Universidade Federal do Rio de Janeiro

Escola Politécnica

MBA em Big Data, Business Intelligence e Business Analytics (MB3B)

TRADUÇÃO AUTOMÁTICA DE LINGUAGEM INFORMAL: UMA ANÁLISE DO IMPACTO DA SELEÇÃO E TRATAMENTO DE DADOS NA EFICÁCIA DE MODELOS DE APRENDIZADO DE MÁQUINA

Autor:

Electronically signed by:
Bernardo Ferreira Marques
Bernardo Ferreira Marques
bernardo.marques93@gmail.com
13/12/2023 00:01:30
ID: DGU2TCMRRGY-GE2DQOBTG4YA-44VUCVXNYBHFL

Bernardo Ferreira Marques

Orientador:

Electronically signed by:
Manoel Villas Boas Junior
mvbjunior@poli.ufrj.br
13/12/2023 16:22:00
ID: FGU2TCMRRGY-GE2DQOBTG4YA-TOPUOI2QDZPX7

Manoel Villas Boas Junior, M. Sc.

Examinador:

Electronically signed by:
Norberto Ribeiro Bellas
Norberto Ribeiro Bellas
norbertobellas@gmail.com
13/12/2023 17:42:14
ID: 0GU2TCMRRGY-GE2DQOBTG4YA-QSGRDEFMA7L5H

Norberto Ribeiro Bellas, M. Sc.

Examinador:

Electronically signed by:
Vinicius Drumond Gonzaga
viniciusdgonzaga@gmail.com
13/12/2023 17:34:41
ID: AGU2TCMRRGY-GE2DQOBTG4YA-YKFMPIBQQMNYG

Vinicius Drumond Gonzaga, M. Sc.

Rio de Janeiro
Agosto de 2023

Declaração de Autoria e de Direitos

Eu, **Bernardo Ferreira Marques** CPF 138.686.797-74, autor da monografia ***TRADUÇÃO AUTOMÁTICA DE LINGUAGEM INFORMAL: UMA ANÁLISE DO IMPACTO DA SELEÇÃO E TRATAMENTO DE DADOS NA EFICÁCIA DE MODELOS DE APRENDIZADO DE MÁQUINA***, subscrevo para os devidos fins, as seguintes informações:

1. O autor declara que o trabalho apresentado na defesa da monografia do curso de Pós-Graduação, Especialização MBA em Big Data, Business Intelligence e Business Analytics da Escola Politécnica da UFRJ é de sua autoria, sendo original em forma e conteúdo.

2. Excetuam-se do item 1 eventuais transcrições de texto, figuras, tabelas, conceitos e ideias, que identifiquem claramente a fonte original, explicitando as autorizações obtidas dos respectivos proprietários, quando necessárias.

3. O autor permite que a UFRJ, por um prazo indeterminado, efetue em qualquer mídia de divulgação, a publicação do trabalho acadêmico em sua totalidade, ou em parte. Essa autorização não envolve ônus de qualquer natureza à UFRJ, ou aos seus representantes.

4. O autor declara, ainda, ter a capacidade jurídica para a prática do presente ato, assim como ter conhecimento do teor da presente Declaração, estando ciente das sanções e punições legais, no que tange a cópia parcial, ou total, de obra intelectual, o que se configura como violação do direito autoral previsto no Código Penal Brasileiro no art.184 e art.299, bem como na Lei 9.610.

5. O autor é o único responsável pelo conteúdo apresentado nos trabalhos acadêmicos publicados, não cabendo à UFRJ, aos seus representantes, ou ao(s) orientador(es), qualquer responsabilização/ indenização nesse sentido.

6. Por ser verdade, firmo a presente declaração.

Rio de Janeiro, 05 de agosto de 2023.

Electronically signed by:
Bernardo Ferreira Marques

Bernardo Ferreira Marques
bernardo.marques93@gmail.com
13/12/2023 00:01:30

ID: DGU2TCMRRGY-GE2DQOBTG4YA-44VUCVXNYBHFL

Bernardo Ferreira Marques

UNIVERSIDADE FEDERAL DO RIO DE JANEIRO

Av. Athos da Silveira, 149 - Centro de Tecnologia, Bloco H, sala - 212,
Cidade Universitária Rio de Janeiro – RJ - CEP 21949-900.

DEDICATÓRIA

Dedico este trabalho aos meus pais, pois seus esforços e trabalho duro me incentivaram a ir além, possibilitando a conclusão deste curso.

AGRADECIMENTO

Agradeço primeiramente a Deus pelo dom da vida e por iluminar o meu caminho durante essa caminhada, me dando força e perseverança para superar todas as dificuldades.

Ao professor Manoel Villas Boas Júnior, obrigado pela orientação, apoio e confiança.

Aos meus familiares, em especial os meus pais e irmãos, pela educação, ensinamentos, confiança e apoio ao longo de toda a minha vida. Obrigado por sempre acreditarem em mim e me incentivarem quando para mim parecia difícil! São as pessoas pelas quais me levanto todo dia para vencer os desafios e deixá-los orgulhosos.

RESUMO

Com o crescente uso da internet para além dos fins comerciais, percebe-se também a maior produção e exposição a conteúdos nos mais diversos idiomas, populares ou não. Quer seja na leitura de *blogs* pessoais ou publicações em redes sociais, quer seja em conversações diretas, tais conteúdos refletem, na medida do possível, a comunicação informal e espontânea que presenciamos fora do ambiente virtual, com o detalhe de que em um mundo mais globalizado e conectado virtualmente as distâncias geográficas possuem cada vez menos impacto. Apesar dos grandes avanços nas áreas de linguística e computação e, consequentemente, nas ferramentas de Tradução Automática, os desafios em detectar padrões objetivos na linguagem informal, como em casos de sarcasmo ou uso de expressões não literais, ainda motivam inúmeros estudos que visam desenvolver ferramentas de tradução que dependam cada vez menos da revisão humana. Tendo em vista a disponibilidade de diversos modelos de aprendizado de máquina, em particular para o Processamento de Linguagem Natural, capazes de detectar padrões quando expostos a uma quantidade razoável de dados, o presente trabalho tem como objetivo focar os esforços em analisar a eficácia de tais modelos frente a bases de dados de maior qualidade e adequação ao que se pretende produzir, ou seja, com dados textuais gerados justamente em contextos informais, e identificar até que ponto esses algoritmos são capazes de detectar padrões nas situações em que se deve traduzir literalmente uma frase, e quais situações exigem maior flexibilidade e análise de contexto para uma tradução mais realista. Apesar de ser capaz de traduzir algumas frases que possuem equivalente parecido na língua inglesa, o modelo desenvolvido gerou também algumas traduções literais do texto original e mostrou não distinguir adequadamente as diferentes traduções que uma palavra pode apresentar a partir de sua função no texto (sujeito, objeto direto ou indireto etc.) e, em alguns casos, gerou frases traduzidas sem um sentido completo. Dentre as limitações do trabalho, encontra-se o fato de que o *corpus* utilizado foi coletado automaticamente utilizando a técnica de *Web Scraping*, o que não assegura a boa qualidade de todas as traduções, e dificulta o tratamento devido dos dados. Por fim, sugere-se ainda a utilização de modelos híbridos para garantir a melhor compreensão e geração de textos, integrando as análises sintática e semântica.

Palavras-Chave: Processamento de Linguagem Natural, *Corpora* bilíngues, Expressões idiomáticas

ABSTRACT

Along with the ever-growing use of internet for commercial purposes, it is possible to notice the production and exposure to content in several languages in a very larger scale. Whether on personal blogs, social media posts, or in direct conversations, such content often mirrors informal and spontaneous communication observed in real-life interactions, albeit with the caveat that in our increasingly globalized and digitally connected world, geographical distances have a diminished impact. Despite recent breakthroughs in the fields of linguistics and computation and, consequently, in the tools of Machine Translation, it is still with much struggle that the machine can detect certain patterns in informal language, as with the cases of sarcasm or the use of non-literal expressions, and it motivates several studies aiming to develop tools that can rely less on human review. Considering the amount of Machine Learning models available, in particular for Natural Language Processing, able to detect patterns when exposed to a reasonable amount of data, the present work focuses rather on analyzing their effectiveness with text data that are more suitable for informal context, and identifying up to what extent these algorithms are able to detect patterns in situations when a sentence must be translated literally and which situations require more flexibility and context analysis for a more assertive and realistic translation. Although the model was able to correctly translate some sentences that contain similar idiom in English, it also generated some literal translations of the original sentence and, apparently, was not able to accurately distinguish the different translations a word might have, depending on how it functions in meaning as well as grammatically within a sentence. In some cases, the model also produced some translated sentences that had no meaning, emphasizing the importance of hybrid models with some predefined rules. One of the limitations is the fact that the *corpus* used was generated automatically through Web Scraping, which makes it difficult to ensure the quality of all the translations and it makes the data cleaning process much harder. In the end, it is also suggested the use of hybrid models to ensure a better text understanding and text generation, integrating syntactic and semantic analysis.

Keywords: Natural Language Processing, Bilingual *corpora*, Idioms

SIGLAS

LSTM	*Long-Short Term Memory*
ML	*Machine Learning*
PLN	Processamento de Linguagem Natural
RNA	Redes Neurais Artificiais
RNN	*Recurrent Neural Network*
TA	Tradução Automática
URL	*Uniform Resource Locator*

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE QUADROS

Sumário

CAPÍTULO 1

Introdução

Apesar da importância da norma culta e do domínio da linguagem formal, a maior parte das interações verbais se dá no contexto informal, seja pela oralidade, seja por meio de conversas virtuais por escrito ou quaisquer outros instrumentos de comunicação. No contexto dessas interações nos deparamos com erros de concordância verbal, numérica, de digitação e muitos outros que, na maioria dos casos, são facilmente contornados pelo raciocínio humano, tendo em vista a espontaneidade das conversas e toda a nossa compreensão do contexto da interação.

