENs foram aprendidas, aumentando assim a amostra de ENs.

Exemplificando uma iteração do aprendizado de relações para a relação localizadaEm, de acordo com Figura 2.3, com as sementes de pares de ENs "São Paulo" e "Brasil", "Nova York" e



**Figura 2.3: Aprendizado de Relações - Relação localizadaEm**

"Estados Unidos", e "Araraquara" e "Brasil" foram aprendidos os PTs de relações "\_ é a cidade mais procurada nos \_" e "\_ é uma das cidades mais bonitas do \_".

Em seguida, com os PTs aprendidos e com as sementes inseridas previamente ("\_ é a maior cidade do \_", "\_ é localizada nos \_" e "\_ é uma das cidades do \_"), foram aprendidos novos pares de ENs: "Ribeirão Preto" e "Brasil", "Nova Iorque" e "Estados Unidos", e "Florianópolis" e "Brasil". Da mesma forma que no aprendizado de categorias, após esse ciclo é iniciada outra iteração.

As probabilidades de instâncias candidatas extraídas pelo CPL são heurísticamente atribuídas pelo uso da fórmula  $1 - 0.5^c$ , na qual  $c$  é o número de padrões promovidos que extraem um candidato.

Na Leitura da Web em inglês, foi dado como entrada ao CPL um *corpus* de 2 bilhões de sentenças, geradas por meio do pacote OpenNLP<sup>1</sup> para extrair *tokens* e *Pos-Tagger* de sentenças de 500 milhões de páginas da web em inglês da base ClueWeb (CALLAN; HOY, 2009).

O segundo componente mais importante para este trabalho, o ConceptResolver (KRISHNAMURTHY; MITCHELL, 2011), foi desenvolvido para lidar com polissemia e também com resolução de sinonímia através da identificação de conceitos latentes aos quais as EN se referem.

No ConceptResolver, o conhecimento de domínio (ontologia) orienta a criação de conceitos na definição de um conjunto de possíveis tipos semânticos para cada um dos conceitos. O processo de identificação do sentido das palavras (*word sense*) é realizado pela inferência de

<sup>1</sup><http://opennlp.sourceforge.net>

um conjunto de tipos semânticos para cada EN. Enquanto a detecção de sinonímia explora informações redundantes para a classificação de vários domínios de forma semissupervisionada.

O aspecto mais importante do ConceptResolver é que ele se baseia principalmente em características semânticas para realizar a resolução de correferências. A ideia básica do ConceptResolver é usar as relações semânticas (descritas na ontologia) como características para um conjunto de modelos de aprendizado.

Considere o exemplo apresentado na Tabela 2.2. Inicialmente, a NELL aprende que uma específica EN se refere a um *atleta* (como *Kobe Bryant*), assim, o sistema tentará aprender instanciações para todas as relações definidas para *atleta*.

Entre as relações definidas na NELL, estão: *athletesSuchAsAthletes*, *athleteHomeStadium*, *athleteCoach*, *athletePlaysInLeague*, *athletePlaysForTeam*, *athleteKnowAs* etc.

**Tabela 2.2: Instâncias de relações usadas como características semânticas para resolução de correferências.**

Noun phrase	Kobe Bryant	Kobe	Bryant	Lebron
<i>athletesSuchAsAthletes</i>	Basketball player	Basketball player	Basketball player	Basketball player
<i>athleteHomeStadium</i>	Staples Center	Staples Center	Staples Center	Quicken Loans Arena
<i>athleteCoach</i>	Byron Scott	Byron Scott	Byron Scott	David Blatt
<i>athletePlaysInLeague</i>	NBA	NBA	NBA	NBA
<i>athletePlaysForTeam</i>	LA Lakers	LA Lakers	LA Lakers	Cleveland Cavaliers
<i>athleteKnowAs</i>	Kobe	Kobe Bryant	Kobe	Lebron James

Com base nessas relações e no exemplo apresentado na Tabela 2.2, os modelos de aprendizado tentarão encontrar padrões para decidir se um par de EN se refere a um conceito, ou não.

No exemplo dado, é possível notar a tendência dos pares de EN (*kobe Bryant*, *Kobe*), (*Kobe Bryant*, *Bryant*) e (*Kobe*, *Bryan*) serem correferentes, referem-se ao mesmo atleta. Por outro lado, os pares de EN (*Kobe Bryant*, *Lebron*), (*Kobe*, *Lebron*) e (*Bryant*, *Lebron*) não são correferentes.

Importante ressaltar que diferentes relações impactam a resolução de correferência de forma diferente, assim, diferentes modelos devem explorar isso. Na Tabela 2.2, por exemplo, a relação *athleteKnownAs* deve ser uma característica mais forte que *athleteHomeStadium*.

Uma das principais características da resolução de correferência, usando KB dinâmica da

NELL, é que nem todos os *slots* são completos. Isso acontece porque a KB não é completa; ela é populada todos os dias com novos fatos (conhecimento), os quais são lidos da web. Apesar disso, não espera-se que a base de conhecimento seja completa. Portanto, é normal encontrar *slots* vazios na KB (como ilustrado na Tabela 2.3)

**Tabela 2.3: Exemplo de *slots* vazios que podem ocorrer com instâncias de relações usadas como características semânticas para a resolução de correferência.**

Noun phrase	Kobe Bryant	Lebron	Steve Nash	Tiger Woods
athletesSuchAsAthletes	Basketball player	Basketball player	Basketball player	Golf player
athleteHomeStadium	Staples Center	Quicken Loans Arena		
athleteCoach	Byron Scott	David Blatt		
athletePlaysInLeague	NBA	NBA	NBA	
athletePlaysForTeam	LA Lakers	Cleveland Cavaliers	LA Lakers	
athleteKnowAs	Kobe	Lebron James		
dateOf Birth				Dec/30/75

A presença de *slots* vazios trazem mais dificuldade para métodos de resolução de correferência baseados em características semânticas. A proposta de trabalho apresentada neste documento levou em consideração essa dificuldade.

## 2.3 Características Linguísticas

Segundo (NG, 2010), as características linguísticas podem ser divididas em quatro grupos: léxico, gramática, semântica, e posicional. O léxico é o conjunto de palavras usadas em uma língua; a gramática refere-se ao conjunto de regras que são usadas em uma determinada língua; a semântica é o significado de uma palavra ou sentença e, por fim, o posicional refere-se ao som gerado.

Esta subseção foi baseada na discussão realizada em (NG, 2010) e são apresentadas algumas características linguísticas comuns na literatura. As características abordadas a seguir foram escolhidas por serem comumente usadas para a resolução de anáforas, sinônimos, ambiguidade e correferência.

- **String-matching** - É uma característica que busca todas as ocorrências de um padrão de texto a partir de uma sequência de caracteres.

Por exemplo, suponha o padrão "Ribeirão" e a seguinte entrada: "**Ribeirão Preto** é uma cidade quente! Conhecida como a capital do chopp, **Ribeirão** é o ponto de pessoas interessadas no agronegócio". As palavras "Ribeirão Preto" e "Ribeirão" foram as ocorrências encontradas a partir do padrão de entrada "Ribeirão".

Em (STRUBE; RAPP; MÜLLER, 2002), a String-matching trata da distância mínima entre as ocorrências, já (CASTAÑO; ZHANG; PUSTEJOVSKY, 2002) tratam a subsequência mais longa. Em (YANG et al., 2004), foi abordada a ideia de *Bag of Words* (BoW) para calcular a similaridade entre dois NPs e obter o valor de TF-IDF (*Term Frequency - Inverse Document Frequency*).

- **Características Sintáticas** - Aqui, as NPs são analisadas a partir de árvores sintáticas. Em (GE; HALE; CHARNIAK, 1998), foi implementada a distância de Hobbs ( $d_H$ ), proposta em (HOBBS, 1986), na qual o antecedente mais provável de um pronome é identificado através de uma árvore sintática. Em (LUO; ZITOUNI, 2005), foram extraídas características a partir de uma árvore sintática para a implementação de *Binding Constraints* (restrições de ligação) (CHOMSKY, 1988).
- **Características gramaticais** - São usadas normalmente em conjunto. Em (NG; CARDIE, 2002), foram aplicadas trinta e quatro características gramaticais (e.g. concordância de gênero, se é um pronome, se começa com a palavra "the", etc.), as quais incluem preferências e restrições. Nesse caso, várias características foram agrupadas visando uma análise mais completa.
- **Características Semânticas** - São características voltadas à semântica. Podem ser usadas utilizando o seguinte método de notação para resolução de correferências: com as propriedades semânticas de um tópico, monta-se uma lista do que pode ou não expressar correferência. Por exemplo: Ator é [-Cidade], [+Celebridade], [0Cantor], etc., em que o sinal de "+" representa o que é presente, o "-" o que não é presente, e zero indica informações insuficientes.

De acordo com (NG, 2010), "preferência de seleção" (DAGAN; ITAI, 1990; KEHLER et al., 2004; YANG; SU; TAN, 2005; HAGHIGHI; KLEIN, 2009) pode ter sido um dos primeiros conhecimentos semânticos usados na resolução de correferência. Nessa abordagem, considera-se que um pronome a ser resolvido pode ter vínculo com outro já resolvido. O vínculo é positivo em caso de ambos os pronomes estarem relacionados ao mesmo verbo e às mesmas características linguísticas.

A WordNet é um exemplo do uso de extração de características semânticas (*Semantic features*) para resolver a similaridade entre dois NPs (PONZETTO; STRUBE, 2006, 2007).

Em (POESIO et al., 2007), foi realizada uma análise de desambiguação na Wikipédia, na qual, para a verificação se dois NPs eram correferentes ambos deveriam constar na lista de categorias e no primeiro parágrafo da página web do NP candidato.

- **Padrões léxico-sintáticos** - Têm sido usados para calcular a probabilidade de dois NPs serem correferentes a partir do relacionamento semântico entre eles.

Padrões léxico-sintáticos podem ser aprendidos a partir de *corpora* anotados (YANG; SU, 2007) e também de *corpora* não anotados, através de *bootstrapping* (BEAN; RILOFF, 2004). A ideia é que quanto mais frequentemente os padrões ocorrerem com dois NPs, mais chances de serem correferentes. Essa abordagem foi usada para resolver anáforas em (MODJESKA; MARKERT; NISSIM, 2003) e (LUO; ZITOUNI, 2005) para encontrar a distância de características lexicais em (POESIO et al., 2004).

- **Discourse-based features** - Característica que identifica o candidato antecedente mais provável, medindo sua distância a NP anáfora a ser resolvida. Em (HIRST, 1981) são apontados vários trabalhos iniciais entre 1970 e 1980.

Em (IDA; INUI; MATSUMOTO, 2009), foi treinado um ranqueador dos candidatos antecedentes. Em (TETREAUULT, 2005), foram usadas a teoria de discurso e a teoria de veias (CRISTEA; IDE; ROMARY, 1998), que a partir de um resolvidor de pronomes baseado em heurísticas identificava e removia candidatos antecedente errados (IDE; CRISTEA, 2000).

- **Concordância de gênero** - Identifica se há concordância de gênero (Masculino/Feminino). Foi uma das trinta e quatro características usadas em (NG; CARDIE, 2002).
- **Concordância de número** - Identifica se há concordância de número (Plural/Singular);
- **Concordância semântica** - Identifica se dois NPs possuem as mesmas categorias semânticas. Exemplo no contexto de cidade: Florianópolis e Floripa.

## 2.4 **Corpora Importantes**

De acordo com (NG, 2010), "Grande parte da popularização do aprendizado de máquina aplicado à resolução de correferência deve-se a disponibilidade de corpus públicos". Atualmente, existem vários corpora disponíveis e os mais importantes para este projeto são citados nesta subseção.

- ClueWeb<sup>2</sup> é um corpus iniciado em 2009 e que continua em crescimento para algumas línguas.

No total, o ClueWeb possui mais de 1 bilhão de páginas web em 10 línguas. Para o português, o ClueWeb possui aproximadamente 40 milhões de páginas web; esse tamanho é fixo, já que a coleta não foi continuada para essa língua. Apesar disso, o ClueWeb foi o maior corpus de páginas web em português encontrado e disponibilizado no início deste projeto.

A extração de páginas para a sua criação foi realizada através de um *crawler*, que executou extrações de diversas bases. O *ClueWeb* está em uso na CMU (Carnegie Mellon University) e foi disponibilizado para a NELL em português. Na Tabela 2.4 são apresentados os números de páginas web extraídas no ClueWeb09.

**Tabela 2.4: Número de páginas do ClueWeb09. Fonte: <http://boston.lti.cs.cmu.edu/clueweb09>**

Language	# Records
English	503,903,810 pages
Chinese	177,489,357 pages
Spanish	79,333,950 pages
Japanese	67,337,717 pages
French	50,883,172 pages
German	49,814,309 pages
Portuguese	37,578,858 pages
Arabic	29,192,662 pages
Italian	27,250,729 pages
Korean	18,075,141 pages

A atual versão desse *corpus* é a ClueWeb12, atual fonte de extração da NELL em inglês. Para essa versão, o número de páginas web em inglês passou para 733.019.372.

- O CETEMPúblico e o CETEMFolha<sup>3</sup> são *corpora* da língua portuguesa não anotados. O CETEMPúblico foi extraído do Jornal Diário Português e possui 7.082.094 frases. Já o CETEMFolha foi extraído da Folha de São Paulo com 1.597.807 frases. Ambos fazem parte da Coleção CHAVE e são referentes aos anos de 1994 e 1995 de ambos os jornais. A partir de 2007, foram disponibilizados textos do Jornal *A Folha de São Paulo* e do *Jornal Público* entre 2004 e 2008 no total de 1456 edições, anotados sintaticamente pelo PALAVRAS (BICK, 2000).
- XLike<sup>4</sup>, atualmente XLime<sup>5</sup>, é um projeto composto por várias instituições: Jozef Stefan Institute (Eslovênia), Karlsruhe Institute of Technology (Alemanha), Universitat Po-

<sup>2</sup><http://lemurproject.org/clueweb09/>

<sup>3</sup><http://www.linguateca.pt/>

<sup>4</sup><http://www.xlike.org>

<sup>5</sup><http://xlime.eu>



litecnica de Catalunya (Espanha), University of Zagreb (Croácia), Tsinghua University (China), iSOCO (Espanha), Bloomberg (USA), Slovenian Press Agency (Eslovênia) e parceiros associados: New York Times e IIT Bombay.

As línguas tratadas são: inglês, alemão, espanhol, chinês, português, esloveno, catalão, entre outras. O objetivo do XLike é agregar diferentes conhecimentos disponíveis na internet e permitir aplicações através da vinculação de várias áreas científicas diferentes: PLN, Aprendizado de Máquina e tecnologias semânticas.

Os recursos do XLike podem ser usados para a sumarização *cross-lingual*, a contextualização, a personalização, a detecção de plágio em conteúdos disponibilizados na internet e, futuramente, para a melhoria da leitura da web em português e para a agregação de outras línguas a NELL.

- O Corpus Brasileiro (CB)<sup>6</sup> pertence ao grupo GELC, sediado no Centro de Pesquisas, Recursos e Informação de Linguagem (CEPRIL) da PUC-SP (Pontifícia Universidade Católica de São Paulo) e apoiado da FAPESP (Fundação de Amparo à Pesquisa do Estado de São Paulo). Possui aproximadamente 1 bilhão de palavras etiquetadas do português brasileiro contemporâneo de vários gêneros textuais como artigo, teses e dissertações, anais, artigos da Wikipédia, entre outros.

## 2.5 Início da Leitura da Web em Português

A *leitura da web em português* é um projeto que teve início em 2009 com o sistema Read The Web in Portuguese (RTWP) apresentado em (DUARTE; NICOLETTE; HRUSCHKA, 2011).

