UNIVERSIDADE FEDERAL DE SÃO CARLOS

CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

IDENTIFICAÇÃO AUTOMÁTICA DE EQUIVALÊNCIA DE CONCEITOS EM DIFERENTES IDIOMAS PARA APRENDIZADO SEM FIM

SILVIO CARLOS MARINO

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

São Carlos - SP
Junho/2019

UNIVERSIDADE FEDERAL DE SÃO CARLOS

CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

IDENTIFICAÇÃO AUTOMÁTICA DE EQUIVALÊNCIA DE CONCEITOS EM DIFERENTES IDIOMAS PARA APRENDIZADO SEM FIM

SILVIO CARLOS MARINO

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de São Carlos, como parte dos requisitos para a obtenção do título de Mestre em Ciência da Computação, área de concentração: Inteligência Artificial
Orientador: Dr. Estevam R. Hruschka Jr.

São Carlos - SP
Junho/2019

UNIVERSIDADE FEDERAL DE SÃO 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 Dissertação de Mestrado do candidato Silvio Carlos Marino, realizada em 24/06/2019:

Profa. Dra. Heloisa de Arruda Camargo
UFSCar

Prof. Dr. Estevam Rafael Hruschka Junior
UFSCar

Prof. Dr. Jaime dos Santos Cardoso
U.Porto

Certifico que a defesa realizou-se com a participação à distância do(s) membro(s) Estevam Rafael Hruschka Junior, Jaime dos Santos Cardoso e, depois das arguições e deliberações realizadas, o(s) participante(s) à distância está(ao) de acordo com o conteúdo do parecer da banca examinadora redigido neste relatório de defesa.

Profa. Dra. Heloisa de Arruda Camargo

Ao meu filho Miguel

AGRADECIMENTO

A Deus, pela vida, pela oportunidade de conhecer suas qualidades e seus propósitos para com o futuro. Obrigado por permitir estudar temas interessantes e desafiadores.

Ao orientador, Estevam Hruschka Jr., que me orienta desde a iniciação científica. Sou muito grato por sua paciência, confiança, cuidado e entusiasmo. Obrigado por ser um modelo de profissional.

Aos meus pais, José e Mercedes, que me deram a vida e fizeram todo o possível para que eu pudesse estudar em uma universidade pública federal. E também, de forma especial, a minha esposa Bruna, pela compreensão em todos os esses anos.

Aos amigos e colegas de trabalho, da Secretaria de Informática da UFSCar, que de alguma forma ajudaram com palavras, motivação, compreensão e por dividir suas experiências.

Obrigado a todos os membros do MaLL (Machine Learning Laboratory) pelo companheirismo, por partilhar um pouco de sua pesquisa e como lidaram com os desafios que enfrentaram.

Aos professores, do Departamento de Computação, por transmitir um pouco do seu conhecimento para nos capacitar a desenvolver os próprios conceitos e habilidades. Aos técnicos que realizam diversas tarefas que facilitam o andamento do trabalho.

O que for que fizerem, trabalhem nisso de toda alma, como para Deus, e não para homens

Colossenses 3:23

RESUMO

A Inteligência Artificial (IA) refere-se a uma máquina capaz de sistematizar e automatizar tarefas que requerem inteligência quando realizadas por humanos. Com IA pode ser possível criar um processo capaz de tomar decisões com uma margem de erro cada vez menor. Neste sentido, foram criados o projeto Read The Web e um sistema de computação de aprendizado sem fim chamado NELL: Never-Ending Language Learning. NELL realiza um processo de aprendizado sem fim para aprender a ler a web cada vez melhor. Com o sucesso de obtenção do conhecimento na língua inglesa, o sistema foi modelado para aprender a extrair fatos de páginas web em outras línguas, tais como: português, francês e espanhol. No entanto, o conhecimento aprendido nos diferentes idiomas não está diretamente relacionado. Sendo assim, a principal motivação da pesquisa é desenvolver um mecanismo capaz de transferir o conhecimento de uma base para outra, utilizando aprendizado de máquina para automaticamente dizer se os conceitos aprendidos em idiomas diferentes são os mesmos. Os resultados dos experimentos, com a utilização de redes neurais, C4.5 e XGBoost como modelo de aprendizado de máquina, mostram um ganho significativo em comparação com o simples uso de dicionários.

Palavras-chave: Inteligência Artificial, Aprendizado de Máquina, Redes Neurais, C4.5, XGBoost, Equivalência de Conceitos.

ABSTRACT

Artificial Intelligence (AI) refers to a machine capable of systematizing and automating tasks that require intelligence when performed by humans. With AI it may be possible to create a decision-making process with an ever-smaller margin of error. In this sense, the project Read the Web and an endless learning computing system called NELL (Never-Ending Language Learning) were created. NELL performs an endless learning process to learn how to read the web better and better. With the success of obtaining knowledge in English, the system was modeled to learn how to extract facts from web pages in other languages, such as Portuguese, French and Spanish. However, the knowledge learned in the different languages is not directly related. Therefore, the main motivation of the research is to develop a mechanism capable of transferring knowledge from one base to another, using machine learning to automatically tell if concepts learned in different languages are the same. The results of the experiments, with the use of neural networks, C4.5 and XGBoost as a model of machine learning, show a significant gain in comparison with the simple use of dictionaries.

Keywords: Artificial Intelligence, Machine Learning, Neural Networks, C4.5, XGBoost, Equivalence of Concepts.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AM – Aprendizado de Máquina

API – *Application Programming Interface*

CDM – Conjunto de Dados Morfossintático

CDS1 – Conjunto de Dados Semântico extraído da Base1 da NELL

CDS2 – Conjunto de Dados Semântico extraído da Base2 da NELL

CDMS1 – Conjunto de Dados Morfossintático e Semântico composto pela junção dos atributos do CDM e do CDS1

CDMS2 – Conjunto de Dados Morfossintático e Semântico composto pela junção dos atributos do CDM e do CDS2

CDM+S1 – Conjunto de Dados Morfossintático e Semântico que armazena o resultado do melhores experimentos no CDM e no CDS1

CML – *Coupled Morphological Learner*

CPL – *Coupled Pattern Learner*

EN – Entidades Nomeadas

IA – Inteligência Artificial

LSTN – *Long Short-Term Memory*

MC – Matriz de Confusão

MLP – *Multilayer Perceptron*

NB – *Naive Bayes*

NEIL – *Never Ending Image Learner*

NELL – *Never Ending Language Learning*

OntExt – *Ontology Extension*

PoS – *Part-of-Speech*

PRA – *Path Ranking Algorithm*

PT – Padrões Textuais

RN – Rede Neural

RNN – *Recurrent Neural Network*

RTW – *Read The Web*

RTWP – *Read The Web in Portuguese*

SEAL – *Set Expander for Any Language*

SUMÁRIO

Capítulo 1

INTRODUÇÃO

1.1 Contextualização

Um bebê com 7 a 11 meses é capaz de engatinhar. Poucos meses depois, normalmente de 11 a 18 meses, a criança é capaz de andar. Com o tempo, aprende a falar, ler, escrever, praticar um esporte, exercer uma profissão, dirigir um veículo, entre outros. De fato, o aprendizado humano é um processo sem fim, que melhora quanto maior a frequência que uma tarefa é realizada.

Durante muitos anos procuramos entender como pensamos, como aprendemos e principalmente como construir máquinas que sejam capazes de tomar decisões para execução de uma tarefa. Quando falamos em Inteligência Artificial (IA), nos referimos a uma máquina capaz de sistematizar e automatizar tarefas "que exigem inteligência quando executadas por pessoas." (KURZWEIL, 1990)

Com a inteligência artificial pode ser possível construir um processo capaz de evoluir constantemente, uma vez que seja possível que, a cada nova informação apresentada as decisões podem ser tomadas com uma margem de erro cada vez menor. Neste sentido foi criado o projeto Read The Web (RTW), um sistema computacional de aprendizado sem fim (chamado NELL: Never-Ending Language Learning), iniciado em 2008.

NELL (MITCHELL et al, 2018) é um sistema que conta com uma base de conhecimento inicial (manualmente definida) composta por categorias e relações. Exemplos de categorias presentes na base de conhecimento da NELL são: cidade, pessoa, empresa, país, produto, atleta, esporte, time esportivo, etc. E alguns

exemplos de relações são: moraEm(pessoa, cidade), produzProduto(empresa, produto), etc.

Com base nas categorias e relações da base de conhecimento inicial, bem como num conjunto de instâncias (entre 10 e 20 instâncias) para cada categoria e cada relação, a NELL realiza seu processo de aprendizado sem fim buscando aprender a ler a web cada vez melhor, para poder extrair mais fatos cada vez com mais precisão. O mecanismo de aquisição de conhecimento se baseia no processo de aprendizagem humana. Inicialmente ocorre a assimilação de informações básicas e com o passar do tempo informações mais complexas são assimiladas, a partir de dados complementares das informações previamente aprendidas.

Com o sucesso de obtenção do conhecimento da NELL em inglês, o sistema foi modelado para aprender a extrair fatos de páginas web em outras línguas. Com isso, atualmente a NELL possui uma base de conhecimento para cada idioma já incorporado (inglês, português, francês e espanhol), com diferentes volumes de dados. No entanto, o conhecimento aprendido nos diferentes idiomas não está diretamente relacionado.

Quando pensamos em uma pessoa que domina diversos idiomas é lógico imaginar que o conhecimento adquirido em uma língua pode ser utilizado em qualquer idioma. Por exemplo, quando lemos um tutorial, em português, sobre como fazer uma torta de maçã, aprendemos a realizar essa atividade em qualquer idioma que dominamos. Isso ocorre porque temos uma forma única de armazenar e relacionar o conhecimento adquirido. Novamente pensando no ser humano, quando conhecemos o conceito associado a uma palavra, em português, e desejamos saber o conceito equivalente, em outro idioma, utilizamos um dicionário. Embora a utilização do dicionário seja bastante útil, há alguns problemas que vale destacar.

O primeiro problema ocorre quando o dicionário não tem a palavra ou expressão que desejamos traduzir. Por exemplo, dificilmente vamos conseguir encontrar no dicionário a tradução para a expressão “Cidade que nunca dorme” equivalente a cidade de “Nova York”. Isso também ocorre na tradução de filmes: “As Patricinhas de Beverly Hills” é a tradução do título do filme em inglês “Clueless”. Alguns nomes de animais também não encontramos com facilidade nos dicionários, por exemplo, “sharp shinned hawk” é conhecido no Brasil como “gavião miúdo”. Além disso, outra situação ocorre quando o dicionário fornece uma lista muito extensa de traduções e não é possível identificar qual delas é a mais adequada para

representar o conceito de interesse. Por último, pode acontecer de a tradução deixar dúvidas no que diz respeito à representação do que desejamos traduzir. Por exemplo, ao procurar pela cidade de Nova York em um dicionário inglês-português é comum encontrar como tradução Nova Iorque. Entretanto, pode haver dúvidas se a tradução é referente a cidade ou ao estado.

1.2 Motivação e Objetivos

A principal motivação deste trabalho de mestrado é ter um mecanismo para preencher uma base de conhecimento, em determinada língua, com o conteúdo de uma base de conhecimento em outra língua.

Tendo em mente os problemas destacados com a simples utilização dos dicionários, foram definidos como objetivos: a definição de uma abordagem, baseada em aprendizado de máquina, capaz de dizer se conceitos aprendidos pela NELL em idiomas diferentes são equivalentes; a implementação de um protótipo computacional que identifica conceitos equivalentes presentes nos conjuntos de dados extraídos das bases de conhecimento da NELL; e a comparação do protótipo com a simples utilização de um dicionário. Como parte deste trabalho de mestrado, serão utilizados apenas dois idiomas: Inglês e Português.

1.3 Metodologia de Desenvolvimento do Trabalho

A metodologia de trabalho de criação de um método capaz de dizer se com dois conceitos (En, Pt), En em inglês e Pt em português, pode-se afirmar que Pt é equivalente a En, envolve as seguintes tarefas principais:

- Escolha das categorias da NELL que serão utilizadas para validação do modelo;
- Definição dos atributos do conjunto de dados com características morfossintáticas e/ou semânticas;
- Preenchimento dos conjuntos com dados;

- Execução de modelos de AM, em uma amostra, para escolha do mais promissor;
- Execução do modelo de aprendizado de máquina nos conjuntos de dados;

1.4 Organização do Trabalho

Este documento é organizado de forma a apresentar, no capítulo 1, o problema, a motivação, as justificativas e os objetivos. No capítulo 2 é considerado a fundamentação teórica, que serve de base para o correto entendimento do projeto. No capítulo 3 são abordadas as tarefas necessárias para a identificação automática de equivalência de conceitos. No capítulo 4 são mostrados os resultados obtidos e suas respectivas análises. Por último, no capítulo 5, são apresentadas as conclusões, limitações e trabalhos futuros.