Sabe-se que a língua, tanto formal quanto informal, muda significativamente com o passar do tempo. Expressões caem em desuso, novas expressões surgem e até mesmo os acordos ortográficos estão sujeitos à revisão, como no caso da língua portuguesa em 2009, além do tempo, o espaço também é uma variável nesse fenômeno. Leite (2010) argumenta que nem mesmo do ponto de vista científico encontra-se respaldo para idealizar a normatização de um único modo de se comunicar por todos os falantes do idioma. Portanto, as variedades linguísticas constituem parte essencial do uso de uma língua que sofre variações como todas as outras. Dentre as possíveis razões listadas pelo autor, encontram-se o convívio com as línguas dos povos colonizados, bem como com as de imigrantes, a distância geográfica entre as diversas regiões onde a língua é falada e a diferença no ritmo com que ocorrem mudanças sociais e culturais em cada região ou mesmo ambientes específicos, como as indústrias ou o campo, por exemplo.

Quando migramos para o contexto da globalização, a complexidade é ainda maior. As comunicações com nativos de outros idiomas têm se tornado cada vez mais frequentes, bem como o acesso a vídeos e músicas estrangeiras dos mais variados gêneros. Nesse cenário, é comum deparar-se com expressões que impõem um desafio muito particular para o processo de tradução, quer seja humana, quer seja de máquina, tendo em vista que, em muitos desses casos, o sentido da frase completa é inferido de forma lógica e não exatamente pelo significado de cada uma das palavras presentes (ALZEEBAREE, 2020). Com isso, a simples tradução literal nem sempre viabiliza uma comunicação adequada e o desafio então passa a ser a transposição do sentimento expresso em uma língua, com

todos os aspectos culturais presentes, para algo equivalente em outra. Um exemplo básico disso é a palavra "saudade", tão presente na língua portuguesa e que não possui um equivalente exato em outros idiomas. O trabalho humano de tradução torna-se essencial nesse momento, sobretudo vindo de alguém que tenha uma experiência significativa com os dois idiomas, porém o resultado desse trabalho ainda é sujeito à subjetividade do profissional e ao seu método de explicar o sentido original do que foi dito em um idioma para pessoas que não tem qualquer conhecimento da língua. Dessa forma, justamente por não haver um método definitivo para tradução de qualquer texto entre quaisquer idiomas, a confiabilidade da automatização de um serviço como esse ainda representa um grande desafio para profissionais das áreas de linguística e computação.

A razão de os textos serem considerados um desafio tão grande para a computação é justamente o fato de que os limites da comunicação humana não são tão previsíveis com regras explícitas como as linguagens de programação ou notações matemáticas (BIRD ET AL., 2009). Conforme dito anteriormente, a linguagem natural sofre inúmeras variações ao longo do tempo e seu uso em diferentes contextos faz com que não haja um único padrão aplicável a um idioma. Nos dias de hoje, em que é necessário lidar com uma massa de dados cada vez maior sem um pré-tratamento específico, textos enquadram-se no que se denomina comumente como dados não estruturados e sua estrutura original é destinada à compreensão humana e não de computadores (PROVOST; FAWCETT, 2013). Assim sendo, os estudos do Processamento de Linguagem Natural (definido na Seção 2) avançam no sentido de automatizar uma melhor compreensão objetiva de dados de textos, apesar do fator humano que dá origem a dados desse tipo. As aplicações são inúmeras, incluindo análises de sentimentos relacionados a bens e serviços que auxiliem na estratégia de *marketing*, detecção de plágio e serviços de tradução.

Atualmente, já estão disponíveis inúmeras ferramentas de tradução automática, muitas das quais são inclusive gratuitas. No entanto, fatores como ambiguidade de palavras e/ou frases ainda exigem alguma edição manual para que se tenha uma tradução adequada (ALZEEBAREE, 2020). Sendo assim, o objetivo do presente trabalho é implementar um modelo de tradução automática treinado para a compreensão de textos informais e que possa ser utilizado principalmente em conversas virtuais, sujeitas a erros de digitação e/ou gramaticais, e especializado na tradução de frases com sentido não literal. A estratégia deste estudo consiste em avaliar a performance do modelo investindo na qualidade da base de dados fornecida e em como ela é adequada ao contexto em que se deseja implementar a ferramenta, mais do que na complexidade do modelo matemático

ou em seus hiperparâmetros. Vale destacar que será considerada apenas a modalidade escrita da língua.

1.1 – Objetivos

O objetivo principal do presente trabalho é o desenvolvimento de um modelo de tradução automática capaz de distinguir expressões comuns de expressões idiomáticas nas línguas inglesa e portuguesa e providenciar uma tradução mais alinhada às necessidades de quem utiliza a tradução automática para comunicação informal.

A estratégia para alcançar o objetivo do projeto consistiu basicamente em: (1) extrair uma base de dados textual com frases em português e sua respectiva tradução em inglês, prezando por frases de origem não necessariamente formal, como letras de músicas de quaisquer gêneros; (2) realizar um tratamento e limpeza dos dados, assegurando a devida separação das frases em cada idioma, remoção de caracteres especiais ou que não acrescentam sentido às frases; (3) treinar o algoritmo com modelos conhecidos de aprendizado de máquina; e, por fim, (4) avaliar a eficácia do modelo na tradução de textos novos, seguindo o mesmo padrão dos que foram utilizados no treinamento.

1.2 – Metodologia

A metodologia da pesquisa foi desenvolvida nas seguintes etapas: revisão bibliográfica dos conceitos relevantes; pesquisa de aplicações recentes no tema para verificação do estado da arte; emprego de técnicas de *Web Scraping*, uma técnica de raspagem de dados *online*, para obtenção de amplas bases de dados contendo textos traduzidos entre as línguas inglesa e portuguesa; e, por fim, desenvolvimento, treino e validação do modelo de tradução automática.

As pesquisas por artigos publicados no tema de tradução automática foram realizadas principalmente no Google Acadêmico, a partir do ano de 2018. Pesquisando pelo termo "*machine translation*", notou-se que os trabalhos de forma geral apresentam um foco maior nos algoritmos utilizados para o treinamento dos modelos, aprofundando as questões de matemática avançada presentes nesses algoritmos. Já com as palavras-chave "*machine translation idiom*", é possível encontrar pesquisas que exploram formas

alternativas de gerar dados que possam ensinar o modelo a distinguir expressões não literais e que necessitem de uma adaptação no momento da tradução. Tendo em vista que o desafio inicial do presente trabalho foi a obtenção de uma quantidade razoável de bases de dados contendo traduções de textos em inglês e português, a pesquisa foi orientada no sentido de organizar a base final de dados de forma que tornasse o mais nítido possível a distinção de frases que podem ser traduzidas de forma literal e quais necessitam de uma análise mais concentrada na semântica e no sentimento expresso no texto. Assim sendo, a escolha do modelo matemático foi baseada nos algoritmos mais utilizados nos trabalhos recentes sem se aprofundar no algoritmo em si.

Para treinar o modelo, foi necessário obter bases de dados contendo frases originais em português com sua respectiva tradução em inglês. Para isso, foi realizada uma coleta de dados do *site* "lyricstranslate.com". Trata-se de uma plataforma em que os usuários podem enviar e consultar letras de músicas de qualquer idioma, além de requisitar e produzir traduções dessas músicas em quaisquer outros idiomas, dentre tantas funcionalidades linguísticas. Se por um lado isso contribui para uma maior diversidade de temas e, consequentemente, vocabulário, por outro, não há garantia de que as traduções estejam sempre corretas ou precisas. A fim de realizar a coleta automática dos dados presentes no *site*, foram filtradas as músicas com a letra original exclusivamente em português e que necessariamente possuíssem tradução completa em inglês. Além disso, ao apresentar a letra de uma determinada música com sua tradução, o *site* ainda lista no final as expressões idiomáticas presentes na música, caso haja.

Para fins de simplificação, foram obtidos apenas *corpora* com original em português e tradução em inglês, e não vice-versa. Tendo em vista que o modelo matemático será alimentado por uma base única de textos em português e inglês, optou-se por evitar o risco de que bases geradas originalmente em línguas distintas pudesse gerar algum tipo de confusão no algoritmo ao tentar reconhecer padrões no processo de tradução.

Para a implementação do modelo de Redes Neurais Artificiais utilizado neste trabalho e realização de experimentos de tradução foi utilizado um *framework* de um modelo de tradução automática construído com as bibliotecas *Tensorflow* e *Keras* na linguagem *Python*. Trata-se de um modelo sequência para sequência (mais conhecido como S*eq2Seq*) que é baseado na arquitetura *Transformer*. O programa, disponível publicamente no site oficial do *Keras* e projetado inicialmente para tradução do inglês para o espanhol, foi adaptado para lidar especificamente com o conjunto de dados gerado

no presente trabalho, sendo necessário ajustar alguns parâmetros como o tamanho do vocabulário em cada idioma e comprimento máximo de uma frase, além prepará-lo para traduzir do português para o inglês.

1.3 – Organização do Documento

O restante do trabalho está organizado da seguinte maneira: A Seção 2 apresenta uma breve contextualização do tema, definições dos termos importantes para compreensão de um trabalho voltado para o Processamento de Linguagem Natural e a revisão bibliográfica com a evolução histórica do tema. Na Seção 3 são explicados os modelos e algoritmos utilizados, bem como os detalhes dos experimentos realizados. Na Seção 4 são apresentados e analisados os resultados obtidos. O estudo é concluído na Seção 5 com as limitações e sugestões de pesquisas futuras na área de tradução automática de expressões idiomáticas.

CAPÍTULO 2

Embasamento Teórico

A fim de compreender os trabalhos já realizados na área de Processamento de Linguagem Natural, bem como de implementar novos modelos e utilizar ferramentas computacionais disponíveis, é essencial compreender algumas das principais terminologias utilizadas por profissionais de tecnologia e linguística. Por esse motivo, nesta seção serão apresentados conceitos importantes para a compreensão deste trabalho, bem como uma breve contextualização do que representam as expressões idiomáticas.

