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

TRADUÇÃO AUTOMÁTICA NEURAL INGLÊS-PORTUGUÊS NO DOMÍNIO DO E-COMMERCE

Lucas Hochleitner da Silva

Orientadora: Profa. Dra. Helena de Medeiros Caseli

São Carlos - SP
17 de junho de 2020

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

TRADUÇÃO AUTOMÁTICA NEURAL INGLÊS-PORTUGUÊS NO DOMÍNIO DO E-COMMERCE

Lucas Hochleitner da Silva

Monografia apresentada ao Curso de Graduação em Ciência da Computação da Universidade Federal de São Carlos, para a obtenção do título de bacharel em Ciência da Computação.
Orientadora: Profa. Dra. Helena de Medeiros Caseli

São Carlos - SP
17 de junho de 2020

Lucas Hochleitner da Silva

TRADUÇÃO AUTOMÁTICA NEURAL INGLÊS-PORTUGUÊS NO DOMÍNIO DO E-COMMERCE

Monografia apresentada ao Curso de Graduação em Ciência da Computação da Universidade Federal de São Carlos, para a obtenção do título de bacharel em Ciência da Computação.

Trabalho aprovado. São Carlos - SP, 17 de junho de 2020:

Profa. Dra. Helena de Medeiros Caseli
DC - Universidade Federal de São Carlos

Prof. Dr. Daniel Lucrédio
DC - Universidade Federal de São Carlos

Prof. Dr. Murilo Coelho Naldi
DC - Universidade Federal de São Carlos

São Carlos - SP
17 de junho de 2020

RESUMO

SILVA, L. H. *Tradução Automática Neural inglês-português no domínio do E-Commerce*. 62 p. Monografia (Graduação em Ciência da Computação) — Universidade Federal de São Carlos, 2019.

A Tradução Automática (TA) é uma das subáreas/aplicações mais antigas e conhecidas do Processamento de Língua Natural (PLN). Após mais de 80 anos de pesquisas nessa área, e mesmo com diversos avanços alcançados principalmente nos últimos 40 anos, a qualidade das traduções automáticas geradas para textos em alguns domínios ainda deixa bastante a desejar. Entre as diversas abordagens propostas para a tradução automática, como a baseada em regras e a estatística, neste trabalho investigou-se o uso da abordagem considerada o estado-da-arte na TA: a tradução automática neural (*Neural Machine Translation* – NMT). Na NMT, redes neurais são treinadas para gerar modelos de tradução capazes de converter uma sentença de entrada em um idioma, em uma sentença de saída equivalente em outro idioma, sem que seja necessário lidar com alinhamento de palavras ou regras de tradução, como em outras abordagens. Este trabalho, portanto, tem como objetivo verificar como os métodos neurais se comportam em um domínio específico: o domínio do *e-commerce*. Diversas peculiaridades desse domínio trazem desafios para as pesquisas em TA como: (i) a falta de estrutura dos títulos dos produtos a venda, (ii) a especificidade de vocabulário para cada uma das várias categorias de produtos a venda e (iii) a grande variabilidade de vocabulário considerando-se todas as categorias de produtos a venda. Assim, a escolha pela investigação da NMT no domínio do *e-commerce* se deu em razão de: (i) sua pouco exploração em PLN, (ii) suas particularidades quando comparado a outros domínios e (iii) à demanda de uma parceria (projeto de extensão) com a empresa B2W Digital[1].

Palavras-chave: Tradução automática neural. E-commerce. Domínio específico. Adaptação de domínio. Português do Brasil.

[1] Disponível em: <https://ri.b2w.digital/> Acesso em: 07 nov. 2019.

ABSTRACT

Machine Translation (MT) is one of the oldest and most known subareas/applications of Natural Language Processing (NLP). After more than 80 years of research in this area, and despite the many advances made mainly in the last 40 years, the quality of the automatically translated texts in some domains is still far from the expected one. Among the various approaches proposed for machine translation, such as the rule-based MT and the statistical MT, this work investigated the use of the state-of-the-art approach in MT: the Neural Machine Translation (NMT). In NMT, neural networks are used to generate translation models capable of converting an input sentence in one language into an equivalent sentence in another language, without having to deal with word alignment or translation rules, as in other approaches. This work aims to verify how neural methods behave in a specific domain: the e-commerce domain. Several peculiarities of this domain pose challenges for MT research, such as: (i) the lack of structure in the titles of the products for sale, (ii) the vocabulary specificity for each of the various categories of products for sale and (iii) the wide vocabulary variability considering all categories of products for sale. Thus, the choice for the research with NMT applied to the e-commerce domain was due to: (i) the lack of works in NLP regaring this domain, (ii) its particularities when compared to other domains and (iii) the demand of a partnership with the company B2W Digital.

Keywords: Neural machine translation. E-commerce. Specific Domain. Domain Adaptation. Brazilian Portuguese.

SUMÁRIO

1 INTRODUÇÃO

A Tradução Automática (TA) é uma aplicação (e subárea) muito importante do Processamento de Línguas Naturais (PLN). O objetivo da TA é gerar uma versão equivalente, em um determinado idioma alvo, de um texto fornecido como entrada em um idioma fonte.

Na versão mais simples desse processo (a tradução direta), a TA consiste da substituição de palavras do idioma fonte para o idioma alvo. Contudo, apenas essa substituição palavra a palavra não é suficiente para a produção de uma boa tradução. Para gerar uma tradução que seja adequada e fluente, são necessários outros processos como a interpretação e a análise dos elementos no texto para saber como cada palavra pode influenciar as outras.

Desde o seu surgimento, diferentes abordagens foram propostas para a TA. A abordagem linguística (também conhecida como abordagem simbólica) foi a mais investigada e empregada nos primórdios das pesquisas em TA. Nela, o conhecimento linguístico profundo nas línguas fonte e alvo é mapeado, geralmente na forma de regras, no que se conhece como TA baseada em regras (*Rule-based Machine Translation* – RBMT)[1]. Ao contrário das abordagens seguintes, a RBMT leva em consideração mais informação linguística específica dos idiomas fonte e alvo, usando regras morfossintáticas e sintáticas, e possivelmente análises semânticas entre os idiomas. Os primeiros sistemas RMBT datam do começo dos anos 1970.

A partir de 1989, abordagens empíricas (também conhecidas como abordagens baseadas em córpus) passaram a ser as mais investigadas. Métodos dessa abordagem se baseiam na suposição de que soluções para a TA podem ser encontradas por meio do processamento de textos já existentes traduzidos por profissionais linguistas (GUIDÈRE, 2002). Usando um córpus paralelo bilíngue[2], os sistemas de TA são capazes de aprender a gerar sentenças no idioma alvo equivalentes às sentenças no idioma fonte. A abordagem empírica é usada até os dias de hoje englobando métodos estatísticos e neurais, considerados o estado da arte na atualidade.

A tradução automática estatística (*Statistical Machine Translation* – SMT) gera modelos que se fundamentam na detecção de padrões dos textos previamente traduzidos por humanos e usam estatística para inferir traduções baseadas nos padrões encontrados. Nessa abordagem, tradicionalmente são gerados dois modelos: um modelo de tradução e um modelo de língua. O primeiro captura o paralelismo entre as línguas fonte e alvo, prezando pela adequação da tradução; enquanto o segundo aprende as estruturas e padrões da língua alvo na tentativa de garantir a fluência da sentença gerada como saída. Exemplos de sistemas de tradução automática estatística não descritos nos trabalhos de Koehn et al. (2007) e Och e Ney (2004).

[1] Um exemplo e uma das primeiras aplicações de RBMT pode ser vista no sistema SYSTRAN. Uma visão detalhada do sistema se encontra no trabalho de Dugast, Senellart e Koehn (2007).

[2] Um córpus paralelo bilíngue é uma coleção de pares de textos escritos em dois idiomas, sendo um texto a tradução do outro.

A SMT foi considerada o estado da arte até 2016, quando a tradução automática neural a sucedeu.

As principais limitações da abordagem estatística englobam a dificuldade de edição do conhecimento representado nos modelos e a dependência em relação ao córpus de treinamento, bem como a incapacidade de generalização dos aspectos estruturais e sintáticos da língua. Em contrapartida, a TA neural surgiu com a proposta de capturar esses aspectos estruturais.

Na TA neural (*Neural Machine Translation* – NMT), redes neurais artificiais são usadas para treinar modelos que capturam as características presentes nos dados de treinamento que são importantes para a realização da TA, como: a morfologia das palavras, suas frequências e contextos de ocorrência. Essas características são o que norteiam o mapeamento para gerar a saída apropriada no idioma alvo, dada uma entrada no idioma fonte. Geralmente, as redes são capazes de mapear sentenças completas em um único modelo integrado.

A NMT permite o treinamento de ponta a ponta (*end-to-end*) de um sistema de tradução sem a necessidade de lidar com alinhamento de palavras, regras de tradução ou algoritmos avançados de decodificação, que são comumente requeridos por métodos estatísticos (KOEHN et al., 2007). O desempenho da tradução automática neural é classificado como o estado da arte em cenários onde há um grande conjunto de recursos para o treinamento (BOJAR RAJEN CHATTERJEE; TURCHI, 2017; NAKAZAWA SHOHEI HIGASHIYAMA; KUROHASHI, 2017). Contudo, reunir um córpus de tamanho suficiente, que seja paralelo e de qualidade, é um grande desafio dependendo do domínio e dos idiomas escolhidos. Experimentos anteriores mostraram que a NMT não obteve um desempenho desejável em cenários onde o conjunto de recursos para treinamento era insuficiente (DUH GRAHAM NEUBIG; TSUKADA, 2013; SENNRICH; ARANSA, 2013; ZOPH DENIZ YURET; KNIGHT, 2016; KOEHN; KNOWLES, 2017).

Nesse cenário de escassez de dados para treinamento, é necessário fazer adaptações para que o modelo neural possa ser aplicado. Um exemplo onde a quantidade de córpus de treinamento pode ser impactante na qualidade de tradução gerada pelo modelo treinado é o da tradução em domínios especializados como a biologia.

Neste projeto objetiva-se investigar a tradução neural aplicada a um domínio específico: o domínio do *e-commerce*. Assim, o foco será na tradução de títulos de produtos a venda em sites da internet, como melhor contextualizado nas próximas seções.

1.1 Motivação

O *e-commerce* é um dos mercados que mais cresce no Brasil e no mundo. Segundo dados divulgados em Agosto/2019 pela Ebit/Nielsen[3], o mercado varejista *online* (*e-commerce*)

[3] Disponível em: <https://www.ecommercebrasil.com.br/noticias/e-commerce-cresce-12-por-cento-webshoppers-i-e-commerce-brasil/>. Acesso em: 28 nov. 2019.

teve um crescimento de 12% no primeiro semestre de 2019. Ainda segundo o estudo divulgado pela Ebit/Nilsen, esse crescimento representa um faturamente de R$ 26,4 bilhões nos últimos seis meses.

A esse crescimento do mercado varejista no Brasil somam-se as vendas de produtos de fornecedores de outros países. Essa nova demanda de comércio *online* multilíngue traz a necessidade de traduzir os títulos e descrições dos produtos ofertados por fornecedores de outros países para o português do Brasil. Contudo, os textos tradicionalmente usados nas páginas de *e-commerce* apresentam peculiaridades que dificultam a tradução. Para ilustrar esse fato, considere os exemplos apresentados na Figura 1.

Figura 1 – Exemplos de sentenças presentes em sites de *e-commerce*. Compara-se a sentença no idioma original, a tradução contida no site, a tradução feita pelo Google Tradutor[a], e a tradução referência gerada pelo autor deste trabalho

Original	*Fashion Men's Winter Warm Overcoat Wool Coat Trench Coat Outwear Long Jacket New.*
Site	Moda masculina de inverno sobretudo quente roupa De Lã Casaco Trench Casaco Jaqueta Longa Nova.
Google Tradutor	Inverno dos homens da moda Casaco Quente de Lã Casaco Trench Coat Outwear Jaqueta Longa Novo.
Tradução Livre	Casaco de inverno masculino da moda Sobretudo quente Casaco de lã Trench coat Jaqueta longa Novo.

Original	*8mm Men Tungsten Carbide Rings Hawaiian Koa Wood Abalone Shell Jewelry Size 7-12.*
Site	Homens 8mm Anéis de Carboneto de tungstênio Havaiana Koa Abalone Shell Joias De Madeira Tamanho 7-12.
Google Tradutor	8mm Homens Anéis De Carboneto De Tungstênio Havaiano Koa Madeira Abalone Shell Jóias Tamanho 7-12.
Tradução Livre	Anéis masculinos de carboneto de tungstênio de 8mm Madeira koa havaiana Concha abalone Joias tamanho 7-12.

Original	*Premium Women's Stretch Ponte Pants - Dressy Leggings - Wear to Work - All Day Comfort*
Site	Calça Ponte Feminina Premium – Leggings Elegantes – Use para Trabalhar – Conforto o dia todo.
Google Tradutor	Calças ponte estiramento para mulher premium - perneiras vistosas - desgaste do trabalho - conforto durante todo o dia.
Tradução Livre	Calça ponte premium feminina - Leggings elegantes - Vista para trabalhar - Conforto o dia todo

[a] Disponível em: <https://translate.google.com/> Acesso em: 07 nov. 2019.

Como é possível notar na Figura 1, erros de tradução podem ser facilmente encontrados nos sites de *e-commerce*. Os exemplos dessa figura também evidenciam os erros induzidos pela falta de estrutura formal nas sentenças originais. A ausência de pontuação, em conjunto com o

emprego de múltiplas descrições do produto – que se assemelham a *tags* – dificultam ainda mais a tradução feita por modelos treinados a partir de estruturas mais formais (geralmente textos jornalísticos). Essa formação incomum da sentença é mais evidente nos dois primeiros exemplos da figura, onde, mesmo no idioma original, a semântica da sentença é prejudicada. O terceiro exemplo demonstra que o mesmo sistema de tradução, se aplicado a uma sentença melhor estruturada, pode produzir bons resultados. Contudo, a análise feita nos sites propostos, aliada ao córpus usado no projeto, demonstram que esse tipo de sentença não é representativa.

Assim, entre as particularidades do domínio de *e-commerce* investigado neste trabalho, destacam-se: (i) a falta de estrutura dos títulos dos produtos, (ii) a especificidade de vocabulário para cada uma das várias categorias de produtos a venda e (iii) a grande variabilidade de vocabulário considerando-se todas as categorias de produtos a venda. Essas particularidades induzem a alguns erros na tradução, como: (i) a quebra de sentido da frase como um todo, (ii) a não tradução de algumas palavras e (iii) a eliminação de relações semânticas entre algumas das palavras. Exemplos dessa variedade textual são apresentados nos trechos retirados do córpus de referência usado neste trabalho, presentes na Figura 1.

Considerando as peculiaridades descritas, o treinamento para o domínio do *e-commerce* se torna especialmente desafiador, e configura uma área que demanda pesquisas direcionadas. Assim, a escolha pela investigação da NMT no domínio do *e-commerce* se deu em razão de: (i) sua pouco exploração em PLN, (ii) suas particularidades quando comparado a outros domínios e (iii) a demanda de uma parceria (projeto de extensão) com a empresa B2W Digital[4].

1.2 Objetivo e Hipótese

Neste contexto, este trabalho tem como **objetivo**:

"Investigar a tradução automática neural aplicada a um domínio específico, de forma a melhorar a qualidade da referida tradução nesse domínio."

Para tal, investigou-se a abordagem apresentada por Artetxe et al. (2018), que se utiliza de um conjunto de técnicas e métodos propostos para tentar melhorar o desempenho de tradutores em domínios específicos.

Uma das técnicas presentes no trabalho citado é a Back-translation (BT) (SENNRICH; BIRCH, 2016; PONCELAS DIMITAR SHTERIONOV; PASSBAN, 2018). A BT faz uso de um córpus denominado *sintético*, formado a partir da tradução reversa da linguagem alvo de volta para a original. No cenário deste trabalho, a tradução inglês-português é a natural, e o seu uso de volta ao inglês compõe o córpus sintético. Essa estratégia, então, reúne o córpus com a tradução natural ao córpus sintético, formando uma mescla de textos genuinamente paralelos e pares de traduções sintéticas.

[4] Disponível em: <https://ri.b2w.digital/> Acesso em: 07 nov. 2019.

O trabalho de Artetxe et al. (2018) também usa modelos de semântica distribucional por meio de *word embeddings* (WE). Os modelos de semântica distribucional (*Distributional Semantic Models*, ou DSMs), também conhecidos como modelos de espaço semântico ou modelos de espaço vetorial (VSM, do inglês *Vector Space Model*), se baseiam na *hipótese distribucional* proposta por Harris (1954), que estabelece que palavras que ocorrem em contextos similares tendem a ter proximidade semântica entre si. Essa ideia foi primeiramente proposta baseando-se na afirmação de que uma palavra é caracterizada pela companhia que possui (FIRTH, 1957). Esse tipo de medida fundamenta-se na conjectura de que o contexto associado a uma palavra pode ser caracterizado pelas palavras que a cercam. Portanto, contextos similares tendem a denotar palavras semanticamente similares. Dessa convicção surgem as *word embeddings*, que, essencialmente, são representações matemáticas de palavras através de um vetor numérico denso em um espaço vetorial. Segundo Lenci (2008), os vetores são as estruturas de dados mais úteis para formalizar as representações contextuais, uma vez que a sequência de números que compõe um vetor codifica a força de associação estatística entre uma palavra e um certo contexto.

Essas duas técnicas propostas por Artetxe et al. (2018), unidas à redes neurais artificiais, permitem que se crie um tradutor automático neural que possa ser treinado de forma não-supervisionada e com um córpus de volume não tão expressivo. Um tradutor neural formado por essas características se torna desejável quando se trabalha com um domínio como o do *e-commerce*.

Dessa forma, este trabalho tem como **hipótese** a de que as estratégias propostas por Artetxe et al. (2018) podem ser aplicadas para melhorar a qualidade da tradução automática neural de inglês para português, no domínio do *e-commerce*, quando comparada à tradução gerada por um modelo tradicional de domínio geral.