Capítulo 2

FUNDAMENTAÇÃO TEÓRICA

2.1 Considerações Iniciais

O Aprendizado de Máquina (AM) busca o desenvolvimento de programas de computador que possam evoluir à medida que são expostos a novas experiências (MITCHELL, 1997). Seu principal objetivo 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).

2.2 Tipos de Aprendizado

O aprendizado de máquina de interesse nesse trabalho é o indutivo. A ideia é que qualquer classe descoberta que aproxima bem um determinado conceito (função alvo) usando um conjunto suficientemente grande de exemplos de treinamento, também aproximará bem a função alvo sobre os outros exemplos não observados (MITCHELL, 1997).

Tradicionalmente na literatura, o aprendizado indutivo é subdividido em: aprendizado supervisionado e o não-supervisionado. O aprendizado supervisionado envolve a aprendizagem de uma função a partir de exemplos de suas saídas. O

aprendizado não-supervisionado envolve a aprendizagem de padrões na entrada, quando não são fornecidos valores de saída específicos.

Em Zhu et al (2003), os autores citam mais um tipo de aprendizado: o semissupervisionado. Uma das definições, para esse tipo de aprendizado, é utilizar um pouco da abordagem supervisionada e um pouco da abordagem não supervisionada. A partir de uma pequena amostra das saídas, ele é capaz de extrair a função (aprendizado supervisionado) para aprendizagem de padrões na entrada (aprendizado não-supervisionado).

2.3 Naive Bayes

Os classificadores Bayesianos são classificadores estatísticos que podem predizer a probabilidade de uma amostra pertencer a uma determinada classe, ou hipótese. Estes classificadores estão fundamentados no teorema de Bayes. Um dos classificadores Bayesianos mais difundidos na literatura é o Naive Bayes (NB). O NB assume que o efeito do valor de um atributo sobre a classe é independente dos valores dos outros atributos. (NIRMALA; VENKATESWARAN; KUMAR, 2017).

2.4 Árvore de Decisão – C4.5

Uma árvore de decisão usa a estratégia dividir para conquistar. Um problema complexo é dividido em problemas mais simples, aos quais recursivamente é aplicada a mesma estratégia. Essa é a ideia básica por trás do C4.5. Uma árvore de decisão abrange todo o espaço de instâncias. Com isso, uma árvore de decisão pode fazer predições para qualquer exemplo de entrada (FACELI et al, 2011).

2.5 Rede Neural

O cérebro humano processa informações de uma forma totalmente diferente de um computador convencional. Pode-se então dizer que o cérebro é um sistema de processamento de informação altamente complexo, não linear e paralelo. Ele consegue organizar seus constituintes estruturais, neurônios, de forma a realizar certos processamentos muito mais rápidos que os computadores convencionais. O cérebro humano realiza de forma rotineira tarefas de reconhecimento, como o reconhecimento de um rosto familiar em uma cena que não é familiar, em aproximadamente 100 a 200 ms, ao passo que tarefas de complexidade muito menor podem levar dias para serem executadas em computador convencional (HAYKIN, 2007).

De modo simples, redes neurais são modelos computacionais inspirados no sistema nervoso dos seres vivos (SILVA; SPATTI; FLAUZINO, 2010). Normalmente a rede é implementada utilizando componentes eletrônicos ou simulada através de programação em um computador digital.

Haykin (1999) oferece a seguinte definição de uma rede neural vista como uma máquina adaptativa: Uma rede neural é um processador maciçamente paralelamente distribuído constituído de unidades de processamento simples, que têm a propensão natural para armazenar conhecimento experimental e torná-lo disponível para o uso. Ela se assemelha ao cérebro em dois aspectos:

- O conhecimento é adquirido pela rede a partir de seu ambiente através de um processo de aprendizagem.
- Forças de conexão entre neurônios, conhecidas como pesos sinápticos, são utilizadas para armazenar o conhecimento adquirido.

O processo de aprendizagem ocorre através da modificação dos pesos sinápticos da rede de forma ordenada para alcançar um determinado objetivo. A rede neural extrai seu poder computacional através de sua estrutura paralelamente distribuída e de sua habilidade de aprender e, consequentemente, de generalizar. A generalização permite a rede produzir saídas adequadas para entradas que não estavam presentes na fase de aprendizagem.

Geralmente, são definidas três classes de arquiteturas ou estruturas de rede: as redes alimentadas adiante com camada única, as redes alimentadas diretamente com múltiplas camadas, e as redes recorrentes.

Nas redes alimentadas adiante com camada única (Figura 2.1), os neurônios estão organizados na forma de camadas. Há uma camada de entrada de nós de fonte que se projeta sobre uma camada de saída. Os nós de fonte não são contabilizados porque não realizam nenhuma computação (HAYKIN, 2007).

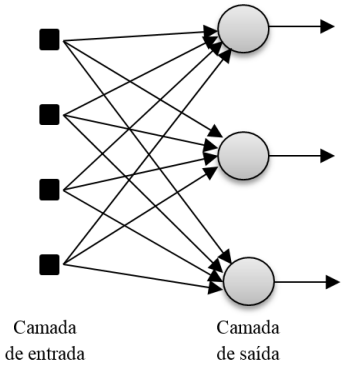

Figura 2.1 Rede com uma única camada.

As redes alimentadas diretamente com múltiplas camadas (Figura 2.2), são constituídas pela presença de uma ou mais camadas escondidas de neurônios. São empregadas na solução de diversos tipos de problemas, tais como aqueles relacionados à aproximação de funções, classificação de padrões, otimização, robótica, etc (SILVA; SPATTI; FLAUZINO, 2010).

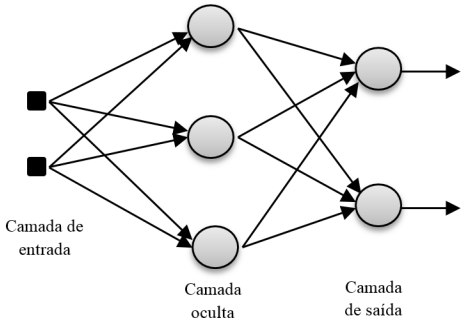

Figura 2.2 Rede com múltiplas camadas

As redes recorrentes (Figura 2.3) se distinguem de uma rede alimentada adiante por ter pelo menos um laço de realimentação (HAYKIN, 1999). Realimentação é uma situação onde a saída de um neurônio é realimentada na sua própria entrada.

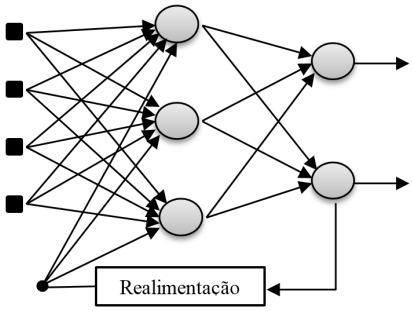

Figura 2.3 Redes recorrentes

2.5.1 Perceptron Camada Única

Haykin (2007) menciona que o perceptron é a forma mais simples de uma rede neural usada para classificação de padrões que se encontram em lados opostos de um hiperplano, ou seja, linearmente separáveis. Com um único neurônio é possível classificar apenas padrões com duas hipóteses. Ao expandir a camada de saída para mais neurônios é possível realizar a classificação de mais classes, desde que elas sejam linearmente separáveis.

O perceptron é composto de um único neurônio com pesos sinápticos que podem ser adaptados de iteração para iteração. O treinamento supervisionado do Perceptron consiste em ajustar os pesos e o bias de modo a se obter a classificação desejada. Bias, também chamado de limiar, é um valor que determina a posição do hiperplano no espaço x (BISHOP, 2005). Para a adaptação do limiar juntamente com os pesos é possível considerá-lo como sendo o peso associado a uma conexão e adaptar o peso relativo a essa entrada.

Uma saída é produzida quando um padrão é apresentado à rede. Após medir a distância entre a resposta atual e a desejada, são realizados os ajustes apropriados nos pesos das conexões de modo a reduzir esta distância, procedimento conhecido como Regra Delta (WIDROW; HOFF, 1960).

2.5.2 Perceptron Múltiplas Camadas

As redes de múltiplas camadas consistem de um conjunto de unidades sensoriais que constituem a camada de entrada, uma ou mais camadas ocultas de nós computacionais e, finalmente, uma camada de saída. O sinal de entrada é propagado para frente através da rede, camada por camada (SILVA; SPATTI; FLAUZINO, 2010).

O treinamento ocorre de forma supervisionada através de um algoritmo conhecido como algoritmo de retropropagação de erro (error back-propagation), que é baseado na regra de aprendizagem por correção de erro. De forma simples, a retropropagação de erro consiste de dois passos através das diferentes camadas da rede: um passo para frente, a propagação, e um passo para trás, a retropropagação. No passo para frente, um padrão de atividade (vetor de entrada) é aplicado aos nós sensoriais e o resultado se propaga para frente através da rede. Nesse passo, os pesos sinápticos são fixos. Depois que um conjunto de saídas é produzido pela rede, ocorre o passo de retropropagação, que ajusta os pesos sinápticos de trás para frente, de acordo com uma regra de correção de erro. A diferença entre os valores de saída produzidos e desejados para cada neurônio da camada de saída indica o erro cometido pela rede. Esse sinal de erro é propagado contra a direção das conexões sinápticas. Os pesos são ajustados para fazer que a resposta real fique mais próxima da resposta desejada (FACELI et al, 2011).

Simon Haykin (2007) afirma que o desenvolvimento do algoritmo de retropropagação representa um marco nas redes neurais, pois fornece um método computacional eficiente para o treinamento de perceptrons de múltiplas camadas. Apesar de não se poder afirmar que o algoritmo de retropropagação forneça uma solução ótima para todos os problemas resolúveis, ele acabou com o pessimismo sobre a aprendizagem de máquinas em múltiplas camadas.

2.6 Never-Ending Language Learning – NELL

Nesta seção será apresentada a definição da NELL, um sistema de aprendizado semissupervisionado que aprende a ler as páginas web. NELL, sigla

para Never-Ending Language Learning, é parte de um projeto de pesquisa chamado de “Read the Web”, ou “Leitura da Web” em português, da universidade de Carnegie Mellon.

Carlson et al. (2010) mencionam que o objetivo principal do projeto é desenvolver um sistema de aprendizado sem fim que roda 24 horas por dia, sete dias por semana. Esse sistema é capaz de extrair informações de páginas web e melhorar seu desempenho a cada dia.

Carlson et al. (2009) explicam que o conhecimento da NELL é organizado em categorias, tais como: animais, cidades, empresas, doenças, entre outros. Os relacionamentos ocorrem em pares dessas categorias, por exemplo, “animal desenvolve doença”.

A Figura 2.1 destaca dois pilares da NELL: a extração de conhecimento (linha magenta) e a base de conhecimento (linha verde). Mitchell et al. (2018) descrevem diversas partes de extração de conhecimento, como: CPL, SEAL, CML, PRA, OntExt, OpenEval, NEIL, entre outros.

CPL (Coupled Pattern Learner) usa estatísticas de co-ocorrência entre frases e padrões contextuais para aprender a extrair padrões para cada predicado de interesse. Depois, esses padrões são utilizados para encontrar instâncias adicionais para cada predicado.

SEAL (Set Expander for Any Language) usa pacotes de instâncias conhecidas para inferir novas instâncias ao redor desses pacotes. Os dois métodos parecem similares. No entanto, no CPL o foco está nos padrões textuais como X comevegetais Y. No CSEAL é considerado os padrões que formam o código HTML. Por isso que o CSEAL é capaz de extrair padrões de qualquer linguagem. Além disso, enquanto CPL extrai padrões de um conjunto estático de páginas web, CSEAL usa pesquisas do Google para buscar informações.

CML (Coupled Morphological Learner) classifica frases nominais baseadas nas características morfológicas. Um exemplo dessas características é o sufixo “burgh” que geralmente é associado a nome de cidades como “Petersburg”, ou “Johannesburg”. Esse componente geralmente é utilizado como um algoritmo de suporte, que confirma os fatos descobertos por outros módulos.

PRA (Path Ranking Algorithm) lida com a extração de conhecimento diretamente da base de dados usando caminhadas aleatórias para inferir possíveis

novos fatos. Devido a constante expansão da base de conhecimento da NELL, esse método tornou-se obsoleto.

OntExt (Ontology Extension) foi adicionado ao processo de extração de conhecimento para preencher a lacuna de descobrimento de novas relações. Primeiro, é feito a extração de sentenças que contém instâncias de duas categorias. Depois, as sentenças são usadas para construir uma matriz de contexto de co-ocorrência na qual é aplicado o agrupamento (que representa o novo relacionamento a ser proposto). Antes de introduzir as novas relações no sistema é preciso fazer uma revisão manual por um especialista.

OpenEval é muito parecido com o CML. Geralmente é utilizado como algoritmo de suporte para assegurar a precisão dos outros módulos. Resumidamente, OpenEval, reúne instâncias da categoria envolvida e as utiliza para pesquisar na web. Depois, extrai o contexto e as palavras-chave das instâncias para assegurar a confiança dos fatos propostos.