2.1 – Processamento de Linguagem Natural

Iniciando pela área do conhecimento que abrange o trabalho como um todo, começamos pela definição do Processamento de Linguagem Natural (PLN). Liddy (2001) define PLN como o "conjunto de técnicas computacionais motivadas teoricamente para analisar e representar naturalmente textos em um ou mais níveis de análise linguística, a fim de alcançar o nível de processamento humano para uma série de tarefas e aplicações". Ao estabelecer como parâmetro de desempenho ótimo a capacidade humana, é importante destacar que o PLN aborda não apenas a capacidade de compreensão da linguagem natural, a qual o ser humano processa por meio da leitura e/ou da audição, mas também da sua capacidade de produção, assim como o ser humano fala e escreve textos.

A fim de otimizar o processo de compreensão de texto, Chopra et al. (2013) lista cinco fases de processamento: análise léxico-morfológica, que abrange a definição do vocabulário presente, separação de frases, parágrafos e palavras, além da identificação e análise da estrutura das palavras presentes; análise sintática, que determina a função de cada termo em uma oração; análise semântica, ou seja, do significado; integração do discurso, dado que o significado de uma oração é influenciado pelas orações precedentes, bem como evoca o sentido das orações seguintes; e, por fim, análise pragmática, relacionada ao objetivo da comunicação.

Tendo em vista a complexidade do processo descrito acima, serão apresentados alguns conceitos básicos e terminologias frequentemente utilizadas nas mais diversas aplicações de PLN:

- ***Corpus*:** palavra em latim significa cujo significado é "corpo". No contexto linguístico, *corpus* é usado para descrever um corpo ou mesmo um conjunto de textos registrados utilizados para estudo e análise. O plural de *corpus* é *corpora*. Para a realização de estudos bilíngues, incluindo modelos de tradução, como neste trabalho, utiliza-se normalmente *corpus* paralelo, contendo textos em um idioma e sua respectiva tradução em outro;
- **N-grama:** muito utilizada para previsão da próxima palavra, o modelo de n-grama assume que a probabilidade da próxima palavra depende apenas das últimas *k* palavras já fornecidas (MANNING; SCHUTZE, 1999). Dessa forma, considera-se não apenas o sentido e a probabilidade de uma determinada palavra isoladamente, mas também o fato de que ela pode estar associada a outras palavras, de forma que n palavras possam aparecer juntas no mesmo texto. Os exemplos mais comuns de n-gramas envolvem bigramas ($n = 2$), trigramas ($n = 3$) e quatro grama ($n = 4$). Um exemplo frequente de bigrama comum seria "semana passada";
- **Lematização:** consiste em reduzir palavras à sua forma raiz, ao seu lema. Os verbos são representados por sua forma no modo infinitivo, já os substantivos e adjetivos são colocados no masculino singular. O foco não está nas declinações ou flexões das palavras, nem mesmo em como determinado verbo deve ser conjugado e sim no seu sentido. Portanto, "fazer", "fará", "feito" representam igualmente o verbo fazer;
- **Tokenização:** uma das etapas do pré-processamento de texto que consiste em dividi-lo em unidades menores, chamadas *tokens*. Dessa forma, de acordo com o escopo da análise, os *tokens* podem ser palavras, orações ou caracteres, por exemplo;
- ***Stop words*:** em português são conhecidas comumente como palavras vazias, ou seja, palavras que normalmente podem ser ignoradas quando a análise é focada em obter informações importantes a respeito de palavras-chave e outros termos importantes, tendo em vista que não adicionam sentido ao texto. Não há um consenso sobre qual seria o conjunto exato de todas essas palavras,

apesar de ser muito comum incluir conjunções, preposições e verbos como "ser" e "poder", por exemplo;

- ***Bag of words*:** em tradução livre significa bolsa/saco de palavras. Quando se considera analisar um conjunto de documentos, cada documento é visto como um conjunto individual, as palavras presentes em todo o conjunto de documentos são consideradas variáveis que, no caso mais básico de aplicação teria o valor igual a 1, caso esteja presente em tal documento, ou igual a 0, caso contrário. Dessa forma, trata-se de representar de forma vetorizada o conjunto de documentos, mapeando potenciais palavras-chave sem levar em consideração a gramática e a ordem em que as palavras aparecem (PROVOST; FAWCETT, 2013);
- ***Word Frequency*:** uma etapa ainda mais avançada que consiste em contabilizar a frequência com que as palavras aparecem em um documento, em vez de classificar apenas como estando presente ou não. Assim é possível uma melhor análise do quanto uma palavra é importante para o documento pelo número de vezes que se repete. É importante que nessa fase sejam realizados alguns pré-processamentos, como a remoção das *stop words*, que não agregam na semântica do texto e técnicas como a de lematização, a fim de agrupar substantivos ou verbos que possuem a mesma forma original, mas que se apresentam em diferentes formas no mesmo documento, seja pela flexão de gênero ou número ou ainda variações na conjugação;
- ***Dependency Parsing*:** a análise de dependências consiste em dividir um documento em palavras e classificar a relação de dependência entre elas. Com isso, é possível determinar a função gramatical de cada palavra no texto, podendo definir quem é o sujeito, o verbo principal, objetos direto e indireto da frase e possibilita, consequentemente, uma melhor análise semântica.

2.2 – Expressões Idiomáticas

A semântica é a área da linguística que estuda o significado de palavras e frases. No entanto, a compreensão de frases ou até mesmo de grupos de palavras na linguagem natural nem sempre se restringe à compreensão individual de cada palavra presente. Nesses casos em que a relação entre o significado das palavras e o de uma frase não é

necessariamente clara, temos as expressões idiomáticas (MANNING; SCHUTZE, 1999).

Xatara (2001) define "expressão idiomática" como "uma lexia complexa indecomponível, conotativa e cristalizada em um idioma pela tradição cultural". Por ser indecomponível, pode ser bem compreendida somente na sua forma completa, assumindo pequenas variações como em gênero número e grau ou, por exemplo, as diferentes conjugações de verbos. No entanto, a autora apresenta ainda um ponto que por representar um desafio ao ser humano, consequentemente torna ainda mais complexa a tarefa de preparar algoritmos de interpretação de texto para lidar com tais expressões: saber o momento em que se está diante de uma expressão idiomática.

Por estar tão relacionada a contextos locais, normalmente um mesmo indivíduo não irá conhecer nem mesmo compreender todas as expressões existentes, mesmo em sua língua materna. Por isso mesmo, o processo de compreensão humana é mais flexível nesse ponto. A menos que já esteja familiarizado com a expressão, um indivíduo buscaria entender a mensagem primeiramente de forma literal e, ao notar alguma incoerência, poderia suspeitar estar diante de uma expressão não literal. Em seguida, resta ainda o desafio de compreender o significado de tal expressão.

A seguir, temos exemplos de algumas expressões da língua portuguesa que, em alguns casos, poderiam ser interpretadas literalmente, mas dentro de determinados contextos um ser humano normalmente identificaria como tendo um significado metafórico:

- **Levar um banho de água fria**: se decepcionar, usada em situações de muitas expectativas seguidas de grande frustração;
- **Chutar o balde**: agir de forma irresponsável em relação a algum problema, desistir de lidar com alguma situação;
- **Dar uma mão**: oferecer ajuda a alguém;
- **Sem pé nem cabeça**: algo que não tem lógica ou que não faz sentido;
- **Sentir dor de cotovelo:** sentir inveja de alguém.

Por estarem muito relacionadas a contextos locais, expressões como essas podem ou não ter equivalentes em outro idioma. No entanto, ainda que possuam equivalente, a mesma expressão no outro idioma pode utilizar outros verbos ou substantivos de forma não literal. Como exemplo, a expressão "santo do pau oco", usada para acusar uma pessoa como falsa ou fingida, poderia ser traduzida apenas como "*false*" ou "*hypocrite*" em inglês, que significam apenas "falso" ou "hipócrita" em português. Já a expressão "não

coloque a carroça na frente dos bois", utilizada para aconselhar alguém a não se afobar, tem seu equivalente em inglês como "*hold your horses*", que literalmente significa, em tradução livre, "segure os seus cavalos".

2.3 – Aprendizado de Máquina

De acordo com Provost e Fawcett (2013), os métodos de Aprendizado de Máquina, no inglês *Machine Learning* (doravante denominado *ML*), consistem em extrair modelos preditivos dos dados. Ao contrário do método tradicional em que a inteligência artificial recebe os dados, é programada a priori para executar determinados procedimentos e, por fim, entregar o resultado buscado pelo usuário, os modelos de *ML* consistem em receber dados de entrada com seus respectivos resultados e retornar uma regra geral (ou uma fórmula matemática) responsável por tal transformação.

No entanto, as técnicas presentes nessa subárea da Inteligência Artificial geralmente trabalham com incertezas e, por isso mesmo, sua implementação requer análises mais profundas de acurácia e outras métricas estatísticas que assegurem um bom desempenho do modelo gerado.

Os desafios que os modelos de *ML* se propõem a resolver são inúmeros, porém quase sempre buscando espelhar o modo com que seres humanos analisam os dados a sua disposição. Como Müller e Guido (2016) destacam, o processo original de desenvolver modelos em que todas as condições eram previstas funcionam para uma variedade de aplicações, porém em outras áreas do conhecimento, necessitariam um trabalho muito mais profundo sobre o processo exato com que especialistas humanos tomam decisões. Dessa forma, precisaríamos traduzir para as linguagens de computação e matemática todo o processamento do raciocínio humano, o que para a maioria dos casos seria inviável.

Um exemplo muito comum em que se torna necessário o uso de técnicas de *ML* é o reconhecimento de imagem. Enquanto o ser humano é capaz de olhar para uma outra pessoa de diversos ângulos diferentes, que provavelmente pode reproduzir inúmeras expressões faciais, e ainda assim reconhecê-la, o computador lê apenas *pixels*. Dessa forma, seria um trabalho muito custoso encontrar todas as combinações possíveis de pixels associadas à imagem de uma mesma pessoa e entregar essas informações ao computador. Para contornar esse problema, é possível, com as técnicas atuais de *ML*, apenas fornecer um número razoável de dados (nesse caso, imagens) para que a máquina

seja capaz de detectar de forma autônoma as características comuns para detectar determinado objeto ou face em imagens distintas, ainda que não necessariamente com 100% de precisão.