RTWP é um sistema criado com base no componente CPL da NELL, o qual aprende entidades nomeadas (EN) e padrões textuais (PT) a partir de páginas web. A primeira versão do RTWP realizava a extração de conhecimento diretamente da web através de *queries* utilizando a API BOSS da Yahoo!<sup>7</sup>.

O objetivo principal da primeira versão do RTWP foi comprovar empiricamente que, com métodos de acoplamento do aprendizado de máquina sem-fim aplicados na NELL, é possível minimizar o desvio de conceito (*semantic-drift*) que acontece em sistemas de aprendizado semissupervisionado.

O desvio semântico ocorre no aprendizado semissupervisionado, por exemplo, quando o

---

<sup>6</sup><http://corpusbrasileiro.pucsp.br/>

<sup>7</sup><https://developer.yahoo.com/search/boss/>

sistema categoriza, de maneira incorreta, um novo exemplo e, a partir daí, passa a utilizá-lo para categorizar novos exemplos (provavelmente provocando incorreções).

O desvio semântico não significa necessariamente um erro. É fato, entretanto, que determinados conceitos podem mudar com o passar do tempo e a identificação das mudanças é crucial para manter a corretude do conhecimento aprendido. Essa característica faz com que, com o passar do tempo, o sistema aprenda mais e mais conceitos incorretos, inviabilizando o seu uso (CURRAN; MURPHY; SCHOLZ, 2007).

De acordo com os resultados apresentados em (DUARTE; NICOLETTE; HRUSCHKA, 2011), com a primeira versão do RTWP foi mostrado, através de evidências empíricas, que a *leitura da web em português* é viável, bem como o uso de métodos de acoplamentos para a minimização do desvio semântico.

Apesar disso, o RTWP ainda não possibilitou a *leitura da web em português* de forma sem fim, como o conceito apresentado na NELL, devido a dois fatores principais:

1. A demora na extração de conhecimento diretamente da web era grande, o que tornava o sistema apto somente para a execução de um número limitado de iterações. Isso acontecia porque o RTWP realizava, de forma conjunta, a tarefa de extração (contagem de todas as ocorrências e co-ocorrências) e a tarefa de aprendizado, diferentemente do CPL.

O CPL conta com um *all-pairs-data*: uma base pré-processada de páginas web com todas as estatísticas necessárias a NELL. A tarefa de extração (pré-processamento) que foi realizada no RTWP não é necessária ao CPL, já que esse atua a partir de um *all-pairs-data*. O uso do *all-pairs-data* diminui muito o tempo do processo de aprendizado, pois o pré-processamento é a tarefa mais demorada devido à identificação e contagem de ENs e PTs.

2. O uso de abordagem inspirada somente em um componente da NELL, o CPL. Como foi apresentado em várias publicações sobre a NELL, tais como (CARLSON et al., 2009) (CARLSON et al., 2010a) e (CARLSON et al., 2010b), a chave do aprendizado sem fim é o uso de métodos e componentes acoplados através do uso do *co-training* (BLUM; MITCHELL, 1998), técnica usada nos acoplamentos propostos na NELL.

Apesar do CPL possuir acoplamentos internos em sua implementação, ele é um componente que aprende somente a partir de padrões textuais, o que torna necessário o uso do sistema todo da NELL, com todos os seus componentes.

Visando o avanço de soluções para tais problemas, a versão seguinte do RTWP foi vol-

tada à exclusão da tarefa de extração. Isso foi possível depois do desenvolvimento de um *all-pairs-data* de aproximadamente 2 milhões de sentenças extraídas de parte da coleção dos textos CHAVE. Essa coleção é disponibilizada pela Linguateca<sup>8</sup>. Os resultados dos experimentos foram apresentados em (HRUSCHKA; DUARTE; NICOLETTI, 2013).

Em (HRUSCHKA; DUARTE; NICOLETTI, 2013), foram testadas novamente as estratégias de acoplamento visando minimizar o impacto do desvio de conceito no RTWP. Nessa abordagem, o problema referente ao tempo de processamento (fator 1) foi resolvido, e a ideia de substituir a extração direta da web por um *all-pairs-data* tornou-se viável.

Por outro lado, o RTWP ainda não podia executar de forma sem fim, devido ao fato de possuir um componente somente e, além disso, o *all-pairs-data* criado mostrou-se pequeno e incapaz de sustentar o aprendizado do sistema por muitas iterações (os testes atingiam até a 10<sup>a</sup> iteração aproximadamente).

Para resolver tais problemas, os seguintes objetivos foram definidos: a criação de uma nova instância da NELL, a tradução da ontologia para o português, algumas alterações no código-fonte da NELL para que fosse possível ler textos com acentuação (não havia tratamento de acentos na NELL) e a criação de um *all-pairs-data* de tamanho suficientemente grande para a execução da NELL em Português de forma sem-fim. Tais tarefas foram parte dos objetivos deste trabalho.

Além disso, para melhorar a confiabilidade da NELL, foi proposta a melhoria abordagem de tratamento de correferência. Essa proposta é relevante pois, além de melhorar a confiabilidade da NELL, independentemente da língua, ela pode auxiliar outros idiomas adicionados a NELL, que provavelmente irão possuir os *all-pairs-data* menores que da NELL em inglês. Quanto menor a redundância de dados, maior a dificuldade em se obter uma alta confiabilidade a partir do uso da macro-leitura.

## 2.6 Trabalhos Correlatos de Resolução de Correferência

Neste trabalho, são usadas as nomenclaturas *microleitura* e *macroleitura*. Define-se *microleitura* como uma análise precisa, na qual é realizada uma leitura detalhada de um texto, do qual se sabe, por exemplo, toda a árvore sintática. Além disso, o acesso ao texto é completo. Já a *macroleitura* é definida como as estatísticas sobre uma grande quantidade de textos. Em outras palavras, na *macroleitura* sabe-se somente as ocorrências e co-ocorrências das combinações de ENs e PTs. Não há acesso ao texto integral ou às árvores sintáticas.

---

<sup>8</sup><http://www.linguateca.pt/>

Como já mencionado, a NELL usa a resolução de correferência para ajudar o sistema a explorar métodos de aprendizado baseados em redundância (métodos que aprendem a classificar um NP com base na frequência com que esse ocorre próximo a um específico PT). O componente responsável por essa tarefa é o ConceptResolver, que explora a resolução de correferência nas relações extraídas de textos da web.

O Resolver, proposto em (YATES; ETZIONI, 2009), é um método não supervisionado e sem conhecimento de domínio para encontrar correferências baseadas em similaridade de *strings* em relações extraídas pelo TextRunner (ETZIONI et al., 2008), o qual não possui ontologia inicial.

Exemplos de abordagens supervisionadas para a resolução de correferência podem ser encontrados em (SINGLA; DOMINGOS, 2006), (LAFFERTY; MCCALLUM; PEREIRA, 2001) e (SNOW et al., 2007). Tais abordagens precisam de um conjunto de dados rotulados, de forma manual, para serem usados como treinamento, o que as torna caras ou inviáveis.

Muitas outras abordagens propõem métodos de resolução de correferência nos quais o usuário provê um domínio de heurísticas de similaridades específicas (WINKLER, 1999), (RAVIKUMAR; COHEN, 2004), (BHATTACHARYA; GETOOR, 2006), (BHATTACHARYA; GETOOR, 2007) e (POON; DOMINGOS, 2007).

A resolução de correferência pode ser abordada de forma mais voltada ao PLN, como a resolução de anáforas (DEEMTER; KIBBLE, 2000). Nesse sentido, ao analisar um documento, um NP pode aparecer em uma frase e a sua anáfora na frase seguinte. A tarefa consiste em, dado um NP1, é preciso identificar outro NP2 usado para se referir a NP1 no mesmo texto.

Note que a resolução de anáfora citada não deve ser uma abordagem adequada ao realizar a resolução de correferência em um domínio em que nenhum documento é dado como entrada (como acontece na NELL).

Em (LIN; MAUSAM; ETZIONI, 2012) foram investigadas várias técnicas a partir do vínculo (*links*) de entidades para a criação de uma KB útil de fatos gerais. Para isso, foi utilizado o REVERB: extrator baseado em regras, o qual identifica relações a partir de restrições sintáticas e lexicais e, em seguida, identifica os pares de NP para cada relação (FADER; SODERLAND; ETZIONI, 2011).

Com o uso do REVERB em 500 milhões de páginas da web, foram atingidas 6 bilhões de extrações como ("Orange Juice", "is rich in", "Vitamin C"). Em seguida, cada palavra foi vinculada (*linking*) à sua entidade correspondente na Wikipédia, conhecida como "entidade de ligação"(ZELENKO et al., 2003).

Em (LIN; MAUSAM; ETZIONI, 2012), foi abordada a microleitura devido ao uso do REVERB

e ao vínculo de cada palavra ao *link* na Wikipédia. A NELL não possui acesso ao texto-fonte, de onde foram extraídas as ENs, logo não há como utilizar o REVERB com o mesmo foco. Além disso, foi usado conhecimento prévio a partir do Freebase e o corpus NGrams do Google Docs, o que influenciaram na confiabilidade *a priori*.

Ainda em (LIN; MAUSAM; ETZIONI, 2012), utilizou-se o cosseno para o cálculo da similaridade entre strings. O cálculo de similaridade, nesse caso, foi usado na microleitura, porém pode ser igualmente aplicado à macroleitura.

A proposta de (LEVIN et al., 2012) aborda a desambiguação de nomes de autores de artigos científicos. A base usada foi a Thomson Reuters' Web of Knowledge<sup>9</sup>, com indexação feita pela ferramenta Lucene<sup>10</sup>. Inicialmente, faz-se o pré-processamento que realiza a quebra dos nomes dos autores em blocos. Em seguida, formam-se grupos a partir da semelhança do sobrenome juntamente com a inicial do primeiro nome (e.g. Hruschka, E.; Hruschka, Estevam).

Depois da criação dos grupos são extraídas regras positivas (referentes aos pares que são possíveis de serem correferentes) e regras negativas (que não estavam ligadas às positivas). Após o pré-processamento, foi usado um algoritmo de desambiguação com características do aprendizado supervisionado e do não supervisionado.

Em (LEVIN et al., 2012), para identificar quais nomes podem ser similares, o nome e o sobrenome foram comparados morfológicamente. Essa é uma característica linguística que pode ser usada na macroleitura da NELL. A diferença entre o uso da abordagem apresentada em (LEVIN et al., 2012) e no tratamento de correferência da NELL é que na NELL não é sabido de onde vem o texto, enquanto em (LEVIN et al., 2012) sabe-se exatamente a fonte do texto (bases de artigos científicos) e o campo a ser utilizado.

No método de resolução de correferência proposto neste projeto de doutorado, comparou-se *strings* morfológicamente para descobrir se são similares o suficiente para serem correferentes, porém o algoritmo não tem conhecimento sobre as *strings*; elas podem ser de qualquer categoria, não há conhecimento prévio, diferentemente do que foi apresentado em (LEVIN et al., 2012).

Em (LEVIN et al., 2012), também é usada a clusterização de características linguísticas, assim como no *ConceptResolver*, porém fazendo uso da microleitura, pois utilizou nomes de autores, co-autores e outras características extraídas de artigos dos quais o formato e os campos exatos de extração eram conhecidos.

---

<sup>9</sup><http://www.webofknowledge.com>)

<sup>10</sup><http://lucene.apache.org>

# Capítulo 3

## A LEITURA DA WEB EM PORTUGUÊS E A RESOLUÇÃO DE CORREFERÊNCIA INDEPENDENTE DE LÍNGUA

---

---

O desenvolvimento da NELL em português iniciou-se com o RTWP, conforme descrito no Capítulo 2. O RTWP foi usado para os testes iniciais no que tange a viabilidade da Leitura da Web em Português. Após a constatação da viabilidade, optou-se pela criação de uma instância completa da NELL para o português.

Neste Capítulo, será abordada a trajetória da Leitura da Web em português, a criação da instância da NELL para o português e o desenvolvimento de um método de resolução de correferência visando a melhoria da confiabilidade da base da NELL, de forma independente de língua.

### 3.1 RWTP & NELL

Foram apontados dois problemas com o RTWP que impossibilitavam a leitura da web em português como já citado no Capítulo 2, Seção 2.5. O primeiro é referente à necessidade de um *all-pairs-data* de tamanho suficientemente grande para a execução do aprendizado da NELL. O segundo refere-se ao fato de o RTWP não possuir todos os acoplamentos necessários para o aprendizado sem-fim. O RTWP é simplesmente uma versão inspirada do CPL.

Visando solucionar tais problemas, foram definidos dois objetivos principais: 1) a criação de um *all-pairs-data* maior em português e 2) a criação e a adaptação de uma nova instância da NELL para o português.

### 3.1.1 All-Pairs-Data

Um *all-pairs-data*, no contexto deste trabalho, significa um corpus pré-processado com base em um *pipeline* PLN (processo de extração de texto através de técnicas linguísticas).

O *all-pairs-data* é importante para a NELL, pois é exatamente a sua fonte de extração de conhecimento, a qual é pré-processada para que contenha todas as estatísticas necessárias ao aprendizado.

Os principais pontos positivos que justificam o desenvolvimento/uso de um *all-pairs-data* são referentes à diminuição do tempo gasto no aprendizado (testes com o RTWP chegaram a diminuir o tempo médio de uma iteração de 1 semana para 1 hora), à separação e à organização dos processos (aprendizado e extração), à necessidade de executar a extração somente uma vez (ou quando da chegada de textos novos), etc.

Para a criação do *all-pairs-data*, inicialmente foi testado parte do conjunto CHAVES e, em seguida, o corpus Brasileiro (ambos citados no Capítulo 2, Seção 2.4).

Em ambos os corpora encontrou-se o mesmo problema: tamanho insuficiente para a execução da NELL. Além disso, o formato em que o corpus Brasileiro foi disponibilizado era de difícil processamento, pois os textos eram todos divididos em vários arquivos, palavra por palavra e organizados por índices. Inicialmente, seria necessário a criação de todos os textos originais e de todo o processo de etiquetagem, extração e contagem de combinações de ENs e PTs, o que demandaria muito tempo para uma pequena quantidade de textos.

Devido à dificuldade de processamento do *corpus* Brasileiro, para teste, foram extraídas aproximadamente 2 milhões de sentenças a partir da página web<sup>1</sup>, na qual o *corpus* está disponível. Para isso, foram usados os padrões de (HEARST, 1992) para buscas textuais disponibilizadas na página do *corpus* Brasileiro. Foram extraídas, aproximadamente, a mesma quantidade de sentenças conseguidas através do CHAVES. A diferença entre os dois *all-pairs-data* foi que, o aprendizado executado pelo RTWP utilizando o *all-pairs-data* do *corpus* Brasileiro, obteve acurácia muito baixa, fazendo com que a sua base pré-processada fosse descartada.

Tendo em vista que o CHAVES (mesmo completo) seria pequeno para o sistema NELL em português, foi realizada nova pesquisa de *corpora* disponíveis em português. Com isso, foi encontrado o ClueWeb, também mencionado no Capítulo 2, Seção 2.4, o qual possui um número maior de páginas que ambos os *corpora* Brasileiro e CHAVES.

Para o pré-processamento do ClueWeb, foi usado um componente disponibilizado pela

---

<sup>1</sup><http://www.sketchengine.co.uk/>

equipe da NELL da Carnegie Mellon University, o qual foi implementado para realizar a extração de textos das páginas web armazenadas em *.warc*. Para o inglês, a extração é realizada somente em UTF-8 e, como o mesmo método foi usado para o Português, foi necessária a alteração do código para a leitura de outros padrões de codificação.