1.3 Organização da monografia

Os próximos capítulos estão organizados como segue. No capítulo 2 é apresentada a fundamentação teórica e os conceitos essenciais para o entendimento deste trabalho. No capítulo 3 são apresentados os trabalhos relacionados a este que visam a tradução automática tanto para domínios gerais como para específicos. No capítulo 4, os materiais utilizados neste trabalho são descritos. No capítulo 5 apresenta-se os experimentos realizados e a análise dos resultados obtidos. Por fim, o capítulo 6 traz as considerações finais, contribuições e possíveis extensões deste trabalho.

2 FUNDAMENTAÇÃO TEÓRICA

A tradução automática neural é considerada, atualmente, o estado da arte. Isso se dá, principalmente, em razão do recente surgimento de novas técnicas (BAHDANAU; CHO; BENGIO, 2014) e também de como um conjunto de dados massivo se tornou acessível, o que era de mais difícil acesso há poucos anos (SKOROKHODOV et al., 2018). Apesar da abundância de conjuntos de dados paralelos para as linguagens e domínios mais populares, há ainda falta de recursos para vários outros domínios e pares de idiomas.

Na TA, um treinamento é considerado supervisionado se o conjunto de dados utilizado no treinamento traz informação de como a tradução é realizada, por exemplo, por meio do uso de um córpus paralelo. Como contraparte, se o conjunto de dados não possui qualquer informação que indique como a tarefa deve ser realizada, o treinamento é considerado não-supervisionado, como no uso exclusivo de córpus monolíngue não paralelo. Ao se utilizar um conjunto de dados que seja apenas parcialmente enriquecido com informação de como executar a tarefa, tem-se um treinamento semi-supervisionado. A estrutura da tradução automática neural, descrita a seguir, permite um treinamento que seja mais flexível quanto ao uso de dados não paralelos se comparada a sistemas RBMT e SMT. Como consequência disso, existe um interesse crescente em técnicas que permitam que o treinamento seja feito de forma semi-supervisionada ou mesmo completamente não-supervisionada.

2.1 Tradução Automática Neural

A tradução automática neural (NMT) (CHO et al., 2014b; SUTSKEVER; VINYALS; LE, 2014; BAHDANAU; CHO; BENGIO, 2014) é uma abordagem proposta recentemente e que tem demonstrado vários avanços para a TA, especialmente em termos de avaliação humana, se comparada com métodos RBMT e SMT (WU et al., 2016). Em sua essência, a NMT usa um conjunto de frases traduzidas anteriormente por seres humanos para estabelecer links – como concordância de gênero e número – entre palavras a nível de frase e, em seguida, aprende como traduzir frases de estrutura semelhante no futuro. Originalmente, foi desenvolvida usando modelos de tradução sequência para sequência (*sequence-to-sequence*) (SUTSKEVER; VINYALS; LE, 2014; CHO et al., 2014b), e depois melhorada através de variantes baseadas em atenção (BAHDANAU; CHO; BENGIO, 2014; LUONG; PHAM; MANNING, 2015), como explicado a seguir.

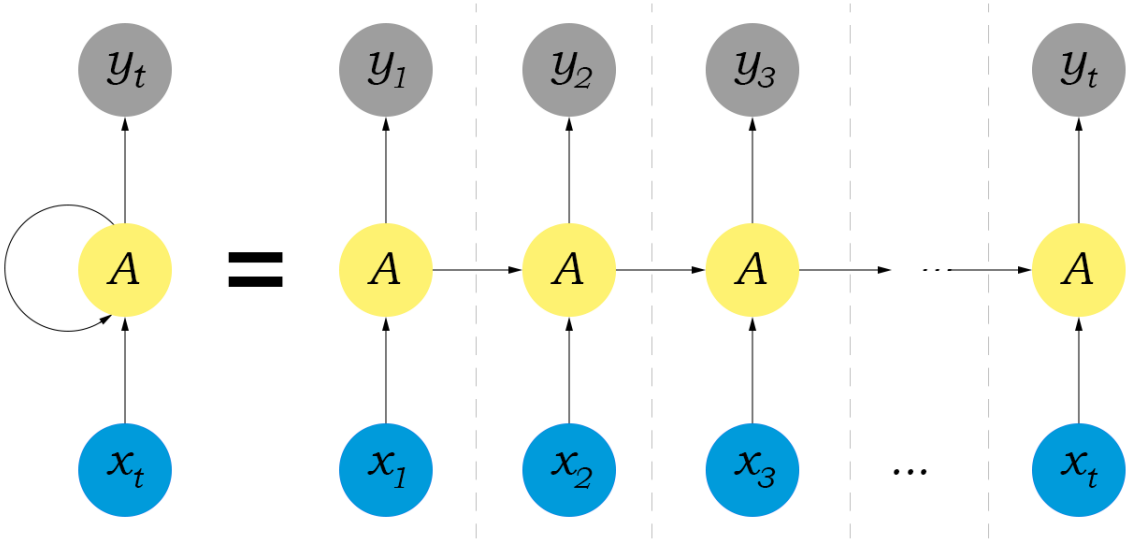
A estrutura tradicional da NMT usa duas redes neurais recorrentes (RNN): uma como *encoder* e a outra como *decoder*. A rede neural *encoder* lê a sentença de entrada no idioma fonte e a codifica em um vetor de tamanho fixo, também conhecido como vetor de contexto.

Já o *decoder* decodifica esse vetor e tem como saída a sentença traduzida no idioma alvo. Esse sistema *encoder-decoder*, para um par de idiomas, é treinado em conjunto a fim de maximizar a probabilidade de uma tradução correta dada uma sentença de entrada (CHO et al., 2014a).

A RNN é um tipo de rede neural artificial que permite uma conexão temporal entre seus nós. Ao contrário de redes neurais tradicionais, ela possibilita que a entrada seja processada de forma sequencial, e não apenas como um vetor de tamanho fixo. Outra diferença é que esse tipo de rede faz com que a informação dada como entrada possa persistir, ao contrário das redes tradicionais, onde a cada nova entrada a informação anterior é perdida. As redes neurais recorrentes surgiram a fim de evitar esses problemas e, para isso, sua estrutura é formada por *loops* que as permitem aprender dependências de longo prazo.

O diagrama da Figura 2 apresenta uma abstração do funcionamento em *loop* de uma RNN. À esquerda, a estrutura da rede demonstra a entrada x_t ligada à rede recorrente A, com a saída y_t. O *loop* garante que a informação dada em um passo t seja passada para o seguinte. Considerando a perspectiva temporal, variando a entrada segundo t, o funcionamento da rede torna-se o que é apresentado na Figura 2, à direita, onde a informação na rede A, no tempo t, é carregada ao processamento da rede no tempo seguinte, de forma sequencial.

Figura 2 – Diagrama de funcionamento de uma RNN



Fonte: Adaptação (Página *Deep Learning Book*[a])

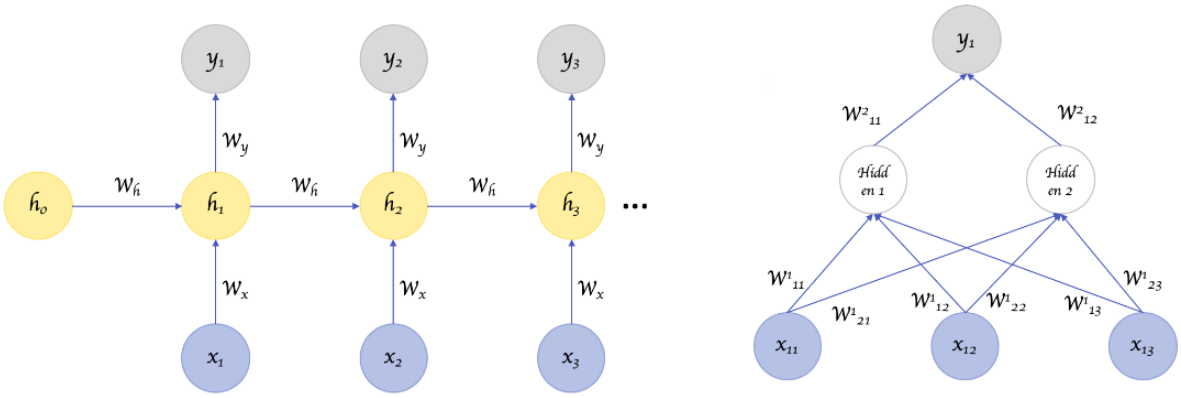
[a] Disponível em:<http://deeplearningbook.com.br/arquitetura-de-redes-neurais-long-short-term-memory/>. Acesso em: 15 de novembro de 2019.

Dessa forma, a RNN é capaz de se lembrar de entradas passadas, e isso é o que torna esse tipo de rede sequencial: cada entrada é processada de forma a ser influenciada por entradas passadas e influencia saídas futuras. Portanto, uma mesma entrada pode produzir

saídas diferentes dependendo de entradas anteriores na sequência. Essa característica a torna especialmente atrativa para aplicações de PLN, que processam a entrada – como texto, áudio e escrita – de forma sequencial.

Uma comparação entre as redes neurais tradicionais e as RNN pode ser observada na Figura 3. À direita, a rede neural tradicional tem como entrada um vetor x_1, de tamanho fixo (nesse caso, 3), e como saída o vetor y_1, também de tamanho fixo (nesse caso, 1). À esquerda tem-se uma RNN, com os vetores x_1, x_2 e x_3 como entradas sequenciais, e as saídas y_1, y_2 e y_3 como as saídas respectivas das entradas (seus tamanhos irrelevantes). Como é possível observar, na RNN, a saída y_2 (da entrada x_2) é influenciada por h_1, um nó que recebeu valores da entrada x_1. Analogamente, o mesmo ocorre com y_3, sendo influenciado por h_2 e h_1, das entradas anteriores. E esse fluxo se repete indefinidamente, independente do tamanho da entrada, diferente da rede tradicional, à direita, onde os vetores de entrada x e saída y sempre serão de tamanho fixo. É essa característica que torna a RNN especialmente interessante quando se trabalha com tradução.

Figura 3 – Comparação de uma RNN e uma rede neural tradicional. À esquerda uma RNN. À direita uma rede neural tradicional



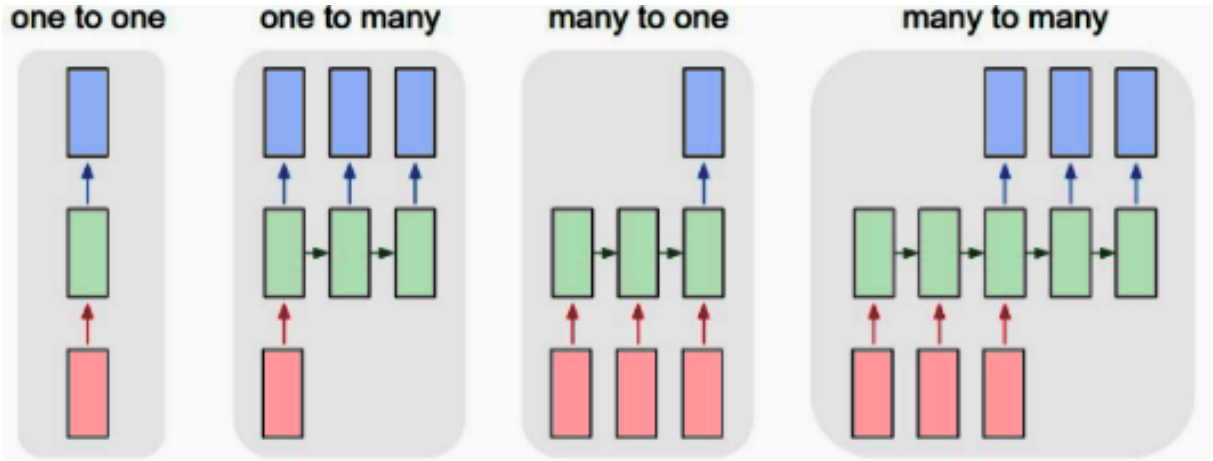
Fonte: Towards Data Science[a]

[a] Disponível em:<https://towardsdatascience.com/recurrent-neural-networks-d4642c9bc7ce>. Acesso em: 15 de novembro de 2019.

A capacidade de uma RNN de processar suas entradas de forma sequencial ainda permite configurações diversas de entrada e saída, não só tendo uma saída para cada entrada, mas podendo ter várias. Algumas são apresentadas na Figura 4, sendo, da esquerda para a direita: (i) um para um, como uma rede neural tradicional, usada em tarefas como a classificação de imagens; (ii) um para muitos, tendo uma entrada de tamanho fixo e uma saída em sequência, usada em geração de legendas de imagens – que, a partir de uma imagem única, exibe uma sequência de palavras como saída; (iii) muitos para um, tendo uma sequência como entrada e uma saída de tamanho fixo, empregada em análise de sentimento – a partir de uma sequência de palavras, classifica como sentimento negativo ou positivo; e (iv) muitos para muitos, tendo

uma sequência tanto como entrada como saída, usada em TA, sendo a entrada e a saída sequências de palavras.

Figura 4 – Algumas variações de entrada e saída de uma RNN



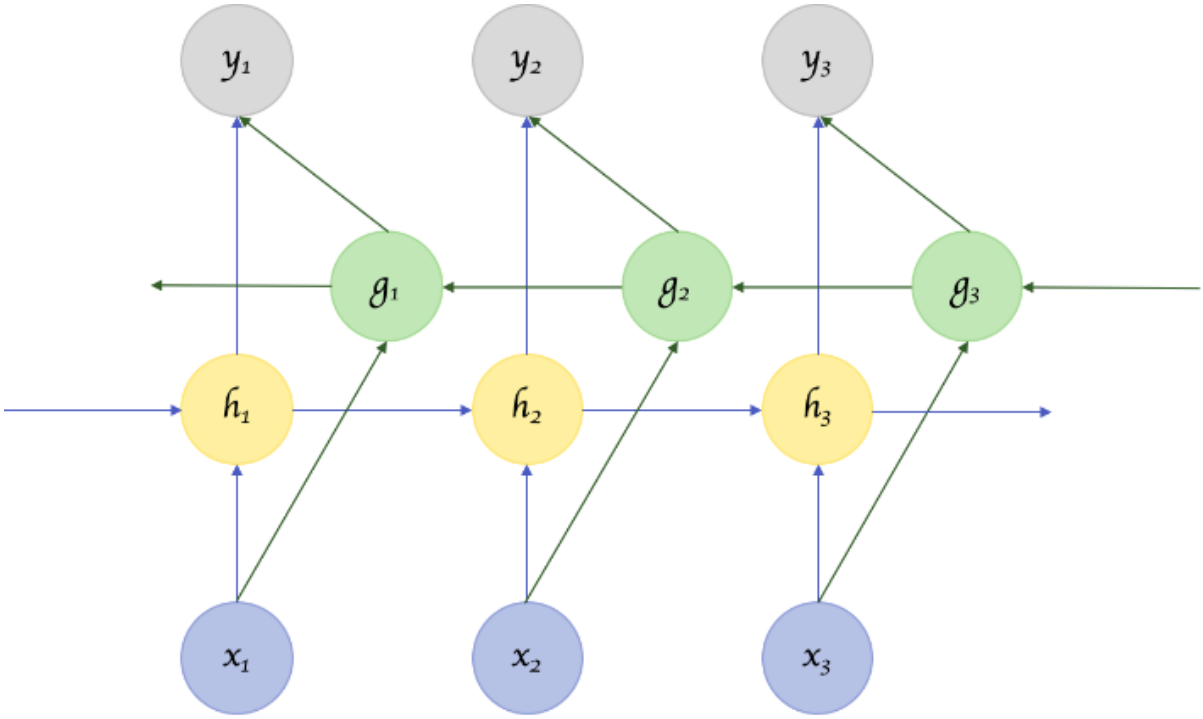
Fonte: Medium[a]

[a] Disponível em:<https://medium.com/explore-artificial-intelligence/an-introduction-to-recurrent-neural-networks-72c97bf0912>. Acesso em: 15 de novembro de 2019.

O conceito fundamental das RNN deu origem a algumas variantes, como as LSTMs (*Long Short Term Memory networks*) (HOCHREITER; SCHMIDHUBER, 1997), que exploraram ainda mais a concepção de dependências de longo prazo e permanência de contexto. Esse tipo de rede introduziu mudanças essenciais na forma como se processam as entradas em uma RNN, tornando-se uma das variantes com melhores resultados e aplicações em PLN. As LSTMs apresentam mudanças nos nós de uma RNN, com novos estados e portas, de forma a garantir que um contexto seja mantido ou resetado, independente da distância entre a entrada atual e o contexto em questão. De fato, esse tipo de rede foi desenvolvido para evitar o problema de se perder contextos em dependências de longo prazo, e lembrar informações por longos períodos é o comportamento que caracteriza uma LSTM.

Dentre as outras variações de RNN, o trabalho apresentado por Artetxe et al. (2018) utiliza RNNs bidirecionais em sua estrutura padrão *encoder-decoder*. Essa variante de rede torna a RNN capaz de usar o contexto não só de entradas passadas, como também levar em conta as entradas futuras na sequência. Essa capacidade da rede ajuda a distinguir o contexto em situações onde a ambiguidade da sentença poderia impedir uma boa compreensão. O diagrama da Figura 5 demonstra esse comportamento. Enquanto x, y e h permanecem similares às apresentadas na Figura 3, esse diagrama introduz g, que carrega informações das entradas seguintes no processamento do t atual. Em alguns casos, a escolha de quantas entradas futuras considerar pode ser um desafio. Em cenários como o reconhecimento de fala, aguardar que diversas palavras sejam faladas antes de se processar a atual pode não ser o ideal. Entretanto, na análise escrita, como o caso da tradução automática, geralmente pode-se considerar uma sentença por completo, visto que ela já se encontra disponível.

