UNIVERSIDADE DE BRASÍLIA
INSTITUTO DE LETRAS
DEPARTAMENTO DE LÍNGUAS ESTRANGEIRAS E TRADUÇÃO

TRADUÇÃO E POPULARIZAÇÃO DE CIÊNCIA: O CAPÍTULO 7 DO LIVRO *MACHINE TRANSLATION FOR EVERYONE*

RAQUEL BRANDÃO NAVARRO

Brasília

2023

RAQUEL BRANDÃO NAVARRO

TRADUÇÃO E POPULARIZAÇÃO DE CIÊNCIA: O CAPÍTULO 7 DO LIVRO *MACHINE TRANSLATION FOR EVERYONE*

Trabalho de Conclusão de Curso apresentado ao Departamento de Línguas Estrangeiras e Tradução da Universidade de Brasília como requisito parcial para a obtenção do título de Bacharel em Letras Tradução - Inglês.

Orientadora: Profª. Drª. Elisa Duarte Teixeira

Brasília

Dezembro de 2023

RESUMO

Entre os diversos mecanismos pelos quais computadores, no geral, e a internet, em particular, facilitaram o acesso ao conhecimento está o advento das traduções automáticas. Não apenas o cidadão comum consegue hoje acessar textos que, em outros tempos, não conseguiria (ou, se conseguisse, seria com grande dificuldade), mas também a ele é permitido ler textos e fazer buscas e compras em dezenas ou centenas de outros idiomas. Essas traduções ainda têm suas limitações, mas representam um grande avanço em relação à ilegibilidade total. Tradutores profissionais também contam com ferramentas de tradução automática para aumentar sua produtividade e a qualidade de seu trabalho, o que é particularmente importante em uma era em que o volume de textos produzidos e traduzidos é significativamente maior do que em qualquer era anterior. Posto isso, o presente trabalho teve por objetivo traduzir o capítulo 7 do livro *Machine Translation for Everyone:* Empowering users in the age of artificial intelligence (KENNY, 2022), que se ocupa de explicar ao público geral que está por trás dos mecanismos de tradução automática baseadas em redes neurais. Entender esse mecanismo é por si só interessante, do ponto de vista da divulgação científica, mas também é relevante para pessoas que estão aprendendo um novo idioma e para profissionais do texto, como tradutores e redatores técnicos, na medida em que permite uma melhor utilização desses tradutores automáticos para melhor produzir ou aprimorar seus textos.

Palavras-chave: Tradução automática. Redes neurais. Divulgação científica. Tradução especializada. Terminologia.

ABSTRACT

Among the various mechanisms by which computers in general, and the internet in particular, have facilitated access to knowledge, is the advent of automatic translations. Not only are ordinary citizens now able to access texts that they wouldn't have been able to in the past (or, if they were, it would have been with great difficulty), but they are also able to read texts, make searches and shopping purchases in dozens or hundreds of languages. These translations still have their limitations, but they represent a major advance over total illegibility. Professional translators also rely on machine translation tools to increase their productivity and quality of ther work, which is particularly important in an era in which the volume of texts produced is significantly higher than in any previous era. Against this backdrop, this paper aims to translate chapter 7 of the book Machine Translation for Everyone, which explains the mechanism of machine translation using neural networks to the general public. Understanding this mechanism is interesting in itself, from the point of view of scientific dissemination, but it is also relevant for people learning a new language, and for text professionals such as translators and technical writers as it allows them to use automatic translators better to produce or improve their own texts.

Keywords: Machine translation. Neural networks. Popularization of Science. Specialized translation. Terminology.

LISTA DE FIGURAS

SUMÁRIO

INTRODUÇÃO

Este trabalho ocupa-se do estudo e da tradução, do inglês para o português, do capítulo 7, Como funciona a tradução automática neural (*How neural machine translation works)*, do livro *Machine Translation for Everyone*, organizado por Dorothy Kenny e publicado pela Language Science Press, em 2022. O livro consiste em uma coletânea de capítulos independentes entre si e escritos por diferentes autores e autoras, muitas vezes em parceria com a própria Dorothy Kenny, que é professora e pesquisadora em Estudos de Tradução pela Universidade de Dublin, na Irlanda. O capítulo 7, selecionado para ser traduzido do presente Trabalho de Conclusão de Curso, foi escrito em conjunto com Juan Antonio Pérez-Ortiz, Mikel Forcada e Felipe Sánchez-Martinez, descrito por Kenny como "o mais técnico do livro" (KENNY, 2022). Ele trata do mecanismo de funcionamento por trás de traduções feitas com redes neurais e abrange as técnicas mais frequentemente utilizadas em sistemas de tradução automática contemporâneos.