Em seguida, foi criado, pela autora desta tese, um processo de etiquetagem do ClueWeb em português, o qual foi enviado (remotamente) para ser executado na CMU. Para isso, foi utilizado o *part-of-speech* LX-Parser, apresentado em (SILVA et al., 2010). Esse processo foi executado na CMU.

Após a etiquetagem do *corpus* ClueWeb, foi implementada a extração de ENs e de PTs, a partir de uma lista de padrões de *tags* linguísticas, criadas empiricamente com base nos experimentos anteriores do RTWP.

Um exemplo de extração de ENs a partir de *tags* linguísticas é a seguinte sentença etiquetada: "X/PNM **é/V** uma/UM cidade/CN bonita/ADJ ...", onde X é a EN a ser extraída. O padrão de *tag* utilizado para localizar a EN está em negrito e o PT são as palavras referentes às *tags* em negrito, que também devem ser extraídas. Para a extração de relações, a idéia é a mesma, a única diferença é que há extração de uma EN no início e de outra no final (um par de ENs).

Após identificar um padrão de *tag* em uma sentença e realizar a sua extração, é necessário remover as *tags* "é uma cidade bonita" e "X". O exemplo dado é bem simples, porém na maioria das sentenças não basta simplesmente remover as *tags*.

Suponha a seguinte sentença: "São Paulo é uma das cidades mais populosas do mundo". A etiquetagem básica seria: "São/STT Paulo/PNM **é/V** uma/UM de\_/PREP as/DA cidades/CN mais/ADV populosas/ADJ de\_/PREP o/DA mundo/CN". Nesse caso, somente a remoção das etiquetas não é o suficiente para o mapeamento da sentença original, pois ficaria: "São Paulo é uma de as cidades mais populosas de o mundo".

O LX-Parser não possui uma ferramenta que realiza o caminho inverso da etiquetagem, ou seja, o retorno ao texto original. Como essa tarefa é necessária após a extração de ENs e de PTs, foi proposto e implementado (em linguagem JAVA) um método que mapeia o texto original a partir de uma sentença etiquetada. Essa foi uma das tarefas mais demoradas do projeto, devido à documentação do LX-Parser não abordar em detalhes todas as *tags* e formatos de etiquetagem utilizados.

A implementação do método ocorreu com base em tabelas e em códigos-fonte encontrados nos arquivos do etiquetador. Após vários testes e melhorias aplicadas ao código, esse foi enviado à equipe de desenvolvimento do LX-Parser caso tivessem interesse. A equipe disponi-



bilizou também uma versão para a execução em 64bits, usada para a etiquetagem do ClueWeb em Português.

A etapa seguinte foi o desenvolvimento de um método de extração de ENs e PTs utilizando o Hadoop, para que fosse possível executá-lo em um *cluster* disponibilizado pelo grupo NELL da Carnegie Mellon University, a fim de agilizar o processo e reduzir o tempo de processamento.

Juntamente com o método de extração de ENs e PTs, foram desenvolvidos mecanismos de limpeza de *stop words* (palavras comuns em uma língua que não são importantes na extração, por exemplo: eu, nós, era, são, é, ali, mas, mesmo, etc.), caracteres estranhos, sentenças mal formadas, etc.

Finalmente, após a extração e a limpeza das sentenças, o cálculo das estatísticas suficientes pôde ser executado e, assim, o *all-pairs-data* foi gerado para o português, a partir do ClueWeb e com o uso do Hadoop.

O processo de criação do *all-pairs-data* pode ser resumido da seguinte forma:

1. Parser HTML - Extração dos textos a partir de páginas web do ClueWeb;
2. Filtragem de spam e de conteúdo adulto usando uma lista de *stop words*;
3. Segmentação dos textos das páginas web em sentenças;
4. Etiquetagem de todas as sentenças a partir do *part-of-speech* LX-Parser (SILVA et al., 2010);
5. Filtro para a eliminação de sentenças indesajáveis ou inúteis (sentenças sem verbo, somente com *stop words*, etc.);
6. Extração de ENs e de PTs;
7. Cálculo das estatísticas suficientes.

A etapa seguinte do projeto foi a adequação da ontologia da NELL em inglês para o português.

### 3.1.2 Criação de Instância do Sistema NELL para o Português e sua Ontologia

A ontologia da NELL é composta por categorias e relações. As categorias referem-se aos tipos de conhecimento especificados na ontologia, por exemplo: país, pessoa, equipe esportiva,

esporte, etc. Já as relações são referentes aos relacionamentos entre as categorias, por exemplo: `pessoaNasceu(pessoa,país)`, `equipeEsportivaAtuaNoEsporte(equipeEsportiva, Esporte)`.

Alguns exemplos de instâncias de categorias são: `país(Brasil)`, `equipeEsportiva(Flamengo)` e `esporte(futebol)`, `jogador(Pablo Armero)`, `atriz(Regina Duarte)`, etc. E alguns exemplos de instâncias de relações são: `pessoaNasceu(Regina Duarte, Brasil)`, `equipeEsportivaAtuaNoEsporte(Flamengo, Futebol)`, `jogadorEquipeEsportiva(Pablo Armero, Flamengo)`, etc. Em outras palavras, as categorias são tratadas de forma unária enquanto as relações, de forma binária.

Este projeto, inicialmente, voltou-se principalmente à investigação do aprendizado de categorias. Essa escolha foi devido à maior facilidade no aprendizado de categorias do que de relações, devido às categorias serem unárias e as relações são binárias. Por outro lado, a probabilidade de acerto no aprendizado de relações é normalmente maior que no aprendizado de categorias, também devido ao formato unário das categorias e binário das relações.

Por *default*, na ontologia de categorias da NELL em inglês, são dadas sementes de ENs em maior número do que de PTs. Enquanto para ENs o número varia entre 10 e 15 na maioria das categorias, para PTs são dadas somente quatro sementes: "`such categoria as _`", "`categoria, including _`", "`categoria such as _`" e "`categoria, such as _`". Tais padrões foram usados na NELL em inglês seguindo os padrões de Hearst (HEARST, 1992).

O primeiro passo para mapear a ontologia em inglês a fim de que fosse também usada em português, foi a tradução de todas as categorias, relações e sementes de categorias. Muitas dessas sementes foram substituídas por outras para que tivessem características da língua portuguesa, especialmente do Brasil. As sementes de PT, dadas por *default* pelo sistema foram mantidas inicialmente, pois dependiam de alteração no código-fonte.

Apesar de mantidas as sementes de PTs em inglês, foram adicionados alguns PTs em português, como sementes para minimizar o impacto do aprendizado pelos PTs em inglês. Além disso, outros PTs foram adicionados de acordo com cada categoria (por exemplo: "`jogou para _`", "`_ categoria conhecida como _`", etc.), pois a partir dos experimentos publicados em (DUARTE; HRUSCHKA, 2014b) foi possível constatar que somente os padrões de Hearst não eram suficientes para o aprendizado em português. Lembrando que, normalmente, o aprendizado extraído por PTs é importante devido à sua assertividade e à sua confiabilidade.

O segundo passo foi criar uma nova instância da NELL. Para que isso fosse possível, realizou-se anteriormente a alteração no código-fonte da NELL visando a leitura de caracteres com acentuação. Tal alteração foi realizada em conjunto com a equipe de desenvolvimento da NELL em inglês.

É importante ressaltar que as sementes de ENs e PTs, bem como os nomes das categorias e das relações, ainda devem ser inseridos sem acentuação. Devido à complexidade da reimplantação do *script* de geração da KB inicial da NELL, esse problema ainda não foi resolvido.

### 3.1.3 Experimentos e Análise

Os experimentos apresentados a seguir foram publicados em (DUARTE; HRUSCHKA, 2014b). Nessa publicação, foram discutidos e comparados os resultados do RTWP (HRUSCHKA; DUARTE; NICOLETTI, 2013) e do CPL, além do impacto do uso de sementes de PTs em português para a melhoria da leitura da web em português.

Os experimentos foram configurados para obter evidências empíricas no que tange a aceitar ou a rejeitar a seguinte hipótese: se é possível construir um *corpus* pré-processado, com base em um *pipeline* PLN, é possível ler a web em português com base na mesma arquitetura e na mesma implementação usada na NELL.

Após a ontologia da NELL ser mapeada para o português, 11 categorias foram escolhidas randomicamente para serem avaliadas manualmente em relação ao número de instâncias corretas aprendidas. Considerando que, os resultados apresentados em (DUARTE; HRUSCHKA, 2014b) foram obtidos a partir da comparação da NELL usando o *all-pairs-data* criado a partir do ClueWeb com a abordagem do RTWP descrita em (HRUSCHKA; DUARTE; NICOLETTI, 2013).

Nesses experimentos, não foi permitido o uso de todos os componentes da NELL, somente do CPL. Considerando tal configuração experimental, é importante salientar que a precisão obtida não é a mesma que a NELL obteria se ele tivesse sido executado com todos os seus componentes e capacidades. Assim, não era esperado uma precisão extremamente alta com o uso do CPL somente. Porém, essa escolha de configuração permitiu uma análise comparativa justa.

A primeira análise dos experimentos está relacionada à hipótese citada no início desta seção. Por isso, foi mantida a configuração original da NELL em inglês e o sistema foi executado tendo como entrada a ontologia em português e o *all-pairs-data* também em português.

Na Tabela 3.1 são mostrados os resultados obtidos nessa primeira análise, os quais levam à aceitação da hipótese e à conclusão de que o *aprendizado de máquina sem-fim* pode ser aplicado para a leitura da web em português, a partir de uma ontologia e de um *all-pairs-data* em português.

Os resultados para categorias como *chefe de cozinha* e *hospital* mostram que o paradigma

**Tabela 3.1: Resultados do primeiro experimento - Sem adição de novas sementes de PT. CI (Correct Instances): Número de Instâncias Corretas; LI (Learned Instances): Número de Instâncias Aprendidas**

		Iterations			
		5th	10th	15th	20th
aeroporto	#LI:	1	1	1	2
	#CI:	1	1	1	1
	%CI:	100.00%	100.00%	100.00%	50.00%
area de esqui	#LI:	69	69	69	69
	#CI:	68	68	68	68
	%CI:	98.55%	98.55%	98.55%	98.55%
arquiteto	#LI:	0	1	1	1
	#CI:	0	1	1	1
	%CI:	0.00%	100.00%	100.00%	100.00%
arranhaceu	#LI:	0	0	2	9
	#CI:	0	0	0	7
	%CI:	0.00%	0.00%	0.00%	63.63%
astronauta	#LI:	18	55	71	72
	#CI:	4	36	47	47
	%CI:	22.22%	65.45%	66.19%	65.27%
blog	#LI:	0	0	0	2
	#CI:	0	0	0	2
	%CI:	0.00%	0.00%	0.00%	100.00%
cidade	#LI:	0	0	0	0
	#CI:	0	0	0	0
	%CI:	0.00%	0.00%	0.00%	0.00%
chefe de cozinha	#LI:	0	0	48	48
	#CI:	0	0	17	17
	%CI:	0.00%	0.00%	35.41%	35.41%
hospital	#LI:	0	0	52	129
	#CI:	0	0	36	80
	%CI:	0.00%	0.00%	69.23%	62.01%
rodovia	#LI:	0	29	30	30
	#CI:	0	11	12	12
	%CI:	0.00%	37.91%	40.00%	40.00%
shoppingCenter	#LI:	1	27	27	27
	#CI:	0	12	12	12
	%CI:	0.0%	42.85%	42.85%	42.85%

de aprendizagem sem-fim pode ser uma maneira eficaz de um sistema aprender a aprender melhor. Em outras palavras, para essas duas categorias, o sistema não aprendeu até a iteração 10. Entretanto, considerando o princípio do aprendizado sem-fim, desde a primeira iteração, o sistema aprendeu novos padrões textuais que podem ser adequados para aprender instâncias nas futuras iterações.

Ao analisar as iterações 15 e 20 de ambas as categorias (*chefe de cozinha* e *hospital*), nota-se que existe um conjunto de PTs aprendidos em todas as iterações anteriores. Tais padrões permitiram que a NELL aprendesse instâncias (após 10 iterações). Com isso, *chefe de cozinha* obteve 17 casos aprendidos corretamente, enquanto *hospital* obteve 36 instâncias aprendidas até a iteração 15 e mais 44 entre a iteração 15 e 20 (totalizando 80 instâncias aprendidas corretamente).

Os resultados obtidos para as categorias *área de ski*, *rodoviária* e *shopping center* mostram que, depois de aprender algumas instâncias corretas nas iterações iniciais, o sistema reduziu o aprendizado, voltando a aprender somente após a supervisão humana. Um prazo maior de execução (maior número de iterações) seria necessário para entender se um platô foi atingido ou não.

Apesar dos resultados consideravelmente bem-sucedidos, uma análise mais cuidadosa pode

ajudar a regular o procedimento para obter um melhor desempenho de aprendizagem. É possível visualizar, por exemplo na Tabela 3.1, que o número de fatos aprendidos é similar aos obtidos em abordagens anteriores (como em (HRUSCHKA; DUARTE; NICOLETTI, 2013)). Nesse sentido, para algumas categorias como *aeroporto*, *arquiteto* e *blog*, apenas um ou dois casos foram aprendidos após 20 iterações.

Essa análise motivou o seguinte questionamento: O aprendizado do CPL em português pode ser melhorado com o aumento de sementes de PTs?

Essa pergunta surgiu porque na versão do CPL não existem muitas sementes de PT para a extração de ENs. Para a categoria *cidade*, por exemplo, as únicas sementes de PT na primeira iteração seriam: "*such categoria as \_*", "*categoria, including \_*", "*categoria such as \_*" e "*categoria, such as \_*".

Considerando que não foram dadas quaisquer sementes de PTs na definição apresentada anteriormente, os resultados da Tabela 3.1 foram obtidos por meio dos mesmos padrões de *Hearst* (HEARST, 1992) em inglês, os quais não são boas sementes para guiar o processo de aprendizagem.

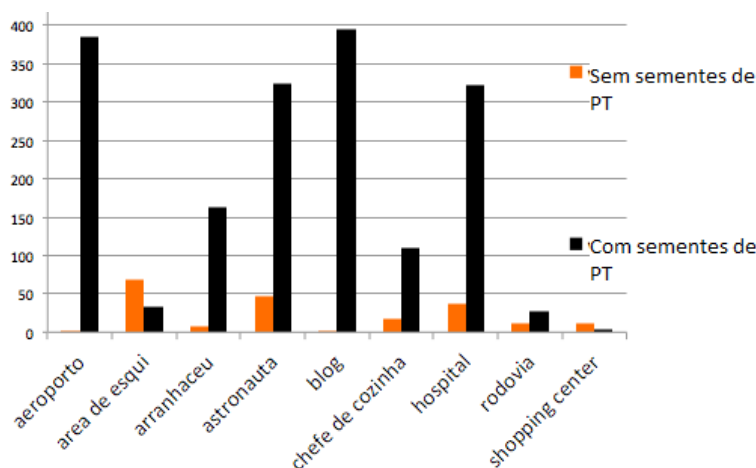
**Tabela 3.2: Resultados do primeiro experimento - Com adição de novas sementes de PTs. CI (Correct Instances): Número de Instâncias Corretas; LI (Learned Instances): Número de Instâncias Aprendidas**

		Iterations			
		5th	10th	15th	20th
aeroporto	#LI:	907	907	907	907
	#CI:	386	386	386	386
	%CI:	42.55%	42.55%	42.55%	42.55%
area de esqui	#LI:	90	90	90	90
	#CI:	34	34	34	34
	%CI:	37.77%	37.77%	37.77%	37.77%
arquiteto	#LI:	170	171	171	171
	#CI:	162	162	162	162
	%CI:	95.29%	94.73%	94.73%	94.73%
arranhaceu	#LI:	954	954	954	954
	#CI:	533	533	533	533
	%CI:	55.87%	55.87%	55.87%	55.87%
astronauta	#LI:	561	568	599	600
	#CI:	320	321	324	324
	%CI:	57.04%	56.51	54.00%	53.91%
blog	#LI:	928	928	928	928
	#CI:	395	395	395	395
	%CI:	42.56%	42.56%	42.56%	42.56%
cidade	#LI:	1829	2715	3607	4485
	#CI:	1373	1859	2366	2854
	%CI:	75.06%	68.47%	65.59%	63.63%
chefe de cozinha	#LI:	171	171	218	243
	#CI:	102	102	106	110
	%CI:	59.64%	59.64%	48.62%	45.26%
hospital	#LI:	915	915	1023	1025
	#CI:	286	286	322	322
	%CI:	31.25%	31.25%	31.47%	31.41%
rodovia	#LI:	109	140	141	141
	#CI:	24	27	27	27
	%CI:	22.01%	19.28%	19.14%	19.14%
shoppingCenter	#LI:	2	72	72	72
	#CI:	0	3	3	3
	%CI:	0.0%	4.16%	4.16%	4.16%

Diante de tais resultados, foi definido o segundo experimento. Nesse, foram dados a NELL conjuntos de cerca de 10 sementes de PTs para cada uma das 11 categorias, as mesmas listadas na Tabela 3.1. Os resultados da segunda configuração são apresentados na Tabela 3.2.