Figura 5 – Diagrama de uma RNN bidirecional



Fonte: Towards Data Science[a]

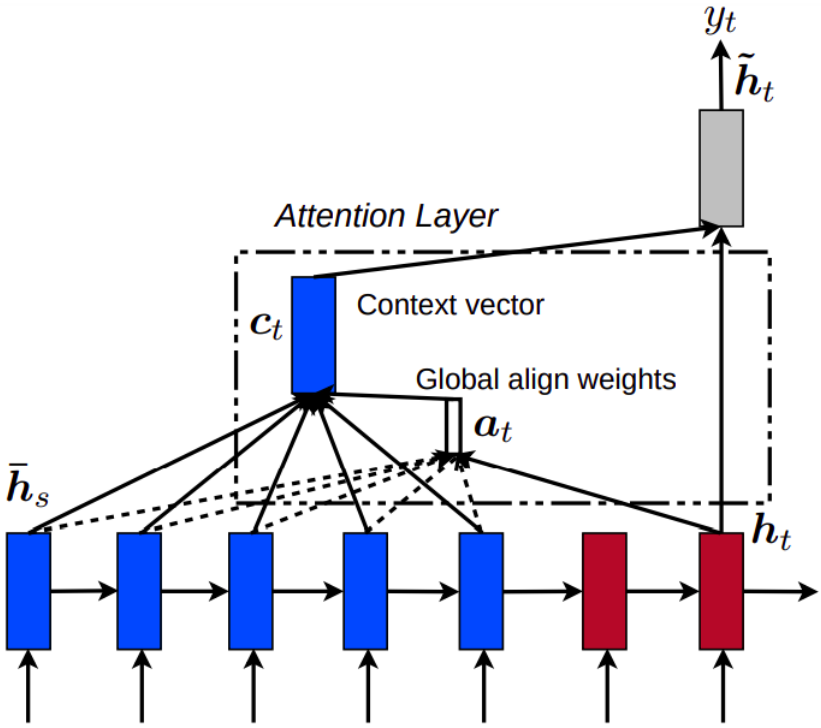
[a] Disponível em:<https://towardsdatascience.com/recurrent-neural-networks-d4642c9bc7ce>. Acesso em: 15 de novembro de 2019.

Outro processo que aprimora o desempenho de uma RNN, como mencionado, é um mecanismo de atenção (LUONG; PHAM; MANNING, 2015). Este, é um método que, a partir de n elementos de entrada e um contexto, retorna um vetor que representa o grau de conexão entre cada entrada e o contexto. Ou seja, retorna um vetor com a média aritmética ponderada da entrada, cujos pesos representam o grau de relevância de cada entrada para o contexto em questão. A estrutura geral da RNN, independente da variante, permanece a mesma. O que ocorre é a produção de um conjunto extra de saídas que são usadas para parametrizar o modelo de atenção, que trabalha sobre um conjunto de tamanho fixo de saídas que é passado à rede neural como uma entrada extra no passo sequencial seguinte. Basicamente, para cada entrada no tempo t, o mecanismo dita o grau de atenção que deve se dar a entradas passadas considerando uma janela de tamanho fixo.

Um mecanismo de atenção comumente utilizado é o modelo global. Nesse modelo, o vetor de atenção considera todas as entradas anteriores e calcula um peso (relevância) para cada uma delas. Esse vetor é, então, usado em conjunto com o vetor de contexto para se prever a saída do modelo. A Figura 6 ilustra esse comportamento. Na figura, o vetor a_t representa o vetor de atenção no tempo t. Esse vetor considera todas as entradas passadas e a atual para inferir o peso que cada uma delas deve ter sobre o contexto c_t atual. Outros mecanismos de atenção variam de acordo com a quantidade de entradas consideras e a função matemática

que calcula os pesos de cada entrada.

Figura 6 – Diagrama de um modelo de atenção global. A cada tempo t, o modelo infere um vetor de alinhamento a_t baseado no estado h_t atual e em todas as entradas anteriores. O vetor de contexto c_t é, então, calculado como a média ponderada, de acordo com os pesos de a_t, das entradas anteriores.



Fonte: (LUONG; PHAM; MANNING, 2015)

2.2 Word Embedding

Independente da variante de rede escolhida, para ser usada nesse modelo, cada palavra precisa ser transformada em uma estrutura de dados que seja compatível com o modelo de aprendizado de máquina, um modelo matemático. Tradicionalmente, cada palavra é transformada em um vetor conhecido como *One Hot Encoding*. Nesse vetor, cada dimensão (posição) representa uma palavra presente no córpus de entrada. Assim, ele possui dimensão igual à quantidade de palavras distintas do córpus de treinamento. Para cada palavra, o vetor é preenchido totalmente com valores *0*, exceto a posição que representa a própria palavra, que é preenchida com *1*. Ao criar esse vetor, um índice é associado a cada palavra. Com esses índices e as palavras tem-se o vocabulário para cada idioma do tradutor. Idealmente, o vocabulário conteria cada palavra única do idioma, porém como o número pode ser muito grande, o vocabulário é limitado às N palavras mais comuns no córpus utilizado. A Figura 7 ilustra essa representação para três palavras (televisor, imagem e nítida) em vetores de 6 dimensões.

Além do *one hot encoding*, outras maneiras de representar uma palavra na forma de vetor podem ser empregadas. Uma das mais utilizadas na atualidade é a representação de

palavras segundo um modelo de espaço vetorial.

Em um Modelo de Espaço Vetorial (MEV) (SALTON, 1971) os textos são representados em forma de vetores numéricos que denotam pontos no espaço, ao invés de sua representação textual original. Existem diversas formas e procedimentos para se converter um texto em um MEV. Esse modelo auxilia no processamento matemático de um texto, visto que transforma textos – podendo ser palavras, sentenças, expressões, documentos – em representações matemáticas que podem ser manipuladas por meio de operações aritméticas.

Assim, a conversão ou mapeamento para um MEV permite que essas unidades de linguagem sejam representadas por posições no espaço, que podem ser comparadas com as posições de outras unidades. Dependendo do método de mapeamento utilizado, a função de proximidade entre cada ponto representa um grau de similaridade entre essas unidades de linguagem. Diversos tratamentos para textos usando MEV foram propostos. No trabalho de Artetxe et al. (2018), *word embeddings* foi o modelo de espaço vetorial escolhido como representação das palavras.

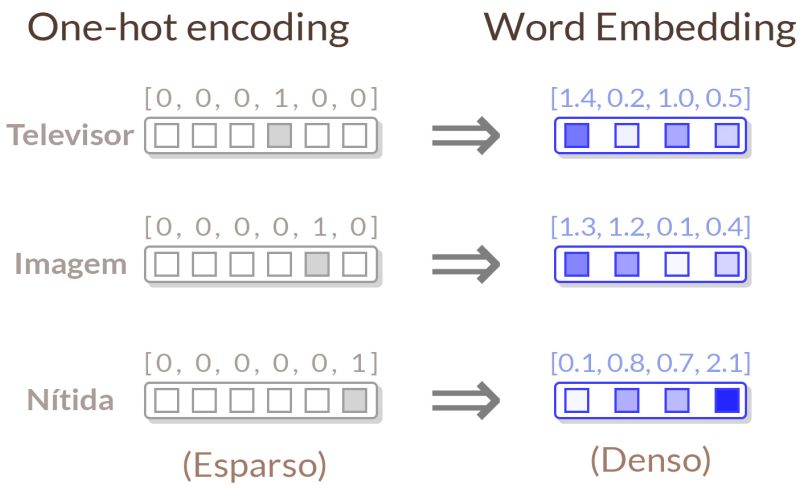
A expressão *Word Embedding* (WE) foi primeiramente usada por Bengio et al. (2003). Seu uso começou a ser melhor explorado após o trabalho de Collobert e Weston (2008), onde WEs se mostraram uma forma eficiente para a realização de tarefas de PLN. Entretanto, foi apenas em 2013 que a representação ganhou destaque, com a criação da ferramenta *word2vec* (MIKOLOV et al., 2013a), capaz de treinar o modelo e utilizá-lo de forma mais eficiente. No ano seguinte, outra ferramenta surgiu para o treinamento de WE, a GloVe (PENNINGTON; SOCHER; MANNING, 2014), demonstrando o auge recente da pesquisa e da utilização dessa representação.

A geração de WE tem sido considerada uma das poucas tarefas de treinamento não-supervisionado a conseguir tamanho sucesso de resultados. Assim, sendo não-supervisionado, pode ser usado no córpus alvo, contendo textos não estruturados – como o domínio do *e-commerce* – para gerar WEs do próprio domínio alvo, de forma automática e eficaz.

A criação de WEs ocorre por meio da transformação de um vetor numérico – que representa uma unidade de linguagem, como uma palavra – para outro vetor numérico de menor dimensão. Geralmente, o vetor inicial é um vetor esparso, ou seja, a maioria dos valores presentes nele são *0*, e pouco representativos. E o vetor de saída é um vetor denso, ou seja, possui uma quantidade razoável de valores que não são *0*, e possivelmente mais representativos. É comum que esse vetor inicial seja o *one hot encoding*, já mencionado. A Figura 7 ilustra essa transformação de um *one hot encoding* em um vetor denso. Na figura é representado um modelo hipotético, onde o córpus de entrada apresenta um vocabulário de apenas 6 palavras. Esse vetor esparso de 6 dimensões é mapeado para um vetor denso de 4 dimensões, constituindo a *word embedding*. Após a transformação do vetor esparso em um vetor de menor dimensão e denso tem-se a *word embedding*. Como o vetor de entrada possui tantas dimensões quanto forem as palavras distintas do córpus de entrada (o vocabulário), é comum que contenham milhares, ou

dezenas de milhares, de dimensões, enquanto as WE geralmente possuem geralmente 100 ou 300 dimensões.

Figura 7 – Transformação de *one hot encoding* em *word embedding*.



Fonte: (SILVA; CASELI, 2017)

Existem várias formas de se transformar um vetor de alta dimensionalidade para outro de menor dimensão. Este trabalho dá enfoque ao modelo utilizado por Artetxe et al. (2018), que foi o método proposto por Mikolov et al. (2013a) com a ferramenta *word2vec*. Esse método traz uma forma eficiente de se produzir WEs a partir de textos de entrada puros (não estruturados), sem pré-processamento.

Dois modelos estão disponíveis na *word2vec* para a geração de WE: *Continuous Bag-of-Words* (CBOW) e *Skip-Gram*. Os modelos trabalham de forma similar, e a principal diferença entre eles é que o CBOW é treinado de forma a tentar prever uma palavra de acordo com o contexto, enquanto o *Skip-Gram* faz o contrário, tenta prever o contexto de acordo com uma palavra, conforme ilustra a Figura 8. Ambos os modelos usam redes neurais para o treinamento. Cada palavra passa por camadas de neurônios um certo número de vezes e, a cada iteração, essa palavra é comparada com as representações vetoriais das outras palavras do vocabulário que se encontram no contexto da palavra atual. Nessa comparação, os vetores das palavras comparadas são alterados de forma a aproximar os vetores (que representam pontos no espaço) das palavras que estão nesse contexto, e afastar, no espaço, as palavras que não estão. O processo de aproximação e afastamento é realizado somando e subtraindo valores desses vetores comparados. Assim, para cada palavra, o algoritmo altera os valores dos vetores das palavras de uma mesma sentença de forma a se tornarem cada vez mais próximos e, por consequência, mais similares. E, ainda, altera os vetores dos vetores das palavras que não estão na mesma sentença de forma a torná-los mais distintos e distantes no espaço vetorial.

Com as WEs produzidas é possível, por exemplo, reduzir sua dimensionalidade e representá-las em um espaço bidimensional ou tridimensional para sua visualização. Nessa

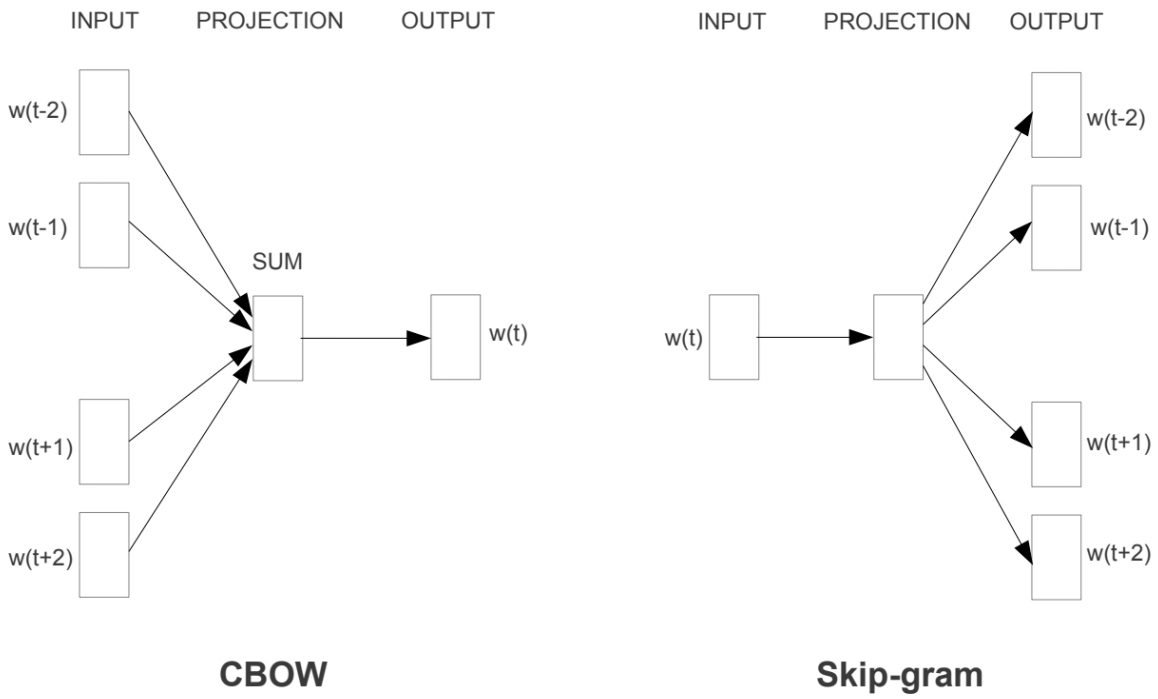


Figura 8 – Abordagens CBOW (à esquerda) e *Skip-gram* (à direita).
Fonte: (MIKOLOV et al., 2013b)

representação gráfica, as palavras (pontos) que se encontram mais próximas tendem a ter sentidos (significados) similares. As *word embeddings* também são capazes de preservar relações semânticas entre pares de palavras. Esse comportamento é ilustrado na Figura 9. Nesse contexto, à esquerda são representados países, que são relacionados às capitais representadas à direita. É possível observar que a distância e direção entre as palavras correspondentes é similar entre os pares de palavras ilustrados, o que demonstra a habilidade do modelo de automaticamente organizar os conceitos e aprender relações implícitas entre eles.

No contexto deste trabalho, as *word embeddings* são treinadas a fim de serem usadas como entrada no modelo neural da tradução. Tal como o trabalho de Artetxe et al. (2018), foram treinadas WEs nos dois idiomas investigados: inglês e português. Em seguinda, as WEs treinadas foram mapeadas em um espaço vetorial compartilhado, usando um método inter linguístico de mapeamento de *word embedding* descrito nos trabalhos de Artetxe, Labaka e Agirre (2018b), Artetxe, Labaka e Agirre (2018a), Artetxe, Labaka e Agirre (2017), Artetxe, Labaka e Agirre (2016), chamado *Vecmap*[1]. Esse processo torna comparáveis as WE de cada idioma, usando métodos que as mapeiam para o mesmo espaço vetorial. O mapeamento feito pelo *Vecmap* permite que, ao escolher um ponto (palavra) p no espaço vetorial de um idioma, a posição de p no espaço vetorial do outro idioma corresponda à tradução – ou aproximação – desta palavra naquele idioma. Assim, é possível relacionar uma WE em um idioma com uma WE em outro idioma por meio da posição dela no espaço vetorial.

[1] Disponível em:<https://github.com/artetxem/vecmap>. Acesso em: 08 de novembro de 2019.

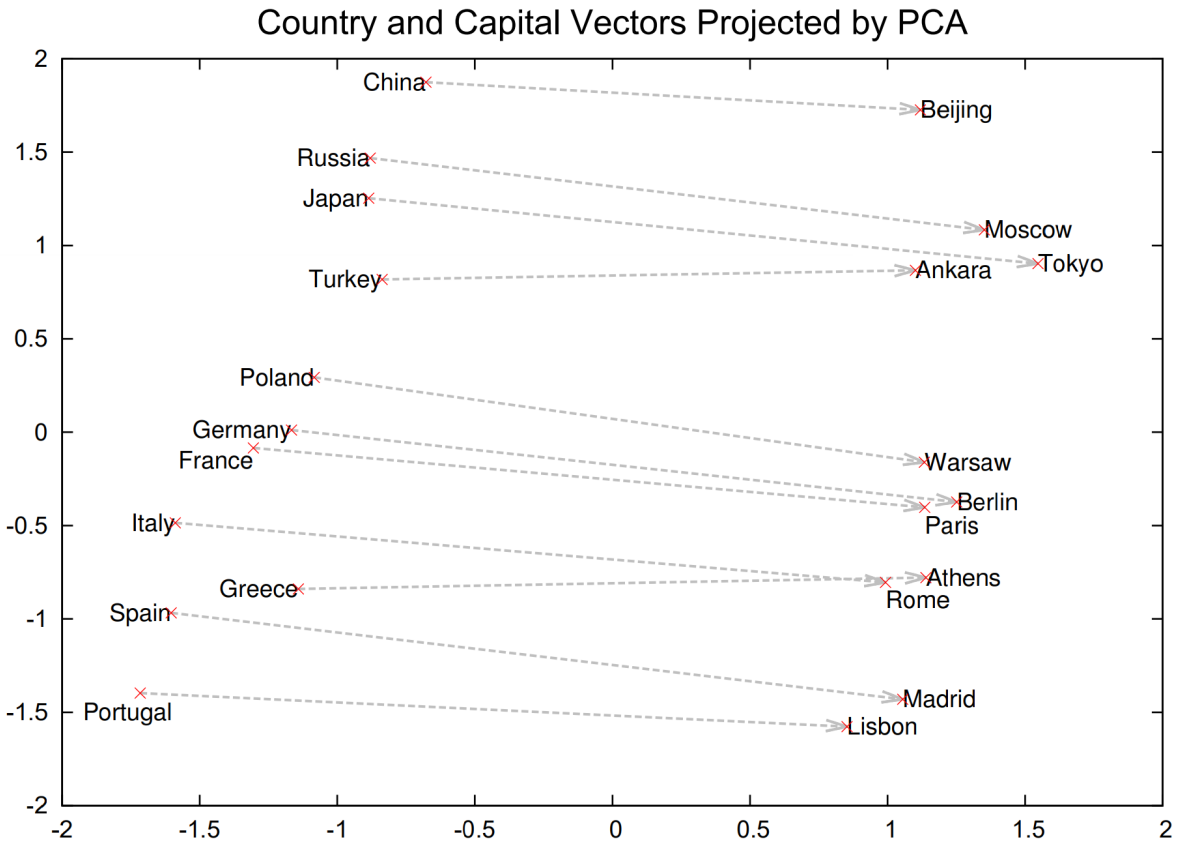


Figura 9 – Preservação de relação semântica entre pares de palavras similares.
Fonte: (MIKOLOV et al., 2013b)

2.3 Medidas de avaliação automática

A fim de avaliar os resultados obtidos por esse trabalho, foram consideradas medidas de avaliação automática como métrica. Estão disponíveis na literatura diversos trabalhos que propõem medidas automáticas para avaliar a qualidade de traduções automáticas (PAPINENI et al., 2002; DODDINGTON, 2002; SNOVER et al., 2006; DENKOWSKI; LAVIE, 2011; POPOVIĆ, 2015).