Dentro do subdomínio do *ML*, encontra-se uma subárea denominada Aprendizado Profundo (do inglês *Deep Learning*), que tem como estrutura as Redes Neurais Artificiais (RNA). Trata-se de modelos baseados no sistema nervoso humano, com propósito de simular a habilidade de aprendizagem humana na obtenção de conhecimento (FACELI et al., 2021).

Inspiradas nas redes neurais biológicas, o elemento básico de uma rede neural artificial é o neurônio, que representa o cálculo de uma função matemática. Geralmente, as RNAs são estruturadas em camadas, cada qual com um dado número de neurônios. Em resumo, o processo que ocorre em cada neurônio é receber dados de entrada, ponderá-los através de alguns pesos, que correspondem às sinapses realizadas nos neurônios biológicos e que qualificam as importâncias de cada entrada, processar tudo através de uma função de soma e, por fim, de uma função de ativação.

As RNAs se dividem em dois tipos principais: Rede Neural *Feed-Forward* e Rede Neural Recorrente (RNN - do inglês *Recurrent Neural Network*) (RUSSELL; NORVIG, 2009). No primeiro, cada neurônio se conecta apenas com o neurônio posterior, enquanto na *RNN*, como o próprio nome sugere, pode haver recorrência de um neurônio consigo mesmo.

Em uma RNA, o processo de otimização da hipótese se baseia na técnica de retropropagação (ou *back-propagation*), onde, no final de cada época, calcula-se o gradiente da função de perda com respeito a cada peso da rede, partindo da camada de saída e propagando-se pelas camadas ocultas de trás para frente. Os gradientes calculados são usados por algoritmos de otimização para atualizar os pesos da rede (RUSSELL; NORVIG, 2009).

O principal diferencial das *RNNs* é o seu mecanismo de memória. Dessa forma, nos casos em que há dependência temporal entre os dados fornecidos, os resultados de cada neurônio são determinados pelos valores das camadas anteriores, bem como o valor passado do próprio neurônio em função da recorrência. Portanto, as *RNNs* são particularmente importantes no reconhecimento de padrões quando os dados do presente são influenciados pelos dados passados, como no caso de séries temporais e PLN (LEANDRO, 2021).

Por fim, uma forma ainda mais avançada e atual de *RNNs* é denominada Redes de Memória de Longo Prazo (*LSTM*, do inglês *Long-Short Term Memory*), proposta por Hochreiter e Schmidhuber (1997). Para entender a necessidade atendida pela *LSTM*, é importante saber que o treinamento das Redes Neurais pode contar com um mecanismo conhecido como *Backpropagation*, a partir do qual os pesos da rede são ajustados de acordo com o erro obtido na iteração anterior, tornando o modelo mais confiável e com maior capacidade de generalização (PIRES, 2019). Sendo assim, os ajustes são realizados de acordo com a mudança ocorrida na camada imediatamente anterior, não atualizando os erros das camadas iniciais, que estão diretamente envolvidas no reconhecimento dos dados de entrada. Dessa maneira, a rede como um todo não é otimizada com a eficiência adequada.

Para solucionar esse problema e alcançar memória de longo prazo, na *LSTM* cada neurônio possui um estado e três portas, a saber: *Input Gate*, que decide o quanto atualizar cada valor no estado do neurônio com novos valores de entrada; *Output Gate*, filtra o valor de saída com base no estado atual do neurônio; e *Forget Gate*, que decide o quanto deve ser esquecido de cada valor no estado do neurônio. Essas portas contém uma função sigmóide de ativação, gerando um valor entre 0 e 1, com 0 representando um nível de ativação completamente fechado, e 1 correspondendo a um nível de ativação completamente aberto. O *LSTM* tem demonstrado bastante eficácia e tem sido a principal forma de implementação de *RNN* para resolução de diversos problemas práticos (LEANDRO, 2021).

2.4 – Evolução da Tradução Automática

De acordo com Hutchins (1994), tradução automática é o termo referente a sistemas computadorizados responsáveis pela produção de traduções com ou sem a assistência de seres humanos. Como parte do processo de evolução da inteligência artificial, das ciências computacionais e linguísticas, o objetivo maior é o de automatizar completamente o processo de tradução semelhante ao modo como é realizado pelo ser humano, embora na prática esse trabalho ainda seja comumente submetido a uma revisão posterior. Embora na língua inglesa seja utilizado o termo "*Machine Translation*" (em tradução livre, tradução de máquina), a terminologia comumente utilizada no Brasil é "Tradução Automática" (TA), que será utilizada ao longo deste trabalho.

2.4.1 – Histórico – de 1947 até os dias atuais

Os primeiros passos no emprego das técnicas numéricas e computacionais nos processos de TA datam de pelo menos 1947, no período pós Segunda Guerra. Um memorando de Warren Weaver enviado ao professor Norbert Wiener marcou o início da pesquisa de TA com a primeira demonstração pública em 1954 de um protótipo de um sistema de tradução Russo-Inglês (ALZEEBAREE, 2020).

O período de 1954 a 1966 marcou uma era de otimismo com relação ao avanço da TA. No entanto, a lógica dos primeiros sistemas consistia basicamente em grandes dicionários bilíngues que ofereciam um ou mais resultados para cada palavra inserida, além de algumas regras bem definidas para a geração da ordem das palavras no idioma final. Um sistema apresentado pela IBM em 1954 consistia de 250 palavras e era capaz de traduzir 49 frases já conhecidas da língua russa para a língua inglesa. Esses e outros sistemas foram surgindo no contexto da Guerra Fria e, consequentemente, aumentando os investimentos em pesquisa de TA.

Apesar das grandes expectativas com os avanços nas áreas da computação e linguística, os pesquisadores não conseguiam encontrar soluções triviais e de regras bem definidas para os problemas de semântica. Sistemas como o Mark II (desenvolvido pela IBM em conjunto com a Universidade de Washington), apesar de atender boa parte da demanda por informação gerada rapidamente, ainda apresentava resultados de baixa qualidade. Em 1966, o governo dos Estados Unidos publica um relatório do Comitê Consultivo de Processamento Automático de Linguagem (ALPAC), que concluía que os programas de TA não eram eficientes em termos de custo. Com base nos resultados já obtidos até então, o documento questionava o fato dos sistemas ainda dependerem significativamente da intervenção humana. Acreditando não haver ainda conhecimento teórico e tecnologia suficiente para o desenvolvimento da TA, foram cortados os investimentos em pesquisa na área e até 1975 já não havia mais nenhum projeto financiado pelo governo (MELO, 2013). No entanto, ainda havia pesquisas em outras partes do mundo e em 1976 a comissão da União Europeia instalou uma versão Inglês-Francês do Systran, iniciado em 1970 e utilizado pela Força Aérea dos Estados Unidos para traduções Inglês-Russo. Além disso, ainda em 1976 surgiu no Canadá o Meteo, outro projeto bem-sucedido utilizado para tradução de previsões do tempo.

No final dos anos 1970 e início dos anos 1980, a demanda por TA não somente aumentou como diversificou. Não mais restrito ao contexto da Guerra Fria e a traduções

Inglês-Russo e Russo-Inglês, a procura pelos serviços de tradução cresceu na Europa, Canadá e Japão em razão do comércio internacional e das demandas de comunidades multilíngues (HUTCHINS, 2014).

A década de 1980 marcou o surgimento de uma grande variedade de novos sistemas de TA em diversos países e com custos cada vez menores. Além disso, foram desenvolvidos novos projetos de pesquisa com novas abordagens à lógica utilizada nas traduções, envolvendo análises semânticas, morfológicas, sintáticas, dentre outras. Alguns dos principais projetos foram: GETA-Ariane (Grenoble), SUSY (Saarbrücken), Mu (Kyoto), DLT (Utrecht), Rosetta (Eindhoven), Carnegie-Mellon University (Pittsburgh), além de projetos multilíngues como o Eurotra (Comunidade Européia) e do CICC (Japão), que contou com participantes na China, Tailândia e Indonésia.

Já na década de 1990, começaram a ser publicados experimentos baseados em métodos estatísticos e em *corpora* com exemplos de tradução. Os novos métodos já não eram baseados em regras pré-definidas, apesar de ainda terem surgido novos projetos com as antigas abordagens. Outra inovação desse período foram as pesquisas em tradução de áudio, integrando módulos de reconhecimento e transcrição de áudios em texto e os métodos de tradução já desenvolvidos até então.

A partir de então, cresceu a comercialização de aplicações práticas em diversos formatos. A demanda vinha desde grandes empresas que necessitavam da tradução de documentos diversos até *softwares* para uso em computadores pessoais. Mais do que a qualidade, a procura era por sistemas capazes de produzir traduções razoavelmente boas praticamente em tempo real, o que deu lugar para o surgimento de ferramentas online como o Google Tradutor.

Da década de 2000 até os dias atuais, os métodos baseados em estatística têm sido cada vez mais dominantes no campo da TA. Uma das principais razões para isso é a crescente disponibilidade de grandes conjuntos de dados, em particular *corpora* bilíngues, o que facilita os algoritmos a identificar padrões semânticos que não seriam transcritos de forma trivial em modelos com regras pré-definidas. Outra vantagem é o fato de esses métodos se aplicarem para a tradução independente das línguas selecionadas. Ainda assim, os métodos híbridos continuam recebendo a devida atenção, tendo em vista a adequação dos métodos baseados em regras pré-definidas para análises sintáticas, morfológicas, transliteração de nomes etc.

2.4.2 – Métodos e abordagens

Conforme visto no histórico, os sistemas de TA se dividem em basicamente dois tipos de acordo com sua abordagem: (1) sistemas baseados em regras; (2) sistemas baseados em *corpus*. O primeiro conta principalmente com a participação de especialistas humanos em linguística capazes de especificar um conjunto de regras que representem o processo de tradução. No segundo, os especialistas precisam fornecer apenas um *corpus* bilíngue suficientemente grande, com exemplos de traduções realizadas manualmente. Nesse caso, trata-se de um trabalho de *ML*, em que são fornecidos os dados de entrada e saída, e o resultado esperado do sistema é uma aproximação do conhecimento necessário para traduzir novas frases, conhecimento esse que anteriormente deveria ser escrito de forma exaustiva pelos especialistas. E, tendo em vista que ambos os métodos são mais adequados para diferentes tipos de análises, é possível também trabalhar com sistemas híbridos, que consistem basicamente na combinação dos dois métodos.