Na Figura 3.1, é apresentado um gráfico com o número das instâncias corretas aprendidas quando o sistema não usou sementes de PTs e quando houve o uso das mesmas para as 11 categorias investigadas.

Os resultados apresentados na Tabela 3.2 revelam que a orientação dos estágios iniciais da NELL em português poderia ajudar a melhorar os resultados da aprendizagem. O uso de sementes de PTs aumentou o número de instâncias aprendidas após 20 iterações em 9 das 11 categorias investigadas. Somente para as categorias *área de esqui* e *shopping center*, as sementes dadas não impulsionaram melhor desempenho.



**Figura 3.1: Instâncias aprendidas corretamente com e sem o uso de sementes de PTs.**

Analisando os resultados apresentados em (HRUSCHKA; DUARTE; NICOLETTI, 2013), é possível concluir que a proposta de PLN usada na NELL/CPL produziu maior número de aprendizado de instâncias corretas.

Apesar da impossibilidade de realizar uma análise comparativa precisa (principalmente porque a ontologia não é exatamente a mesma), nota-se que nas 5 iterações iniciais (HRUSCHKA; DUARTE; NICOLETTI, 2013) foram aprendidas 23,6 instâncias corretas por categoria, enquanto que a combinação do *pipeline* de PLN + NELL alcançou em média 328,6 de instâncias corretas.

## 3.2 Leitura da Web em Português: A NELL em Português

Para que fosse possível a execução do sistema NELL completo (com todos os componentes) para a leitura da web em português, surgiu mais uma etapa: a criação de sementes de PTs para toda a ontologia.

Visando uma padronização, a geração das sementes de PTs foi baseada nos padrões de Hearst (HEARST, 1992) e em algumas variações do português (gênero e número), como: "\_ é a categoria", "\_ é uma categoria", entre outros. Além disso, foram adicionados padrões para a extração de correferências, como: "categoria conhecida como \_", "categoria apelidada de \_", entre outros.

Após a geração da nova ontologia, foram executados testes com a NELL completa em português. Os experimentos prévios indicaram uma grande tendência de instâncias serem aprendidas em inglês, mesmo com o sistema utilizando a ontologia em português.

Isso aconteceu devido à configuração padrão do CSEAL, que realiza extrações a partir de padrões HTML em uma base pré-processada de páginas web em inglês. Devido à extração não ser diretamente da web, os primeiros experimentos não obtiveram bons resultados, pois o aprendizado tendeu muito à língua inglesa, chegando a um platô.

Para que esse problema fosse sanado, a equipe da NELL em inglês permitiu que o componente CSEAL, da instância da NELL em português, tivesse acesso *on-line* às buscas de páginas web. Tal busca foi realizada utilizando uma *app* do Google a partir de um endereço ip liberado para o projeto NELL, o que não limita a extração de páginas web. Além disso, o CSEAL foi configurado para acessar apenas páginas em português.

Para que a NELL em português tivesse todas as características necessárias ao aprendizado sem-fim aplicadas na NELL em inglês, a supervisão humana também foi aplicada, atividade que não havia sido abordada em nenhum dos experimentos anteriores. Tal supervisão da NELL em inglês ocorre semanalmente, com somente 5 minutos de duração por usuário (professores e colaboradores do projeto NELL em inglês). Nesses 5 minutos, o supervisor humano verifica se o aprendizado promovido pela NELL está correto ou não.

Como já citado anteriormente, a NELL em português possui um *all-pairs-data* menor, e para que o aprendizado pudesse evoluir com maior confiança, iniciou-se a supervisão de forma mais intensa que na NELL em inglês.

Na etapa de supervisão humana houve a participação de todo o laboratório MaLL (Machine Learning Laboratory). Ela foi realizada em 2 momentos: 1) após a iteração 50 e 2) após a itera-

ção 80. Além disso, a supervisão foi tratada com amostragem: em 1) foram supervisionados por volta de 100 fatos corretos (ENs) ou por volta de 300 fatos no total, entre corretos e incorretos e, em 2) foram supervisionados por volta de 100 fatos corretos (ENs) ou por volta de 150 fatos o total.

Nessa etapa, o sistema foi parado após a iteração 50 e após a iteração 80 para a execução das supervisões.

Algumas categorias tiveram um maior número de supervisões, pois acreditou-se que as que estavam aprendendo mais poderiam aprender ainda mais sendo supervisionadas mais constantemente.

### 3.2.1 Experimentos e Análise

O experimento foi realizado a partir da NELL em português completo, o qual possui todos os componentes em uso pela NELL (CPL, CSEAL, PRA, CMC), o *all-pairs-data* em português extraído do ClueWeb e a ontologia mapeada para o português com sementes de ENs e de PTs.

O experimento foi executado com ambos os aprendizados, de categorias e de relações, porém não foram inseridas sementes iniciais de PTs para o aprendizado de relações, somente para categorias.

A diferença na inserção de sementes de PTs e ENs no sistema se deveu à hipótese de que as categorias ajudariam no desenvolvimento das relações. Além disso, uma vez inseridas sementes em categorias que se relacionam, essas também são automaticamente adicionadas às sementes de relações.

É importante ressaltar que o CSEAL foi configurado para realizar extrações de padrões HTML diretamente da web em português e o sistema todo foi alterado para a leitura de caracteres com acentuação.

As categorias abordadas foram as mesmas apresentadas em (DUARTE; HRUSCHKA, 2014b). Além dos diferentes componentes, a principal e grande diferença entre os experimentos apresentados em (DUARTE; HRUSCHKA, 2014b) e os apresentados nesta seção, a NELL completo em português, foi a configuração de validação. Em (DUARTE; HRUSCHKA, 2014b) a taxa de aprendizado foi muito baixa e, por isso, as análises realizadas consideraram igualmente o aprendizado promovido e o candidato. Em outras palavras, não houve diferenciação entre as ENs candidatas e as promovidas; todas foram tratadas como promovidas.

Diferentemente de (DUARTE; HRUSCHKA, 2014b), nos experimentos da NELL completa



em português, somente as ENs promovidas foram considerados na avaliação. Já as candidatas somente foram consideradas para a tarefa de supervisão em que essas foram selecionadas como corretas ou incorretas (as incorretas podiam ser apontadas como exemplos negativos também).

Outra diferença entre os experimentos é referente à análise dos resultados. Em (DUARTE; HRUSCHKA, 2014b), a análise foi realizada de forma a considerar pequenos erros que ocorreram devido à etiquetagem incorreta do *part-of-speech* ou na extração de sentenças para a criação do *all-pairs-data*. A avaliação foi realizada com base no processo de aprendizado sem-fim, por exemplo "Viena e Munich" é uma instância incorreta para a categoria *cidade*, porém foi considerada correta, pois o processo de aprendizado identificou o contexto de uma cidade.

Antes da execução desse experimento, foi realizada uma melhoria no processo de criação do *all-pairs-data*, visando tratar casos como no exemplo apontado acima, no qual duas instâncias ligadas por "e" não eram separadas. Além disso, foram melhorados os mecanismos de identificação de ENs (implementação abordada com mais detalhes no Apêndice A).

Os resultados obtidos a partir desse experimento pelo aprendizado de categorias são apresentados na Tabela 3.3.

**Tabela 3.3: Leitura da Web em Português com Supervisão Humana para categorias - Resultados Cumulativos**

Categorias	Iterações de Supervisão Humana	
	51	81
aeroporto	115 0	190 0
area de ski	469 36	581 46
arquiteto	221 0	231 0
arranha-céu	62 0	174 0
astronauta	593 426	703 426
blog	184 5	244 5
cidade	1872 148	2009 2137
chefe de cozinha	213 1	245 1
hospital	105 9	227 9
rodovia	83 2	85 2
shopping center	41 0	41 0

Quantidade de Supervisão  
Promoções Corretas

Os resultados apresentados na Tabela 3.3 estão organizados da seguinte forma: para cada

categoria, são apresentadas as quantidades de supervisões realizadas nas 2 iterações, 51 e 81, as quais são tratadas de forma cumulativa. Além disso, são apresentadas as quantidades de promoções realizadas, antes e depois da supervisão na iteração 51.

Por exemplo, na primeira linha da categoria *aeroporto* na iteração 51, foram supervisionadas 115 instâncias corretas e mais 75 na iteração 81, totalizando 190 instâncias corretas provenientes da supervisão humana.

Na segunda linha da categoria *área de ski*, apresenta-se a quantidade de aprendizado promovido corretamente até a iteração 50, ou seja, sem supervisão. Em seguida, a quantidade de aprendizado promovido corretamente após a supervisão em 51 e antes da supervisão em 81. Para *área de ski* foram aprendidas, sem o uso da supervisão, 36 instâncias até a iteração 50 e, após a supervisão, mais 10, totalizando 46 instâncias.

O resultados referentes ao aprendizado realizado após a última supervisão humana na iteração 81 serão apresentados em um artigo a ser publicado em breve.

A partir da análise dos resultados, comparando-se com (DUARTE; HRUSCHKA, 2014b), houve uma clara melhoria em quantidade de aprendizado. Por exemplo, para a categoria *cidade* em (DUARTE; HRUSCHKA, 2014b) foram aprendidas 2.854 instâncias corretas (entre promovidas e candidatas). No experimento com a NELL em português completa, foram aprendidas (promovidas) 2.137 instâncias corretas ao final do processo e, ao se analisar as candidatas (sistema em execução, sobre o qual os resultados ainda serão apresentados em um artigo), nota-se que ainda existem instâncias corretas, que podem ser promovidas.

Em todas as iterações em que houve supervisão, todas as promoções foram avaliadas, mantendo assim a confiabilidade em 100% até a iteração 81. Isso ocorreu porque as instâncias incorretas foram excluídas e/ou adicionadas como exemplos negativos no sistema.

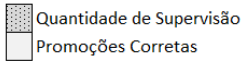
A tarefa de supervisão também foi executada no aprendizado de relações e os resultados são apresentados na Tabela 3.4. As relações que tiveram supervisão humana atingiram um baixo aprendizado, seguindo o mesmo comportamento do aprendizado de categorias.

Pode-se notar pela Tabela 3.4 que nenhum dos resultados do aprendizado de relações teve um bom desempenho, pois a quantidade aprendida foi muito baixa. Isso já era esperado, e uma das prováveis justificativas recai sobre o *all-pairs-data*, o qual ainda é pequeno (possui menos de 10% do tamanho da base inicial utilizada pela NELL em inglês). Além disso, há falhas de etiquetagem no texto, o que impediu que o mesmo fosse aproveitado da forma apropriada.

Um ponto a ser abordado futuramente para a melhoria no aprendizado de relações é a inicialização do sistema NELL com sementes de PTs, assim como foi realizado para as categorias.

**Tabela 3.4: Leitura da Web em Português com Supervisão Humana para relações semânticas - Resultados Cumulativos**

Relações	Iterações de Supervisão Humana	
	51	81
ator atuou no filme	29	34
	6	27
tem cônjuge	0	0
	0	0
trabalha para organização	23	23
	16	16
atleta joga na equipe	25	25
	0	0
equipe joga contra equipe	0	0
	2	4
esporte tem fãs no país	0	0
	0	4
rio corre na cidade	0	0
	1	1



Quantidade de Supervisão  
Promoções Corretas

Além disso, uma investigação para a melhoria dos pares de ENs deve trazer melhores resultados, desde que o *all-pairs-data* seja maior.

Embora o resultado das relações tenha sido muito baixo, vale ressaltar que não ocorreu promoção de relações nos experimentos anteriores.

A conclusão de que o *all-pairs-data* em português ainda é pequeno deve-se à comparação com o tamanho necessário para que a Leitura da Web em inglês fosse executada. Enquanto a NELL em português possui quase 40 milhões de páginas web, a versão em inglês iniciou com 500 milhões. Além disso, atualmente a NELL em inglês atua com mais uma base de 700 milhões de páginas web. Todas essas informações são apresentadas em detalhes no Capítulo 2, subseção 2.4.

Além das melhorias citadas, futuramente o objetivo principal será desenvolver um novo componente voltado às versões da NELL que possuam um *all-pairs-data* pequeno: o *Dictionary Tinker Toy* (DTT).

O DTT traduzirá todas as ENs e pares de ENs de inglês para a nova língua da NELL. A ideia é que o aprendizado promovido da NELL em inglês ajude novas línguas a se desenvolverem mais facilmente, já que essas poderão utilizar o que a NELL já aprendeu como uma "irmã mais velha".

A falta de um *all-pairs-data* suficientemente grande fez com que a NELL em português encontrasse algumas dificuldades e demandasse muita supervisão humana. Apesar disso, essa versão mostrou-se muito útil no que tange apontamentos das melhorias necessárias.

Ainda não se pode afirmar que a Leitura da Web em Português foi atingida de forma sem-

fim, mas os próximos resultados a serem publicados poderão responder melhor esta pergunta.

Além disso, as investigações, os estudos, os experimentos e toda a pesquisa apontaram dificuldades que quaisquer versões da NELL em qualquer língua encontrarão muito provavelmente durante o seu desenvolvimento, tendo em vista que a maior quantidade de textos disponíveis na web é em inglês.

### 3.3 Resolução de Correferência independente de língua na NELL

A ideia de atuar na resolução de correferência na NELL surgiu devido a dois fatores: (1) mesmo que o *all-pairs-data* da parcela em português do ClueWeb tenha sido desenvolvido, ele ainda possui tamanho inferior ao usado pela NELL em inglês e (2) a NELL apesar de possuir o componente de tratamento de correferência ConceptResolver (KRISHNAMURTHY; MITCHELL, 2011), o tratamento é executado com base somente em relações semânticas, e não categorias.

A diferença no tamanho dos *all-pairs-data* (português e inglês) impacta negativamente tanto na quantidade quanto na acurácia dos resultados do português. Além disso, a resolução de correferência torna-se mais difícil, pois não existem relações suficientes para a execução adequada do processo do ConceptResolver. (detalhes na Seção 2.2).

Tendo em vista que a NELL em português sofre um impacto maior no aprendizado devido ao tamanho do *all-pairs-data* e à dificuldade no tratamento de correferência por conta da ausência de relações, foi proposto o tratamento de correferência de forma a unir a abordagem do ConceptResolver a uma nova abordagem voltada ao aprendizado de categorias, de forma independente de língua.