Antes, contudo, é necessário explicar os conceitos de *token* e n-grama. Um *token* é uma unidade mínima de avaliação que, no caso das medidas utilizadas neste trabalho, pode ser uma palavra, um número ou mesmo um caractere de pontuação. Um n-grama é uma sequência de n *tokens*, ou seja, um unigrama é apenas um *token*, enquanto um bigrama é uma sequência de dois *tokens*, um trigrama, de três *tokens* e assim por diante.

Papineni et al. (2002) apresentam uma medida de avaliação automática chamada de BLEU. A medida é calculada para segmentos individuais traduzidos – geralmente sentenças – ao compará-los com um conjunto de traduções boas (referência), traduzidas por humanos. Essa medida tem como base o cálculo da precisão em que o número de palavra presentes na sentença candidata (produzida pelo tradutor automático) também presentes na sentença referência

(produzida por um tradutor humano) é dividido pelo número total de palavras da sentença candidata. Como limitações dessa proposta simples, os autores apontam o fato dela favorecer a presença de palavras repetidas na sentença candidata. Para solucionar este problema, Papineni et al. (2002) indicam o cálculo de *modified unigram precison* que ignora as palavras com um casamento já identificado entre a sentença candidata e a referência.

O cálculo da *modified unigram precison* é apresentado na equação 2.1. Ele leva em consideração cada palavra distinta w presente na sentença candidata s.

$$\text{score}(s) = \frac{\sum_{w \in s} \min\{c(w,s), c(w,r)\}}{\|s\|} \tag{2.1}$$

A primeira etapa é a contagem de ocorrências da palavra em questão nas sentenças de referência r e a candidata. Depois, o número de ocorrências da palavra em questão na sentença candidata ($c(w,s)$) é comparado com o valor obtido para a referência ($c(w,r)$) e o mínimo entre os dois valores é escolhido. A escolha do valor mínimo permite à medida ignorar ocorrências múltiplas da palavra que não possuam um determinado casamento com a referência. O procedimento é realizado para cada palavra distinta na sentença candidata, sendo as quantidades somadas. Ao final, o valor é normalizado com relação ao número de *tokens* da sentença candidata.

A medida também leva em consideração o tamanho da sentença candidata em comparação à referência. Quando o tamanho da sentença produzida é menor que o da sentença de referência, calcula-se uma penalidade (BP, do inglês *brevity penalty*) exponencial – $BP = \exp\left(1 - \frac{\|s\|}{\|r\|}\right)$ – referente ao tamanho das sentenças. Caso contrário, não há penalidade, isto é $BP = 1$.

Desta forma, o cálculo final da medida BLEU é apresentado na equação 2.2, sendo p_n o valor de precisão calculada para os n-gramas de tamanho n, variando de 1 até um valor pré-determinado máximo N (geralmente 4). Os valores dos pesos w_n devem somar 1 e os autores da medida indicam a utilização de valores uniformes $\frac{1}{N}$.

$$\text{BLEU} = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{2.2}$$

BLEU foi uma das primeiras medidas de avaliação automática a representar uma boa correlação com a ideia humana, e subjetiva, de qualidade, e continua sendo uma das mais aplicadas até hoje. O valor de BLEU varia de 0 a 1 (ou a 100, como se convencionou representar). Esse valor indica a similaridade da sentença candidata com a(s) referência(s), e quanto maior esse valor, melhor considera-se a qualidade da sentença candidata gerada pelo tradutor automático.

Doddington (2002) propõe alterações nos cálculos de BLEU, resultando em uma nova medida chamada NIST. O autor indica que a técnica de combinação dos valores de precisão

para os diferentes n-gramas utilizado na BLEU (média geométrica) é muito sensível a pequenas contagens de coocorrências, bastante comuns para n-gramas maiores. Para solucionar esse problema, Doddington (2002) indica o uso da média aritmética ao invés da média geométrica.

A métrica ainda calcula o quão informativo é um n-grama, ou seja, quando um n-grama correto é encontrado, quanto mais raro (ou menos frequente) ele é, mais peso ele recebe no cálculo. Para isso, é utilizado o cálculo de informação conforme mostra a equação 2.3 em que a informação de um determinado n-grama $w_1, \ldots, w_n$ é calculado através de contagens como representado pela função $c(\cdot)$.

$$\mathrm{Info}(w_1, \ldots, w_n) = \log_2 \left(\frac{c(w_1, \ldots, w_{n-1})}{c(w_1, \ldots, w_n)} \right) \tag{2.3}$$

Os valores de informação são utilizados para calcular a medida NIST para um par de sentença candidata s e sentença de referência r, assim como para a medida de BLEU. A equação 2.4 apresenta o cálculo da NIST. $\mathcal{W}$ representa todos os n-gramas de tamanho n que possuem correspondência na sentença referência. Os cálculos são realizados para todos os n-gramas de tamanho até N, pré-determinado anteriormente ao cálculo.

$$\mathrm{NIST} = BP \cdot \sum_{n=1}^{N} \left\{ \frac{\sum_{w \in \mathcal{W}} \mathrm{Info}(w)}{\|s\|} \right\} \tag{2.4}$$

Doddington (2002) também mantém o uso de uma penalidade a sentenças curtas (BP). Apesar disso, o cálculo do BP foi alterado pois, segundo o autor, o cálculo utilizado na BLEU possuía variações pequenas, o que pode impactar a medida final. O método de cálculo do BP implementado na medida NIST é apresentado na equação 2.5.

$$BP = \exp \left\{ \beta \log^2 \left[\min \left(\frac{\|s\|}{\|r\|}, 1 \right) \right] \right\} \tag{2.5}$$

Os valores de NIST iniciam em 0, mas, diferente de BLEU, não possuem um limite máximo. Da mesma forma que BLEU, quanto maior o valor de NIST, melhor é a qualidade da tradução gerada.

Denkowski e Lavie (2011) também propõem uma medida automática de avaliação da tradução automática chamada METEOR. Esta medida, assim como a BLEU, é baseada no casamento das palavras entre a sentença candidata e a sentença de referência, porém a METEOR incorpora diversos métodos de *match*, sendo estes:

- Exato: O casamento entre as formas das palavras é exato;
- Raíz: Realiza o casamento entre as raízes das palavras, através da aplicação do método de *stemming* desenvolvido por Porter (2001);

- Sinônimos: Realiza o casamento entre sinônimos disponíveis na WordNet (MILLER, 1995);
- Paráfrases: Realiza o casamento entre paráfrases disponíveis em tabelas desenvolvidas pelos próprios autores em seu trabalho.

Por conta de limitações dos recursos, a utilização de sinônimos está limitada ao idioma inglês e as tabelas de paráfrases aos idiomas inglês, francês, alemão e espanhol.

Para realizar o cálculo da medida, Denkowski e Lavie (2011) diferenciam as chamadas *content words* (h_c na sentença candidata e r_c na referência) das chamadas *function words* (h_f e r_f), com pesos diferentes através de um parâmetro δ. Cada método de casamento m_i apresentado anteriormente é levado em consideração com pesos diferentes w_i. A partir dessas informações é possível calcular tanto a precisão quanto a cobertura, como indica a equação 2.6. A precisão é calculada com relação a quantas palavras na sentença candidata foram alinhadas através do casamento, já a cobertura utiliza como base de cálculo o número de palavras alinhadas na sentença de referência.

$$\begin{aligned} \text{Precisao} &= \frac{\sum_i w_i \cdot (\delta \cdot m_i(h_c) + (1-\delta) \cdot m_i(h_f))}{\delta \cdot |h_c| + (1-\delta) \cdot |h_f|} \\ \text{Cobertura} &= \frac{\sum_i w_i \cdot (\delta \cdot m_i(r_c) + (1-\delta) \cdot m_i(r_f))}{\delta \cdot |r_c| + (1-\delta) \cdot |r_f|} \end{aligned} \tag{2.6}$$

Com os valores de precisão e cobertura é possível calcular uma média harmônica parametrizada entre eles, como é indicado na equação 2.7. Este cálculo possui um parâmetro α indicado no momento do cálculo da medida.

$$F = \frac{\text{Precisao} \cdot \text{Cobertura}}{\alpha \cdot \text{Precisao} + (1-\alpha) \cdot \text{Cobertura}} \tag{2.7}$$

Denkowski e Lavie (2011) também calculam uma penalidade que considera diferenças na ordem das palavras conforme a equação 2.8. Este valor é calculado a partir do número total de palavras alinhadas m, calculado como a média do número de palavras alinhadas na sentença candidata e na sentença referência. Este cálculo também leva em consideração o número total de *chunks* ch, sendo um *chunk* uma sequência de *tokens* alinhados contiguamente e ordenados de forma idêntica nas duas sentenças. β é um parâmetro estabelecido previamente ao cálculo.

$$\text{Penalidade} = \gamma \cdot \left(\frac{ch}{m}\right)^{\beta} \tag{2.8}$$

A partir destes valores, é possível calcular o valor final da medida METEOR, conforme é indicado na equação 2.9. Assim como a BLEU, os valores de METEOR variam de 0 a 1 (ou

100) e quanto maior for este valor, melhor considera-se a qualidade da sentença candidata gerada pelo tradutor automático.

$$\text{METEOR} = (1 - \text{Penalidade}) \cdot F \tag{2.9}$$

3 TRABALHOS RELACIONADOS

Como mencionado, alguns domínios apresentam o desafio de possuir um conjunto de textos insuficiente na construção de um bom modelo de tradução. Essa dificuldade motivou diversas pesquisas direcionadas a atacar esse problema, e os trabalhos predominantes sobre adaptação de domínio foram desenvolvidos para SMT. Os métodos propostos por esses trabalhos podem ser divididos em duas categorias principais: centrados em dados e centrados em modelo. Os centrados em dados são baseados na seleção do córpus de treinamento a ser utilizado pela tradução (MOORE; LEWIS, 2010; AXELROD; HE; GAO, 2011; DUH et al., 2013; CUONG; SIMA'AN, 2014; DURRANI et al., 2015; CHEN et al., 2016; UTIYAMA; ISAHARA, 2003; WANG et al., 2014; CHU, 2015; WANG et al., 2016; MARIE; FUJITA, 2017). Enquanto os centrados em modelo foram desenvolvidos a fim de alterar a estrutura de treinamento (SENNRICH; ARANSA, 2013; DURRANI et al., 2015; IMAMURA; SUMITA, 2017; MATSOUKAS; ROSTI; ZHANG, 2009; FOSTER; GOUTTE; KUHN, 2010; SHAH; BARRAULT; SCHWENK, 2010; ROUSSEAU et al., 2011; ZHOU; CAO; ZHAO, 2015).

Vários trabalhos foram propostos com o intuito de realizar a tradução automática em um domínio específico em SMT. Contudo, essa área de pesquisa na NMT é relativamente nova. Alguns estudos realizados para SMT sobre adaptação de domínio são usados como base para outros focados na tradução automática neural, no entanto, devido a diferentes características entre SMT e NMT, muitos dos métodos desenvolvidos para a tradução estatística não podem ser aplicados diretamente à NMT. Ainda assim, tal como na SMT, os métodos desenvolvidos para adaptação de domínio focados na tradução automática neural podem ser divididos nas mesmas duas categorias principais já mencionadas: centrados em dados e centrados em modelo.

Para os centrados em dados, os métodos desenvolvidos focam-se especialmente na utilização de (i) córpus monolíngue (ZHANG; ZONG, 2016; CHENG et al., 2016; CURREY; BARONE; HEAFIELD, 2017; DOMHAN; HIEBER, 2017), (ii) córpus sintético (SENNRICH; BIRCH, 2016; ZHANG; ZONG, 2016; PARK; SONG; YOON, 2017), ou (iii) córpus paralelo fora do domínio (CHU; DABRE; KUROHASHI, 2017; SAJJAD et al., 2017; BRITZ; LE; PRYZANT, 2017; WANG et al., 2017a; WEES; BISAZZA; MONZ, 2017). Já nos centrados em modelo, as alterações ocorrem principalmente (i) na estrutura de treinamento (LUONG; MANNING, 2015; SENNRICH; BIRCH, 2016; SERVAN; CREGO; SENELLART, 2016; FREITAG; AL-ONAIZAN, 2016; WANG et al., 2017b; CHEN et al., 2017; VARGA, 2017; DAKWALE; MONZ, 2017; CHU; DABRE; KUROHASHI, 2017; BARONE et al., 2017), (ii) na arquitetura das redes neurais (KOBUS; CREGO; SENELLART, 2016; GULCEHRE et al., 2015; BRITZ; LE; PRYZANT, 2017), ou (iii) na etapa de *decoding* (GULCEHRE et al., 2015; DAKWALE; MONZ, 2017; KHAYRALLAH et al., 2017). Intersecções entre essas diversas abordagens são comuns, tanto em métodos da mesma categoria quanto entre os da outra categoria.

3.1 Centrados em Dados

A seguir são descritos os trabalhos de adaptação de domínio centrados em dados, aplicados na NMT, separados pelo tipo de dados que usaram.

3.1.1 Córpus Monolíngue

O trabalho de Gulcehre et al. (2015) enfatiza o uso de córpus monolíngue de domínio específico ao treinar um modelo de linguagem[1] em RNN (RNN *Language Model*, ou simplesmente RNNLM) em um conjunto de dados monolíngue. Posteriormente, o autor funde (usando Fusão Profunda e Rasa – descritas na seção 3.2.2) esse modelo de linguagem com um modelo NMT. Modelos distintos foram treinados para 4 pares de idiomas: chinês-inglês, turco-inglês, tcheco-inglês, e alemão-inglês. O córpus monolíngue de domínio específico utilizado foi o *English Gigaword*[2], composto por 1.756.504 *tokens*, e contém textos de notícias jornalísticas curtas (conhecidas como *newswire*). Dentre os modelos treinados, comparou-se aqueles que fizeram uso do córpus monolíngue através da RNNLM aos que usaram um treinamento NMT padrão – usando córpus paralelo bilíngue do domínio específico. O uso do córpus monolíngue aumentou até 1,96 pontos na medida BLEU (de 0 a 100 neste trabalho), com média de aumento de 1,65 pontos BLEU considerando todos os modelos. A partir dos resultados relatados pelos autores, eles concluem que o uso de córpus monolíngue do domínio pode melhorar a qualidade da tradução.

Em Currey, Barone e Heafield (2017), a abordagem proposta se baseia na cópia do córpus monolíngue do idioma alvo para o idioma fonte. Assim, cria-se um córpus paralelo onde cada sentença é idêntica. Esse conjunto é mesclado com o córpus paralelo bilíngue, e nenhuma distinção é feita entre o que foi copiado e o que de fato é tradução. Os modelos foram avaliados em 3 pares de idiomas: inglês-turco, inglês-romeno, e inglês-alemão. O domínio do córpus monolíngue utilizado foi o de notícias, e composto por 414.746 sentenças no idioma turco, 608.320 sentenças no romeno, e 10.000.000 sentenças no alemão. Comparando os modelos que usaram e os que não usaram cópia do córpus monolíngue como córpus paralelo, o aumento na métrica BLEU foi de até 1,2 pontos, com média de aumento de 0,48 entre os modelos .

Cheng et al. (2016) usa córpus monolíngue tanto do idioma fonte como do idioma alvo para o treinamento da NMT através da reconstrução do córpus monolíngue usando a NMT como um *autoencoder*[3]. A ideia central é reconstruir o córpus monolíngue usando um modelo *encoder-decoder*, onde o *encoder* é um modelo de tradução do idioma fonte para o alvo, e o *decoder* é um modelo do idioma alvo para o idioma fonte. Assim, o modelo por completo tem como entrada o idioma fonte, traduz para o idioma alvo e de novo para o idioma fonte, a fim de manter o texto original. O modelo também faz o processo inverso, ou seja, além do

[1] Modelos de linguagem associam valores de probabilidade a sequências de palavras

[2] Disponível em:<https://catalog.ldc.upenn.edu/LDC2003T05>. Acesso em: 08 de novembro de 2019.

[3] *Autoencoder* são redes neurais treinadas com o objetivo de reproduzir sua entrada para sua saída.

autoencoder fazer o processamento do idioma fonte para ele mesmo, ele também faz para o idioma alvo. O tradutor foi treinado com os córpus monolíngues Gigaword[4] com 18.75 milhões de sentenças em chinês e 22.32 milhões em inglês. O modelo foi comparado com um treinado em SMT e outro em NMT com RNNs, e superou ambos os modelos com 3,5 e 4,7 pontos na medida BLEU, respectivamente.

3.1.2 Córpus Paralelo Sintético

Modelos NMT têm a capacidade de aprender modelos de linguagem, portanto o próprio modelo da tradução pode ser aplicado a fim de aprimorar a rede *decoder* (SENNRICH; BIRCH, 2016), com córpus no idioma alvo, ou a rede *encoder* (ZHANG; ZONG, 2016), com córpus no idioma fonte. Uma técnica empregada na tentativa de melhorar a qualidade dos modelos de linguagem é a *back translation*, na qual gera-se um córpus paralelo sintético.