A escolha da tradução deste capítulo se dá por uma miríade de fatores. Primeiramente, pela prevalência do uso de ferramentas de tradução automáticas por usuários de internet: o aplicativo para celular do Google Tradutor, por exemplo, já foi baixado mais de um bilhão de vezes (GU, 2023). Além disso, apesar de estarem muito presentes na vida de uma grande parcela da população, as ferramentas de tradução automática e seu funcionamento são relativamente desconhecidos. Do ponto de vista de seus usuários, elas funcionam como caixas pretas: basta digitar uma frase em um idioma no campo de entrada e receber sua tradução no campo de saída.

Por ser um assunto técnico e especializado, poucas pessoas sabem como funcionam redes neurais para tradução, o que é em si suficiente para justificar a ampliação de conhecimento da área em português brasileiro. Um terceiro motivo é o fato de o tema ser de interesse profissional para tradutores que desejem aprimorar-se no ofício. Entender o

funcionamento de traduções automáticas permite aumentar sua produtividade, por contar com uma nova ferramenta em seu arsenal, bem como expandir seu entendimento do processo tradutório, ao reconhecer vieses e limitações dos softwares atuais e, com isso, melhorar traduções existentes.

Por fim, a tradução deste capítulo de livro, do gênero "divulgação científica", apresenta desafios únicos, que se assemelham aos enfrentados pelos divulgadores científicos que escrevem obras do gênero: atuam, de certa forma, como intérpretes de pesquisadores e pesquisadoras, que se comunicam entre si por meio de uma linguagem técnica altamente especializada, transmitindo esse conhecimento para o público geral. Tradutores, por sua vez, ampliam o alcance da divulgação científica para falantes de outros idiomas e procuram preservar, no texto traduzido, a inteligibilidade e propósito comunicativo do texto de partida, sem introduzir nele erros e imprecisões que prejudicam a divulgação científica. Mas este trabalho visa não apenas traduzir conceitos especializados, mas também combater a prevalência de anglicismos na terminologia da área, procurando promover uma maior inclusão tecnológica, linguística e cultural.

O trabalho está dividido em duas partes principais. A primeira consiste em uma discussão teórica sobre o projeto de tradução apresentado. Com relação ao gênero da divulgação científica, aos temas de tradução no geral, à tradução automática em particular, e à obra escolhida, são tecidos alguns comentários e apresentados alguns princípios que norteiam as decisões tradutórias. Entretanto, é possível adiantar que, na medida do possível, os critérios preponderantes adotados para a tradução foram a simplicidade e a naturalidade do texto de chegada, para atender ao público brasileiro. A segunda parte do trabalho trata de apresentar, em detalhes, a metodologia utilizada para o processo tradutório e um relatório de tradução, nos quais se registram os aprendizados, desafios e soluções envolvidos na pesquisa e na prática de traduzir o texto em discussão.

1 JUSTIFICATIVA

O livro utilizado como texto de partida, *Machine Translation for Everyone:* Empowering users in the age of artificial intelligence (KENNY, 2022), tem como público-alvo todas as pessoas interessadas em fazer uso de traduções automáticas. Ele busca abordar um amplo repertório de questões relacionadas ao tema, tais como a forma de avaliar uma tradução automática (cap. 3), meios de preparar textos para facilitar traduções automáticas futuras (cap. 4), questões éticas pertinentes ao campo (cap. 6) e formas de editar textos traduzidos automaticamente (cap. 5), entre outros. O livro não pressupõe conhecimento prévio de teorias da tradução, redes neurais, computação ou matemática, podendo ser usado com proveito por qualquer pessoa interessada em aprender ou ensinar sobre o ato de traduzir com o auxílio da tradução por máquina. O excerto escolhido, o capítulo 7, *Como funciona a tradução automática neural*, trata do mecanismo de funcionamento de tradutores automáticos baseados em redes neurais.

A tradução automática neural, do inglês *Neural Machine Translations* (NMT), é um fenômeno relativamente recente na história da tecnologia e até mesmo na história da tradução automática. Até há pouco, vigoravam modelos estatísticos – também do inglês, *Statistical Machine Translation* (SMT). O primeiro trabalho científico a respeito de traduções feitas por redes neurais foi publicado em meados da década passada (BAHDANAU, 2014) e, desde então, a popularidade da NMT só fez crescer.