2.4.2.1 – Sistemas baseados em regras

Principal abordagem dos primeiros sistemas de TA, o sistema de TA baseado em regras consiste em utilizar regras linguísticas e gramaticais relacionadas às regularidades sintáticas, morfológicas e semânticas de dois idiomas. Com tais regras definidas a priori, o sistema deveria ser capaz de gerar, a partir de um conjunto de frases em um idioma, uma frase completa traduzida, respeitando as regras gramaticais do idioma alvo. A metodologia geralmente se aplica em três fases: análise, transferência e geração. A análise é realizada na língua original, determinando sua estrutura gramatical. Em seguida, a estrutura resultante é transformada em uma estrutura adequada para que seja gerada a frase na língua de destino com um ordenamento que respeite sua estrutura.

Em tradução livre dos seus nomes originais em inglês, as principais abordagens baseadas em regras são: Tradução Automática Direta; Tradução Automática mediante Transferência; e Tradução Automática mediante Língua Intermediária. A qualidade de cada uma se diferencia na profundidade da análise da língua original e a capacidade de representar o significado independente dos idiomas envolvidos. A abordagem Direta é a mais superficial, atuando principalmente no nível palavra por palavra com alguns ajustes gramaticais. Por não passar por uma fase intermediária (transferência), normalmente é

orientada para apenas um idioma de destino e, portanto, precisaria de maiores adaptações caso seja desejada a tradução para uma outra língua. No método interlíngua, o significado do texto original passa primeiramente por uma representação em uma língua "intermediária", tendo como vantagem maior capacidade de geração do texto traduzido em mais idiomas. Já o método mediante transferência difere da interlíngua em uma dependência maior das línguas envolvidas, tendo em vista que é realizada uma análise mais avançada da estrutura sintática da língua original e morfológica da língua de destino no momento de gerar a frase traduzida.

De maneira geral, os sistemas baseados em regras apresentam como limitações o desafio de encontrar uma língua intermediária que, de fato, possa representar a semântica de qualquer idioma. A própria representação do significado de uma frase em um idioma específico já não é uma tarefa trivial. Além disso, definir regras para cada etapa (análise, transferência e geração) do processo de tradução que possa ser aplicado para quaisquer idiomas geraria um trabalho manual significativo.

2.4.2.2 – Sistemas baseados em *corpus*

Usado como alternativa aos métodos tradicionais ou ainda como complemento, no caso dos métodos híbridos, os sistemas baseados em *corpus* alcançam cada vez mais relevância na era do *Big Data*, dado o crescimento exponencial da geração de dados, em particular os textuais, e da maior capacidade computacional de armazená-los e processá-los. Basicamente, o conhecimento gerado para esses sistemas é originado de *corpora* paralelos e bilíngues. Quanto maior a quantidade de exemplos de textos traduzidos de um idioma para o outro, maior a facilidade do sistema de adquirir o conhecimento necessário para tradução. Tal abordagem é dividida em duas sub-abordagens: Tradução Automática Estatística e Tradução Automática baseada em Exemplos.

Na tradução estatística, a premissa é de que com alguma probabilidade qualquer frase no idioma de destino pode ser a tradução de uma determinada frase no idioma de origem, sendo aquela com maior probabilidade a mais apropriada e escolhida pelo modelo. Apesar de alguns avanços, esses modelos também apresentam algumas limitações. O fato de cada idioma ter sua estrutura própria no ordenamento de sujeito, verbo, objeto e alguns classificadores impõe um desafio particular a esses algoritmos que,

no caso de traduções entre línguas asiáticas e europeias, por exemplo, não apresentam um desempenho satisfatório (OKPOR, 2014).

Nos sistemas de tradução baseados em exemplos, o modelo de tradução é por analogia, ou seja, a frase que se deseja traduzir é associada a frases similares fornecidas no *corpus* utilizado para treinar o modelo. Assim como o método estatístico, uma das principais vantagens dessa abordagem é prescindir da necessidade de configurar regras manualmente. Já em relação aos desafios, um dos principais é requerer maior eficiência computacional, sobretudo para bases muito grandes de textos, o que pode ser resolvido com técnicas de computação paralela.

CAPÍTULO 3

Implementação do Modelo

Para aplicação de técnicas de Aprendizado de Máquina na construção de um algoritmo de Tradução Automática, o estudo de caso foi estruturado na seguinte ordem: (1) raspagem de dados em um site de músicas traduzidas e consolidação dos dados textuais em um bloco de texto no formato de um *corpus* paralelo e bilíngue; (2) tratamento dos dados e vetorização para que possam ser usados em um modelo matemático; e, por fim, (3) aplicação de um modelo conhecido de RNA para o treinamento, validação e teste do algoritmo desenvolvido.

Para rodar a aplicação inteira, desde a leitura, passando pelo tratamento, até o treinamento, validação e teste do modelo de Redes Neurais, foi utilizado um *framework* disponível no site oficial do *Keras*. O trabalho foi desenvolvido originalmente para tradução de textos originalmente em inglês para o espanhol. Apesar de algumas adaptações terem sido necessárias para treinar os dados utilizados neste trabalho, os principais parâmetros utilizados originalmente foram mantidos, seguindo a recomendação dos autores para que um modelo de TA seja devidamente treinado.

3.1 – Ambiente de Desenvolvimento

Todos os experimentos foram executados em um *notebook* Asus F555L, com processador Intel Core TM i5, sistema operacional *Parrot OS*, 64 *bits* e 5GB de memória RAM. A programação de todas as etapas foi realizada na linguagem *Python* versão 3.9.13, em arquivos no formato *Jupyter notebook*. Para auxiliar na programação em diversas etapas, a linguagem conta com bibliotecas de código aberto, das quais as seguintes foram utilizadas:

- ***Numpy*:** biblioteca que permite a criação e manipulação de matrizes multidimensionais. É bastante usada para executar cálculos entre suas matrizes ou vetores, além de operações matemáticas diversas e de estatística básica em geral (NUMPY COMMUNITY, 2022).

- **Re**: do inglês *Regular Expressions*, este módulo fornece operações para correspondência de expressões regulares. *Regular expressions* também são frequentemente chamadas de *regex* (PYTHON SOFTWARE FOUNDATION, 2022). Esse módulo foi utilizado para que na leitura dos textos, todos os caracteres pudessem ser lidos como literais, inclusive os metacaracteres do *Python*. Metacaracteres são símbolos que podem vir dentro de arquivos de textos e que possuem um significado especial para a linguagem de programação. Por exemplo, o símbolo "\n" pode ser usado para indicar uma nova linha. No caso das frases utilizadas neste trabalho, entende-se que todos os caracteres foram originalmente escritos de forma literal, portanto é fundamental especificar isso para que o programa não interprete algumas sequências de caracteres como algum comando específico.
- ***String***: biblioteca usada para realizar operações com variáveis de texto. Esse módulo foi utilizado para filtrar e remover os caracteres de pontuação de todas as frases.
- ***Random***: módulo que implementa geradores de números pseudoaleatórios para várias distribuições. Para números inteiros, há uma seleção uniforme de um intervalo. Para sequências, há uma seleção uniforme de um elemento aleatório, uma função para gerar uma permutação aleatória de uma lista internamente e uma função para amostragem aleatória sem substituição (PYTHON SOFTWARE FOUNDATION, 2022). Neste trabalho, esta biblioteca foi utilizada para embaralhar o conjunto de frases.
- ***Selenium***: módulo designado para automatizar interações em navegador *web* com o *Python*, sendo capaz de acessar *websites*, clicar nos *links* disponíveis, localizar itens em uma página, preencher formulários ou qualquer tarefa que poderia ser executada normalmente através de um *mouse* e/ou teclado. Foi utilizado neste trabalho para pesquisar e acessar as páginas, bem como obter os textos original e traduzido de cada música encontrada.
- ***Time***: biblioteca com funções relacionadas à tempo. Tendo em vista a necessidade de acessar e carregar páginas na *web* com o *Selenium*, foi utilizada principalmente a função *sleep*, a fim de assegurar que a cada clique a página pudesse ter tempo suficiente para carregar, antes de seguir para a próxima interação.

- ***Webdriver Manager***: utilizada em conjunto com o *Selenium* para abrir uma janela do navegador Google Chrome e, em seguida, iniciar as interações necessárias para a coleta dos dados.
- ***Tensorflow***: biblioteca de código aberto para desenvolver e criar modelos de Aprendizado de Máquina.
- ***Keras***: é uma *Application Programming Interface* (API) desenvolvida em *Python* para criação de modelos de Aprendizado Profundo (do inglês *Deep Learning*) e é executada na plataforma *TensorFlow*. Seu objetivo é ser uma ferramenta simples, flexível e ao mesmo tempo poderosa (KERAS, 2022). Seu uso neste trabalho está voltado para a criação do modelo de RNA e configuração de sua arquitetura.

3.2 – Coleta de Dados

Para a estruturação de dados textuais, foi desenvolvido um *script* na linguagem de programação *Python* para automatizar a busca de versos de músicas originais em português e suas respectivas traduções em inglês. A figura 3.1 mostra a página de onde o algoritmo inicia o processo. Manualmente, foi selecionado o português como língua de origem e o inglês como o da versão traduzida, o que já é suficiente para gerar uma *URL* nova. A partir daí, o programa clica na opção *settings*, destacada na figura 3.1, e realiza o filtro de exibir apenas canções que a letra completa esteja em língua portuguesa, selecionando a opção *exact* e clicando, por fim, em *save* conforme a figura 3.2.