A motivação para a investigação de uma abordagem para a resolução de correferência independente de língua e baseada nas categorias, foi que, além das categorias ajudarem a melhorar as relações, o ConceptResolver não atua sobre elas. Visto que as categorias são mais fáceis de serem aprendidas, espera-se que ocorram mais ENs correferentes extraídas pelo aprendizado de categorias do que pares de ENs extraídas pelo aprendizado de relações.

Além disso, a melhoria no processo de resolução de correferência impacta positivamente na Leitura da Web a partir de um *all-pairs-data* que não seja tão grande quanto o utilizado pela NELL inglês. Quanto maior a confiabilidade de uma base de conhecimento, maiores são as chances de continuidade do aprendizado, desde que haja uma fonte adequada de extração.

Como já citado na Seção 2.2, uma das principais limitações do ConceptResolver está relaci-

onada à incompletude da base da NELL. Assim, sempre que o método encontra muitos espaços vazios na KB ele enfrenta dificuldades para identificar correferências.

Como exemplo, considere o par de ENs (*kobe Bryant, Kobe*) apresentado na Tabela 3.5. Nesse caso, a NELL aprendeu apenas algumas relações para essas duas entidades e o ConceptResolver pode não ter evidências suficientes para identificar o par de sintagmas nominais como correferentes.

**Tabela 3.5: Instâncias de relações usadas como características semânticas para a resolução de correferência.**

Noun phrase	Kobe Bryant	Kobe	Bryant	Lebron
athletesSuchAsAthletes	Basketball player	Basketball player		Basketball player
athleteHomeStadium	Staples Center		Staples Center	Quicken Loans Arena
athleteCoach	Byron Scott		Byron Scott	David Blatt
athletePlaysInLeague	NBA			NBA
athletePlaysForTeam		LA Lakers		Cleveland Cavaliers
athleteKnowAs			Kobe	Lebron James

Nesse mesmo cenário, considera-se que algumas características morfológicas (como similaridade de *string*, por exemplo) possam ser acrescentadas ao modelo de resolução de correferência. Assim, o modelo pode ter mais evidências de um par correferente de ENs, mesmo quando a NELL não tiver aprendido muito ainda sobre tais menções.

Então, uma abordagem híbrida que combina características linguísticas, semânticas e morfológicas foi implementada. Os resultados obtidos a partir da experimentação na NELL em inglês foram publicados em (DUARTE; HRUSCHKA, 2014a).

Dado um par de ENs ( $EN1, EN2$ ), as características morfológicas utilizadas para a abordagem apresentada neste trabalho são as seguintes:

1. DNL - Diferente Número de Letras: esse atributo binário tem valor 1 se  $EN1$  tem menos letras que  $EN2$  ou vice-versa. Se ambas as ENs tiverem o mesmo número de letras, o DNL recebe o valor 0. Assim,  $DNL(Kobe Bryant, Kobe) = 1$  e  $DNL(Bryant, Lebron) = 0$ ;
2. Subconjunto: esse atributo binário recebe valor 1 se  $EN1$  fizer parte de  $EN2$  ou vice-versa. Caso contrário, o atributo recebe valor 0. Assim,  $Subconjunto(Kobe, Kobe Bryant) = 1$  e  $Subconjunto(Kobe Bryant, Lebron) = 0$ .

3. Similaridade de *String*: é um atributo de valor real (normalizado de 0 a 1), obtido por meio de uma simples medida de similaridade entre *strings* na qual quanto mais próximo de 1, mais similares são EN1 e EN2.
4. Apelido: esse atributo recebe valor 1 se a menor EN é formado pelas letras iniciais e finais, ou somente iniciais, da EN de tamanho maior. Assim,  $Apelido(New\ York,\ NY) = 1$  e  $Apelido(Kobe\ Bryant,\ Kobe) = 0$ .
5. Covington *String Similarity* (CSS): é um atributo de valor real (normalizado de 0 a 1), obtido por meio da aplicação de um método de similaridade entre *strings* baseado no algoritmo de Covington (COVINGTON, 1996).
6. Proximidade: atributo de valor real (normalizado entre 0 e 1), obtido através do método baseado no algoritmo de proximidade JaroWinkler, disponível na biblioteca LingPipe (CARPENTER, 2007).

Essas 6 características linguísticas foram baseadas nas características da função objetivo, que consiste em detectar ENs correferentes na base da NELL.

Considerando a análise e a discussão do impacto da adição de características morfológicas no método atual de resolução de correferências da NELL, baseado em características semânticas (relações), havia duas alternativas para tal experimentação:

- **Primeira:** PAs 6 características linguísticas poderiam ser adicionadas ao ConceptResolver e, com isso, duas versões seriam executadas: uma somente com características semânticas (somente ConceptResolver) e outra somente com características morfológicas.
- **Segunda:** Assim como as características morfológicas, as semânticas poderiam ser usadas em um simples modelo de resolução de correferência para analisar o impacto da adição de novos recursos a NELL, sem ter que gastar muito tempo na adaptação do ConceptResolver para trabalhar com as características morfológicas.

A segunda alternativa foi escolhida como a mais adequada para a experimentação e 2 características (7<sup>a</sup> e 8<sup>a</sup>) foram definidas:

7. Número de relações que compartilham o mesmo valor de instância (CompartilhandoRelações): é um atributo de valor inteiro que simplesmente soma as instâncias de relações que compartilham o mesmo valor para ambos os sintagmas nominais EN1 e EN2. Considerando a Tabela 2.2 como exemplo, o cálculo seria:  $CompartilhandoRelações(Kobe$

$Bryant, Kobe) = 5$ ,  $CompartilhandoRelações(Kobe Bryant, Lebron) = 2$  e  $CompartilhandoRelações(Kobe, Bryant) = 6$ .

8. Média de relações que compartilham o mesmo valor de instância ( $MediaCompartilhamentoRelações$ ): é um atributo de valor real (normalizado de 0-1) que consiste na média das instâncias de relações que compartilham o mesmo valor nos dois sintagmas nominais, EN1 e EN2 em relações não vazias. Em outras palavras, para todas as  $n$  relações em que a NELL aprendeu algo para ambas ENs, o número de combinações é dividido por  $n$ . Considerando a Tabela 2.2 como exemplo, o cálculo seria:  $MediaCompartilhamentoRelações(Kobe Bryant, Kobe) = 5/6 = 0.83$ ,  $MediaCompartilhamentoRelações(Kobe Bryant, Lebron) = 2/6 = 0,33$  e  $MediaCompartilhamentoRelações(Kobe, Bryant) = 6/6 = 1$ .

### 3.3.1 Experimentos e Análise

Para os experimentos, foi construído um conjunto de dados a partir de extração manual de ENs da KB da NELL.

Esse conjunto é formado por 200 pares de ENs, para os quais o ConceptResolver detectou correferência. Metade do conjunto (100 pares de EN) são pares que o ConceptResolver classificou corretamente como sendo correferentes (verdadeiros positivos) e, os outros 100 pares, ele classificou erroneamente (falsos positivos).

Os falsos positivos e os falsos negativos foram verificados manualmente e rotulados. Usando a estratégia de validação cruzada (10-fold cross-validation), definiu-se uma tarefa de classificação binária, na qual o classificador deveria ser treinado para identificar os pares de sintagmas nominais como correferentes (sim) ou não correferentes (não).

Para executar os experimentos, foram utilizados 10 classificadores diferentes implementados pelo WEKA (HALL et al., 2009). Na Tabela 3.6, encontra-se a lista dos 10 classificadores e os resultados dos três experimentos:

- Exp.1: classificadores híbridos utilizam tanto características linguísticas como morfológicas. Os 8 atributos são utilizados.
- Exp.2: classificadores semânticos utilizam somente características semânticas. Somente 2 atributos são usados (7 e 8).
- Exp.3: classificadores morfológicos utilizam somente características morfológicas. Somente 6 atributos são usados (1 a 6).

**Tabela 3.6: Sumário dos resultados empíricos para as 3 configurações de experimentos. Os resultados são as médias obtidas utilizando o 10-fold cross-validation. Os melhores resultados estão em negrito**

Algorithms	Semantic	Morphologic	Confusion Matrix				F-score
			YES		NO		
			Right	Wrong	Right	Wrong	
LibSVM Linear	Exp.1		89	11	92	8	<b>0.905</b>
	Exp.2		82	18	91	9	0.865
		Exp.3	66	34	95	5	0.801
LibSVM Polinomial	Exp.1		86	14	90	10	<b>0.88</b>
	Exp.2		82	18	90	10	0.86
		Exp.3	57	43	100	0	0.775
Bayesian Logistic Regression	Exp.1		94	6	86	14	<b>0.9</b>
	Exp.2		92	8	81	19	0.865
		Exp.3	73	27	94	6	0.833
Bayes Net	Exp.1		91	9	93	7	<b>0.9</b>
	Exp.2		89	11	89	11	0.89
		Exp.3	72	28	96	4	0.838
Simple Logistic	Exp.1		89	11	91	9	<b>0.9</b>
	Exp.2		85	15	89	11	0.87
		Exp.3	72	28	93	7	0.823
Logistic Regression	Exp.1		89	11	93	7	<b>0.91</b>
	Exp.2		87	13	90	10	0.885
		Exp.3	74	26	93	7	0.833
Naïve Bayes	Exp.1		89	11	91	9	<b>0.9</b>
	Exp.2		94	6	86	14	<b>0.9</b>
		Exp.3	75	25	91	9	0.829
Voted Perceptron	Exp.1		86	14	86	14	<b>0.86</b>
	Exp.2		80	20	88	12	0.84
		Exp.3	77	23	93	7	0.849
Random Forest	Exp.1		92	8	93	7	<b>0.925</b>
	Exp.2		95	5	80	20	0.874
		Exp.3	76	24	90	10	0.829
J48	Exp.1		86	14	89	11	0.875
	Exp.2		94	6	84	16	<b>0.89</b>
		Exp.3	73	27	92	8	0.823

Importante ressaltar que o intuito dos experimentos não foi mostrar ou comparar o desempenho do ConceptResolver, mas coletar evidências empíricas para sustentar a hipótese de que uma abordagem híbrida de características semânticas e morfológicas pode ajudar a NELL a ser mais precisa na identificação de sintagmas nominais correferentes.

Os resultados apresentados na Tabela 3.6 revelam que os classificadores que utilizam apenas características morfológicas tendem a apresentar pior desempenho, com exceção do classificador *Voted Perceptron* Exp.3 (um em cada dez).

Por outro lado, os classificadores que usam características semânticas e morfológicas obtiveram o melhor desempenho em nove de cada dez classificadores. Já os classificadores que utilizam somente características semânticas apresentaram resultados intermediários, na maioria dos experimentos.

Vale notar que os classificadores morfológicos (Exp.3) produziram menos falsos negativos. Em sete, de dez execuções, tais classificadores foram os mais precisos na identificação de sin-



tagmas nominais não correferentes. Essa característica deve ser melhor explorada e pode ajudar a NELL a evitar decisões erradas.

Com base nos experimentos, é possível afirmar que há evidência empírica de que o uso de características semânticas e morfológicas pode impactar positivamente na base de conhecimento dinâmica da NELL. Além disso, é possível afirmar que nenhum classificador apresentou um desempenho muito melhor em relação aos outros. A maioria dos resultados são parecidos entre diferentes classificadores. Com isso, um trabalho futuro consiste no desenvolvimento de uma abordagem híbrida acoplada, de forma a explorar a independência entre ambas as abordagens.

# Capítulo 4

## CONCLUSÕES

---

---

O Never-Ending Learning (NEL) é um novo paradigma de aprendizado de máquina usado para resolver problemas complexos.

A NELL (Never-Ending Language Learning) está em execução, sem parar, desde 2010 visando aprender cada vez mais a língua inglesa. O objetivo deste trabalho foi o de desenvolver, através dos mesmos mecanismos, a NELL em português.

Em outras palavras, a NELL, também chamada de Leitura da Web, foi a motivação para a criação da Leitura da Web em português.

Tendo como objetivo principal a criação da Leitura da Web em Português, o primeiro objetivo específico foi a criação de uma base pré-processada (*all-pairs-data*) através de uma tarefa baseada em PLN.

Foram realizadas várias pesquisas, testes e implementações que possibilitaram a criação de uma *all-pairs-data* para a execução da NELL em português. Isso foi possível devido à disponibilização de 40 milhões de páginas web em português do *corpus* ClueWeb pelo grupo NELL da Carnegie Mellon University.

Inicialmente, foram realizados experimentos e análises somente no CPL (resultados publicados em (DUARTE; HRUSCHKA, 2014b)), os quais proporcionaram evidências empíricas suficientes para a continuidade do projeto.

Levantou-se também a necessidade de melhoria do processo de resolução de correferência de forma a tratar o aprendizado de categorias e de relações conjuntamente. A resolução de correferência já é realizada na NELL pelo ConceptResolver (KRISHNAMURTHY; MITCHELL, 2011), porém de forma voltada a somente relações.

A abordagem híbrida (DUARTE; HRUSCHKA, 2014a) foi desenvolvida e experimentada na

NELL em inglês e obteve bons resultados no que tange a união de ambas as abordagens (categorias e relações). A ideia não era substituir o ConcepResolver, mas apontar uma mudança necessária no processo já executado pelo componente.

Em seguida, foram realizados experimentos com todos os componentes da NELL. Os resultados apresentados também proporcionaram evidências empíricas da validade da proposta da Leitura da Web em português. Entretanto, o *all-pairs-data* gerado não permitiu com que o sistema evoluísse por muito tempo. O tamanho limitado do *all-pairs-data*, a não adição de novas páginas e os erros na etiquetação que levaram a extrações incorretas são as razões que justificam tais resultados.

Os resultados após a supervisão humana não foram atingidos durante o fechamento deste documento.

## 4.1 **Objetivos Alcançados**

Todos os objetivos, principais e específicos, foram alcançados conforme planejados. O principal foi a implementação da NELL em português. Para isso, foi desenvolvido um *all-pairs-data* a partir da extração de sentenças do *corpus* ClueWeb (primeiro objetivo específico) e uma abordagem híbrida para o tratamento de correferência de forma independente de língua (segundo objetivo específico).

Através de experimentos foram obtidas evidências empíricas da validade de ambas propostas. Parte dos resultados foram publicados em (DUARTE; HRUSCHKA, 2014b) e (DUARTE; HRUSCHKA, 2014a). O restante dos resultados serão publicados em breve e já constam neste documento.

## 4.2 **Contribuições e Limitações**

Como contribuições, seguem algumas publicações em anais de congressos:

- DUARTE, M. C.; HRUSCHKA, E. R. Exploring two views of coreference resolution in a never-ending learning system. In: *Proceedings of the 14th International Conference on Hybrid Intelligent*. Kuwait, 2014. (HIS'14), p. 261-266.
- DUARTE, M. C.; HRUSCHKA, E. R. How to read the web in portuguese using the never ending language learner's principles. In: *Proceedings of the 14th International Confe-*

*rence Intelligent Systems Design and Applications*. Okinawa, Japan, 2014. (ISDA'14), p. 162-167.

- DUARTE, M. C.; NICOLETTE, M. d. C.; HRUSCHKA, E. R. H. Minimização do impacto do problema de desvio de conceito por meio de acoplamento em ambiente de aprendizado sem fim. In: *Proceedings of the 8th Brazilian Symposium in Information and Human Language Technology*. Cuiabá, Brasil, 2011. (STIL'11), p. 134-143.

E a seguinte publicação em periódico:

- HRUSCHKA, E. R. H.; DUARTE, M. C.; NICOLETTI, M. C. Coupling as strategy for reducing concept-drift in never-ending learning environments. *Fundamenta Informaticae*, IOS Press, v. 124, n. 1, p. 47-61, 2013.