Never Ending Image Learner (NEIL), distanciou-se da NELL e possui um site de projeto separado (http://neil-kb.com/). NEIL adota uma abordagem diferente das técnicas de extração de informações da NELL. Em vez de texto na web, NEIL concentra sua atenção na imagem visual e nas relações observadas entre diferentes objetos identificados nessas imagens. Tal como nos outros componentes, os relacionamentos entre instâncias podem ser forçados a satisfazer novos relacionamentos. Nas edições de rotulagem de carros, por exemplo, esperamos que as rodas estejam presentes na imagem extraída, já que a relação "Roda é parte do carro" foi previamente estabelecida. A aquisição semântica de imagens é feita através de ferramentas de indexação baseadas em texto, como o Google Image Search.

De forma geral, é possível dizer que o retângulo verde, da Figura 2.1, contém os módulos responsáveis pela extração de conhecimento e inserção nos candidatos a fato; e o quadrado magenta contém os módulos que lidam com a criação de novas instâncias na base de conhecimento. As novas instâncias dos candidatos a fatos são coletadas e analisadas pelo integrador de conhecimento, que decide se elas devem ser promovidas para fatos. Note que, a adição de novos candidatos não significa que eles serão imediatamente promovidos a fatos na base de conhecimento. Pode ser necessário mais de uma iteração até o integrador de conhecimento ter evidências suficientes para a promoção. Na Figura 2.4 também é possível ver que, promoções

de novas informações, dão início a uma nova iteração de extração. Isso significa que o foco não é aprender tudo de uma única vez.

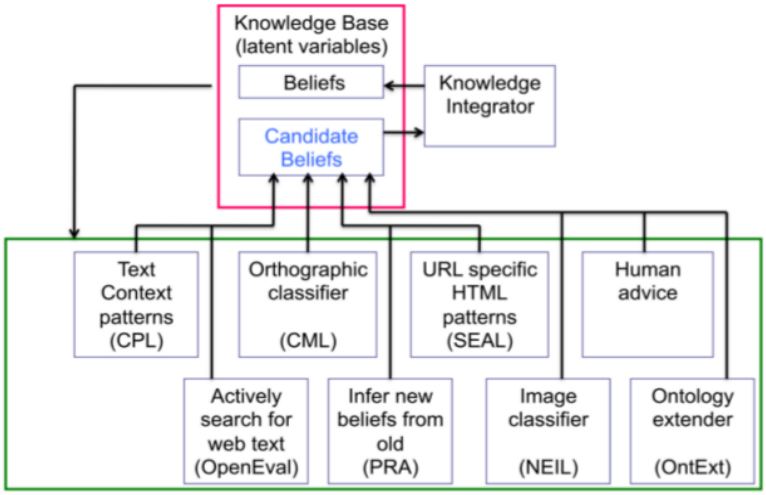

Figura 2.4 Arquitetura da NELL. A informação que foi coletada pelos módulos é sugerida como candidato a fato e depois (pode ou não ser) promovida a fato pelo integrador de conhecimento. De (MITCHELL et al, 2018).

Como qualquer sistema de aprendizado de longa duração, a NELL enfrenta diversos obstáculos como inconsistência de dados e desvio semântico, que causam uma acumulação de erros. Esses obstáculos são descritos pelos autores (CURRAN; MURPHY; SCHOLZ, 2007) e podem ser mitigados por restringir o processo de aprendizado. Uma forma de fazer isso, é por forçar os diversos componentes de extração a concordar um com o outro. Além disso, é possível verificar o tipo das relações, para assegurar que suas instâncias, de fato, pertencem às categorias marcadas para elas.

Na NELL, a extração de informação é feita utilizando aprendizado semissupervisionado e as instâncias de uma categoria são utilizadas como sementes para futuras extrações de novas informações da mesma categoria, conforme detalhado em (WANG; COHEN, 2009) e (CARLSON et al, 2009). O primeiro passo da extração de informação, lida com a recuperação da informação. Isso é feito através de um conjunto estático de 500 milhões de páginas web e 100 mil buscas diárias no Google, conforme explanado por (MITCHELL et al, 2015). O segundo passo, lida com a transformação da informação em conhecimento estruturado. Isso pode ser feito utilizando técnicas de Part-of-Speech (POS). Cai et al (2018) mencionam que na técnica de POS, a sentença é dividida em estruturas do

discurso, tais como pronomes, verbos, adjetivos, entre outros; e características semânticas, como sujeito, objeto, predicado, etc. Nesse ponto, o conteúdo das informações sofre de um problema chamado correferência, que envolve a capacidade das palavras representarem conceitos diferentes. Em (FLACH, 2012) o autor ataca esse problema usando uma abordagem de aprendizado supervisionado. Entretanto, na NELL, esse problema é resolvido utilizando aprendizado semissupervisionado, conforme apresentado por (KRISHNAMURTHY; MITCHELL, 2011).

Em 2009, teve início o projeto de leitura da web em português, com o sistema Read The Web in Portuguese (RTWP) apresentado em (DUARTE, 2011). RTWP é um sistema criado com base na NELL, que aprende entidades nomeadas (EN) e padrões textuais (PT) a partir de páginas web. No trabalho, EN abrange palavras da classe dos substantivos. Além disso, a autora define que PT são os padrões textuais que acompanham uma EN, após ou antes. Os resultados da primeira versão do RTWP mostraram, através de evidências empíricas, que a leitura da web em português é viável.

2.6.1 Base de Dados

Nesse trabalho foram utilizados os dados de duas bases da NELL. O valor de cada atributo das bases da NELL é composto por uma palavra ou um conjunto de palavras. Quando há mais de uma palavra, o caractere “_” é utilizado para separação.

2.6.1.1 Base1 da NELL

A Base1 da NELL, possui apenas informações que são consideradas como fatos. Dentre todos os atributos, foram utilizados apenas “Entity”, “Relation” e “Value”. A Base1 em Inglês contém mais de 2,3 milhões de linhas. A Base1 em português contém pouco mais de 850 mil linhas. A Tabela 2.1 apresenta uma amostra dos dados da Base1, em inglês, para “cats”.

Tabela 2.1 – Dados da Base1, em inglês, para o conceito "cats".

Base1 – amostra de dados para cats		
Entity	Relation	Value
concept:mammal:cats	concept:animaldevelopdisease	concept:disease:arthritis
concept:mammal:cats	concept: animaleatfood	concept:condiment:chocolate
concept:mammal:cats	concept: animaleatvegetable	concept:vegetable:lettuce
concept:mammal:cats	concept: animalistypeofanimal	concept:animal:house_pets
concept:mammal:cats	concept:animaldevelopdisease	concept:disease:hyperthyroidism

2.6.1.2 Base2 da NELL

A Base2 da NELL, contém informações que ainda estão esperando por mais evidências para se tornarem fatos. A base possui apenas os atributos: "Subject", "Verb" e "Object". A Base2, em inglês, contém mais de 185 milhões de linhas. A Base2, em português, contém pouco 50 milhões de linhas A Tabela 2.2 apresenta uma amostra dos dados da Base2, em inglês, para "cats".

Tabela 2.2 – Dados da Base2, em inglês, para o conceito "cats".

Base2 – amostra de dados para cats		
Subject	Verb	Object
cat	fancy in	southern africa
cat	torture	mouse
cat	was really	furry thing
cat	pages when	kittie

2.7 Pesquisas Relacionadas

Há diversas pesquisas que abordam um ou mais temas relacionados ao objeto de estudo deste trabalho. Nesta seção são destacados apenas os que mais contribuem para a pesquisa.

2.7.1 XGBoost – eXtreme Gradient Boosting

Em (CHEN; GUESTRIN, 2016), os autores descrevem que árvores de aumento de gradiente são muito utilizadas como método de aprendizado de máquina. Apresentam um sistema de árvore de aumento de gradiente, chamado de XGBoost, que tem sido utilizado, com êxito, por vários pesquisadores para alcançar excelentes resultados na solução de problemas desafiadores de AM. XGBoost é um algoritmo que lida muito bem com dados esparsos. Além disso, os autores mencionam que XGBoost consegue ser escalonado para bilhões de exemplos usando menos recursos do que os sistemas existentes.

2.7.2 Aumento de Gradiente no Processamento de Linguagem Natural

Em (COLLINS; KOO, 2005), os autores descrevem como um algoritmo baseado em aumento de gradiente é utilizado para problemas de ranking em processamento de linguagem natural. O algoritmo aproveita a natureza esparsa dos dados utilizados. No final houve um ganho relativo de 13% na redução de erro para o processamento dos dados do Wall Street Journal.

2.7.3 Uso de Aprendizado Profundo para Entender Comandos Fornecidos através de Linguagem Natural para Robôs

Em (MARTINS; CUSTÓDIO; VENTURA, 2018) os autores abordam que utilizar linguagem natural para dar instruções para robôs é uma tarefa desafiadora. No entanto, esse objetivo foi alcançado utilizando, utilizando Recurrent Neural Network (RNN) e Long Short-Term Memory (LSTN), duas arquiteturas de aprendizado profundo.

2.7.4 Rede Neural para Processamento de Palavras em Inglês

Em (KE; HAGIWARA, 2015), os autores apresentam como utilizaram redes neurais para processar textos em inglês. Foi utilizado uma rede neural com 5 camadas ocultas para extrair do texto sentenças, frases, palavras e conceitos com objetivo de responder a um questionário igual ao aplicado a alunos. Dos 495 pontos

possíveis, a rede neural conseguiu uma média de 276 pontos, enquanto a média dos alunos foi de 263.

2.7.5 Rede Neural para Processamento de Linguagem Natural

Goldberg (2016), aborda a aplicação de redes neurais para tarefas de processamento de linguagem natural. Em seu trabalho, o autor conclui que redes neurais conseguem aprender muito bem. Dessa forma, os pesquisadores interessados em linguagem natural podem aproveitar as vantagens do poder das redes neurais em seus trabalhos.

2.7.6 Combinação de Bases de Conhecimento em Diferentes Idiomas

Em (GONZÁLEZ J.; HRUSCHKA E. R.; MITCHELL T. M., 2017), os autores descrevem como lidaram com o problema de identificar conceitos equivalentes, nos idiomas português e inglês, assumindo que eles compartilham a mesma ontologia de categorias e relações. No estudo, foram usadas duas bases de conhecimento aprendidas independentemente a partir de diferentes páginas web escritas respectivamente na língua inglesa e na língua portuguesa. Foi utilizada uma abordagem baseada no PageRank personalizado (que pode ser considerado como a medida de similaridade que caracteriza a vizinhança de um nó X em um grafo) e uma técnica de inferência para descobrir caminhos relevantes comuns através das bases de conhecimento. Constatou-se que a técnica de inferência proposta identifica eficientemente caminhos relevantes com resultados melhores que a simples utilização de dicionários na maioria das categorias testadas.

Para facilitar a comparação entre os resultados do PageRank personalizado com o objeto de estudo desse trabalho, foram feitos experimentos adicionais que usam as mesmas categorias e os mesmos exemplos, positivos e negativos, que os autores utilizaram.

2.7.7 Análise de Correferência

Quando analisamos o conhecimento adquirido em um determinado idioma, é possível encontrar casos em que o objeto, que representa uma informação

aprendida, acaba sendo nomeado de diversas maneiras. Estes casos são chamados de correferentes e são muito relevantes para o aprendizado sem fim.

Em (DUARTE M.C.; HRUSHCKA E.R., 2014) os autores descrevem como as características semânticas e morfológicas (dentre elas a diferença no número de palavras e a similaridade de *strings*) foram combinadas na resolução do problema de correferência da base de conhecimento da NELL. Evidências empíricas foram obtidas para mostrar que combinar características morfológicas e semânticas em um modelo híbrido pode impactar de forma positiva a base de conhecimento da NELL.

Em (MANSANO A.F.; HRUSHCKA E.R.; PAPA J.P., 2018) é abordado que o conhecimento sobre certa entidade em uma base textual pode estar distribuído entre suas diversas denominações e por isso a análise de correferência no paradigma de aprendizagem da NELL é muito importante. Os autores apresentam um método para combinar diferentes vetores de recursos como uma tarefa de otimização, executado por técnicas meta-heurísticas e redes neurais, a fim resolver o problema de resolução de referência. Os experimentos mostraram que a metodologia obtém melhores resultados quando comparado à performance de extração de características individuais.

2.7.8 Mapeamento de Verbos em Diferentes Línguas para Relações de Base de Conhecimento

Em (WIJAYA D.T.; MITCHELL T.M, 2016) os autores ressaltam a importância de um recurso verbal que faça o mapeamento dos verbos em diferentes linguagens para relações da base de conhecimento em diferentes idiomas. O artigo apresenta uma abordagem escalável para construir esse recurso através de um corpus de texto da web. Dado um corpus de texto em qualquer idioma e qualquer base de conhecimento é possível produzir um mapeamento das frases verbais do idioma para as relações da base de conhecimento. O recurso foi aplicado de forma eficaz na base da NELL, em inglês e português.