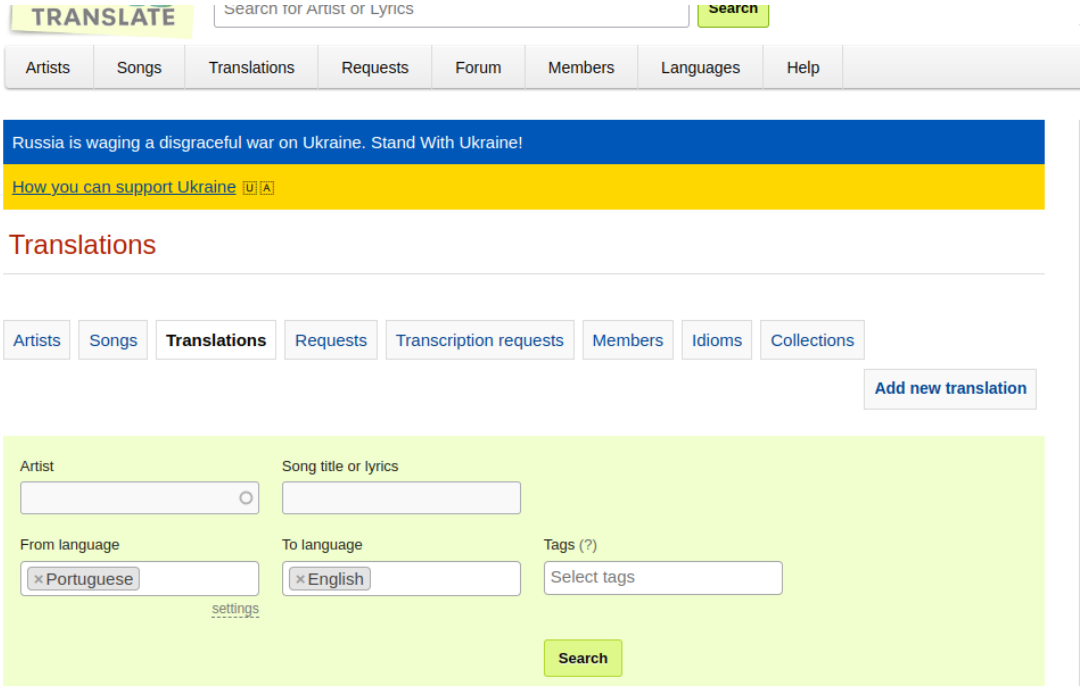


Figura 3.1 – Tela inicial do website utilizado na busca de letras, já com os idiomas selecionados.
Fonte: LyricsTranslate, 2022

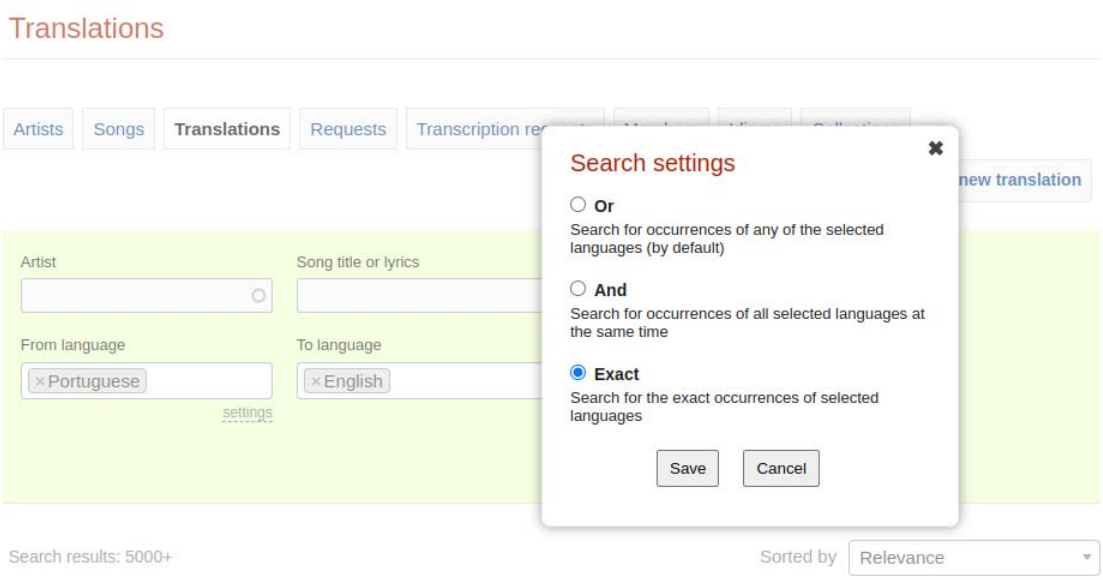
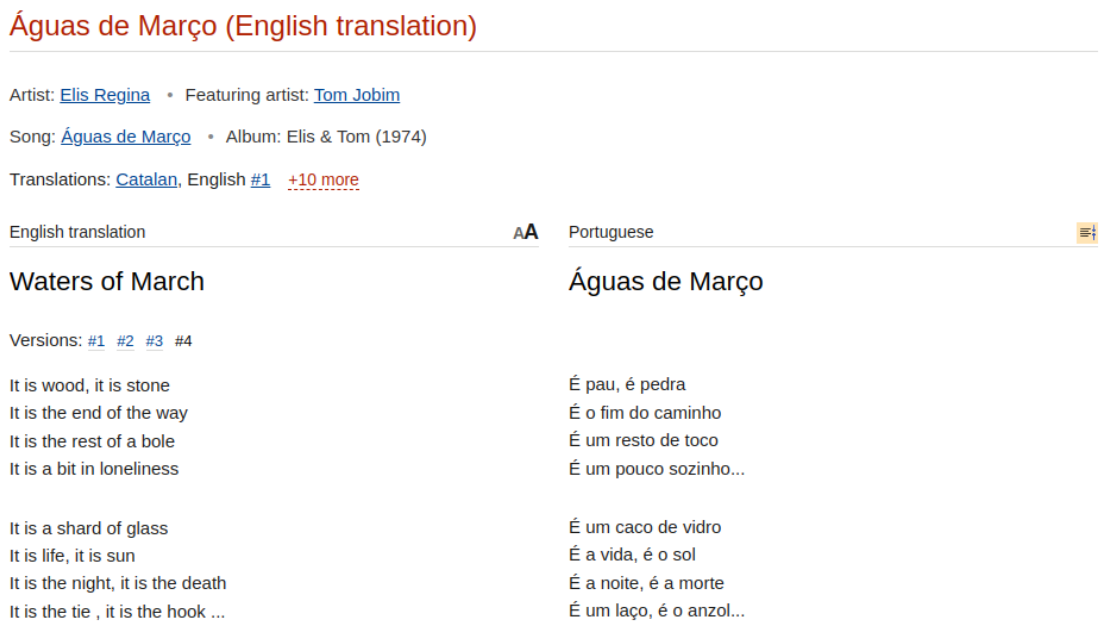


Figura 3.1 - Primeiro comando executado pelo código, filtrando músicas exclusivamente em português.
Fonte: LyricsTranslate, 2022

Utilizando apenas as músicas exibidas na primeira página da lista, o programa entra num ciclo de abrir a página de cada música e extrair os versos nos dois idiomas. A figura 3.3 mostra um exemplo de uma das letras obtidas. Alguns detalhes ainda foram necessários no código para ajustar a página e assegurar que as duas versões sejam exibidas

e os versos estejam alinhados. O processo foi repetido em todas as músicas e o resultado armazenado em um arquivo de texto.



Águas de Março (English translation)

Artist: Elis Regina • Featuring artist: Tom Jobim

Song: Águas de Março • Album: Elis & Tom (1974)

Translations: Catalan, English #1 +10 more

English translation	Portuguese
Waters of March	**Águas de Março**
Versions: #1 #2 #3 #4	
It is wood, it is stone	É pau, é pedra
It is the end of the way	É o fim do caminho
It is the rest of a bole	É um resto de toco
It is a bit in loneliness	É um pouco sozinho...
It is a shard of glass	É um caco de vidro
It is life, it is sun	É a vida, é o sol
It is the night, it is the death	É a noite, é a morte
It is the tie , it is the hook ...	É um laço, é o anzol...

Figura 3.2- Exemplo de uma das músicas com tradução disponível para o inglês
Fonte: LyricsTranslate, 2022

3.3 – Tratamento dos Dados

Para que os dados textuais obtidos possam ser utilizados de maneira eficaz em um modelo de aprendizado de máquina, é necessário seguir algumas etapas de limpeza, tratamento e pré-processamento, visando melhorar a qualidade dos dados, eliminar ruídos e assegurar que o modelo aprenda informações relevantes desses dados.

Após reunir todos os versos em português e inglês, foram removidos espaços duplicados, quebras de linhas e pontuações presentes em todo o conjunto de dados. Visando manter a naturalidade e evitar que o contexto pudesse se perder, as chamadas *Stop words* foram mantidas em ambos os idiomas. Para garantir uma maior confiabilidade, foram geradas amostras aleatórias de pares de versos, que foram conferidos manualmente se de fato o verso em inglês correspondia ao verso em português. Além da limpeza realizada, para cada frase no idioma de destino são adicionados *tokens* no início e no final, para que o modelo possa aprender onde começar e onde encerrar uma frase no texto traduzido.

Na etapa de pré-processamento, é realizada inicialmente a vetorização dos dados textuais. Utilizando duas instâncias da camada *TextVectorization* do *Keras*, uma para cada idioma, transformando cada verso em uma sequência de números inteiros. Para isso, é definido um vocabulário com todas as palavras presentes no conjunto de dados e, em seguida, a cada palavra é atribuído um número inteiro como índice. Esse processo é aplicado separadamente para cada um dos dois idiomas.

Nessa fase, as principais informações que são passadas ao algoritmo, com base na documentação oficial do *Keras Tensorflow*, são: (1) o tamanho do vocabulário, ou seja, a quantidade de palavras diferentes presentes no *corpus* utilizado; (2) a frase com o maior número de caracteres; (3) o tipo de normalização a ser aplicada.

A fim de padronizar os vetores numéricos representantes das frases, o verso de maior tamanho é usado como parâmetro. Desse modo, se a maior frase possui 20 palavras, por exemplo, todas as frases serão representadas por um vetor de dimensão igual a 20. No caso dos textos menores, os espaços restantes são preenchidos com "0" no final, tendo em vista que o modelo espera dos dados de entrada a mesma dimensão, independentemente do tamanho real da frase. A normalização pode ser realizada apenas transformando todas as letras em minúsculas, ou removendo pontuações, ou ambos, como é o caso deste trabalho. Por fim, é possível especificar ao algoritmo a necessidade de gerar bigramas, trigramas etc., para que o modelo entenda que alguns termos podem apresentar um significado específico quando em conjunto.