Além disso, seguem contribuições relacionadas ao primeiro e segundo objetivos específicos (a criação da base pré-processada, o *all-pairs-data*):

- proposta e discussão da tarefa de processamento utilizando PLN para o pré-processamento de páginas web em português obtidas através do ClueWeb;
- aplicação da proposta de pré-processamento baseada em PLN;
- criação de um *all-pairs-data* (fonte de extração de conhecimento da NELL) baseado em PLN;
- apresentação de análises comparativas empíricas com as abordagens anteriores reportadas na literatura;
- investigação sobre a criação de uma nova instância da NELL;
- discussões e análises sobre o *all-pairs-data* em português, que possui tamanho muito inferior ao do inglês. Essa contribuição será útil para futuras implementações que também possuam um *all-pairs-data* pequeno.
- apontamentos de melhorias no desenvolvimento do *all-pairs-data* em português;

Relacionadas ao terceiro objetivo específico (tratamento de correferência independente de língua), seguem as contribuições:

- proposta e discussão de abordagem híbrida de resolução de correferência independente de língua, baseada em características morfológicas, a partir de categorias, e em características semânticas, a partir de relações;
- apresentação de análises comparativas empíricas com abordagens anteriores reportadas na literatura;
- levantamento de evidências empíricas sobre a melhoria de resultados utilizando a abordagem híbrida;

### 4.3 **Trabalhos Futuros**

O primeiro trabalho futuro refere-se a análise dos resultados a serem obtidos com a supervisão humana, já em andamento. Após a análise será submetido um artigo sobre o assunto.

O trabalho futuro de maior prioridade é o desenvolvimento do componente DDT, com o qual objetiva-se melhorar o aprendizado em português e verificar a cobertura alcançada com o atual *all-pairs-data* em português mais detalhadamente.

O próximo trabalho futuro é o desenvolvimento de uma abordagem híbrida que utilize o acoplamento de características semânticas e morfológicas para o tratamento de correferências. Tal abordagem irá auxiliar o ConceptResolver a obter melhores resultados. Além disso, a versão híbrida com o uso de acoplamento será desenvolvida de forma a poder ser usada na NELL.

Paralelamente, se dará continuidade às investigações sobre melhores etiquetadores da língua portuguesa e o processo de *part-of-speech* na criação do *all-pairs-data* será aperfeiçoado objetivando a identificação de um maior número de sentenças de qualidade no *corpus*.

Os trabalhos futuros continuam na linha de investigar meios para que a NELL em português evolua o seu aprendizado. A continuidade dessa investigação, além de auxiliar na evolução do aprendizado sem-fim em português, também é útil para as próximas línguas a serem adicionadas a NELL.

## REFERÊNCIAS

---

---

- APPEL, A.; HRUSCHKA, E. Prophet – a link-predictor to learn new rules on nell. In: *Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on*. [S.l.: s.n.], 2011. p. 917–924.
- BEAN, D.; RILOFF, E. Unsupervised learning of contextual role knowledge for coreference resolution. In: *Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. [S.l.: s.n.], 2004. p. 297–304.
- BETTERIDGE, J. et al. Toward never ending language learning. In: *AAAI Spring Symposium*. [S.l.: s.n.], 2009.
- BHATTACHARYA, I.; GETOOR, L. A latent dirichlet model for unsupervised entity resolution. In: *SIAM INTERNATIONAL CONFERENCE ON DATA MINING*. [S.l.: s.n.], 2006.
- BHATTACHARYA, I.; GETOOR, L. Collective entity resolution in relational data. *ACM Trans. Knowl. Discov. Data*, ACM, New York, NY, USA, v. 1, n. 1, mar. 2007. ISSN 1556-4681. Disponível em: <<http://doi.acm.org/10.1145/1217299.1217304>>.
- BICK, E. *The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Aarhus University Press, 2000. ISBN 9788772889108. Disponível em: <<http://books.google.com.br/books?id=ISUGDvPg7hcC>>.
- BLUM, A.; MITCHELL, T. Combining labeled and unlabeled data with co-training. In: *Proc. of COLT*. [S.l.: s.n.], 1998.
- CALLAN, J.; HOY, M. *ClueWeb09 Data Set*. 2009. [Http://boston.lti.cs.cmu.edu/Data/clueweb09/](http://boston.lti.cs.cmu.edu/Data/clueweb09/).
- CARLSON, A. et al. Coupling semi-supervised learning of categories and relations. In: *Proc. of the NAACL HLT 2009 Workshop on Semi-supervised Learning for Natural Language Processing*. [S.l.: s.n.], 2009.
- CARLSON, A. et al. Toward an architecture for never-ending language learning. In: *Proceedings of the Twenty-Fourth Conference on Artificial Intelligence (AAAI 2010)*. [S.l.: s.n.], 2010.
- CARLSON, A. et al. Coupled semi-supervised learning for information extraction. In: *Proc. of WSDM*. [S.l.: s.n.], 2010.
- CARPENTER, B. Lingpipe for 99.99 *Proceedings of the 2nd BioCreative workshop*. Valencia, Spain: [s.n.], 2007.

- CASTAÑO, J.; ZHANG, J.; PUSTEJOVSKY, J. Anaphora resolution in biomedical literature. In: *In Proceedings of the 2002 International Symposium on Reference Resolution*. [S.l.: s.n.], 2002.
- CHEN, X.; SHRIVASTAVA, A.; GUPTA, A. Neil: Extracting visual knowledge from web data. In: *Computer Vision (ICCV), 2013 IEEE International Conference on*. [S.l.: s.n.], 2013. p. 1409–1416. ISSN 1550-5499.
- CHOMSKY, N. *Language and Problems of Knowledge*. [S.l.]: The Mit Press, 1988. 5–33 p.
- CLARK, P. et al. Reading to learn: An investigation into language understanding. In: *Proceedings of the 2007 AAAI Spring Symposium*. Menlo Park, California: The AAAI Press, 2007.
- COVINGTON, M. A. An algorithm to align words for historical comparison. *Comput. Linguist.*, MIT Press, Cambridge, MA, USA, v. 22, n. 4, p. 481–496, dez. 1996. ISSN 0891-2017. Disponível em: <<http://dl.acm.org/citation.cfm?id=256329.256333>>.
- CRISTEA, D.; IDE, N.; ROMARY, L. Veins theory: A model of global discourse cohesion and coherence. In: *In Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and of the 17th International Conference on Computational Linguistics (COLING/ACL98)*. [S.l.: s.n.], 1998. p. 281–285.
- CURRAN, J. R.; MURPHY, T.; SCHOLZ, B. Minimising semantic drift with mutual exclusion bootstrapping. *Proceedings of the Conference of the Pacific Association for Computational Linguistics*, p. 172–180, 2007.
- DAGAN, I.; ITAI, A. Automatic processing of large corpora for the resolution of anaphora references. In: *Proceedings of the 13th conference on Computational linguistics - Volume 3*. Stroudsburg, PA, USA: Association for Computational Linguistics, 1990. (Conference on Computational Linguistics'90), p. 330–332.
- DEEMTER, K. van; KIBBLE, R. On coreferring: Coreference in muc and related annotation schemes. *Comput. Linguist.*, MIT Press, Cambridge, MA, USA, v. 26, n. 4, p. 629–637, dez. 2000. ISSN 0891-2017. Disponível em: <<http://dl.acm.org/citation.cfm?id=971882.971888>>.
- DUARTE, M. C.; HRUSCHKA, E. R. Exploring two views of coreference resolution in a never-ending learning system. In: *Proceedings of the 14th International Conference on Hybrid Intelligent*. Kuwait: [s.n.], 2014. (HIS'14), p. 261–266.
- DUARTE, M. C.; HRUSCHKA, E. R. How to read the web in portuguese using the never-ending language learner's principles. In: *Proceedings of the 14th International Conference on Intelligent Systems Design and Applications*. Okinawa, Japan: [s.n.], 2014. (ISDA'14), p. 162–167.
- DUARTE, M. C.; NICOLETTE, M. d. C.; HRUSCHKA, E. R. Minimização do impacto do problema de desvio de conceito por meio de acoplamento em ambiente de aprendizado sem fim. In: *Proceedings of the 8th Brazilian Symposium in Information and Human Language Technology*. Cuiabá, Brasil: [s.n.], 2011. (STIL'11), p. 134–143.
- ETZIONI, O.; BANKO, M.; CAFARELLA, M. J. Machine reading. In: *Proceedings of the 2007 AAAI Spring Symposium*. Menlo Park, California: The AAAI Press, 2007.

ETZIONI, O. et al. Open information extraction from the web. *Commun. ACM*, ACM, New York, NY, USA, v. 51, n. 12, p. 68–74, dez. 2008. ISSN 0001-0782. Disponível em: <<http://doi.acm.org/10.1145/1409360.1409378>>.

FADER, A.; SODERLAND, S.; ETZIONI, O. Identifying relations for open information extraction. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2011. (EMNLP '11), p. 1535–1545. ISBN 978-1-937284-11-4. Disponível em: <<http://dl.acm.org/citation.cfm?id=2145432.2145596>>.

GARDNER, M. et al. Incorporating vector space similarity in random walk inference over knowledge bases. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar.: Association for Computational Linguistics, 2014.

GARDNER, M. et al. Improving learning and inference in a large knowledge-base using latent syntactic cues. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP 2013)*, 2013.

GE, N.; HALE, J.; CHARNIAK, E. A statistical approach to anaphora resolution. In: *In Proceedings of the Sixth Workshop on Very Large Corpora*. [S.l.: s.n.], 1998. p. 161–170.

HAGHIGHI, A.; KLEIN, D. Simple coreference resolution with rich syntactic and semantic features. In: *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3 - Volume 3*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2009. (Conference on Empirical Methods in Natural Language Processing'09), p. 1152–1161. ISBN 978-1-932432-63-3.

HALL, M. et al. The weka data mining software: An update. *SIGKDD Explor. Newsl.*, ACM, New York, NY, USA, v. 11, n. 1, p. 10–18, nov. 2009. ISSN 1931-0145. Disponível em: <<http://doi.acm.org/10.1145/1656274.1656278>>.

HEARST, M. A. Automatic acquisition of hyponyms from large text corpora. In: *Proc. of COLING*. [S.l.: s.n.], 1992.

HIRST, G. Discourse-oriented anaphora resolution in natural language understanding: a review. *Comput. Linguist.*, MIT Press, Cambridge, MA, USA, v. 7, n. 2, p. 85–98, abr. 1981. ISSN 0891-2017.

HOBBS, J. Readings in natural language processing. In: GROSZ, B. J.; SPARCK-JONES, K.; WEBBER, B. L. (Ed.). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1986. cap. Resolving pronoun references, p. 339–352. ISBN 0-934613-11-7.

HRUSCHKA, E. R.; DUARTE, M. C.; NICOLETTI, M. C. Coupling as strategy for reducing concept-drift in never-ending learning environments. *Fundamenta Informaticae*, IOS Press, v. 124, n. 1, p. 47–61, 2013.

IDE, N.; CRISTEA, D. A hierarchical account of referential accessibility. In: *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2000. (Association for Computational Linguistics'00), p. 416–424.



- IIDA, R.; INUI, K.; MATSUMOTO, Y. Capturing salience with a trainable cache model for zero-anaphora resolution. In: *Proceedings of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - Volume 2*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2009. (Association for Computational Linguistics'09), p. 647–655. ISBN 978-1-932432-46-6.
- KEHLER, A. et al. The (Non)Utility of Predicate-Argument Frequencies for Pronoun Interpretation. In: *Proceedings of the 2004 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*. [S.l.: s.n.], 2004. p. 289–296.
- KRISHNAMURTHY, J.; MITCHELL, T. M. Which noun phrases denote which concepts. In: *Proceedings of the Forty Ninth Annual Meeting of the Association for Computational Linguistics*. [S.l.: s.n.], 2011.
- LAFFERTY, J.; MCCALLUM, A.; PEREIRA, F. Conditional random fields: probabilistic models for segmenting and labeling sequence data. In: *ICML*. [S.l.: s.n.], 2001.
- LEVIN, M. et al. Citation-based bootstrapping for large-scale author disambiguation. *Journal of the American Society for Information Science and Technology*, v. 63, n. 5, p. 1030–1047, 2012. ISSN 1532-2890. Disponível em: <<http://dx.doi.org/10.1002/asi.22621>>.
- LIN, T.; MAUSAM; ETZIONI, O. Entity linking at web scale. In: *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction (AKBC-WEKEX)*. Montréal, Canada: Association for Computational Linguistics, 2012. p. 84–88. Disponível em: <<http://www.aclweb.org/anthology/W12-3016>>.
- LUO, X.; ZITOUNI, I. Multi-lingual coreference resolution with syntactic features. In: *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2005. (HLT '05), p. 660–667. Disponível em: <<http://dx.doi.org/10.3115/1220575.1220658>>.
- MITCHELL, T. et al. Never-ending learning. In: *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI-15)*. [S.l.: s.n.], 2015.
- MITCHELL, T. M. *Machine Learning*. 1. ed. New York, NY, USA: McGraw-Hill, Inc., 1997. ISBN 0070428077, 9780070428072.
- MITCHELL, T. M. *The discipline of machine learning*. White paper, cmu-ml-06-108. [S.l.], June 2006. Disponível em: <<http://www.cs.cmu.edu/~tom/pubs/MachineLearning.pdf>>.
- MITCHELL, T. M. et al. Theo: A framework for self-improving systems. *Arch. for Intelligence*, Lawrence Erlbaum Associates, p. 323–356, 1991.
- MODJESKA, N. N.; MARKERT, K.; NISSIM, M. Using the web in machine learning for other-anaphora resolution. In: *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2003. (EMNLP '03), p. 176–183. Disponível em: <<http://dx.doi.org/10.3115/1119355.1119378>>.

- MOHAMED, T. P. et al. Discovering relations between noun categories. In: *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. Edinburgh, Scotland, UK.: Association for Computational Linguistics, 2011. p. 1447–1455. Disponível em: <<http://www.aclweb.org/anthology/D11-1134>>.
- NG, V. Supervised noun phrase coreference research: The first fifteen years. In: *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2010. (ACL 10), p. 1396–1411. Disponível em: <<http://dl.acm.org/citation.cfm?id=1858681.1858823>>.
- NG, V.; CARDIE, C. Improving machine learning approaches to coreference resolution. In: *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2002. (Association for Computational Linguistics'02), p. 104–111.
- OLIVERIO, V.; JR, E. R. H. Contradiction detection and ontology extension in a never-ending learning system. In: *Advances in Artificial Intelligence–IBERAMIA 2012*. [S.l.]: Springer Berlin Heidelberg, 2012. p. 1–10.
- POESIO, M. et al. *ELERFED: Final report of the research group on Exploiting Lexical and Encyclopedic Resources For Entity Disambiguation*. [S.l.], 2007.
- POESIO, M. et al. Learning to resolve bridging references. In: *Proceedings of the 42Nd Annual Meeting on Association for Computational Linguistics*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2004. (ACL '04). Disponível em: <<http://dx.doi.org/10.3115/1218955.1218974>>.
- PONZETTO, S. P.; STRUBE, M. Exploiting semantic role labeling, wordnet and wikipedia for coreference resolution. In: *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2006. (Human Language Technologies-North American Chapter of the Association for Computational Linguistics'06), p. 192–199.
- PONZETTO, S. P.; STRUBE, M. Knowledge derived from wikipedia for computing semantic relatedness. *J. Artif. Int. Res.*, AI Access Foundation, USA, v. 30, n. 1, p. 181–212, out. 2007. ISSN 1076-9757.
- POON, H.; DOMINGOS, P. Joint inference in information extraction. In: *Proceedings of the 22Nd National Conference on Artificial Intelligence - Volume 1*. AAAI Press, 2007. (AAAI'07), p. 913–918. ISBN 978-1-57735-323-2. Disponível em: <<http://dl.acm.org/citation.cfm?id=1619645.1619792>>.
- QUINLAN, J. R.; CAMERON-JONES, R. M. Foil: A midterm report. In: *Proc. of ECML*. [S.l.: s.n.], 1993.
- RAVIKUMAR, P.; COHEN, W. W. A hierarchical graphical model for record linkage. In: *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*. Arlington, Virginia, United States: AUAI Press, 2004. (UAI '04), p. 454–461. ISBN 0-9749039-0-6. Disponível em: <<http://dl.acm.org/citation.cfm?id=1036843.1036898>>.