Capítulo 3

IDENTIFICAÇÃO AUTOMÁTICA DE EQUIVALÊNCIA DE CONCEITOS

3.1 Considerações Iniciais

Neste Capítulo, será abordada a trajetória para a criação do método capaz de dizer se dois conceitos (En, Pt), En em inglês e Pt em português, são equivalentes. Para isso, foi necessário escolher as categorias, realizar a manipulação dos dados da NELL, definir a estrutura dos conjuntos de dados, preencher os conjuntos com dados e escolher o modelo de aprendizado de máquina.

3.2 Categorias

O trabalho apresentado no item 2.7.6 – Combinação de Bases de Conhecimento em Diferentes Idiomas faz uso das categorias: animal, ator, cidade, escritor, esporte, filme, país e pessoa. Para facilitar a comparação dos modelos, as mesmas categorias foram utilizadas. A única exceção é país, que por possuir poucos exemplos positivos foi substituído pela categoria alimento. A Tabela 3.1 apresenta a quantidade de relações de cada categoria.

Tabela 3.1 – Total de relações em cada categoria.

Total de relações por categoria	
Categoria	Total de Relações
Alimento	50
Animal	60
Ator	123
Cidade	146
Escritor	121
Esporte	38
Filme	34
Pessoa	118

3.3 Definição da Estrutura dos Conjuntos de Dados para o Modelo de Aprendizado de Máquina

Com base nos trabalhos reportados na literatura, apresentados no item 2.7.5 - Rede Neural para Processamento de Palavras em Inglês e no item 2.7.6 - Combinação de Bases de Conhecimento em Diferentes Idiomas, propôs-se a utilização de três conjuntos de dados para treinamento do modelo de aprendizado de máquina: morfossintática, semântica e morfossintática com semântica.

O primeiro conjunto de dados, chamado de CDM, tem um número fixo de atributos, construídos a partir das características morfossintáticas. O CDS1 possui uma quantidade variável de atributos, extraídos a partir das características semânticas, existentes na Base1. O CDS2 possui uma quantidade fixa de atributos, provenientes das características semânticas, existentes na Base2. Os últimos conjuntos de dados, CDMS1 e CDMS2, possuem os atributos do CDM junto com os atributos do conjunto semântico, existentes no CDS1 e CDS2 respectivamente.

3.3.1 Conjunto de Dados com Características Morfossintáticas - CDM

O conjunto de atributos com características morfossintáticas foi definido com base no comportamento humano. Geralmente, tentamos encontrar algumas regras para dizer se duas coleções de palavras, em idiomas diferentes, são equivalentes. Alguns dos parâmetros que podemos utilizar são: a quantidade de palavras entre os dois candidatos, o resultado da consulta a um dicionário e a similaridade entre eles.

A diferença na quantidade de palavras entre os dois candidatos é um fator que pode ser utilizado para tomar a decisão. No geral, esperamos que um candidato em inglês com uma única palavra possua, aproximadamente, o mesmo número de palavras em português.

Além do que, a tarefa ficará mais fácil se um dicionário puder ser consultado. Se o candidato em português aparecer na lista de traduções, há mais segurança para dizer que a equivalência é verdadeira.

Finalmente, a similaridade entre os candidatos pode ser comparada. Há várias palavras em inglês e português que possuem grafias idênticas (“banana”, “banana”) ou semelhantes (“cat”, “gato”). Isso ocorre com nomes de pessoas, lugares, animais, frutas, objetos, entre outros.

Dessa forma, o conjunto de dados morfossintático possui os seguintes atributos: “En”, “Pt”, “Dif Qtd Palavras En - Pt”, “Posição no Dicionário”, “Similaridade” e “Equivalente”. A Tabela 3.2 apresenta uma descrição para cada um dos atributos.

Tabela 3.2 – Detalhamento dos atributos do conjunto de dados morfossintático – CDM.

Detalhamento dos atributos do CDM	
Atributo	Descrição
En	Candidato em inglês.
Pt	Candidato em português.
Dif Qtd Palavras En - Pt	Diferença da quantidade de palavras do candidato em inglês e o candidato em português.
Posição no Dicionário	Valor numérico que informa a posição que o candidato em português aparece na lista de traduções (do candidato em inglês) no dicionário. Se não for encontrado, o valor retornado é -1.
Similaridade	Valor obtido através do algoritmo que compara similaridade entre cadeias de caracteres. Retorna um valor indicativo (0 até 1) de semelhança.
Equivalente	Para exemplos positivos, o valor do campo é 1. Para exemplos negativos, o valor do campo é 0.

3.3.2 Conjuntos de Dados com Características Semânticas

Os conjuntos de dados com características semânticas, CDS1 e CDS2, observam o domínio em que uma palavra está inserida. Por exemplo, para o conceito “gatos”, em português, é esperado encontrar em seu domínio: “animal de estimação”, “doença”, “peludo”, “rato”, entre outros. Supõe-se que algo similar acontece com a tradução, “cats”.

Cada categoria da NELL possui um conjunto específico de relações em seu domínio. Como a maioria das relações existentes na base em inglês possui seu correspondente na base em português, a ideia é utilizar cada relação existente na Base1 como atributo do conjunto de dados semânticos CDS1. A Tabela 3.3 apresenta uma amostra das relações do domínio “animals”, em inglês, e seu equivalente “animais”, em português.

Tabela 3.3 – Amostra das relações do domínio "animals", da base em inglês, e seu equivalente "animais", da base em português.

Amostra das relações dos domínios animals e animais	
Animals	Animais
animaleatfood	animalcomealimento
animaleatvegetable	animalcomevegetais
animalistypeofanimal	animaledamesmaracadoanimal
animaldevelopdisease	animaldesenvolvedoenca
agentcontributedtocreativework	agentcontributedtocreativework
agentcontrols	agentcontrols
agentcreated	agentecriou
agenthaswebsite	agentetemwebsite
agenthierarchicallyaboveagent	agenthierarchicallyaboveagent
agenthierarchicallybelowagent	-
synonymfor	sinonimode
targetinbombing	-
wascollidedin	-

A partir da tabela, é possível observar que a maioria das relações do domínio em inglês possui seu equivalente na base em português. As relações que não possuem tradução para o português, foram tratadas durante o preenchimento dos dados.

A Tabela 3.4 apresenta a quantidade de atributos do CDS1, extraídos a partir da Base1, para cada categoria.

Tabela 3.4 – Total de atributos do conjunto de dados CDS1 para cada categoria.

Atributos do conjunto de dados CDS1 por categoria	
Categoria	Total de Atributos
Alimento	53
Animal	63
Ator	126
Cidade	149
Escritor	124
Esporte	41
Filme	37
Pessoa	121

O conjunto de dados que contém todas as categorias reúne também todas as relações. Após remover as repetições, o conjunto ficou com 277 atributos.

O conjunto de dados semânticos (CDS2), extraído a partir das características semânticas da Base2, possui os seguintes atributos: “En”, “Pt”, “Sujeitos”, “Verbos”, “Objetos” e “Equivalente”. A Tabela 3.5 apresenta uma descrição de cada um dos atributos.

Tabela 3.5 – Detalhamento dos atributos do conjunto de dados semântico – CDS2.

Detalhamento dos atributos do CDS2	
Atributo	Descrição
En	Candidato em inglês.
Pt	Candidato em português.
Sujeitos	Quantidade de sujeitos existentes na base em inglês que a tradução existe na base em português.
Verbos	Quantidade de verbos existentes na base em inglês que a tradução existe na base em português.
Objetos	Quantidade de objetos existentes na base em inglês que a tradução existe na base em português.
Equivalente	Para exemplos positivos, o valor do campo é 1. Para exemplos negativos, o valor do campo é 0.

3.3.3 Conjunto de Dados com Características Morfossintáticas e Semânticas

Os conjuntos de dados morfossintático e semântico, CDMS1 e CDMS2, possuem os atributos do CDM junto com os atributos do conjunto semântico, existentes no CDS1 e CDS2 respectivamente.

3.3.4 Conjunto de Dados com o Resultado dos Experimentos no CDM e no CDS1 – CDM+S1

Foi montado um conjunto de dados que utiliza como entrada a saída dos experimentos, no CDM e no CDS1. A Tabela 3.6 apresenta uma descrição de cada um dos atributos.

Tabela 3.6 – Detalhamento dos atributos do conjunto que utiliza a saída dos experimentos no CDM e no CDS1 – CDM+S1.

Detalhamento dos atributos do CDM+S1	
Atributo	Descrição
En	Candidato em inglês.
Pt	Candidato em português.
Morfossintático	Armazena o resultado fornecido pelo modelo de aprendizado de máquina no conjunto de dados com características morfossintáticas.
Semântico	Armazena o resultado fornecido pelo modelo de aprendizado de máquina no conjunto de dados com características semânticas.
Equivalente	Para exemplos positivos, o valor do campo é 1. Para exemplos negativos, o valor do campo é 0.

3.4 Dicionários

Foram utilizadas as seguintes APIs de dicionários: Wikipedia, Wordnet e Oxford. A partir de uma amostra de dados, foi verificado que a API da Oxford

conseguiu trazer o maior número de traduções. O resultado de cada palavra ou expressão traduzida é armazenado em arquivo no formato json.

3.5 Preenchimento dos Conjuntos de Dados

O preenchimento dos conjuntos de dados foi realizado com informações de algumas categorias da NELL, conforme descrito no item 3.2. Para cada exemplo positivo, em que o candidato em português é equivalente ao candidato em inglês, foi gerado manualmente um exemplo negativo. Os exemplos negativos foram gerados a partir de dados da mesma categoria. A Tabela 3.7 apresenta a quantidade de exemplos positivos e negativos de cada categoria.

Tabela 3.7 – Conjunto de dados particionado pelas oito categorias.

Conjunto de dados particionado por categoria		
Categoria	Quantidade de Exemplos Positivos	Quantidade de Exemplos Negativos
Alimento	600	600
Animal	600	600
Ator	600	600
Cidade	600	600
Escritor	600	600
Esporte	600	600
Filme	600	600
Pessoa	600	600
Total	4800	4800

Foi desenvolvido um algoritmo que, a partir dos exemplos positivos e negativos, realiza o preenchimento de cada atributo dos conjuntos de dados. Para evitar criar blocos com um único tipo de exemplo, é lido uma linha no arquivo de exemplos positivos e uma linha no arquivo de exemplos negativos. No final, para cada tipo (morfossintático, semântico e morfossintático com semântico), há um conjunto por categoria (8 no total) e mais um com todas as categorias juntas.

3.5.1 Preenchimento do Conjunto de Dados com Características Morfossintáticas - CDM

Inicialmente o candidato em inglês, o candidato em português e o valor que informa se a equivalência é verdadeira ou falsa são armazenados, respectivamente, nos atributos “En”, “Pt” e “Equivalente. Na sequência, é contada a quantidade de palavras do candidato em inglês e do candidato em português. A diferença entre eles é guardada no atributo “Dif Qtd Palavras En - Pt”. Se o candidato em português existir no dicionário, como tradução do candidato em inglês, a posição que ele aparece é salva no atributo “Posição no Dicionário”. Caso contrário, o atributo recebe o valor -1. Por último, o valor retornado pela função SequenceMatcher, existente na biblioteca difflib disponível para a linguagem Python e responsável por verificar a similaridade entre duas cadeias de caracteres, é inserido no atributo “Similaridade”. A Tabela 3.8 apresenta uma amostra do conjunto de dados com características morfossintáticas.

Tabela 3.8 – Amostra do conjunto de dados com atributos morfossintáticos – CDM.

Amostra do conjunto de dados CDM					
En	Pt	Dif Qtd Palavras En - Pt	Posição no Dicionário	Similaridade	Equivalente
goose	ganso	0	4	0.4	1
cat	golfinho	0	-1	0.0	0
tortoises	iguanas	0	-1	0.25	0
pumas	pumas	0	1	1.0	1
pests	gato	0	-1	0.43	0
insects	rato	0	-1	0.25	0

3.5.2 Preenchimento do Conjunto de Dados com Características Semânticas – CDS1

Inicialmente o candidato em inglês, o candidato em português e o valor que informa se a equivalência é verdadeira ou falsa são armazenados, respectivamente, nos atributos “En”, “Pt” e “Equivalente.

Depois, todas as relações do candidato em inglês são recuperadas. Na sequência, é verificado se a tradução, da relação em inglês, aparece no domínio do candidato em português. Caso isso aconteça, cada valor associado à relação em inglês é obtido. Por último, é verificado se a tradução do valor, da relação em inglês, aparece entre os valores da relação do candidato em português.

Antes de exemplificar como é feito o processamento do conjunto de dados com atributos variáveis, vale observar o conteúdo da Base1 em português. A Tabela 3.9 apresenta esses valores para o conceito “gatos”.