O córpus paralelo sintético é um conjunto de textos paralelos com, normalmente, traduções do idioma fonte para o idioma alvo. A diferença dele para o bilíngue está na origem da tradução. Enquanto um córpus paralelo bilíngue é traduzido por humanos, o sintético é traduzido por outro sistema de TA. A tradução pode ser feita tanto do idioma fonte (humano) para o alvo (sintético) – sendo usado para aprimorar o *decoder* –, como do idioma alvo (humano) para o idioma fonte (sintético) – aplicado na melhora do *encoder*. O modelo de TA usado nessa tradução sintética geralmente é outro tradutor NMT, ou mesmo um SMT. Cada autor usa uma pequena variação no modelo para a tradução sintética, mas todos são baseados no treinamento padrão de cada sistema, usando córpus paralelo – podendo ser de domínio geral ou domínio específico, dependendo da disponibilidade.

Alguns trabalhos demonstraram que a aplicação de córpus sintético ao treinamento apresenta bons resultados para adaptação de domínio. Neste contexto, é possível usar córpus monolíngue do idioma alvo (SENNRICH; HADDOW; BIRCH, 2016), ou córpus monolíngue do idioma fonte (ZHANG; ZONG, 2016), ou mesmo gerar o córpus paralelo de ambos (PARK; SONG; YOON, 2017).

Em Sennrich, Haddow e Birch (2016), experimentos são conduzidos a fim de aprimorar a tradução de palavras raras ou inexistentes no córpus de treinamento. Para isso, o trabalho gera um córpus paralelo sintético a partir de uma coleção de textos monolíngue no idioma alvo traduzindo-os automaticamente para o idioma fonte. Foram usados 2 pares de idiomas: inglês-alemão e inglês-russo. Ambos foram treinados com um córpus do domínio de notícias jornalísticas. Para o par inglês-alemão, o conjunto de treinamento continha 4,2 milhões de sentenças e aproximadamente 100 milhões de *tokens*, enquanto para o par inglês-russo o córpus possuia 2,6 milhões de sentenças e 50 milhões de *tokens*. O modelo foi comparado com outros que não usam o *back-translation* e algumas variações de treinamento, nos dois pares de idiomas.

[4] Disponível em:<https://catalog.ldc.upenn.edu/>. Acesso em: 08 de novembro de 2019.

O modelo proposto superou a maioria dos modelos, com o aumento na pontuação da medida BLEU entre 0,3 e 1,3.

A pesquisa realizada por Zhang e Zong (2016) envolvia a exploração de duas abordagens com o propósito de investigar como a aplicação de córpus monolíngue no idioma fonte pode melhorar modelos em NMT. A primeira abordagem envolvia usar o modelo de treinamento para gerar o córpus sintético paralelo a partir de um córpus de grande volume no idioma fonte, como investigado no trabalho de Sennrich, Haddow e Birch (2016), que usou no idioma alvo. Assim, foi construído um sistema de tradução base com o córpus paralelo disponível, e então usado este modelo para criar um córpus paralelo sintético no conjunto de textos monolíngue do idioma fonte.

A segunda abordagem aplica aprendizado multitarefa para prever a tradução no idioma alvo e, ao mesmo tempo, usa sentenças reordenadas no idioma fonte. Dessa forma, essa abordagem usa duas NMTs: uma treinada com córpus paralelo bilíngue para prever uma sentença do idioma alvo a partir de uma do idioma fonte, e a outra treinada no córpus monolíngue do idioma fonte para prever a sentença corretamente ordenada a partir da sentença original. O reordenamento proposto envolve alterar a ordem das palavras na sentença no idioma fonte para que a ordem seja próxima àquela pretendida no idioma alvo. O trabalho também investiga a aplicação de um córpus reduzido comparado a um córpus de grande volume. O par de idiomas usado foi o inglês-chinês, e o conjunto de dados paralelo continha 0,63 milhões de sentenças no córpus reduzido, e 2,1 milhões no córpus volumoso. Já o conjunto de dados monolíngue era composto por 6,5 milhões de sentenças no córpus reduzido e 12 milhões no volumoso. Todo o conjunto de dados foi retirado do domínio de notícias jornalísticas. O modelo foi treinado com diferentes variações e comparado com um treinado em NMT padrão e outro em SMT. O modelo tem o melhor desempenho quando as sentenças são ordenadas de acordo com a cobertura das palavras, e se usa o córpus volumoso com 50% do córpus total, excluindo as sentenças de menor cobertura. Esse modelo superou todos os modelos com os quais foi comparado, com aumento de até 4,05 pontos na métrica BLEU.

Por fim, o trabalho de Park, Song e Yoon (2017) visa investigar o uso exclusivo de córpus paralelo sintético no treinamento de um modelo NMT. Os autores propõem um novo tipo de córpus paralelo formado unicamente por traduções sintéticas, batizado de *Pseudo mix*. Nesse modelo, os córpus monolíngues em ambos os idiomas são traduzidos sinteticamente e, assim, são produzidos dois conjuntos de sentenças paralelas sintéticas. Os dois córpus paralelos sintéticos resultantes são mesclados, formando um único córpus onde cada texto pode ser a sentença original (fonte) pareada com a tradução sintética (alvo), ou o inverso, tradução sintética (fonte) e sentença original (alvo). Assim, o córpus resultante é uma combinação da união dos dois córpus sintéticos.

Os pares de idiomas escolhidos para os experimentos foram o francês-alemão e o tcheco-alemão, usando um córpus com notícias jornalísticas, com 1,45 milhões de sentenças

monolíngues tanto no idioma francês como no alemão, para o primeiro experimento (realizado apenas no par francês-alemão). Experimentos posteriores, usando os dois pares de idiomas, foram conduzidos com córpus mais volumosos. O córpus *pseudo mix* do primeiro experimento foi formado a partir da exclusão aleatória de metade das sentenças de cada córpus sintético, resultando em um conjunto com 1,45 milhões de sentenças paralelas sintéticas. No segundo experimento, a criação do córpus foi similar, para os dois pares de idiomas, e resultou em um córpus *pseudo mix* de 3,5 milhões no par tcheco-alemão e 3,7 milhões no par francês-alemão.

Os modelos NMT criados foram comparados com diversos modelos, incluindo modelos usando tradução sintética apenas em um idioma (tanto no alvo como no fonte), modelos de SMT, modelos com córpus paralelo bilíngue (traduzido por humanos), variações de tamanho de córpus, modelos com córpus paralelo sintético usando *back-translation* e modelos com córpus paralelo sintético gerado através de idioma *pivot*[5]. Experimentos também foram conduzidos ao incrementar o córpus paralelo sintético com traduções feitas por humanos. Os autores do trabalho concluem que o modelo *pseudo mix* proposto apresenta melhora substancial considerando os modelos comparados, superando em até 5,11 pontos BLEU[6] com o córpus aprimorado por traduções humanas, e, sem o uso de traduções feitas por humanos, o modelo superou o *baseline* em até 4,43 pontos BLEU.

3.1.3 Córpus Paralelo Fora do Domínio

Em TA, um córpus fora do domínio é considerado um conjunto de textos formado por sentenças cujo vocabulário não é típico do domínio pretendido do treinamento. Esse córpus pode ser de domínio geral, ou mesmo de outro domínio específico diferente do domínio pretendido. Esse tipo de córpus é geralmente usado porque permite ao sistema expandir o vocabulário e o conhecimento de estrutura do idioma utilizado, desde que a estrutura do domínio específico seja similar à do córpus fora do domínio.

Ao usar córpus de domínio específico e córpus fora do domínio, torna-se necessário o treinamento de um sistema de domínio mesclado, que possa melhorar a tradução de sentenças no domínio específico sem prejudicar a tradução de sentenças fora do domínio. Métodos que realizam esse tipo de treinamento são conhecidos como métodos multidomínio e apresentam resultados interessantes para a TA.

Chu, Dabre e Kurohashi (2017) desenvolve um método multidomínio em que as sentenças são marcadas com *tags* de identificação de domínio no treinamento do modelo. A abordagem primeiro treina um modelo usando córpus fora do domínio, e depois aprimora o

[5] Nesse modelo, cada tradução sintética foi gerada usando um tradutor NMT com um idioma intermediário (no trabalho, o inglês). Assim, a tradução foi feita traduzindo a sentença do idioma fonte para o inglês e, em seguida, do inglês para o idioma alvo. Esse tipo de geração de córpus paralelo sintético superou a aplicação de simples *back-translation*.

[6] Modelo *pseudo mix* com tradução por idioma *pivot* com o melhoramento por traduções feitas por humanos, se comparado com o modelo *baseline* usando *back-translation*, na tradução de alemão para tcheco, considerando os experimentos com córpus volumoso.

modelo usando córpus paralelo de domínio específico mesclado com córpus paralelo de fora do domínio. Dois experimentos foram realizados neste trabalho: um com um conjunto de textos de domínio específico que os autores identificaram como "de alta qualidade", e outro com textos de "baixa qualidade". O primeiro experimento foi aplicado ao par de idiomas chinês-inglês, usando um córpus de domínio geral com 1 milhão de sentenças paralelas e um de domínio específico[7] com 209.491 sentenças paralelas. O segundo experimento usou o par de idiomas chinês-japonês, utilizando um córpus de domínio geral com 136.013 pares de sentença e um de domínio específico com 672.315 sentenças. Diversas variações dos modelos foram desenvolvidas para os dois experimentos. Os resultados demonstram que a aplicação de *tags* no treinamento multidomínio aprimora os modelos se comparado com o mesmo treinamento sem as *tags*. Para o primeiro experimento, com o par chinês-inglês, o modelo proposto, com *tags*, superou o sem *tags* em até 0,67 pontos BLEU, e para o segundo experimento, no par chinês-japonês, superou em até 0,7 pontos BLEU.

O trabalho de Sajjad et al. (2017) realiza experimentos usando diversas abordagens para modelos multidomínio. São aplicados 4 métodos nos testes:

1. **Concatenação de dados** – treina-se um modelo com base na concatenação de todo o córpus de domínio específico com o de domínio geral;

2. **Empilhamento de modelos** – constróe-se um tradutor NMT usando variações de domínio, começando pelo domínio menos similar ao pretendido, em seguida aprimorando-o com domínios mais similares e, finalmente, aprimorando ainda mais com o córpus de domínio específico;

3. **Seleção de dados** – seleciona-se um determinado percentual do córpus de domínio geral que seja mais semelhante ao córpus de domínio específico e usa essa seleção para o treinamento;

4. **Conjunto multidomínio** – treinam-se modelos separados para cada domínio disponível e depois esses modelos são combinados no passo *decoder* do modelo usando algum sistema de médias para cada modelo.

Os sistemas treinados seguindo os métodos descritos foram aplicados aos pares de idiomas árabe-inglês e alemão-inglês. Como córpus de domínio específico, ambos os pares de idioma usam sentenças traduzidas transcritas de *TED talks*, sendo 229.000 sentenças para o par árabe-inglês e 209.000 para alemão-inglês. Para os córpus de domínio geral são usados dois córpus para cada par de idiomas: um com 18,3 milhões de sentenças paralelas e outro com 22,4 milhões para o par árabe-inglês, e, para o par alemão-inglês, um com 1,9 milhões de sentenças paralelas e outro com 2,3 milhões. Os resultados dos experimentos demonstram que

[7] Tradução das transcrições de *TED talks*

o sistema com concatenação, aprimorado com córpus de domínio específico, entre todos os métodos propostos, foi o que apresentou melhores resultados atingindo 38,0 pontos BLEU no par árabe-ingês e 38,1 no par alemão-inglês. Resultados também comprovam que a seleção dos dados pode diminuir o tempo de treinamento em troca de baixa perda de qualidade da tradução.

3.2 Centrados em Modelo

Nessa seção são sumarizados os trabalhos relacionados que utilizaram diferentes estratégias no treinamento das redes ou propõem alterações na arquitetura propriamente dita.

3.2.1 Estratégia de Treinamento

Foram desenvolvidas diversas estratégias para aprimorar as etapas de treinamento de um modelo NMT. Como visto nos experimentos descritos anteriormente, o aprimoramento de modelo é uma das estratégias mais utilizadas na adaptação de domínio (LUONG; MANNING, 2015; SENNRICH; BIRCH, 2016; SERVAN; CREGO; SENELLART, 2016; FREITAG; AL-ONAIZAN, 2016). Na estratégia de aprimoramento do modelo, o modelo de NMT é treinado em um córpus volumoso de domínio geral, e, em seguida, é aprimorado, geralmente, com o córpus de domínio específico, comumente de menor volume e qualidade se comparado ao de domínio geral.

Outra estratégia de treinamento aplicada é o aprimoramento mesclado, que unifica o aprimoramento de modelo ao treinamento multidomínio. Essa estratégia é aplicada em duas etapas: primeiro treina-se um modelo NMT em um córpus de domínio específico, depois o modelo é aprimorado usando uma mesclagem do córpus de domínio específico com outro de domínio geral. Nesse contexto, o trabalho de Chu, Dabre e Kurohashi (2017) demonstra que o aprimoramento mesclado apresenta melhores resultados se comparado com os modelos que usam unicamente multidomínio ou aprimoramento de modelo.

3.2.2 Arquitetura da Rede Neural

Nesse tipo de abordagem, a arquitetura da rede neural usada para treinar o tradutor neural é adaptada para melhorar a tradução do domínio específico. Uma das técnicas propostas para esse fim é a Fusão Profunda. Nela, um RNNLM é treinado para o *decoder* da rede neural em córpus monolíngue de domínio específico. Esse modelo de linguagem é, então, combinado – ou fundido – com um modelo de NMT padrão. O trabalho proposto por Gulcehre et al. (2015), já citado, usa essa técnica em um dos modelos treinados, e este é o que apresenta melhores resultados dentre os modelos treinados, em todos os cenários. Outra técnica similar a essa é a Fusão Rasa. Nela, ao invés de fundir os modelos de NMT e RNNLM, eles são considerados separadamente. Gulcehre et al. (2015) também aplica essa técnica, mas a Fusão Profunda

ainda foi a de melhor resultado nos experimentos reportados por eles. O trabalho de Domhan e Hieber (2017) introduz uma técnica similar à Fusão Profunda. Mas, ao contrário do que proposto por Gulcehre et al. (2015), os autores treinam os modelos RNNLM e NMT padrão de forma conjunta, e não separadamente.

Outra técnica, proposta por Britz, Le e Pryzant (2017), é a Discriminação de Domínio. Nessa técnica discriminativa, uma rede neural *feed-forward*[8] é adicionada ao *encoder* como discriminadora. Essa rede, então, usa o mecanismo de atenção para prever qual é o domínio da sentença de entrada. Assim, além de traduzir a sentença, o modelo procura identificar a qual domínio ela pertence.

Controle de Domínio é outra técnica usada para aprimorar a adaptação de domínio. Kobus, Crego e Senellart (2016) propõem a adição de atributos a nível de palavra (ou a nível de sentença) na camada de *embedding* da NMT, com o propósito de controlar os domínios. No trabalho, os autores adicionam uma tag de domínio para cada palavra do córpus.

No trabalho de Artetxe et al. (2018), investigado nesse trabalho, os autores também propõem alterações da estrutura da rede neural. A estrutura do modelo NMT é modificada em dois aspectos principais:

1. Os autores usam um *encoder* compartilhado, que aceita como entrada sentenças em ambos os idiomas. O compartilhamento do *encoder* é possível em virtude da utilização de *word embeddings*.

2. O uso de *encoder* compartilhado permite, segundo os autores, que a rede seja treinada e usada, simultaneamente, para tradução entre os idiomas em ambas as direções. Dessa forma, a estrutura da rede é composta de um *encoder* compartilhado e dois *decoders*, um para cada idioma. Assim, o *encoder* recebe a sentença de entrada e, usando um vocabulário específico para cada idioma, identifica qual *decoder* deve ser usado na tradução. Ao final do processo, a ferramenta proposta pelos autores tem como resultado 4 tradutores: alvo-fonte, fonte-alvo, alvo-alvo e fonte-fonte. Para o presente trabalho, focou-se no sistema alvo-fonte.

Mais detalhes sobre a estrutura da rede usada pelos autores se encontra na seção 3.3.1.

3.2.3 Estrutura do *Decoder*

Técnicas baseadas na estrutura do *decoder* são focadas na alteração da rede neural usada como *decoder* na NMT. Essa estratégia é complementar às apresentadas anteriormente , centradas em modelo.

[8] É um tipo de rede neural que, ao contrário da RNN, não possui loops e a informação é transportada apenas para frente nos nós.

Uma delas é a, já mencionada, Fusão Rasa, proposta por Gulcehre et al. (2015). Os modelos de linguagens do idioma alvo são usados no *decoder* da NMT a fim de aprimorar a tradução. A função dos RNNLM no *decoder* é associar uma nova pontuação à palavra prevista pelo modelo NMT baseada na soma ponderada da saída da NMT com a do RNNLM.

Khayrallah et al. (2017) propõem um algoritmo de *decoder* baseado em pilhas aplicado a reticulados[9] das palavras. Esses reticulados são gerados por um modelo SMT. Os experimentos foram realizados em 4 domínios distintos, e apresentaram melhoras na pontuação BLEU de 0,9 a 3,1, comparado com modelos que não usam os reticulados.

3.3 Abordagem adotada neste trabalho: UNdreaMT

Considerando as técnicas e abordagens apresentadas até aqui, este trabalho pretende investigar a aplicação proposta por Artetxe et al. (2018). No trabalho, os autores propõem um método que treina uma NMT de forma totalmente não-supervisionada e apenas em córpus monolíngue. A proposta é interessante do ponto de vista acadêmico porque mescla diversas das estratégias apresentadas anteriormente como detalhado a seguir.

3.3.1 Estrutura

Quanto à estrutura aplicada, o trabalho de Artetxe et al. (2018) é baseado em uma estrutura padrão *encoder-decoder* com um mecanismo de atenção (BAHDANAU; CHO; BENGIO, 2014). Para o *encoder* e *decoder* são usadas duas RNNs bidirecionais de duas camadas. Como entrada do modelo são usadas *word embeddings* de 300 dimensões, e o mecanismo de atenção usado é o método global proposto por Luong, Pham e Manning (2015), já ilustrado pela Figura 6. Entretanto, o método proposto se diferencia de um modelo NMT padrão em três aspectos principais, que proporcionam ao modelo a capacidade de ser treinado de forma totalmente não-supervisionada:

- **Estrutura dupla** – Sistemas NMT são comumente construídos para uma direção específica de tradução no par de idiomas. Na proposta, os autores usam uma estrutura capaz de realizar a tradução nas duas direções.