SAMADI, M.; VELOSO, M. M.; BLUM, M. Openeval: Web information query evaluation. In: *AAAI*. [S.l.: s.n.], 2013.

SILVA, J. a. et al. Out-of-the-box robust parsing of portuguese. In: *Proceedings of the 9th International Conference on Computational Processing of the Portuguese Language*. Berlin, Heidelberg: Springer-Verlag, 2010. (PROPOR'10), p. 75–85. ISBN 3-642-12319-8, 978-3-642-12319-1. Disponível em: <[http://dx.doi.org/10.1007/978-3-642-12320-7\\_10](http://dx.doi.org/10.1007/978-3-642-12320-7_10)>.

SINGLA, P.; DOMINGOS, P. Entity resolution with markov logic. In: *Proceedings of the Sixth International Conference on Data Mining*. Washington, DC, USA: IEEE Computer Society, 2006. (ICDM '06), p. 572–582. ISBN 0-7695-2701-9. Disponível em: <<http://dx.doi.org/10.1109/ICDM.2006.65>>.

SNOW, R. et al. Learning to merge word senses. In: *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*. [s.n.], 2007. Disponível em: <<http://aclweb.org/anthology/D07-1107>>.

STRUBE, M.; RAPP, S.; MÜLLER, C. The influence of minimum edit distance on reference resolution. In: *Proceedings of the Association for Computational Linguistics'02 conference on Empirical methods in natural language processing - Volume 10*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2002. (Conference on Empirical Methods in Natural Language Processing'02), p. 312–319.

TETREAULT, J. *Empirical evaluations of pronoun resolution*. Tese (Doutorado), Rochester, NY, USA, 2005. AAI3156835.

VIEIRA, T. *Never-Ending Language Metalearning: model management for CMU's Read The Web project*. Tese (Doutorado) — FEUP - Faculdade de Engenharia - Universidade Porto, 2015.

WANG, R. C.; COHEN, W. W. Character-level analysis of semi-structured documents for set expansion. In: *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3 - Volume 3*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2009. (EMNLP '09), p. 1503–1512. ISBN 978-1-932432-63-3. Disponível em: <<http://dl.acm.org/citation.cfm?id=1699648.1699697>>.

WINKLER, W. E. *The state of record linkage and current research problems*. [S.l.], 1999.

YANG, X.; SU, J. Coreference resolution using semantic relatedness information from automatically discovered patterns. In: *Annual Meeting-Association for Computational Linguistics*. [S.l.: s.n.], 2007. v. 45, n. 1, p. 528.

YANG, X.; SU, J.; TAN, C. L. Improving pronoun resolution using statistics-based semantic compatibility information. In: *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2005. (Association for Computational Linguistics'05), p. 165–172.

YANG, X. et al. Improving noun phrase coreference resolution by matching strings. In: *Proceedings of the First international joint conference on Natural Language Processing*. Berlin, Heidelberg: Springer-Verlag, 2004. (IJCNLP'04), p. 22–31.

---

YATES, A.; ETZIONI, O. Unsupervised methods for determining object and relation synonyms on the web. *Journal of Artificial Intelligence Research*, 2009.

ZELENKO, D. et al. Kernel methods for relation extraction. *Journal of Machine Learning Research*, v. 3, 2003.

ZHU, X. et al. Semi-supervised learning using gaussian fields and harmonic functions. In: *MACHINE LEARNING-INTERNATIONAL WORKSHOP THEN CONFERENCE*-. [S.l.: s.n.], 2003. v. 20, n. 2, p. 912.

## GLOSSÁRIO

---

---

**AMSF** – *Aprendizado de Máquina Sem-Fim.*

**All-Pairs-Data** – *Base pré-processada de páginas web. Nela são armazenadas todas as estáticas suficientes para a execução da NELL.*

**CB** – *Corpus Brasileiro - Possui aproximadamente 1 bilhão de palavras etiquetas no português brasileiro contemporâneo. Pertence à PUC-SP (Pontifícia Universidade Católica de São Paulo) e é apoiado pela FAPESP (Fundação de Amparo à Pesquisa do Estado de São Paulo). Foi usado parcialmente em testes iniciais para a construção do all-pairs-data.*

**CETEMFolha** – *é um corpus de textos da Folha de S. Paulo, processado pelo NILC (Núcleo Interinstitucional de Linguística Computacional de São Carlos), possui aproximadamente 24 milhões de palavras em português brasileiro.*

**CETEMPúblico** – *é um corpus com aproximadamente 180 milhões de palavras em português europeu. Os textos foram coletados do Jornal Público de Portugal.*

**CHAVE** – *Coleção de textos à qual o CETEMPúblico e o CETEMFolha fazem parte. Parte da coleção CHAVE foi utilizada em experimentos do RTWP. A coleção CHAVE é disponibilizada pela Linateca (<http://www.linateca.pt>)*

**CLN** – *Compreensão da Linguagem Natural.*

**CMU** – *Carnegie Mellon University.*

**CMU** – *Coupled Morphological Classifier - Componente da NELL que atua como classificador morfológico a partir dos resultados obtidos pelo CPL e CSEAL.*

**CPL** – *Coupled Pattern Learning - Componente da NELL que realiza o aprendizado a partir da extração de ENs e PTs em páginas web.*

**CSEAL** – *Coupled SEAL - Componente da NELL que realiza o aprendizado a partir de queries de padrões HTML na web.*

**Categoria** – São os classes da ontologia, em outras palavras, são os tipos de conhecimentos que o sistema deve aprender. Exemplos: pessoa, equipeEsportiva, fruta, emoção, etc.

**ClueWeb** – Corpus usado para a criação do all-pairs-data da NELL em português. Possui aproximadamente 40 milhões de páginas em português. Seu tamanho é considerado pequeno quando comparado à quantidade de páginas web utilizadas na NELL em inglês, aproximadamente 1.5 bilhões.

**ConceptResolver** – Componente da NELL projetado para atuar na resolução de referência a partir de características linguísticas semânticas.

**Corpora** – Conjunto de vários Corpus.

**Corpus** – Conjunto de textos.

**DC** – Departamento de Computação.

**DTT** – Dictionary Tinker Toy - Componente que visa traduzir todas as instâncias promovidas de EN da NELL em inglês para português, inicialmente, e em seguida para outras línguas.

**EN** – Significa Entidade Nomeada. Neste documento uma EN é definida como um substantivo. Exemplos: "São Paulo", "Brasil", "gripe", "futebol", etc.

**IA** – Inteligência Artificial.

**KB** – Knowledge Base (Base de Conhecimento).

**KI** – Componente da NELL que analisa os resultados obtidos de cada componente e decide quais candidatos serão promovidos.

**LX-Parser** – Etiquetador textual utilizado para a etiquetagem do ClueWeb no desenvolvimento do all-pairs-data.

**MR** – Machine Reading.

**MaLL** – Machine Learning Laboratory.

**Macroleitura** – Neste projeto, a macroleitura é definida como as estáticas sobre um texto, sem a preocupação da compreensão total desse texto.

**Microleitura** – Neste projeto, a microleitura é definida como uma análise precisa e detalhada de um texto.

**NELL** – Never-Ending Language Learning.

**NEL** – Never-Ending Learning.

**NP** – *Noun Phrase*.

**Ontologia** – *Estrutura de organização da KB da NELL*.

**PLN** – *Processamento de Linguagem Natural*.

**PPG-CC** – *Programa de Pós-Graduação em Ciência da Computação*.

**PRA** – *Componente da NELL o qual infere novas crenças a partir de análise de caminhos percorridos*.

**PT** – *Significa Padrões Textuais. Neste documento um PT refere-se a um conjunto de palavras encontrado antes ou depois de uma EN no aprendizado de categorias. No aprendizado de relações refere-se ao conjunto de palavras encontrado entre um par de ENs. Exemplos de PTs para categorias: "\_ é uma cidade", "países como \_", "\_ é uma doença", etc. Exemplos de PTs para relações: "\_ é um dos esportes mais famosos do \_", "\_ é a capital do \_", "\_ também conhecido como \_", etc.*

**RL** – *Rule Learner - Componente da NELL que infere cláusulas de Horn a partir da KB*.

**RTWP** – *Read The Web in Portuguese - Leitura da Web em Português. Componente desenvolvido inspirado no CPL para a leitura da web em português*.

**Relações** – *São os relacionamentos entre as categorias na ontologia. Exemplos: atletaJogaParaEquipeEsportiva(atleta, equipeEsportiva), musicoTocaInstrumento(músico, instrumento), etc.*

**Token** – *Segmento de texto. Exemplos: "Tom", "Mitchell", "futebol", "de", "a", "capital", etc.*

**UFSCar** – *Universidade Federal de São Carlos*.

**XLike** – *Atualmente chamado de XLime, é um grupo europeu que, dentre outras várias áreas, atua na extração de informação a partir da web. Juntamente com o grupo NELL da CMU, o XLime disponibilizou o corpus ClueWeb para o desenvolvimento deste projeto*.

# Apêndice A

## PRÉ-PROCESSAMENTO E CRIAÇÃO DE ALL-PAIRS-DATA A PARTIR DO CLUEWEB

---

---

### A.1 Contextualização

*All-pairs-data* é um conjunto de estatísticas suficientes sobre um corpus. No caso da Leitura da Web, o *all-pairs-data* é responsável por conter todas as estatísticas sobre ENs e PTs, tanto para categorias, quanto para relações.

Este apêndice especifica as tarefas executadas e as implementações necessárias durante o processo de criação do *all-pairs-data* em português.

O *All-Pairs-Data* apresentado neste projeto foi criado a partir do processamento de 40 milhões de páginas em português, obtidas através do *corpus* ClueWeb, disponibilizado pela Carnegie Mellon University (CMU).

Inicialmente, as páginas foram etiquetadas pelo part-of-speech LX-Parser ((SILVA et al., 2010)) e, em seguida, foram extraídas as ENs e os PTs para categorias e relações. Por fim, foram contabilizadas as ocorrências e as co-ocorrências das combinações de ENs e PTs para as categorias e os pares de ENs e de PTs para as relações.

### A.2 *Corpus* ClueWeb - Coleta de Páginas Web

A versão utilizada do ClueWeb é a ClueWeb09, que contém páginas web coletadas em 2009. Existe também uma versão mais recente, a ClueWeb12, referente às coletas realizadas em 2012. A opção pela versão ClueWeb09 é devido ao fato de essa versão somente possuir páginas em



português. A captura de páginas do ClueWeb09 foi executada pelo *crawler Sapphire*<sup>1</sup> entre Janeiro e fevereiro de 2009.

Foram usadas 29 milhões de URLs obtidas com duas técnicas diferentes na extração realizada pelo *Sapphire*: 1) dessas URLs, 20 milhões foram as URLs com pontuações maiores em uma lista de 200 milhões de páginas em inglês, extraídas por *crawler* de Janeiro à Junho de 2008; 2) 9 milhões de páginas foram obtidas a partir de pesquisa das que estariam no topo do *ranking* de motores de busca.

De forma resumida, as técnicas utilizadas pelo *crawler Sapphire* foram:

- AOL Query Log: as 1050 *queries* mais frequentes foram selecionadas do log do AOL. Mais 1050 *queries* foram selecionadas de forma randômica do log do AOL, de acordo com a frequência relativa. A maioria das páginas web extraídas eram em inglês, consequentemente a maioria das sementes adquiridas também. Das *queries* citadas, foram selecionados os 500 melhores (top 500) resultados retornados pelo Google, Yahoo! e MSN;
- DMOZ Category Names: foi criada uma lista de 2000 *queries* no DMOZ a partir das categorias de nomes. As 2000 categorias DMOZ mais largas, com até 3 de profundidade, foram usadas. Assim como no item anterior, foram selecionados os 500 melhores (top 500) resultados das páginas web obtidas pelo DMOZ a partir da verificação nos motores de busca Google, Yahoo! ou MSN.
- Queries* traduzidas: as *queries* do AOL e do DMOZ foram automaticamente traduzidas do inglês para 9 outras línguas através do Google Translate. Foram usados três buscadores Web dependendo do suporte que forneciam para tais línguas: Baidu, Google e Yahoo!. Para o português, somente o Google foi usado. Os 200 melhores (top 200) resultados para cada query foram selecionados.
- Queries* multilíngue Yahoo!: complementando as *queries* traduzidas citadas no item anterior e, como um recurso mais realista de *queries* multilíngues, foi criado um conjunto de 1.000 consultas mais frequentes em cada idioma, exceto árabe, a partir do *Yahoo Research Webscope Program*. As consultas foram coletadas pelo Yahoo! durante três meses em 2008.
- Queries* em Chinês: o chinês é segunda maior língua da coleção, devido à diferença foram realizadas consultas de páginas de Hong Kong e Taiwan, ao invés da China Continental.

---

<sup>1</sup><http://boston.lti.cs.cmu.edu/crawler/>

As páginas web em inglês foram codificadas em UTF-8. As demais páginas foram lidas e armazenadas de acordo com as codificações originais. Devido a isso, houve a necessidade da alteração da tarefa (1), citada na seção seguinte.

Além disso, páginas de algumas línguas continuam a ser extraídas, como é o caso do inglês. Já para português a coleta foi descontinuada devido à baixo número de páginas encontradas.

O ClueWeb foi criado pelo projeto Lemur, desenvolvido pelo Instituto de Tecnologia da Linguagem<sup>2</sup> da CMU - Estados Unidos.

### A.3 Desenvolvimento do *All-Pairs-Data* do corpus ClueWeb

Basicamente, um sistema de aprendizado sem-fim necessita de dois processos distintos: 1) a extração e o pré-processamento de páginas web e 2) o aprendizado. O *all-pairs-data* é responsável por armazenar as estatísticas finais, após a extração e o pré-processamento das páginas web. Sendo assim, o sistema de aprendizado sem-fim tem foco somente na tarefa de aprendizado.

O pré-processamento é a etapa de preparação das sentenças extraídas de páginas web armazenadas em formato .warc (formato fornecido pelo ClueWeb). Tanto para a extração de categorias como para a de relações semânticas, o pré-processamento pode ser dividido nas seguintes etapas:

- 1.Extração do texto puro a partir de arquivo .warc (as páginas são armazenadas em HTML);
- 2.Etiquetagem do texto extraído em (1) com o Tagger LX-Parser;
- 3.Criação de *tags* linguísticas para a extração de categorias e de relações;
- 4.Extração das *tags* encontradas;
- 5.Armazenamento de todas as estatísticas no *all-pairs-data*;

### A.4 Fluxogramas de processo de execução do pré-processamento do ClueWeb

Para melhor entender a **quarta etapa**, foram criados dois fluxogramas; um sobre o pré-processamento do *all-pairs-data* para as categorias e outro para as relações semânticas. Em

<sup>2</sup><http://boston.lti.cs.cmu.edu/Data/web08-bst/planning.html>

ambos a prioridade é encontrar ENs com a etiqueta PNM (Part of Name), as quais são normalmente nomes próprios. Em seguida, caso nenhuma PNM tenha sido extraída, tenta-se encontrar ENs que não foram identificadas como PNM.

### A.4.1 Fluxograma de processo para as categorias

O fluxograma para as categorias é apresentado na Figura A.1.