Tabela 3.9 – Dados da Base1, em português, para o conceito “gatos”.

Base1 – dados para gatos		
Entidade	Relação	Valor
concept:mamifero:gatos	concept:animaldesenvolvedoenca	concept:doenca:hipertireoidismo

Inicialmente o algoritmo recupera todas as relações existentes para o conceito em inglês. Conforme pode ser observado na Tabela 2.1 há quatro relações (atributo Relation) diferentes para “cats”: “animaleatfood”, “animaleatvegetable”, “animalistypeofanimal” e “animaldevelopdisease”.

Na Tabela 3.3 é possível verificar que as traduções para as relações em inglês “animaleatfood”, “animaleatvegetable”, “animalistypeofanimal” e “animaldevelopdisease” podem existir na Base1 em português, respectivamente, como “animalcomealimento”, “animalcomevegetais”, “animaledamesmaracadoanimal” e “animaldesenvolvedoenca”.

De acordo com a Tabela 3.9 é possível observar que a tradução das relações “animaleatfood”, “animaleatvegetable”, “animalistypeofanimal” não existem no atributo “Relação” do conceito em português. Sendo assim, esses atributos são preenchidos, com zero, no conjunto de dados CDS1.

A última relação em inglês, “animaldevelopdisease” possui sua tradução, “animaldesenvolvedoenca”, no atributo “Relação” do conceito em português, conforme pode ser observado na Tabela 3.9. Quando isso ocorre, o atributo recebe uma pontuação mínima que poderá ser incrementada caso aconteçam mais semelhanças com os valores da relação.

Dessa forma, o próximo passo é recuperar os dados do atributo “Value” para os casos que a relação em inglês é “animaldevelopdisease”. Novamente, ao observar os dados da Tabela 2.1 obtem-se: “arthritis” e “hyperthyroidism”.

O dicionário é utilizado para tentar traduzir os valores da relação em inglês. Cada valor traduzido é procurado nos dados do atributo “Valor” para a relação em português. Finalmente, ao observar a Tabela 3.9, é possível notar que nesse caso, apenas “hipertireoidismo”, tradução de “hyperthyroidism”, existe no atributo “Valor” da relação “animaldesenvolvedoenca”. Quando isso acontece o valor do atributo é incrementado em uma unidade. A Tabela 3.10 apresenta uma amostra do conjunto de dados CDS1. Nela é possível observar que cada relação, da categoria animals em inglês, é utilizada como atributo do conjunto de dados.

Tabela 3.10 – Amostra do conjunto de dados com atributos semânticos – CDS1

Amostra do conjunto de dados CDS1					
En	Pt	animalDevelopDesease	...	animalEatFood	Equivalente
cat	gato	5	...	4	1
cat	urso	0	...	1	0
dog	rato	1	...	0	0

3.5.3 Preenchimento do Conjunto de Dados com Características Semânticas – CDS2

Inicialmente o candidato em inglês, o candidato em português e o valor que informa se a equivalência é verdadeira ou falsa são armazenados, respectivamente, nos atributos “En”, “Pt” e “Equivalente. Na sequência, são recuperados todos os valores do atributo “Sujeito”, na Base2 em português, que o candidato existe. O mesmo é feito para o atributo “Subject”, na Base2 em inglês. Depois, é verificado se existe um arquivo json com as traduções, para o português, dos valores de “Subject” recuperados no passo anterior. Se não existir, a API do dicionário é consultada para

recuperar as traduções. Cada tradução é pesquisada nos valores do atributo "Sujeito". Quando a tradução é encontrada, o atributo "Sujeito", do CDS2, tem seu valor incrementado em uma unidade. O mesmo procedimento é realizado para os atributos "Verbos" e "Objetos". A Tabela 3.11 apresenta uma amostra do conjunto de dados semânticos CDS2.

Tabela 3.11 – Amostra do conjunto de dados com atributos semânticos – CDS2

Amostra do conjunto de dados CDS2					
Em	Pt	Sujeitos	Verbos	Objetos	Equivalente
cats	gatos	36	0	0	1
dog	aruanã	0	0	0	0
turtles	tartarugas	299	642	0	1

3.5.4 Preenchimento do Conjunto de Dados com Características Morfossintáticas e Semânticas – CDMS1 e CDMS2

Os conjuntos de dados morfossintático e semântico, CDMS1 e CDMS2, foram preenchidos a partir da simples junção dos dados que foram inseridos no conjunto CDM e dos dados que foram inseridos nos conjuntos CDS1 e CDS2, respectivamente.

3.5.5 Preenchimento do Conjunto de Dados com o Resultado dos Experimentos em CDM e CDS1 – CDM+S1

O atributo "Morfossintático" do CDM+S foi preenchido com os resultados, do melhor cenário, dos experimentos em CDM. O atributo "Semântico" foi preenchido com os resultados, do melhor cenário, dos experimentos em CDS1.

3.6 Modelo de Aprendizado de Máquina

Uma amostra aleatória dos dados do CDM foi utilizada para escolha do modelo de aprendizado de máquina a ser utilizado na pesquisa. Há 339 exemplos

positivos e 261 exemplos negativos. Os modelos testados foram C4.5, Naive Bayes, Rede Neural e XGBoost. A Tabela 3.12 contém o resultado de cada modelo.

Tabela 3.12 – Resultado dos modelos de aprendizado de máquina

Resultado dos testes com diferentes algoritmos de aprendizado de máquina	
Modelo	Precisão (%)
Naive Bayes	89,6
XGBoost	92,0
C4.5	92,3
Rede Neural	92,5

A partir dos dados da Tabela 3.12, é possível verificar que todos os modelos tiveram aproximadamente 90% de acerto. Apesar de não ter sido realizado testes de significância estatística, os resultados não apresentaram uma diferença muito relevante. Todavia, a Rede Neural foi o modelo que obteve o melhor resultado: 92,5%. Com base na revisão da literatura (descrito na seção 2.7) e nos testes com os diferentes algoritmos, a Rede Neural foi escolhida como modelo de aprendizado de máquina.

A rede foi treinada utilizando validação cruzada 5-fold. Isso significa que o conjunto de dados foi dividido em cinco partes e a rede foi executada cinco vezes. A cada iteração, uma parte é removida para ser utilizada como prova ou teste, e as outras quatro são utilizadas como conjunto de treinamento.

Os experimentos foram realizados variando a taxa de aprendizado, o número de camadas ocultas e número de neurônios. Antes de executar a rede neural, os dados foram normalizados, isto é, média zero e desvio padrão um.

O C4.5 e o XGBoost tiveram desempenho muito parecido com a rede neural. Sendo assim, foram realizados experimentos adicionais em que o C4.5 e o XGBoost substituem a rede como modelo de aprendizado de máquina. Nesses experimentos também foi utilizado validação cruzada 5-fold. Taxa de aprendizado e profundidade, foram os parâmetros variados nos experimentos com XGBoost.

Capítulo 4

EXPERIMENTOS E ANÁLISES

4.1 Considerações Iniciais

Foram realizados experimentos com o conjunto de dados particionados por categoria, bem como, com o conjunto unificado, que reúne todas as categorias. A Tabela 4.1 apresenta os experimentos realizados.

Tabela 4.1 – Experimentos realizados

Experimentos realizados	
Categoria	
Unificada	Particionada
Morfossintática Unificada	Morfossintática Particionada
Semântica Unificada	Semântica Particionada
Morfossintática + Semântica Unificada	Morfossintática + Semântica Particionada

A Rede Neural foi executada nos conjuntos variando o número de camadas ocultas, o número de neurônios em cada camada e a taxa de aprendizado. Foram realizados experimentos com até 5 camadas ocultas. O número de neurônios, inicia com o mesmo valor do número de atributos do conjunto de dados. A cada nova camada, o número de neurônios é diminuído pela metade. Em todos os experimentos, os melhores resultados foram encontrados com uma única camada. Por último, a taxa de aprendizagem foi variada de 0.01 até 1.00.

Para comparar os resultados obtidos pela rede neural, foram criados três casos de utilização do dicionário inglês-português. No primeiro, ao encontrar o

candidato em inglês no dicionário, a primeira tradução disponível é comparada. No segundo, a tradução é escolhida aleatoriamente. Por último, para dar mais chances para o dicionário, é verificado se o candidato, em português, está na lista de traduções do candidato, em inglês.

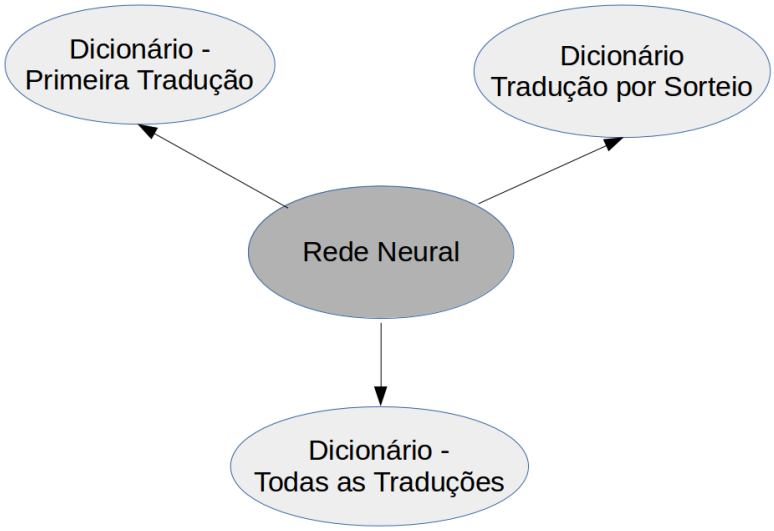

Figura 4.1 Dados da rede neural comparado com os três casos de utilização do dicionário.

As análises dos resultados terão como base as matrizes de confusão (MC). Assim, temos interesse em capturar os falsos negativos, os verdadeiro negativos, os falsos positivos e os verdadeiros positivos. Outras técnicas que exploram diferentes combinações dos dados apresentados na MC (F-1, Curva ROC, Área abaixo da curva, sensibilidade, especificidade, etc.) não serão especificamente abordadas. Recomenda-se que o leitor, interessado em métricas mais específicas, utilize os dados aqui reportados nas MC para realizar os cálculos da métrica desejada.

4.2 Experimento com os Dicionários

Foram realizados experimentos para obter um parâmetro inicial de quantas equivalências os dicionários são capazes de encontrar. Para isso, foram utilizados apenas os atributos “En”, “Pt” e “Equivalência” do conjunto de dados CDM. Foi realizado uma consulta do candidato em inglês, no dicionário, e a tradução foi

escolhida, conforme o tipo de utilização. A Tabela 4.2 apresenta a matriz de confusão para: primeira tradução, tradução aleatória e todas as traduções.

Tabela 4.2 – Matriz de confusão dos dicionários

Matriz de confusão de utilização dos dicionários											
Primeira Tradução				Tradução Aleatória				Todas as Traduções			
X		Previsto		X		Previsto		X		Previsto	
		0	1			0	1			0	1
Real	0	4800	0	Real	0	4800	0	Real	0	4800	0
	1	4608	192		1	4654	146		1	4579	221

É válido mencionar que, para cada experimento, ao somar os valores da linha 0, obtém-se os 4800 exemplos negativos do conjunto de dados. Ao somar os valores da linha 1, obtém-se os 4800 exemplos positivos do conjunto de dados.

Ao observar a tabela, é possível verificar que o dicionário tem excelente precisão, para os exemplos negativos. Apesar disso, o desempenho é ruim, 221 acertos no melhor cenário, para os exemplos positivos.

Isso implica que a simples utilização de dicionário, no máximo, conseguirá transferir aproximadamente 4.6% (221/4800) do conhecimento da base em inglês para a base em português.

4.3 Experimentos no CDM – Unificado

A Tabela 4.3 apresenta a matriz de confusão da rede neural, para o melhor cenário, no CDM unificado: 1 camada oculta, 3 neurônios e taxa de aprendizagem de 0.19.

Tabela 4.3 – Matriz de confusão da rede neural no CDM unificado

Matriz de confusão da rede neural no CDM unificado – Melhor cenário						
X		Previsto		Melhor Cenário		
		0	1	Camada(s)	Neurônio(s)	Aprendizagem
Real	0	1920	2880	1	3	0.19
	1	1878	2922			

A rede neural acertou 40% (1920/4800) dos exemplos negativos. Resultado bem inferior ao dicionário. Contudo, a taxa de acertos dos exemplos positivos foi de 61% (2922/4800). Isso significa que a RN foi 13 vezes superior ao dicionário.

4.4 Experimentos no CDM – Particionado

Os mesmos experimentos que foram realizados no CDM unificado, foram executados no conjunto de dados particionado pelas categorias. A Tabela 4.4 apresenta a matriz de confusão da rede neural, para o melhor cenário de cada categoria.