Em 2016, durante a WMT (*Conference on Machine Translation*, previamente *Workshop on Machine Translation*), uma conferência anual organizada pela *Machine Translate* na qual pesquisadores da área avaliam modelos desenvolvidos por outros pesquisadores em diversas tarefas tradutórias, 90% das propostas vencedoras eram de modelos de NMTs (BOJAR et al., 2016). Ainda em 2016, o Google anunciou que adotaria um modelo de NMT para as traduções feitas pelo Google Translate (TUROVSKY, 2016). Em 2017, foi lançado para o

público outra ferramenta popular de tradução automática, DeepL[1], construída a partir de um modelo desenvolvido por uma equipe dentro do Linguee, provedora de um dicionário online de mesmo nome.

O rápido crescimento da popularidade de NMT se deu pelo fato de, grosso modo, requererem significativamente menos memória para funcionar. Os métodos de SMT dependem de grandes corpora bilíngues para produzir resultados aceitáveis, muitos dos quais são raros para uma grande quantidade de pares linguísticos.

A prevalência do uso de traduções feitas por computadores, sejam elas por redes neurais ou não, assim como a popularidade de NMT e diversas áreas correlatas, como grandes modelos de linguagem (*Large Language Models*, ou LLMs), redes neurais em um contexto mais amplo, aprendizado de máquina e inteligência artificial é refletida e reflete os avanços que ocorrem na pesquisa acadêmica, responsável por publicar um grande volume de estudos acerca desses temas todos os anos e em diversas partes do mundo. Essas publicações são majoritariamente em inglês, que é considerado atualmente a *lingua franca* da comunidade acadêmica internacional (JENKINS, 2013).

A fim de ficar a par dos avanços da ciência, o público internacional não especializado depende da atuação de meios de divulgação científica e tradutores, que, juntos, ajudam a suprir uma alta demanda por temas pelos quais há crescente interesse, e que estão em constante processo de atualização. Nesse contexto de novidades frequentes, a tradução especializada ganha ainda mais importância. É essencial destacar que o tradutor não apenas transmite um conhecimento já construído, mas também atua como criador de significados, pois ele geralmente participa da construção terminológica e cultural do novo tema em sua língua-alvo (RUSH HOVDE, 2010).

[1] Disponível em: <https://www.deepl.com/translator>

A obra em estudo é, portanto, relevante para o profissional que deseja entender melhor o funcionamento de uma ferramenta útil para o seu fluxo de trabalho, bem como para quem ensina tradução. Além disso, é importante para o público em geral, que deseja se informar a respeito de um dos usos mais bem-sucedidos da tecnologia na atualidade, de grande prevalência e inúmeras outras aplicações, que são as redes neurais artificiais.

Ademais, uma nova tradução na área de tecnologia contribui significativamente para o enriquecimento do vocabulário na língua-alvo. Ela permite a criação e o uso de termos especializados nativos do português, que podem ser mais facilmente compreendidos pelos falantes da língua. Portanto, o presente trabalho pode até mesmo contribuir, com a consulta a especialistas da área, para a construção de um vocabulário terminológico especializado na área da tecnologia em questão, o que contribuiria para reduzir a dependência da área por anglicismos, muito comuns nas áreas de tradução automática, programação e tecnologia em geral.

Vale lembrar que nem sempre há equivalentes diretos comumente usados para diversos termos técnicos da área de tecnologia para língua portuguesa, o que acaba privilegiando o uso de anglicismos pelo público especializado. Isso leva a desafios na tradução e na comunicação eficaz de conceitos especializados. Além disso, essa escolha por termos em inglês pode resultar em palavras ou frases mais longas e menos práticas em português, nos livros e materiais de divulgação da área, prejudicando a disseminação de conhecimento e a preservação da cultura e do português brasileiro. Um dos objetivos deste trabalho de tradução, portanto, é propor alternativas para os anglicismos na área que, quiçá, passarão a ser usados pelos especialistas da área de ferramentas de tradução por máquina.

Na seção a seguir, serão explicadas as teorias e princípios que embasaram e nortearam as decisões tradutórias do capítulo escolhido, entre as quais destacaremos a Tradução

Especializada como área de pesquisa no âmbito dos Estudos da Tradução, e a Terminologia, disciplinas que dependem uma da outra e se influenciam mutuamente.

2 FUNDAMENTAÇÃO TEÓRICA

Nesta seção, a fim de explicar alguns dos princípios norteadores que se manifestam concretamente por meio das decisões tradutórias tomadas ao longo da tradução feita neste trabalho, do Capítulo 7 do livro *Machine translation for everyone* (KENNY, 2022). O arcabouço teórico que embasou essas decisões tem suas raízes em três áreas fundamentais: a Terminologia bilíngue, a tradução especializada, e o conceito de inglês como *lingua franca* da comunidade acadêmica internacional.