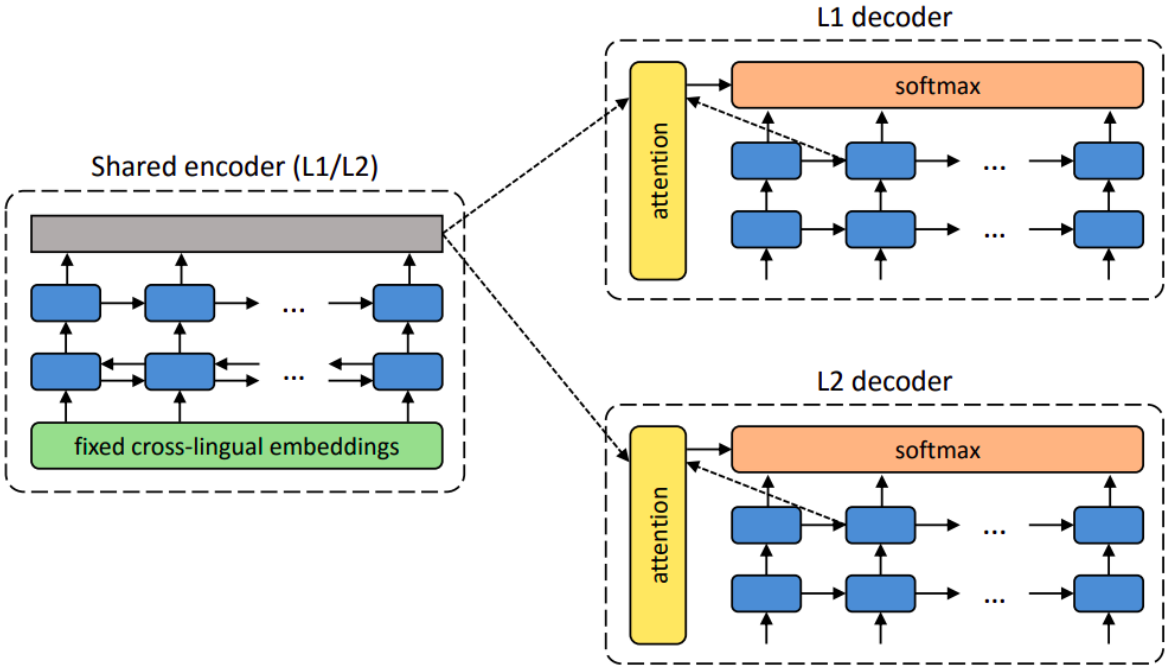
- ***Encoder* compartilhado** – A proposta faz uso de um único *encoder*, que é compartilhado para ambos os idiomas. Esse *encoder* é usado a fim de produzir uma representação da sentença de entrada que seja independente do idioma, por isso faz uso de *word embeddings* em um espaço compartilhado.

[9] Estrutura matemática abstrata que consiste de um conjunto parcialmente ordenado onde cada elemento apresenta um supremo e um ínfimo.

- ***Embeddings* de tamanho fixo no *encoder*** – Na maioria dos sistemas NMT, os vetores de palavra de entrada são inicializados de acordo com a entrada, e atualizados conforme o treinamento. No método proposto, são usadas *embeddings* de tamanho fixo.

A Figura 10 apresenta um diagrama ilustrando a estrutura do sistema proposto pelos autores. Assim, o modelo apresenta as 3 redes neurais usadas: (i) *encoder* compartilhado, (ii) *decoder* do idioma L1, e (iii) *decoder* do idioma L2.

Figura 10 – Arquitetura do sistema proposto por Artetxe et al. (2018). O *encoder* compartilhado alimenta os *decoders* de ambos os idiomas – L1 e L2. O sistema treina, então, 4 modelos simultaneamente: ao receber uma sentença no idioma L1, o vetor de contexto gerado pelo *encoder* compartilhado envia sua saída para o *decoder* de L1 (treinando o tradutor L1-L1), e para o *decoder* de L2 (treinando o tradutor L1-L2); analogamente, recebendo sentença no L2 envia para os *decoders* de L1 (treinando o tradutor L2-L1) e L2 (treinando o tradutor L2-L2)



Fonte: (ARTETXE et al., 2018)

Como a proposta visa o treinamento usando apenas córpus monolíngue de forma não-supervisionada, os autores usam duas estratégias para suprir a falta de córpus paralelo.

A primeira é a Remoção de Ruídos. Segundo os autores, ao usar um *encoder* compartilhado e uma estrutura dupla, o sistema proposto pode ser treinado para reconstruir na saída a própria entrada recebida, similar ao *autoencoder*. O sistema pode ser otimizado para receber uma sentença de entrada em um idioma, codificá-la usando o *encoder* compartilhado, e reconstruir a sentença original usando o *decoder* do idioma escolhido. Como são usadas *embeddings* pré treinadas como entrada do *encoder* compartilhado, o *encoder* deve ser capaz de aprender a compor as *embeddings* de ambos os idiomas de uma forma que ela seja inde-

pendente de idioma. E, ainda, cada *decoder* deve ser capaz de decompor as *embeddings* em representações correspondentes em seu idioma. Tendo isso em vista, esse comportamento é prejudicado pelo fato de que o processo resultante se torna uma simples tarefa de cópia. Dessa forma, a solução proposta pelos autores para esse problema é a adição de ruído aleatório nas sentenças de entrada. O princípio é similar à remoção de ruídos de *autoencoders* (VINCENT et al., 2010), onde o sistema é treinado para reconstruir a versão original de uma sentença de entrada corrompida. Dessa forma, a proposta dos autores é a alternação da ordem das palavras da sentença de entrada realizando trocas aleatórias entre palavras consecutivas. Explicitamente, para uma sequência de N elementos, são realizadas $N/2$ trocas desse tipo. Assim, os autores pretendem que o sistema seja capaz de aprender a estrutura interna dos idiomas envolvidos para que seja possível a reconstrução da ordem correta das palavras. Adicionalmente, ao fazer com que o sistema se baseie menos na ordem da sequência de entrada, o modelo pode sofrer menos com a divergência da ordem das palavras nos idiomas.

A segunda estratégia usada para suprir a falta de córpus paralelo é a de *back-translation*. A diferença aqui é que esta é feita em tempo de treinamento. Os autores adaptam o *back-translation* tradicional de tal forma que: dada uma sentença de entrada em um idioma, o sistema usa o *encoder* compartilhado e aplica o *decoder* do outro idioma para criar uma sentença paralela sintética. Essa adaptação do método de *back-translation* se diferencia do tradicional quanto ao córpus uma vez que, enquanto a tradução sintética do tradicional é feita anteriormente ao processo de treinamento, esse novo método permite que a qualidade da tradução sintética seja aprimorada conforme o modelo é treinado.

3.3.2 Resultados reportados no trabalho original

Os experimentos foram conduzidos nos pares de idiomas francês-inglês e alemão-inglês, com córpus do domínio de notícias jornalísticas. Para o francês foram usados aproximadamente 749 milhões de *tokens*, para o alemão 1.606 milhões de *tokens*, e para o inglês 2.238 milhões de *tokens*. Os autores experimentaram 3 estratégias diferentes:

1. **Treinamento não supervisionado** – essa é a proposta principal do trabalho, na qual o treinamento usa apenas córpus monolíngue;

2. **Treinamento semi-supervisionado** – para comparar com cenários onde uma pequena quantidade de córpus paralelo do domínio específico se encontra disponível, os autores realizaram o treinamento de um modelo com 10 mil sentenças paralelas de domínio específico e outro com 100 mil;

3. **Treinamento supervisionado** – esse é o cenário tradicional da NMT, no qual estão disponíveis grandes quantidades de textos para o treinamento, e esse modelo é usado como o limite superior que o modelo não-supervisionado pode atingir.

A essas variantes de estratégias, somam-se aos modelos também a presença apenas da remoção de ruídos, ou apenas de *back-translation*.

Os autores atestam que, considerando que o modelo proposto foi treinado exclusivamente com córpus monolíngue, a abordagem obtém resultados satisfatórios nesse cenário. O modelo treinado atinge de 14 a 15 pontos BLEU para o par francês-inglês e 6 a 10 pontos BLEU no par alemão-inglês, dependendo da variante do modelo. Os resultados também demonstram que a técnica de *back-translation* é essencial para que o modelo apresente resultados melhores, e demonstra uma vantagem superior ao uso da remoção de ruídos.

Adicionalmente, os resultados do sistema semi-supervisionado demonstram que o modelo proposto se beneficia muito de um córpus paralelo bilíngue, mesmo que reduzido. O experimento com uso de 10 mil sentenças paralelas demonstrou um aumento de 3 pontos BLEU no par francês-inglês e de 0,97 a 1,30 no par alemão-inglês. Já com o uso de 100 mil sentenças paralelas, o aumento foi de 6,25 a 7,38 no par francês-inglês e 4,06 a 5,08 no par alemão-inglês. A melhora proporcionada pelo uso do método proposto pelos autores em conjunto com um córpus paralelo supera até mesmo o modelo supervisionado, treinado em um córpus volumoso, em até 1,85 pontos BLEU. O modelo treinado com o método proposto e com a adição de 100 mil sentenças paralelas, ao ser comparado com o modelo supervisionado, obteve uma melhor pontuação BLEU, sendo 21,74 pontos e 19,89 pontos, respectivamente[10].

Considerando a variedade de estratégias aplicadas e os resultados obtidos por Artetxe et al. (2018), o presente trabalho visou seguir a mesma abordagem e investigar seu comportamento no domínio do *e-commerce*, geralmente mais desafiador de ser trabalhado, se comparado com o usado no trabalho original e na maioria dos trabalhos citados nesta seção: o de notícias jornalísticas.

[10] Comparação realizada no par de idiomas francês-inglês, apenas da direção inglês → francês. Para a direção francês → inglês a diferença comparado com o sistema supervisionado foi de 1,33 pontos BLEU.

4 MATERIAIS

A seguir são descritos os materiais utilizados nesse trabalho: o córpus (na seção 4.1) e as ferramentas auxiliares (na seção 4.2).

4.1 Córpus

O córpus utilizado para treinamento dos modelos de tradução neural investigados neste trabalho é composto por textos de domínio geral e de domínio específico.

O córpus de domínio geral é formado por 2 coleções de textos paralelos (inglês e português): (i) o córpus FAPESP[1] (AZIZ; SPECIA, 2011), que é composto por textos extraídos da revista *Pesquisa FAPESP*; (ii) alguns manuais de softwares, cedidos pela B2W Digital.

Já o córpus de domínio específico é a reunião de títulos de produtos de alguns sites de *e-commerce*, sendo um conjunto de títulos paralelos e dois conjuntos monolíngues (um inglês e outro português). Um *crawler*[2] foi desenvolvido a fim de coletar os títulos dos produtos. O primeiro conjunto de pares de títulos paralelos foram coletados da Amazon brasileira[3] (português) e pareados com os títulos em inglês da Amazon internacional[4] (inglês). O segundo conjunto foi coletado do ebay[5], tanto os títulos em português como os em inglês. Os dois conjuntos monolíngues, nos dois idiomas, foram cedidos pela B2W Digital.

Também foi utilizado um córpus de teste, composto por títulos paralelos em inglês e português para a avaliação do modelo de tradução, com títulos retirados da Amazon Brasileira e da Internacional.

Vale salientar que o pré-processamento aplicado ao córpus realiza os passos tradicionais de: (i) exclusão de pontuação - com exceção do ponto em números decimais - e símbolos fora do padrão alfanumérico, (ii) remoção de acentos e (iii) conversão do texto para letras minúsculas. A contagem de *tokens* foi realizada após o pré-processamento. A Tabela 1 resume as quantidades de sentenças e *tokens* presentes nas versões de treinamento e de teste das coleções que compõem o córpus.

O treinamento dos modelos de tradução, descritos no Capítulo 5, utilizam em parte ou na totalidade o conjunto de textos aqui mencionados.

---

[1] Disponível em:<http://www.nilc.icmc.usp.br/nilc/tools/Fapesp%20Corpora.htm>. Acesso em: 08 de novembro de 2019.

[2] Algoritmo cuja função é analisar o código de páginas web e extrair as informações requeridas.

[3] Disponível em:<https://www.amazon.com.br/>. Acesso em: 08 de novembro de 2019.

[4] Disponível em:<https://www.amazon.com/>. Acesso em: 08 de novembro de 2019.

[5] Disponível em:<https://www.ebay.com/>. Acesso em: 08 de novembro de 2019.

Tabela 1 – Quantidade de sentenças e de *tokens* nas versões de treinamento e de teste do córpus de domínio geral e específico

	Quantidade sentenças		Quantidade *tokens*	
	pt	en	pt	en
Treinamento (domínio geral)				
Paralelo - FAPESP	160.975	160.975	3.564.654	3.837.924
Paralelo - B2W	258.880	258.880	2.070.223	2.374.051
Total	**419.855**	**419.855**	**5.634.877**	**6.211.975**
Treinamento (domínio específico)				
Paralelo - Amazon	27.892	27.892	285.779	252.527
Paralelo - ebay	29.566	29.566	339.599	382.200
Monolíngue - B2W	318.816	372.386	5.530.874	3.569.235
Total	**376.274**	**429.844**	**6.156.252**	**4.203.962**
Teste (domínio específico)				
B2W	1.256	1.256	13.965	10.986

4.2 Ferramentas

Algumas ferramentas foram utilizadas para a realização deste trabalho a fim de auxiliar nas etapas de coleta, pré-processamento, geração de *word embedding*, treinamento e teste.

Na primeira etapa, de coleta, foi desenvolvido um *crawler*, como mencionado, utilizando a ferramenta *Scrapy*[6] através da linguagem Python. *Scrapy* é um *framework* de código aberto para extração de dados a partir de *websites*. O pré-processamento dos textos obtidos foi realizado com o auxílio da ferramenta NLTK[7](BIRD; LOPER; KLEIN, 2009). NLTK é um conjunto de bibliotecas, programas e dados, implementado em Python, usado em aplicações de PLN.

Para treinar os modelos de tradução, como orientado por Artetxe et al. (2018), *word embeddings* precisaram ser geradas para o córpus, nos dois idiomas. Para tal, foi utilizada a

[6] Disponível em:<https://scrapy.org/>. Acesso em: 08 de novembro de 2019.
[7] Disponível em:<https://www.nltk.org/>. Acesso em: 08 de novembro de 2019.

ferramenta Gensim[8](ŘEHŮŘEK; SOJKA, 2010), uma biblioteca que contém diversas funções úteis para PLN. Foi implementada em Python e Cython e desenvolvida visando o eficiente manuseio de grandes coleções de textos. A ferramenta oferece uma aplicação do método desenvolvido por Mikolov, Yih e Zweig (2013) através do *word2vec*. Assim, as WE utilizadas nesse trabalho foram treinadas pelo método *word2vec* implementado pela ferramenta *Gensim*.

Após o treinamento das *word embeddings*, foi necessário o mapeamento delas para um espaço vetorial compartilhado, como descrito na seção 2.2. A ferramenta utilizada foi a Vecmap[9] (ARTETXE; LABAKA; AGIRRE, 2018b; ARTETXE; LABAKA; AGIRRE, 2018a; ARTETXE; LABAKA; AGIRRE, 2017; ARTETXE; LABAKA; AGIRRE, 2016). Este, é um *framework* de código aberto usado para aprender o mapeamento interlinguístico de *word embeddings* através de um treinamento não-supervisionado. É uma ferramenta desenvolvida pelos mesmos autores do trabalho Artetxe et al. (2018), objeto de investigação do presente trabalho.

Finalmente, para a aplicação e treinamento dos recursos coletados, foi empregada a abordagem proposta por Artetxe et al. (2018), através da ferramenta UNdreaMT[10](ARTETXE et al., 2018). A ferramenta é o resultado do trabalho mencionado, sendo uma implementação em Python do modelo não-supervisionado de NMT segundo os métodos propostos. A ferramenta se propõe a produzir traduções usando apenas córpus monolíngue nas duas línguas, mas demonstra que a adição de um córpus paralelo, mesmo que relativamente pequeno, impulsiona muito os resultados. O trabalho foi descrito com maior profundidade na seção 3.3.

[8] Disponível em:<https://radimrehurek.com/gensim/>. Acesso em: 08 de novembro de 2019.
[9] Disponível em:<https://github.com/artetxem/vecmap>. Acesso em: 10 de novembro de 2019.
[10] Disponível em:<https://github.com/artetxem/undreamt>. Acesso em: 15 de novembro de 2019.

5 EXPERIMENTOS E AVALIAÇÃO

Esse capítulo descreve os experimentos realizados a fim de comprovar ou refutar a hipótese apresentada para este trabalho de que: as estratégias de adaptação de domínio propostas por Artetxe et al. (2018) podem ser aplicadas para melhorar a qualidade da tradução automática neural de inglês para português, no domínio do *e-commerce*, quando comparada à tradução gerada por um modelo tradicional de domínio geral.

Para tanto, 3 modelos foram treinados seguindo as estratégias propostas por Artetxe et al. (2018):

1. **Modelo não-supervisionado** – O primeiro modelo foi treinado usando o córpus completo. Contudo, o paralelismo das sentenças foi descartado ao trocar a ordem das sentenças de forma aleatória e, nas configurações da ferramenta UNdreaMT, não usar a opção de córpus paralelo. Assim, esse modelo foi treinado com 429.844 sentenças em inglês, 376.274 sentenças em português – ambas monolíngues de domínio específico –, e dois conjuntos de domínio geral monolíngues, um em inglês e o outro em português, com 419.855 sentenças cada. Esse modelo pretendia analisar a proposta original de Artetxe et al. (2018), baseando-se apenas em córpus monolíngue nas duas linguagens. Esse modelo também faz uso dos melhoramentos demonstrados por Sajjad et al. (2017) ao usar um córpus de domínio geral concatenado ao de domínio específico.

2. **Modelo supervisionado de domínio geral** – O segundo modelo foi treinado a partir do córpus de domínio geral apenas, considerando o paralelismo das sentenças. Ao todo, 419.855 sentenças paralelas bilíngues de domínio geral. Esse modelo visou a investigação de como a ausência de um córpus de domínio específico no treinamento pode influenciar a qualidade do modelo. Apesar de usar, aproximadamente, apenas metade do córpus utilizado nos outros modelos, esse experimento assume que o uso de sentenças paralelas traduzidas por humanos possa compensar o córpus reduzido e ser um modelo com resultados comparáveis aos demais.