Tabela 4.4 – Matriz de confusão da rede neural no CDM particionado

Matriz de confusão da rede neural no CDM particionado – Melhor Cenário							
Categoria	X		Previsto		Melhor Cenário		
			0	1	Camada(s)	Neurônio(s)	Aprendizagem
Alimento	Real	0	475	125	1	1	0.01
		1	256	344			
Animal	Real	0	244	356	1	1	0.01
		1	95	505			
Ator	Real	0	600	0	1	1	0.08
		1	0	600			
Cidade	Real	0	596	4	1	1	0.06
		1	6	594			
Escritor	Real	0	600	0	1	1	0.07
		1	0	600			
Esporte	Real	0	548	52	1	1	0.01
		1	147	453			
Filme	Real	0	539	61	1	1	0.02
		1	310	290			
Pessoa	Real	0	600	0	1	1	0.12
		1	0	600			
Total	Real	0	4202	598			
		1	814	3986			

No total, a rede neural acertou 88% (4202/4800) dos exemplos negativos. Resultado ainda inferior ao dicionário. Não obstante, notamos que houve ganho nos experimentos do CDM, separado por categoria, em relação aos experimentos com todas as categorias juntas. A taxa de acertos, para os exemplos negativos, aumentou mais de 2 vezes.

A rede neural acertou 83% (3986/4800) dos exemplos positivos. Resultado continua muito superior ao dicionário. Além do mais, notamos que houve ganho nos experimentos do CDM, separado por categoria, em relação aos experimentos com todas as categorias juntas. A taxa de acertos, para os exemplos positivos, aumentou 22 pontos percentuais (83-61).

Ao observar os resultados por categoria, verificamos que o modelo de AM teve facilidade para encontrar as equivalências para nomes de pessoas, atores, cidades e escritores. Porém, houve dificuldade para encontrar as equivalências nos nomes de filmes, alimentos, esportes e animais.

4.5 Experimentos no CDS1 – Unificado

A Tabela 4.5 apresenta a matriz de confusão da rede neural, para o melhor cenário, no CDS1 unificado: 1 camada oculta, 1 neurônio e taxa de aprendizagem de 1.00.

Tabela 4.5 – Matriz de confusão da rede neural no CDS1 unificado

Matriz de confusão da rede neural no CDS1 unificado – Melhor cenário						
X		Previsto		Melhor Cenário		
		0	1	Camada(s)	Neurônio(s)	Aprendizagem
Real	0	4800	0	1	1	1.00
	1	2880	1920			

Ao analisar os 4800 acertos de exemplos negativos, é possível dizer que, para as situações que não era a tradução, a rede neural teve a mesma precisão que o dicionário. Entretanto, a rede acertou 40% (1920/4800) dos exemplos positivos.

Esse resultado continua superior ao dicionário. Todavia, é inferior aos 83% obtidos com a RN no CDM particionado.

4.6 Experimentos no CDS1 – Particionado

Os mesmos experimentos que foram realizados no CDS1 unificado, foram executados no conjunto de dados particionado pelas categorias. A Tabela 4.6 apresenta a matriz de confusão da rede neural, para o melhor cenário de cada categoria.

Tabela 4.6 – Matriz de confusão da rede neural no CDS1 particionado

Matriz de confusão da rede neural no CDS1 particionado – Melhor Cenário							
Categoria	X		Previsto		Melhor Cenário		
			0	1	Camada(s)	Neurônio(s)	Aprendizagem
Alimento	Real	0	360	240	1	1	1.00
		1	360	240			
Animal	Real	0	600	0	1	2	0.13
		1	598	2			
Ator	Real	0	598	2	1	2	0.01
		1	564	36			
Cidade	Real	0	120	480	1	1	1.00
		1	120	480			
Escritor	Real	0	480	120	1	1	0.14
		1	477	123			
Esporte	Real	0	594	6	1	2	0.01
		1	591	9			
Filme	Real	0	480	120	1	1	1.00
		1	280	320			
Pessoa	Real	0	599	1	1	1	0.13
		1	597	3			
Total	Real	0	3831	969			
		1	3587	1213			

No total, a rede neural acertou 80% (3831/4800) dos exemplos negativos. Resultado inferior ao dicionário e aos experimentos com todas as categorias juntas. Ademais, a precisão para os exemplos positivos foi de 25% (1213/4800). Esse resultado é superior ao obtido com os dicionários, mas inferior a todos os experimentos realizados.

Ao observar os resultados por categoria, verificamos que o modelo de AM teve facilidade para encontrar as equivalências para nomes de cidades, filmes, alimentos e escritores. Ainda assim, quase não conseguiu encontrar as equivalências para os nomes de atores, esportes, pessoas e animais.

4.7 Experimentos no CDMS1 – Unificado

A Tabela 4.7 apresenta a matriz de confusão da rede neural, para o melhor cenário, no CDMS1 unificado: 1 camada oculta, 2 neurônios e taxa de aprendizagem de 0.05.

Tabela 4.7 – Matriz de confusão da rede neural no CDMS1 unificado

Matriz de confusão da rede neural no CDMS1 unificado – Melhor cenário						
X		Previsto		Melhor Cenário		
		0	1	Camada(s)	Neurônio(s)	Aprendizagem
Real	0	1151	3649	1	2	0.05
	1	972	3828			

A rede neural acertou 24% (1151/4800) dos exemplos negativos. Resultado inferior ao dicionário. Não obstante, a precisão para os exemplos positivos foi de 80% (3828/4800). Esse resultado continua melhor que o dicionário. Mas, é inferior aos 83% obtidos com a RN no CDM particionado. Observamos que, utilizar as características morfossintáticas com semânticas, ainda não trouxe melhorias.

4.8 Experimentos no CDMS1 – Particionado

Os mesmos experimentos que foram realizados no CDMS1 unificado, foram executados no conjunto particionado pelas categorias. A Tabela 4.8 apresenta a matriz de confusão da rede neural, para o melhor cenário de cada categoria.

Tabela 4.8 – Matriz de confusão da rede neural no CDMS1 particionado

Matriz de confusão da rede neural no CDMS1 particionado – Melhor Cenário							
Categoria	X		Previsto		Melhor Cenário		
			0	1	Camada(s)	Neurônio(s)	Aprendizagem
Alimento	Real	0	473	127	1	1	0.01
		1	253	347			
Animal	Real	0	465	135	1	4	0.01
		1	122	478			
Ator	Real	0	600	0	1	3	0.01
		1	0	600			
Cidade	Real	0	594	6	1	2	0.01
		1	7	593			
Escritor	Real	0	600	0	1	1	0.07
		1	0	600			
Esporte	Real	0	548	52	1	1	0.01
		1	131	469			
Filme	Real	0	539	61	1	1	0.02
		1	310	290			
Pessoa	Real	0	600	0	1	1	0.1
		1	0	600			
Total	Real	0	4419	381			
		1	823	3977			

No total, a rede neural acertou 92% (4419/4800) dos exemplos negativos. Resultado inferior ao dicionário. Apesar disso, ao comparar o resultado obtido com a precisão de acertos dos exemplos negativos do CDM particionado, notamos que a introdução das características semânticas do CDMS1, trouxe ganho de 5 pontos percentuais. Notamos também que houve ganho nos experimentos no CDMS1,

separado por categoria, em relação aos experimentos com todas as categorias juntas. A taxa de acertos, para os exemplos negativos, aumentou quase 4 vezes.

O modelo de AM acertou 83% (3977/4800) dos exemplos positivos. Resultado melhor que o dicionário e praticamente igual aos experimentos no CDM, separado por categoria.

Ao observar os resultados por categoria, verificamos que o modelo de AM teve facilidade para encontrar as equivalências para nomes de pessoas, atores, cidades e escritores. Todavia, houve dificuldade para encontrar as equivalências nos nomes de filmes, alimentos, esportes e animais.

4.9 Experimentos no CDM+S1 – Unificado

A Tabela 4.9 apresenta a matriz de confusão da rede neural, para o melhor cenário, no CDM+S1 unificado: 1 camada oculta, 2 neurônios e taxa de aprendizagem de 0.05.

Tabela 4.9 – Matriz de confusão da rede neural no CDM+S1 unificado

Matriz de confusão da rede neural no CDM+S1 unificado – Melhor cenário						
X		Previsto		Melhor Cenário		
		0	1	Camada(s)	Neurônio(s)	Aprendizagem
Real	0	1920	2880	1	2	0.05
	1	1920	2880			

A rede neural acertou 40% (1920/4800) dos exemplos negativos, resultado inferior ao dicionário. Além disso, a precisão para os exemplos positivos foi de 60% (2880/4800). Esse resultado continua melhor do que o dicionário. Porém, é inferior aos 83% obtidos com a RN no CDMS1 particionado.

4.10 Experimentos no CDM+S1 – Particionado

Os mesmos experimentos que foram realizados no CDM+S1 unificado, foram executados no conjunto particionado pelas categorias. A Tabela 4.10 apresenta a matriz de confusão da rede neural, para o melhor cenário de cada categoria.

Tabela 4.10 – Matriz de confusão da rede neural no CDM+S1 particionado

Matriz de confusão da rede neural no CDM+S1 particionado – Melhor Cenário							
Categoria	X		Previsto		Melhor Cenário		
			0	1	Camada(s)	Neurônio(s)	Aprendizagem
Alimento	Real	0	401	199	1	1	0.16
		1	213	387			
Animal	Real	0	225	375	1	1	0.02
		1	141	459			
Ator	Real	0	600	0	1	1	0.15
		1	0	600			
Cidade	Real	0	596	4	1	1	0.15
		1	6	594			
Escritor	Real	0	600	0	1	1	0.09
		1	0	600			
Esporte	Real	0	548	52	1	1	0.05
		1	147	453			
Filme	Real	0	548	52	1	1	0.17
		1	370	230			
Pessoa	Real	0	580	20	1	1	0.1
		1	0	600			
Total	Real	0	4098	702			
		1	877	3923			

No total, a rede neural acertou 85% (4098/4800) dos exemplos negativos. Resultado inferior ao dicionário. Ademais, a precisão para os exemplos positivos foi de 82% (3923/4800). Esse resultado continua melhor do que o dicionário. Ainda assim, é inferior aos 83% obtidos com a RN no CDMS1 particionado.

4.11 Experimentos no CDS2 – Unificado

A Tabela 4.11 apresenta a matriz de confusão da rede neural, para o melhor cenário, no CDS2 unificado: 1 camada oculta, 1 neurônio e taxa de aprendizagem de 0.05.

Tabela 4.11 – Matriz de confusão da rede neural no CDS2 unificado

Matriz de confusão da rede neural no CDS2 unificado – Melhor cenário						
X		Previsto		Melhor Cenário		
		0	1	Camada(s)	Neurônio(s)	Aprendizagem
Real	0	1785	3015	1	1	0.05
	1	1716	3084			

A rede neural acertou 37% (1785/4800) dos exemplos negativos. Resultado inferior aos experimentos com o dicionário e no CDS1 unificado.

A precisão para os exemplos positivos foi de 64% (3084/4800). Esse resultado continua melhor do que o dicionário. Além do que, ao comparar a precisão dos exemplos positivos do CDS2 unificado com os experimentos no CDS1 unificado, observamos que extrair as características semânticas da base da NELL com mais dados, trouxe um ganho de 24 pontos percentuais (64-40). Todavia, esse resultado não foi capaz de superar os 83% de acertos da rede no CDMS1 particionado.

4.12 Experimentos no CDS2 – Particionado

Os mesmos experimentos que foram realizados no CDS2 unificado, foram executados no conjunto de dados particionado pelas categorias. A Tabela 4.12 apresenta a matriz de confusão da rede neural, para o melhor cenário de cada categoria.

Tabela 4.12 – Matriz de confusão da rede neural no CDS2 particionado

Matriz de confusão da rede neural no CDS2 particionado – Melhor Cenário							
Categoria	X		Previsto		Melhor Cenário		
			0	1	Camada(s)	Neurônio(s)	Aprendizagem
Alimento	Real	0	577	23	1	2	0.01
		1	533	67			
Animal	Real	0	346	254	1	2	0.29
		1	292	308			
Ator	Real	0	526	74	1	1	0.01
		1	268	332			
Cidade	Real	0	407	193	1	2	0.01
		1	306	294			
Escritor	Real	0	553	47	1	1	0.01
		1	471	129			
Esporte	Real	0	465	135	1	3	0.01
		1	431	169			
Filme	Real	0	463	137	1	3	0.21
		1	461	139			
Pessoa	Real	0	595	5	1	1	0.01
		1	558	42			
Total	Real	0	3932	868			
		1	3320	1480			

A rede neural acertou 82% (3932/4800) dos exemplos negativos. Resultado inferior ao dicionário. Contudo, ao comparar o resultado com os experimentos no CDS1 particionado, observamos que extrair as características semânticas de uma base da NELL com mais dados, trouxe um pequeno ganho de 2 pontos percentuais (82-80).