3.4 – Treinamento, Validação e Teste do Modelo

Na etapa seguinte, o conjunto de dados deve ser formatado de acordo com a arquitetura utilizada no modelo. Trata-se de um modelo *Seq2Seq* (*Sequence-to-Sequence*), uma arquitetura de rede neural projetada para lidar com tarefas de sequências. Nesse caso, o treinamento do algoritmo segue algumas etapas, e em cada uma delas o modelo tentará prever a próxima palavra a ser gerada no idioma de destino (dado de saída), com base nas palavras já previstas nas etapas anteriores e na frase completa do idioma de origem (dados de entrada).

Seguindo as especificações predefinidas no modelo desenvolvido pelo *Keras*, inicialmente as frases foram embaralhadas. Essa etapa é particularmente importante neste trabalho, tendo em vista que as frases vieram ordenadas por música e, portanto, os versos

de uma mesma música estavam sempre juntos. Após o reordenamento, 70% das frases foram alocadas ao conjunto de treino, como padrão em boa parte dos modelos de aprendizado de máquina, e o restante, ao conjunto de validação.

Para a construção do modelo, foi necessário configurar um codificador, um decodificador e um *Positional Embedding* (em tradução livre "incorporação posicional"). O codificador é responsável por processar a frase no idioma original e enviar uma nova representação matemática dessa sequência junto à sequência já prevista no idioma de destino até o momento para o decodificador, que irá prever as próximas palavras. Essa etapa é crítica em modelos de aprendizado de máquina, pois deve-se assegurar que o modelo tenha acesso apenas a dados que estarão disponíveis no momento da inferência, portanto o decodificador não pode ter acesso às palavras ainda não previstas durante o treinamento. Já o *Positional Embedding* é uma camada adicional responsável por fornecer ao modelo informações sobre a ordem e posição das palavras, capturando a relação sequencial delas na frase. A figura 3.4 mostra, resumidamente, a arquitetura *Seq2Seq* gerada, bem como a especificação de cada camada e suas dimensões.

```
Model: "transformer"
__________________________________________________________________________________________________
 Layer (type)                   Output Shape         Param #     Connected to
==================================================================================================
 encoder_inputs (InputLayer)    [(None, None)]       0           []

 positional_embedding (Position  (None, None, 256)   3845120     ['encoder_inputs[0][0]']
 alEmbedding)

 decoder_inputs (InputLayer)    [(None, None)]       0           []

 transformer_encoder (Transform  (None, None, 256)   3155456     ['positional_embedding[0][0]']
 erEncoder)

 model_1 (Functional)           (None, None, 15000)  12959640    ['decoder_inputs[0][0]',
                                                                  'transformer_encoder[0][0]']

==================================================================================================
Total params: 19,960,216
Trainable params: 19,960,216
Non-trainable params: 0
```

Figura 3.4 – Resumo da Arquitetura *Seq2Seq*
Fonte: elaborado pelo autor, 2022

Por fim, é necessário repassar como argumentos os atributos que possibilitem um treinamento eficiente e eficaz do modelo, dada a arquitetura já configurada e a separação dos dados já realizada. A tabela 3.1 mostra a especificação dos hiperparâmetros utilizados no programa. *Loss* é a função de perda/custo, a qual o modelo busca minimizar até o fim do treinamento. *Optimizer* é a função de otimização definida e que será responsável pelo ajuste dos pesos e taxa de aprendizado, assegurando a minimização da função *Loss*.

Batch_size (ou tamanho do lote) representa o número de amostras processadas em paralelo antes de uma atualização dos pesos. E *epochs* (ou épocas) representa o número de iterações durante o treino do modelo (KERAS, 2022).

Tabela 3.1 – Hiperparâmetros utilizados na configuração do modelo de Redes Neurais

Hiperparâmetro	**Configuração**
Loss function	*Sparse categorical crossentropy*
Optimizer	*RMS Prop*
Batch Size	64
Epochs	50

Fonte: Elaborado pelo autor, 2022

CAPÍTULO 4

Resultados Obtidos

Apesar de modelos de PLN contarem com métricas quantitativas de desempenho, dentre as quais as funções listadas no quadro 4.1, foi analisada principalmente a qualidade da tradução gerada pelo modelo, sua capacidade de reconhecer gírias ou expressões que tenham estado presentes nos dados de treino e avaliar possíveis causas dos principais erros de tradução. A baixa confiabilidade das traduções, bem como a diversidade de autores responsáveis pelas traduções, explica o menor foco nas métricas quantitativas propostas na literatura, tendo em vista que os dados não estavam necessariamente na condição ideal para o treinamento do modelo.

A figura 4.1 mostra o resultado numérico gerado após o treinamento, em particular o resultado da última iteração (*epoch*). Tendo em vista o número de *epochs* e a quantidade de frases utilizadas para treinamento, teste e validação (aproximadamente 55.000 frases), o processo de treinamento durou cerca de 12 horas, com acurácia de aproximadamente 76,4%.

```
Epoch 50/50
606/606 [==============================] - 931s 2s/step - loss: 0.8013 - accuracy: 0.7643 - val_loss: 1.8952 - v
al_accuracy: 0.4588
```

Figura 4.1 - Resultados numéricos do treinamento
Fonte: elaborado pelo autor, 2022

O quadro 4.1 mostra alguns exemplos de frases testadas para tradução. Iniciando pelas frases mais simples, com a ordem sujeito + verbo + objeto, percebe-se que o modelo foi capaz de aprender corretamente a traduzi-las, sem, necessariamente, atentar à conjugação mais adequada. No segundo exemplo, o tradutor unificou o verbo querer (em inglês, *want*) e a preposição "*to*", o que é bem particular da linguagem informal e não previsto na norma culta da língua inglesa. Na frase "Eu fico triste", o verbo foi traduzido com o sentido mais comum que apresenta no português, o de "permanecer", sendo que nesse caso o sentido está mais relacionado a uma mudança de estado/humor. Por fim, no último exemplo percebe-se a adição do advérbio "muito" na tradução da palavra "alegre", o que possivelmente ocorreu pela frequência com que as duas palavras podem ter aparecido juntas nos dados do treinamento.

Quadro 4.1 - Frases com a ordem Sujeito + verbo + objeto

Frase original (Português)	Tradução gerada (Inglês)
Eu amo ela	*I love her*
Eu quero saber seu nome	*I wanna know your name*
Eles são um casal	*They are a love*
Eles gostam de música	*They like music*
Eu fico triste	*I stay sad*
Ele não sabe sorrir	*He can't smile*
Ela está alegre	*She is very happy*

Fonte: elaborado pelo autor, 2022

Já no quadro 4.2 foram testadas algumas das mesmas frases, porém com a adição do advérbio de intensidade "muito" e o adjetivo "lindo", ambos sublinhados nos exemplos mostrados. É possível notar que em um dos casos (na frase "Eles gostam muito de música") o advérbio é simplesmente ignorado na tradução, sendo indiferente sua utilização ou não na frase original em português. Na frase "Eu quero muito saber seu nome", a intenção é enfatizar o verbo querer, enquanto na tradução gerada a ênfase é dada no verbo saber, com o sentido de "saber exatamente o nome da pessoa". Apesar de inúmeros exemplos utilizados no treinamento, faltou passar para o modelo de forma mais objetiva, sobretudo na vetorização dos dados de texto, algum componente que indicasse a alteração semântica que esses advérbios produzem no texto. Já no caso da frase com o adjetivo "lindo", o modelo aparentemente mudou o sentido também da palavra "casal".

Quadro 4.2 - Frases com adição de adjetivo e advérbio

Frase original (Português)	Tradução gerada (Inglês)
Eu amo muito ela	*I love her too much*
Eu quero muito saber seu nome	*I wanna know very well your name*
Eles são um lindo casal	*They are a beautiful I look*
Eles gostam muito de música	*They like music*
Eu fico muito triste	*I get much very sad*
Ela está muito alegre	*She is very very happy*

Fonte: elaborado pelo autor, 2022

Além do sentido produzido na tradução, foi verificado se de alguma forma o modelo aprendeu a diferença dos pronomes de acordo com a função exercida na oração. O quadro 4.3 demonstra que para a tradução isolada de cada pronome, o modelo não necessariamente aprendeu a forma no caso reto. No caso do pronome "eles", a tradução encontrada no inglês (*them*) corresponde ao pronome utilizado com a função de objeto, provavelmente por ter sido a forma mais frequente nos exemplos utilizados para o treinamento. Desse modo, apesar das variações de cada pronome de acordo com sua

função na frase, o modelo parece ter mapeado apenas uma palavra no idioma de destino para alguns pronomes. Apesar disso, no primeiro exemplo do Quadro 4.1, o pronome "ela" como objeto direto foi devidamente traduzido por "*her*", enquanto no caso em que foi traduzido isoladamente (Quadro 4.3), o modelo obteve a tradução esperada, utilizando o pronome no caso reto.

Quadro 4.3 - Tradução dos pronomes no caso reto

Frase original (Português)	Tradução gerada (Inglês)
Eu	*I*
Tu/Você	*You*
Ele/Ela	*He/She*
Nós	*We*
Vocês	*You*
Eles/Elas	*Them*

Fonte: elaborado pelo autor, 2022

Por fim, foi analisada a tradução de algumas frases mais coloquiais, com as expressões sublinhadas na oração original no Quadro 4.4. Na tabela são apresentadas as explicações de cada frase original em português, a tradução esperada de um ser humano e a tradução gerada pelo modelo. Apesar de em alguns poucos casos a tradução ter se aproximado da tradução esperada, trata-se de casos em que o equivalente em inglês também utiliza expressões parecidas, como no caso de "mar de rosas". Em outros casos como na expressão "Pelo amor de Deus", o sentido é mais intuitivo, portanto, a tradução não se afastou muito do esperado.