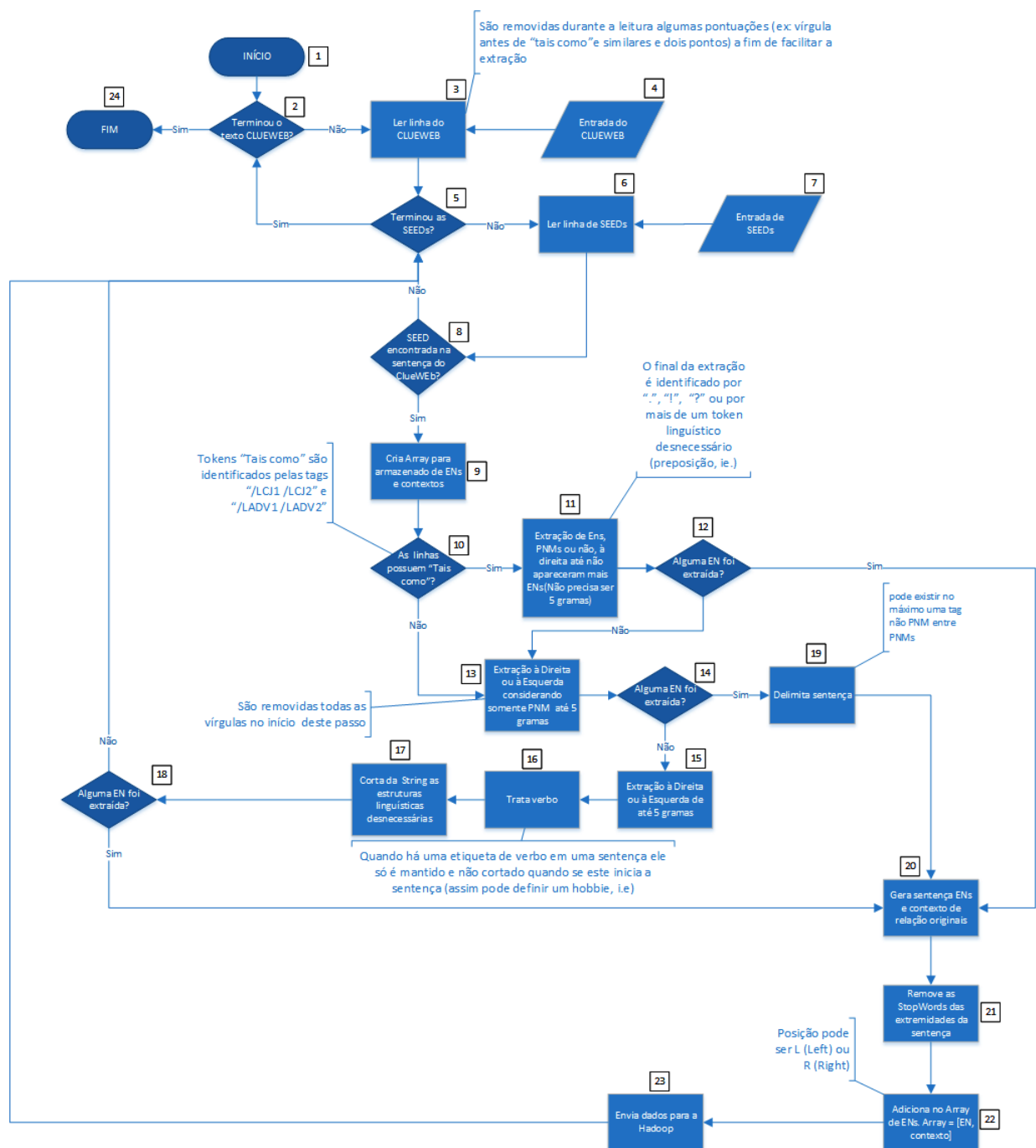


Figura A.1: Fluxograma de processo na extração de categorias

O processo é iniciado com duas entradas: o texto puro do ClueWeb, etiquetado pelo LX-Parser e as sementes de *tags* de categorias para a extração baseada em padrões linguísticos (lista criada na **terceira etapa**).

No fluxograma, a execução da terceira etapa é realizada em 1, 2, 3, 4, 5, 6 e 7. De 1 a 7, executa-se um *loop* dentro de outro, a fim de buscar ocorrências de *tags* linguísticas (sementes) na sentença lida. Em 3, algumas pontuações, como dois pontos, são removidas para facilitar a extração de ENs. Quando uma semente é encontrada na sentença lida (8), cria-se um vetor para que as ENs e os contextos sejam armazenados(9);

Em 10, verifica-se a existência de padrões que definam as expressões “tal como”, “tais como”, etiquetadas como LCJ1 e LCJ2 (*Multi-Word Adverbs*) ou LADV1 e LADV2 (*Multi-Word Conjunctions*). Caso as *tags* LCJ1 e LCJ2 ou LADV1 e LADV2 sejam encontradas, em 11, tenta-se extrair uma ou mais ENs à direita (sem tamanho definido), sejam elas PNM ou não. O corte é realizado quando se encontra um ponto final, interrogação, exclamação ou quando são encontrados dois *tokens* de PREP (preposição).

Caso não seja encontrada nenhuma das sementes solicitadas em 10 ou não tenha sido extraída nenhuma EN em 12, tenta-se, em 13, extrair uma EN com *tag* PNM com limite de 5 gramas. Se alguma EN for extraída em (14), realiza-se a delimitação de PNM em 19. A delimitação é um processo que não permite mais de um *token* com *tag* diferente de PNM entre *tokens* PNM.

O processo de delimitação de ENs ocorre da seguinte forma: como não é permitida a existência de dois ou mais *tokens* seguidos que não sejam PNM, caso sejam encontrados, realiza-se o corte de acordo com a direção de extração. Se a direção da extração é à direita, corta-se a sentença a partir do primeiro *token* seguido até o fim. Se a direção da extração é à esquerda, corta-se a sentença do início até o último. Cortar, aqui, significa remover.

Se nenhuma EN com *tag* PNM for extraída em 14, em 15, extrai-se uma subsentença, que tem como ponto de corte o número de gramas igual a 5 ou a ocorrência de pontos de exclamação, de interrogação ou final. Em 16, trata-se a sentença que possui algum *token* com etiqueta de verbo. Essa só é mantida se o verbo iniciar a sentença, caso contrário a sentença é cortada até o verbo.

Em 17 são cortadas as estruturas linguísticas desnecessárias (preposições, dígitos, etc.). Além de extrair tais etiquetas das extremidades da sentença, caso exista duas etiquetas indesejadas na sequência, essa também será cortada.

Se ocorrer a extração de ENs em 12, 14 ou 18, o próximo passo acontecerá em 20, com

exceção de 14, que antes delimita a sentença em 19. A partir de 20, gera-se a EN original, removendo-se a etiquetação. Em 21, são removidas as *StopWords* das extremidades das ENs e, em 22, são adicionadas as combinações de ENs e de contexto no vetor criado em 9, passando, então, para o envio das sentenças do vetor para o processo de contagem do Hadoop em 23.

Finalmente, o processo se reinicia em 5 com a leitura de outras sementes e, se não houver saída em 5, outra sentença do ClueWeb é lida em 3, o que reinicia a leitura de *tags*. Se as sentenças do ClueWeb estiverem acabado (2), o processo é finalizado em 22.

### A.4.2 Fluxograma de Relações Semânticas

O fluxograma de relações semânticas não difere muito do fluxograma de categorias, porém há algumas diferenças em sua estrutura devido ao fato de extrair um par de ENs de uma sentença, tendo um contexto de relação entre elas (PTs de relação). Além disso, não há extração de padrões de “tais como” e similares. O fluxograma é apresentado na Figura A.2.

A estrutura de um *loop* dentro de outro é a mesma, sendo assim, os números 1, 2, 3, 4, 5, 6 e 7 representam a entrada de texto ClueWeb e a de *tags*/sementes (*seeds*). Vale lembrar que a leitura de sementes é reiniciada a cada nova sentença lida do ClueWeb e o final do processo acontece quando não há mais sentenças a serem lidas(22).

Quando a semente lida é encontrada na sentença do ClueWeb em 8, um vetor para o armazenamento das ENs e contextos de relações semânticas é criado em 9.

Em 10, tenta-se a extração das ENs da direita e da esquerda (par de ENs) que sejam PNMs. Caso ambas sejam extraídas (11) ou apenas uma delas (13), realiza-se a delimitação da sentença, processo que não permite a existência de duas ou mais etiquetas que não sejam PNMs que estejam na sequência, caso contrário, realiza-se o corte.

Em 14, caso alguma EN não seja PNM (13), acontece a extração de sentença(s) de até 5 gramas, havendo a delimitação de pontuação (interrogação, exclamação e ponto final).

Já em 15, o *token* etiquetado como verbo só é mantido caso esteja no início da sentença. E, em 16, cortam-se as estruturas linguísticas desnecessárias, as mesmas citadas para as categorias.

Se ambas as ENs forem extraídas, as sentenças originais são geradas em 18 e, em 19, são removidas as *StopWords* de ambas as ENs. Em 20, as ENs e o contexto de relações são adicionados ao vetor criado em 9.



## A.5 Uso do Hadoop

O Hadoop foi usado na implementação da quarta etapa do pré-processamento para a criação do *All-Pairs-Data*. A implementação básica do Hadoop pode ser dividida em duas partes: *Map* e *Reduce*.

O *Map* é responsável por organizar o dado que será contabilizado, neste caso, as ocorrências de ENs e de PTs de categorias, além de pares de ENs e PTs de relações.

O *Reduce* é responsável por contar as ocorrências organizadas em *Map*.

A abordagem usada foi simples. Considerou-se cada linha como uma sentença única, sem divisão entre ENs e PTs ou pares de ENs e PTs, ficando assim: [EN, PT, X ] para categoria ou [ENLeft, PT, ENRight, X] para relação semântica (etapa do *Map*).

Em seguida, cada sentença é contada, na qual X passa a valer o número de ocorrência a cada iteração do Hadoop (etapa do *Reduce*).

As implementações *Map* e *Reduce* para categorias são apresentada na Figura A.3 e Figura A.4, respectivamente:

```

25 public static class ContextMatcher extends Mapper<LongWritable, Text, Text, LongWritable> {
26
27
28     private final static LongWritable one = new LongWritable(1);
29
30     public void map(LongWritable key, Text value, Context hadoopContext) throws IOException, InterruptedException {
31         //matches and count matches
32
33         //Read the line
34         ContextReader contextReader = new ContextReader();
35         //Read the line to context - class Context has particularities of context
36         ToCategory.Context context = new ToCategory.Context();
37
38         while ((context = contextReader.getNextContext()) != null) {
39
40             //This replace prepare the "tag" ON "tag" to be extracted if the sentence has a "," before.
41             ContextInTaggedSentence cits = new ContextInTaggedSentence(context,
42                 HelperMethods.RemoveDuplicateBlanks(value.toString())
43                 .replace(" ,//PNT tag/LCJ1 tag/LCJ2", " tag/LCJ1 tag/LCJ2")
44                 .replace(" ,//PNT tag/LADV1 tag/LADV2", " tag/LADV1 tag/LADV2")
45                 .replace(" ,//PNT tag/LCJ1 tag/LCJ2", " tag/LCJ1 tag/LCJ2")
46                 .replace(" ,//PNT tag/LADV1 tag/LADV2", " tag/LADV1 tag/LADV2")
47                 .replace("LADV2 ://PNT", "LADV2")
48                 .replace("LCJ2 ://PNT", "LADV2")
49                 .replace("//PNT .*//PNT", "//PNT"));
50
51             if (cits.isContextInSentence()) {
52
53                 //Here the "while" is necessary because a seed can extract more than one EN of the same sentence
54                 List<ObjectToCollectStringsToCategory> namedEntity = new ArrayList();
55                 namedEntity = cits.getNamedEntity();
56
57                 if (namedEntity.size() > 0) {
58                     int i = 0;
59                     while (i < namedEntity.size()) {
60                         hadoopContext.write(new Text(namedEntity.get(i).getNamedEntity()
61                             + " " + cits.showStringThatMatchesContext(namedEntity.get(i).getContext(), namedEntity.get(i).getEntityOrientation())
62                             + " "), one);
63                         i++;
64                     }
65                 }
66             }
67         }
68     }
69 }

```

Figura A.3: Implementação em Java do *Map* (Hadoop) para categorias

No código apresentado na Figura A.3, entre linhas 29 e 69 é processado o que é necessário para a criação da sentença [EN, PT, X ]. O mais importante acontece nas linhas 59, 60 e 61, pois nelas a sentença é enviada ao Hadoop pela variável "Text" em ("new Text..."), como pode ser visto na linha 59 e na variável da chamada do método na linha 29 ("Text value"). As linhas entre 41 e 48 são necessárias no tratamento do "tais como" (fluxograma de categorias): se houver uma vírgula antes de sementes que definam o "tais como", ela é removida para que o processamento seja simplificado, pois as ENs a serem extraídas podem ser demarcadas com base na vírgula.

```
58 public static class ContextCounter extends Reducer<Text, LongWritable, Text, LongWritable> {
59
60     private LongWritable result = new LongWritable();
61
62     //sums all matches
63     public void reduce(Text key, Iterable<LongWritable> values, Context hadoopContext) throws IOException, InterruptedException {
64
65         int sum = 0;
66
67         for (LongWritable val : values) {
68             sum += val.get();
69         }
70
71         result.set(sum);
72         hadoopContext.write(key, result);
73     }
74 }
75
76
```

**Figura A.4: Implementação em Java do *Reduce* (Hadoop) para categorias**

A variável "Text" é recebida no *Reduce*, como pode ser visto na Figura A.4 (linha 63) e, nesse método, a cada iteração do Hadoop (que depende do número de núcleos e do número de sentenças obtidas), a contagem das sentenças é executada.

A implementação do Hadoop para relações semânticas não possui grandes alterações de categorias, tendo mudanças apenas no formato do vetor que envia a sentença do Hadoop, e não há tratamento de "tais como".



```
23 public class AllPairs_Relation {
24
25     public static class ContextMatcher extends Mapper<LongWritable, Text, Text, LongWritable> {
26
27         private final static LongWritable one = new LongWritable(1);
28
29         public void map(LongWritable key, Text value, Context hadoopContext) throws IOException, InterruptedException {
30             //matches and count matches
31
32             ContextReader contextReader = new ContextReader();
33             ToSemanticRelation.RelationContext relationContext = new ToSemanticRelation.RelationContext();
34
35             while ((RelationContext = contextReader.getNextRelationContext()) != null) {
36
37                 RelationContextInTaggedSentence cits = new RelationContextInTaggedSentence(RelationContext,
38                     HelperMethods.RemoveDuplicateBlanks(value.toString().replace("/PNDM .*//PNT", "/PNDM")));
39
40                 if (cits.isContextInSentence()) {
41                     List<ObjectToCollectStringsToRelation> namedEntity = new ArrayList();
42                     namedEntity = cits.getNamedEntity();
43
44                     if (namedEntity.size() > 0) {
45                         int i = 0;
46                         while (i < namedEntity.size()) {
47                             hadoopContext.write(new Text(namedEntity.get(i).getNamedEntity_left()
48                                 + " " + cits.showStringThatMatchesContextOfRelations(namedEntity.get(i).getRelationContext())
49                                 + " " + namedEntity.get(i).getNamedEntity_right() + " "), one);
50                             i++;
51                         }
52                     }
53                 }
54             }
55         }
56     }
}
```

**Figura A.5: Implementação em Java do Map (Hadoop) para relações semânticas**

Na Figura A.5, apresenta-se a implementação para as relações semânticas do *Map*, já o *Reduce* é idêntico à Figura A.4, pois a sentença é tratada da mesma forma; sendo única, sem divisão.