3. **Modelo semi-supervisionado** – Para o terceiro modelo, todo o córpus disponível foi utilizado. Ou seja, 419.855 sentenças paralelas de domínio geral, e 57.458 sentenças paralelas de domínio específico, além das 372.386 sentenças monolíngues do inglês e 318.816 do português. O objetivo desse modelo era analisar a aplicação do método proposto considerando todo o córpus coletado. Esse modelo segue a mesma proposta semi-supervisionada apresentada em Artetxe et al. (2018), porém com o córpus coletado aqui descrito.

Os 3 modelos foram treinados usando as mesmas configurações daqueles que obtiveram melhores resultados no trabalho investigado, ou seja, todos usam *back-translation* com a remoção de ruídos.

Como a escassez de recursos no domínio do *e-commerce* faz parte do foco deste trabalho, o volume do córpus aplicado nos experimentos é relativamente pequeno, se comparado a outros trabalhos mencionados aqui. Assim, esse trabalho também investiga como o método se comporta em cenários com volume reduzido de conjuntos de texto.

Os modelos foram treinados na ferramenta UNdreaMT usando as configurações pré-definidas. O treinamento das *word embeddings* foi realizado conforme descrito pelos autores do UNdreaMT nos experimentos conduzidos por eles, com WE de dimensão 300. O mapeamento das WE para um espaço vetorial compartilhado também foi realizado usando as configurações estabelecidas pela ferramenta VecMap. A duração do treinamento de cada modelo variou de 3 (modelo 2) a 6 dias (modelo 3) no *hardware* disponível: com uma CPU intel core i7, memória RAM de 8GB e uma GPU GTX 1070 com 8GB.

5.1 Resultados

Para análise quantitativa dos resultados, foram empregadas as métricas BLEU, NIST e METEOR na tradução do conjunto de teste selecionado. A Tabela 2 sintetiza os valores de todos os modelos que foram treinados nesse trabalho. Além dos valores das medidas calculadas para cada modelo, também são apresentados os valores obtidos por (i) a tradução gerada pelo Google Tradutor[1] e (ii) a tradução gerada pelo DeepL[2]. Os dois últimos são tradutores de domínio geral bastante utilizados na atualidade, e as medidas foram calculadas com base no mesmo córpus de teste usado nos modelos treinados.

Com os resultados, estima-se que o baixo desempenho do tradutor deu-se principalmente em razão da baixa qualidade e volume do córpus usado. Contudo, comparando os valores obtidos em cada medida, o resultado obtido para o modelo 3 não é tão inferior ao Google Tradutor e ao DeepL. Considerando que esses últimos são treinados em conjuntos de dados com pelo menos dezenas de milhões de sentenças, o modelo treinado ainda apresentou um resultado considerado razoável, e os baixos valores obtidos por esses sistemas de maior escala, porém de domínio geral, evidenciam ainda mais os desafios que o domínio do *e-commerce* apresenta para o PLN.

A Figura 11 apresenta alguns exemplos de sentenças traduzidas pelos três modelos treinados, além das traduções geradas pelo Google Tradutor e pelo DeepL, bem como a tradução referência fornecida pela B2W Digital através do córpus de teste.

É possível observar a clara disparidade entre os modelos nas sentenças selecionadas. Na

[1] Disponível em: <https://translate.google.com.br/>. Acesso em: 04 dez. 2019.
[2] Disponível em: <https://www.deepl.com/>. Acesso em: 04 dez. 2019.

Tabela 2 – Resultados das medidas avaliativas aplicadas: BLEU[a], NIST e METEOR. Compara-se os 3 modelos treinados, e traduções feitas pelo Google Tradutor e a ferramenta DeepL.

	BLEU	BLEU-1	BLEU-2	BLEU-3	NIST	METEOR
Modelo 1	3,53	22,04	13,74	7,16	2,4033	0,10921
Modelo 2	2,11	15,32	8,59	4,48	1,5021	0,09100
Modelo 3	5,81	27,70	17,18	10,05	2,7568	0,16871
DeepL	7,61	28,01	17,18	11,01	3,0724	0,21977
Google Tradutor	10,44	32,05	21,23	14,34	3,5670	0,25037

[a] As medidas BLEU-1, BLEU-2, e BLEU-3 representam a medida BLEU usando unigrama, bigrama e trigrama, respectivamente. Enquanto a BLEU simboliza o que seria a BLEU-4, que é a medida padrão.

figura, o modelo 1 demonstra resultados desconexos tanto com a sentença original como entre as palavras da própria tradução. Presume-se que isso se dá em razão da falta de paralelismo do córpus usado no treinamento (uma vez que este é o modelo não-supervisionado), unido à pequena quantidade de sentenças disponíveis no treinamento, que resulta em um mapeamento de *word embedding* pouco confiável.

Já no modelo 2, é notável a ausência de palavras no domínio específico, resultado da privação de córpus de domínio específico no treinamento (uma vez que este é o modelo supervisionado de domínio geral). Apesar disso, as sentenças ainda apresentam certa fluidez, mesmo que não adequadas à sentença original. Acredita-se que o paralelismo do córpus seja o principal responsável por essa fluidez, visto que garante um mapeamento de *word embeddings* mais preciso. Assim, mesmo que a tradução não corresponda à sentença original, ela ainda é mais coesa que a sentença traduzida pelo modelo 1, apesar de este conter palavras do domínio específico, por isso a pontuação superior na medida BLEU.

Quanto ao terceiro modelo, ele apresenta uma melhor qualidade na tradução tanto em termos de coesão quanto do vocabulário de domínio específico nos exemplos apresentados. Contudo, apesar de superior aos outros modelos, pela pontuação na medida BLEU é possível inferir que a qualidade geral da tradução continua longe de uma aplicação prática.

A Figura 12 ilustra algumas das traduções feitas pelo modelo 3 que impedem que ele seja aplicado em um cenário real. O primeiro exemplo da figura mostra uma tradução onde uma característica primordial do objeto vendida foi trocada por outra de igual valor semântico, porém que modifica o produto vendido. Neste exemplo, a película protetora para um modelo específico de *smartphone* foi trocada pelo modelo de outro *smartphone*, mudando, assim, a

Figura 11 – Exemplos de traduções de sentenças de inglês (original) para português usando os 3 modelos, bem como sentenças para comparação: referência (considerada a tradução correta) e geradas por tradutores de domínio geral amplamente utilizados na atualidade.

Original	*barbie careers doctor doll*
Referência	boneca barbie profissoes doutora
Google Tradutor	barbie carreiras médico boneca
DeepL	barbie carreiras doutor boneca
Modelo 1	meus meus meus steps
Modelo 2	carrinho de cama classic pecas
Modelo 3	boneca barbie profissoes

Original	*3m privacy filter for 22 diagonal widescreen monitor protects your confidential information black out side views 16 10 pf220w1b*
Referência	filtro de privacidade 3m filtros de privacidade e de tela para notebooks
Google Tradutor	O filtro de privacidade de 3m para o monitor widescreen 22 diagonal protege suas informações confidenciais ocultar as vistas laterais 16 10 pf220w1b
DeepL	3m filtro de privacidade para 22 diagonal widescreen monitor protege suas informações confidenciais black out side views 16 10 10 pf220w1b
Modelo 1	fita adesiva para filtro de agua 12 polegadas para monitorar todos os dados
Modelo 2	fita digital para filtro para
Modelo 3	filtro de privacidade 3m 18 polegadas 22 polegadas 22 polegadas protege a sua informacao

Original	*jansport right pack*
Referência	mochila jansport right pack
Google Tradutor	pacote certo jansport
DeepL	jansport right pack
Modelo 1	12 pack
Modelo 2	mochila escolar direito
Modelo 3	mochila jansport right pack

natureza do produto. No segundo exemplo vê-se uma tradução onde foram inseridas duas marcas concorrentes que vendem o mesmo produto (canon e nikon). E no terceiro, a natureza do produto é totalmente perdida. Esses erros de tradução não são apenas ruins para o modelo,

mas, ao serem aplicados em cenários reais, gera uma série de problemas legais sobre o uso desta tradução.

Figura 12 – Exemplos de traduções de qualidade insatisfatória geradas pelo Modelo 3.

Original	*amfilm glass screen protector for samsung galaxy s9 plus 3d curved tempered glass dot matrix with easy installation tray case friendly black*
Referência	pelicula de vidro temperado 3d curva para samsung galaxy s9 plus
Modelo 3	vidro temperado tela vidro temperado para samsung galaxy j7 prime g610

Original	*canon 90d digital slr camera body only*
Referência	camera canon eos 90d dslr corpo
Modelo 3	cameras canon digital nikon coolpix

Original	*case logic slrc 206 slr camera and 15.4 inch laptop backpack black*
Referência	mochila p camera laptop slr case logic slrc206 case logic acessorios para cameras digitais preta
Modelo 3	bateria para notebook

Ainda vale ressaltar que os córpus utilizados, tanto para treinamento como para teste, apresentavam ruídos. Algumas das traduções presentes não eram traduções diretas, mas descrições do mesmo produto de forma a não ser uma tradução propriamente paralela. Outro problema com o córpus foi a não tradução de algumas das sentenças do córpus paralelo (quando o título do produto na loja nacional é uma cópia do título em inglês, ou vice-versa para produtos brasileiros vendidos na loja internacional). Os dois problemas são justificados pelo fato de o domínio do *e-commerce* ser bastante heterogêneo (a estrutura da sentença varia muito entre os *websites*), e conseguir um córpus de qualidade torna-se bastante difícil. Entretanto, mesmo com essa configuração de córpus, os resultados se mostraram razoavelmente satisfatórios, comparados a sistemas bem estabelecidos no mercado atual.

Assim, a partir dos experimentos realizados com as estratégias propostas por Artetxe et al. (2018) para a tradução neural com adaptação de domínio e os resultados apresentados aqui, pode-se concluir que, no que se refere ao modelo semi-supervisionado, a hipótese de pesquisa foi comprovada, ou seja: o modelo semi-supervisionado (modelo 3) proposto por Artetxe et al. (2018), quando aplicado na tradução inglês-português no domínio do *e-commerce*, teve resultado superior ao modelo tradicional (supervisionado, modelo 2) de domínio geral.

6 CONSIDERAÇÕES FINAIS

Ainda que diversos trabalhos tenham sido desenvolvidos para a realização da tarefa de Tradução Automática (TA), ainda não é possível obter automaticamente traduções completamente corretas. Essa tarefa torna-se ainda mais desafiadora quando são considerados domínios pouco convencionais, como o do *e-commerce*.

Com o surgimento da tradução automática neural (NMT), e desde que foi considerada o estado da arte na TA, abordagens vêm sendo propostas para o problema de adaptação de domínio.

Neste trabalho, investigou-se a aplicação das abordagens propostas por Artetxe et al. (2018), que empregam diversas técnicas interessantes a serem testadas no domínio proposto. O trabalho investigou a aplicação de *back-translation*, *word embeddings* e variação do córpus de treinamento, além de usar uma estrutura de rede neural que se diferencia da convencional *encoder-decoder* em diversos aspectos.

Empregando as técnicas e métodos utilizados por Artetxe et al. (2018), o presente trabalho treinou 3 modelos de tradutores distintos: (i) modelo não-supervisionado, (ii) modelo supervisionado de domínio geral, e (iii) modelo semi-supervisionado. Os experimentos foram conduzidos no par de idiomas inglês-português, no domínio do *e-commerce*. Os modelos foram treinados tanto com córpus paralelo de domínio geral – com cerca de 400 mil pares de sentenças – quanto em córpus de domínio específico não paralelo – também com cerca de 400 mil sentenças em cada idioma – e paralelo – com cerca de 57 mil pares de sentenças. O conjunto de domínio geral foi composto pelo córpus FAPESP (AZIZ; SPECIA, 2011) e por um conjunto de sentenças fornecidas pela B2W (apoiadora deste trabalho). Já o conjunto de textos de domínio específico é composto por títulos retirados de páginas de *e-commerce* muito utilizadas atualmente: Amazon brasileira[1] (para o córpus em português), Amazon internacional[2] (para o córpus em inglês), ebay[3] (córpus em inglês e português), além de um conjunto de títulos disponibilizados pela B2W.

Os modelos treinados foram, então, comparados e, adicionalmente, dois sistemas de tradução muito utilizados na atualidade também foram considerados: Google Tradutor[4] e DeepL[5]. Os resultados apontam que o melhor modelo foi o semi-supervisionado (modelo 3) que utiliza tanto córpus paralelo de domínio geral, como córpus de domínio específico (paralelo e não paralelo) no treinamento.

---

[1] Disponível em:<https://www.amazon.com.br/>. Acesso em: 08 de novembro de 2019.
[2] Disponível em:<https://www.amazon.com/>. Acesso em: 08 de novembro de 2019.
[3] Disponível em:<https://www.ebay.com/>. Acesso em: 08 de novembro de 2019.
[4] Disponível em: <https://translate.google.com.br/>. Acesso em: 04 dez. 2019.
[5] Disponível em: <https://www.deepl.com/>. Acesso em: 04 dez. 2019.

A partir dos experimentos e resultados apresentados neste trabalho, pode-se concluir que a hipótese de pesquisa foi comprovada, ou seja, o modelo semi-supervisionado proposto por Artetxe et al. (2018), quando aplicado na tradução inglês-português no domínio do *e-commerce* teve resultado superior ao modelo tradicional de domínio geral.

6.1 Contribuições desta Pesquisa

Como principal contribuição desta pesquisa tem-se a investigação das técnicas apresentadas por Artetxe et al. (2018) e seu desempenho como modelo de adaptação de domínio para o *e-commerce*. Até onde se sabe, este é primeiro trabalho a abordar a adaptação de domínio na tradução neural para o *e-commerce*, ao menos no que se refere ao português do Brasil.

Outra contribuição deste trabalho são os três modelos treinados, com especial interesse para o modelo semi-supervisionado, que obteve os melhores resultados dentre os outros modelos desenvolvidos nesta pesquisa.

Além das contribuições mencionadas, a principal contribuição científica deste trabalho está relacionada à comprovação da hipótese de pesquisa, que o uso de *back-translation*, *word embeddings* e uma nova estrutura de tradutor, dentre outros métodos e abordagens usados em Artetxe et al. (2018), podem melhorar a tradução automática neural de títulos de produtos do domínio do *e-commerce*.

6.2 Trabalhos Futuros

Algumas propostas de trabalhos futuros decorrem desta pesquisa. A principal está relacionada aos recursos utilizados no treinamento do tradutor. Acredita-se, com base nos valores obtidos nos experimentos, que o desempenho do modelo 3 pode ser melhorado se um córpus maior e de melhor qualidade, no domínio específico, for utilizado. Considerando que o modelo 3, treinado com um córpus de domínio específico com menos de 500 mil sentenças para cada idioma, com ruídos, obteve um desempenho de metade do valor BLEU alcançado pelo tradutor do Google (de domínio geral, mas considerado o estado-da-arte); um refinamento desse conjunto de textos pode atingir resultados ainda melhores.

Como perspectiva para trabalhos futuros, também pode-se propor o uso de algumas técnicas de adaptação de domínio adicionais às propostas por Artetxe et al. (2018). Especialmente aquelas que se referem à modificação do córpus, como a abordagem multidomínio ou usando córpus paralelo fora do domínio.

Também é possível a investigação da aplicação de outros trabalhos da literatura para adaptação ao domínio do *e-commerce*, usando diferentes estruturas no sistema do tradutor.

REFERÊNCIAS

ARTETXE, M.; LABAKA, G.; AGIRRE, E. Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics, 2016. p. 2289–2294. Disponível em: <https://www.aclweb.org/anthology/D16-1250>. Citado 2 vezes nas páginas 25 e 45.

ARTETXE, M.; LABAKA, G.; AGIRRE, E. Learning bilingual word embeddings with (almost) no bilingual data. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics, 2017. p. 451–462. Disponível em: <https://www.aclweb.org/anthology/P17-1042>. Citado 2 vezes nas páginas 25 e 45.

ARTETXE, M.; LABAKA, G.; AGIRRE, E. Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations. 2018. Disponível em: <https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16935/16781>. Citado 2 vezes nas páginas 25 e 45.

ARTETXE, M.; LABAKA, G.; AGIRRE, E. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, 2018. p. 789–798. Disponível em: <https://www.aclweb.org/anthology/P18-1073>. Citado 2 vezes nas páginas 25 e 45.

ARTETXE, M. et al. Unsupervised neural machine translation. In: *Proceedings of the Sixth International Conference on Learning Representations*. [S.l.: s.n.], 2018. Citado 16 vezes nas páginas 14, 15, 20, 23, 24, 25, 38, 39, 40, 42, 44, 45, 47, 51, 53 e 54.

AXELROD, A.; HE, X.; GAO, J. Domain adaptation via pseudo in-domain data selection. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2011. (EMNLP '11), p. 355–362. ISBN 978-1-937284-11-4. Disponível em: <http://dl.acm.org/citation.cfm?id=2145432.2145474>. Citado na página 31.

AZIZ, W.; SPECIA, L. Fully automatic compilation of a Portuguese-English parallel corpus for statistical machine translation. In: *STIL 2011*. Cuiabá, MT: [s.n.], 2011. Citado 2 vezes nas páginas 43 e 53.

BAHDANAU, D.; CHO, K.; BENGIO, Y. Neural machine translation by jointly learning to align and translate. In: . [s.n.], 2014. abs/1409.0473. Disponível em: <http://arxiv.org/abs/1409.0473>. Citado 2 vezes nas páginas 17 e 39.

BARONE, A. V. M. et al. Regularization techniques for fine-tuning in neural machine translation. In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, 2017. p. 1489–1494. Disponível em: <https://www.aclweb.org/anthology/D17-1156>. Citado na página 31.

BENGIO, Y. et al. A neural probabilistic language model. *JMLR - Journal of Machine Learning Research*, p. 1137–1155, 2003. Citado na página 23.

BIRD, S.; LOPER, E.; KLEIN, E. *Natural Language Processing with Python*. [S.l.]: O'Reilly Media Inc., 2009. Citado na página 44.