Quadro 4.4 - Frases com sentidos não literais

Frase original (Português)	Significado da expressão	Tradução esperada	Tradução gerada (Inglês)
Fala sério	Usado para expressar espanto/incredulidade com o que foi dito	*Come on*!	*Say that* (Literalmente *Diga isso*)
A vida é um mar de rosas	Vida sem adversidades, tranquila	*Life is a bed of roses.* (tradução similar, porém com a palavra "cama" ao invés de "mar"	*Life is a beautiful life*. (Literalmente *A vida é uma vida bela*)
Ele vai encher a cara	Ficar embriagado	*He will get drunk*	*He will walk the face*
Novo em folha	Muito novo	*Brand new*	*New*

			(traduzido sem nenhuma ênfase)
Ele saiu da linha	Comportar-se de maneira inconveniente, inesperada	*He stepped out of line*	*He has gone from the line* (Literalmente “Ele foi da linha”, no sentido de partir, sair andando
Caiu a ficha	Usado quando algo começa a fazer sentido	*To sink in* (frase no infinitivo)	*It was the* (frase traduzida sem um sentido completo)
Pelo amor de Deus	Expressão usada em diferentes contextos, seja indicando raiva, espanto etc.	*For God’s sake*	*For the love of God* (Nesse caso foi gerada a tradução exata de cada palavra da frase original)
Isso não é minha praia	Usado quando um sujeito não entende ou não gosta muito de algo	*It’s not my thing/my cup of tea*	*It is not my beach* (Nesse caso foi gerada a tradução literal)
Eu abro mão disso	Renunciar/desistir de algo	*I give up on it*.	*I open my hand.* (Nesse caso foi gerada a tradução literal)
Ele sabe de cor	Saber bem, sem precisar consultar nenhuma anotação	*He knows it by heart*.	*He knows of your own.* (Nesse caso, a tradução não somente perdeu o sentido original, como introduziu o pronome “teu” indevidamente

Fonte: elaborado pelo autor, 2022

CAPÍTULO 5

Conclusão e Trabalhos Futuros

O estudo realizado buscou aplicar algumas das técnicas e conceitos aprendidos durante o curso na análise de dados textuais e buscando otimizar o processo de tradução, cada vez mais presente no cotidiano social e profissional das pessoas no mundo inteiro. Assim como nos problemas de negócios em geral, torna-se evidente que os avanços tecnológicos e, em particular, da Inteligência Artificial são viáveis somente com a junção de especialistas de diversas disciplinas e áreas do conhecimento.

Conforme análise da evolução histórica e das limitações ainda presentes no campo da Tradução Automática, percebe-se que o real desafio do ser humano ao buscar automatizar e otimizar o tempo despendido em tarefas de análise é, na verdade, o desafio de registrar de forma objetiva todas as etapas e atividades, até os mínimos detalhes, envolvidas no processo. Conclui-se daí que o fluxo de atividades em um processo analítico e de tomada de decisão nem sempre é trivial ou perfeitamente definido, e em muitos casos não há previsões para todos os cenários possíveis, logo entende-se a importância de se utilizar de ferramentas estatísticas para trabalhar com as incertezas do mundo real.

Dentre as limitações do projeto, algumas poderiam ser resolvidas utilizando modelos híbridos de Tradução Automática. Como exemplo, a etapa de tratamento não contou com verificação e correção ortográfica, o que possivelmente confunde o algoritmo que pode vir a tratar uma mesma palavra como duas ou mais distintas ao lê-la com erros de digitação em algumas frases e escrita adequadamente em outras. Além disso, encontram-se disponíveis atualmente inúmeras ferramentas, inclusive no *Python*, para análise sintática e semântica de textos, sendo capazes de classificar devidamente cada termo de uma oração, bem como determinar sua função naquele texto em específico. Cabe destacar, ainda, que não foram utilizadas bibliotecas de PLN que pudessem ampliar o vocabulário do modelo, nem mesmo apresentar todas as formas em gênero, número e grau em que as palavras podem ocorrer.

Dado o volume de dados que foi necessário extrair, buscou-se minimizar também o trabalho com tratamento. Dessa forma, foi adotada a premissa de que cada verso isolado

possuía um sentido completo e poderia ser utilizado como exemplo para alimentar o modelo, o que nem sempre é o caso. Em muitas das vezes, é possível que um verso contenha algum pronome, por exemplo, que faça referência a algum agente mencionado em outro verso. Dessa forma, é sugerido que na fase de tratamento seja assegurado que cada frase possui um sentido completo em si própria, visto que o modelo reconhece cada uma como independente da outra.

Outro fator que comprometeu a etapa de coleta de dados é o fato de que o website utilizado para realizar as buscas por traduções de músicas não possui uma estrutura fixa em todas as páginas. Dessa forma, foram necessárias inúmeras adaptações no código desenvolvido para automatizar tais buscas de forma que, pelo menos, a maior parte das frases coletadas estivessem alinhadas, ou seja, para que no arquivo final o texto original estivesse bem ao lado de sua respectiva tradução.

Dentre as etapas citadas nos sistemas de TA, é importante ainda considerar algumas regras para a geração de frases para o usuário final, sobretudo quando se desenvolve um algoritmo para fins comerciais. Muitos já são os desafios em se gerar uma tradução que necessite de pouco ou nenhum ajuste humano, logo é importante que pelo menos as regras gramaticais e de concordância verbal estejam já ajustadas na resposta gerada pelo modelo para o usuário final, o que evidencia uma vantagem dos modelos híbridos.

Referências Bibliográficas

ALZEEBAREE, Yaseen. ***Machine Translation and Issues of Multiword Units: Idioms and Collocations.*** *Eastern Journal of Languages, Linguistics and Literatures (EJLLL)*, 1(2), p. 11-33, 2020.

BIRD, Steven; KLEIN, Ewan; LOPES, Edward. ***Natural Language Processing with Python - Analyzing Text with the Natural Language Toolkit..*** O'Reilly Media, Inc., 1ª Ed., 2009.

CHOPRA, Abhimanyu; PRASHAR, Abhinav; SAIN, Chandresh. ***Natural language processing***. *International Journal of Technology Enhancements and Emerging Engineering Research*, v.1, n.4, p. 131-134, 2013.

FACELI, Katti *et al*. **Inteligência Artificial: Uma Abordagem de Aprendizado de Máquina**. LTC, 2021.

HOCHREITER, Sepp; SCHMIDHUBER, Jurgen. ***Long Short-Term Memory***. *Neural Computation*, 9(8), p. 1735-1780, 1997.

HUTCHINS, John. ***The History of Machine Translation in a Nutshell***. 2014. Disponível em: < https://aclanthology.org/www.mt-archive.info/10/Hutchins-2014.pdf >. Acesso em 10 jan. 2022.

HUTCHINS, William John. ***Machine Translation: History and General Principles***. *The Encyclopedia of Languages and Linguistics***,** n. 5, p. 2322-2332, 1994.

KERAS. ***About Keras***. Disponível em: <https://keras.io/about/>. Acesso em: 01 jul. 2022.

KERAS. ***English-to-Spanish translation with a sequence-to-sequence Transformer***. Disponível em: < https://keras.io/examples/nlp/neural_machine_translation_with_transformer/>. Acesso em: 01 jul. 2022.

LEANDRO, Jhonatan Correa. **Aplicação de Redes Neurais LSTM para Previsão de Séries Temporais Financeiras.** Trabalho de Conclusão de Curso – Bacharel em Engenharia de Computação, Faculdade de Ciências Exatas e Tecnologia, Universidade Federal da Grande Dourados, Mato Grosso do Sul, 2021.

LEITE, Jan Edson Rodrigues. **Fundamentos de Linguística.** Graduação de Letras, UFMG, 2010. Disponível em

<https://grad.letras.ufmg.br/arquivos/monitoria/LEITE_2010.pdf>. Acesso em: 10 jan. 2022.

LIDDY, Elizabeth D. ***Natural Language Processing***. *Encyclopedia of Library and Information Science*, 2ª Ed. NY. Marcel Decker, Inc., 2001.

MANNING, Christopher D.; SCHUTZE, Hinrich. ***Foundations of Statistical Natural Language Processing*.** Cambridge, MA, USA: The MIT Press, 1999.

MELO, Sheila de Souza Corrêa de. **Tradução Automática e Competência Tradutória: Repensando Interseções**. Rónai: Revista de Estudos Clássicos e Tradutórios**,** p. 60-72, 2013.

MÜLLER, Andreas C.; GUIDO, Sarah. ***Introduction to Machine Learning with Python: A Guide for Data Scientists.*** 1. ed. [S.l.]. O'Reilly Media, 2016.

NUMPY COMMUNITY. ***NumPy User Guide***, Release 1.23.0, 2022**.** Disponível em: https://numpy.org/doc/stable/numpy-user.pdf. Acesso em 01 jul. 2022.

OKPOR, Margaret Dumebi. ***Machine Translation Approaches: Issues and Challenges***, *IJCSI International Journal of Computer Science Issues*, Vol. 11, n. 5, 2014.

PIRES, Letícia Florentino. **Modelo de LSTM Aplicado na Previsão das Séries de Radiação Solar de Burkina Faso.** Trabalho de Conclusão de Curso – Bacharel em Ciência da Computação, Instituto de Ciências Exatas, Universidade Federal de Juiz de Fora, Minas Gerais, 2019.

PROVOST, Foster; FAWCETT, Tom. ***Data Science for Business*: *What you need to know about data mining and data-analytic thinking.*** O'Reilly Media, Inc., 1ª Ed., 2013.

PYTHON SOFTWARE FOUNDATION. ***Random – Generate Pseudo-Random Numbers*.** Disponível em: <https://docs.python.org/3/library/random.html>. Acesso em 01. jul 2022.

PYTHON SOFTWARE FOUNDATION. ***Re – Regular Expression Operations*.** Disponível em: <https://docs.python.org/3/library/re.html#>. Acesso em 01 jul. 2022.

RUSSELL, Stuart; NORVIG, Peter. ***Artificial Intelligence: A Modern Approach*.** *Prentice Hall*, 3ª Ed., 2009.

XATARA, Cláudia Maria. **O Ensino do Léxico: As Expressões Idiomáticas**, Trab. Ling. Apl., Campinas**,** (37):49-59, Jan./Jun. 2001.