O modelo de AM acertou 31% (1480/4800) dos exemplos positivos. Resultado superior ao dicionário. Além do que, ao comparar o resultado com os experimentos no CDS1 particionado, observamos que extrair as características semânticas da base da NELL com mais informações, trouxe um ganho de 6 pontos percentuais (31-25. Porém, esse resultado não foi suficiente para superar o encontrado nos experimentos do CDMS1, particionado por categoria.

Ao observar os resultados por categoria, verificamos que o modelo de AM teve facilidade para encontrar as equivalências para nomes de animais, atores, cidades e esportes. Contudo, o desempenho foi ruim para encontrar as equivalências nos nomes de alimentos, escritores, pessoas e filmes.

4.13 Experimentos no CDMS2 – Unificado

A Tabela 4.13 apresenta a matriz de confusão da rede neural, para o melhor cenário, no CDMS2 unificado: 1 camada oculta, 3 neurônios e taxa de aprendizagem de 0.96.

Tabela 4.13 – Matriz de confusão da rede neural no CDMS2 unificado

Matriz de confusão da rede neural no CDMS2 unificado – Melhor cenário						
X		Previsto		Melhor Cenário		
		0	1	Camada(s)	Neurônio(s)	Aprendizagem
Real	0	2735	2065	1	3	0.96
	1	2612	2188			

A rede neural acertou 57% (2735/4800) dos exemplos negativos. Desempenho inferior ao dicionário. Além disso, a precisão para os exemplos positivos foi de 46% (2188/4800). Embora, esse resultado é melhor do que o dicionário, é inferior aos 83% obtidos com a RN no CDMS1 particionado.

4.14 Experimentos no CDMS2 – Particionado

Os mesmos experimentos que foram realizados no CDMS2 unificado, foram executados no conjunto de dados particionado pelas categorias. A Tabela 4.14 apresenta a matriz de confusão da rede neural, para o melhor cenário de cada categoria.

Tabela 4.14 – Matriz de confusão da rede neural no CDMS2 particionado

Matriz de confusão da rede neural no CDMS2 particionado – Melhor Cenário							
Categoria	X		Previsto		Melhor Cenário		
			0	1	Camada(s)	Neurônio(s)	Aprendizagem
Alimento	Real	0	413	187	1	6	0.2
		1	192	408			
Animal	Real	0	454	146	1	4	0.45
		1	314	286			
Ator	Real	0	600	0	1	1	0.01
		1	0	600			
Cidade	Real	0	594	6	1	4	0.01
		1	7	593			
Escritor	Real	0	600	0	1	2	0.01
		1	0	600			
Esporte	Real	0	373	227	1	4	0.02
		1	270	330			
Filme	Real	0	534	66	1	1	0.01
		1	315	285			
Pessoa	Real	0	600	0	1	1	0.12
		1	0	600			
Total	Real	0	4168	632			
		1	1098	3702			

No total, a rede neural acertou 87% (4168/4800) dos exemplos negativos. Resultado inferior ao dicionário. Mas, notamos que houve ganho nos experimentos no CDMS2, separado por categoria, em relação aos experimentos com todas as categorias juntas. Ainda assim, esse resultado não foi suficiente para superar o obtido nos experimentos no CDMS1 particionado.

O modelo de AM acertou 77% (3702/4800) dos exemplos positivos. Embora, esse resultado é melhor do que o dicionário, é inferior aos 83% obtidos com a RN no CDMS1 particionado.

Ao observar os resultados por categoria, verificamos que o modelo de AM teve facilidade para encontrar as equivalências para nomes de pessoas, atores,

cidades e escritores. Não obstante, houve dificuldade para encontrar as equivalências nos nomes de filmes, alimentos, esportes e animais.

4.15 Experimentos para Comparação com PageRank Personalizado

Experimentos adicionais foram feitos preenchendo o CDM, CDS1, CDMS1, CDS2 e CDMS2 com os dados que os autores utilizaram no trabalho de (GONZÁLEZ J.; HRUSCHKA E. R.; MITCHELL T. M., 2017).

As categorias utilizadas foram: animal, ator, cidade, escritor, esporte, filme, país e pessoa. Nesse conjunto de dados são preenchidos 2 exemplos negativos para cada exemplo positivo (visto que esta foi a estratégia usada em GONZÁLEZ J.; HRUSCHKA E. R.; MITCHELL T. M., 2017). Além do mais, os autores utilizam um número diferente instâncias para cada categoria. A Tabela 4.15 apresenta a quantidade de exemplos positivos e negativos por categoria.

Tabela 4.15 – Conjunto de dados do PageRank personalizado particionado pelas oito categorias.

Conjunto de dados do PageRank personalizado particionado por categoria		
Categoria	Quantidade de Exemplos Positivos	Quantidade de Exemplos Negativos
Animal	60	120
Ator	510	1020
Cidade	510	1020
Escritor	60	120
Esporte	170	340
Filme	40	80
País	110	220
Pessoa	1560	3120
Totais	3020	6040

Os resultados obtidos com a utilização do PageRank personalizado, para cada categoria, estão dispostos na Tabela 4.16.

Tabela 4.16 – Precisão do PageRank personalizado por categoria.

Precisão do PageRank personalizado por categoria	
Categoria	Precisão (%)
Animal	22
Ator	70
Cidade	40
Escritor	80
Esporte	45
Filme	50
País	38
Pessoa	48

A Rede Neural foi executada em todos os conjuntos de dados, variando o número de camadas ocultas, o número de neurônios em cada camada e a taxa de aprendizado. O número de neurônios, inicia com o mesmo valor do número de atributos do conjunto de dados. A cada nova camada, o número de neurônios é diminuído pela metade.

O melhor resultado foi obtido no CDM separado por categoria. A Tabela 4.17 apresenta a matriz de confusão da rede neural, para o melhor cenário de cada categoria, que ocorreu com 1 camada oculta, 1 neurônio e taxa de aprendizagem de 0.01.

Tabela 4.17 – Matriz de confusão da rede neural no CDM do PageRank personalizado por categoria

Matriz de confusão da RN no CDM do PageRank personalizado por categoria					
Categoria	X		Previsto		Precisão (%)
			0	1	
Animal	Real	0	114	6	89
		1	14	46	
Ator	Real	0	1020	0	100
		1	0	510	
Cidade	Real	0	1019	1	100
		1	0	510	
Escritor	Real	0	118	2	99
		1	0	60	
Esporte	Real	0	327	13	87
		1	52	118	
Filme	Real	0	79	1	99
		1	0	40	
País	Real	0	211	9	93
		1	15	95	
Pessoa	Real	0	3115	5	100
		1	0	1560	

Ao comparar os resultados da Tabela 4.17 com a Tabela 4.16 é possível notar que, em todas as categorias, a rede neural teve desempenho superior ao PageRank personalizado.

4.16 Experimentos com C4.5 no CDMS2 – Unificado

Foram realizados experimentos que trocaram o uso da rede neural, como modelo de aprendizado de máquina pelo C4.5 (segundo melhor resultado nos experimentos para escolha do modelo de AM). Como o objetivo era apenas verificar se a RN poderia ser substituída pelo modelo, o CDMS2 foi escolhido para realizar os

experimentos. CDMS2 permite explorar tanto as características morfossintáticas como as semânticas e facilita a construção dos experimentos por possuir um número fixo de colunas. A Tabela 4.18 apresenta a matriz de confusão do C4.5.

Tabela 4.18 – Matriz de confusão do C4.5 no CDMS2 unificado

Matriz de confusão do C4.5 no CDMS2			
X		Previsto	
		0	1
Real	0	4747	53
	1	1716	3084

O C4.5 acertou 99% (4747/4800) dos exemplos negativos. Desempenho quase igual ao dicionário. A precisão para os exemplos positivos foi de 64% (3084/4800). Esse resultado continua melhor do que o dicionário. No entanto, é inferior aos 83% obtidos com a RN no CDMS1 particionado.

4.17 Experimentos com C4.5 no CDMS2 – Particionado

Os mesmos experimentos que foram realizados no CDMS2 unificado, foram executados no conjunto de dados particionado pelas categorias. A Tabela 4.19 apresenta a matriz de confusão do C4.5.

Tabela 4.19 – Matriz de confusão do C4.5 no CDMS2 particionado

Matriz de confusão do C4.5 no CDMS2 particionado				
Categoria	X		Previsto	
			0	1
Alimento	Real	0	464	136
		1	284	316
Animal	Real	0	494	106
		1	179	421
Ator	Real	0	599	1
		1	0	600
Cidade	Real	0	596	4
		1	6	594
Escritor	Real	0	599	1
		1	0	600
Esporte	Real	0	571	29
		1	182	418
Filme	Real	0	577	23
		1	365	235
Pessoa	Real	0	599	1
		1	0	600
Total	Real	0	4499	301
		1	1016	3784

No total, o C4.5 acertou 94% (4499/4800) dos exemplos negativos. Desempenho inferior ao dicionário. Apesar disso, esse resultado foi igual ao obtido nos experimentos no CDMS1 particionado.

O novo modelo de AM acertou 79% (3784/4800) dos exemplos positivos. Dessa forma, notamos que houve ganho nos experimentos no CDMS2, separado por categoria, em relação aos experimentos com todas as categorias juntas.

Ao observar os resultados por categoria, verificamos que o novo modelo de AM teve facilidade para encontrar as equivalências para nomes de pessoas, atores, cidades e escritores. Não obstante, houve dificuldade para encontrar as equivalências nos nomes de filmes, alimentos, esportes e animais.

4.18 Experimentos com XGBoost no CDMS2 – Unificado

Foram realizados experimentos que trocaram o uso da rede neural, como modelo de aprendizado de máquina, pelo XGBoost (terceiro melhor resultado nos experimentos para escolha do modelo de AM). Como o objetivo era apenas verificar se a RN poderia ser substituída pelo XGBoost, o CDMS2 foi escolhido para realizar os experimentos. CDMS2 permite explorar tanto as características morfossintáticas como as semânticas e facilita a construção dos experimentos por possuir um número fixo de colunas. Os parâmetros do XGBoost foram alterados variando o número máximo de profundidade (de 3 até 10) e a taxa de aprendizado (de 0.01 até 0.20). A Tabela 4.20 apresenta a matriz de confusão do XGBoost, para o melhor cenário: numero máximo de profundidade 3 e taxa de aprendizagem de 0.01.

Tabela 4.20 – Matriz de confusão do XGBoost no CDMS2 unificado

Matriz de confusão do XGBoost no CDMS2 unificado – Melhor cenário					
X		Previsto		Melhor Cenário	
		0	1	Profundidade	Aprendizagem
Real	0	4416	384	3	0.01
	1	1268	3532		

O XGBoost acertou 92% (4416/4800) dos exemplos negativos. Desempenho inferior ao dicionário. Ademais, a precisão para os exemplos positivos foi de 74% (3532/4800). Esse resultado continua melhor do que o dicionário. Apesar disso, é inferior aos 83% obtidos com a RN no CDMS1 particionado.

4.19 Experimentos com XGBoost no CDMS2 – Particionado

Os mesmos experimentos que foram realizados no CDMS2 unificado, foram executados no conjunto de dados particionado pelas categorias. A Tabela 4.21 apresenta a matriz de confusão do XGBoost, para o melhor cenário de cada categoria.

Tabela 4.21 – Matriz de confusão do XGBoost no CDMS2 particionado

Matriz de confusão do XGBoost no CDMS2 particionado – Melhor Cenário						
Categoria	X		Previsto		Melhor Cenário	
			0	1	Profundidade	Aprendizagem
Alimento	Real	0	410	90	4	0.01
		1	93	407		
Animal	Real	0	448	52	3	0.01
		1	41	459		
Ator	Real	0	600	0	3	0.2
		1	0	600		
Cidade	Real	0	593	7	3	0.01
		1	4	596		
Escritor	Real	0	600	0	3	0.2
		1	0	600		
Esporte	Real	0	525	75	3	0.01
		1	120	480		
Filme	Real	0	453	47	3	0.01
		1	260	340		
Pessoa	Real	0	600	0	3	0.2
		1	0	600		
Total	Real	0	4229	571		
		1	718	4082		

No total, o XGBoost acertou 88% (4229/4800) dos exemplos negativos. Desempenho inferior ao dicionário. Além disso, esse resultado foi 4 pontos percentuais (88-92) inferior ao obtido nos experimentos no CDMS1 particionado.

O novo modelo de AM acertou 85% (4082/4800) dos exemplos positivos. Dessa forma, notamos que houve ganho nos experimentos no CDMS2, separado por categoria, em relação aos experimentos com todas as categorias juntas. Ademais, esse resultado supera, em 2 pontos percentuais (85-83), o obtido com a rede neural nos experimentos no CDMS1 particionado.

Ao observar os resultados por categoria, verificamos que o novo modelo de AM teve facilidade para encontrar as equivalências para nomes de pessoas, atores,

cidades e escritores. Não obstante, houve dificuldade para encontrar as equivalências nos nomes de filmes, alimentos, esportes e animais.