BOJAR RAJEN CHATTERJEE, C. F. Y. G. B. H. S. H. M. H. P. K. Q. L. V. L. C. M. M. N. M. P. R. R. L. S. O.; TURCHI, M. Findings of the 2017 conference on machine translation (wmt17). In: *Proceedings of the Second Conference on Machine Translation*. [S.l.]: Association for Computational Linguistics, 2017. v. 1, p. 169—-214. Citado na página 12.

BRITZ, D.; LE, Q.; PRYZANT, R. Effective domain mixing for neural machine translation. In: *Proceedings of the Second Conference on Machine Translation*. Copenhagen, Denmark: Association for Computational Linguistics, 2017. p. 118–126. Disponível em: <https://www.aclweb.org/anthology/W17-4712>. Citado 2 vezes nas páginas 31 e 38.

CHEN, B. et al. Cost weighting for neural machine translation domain adaptation. In: *Proceedings of the First Workshop on Neural Machine Translation*. Vancouver: Association for Computational Linguistics, 2017. p. 40–46. Disponível em: <https://www.aclweb.org/anthology/W17-3205>. Citado na página 31.

CHEN, B. et al. Bilingual methods for adaptive training data selection for machine translation. In: *Proc. of AMTA*. [S.l.: s.n.], 2016. p. 93–103. Citado na página 31.

CHENG, Y. et al. Semi-supervised learning for neural machine translation. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics, 2016. p. 1965–1974. Disponível em: <https://www.aclweb.org/anthology/P16-1185>. Citado 2 vezes nas páginas 31 e 32.

CHO, K. et al. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2014. p. 1724–1734. Disponível em: <http://aclweb.org/anthology/D14-1179>. Acesso em: 16 maio 2017. Citado na página 18.

CHO, K. et al. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. [S.l.]: Association for Computational Linguistics, 2014. v. 1, p. 1724–1734. Citado na página 17.

CHU, C. *Integrated Parallel Data Extraction from Comparable Corpora for Statistical Machine Translation*. 2015. Citado na página 31.

CHU, C.; DABRE, R.; KUROHASHI, S. An empirical comparison of domain adaptation methods for neural machine translation. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Vancouver, Canada: Association for Computational Linguistics, 2017. p. 385–391. Disponível em: <https://www.aclweb.org/anthology/P17-2061>. Citado 3 vezes nas páginas 31, 35 e 37.

COLLOBERT, R.; WESTON, J. A unified architecture for natural language processing: deep neural networks with multitask learning. In: COHEN, W. W.; MCCALLUM, A.; ROWEIS, S. T. (Ed.). *ICML*. ACM, 2008. (ACM International Conference

Proceeding Series, v. 307), p. 160–167. ISBN 978-1-60558-205-4. Disponível em: <http://dblp.uni-trier.de/db/conf/icml/icml2008.html#CollobertW08>. Citado na página 23.

CUONG, H.; SIMA'AN, K. Latent domain translation models in mix-of-domains haystack. In: *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. Dublin, Ireland: Dublin City University and Association for Computational Linguistics, 2014. p. 1928–1939. Disponível em: <https://www.aclweb.org/anthology/C14-1182>. Citado na página 31.

CURREY, A.; BARONE, A. V. M.; HEAFIELD, K. Copied monolingual data improves low-resource neural machine translation. In: *Proceedings of the Second Conference on Machine Translation*. Copenhagen, Denmark: Association for Computational Linguistics, 2017. p. 148–156. Disponível em: <https://www.aclweb.org/anthology/W17-4715>. Citado 2 vezes nas páginas 31 e 32.

DAKWALE, P.; MONZ, C. Fine-tuning for neural machine translation with limited degradation across in-and out-of-domain data. *Proceedings of the XVI Machine Translation Summit*, p. 117, 2017. Citado na página 31.

DENKOWSKI, M.; LAVIE, A. Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the sixth workshop on statistical machine translation*. [S.l.], 2011. p. 85–91. Citado 3 vezes nas páginas 26, 28 e 29.

DODDINGTON, G. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In: MORGAN KAUFMANN PUBLISHERS INC. *Proceedings of the second international conference on Human Language Technology Research*. [S.l.], 2002. p. 138–145. Citado 3 vezes nas páginas 26, 27 e 28.

DOMHAN, T.; HIEBER, F. Using target-side monolingual data for neural machine translation through multi-task learning. In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, 2017. p. 1500–1505. Disponível em: <https://www.aclweb.org/anthology/D17-1158>. Citado 2 vezes nas páginas 31 e 38.

DUGAST, L.; SENELLART, J.; KOEHN, P. Statistical post-editing on SYSTRAN's rule-based translation system. In: *Proceedings of the Second Workshop on Statistical Machine Translation*. Prague, Czech Republic: Association for Computational Linguistics, 2007. p. 220–223. Disponível em: <https://www.aclweb.org/anthology/W07-0732>. Citado na página 11.

DUH GRAHAM NEUBIG, K. S. K.; TSUKADA, H. Adaptation data selection using neural language models: Experiments in machine translation. In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*. [S.l.]: Association for Computational Linguistics, 2013. v. 2 : Short Papers, p. 678—683. Citado na página 12.

DUH, K. et al. Adaptation data selection using neural language models: Experiments in machine translation. In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Sofia, Bulgaria: Association for Computational Linguistics, 2013. p. 678–683. Disponível em: <https://www.aclweb.org/anthology/P13-2119>. Citado na página 31.

DURRANI, N. et al. Using joint models for domain adaptation in statistical machine translation. In: . [S.l.: s.n.], 2015. Citado na página 31.

FIRTH, J. A synopsis of linguistic theory 1930-1955. *Studies in linguistic analysis*, Oxford: Blackwell, p. 1–32, 1957. Citado na página 15.

FOSTER, G.; GOUTTE, C.; KUHN, R. Discriminative instance weighting for domain adaptation in statistical machine translation. In: *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Cambridge, MA: Association for Computational Linguistics, 2010. p. 451–459. Disponível em: <https://www.aclweb.org/anthology/D10-1044>. Citado na página 31.

FREITAG, M.; AL-ONAIZAN, Y. *Fast Domain Adaptation for Neural Machine Translation*. 2016. Citado 2 vezes nas páginas 31 e 37.

GUIDÈRE, M. Toward corpus-based machine translation for standard arabic. In: . [S.l.: s.n.], 2002. Citado na página 11.

GULCEHRE, C. et al. *On Using Monolingual Corpora in Neural Machine Translation*. 2015. Citado 5 vezes nas páginas 31, 32, 37, 38 e 39.

HARRIS, Z. Distributional structure. *Word*, v. 10, n. 23, p. 146–162, 1954. Citado na página 15.

HOCHREITER, S.; SCHMIDHUBER, J. Long short-term memory. In: . Cambridge, MA, USA: MIT Press, 1997. v. 9, n. 8, p. 1735–1780. ISSN 0899-7667. Disponível em: <http://dx.doi.org/10.1162/neco.1997.9.8.1735>. Citado na página 20.

IMAMURA, K.; SUMITA, E. Multi-domain adaptation for statistical machine translation based on feature augmentation. *Journal of Natural Language Processing*, v. 24, n. 4, p. 597–618, 2017. Citado na página 31.

KHAYRALLAH, H. et al. Neural lattice search for domain adaptation in machine translation. In: *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Taipei, Taiwan: Asian Federation of Natural Language Processing, 2017. p. 20–25. Disponível em: <https://www.aclweb.org/anthology/I17-2004>. Citado 2 vezes nas páginas 31 e 39.

KOBUS, C.; CREGO, J.; SENELLART, J. *Domain Control for Neural Machine Translation*. 2016. Citado 2 vezes nas páginas 31 e 38.

KOEHN, P. et al. Moses: Open source toolkit for statistical machine translation. In: *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2007. (ACL '07), p. 177–180. Disponível em: <http://dl.acm.org/citation.cfm?id=1557769.1557821>. Citado 2 vezes nas páginas 11 e 12.

KOEHN, P.; KNOWLES, R. Six challenges for neural machine translation. In: *Proceedings of the First Workshop on Neural Machine Translation*. [S.l.]: Association for Computational Linguistics, 2017. v. 1, p. 28–39. Citado na página 12.

LENCI, A. Distributional semantics in linguistic and cognitive research. *Italian journal of linguistics*, v. 20, n. 1, p. 1–31, 2008. Citado na página 15.

LUONG, M.-T.; MANNING, C. D. Stanford neural machine translation systems for spoken language domains. In: . [S.l.: s.n.], 2015. Citado 2 vezes nas páginas 31 e 37.

LUONG, M.-T.; PHAM, H.; MANNING, C. D. *Effective Approaches to Attention-based Neural Machine Translation*. 2015. Citado 4 vezes nas páginas 17, 21, 22 e 39.

MARIE, B.; FUJITA, A. Efficient extraction of pseudo-parallel sentences from raw monolingual data using word embeddings. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Vancouver, Canada: Association for Computational Linguistics, 2017. p. 392–398. Disponível em: <https://www.aclweb.org/anthology/P17-2062>. Citado na página 31.

MATSOUKAS, S.; ROSTI, A.-V. I.; ZHANG, B. Discriminative corpus weight estimation for machine translation. In: *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*. Singapore: Association for Computational Linguistics, 2009. p. 708–717. Disponível em: <https://www.aclweb.org/anthology/D09-1074>. Citado na página 31.

MIKOLOV, T. et al. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013. Citado 2 vezes nas páginas 23 e 24.

MIKOLOV, T. et al. Distributed representations of words and phrases and their compositionality. *CoRR*, abs/1310.4546, 2013. Disponível em: <http://arxiv.org/abs/1310.4546>. Citado 2 vezes nas páginas 25 e 26.

MIKOLOV, T.; YIH, W.-t.; ZWEIG, G. Linguistic regularities in continuous space word representations. In: *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. [S.l.: s.n.], 2013. p. 746–751. Citado na página 45.

MILLER, G. A. Wordnet: a lexical database for english. *Communications of the ACM*, ACM, v. 38, n. 11, p. 39–41, 1995. Citado na página 29.

MOORE, R. C.; LEWIS, W. Intelligent selection of language model training data. In: *Proceedings of the ACL 2010 Conference Short Papers*. Uppsala, Sweden: Association for Computational Linguistics, 2010. p. 220–224. Disponível em: <https://www.aclweb.org/anthology/P10-2041>. Citado na página 31.

NAKAZAWA SHOHEI HIGASHIYAMA, C. D. H. M. I. G. H. K. Y. O. G. N. T.; KUROHASHI, S. Overview of the 4th workshop on asian translation. In: *Proceedings of the 4th Workshop on Asian Translation (WAT2017)*. [S.l.]: Asian Federation of Natural Language Processing, 2017. v. 1, p. 1—-54. Citado na página 12.

OCH, F. J.; NEY, H. The alignment template approach to statistical machine translation. *Computational Linguistics*, v. 30, n. 4, p. 417–449, 2004. Disponível em: <https://doi.org/10.1162/0891201042544884>. Citado na página 11.

PAPINENI, K. et al. Bleu: A method for automatic evaluation of machine translation. In: *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2002. (ACL '02), p. 311–318. Disponível em: <http://dx.doi.org/10.3115/1073083.1073135>. Citado 2 vezes nas páginas 26 e 27.

PARK, J.; SONG, J.; YOON, S. *Building a Neural Machine Translation System Using Only Synthetic Parallel Data*. 2017. Citado 3 vezes nas páginas 31, 33 e 34.

PENNINGTON, J.; SOCHER, R.; MANNING, C. D. Glove: Global vectors for word representation. In: *EMNLP*. [S.l.: s.n.], 2014. v. 14, p. 1532–1543. Citado na página 23.

PONCELAS DIMITAR SHTERIONOV, A. W. G. M. d. B. W. A.; PASSBAN, P. Investigating backtranslation in neural machine translation. In: *Proceedings of the 21st Annual Conference of the European Association for Machine Translation*. [S.l.]: Association for Computational Linguistics, 2018. v. 1, p. 249–258. Citado na página 14.

POPOVIĆ, M. chrf: character n-gram f-score for automatic mt evaluation. In: *Proceedings of the Tenth Workshop on Statistical Machine Translation*. [S.l.: s.n.], 2015. p. 392–395. Citado na página 26.

PORTER, M. F. *Snowball: A language for stemming algorithms*. 2001. Citado na página 28.

ŘEHŮŘEK, R.; SOJKA, P. Software Framework for Topic Modelling with Large Corpora. In: *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Valletta, Malta: ELRA, 2010. p. 45–50. <http://is.muni.cz/publication/884893/en>. Citado na página 45.

ROUSSEAU, A. et al. Lium's systems for the iwslt 2011 speech translation tasks. In: *International Workshop on Spoken Language Translation (IWSLT) 2011*. [S.l.: s.n.], 2011. Citado na página 31.

SAJJAD, H. et al. *Neural Machine Translation Training in a Multi-Domain Scenario*. 2017. Citado 3 vezes nas páginas 31, 36 e 47.

SALTON, G. (Ed.). *The SMART Retrieval System Experiments in Automatic Document Processing*. Englewood Cliffs: Prentice-Hall: [s.n.], 1971. Citado na página 23.

SENNRICH, B. H. R.; BIRCH, A. Improving neural machine translation models with monolingual data. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. [S.l.]: Association for Computational Linguistics, 2016. v. 1: Long Papers, p. 86—-96. Citado 4 vezes nas páginas 14, 31, 33 e 37.

SENNRICH, H. S. R.; ARANSA, W. A multi-domain translation model framework for statistical machine translation. In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*. [S.l.]: Association for Computational Linguistics, 2013. v. 1: Long Papers, p. 832—840. Citado 2 vezes nas páginas 12 e 31.

SENNRICH, R.; HADDOW, B.; BIRCH, A. Neural machine translation of rare words with subword units. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics, 2016. p. 1715–1725. Disponível em: <https://www.aclweb.org/anthology/P16-1162>. Citado 2 vezes nas páginas 33 e 34.

SERVAN, C.; CREGO, J.; SENELLART, J. *Domain specialization: a post-training domain adaptation for Neural Machine Translation*. 2016. Citado 2 vezes nas páginas 31 e 37.

SHAH, K.; BARRAULT, L.; SCHWENK, H. Translation model adaptation by resampling. p. 392–399, 07 2010. Citado na página 31.

SILVA, L. H. *Tradução Automática Neural inglês-português no domínio do E-Commerce*. 62 p. Monografia (Graduação em Ciência da Computação) — Universidade Federal de São Carlos, 2019. Citado na página 5.

SILVA, L. H.; CASELI, H. M. Word embeddings para cálculo de similaridade semântica entre textos de e-commerce. In: *Anais do V Workshop de Iniciação Científica em Tecnologia da Informação e da Linguagem Humana (TILIC 2017)*. [s.n.], 2017. p. 1–7. Disponível em: <https://sites.google.com/view/tilic2017/trabalhos>. Citado na página 24.

SKOROKHODOV, I. et al. Semi-supervised neural machine translation with language models. In: *Proceedings of the AMTA 2018 Workshop on Technologies for MT of Low Resource Languages (LoResMT 2018)*. [S.l.: s.n.], 2018. p. 37–44. Citado na página 17.

SNOVER, M. et al. A study of translation edit rate with targeted human annotation. In: *Proceedings of association for machine translation in the Americas*. [S.l.: s.n.], 2006. v. 200, n. 6. Citado na página 26.

SUTSKEVER, I.; VINYALS, O.; LE, Q. V. Sequence to sequence learning with neural networks. In: *Proceedings of the 27th International Conference on Neural Information Processing Systems*. Cambridge, MA, USA: MIT Press, 2014. (NIPS'14), p. 3104–3112. Disponível em: <http://dl.acm.org/citation.cfm?id=2969033.2969173>. Acesso em: 16 maio 2017. Citado na página 17.

UTIYAMA, M.; ISAHARA, H. Reliable measures for aligning japanese-english news articles and sentences. In: *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2003. (ACL '03), p. 72–79. Disponível em: <https://doi.org/10.3115/1075096.1075106>. Citado na página 31.

VARGA, A. C. *Domain adaptation for multilingual neural machine translation*. Tese (Doutorado) — Master Thesis, Saarlandes University, 2017. Citado na página 31.

VINCENT, P. et al. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *J. Mach. Learn. Res.*, JMLR.org, v. 11, p. 3371–3408, dez. 2010. ISSN 1532-4435. Disponível em: <http://dl.acm.org/citation.cfm?id=1756006.1953039>. Citado na página 41.

WANG, R. et al. Sentence embedding for neural machine translation domain adaptation. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Vancouver, Canada: Association for Computational Linguistics, 2017. p. 560–566. Disponível em: <https://www.aclweb.org/anthology/P17-2089>. Citado na página 31.

WANG, R. et al. Instance weighting for neural machine translation domain adaptation. In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, 2017. p. 1482–1488. Disponível em: <https://www.aclweb.org/anthology/D17-1155>. Citado na página 31.

WANG, R. et al. Neural network based bilingual language model growing for statistical machine translation. In: *EMNLP*. [S.l.: s.n.], 2014. Citado na página 31.

WANG, R. et al. *Connecting Phrase based Statistical Machine Translation Adaptation*. 2016. Citado na página 31.

WEES, M. van der; BISAZZA, A.; MONZ, C. *Dynamic Data Selection for Neural Machine Translation*. 2017. Citado na página 31.

WU, Y. et al. Google's neural machine translation system: Bridging the gap between human and machine translation. In: . [s.n.], 2016. abs/1609.08144. Disponível em: <http://arxiv.org/abs/1609.08144>. Acesso em: 19 maio 2017. Citado na página 17.

ZHANG, J.; ZONG, C. Exploiting source-side monolingual data in neural machine translation. In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics, 2016. p. 1535–1545. Disponível em: <https://www.aclweb.org/anthology/D16-1160>. Citado 3 vezes nas páginas 31, 33 e 34.

ZHOU, X.; CAO, H.; ZHAO, T. Domain adaptation for smt using sentence weight. In: *CCL*. [S.l.: s.n.], 2015. Citado na página 31.

ZOPH DENIZ YURET, J. M. B.; KNIGHT, K. Transfer learning for low-resource neural machine translation. In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. [S.l.: s.n.], 2016. v. 1, p. 1568—1575. Citado na página 12.