Capítulo 5

CONCLUSÃO

5.1 Objetivos Alcançados

Resultados animadores foram obtidos com a ideia de utilizar um modelo de aprendizado de máquina, para ajudar no processo de descoberta de conceitos equivalentes. Os resultados gerados por meio do experimento no conjunto de dados morfossintático foram fundamentais para verificação da viabilidade da pesquisa. Somente após obter bons resultados com ele, é que se cogitou a expansão na quantidade de atributos.

Inicialmente, foi pensado na utilização apenas dos atributos semânticos, com a expectativa de que a inserção dessas informações, específicas de cada domínio, fosse capaz de auxiliar o modelo de aprendizado de máquina. Mas, os experimentos mostraram que, isolados, eles não foram capazes de melhorar a precisão do modelo de AM. Sendo assim, pensou-se no uso de atributos semânticos junto os morfossintáticos. De fato, os experimentos no CDMS1 particionado trouxeram evidência empírica de um aumento, de 5 pontos percentuais no acerto dos exemplos negativos, quando as informações semânticas foram utilizadas.

Resultados interessantes foram obtidos com os experimentos nos conjuntos de dados, construídos a partir de informações que ainda não são consideradas fato, mas que apresentam um nível elevado de confiança de que estão corretos. Os experimentos no CDS2 unificado mostraram um aumento de 24 pontos percentuais na quantidade de acertos dos exemplos positivos, em comparação com os mesmos experimentos executados no CDS1. Ademais, os experimentos no CDS2

particionado apresentaram um aumento, de 6 pontos percentuais, no número de acerto dos exemplos positivos, em comparação com os mesmos experimentos realizados no CDS1.

Na comparação do modelo de aprendizado de máquina com o PageRank personalizado, foi possível observar que o primeiro teve um desempenho muito superior ao segundo.

A utilização do C4.5 e do XGBoost confirmam a ideia de que vários modelos podem ser utilizados para descoberta das equivalências de conceitos. O XGBoost, no CDMS2, trouxe desempenho de 4 pontos percentuais inferior a rede neural para exemplos negativos. Contudo, ele obteve precisão de 2 pontos percentuais superior, para exemplos positivos.

Com base nos experimentos realizados, é possível dizer que foram obtidas evidências empíricas de que o modelo de aprendizado de máquina pode ser utilizado se o objetivo for: ter a menor quantidade exemplos negativos inseridos por engano, a maior quantidade de inserção de exemplos positivos ou a maior precisão possível de exemplos positivos e negativos. Experimentos que executaram a rede neural no CDS1 unificado, apresentaram o 0% de erro nos exemplos negativos e acertaram 40% dos exemplos positivos. Experimentos que fizeram uso do XGBoost no CDMS2 particionado, trouxeram 12% de erro nos exemplos negativos e 85% de acerto nos exemplos positivos. Finalmente, os experimentos com a rede neural no CDMS1 particionado, exibiram 8% de erro para exemplos negativos e 83% de acerto nos exemplos positivos.

Conforme mencionado, a Base1 da NELL, em inglês, tem 2,3 milhões de fatos. A base equivalente, em português, tem 850 mil. A Base2, em inglês, tem 185 milhões de instâncias, que se tem elevada confiança que estão corretos, mas ainda não são fatos. A base equivalente, em português, tem 50 milhões.

Supondo que todo conhecimento da Base1, em português, exista na base em inglês, restam 1,45 milhões de instâncias que estão apenas na base em inglês. O melhor cenário, utilizando apenas os dicionários seria capaz de trazer apenas 66 mil (4,6% dos 1,45 milhões). Amparado nos experimentos, a rede neural poderia ser utilizada para transferir esses dados para a base em português. No cenário conservador, 580 mil (40% dos 1,45 milhões) novos fatos poderiam ser adicionados na base, sem trazer muitos falsos negativos. Nessa base, a escolha do modelo

conservador é mais indicada, pois, se essas informações forem promovidas, elas serão utilizadas na descoberta de novos fatos.

Por outro lado, supondo que todo conhecimento da Base2, em português, exista na base em inglês, restam 135 milhões de instâncias que existem apenas na base em inglês. O melhor cenário, utilizando apenas os dicionários seria capaz de trazer apenas 6 milhões (4,6% dos 135 milhões). Apoiado nos experimentos, o XGBoost poderia ser utilizado para transferir esses dados para a base em português. No cenário agressivo, 114,75 milhões de instâncias (85% dos 135 milhões) seriam adicionados na base, junto com 16,2 milhões (12% dos 135 milhões) de falsos negativos. Nessa base, a escolha do cenário agressivo é mais indicada. Uma vez que, os falsos negativos poderiam ser descartados pelos outros componentes da NELL, antes de os promover para fatos.

Vale lembrar, conforme mencionado na seção 2.6, que os resultados não são imediatamente utilizados pela NELL, eles serão propostos como candidatos a fato. Assim, mesmo que tiver ruídos nos resultados, o integrador de conhecimento poderá recusar a promoção desses candidatos.

5.2 Limitações

Podemos citar alguns fatores que estão diretamente relacionados com os resultados dos experimentos: quantidade de dados na base, o dicionário utilizado e o modelo de aprendizado de máquina.

Ao comparar a quantidade de linhas apresentadas na Tabela 2.1, que é uma amostra representativa da base de NELL em inglês, com a quantidade de linhas exibidas na Tabela 3.9, que é uma amostra representativa da base de NELL em português, é possível notar que a base em inglês contém muito mais informações. Isso ocorre porque a base em inglês existe há muito mais tempo do que a base em português. Além do que, o número de páginas, escritas em inglês, é maior do que o de páginas em português. Dessa forma, há menos informações na base em português. Isso implica que muitos candidatos em inglês, que possuem tradução para o português nos dicionários, podem não existir na base em português e invalidar as comparações.

Outro fator que influencia o resultado obtido nos conjuntos de dados com características semânticas, está relacionado com a dependência do dicionário. Se a tradução do candidato em inglês existir na base em português, mas não existir no dicionário, a comparação é descartada.

Por último, foram escolhidos apenas alguns modelos de aprendizado de máquina para testes. Testes preliminares, com os diferentes modelos indicaram uma precisão de aproximadamente 90%, que foi comprovada com a execução da rede neural no CDMS1 particionado (87,5%). Mas, não é possível esgotar todos os modelos existentes. Talvez, a escolha de um modelo, que não estava na lista dos avaliados, poderia ser mais adequada para essa situação.

5.3 Trabalhos Futuros

Os seguintes trabalhos futuros podem ser realizados:

- Utilizar versões novas das bases de conhecimento. A NELL roda 24 horas por dia, sete dias por semana. É possível que, a cada dia, novos fatos sejam adicionados. Sendo assim, extrair conjuntos de dados das bases mais novas e executar novamente os experimentos, pode encontrar equivalências que não existiam nas antigas;

- Criação de um novo conjunto de dados que utiliza características semânticas geradas com aprendizado profundo para embedding. Nessa abordagem, cada candidato possui um vetor de características. Vetores similares, ajudarão a encontrar equivalências entre dois candidatos, em idiomas diferentes;

- Usar um modelo de aprendizado de máquina, como a rede neural ou XGBoost, para preencher a base em português com fatos que foram aprendidos na base em inglês;

- Adicionar um modelo de aprendizado de máquina, como a rede neural ou o XGBoost, na lista de componentes que julgam se um dado, aprendido pela NELL, deve ser promovido a fato.

REFERÊNCIAS

BISHOP C. M. Neural Networks for Pattern Recognition. New York, United States: Oxford University Press, 2005.

CAI X. *et al.* A deep learning model incorporating part of speech and self-matching attention for named entity recognition of Chinese electronic medical records. In: 4TH CHINA HEALTH INFORMATION PROCESSING CONFERENCE, 2018.

CHEN T.; GUESTRIN C. XGBoost: A scalable Tree Boosting System. In: PROCEEDINGS OF THE 22ND ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, 2016. p. 785-794.

CARLSON A. *et al.* Coupling semi-supervised learning of categories and relations. In: PROCEEDINGS OF THE NAACL HLT 2009 WORKSHOP ON SEMI-SUPERVISED LEARNING FOR NATURAL LANGUAGE PROCESSING, 2009. p. 1-9.

CARLSON A. *et al.* Toward an architecture for never-ending language learning. In: PROCEEDINGS OF THE CONFERENCE ON ARTIFICIAL INTELLIGENCE (AAAI), 2010, p. 1306–1313.

COLLINS M; KOO T. Discriminative Reranking for Natural Language Parsing. In: COMPUTATIONAL LINGUISTICS, 2005. v. 31.

CURRAN J; MURPHY T; SCHOLZ B. Minimizing semantic drift with mutual exclusion Bootstrapping. In: PROCEEDINGS OF THE 10TH CONFERENCE OF THE PACIFIC ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 2007. v. 6.

DUARTE M. C. Aprendizado Semissupervisionado através de técnicas de acoplamento. 2011. Disponível em: https://repositorio.ufscar.br/bitstream/handle/ufscar/474/3777.pdf.

DUARTE M. C. Exploring two Views of Coreference Resolution in a Never-Ending Learning System. In: INTERNATIONAL CONFERENCE ON HYBRID INTELLIGENT SYSTEMS (HIS), 2014

FACELI K. *et al.* Inteligência Artificial Uma Abordagem de Aprendizado de Máquina. Rio de Janeiro: LTC – Livros Técnicos e Científicos Editora Ltda., ed. 1, 2011.

FLACH P. Machine learning: the art and science of algorithms that make sense of data. Cambridge University Press, 2012.

GOLDBERG Y. A primer on neural network models for natural language processing. In: JOURNAL OF ARTIFICIAL INTELLIGENCE RESEARCH, 2016, v. 57, p. 345-420.

GONZÁLEZ J.; HRUSCHKA E. R.; MITCHELL T. M. Merging Knowledge bases in different languages. 2017. Disponível em: http://www.aclweb.org/anthology/W17-2403

HAYKIN S. Neural Networks and Learning Machines. Hamilton, Ontario, Canada: Pearson, ed. 3, 1999.

HAYKIN S. Redes Neurais Princípios e Práticas. São Paulo: Bookman, ed. 2, 2007.

KE Y.; HAGIWARA M. A Natural Language Processing Neural Network Comprehending English. In: INTERNATIONAL JOINT CONFERENCE ON NEURAL NETWORKS (IJCNN). 2015. Disponível em: https://ieeexplore.ieee.org/abstract/document/7280492/

KRISHNAMURTHY J.; MITCHELL T. M. Which noun phrases denote which concepts? In: PROCEEDINGS OF THE 49TH ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS: HUMAN LANGUAGE TECHNOLOGIES, 2011. v. 1 p. 570–580.

KURZWEIL, R. The Age of Spiritual Machines. Massachusetts: The MIT Press, 1990.

MANSANO A. F.; HRUSHCKA E.R.; PAPA J.P. Co-reference Analysis Through Descriptor Combination. In: EUROPEAN CONGRESS ON COMPUTATIONAL METHODS IN APPLIED SCIENCES AND ENGINEERING, 2017. p. 525-534.

MARTINS P.; CUSTÓDIO L.; VENTURA R. A deep learning approach for understanding natural language commands for mobile service robots. Cornel University, 2018.

MITCHELL, T. M. Machine Learning. 1. ed. New York, NY, USA: McGraw-Hill, Inc., 1997.

MITCHELL, T. M. The discipline of machine learning. White paper, cmu-ml-06-108. [S.l.], July 2006. Disponível em: http://www.cs.cmu.edu/tom/pubs/MachineLearning.pdf

MITCHELL, T. M. *et al.* Never-Ending learning. In: PROCEEDINGS OF THE AAAI CONFERENCE ON ARTIFICIAL INTELLIGENCE (AAAI), 2015.

MITCHELL, T. M. *et al.* Never-ending learning. In: COMMUNICATIONS OF THE ACM, 2018. v. 61 p. 103-115.

NIRMALA K.; VENKATESWARAN N. L.; KUMAR C. V. HoG based Naive Bayes classifier for glaucoma detection. In: IEEE REGION 10 ANNUAL INTERNATIONAL CONFERENCE, Proceedings/TENCON, 2017. v. 2017 p. 2331-2336.

RUSSELL S.; NORVIG P. Inteligência Artificial. 2. Ed. Rio de Janeiro, RJ, Brasil: Elsevier Ltda, 2003. p. 629-630.

SILVA I. N.; SPATTI D. H.; FLAUZINO R. A. Redes Neurais Artificiais para engenharia e ciências aplicadas. 1. ed. São Paulo, SP, Brasil: Artliber Editora Ltda, 2010. p. 24.

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, 2009, v. 3, p. 1503-1512.

WIDROW B.; HOFF M. Adaptative switching circuits. Institute of Radio Engineers. Western Electronic Show and Convention, 1960.

WIJAYA D. T.; MITCHELL T.M. Mapping Verbs in Different Languages to Knowledge Base Relations using Web Text as Interlingua. HLT-NAACL, 2016.

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.