A tradução foi feita tendo em mente um esforço crítico para a democratização da terminologia, no português brasileiro, relacionada à tradução automática, inteligência artificial e tecnologia em geral. Para fundamentar esse esforço, é necessário abordar, primeiramente, a problemática da predominância do inglês como língua franca, principalmente na área de tecnologia, visto que essa predominância impõe desafios comunicativos e tradutórios, tanto para a divulgação científica quanto para outros trabalhadores da área, inclusive tradutores.

Por outro lado, não há como falar de tradução especializada sem falar em Terminologia, o estudo do léxico especializado. Da perspectiva dos Estudos de Tradução, a Terminologia é, segundo (CABRÉ, 1999), a "disciplina que se ocupa do estudo e da compilação de termos especializados". É ela que sistematiza os princípios que guiam a compilação de termos próprios a uma dada ciência, arte, técnica ou profissão (BEVILACQUA, 2017), conforme discutimos nos subitens a seguir.

2.1 Tradução e Terminologia

Tradução e Terminologia, enquanto campos do conhecimento e disciplinas acadêmicas, são distintas, mas interrelacionadas (CABRÉ, 1999). Uma grande superfície de contato entre essas disciplinas ocorre no campo da tradução especializada. Se a tradução, em termos gerais, se ocupa de transmitir, em uma segunda língua, ideias originalmente expressas em outra, é fácil

ver no campo da tradução especializada a influência da terminologia (grafada aqui com minúsculas, para se referir ao conjunto de termos de uma dada área especializada), uma vez que as ideias de um texto técnico só podem ser transmitidas adequadamente se os termos técnicos utilizados em ambas as línguas puderem ser considerados equivalentes aceitáveis, ainda que essa equivalência seja parcial.

O crescimento contínuo nas interações linguísticas, avanços tecnológicos constantes e um ritmo acelerado no desenvolvimento científico intensificam a necessidade de uma conversão precisa de termos entre diferentes idiomas e áreas do conhecimento. Tais mudanças sociais importantes ocorridas nas últimas décadas serviram para aumentar ainda mais o contato entre as duas disciplinas, dado que ambas se ocupam, primariamente, dos usos da língua feitos em contextos profissionais e especializados reais. Cabré (1999) explica como as mudanças sociais ocorridas nas últimas décadas afetaram o desenvolvimento histórico da Terminologia, algumas delas diretamente relacionadas com a tradução especializada. A autora explica que o desenvolvimento acelerado da tecnologia faz com que surja um grande número de novos conceitos, que recebem termos próprios para sua designação. Em particular, o desenvolvimento das tecnologias de comunicação envolve um número cada vez maior de pessoas e o contato entre elas faz surgir novas necessidades linguísticas.

O volume de informações produzidas e a frequência com que elas necessitam de atualizações torna cada vez mais relevante a atividade da tradução especializada. Rogers (2015) estima que 80% a 90% da demanda profissional por tradução seja voltada à tradução técnica. A disseminação e a acessibilidade a novos produtos tecnológicos, como celulares e a internet, catalisa os fenômenos apontados, na medida em que a língua é o que materializa a faculdade da linguagem.

Mas a comunicação em ambientes profissionais não é uniforme. Especialistas podem se comunicar com seus pares, com um público semi-especializado (como especialistas de áreas correlatas ou aprendizes) ou o público geral, geralmente por meio de veículos de comunicação em massa, quando um vocabulário mais simples precisa ser usado para que haja compreensão dos conceitos e fenômenos complexos pesquisados. A este tipo de comunicação dá-se o nome de divulgação científica, além disso, vale destacar que a divulgação científica varia em seu grau de especialização, fazendo parte do dia a dia das pessoas de maneiras distintas. No item a seguir, são expostos alguns fatores importantes a se considerar na tradução de textos de divulgação científica.

2.2 Tradução de textos de divulgação científica

Traduzir textos transcende a mera conversão de palavras de um idioma para outro, abrindo caminhos para o acesso ao conhecimento por um público diversificado que talvez não domine o idioma original. Esta perspectiva nos convida a considerar a interconexão entre Tradução e Terminologia no universo dos textos especializados. Embora seja tentador pensar que a linguagem especializada é o que liga a Tradução à Terminologia, a classificação de um texto como "técnico" ou como "divulgação científica" não é estritamente binária, mas um continuum de níveis de especialização. Os contextos em que a população em geral entra em contato com textos técnicos são bastante diversos, como, por exemplo, a consulta a um manual de eletrodoméstico, ou a leitura de uma legislação para requerer algum direito junto a órgãos públicos. Por meio de textos especialmente criados para esse fim, o público pode aprender a respeito de temas científicos especializados com maior facilidade – o que não seria possível por meio dos textos científicos, escritos por especialistas para serem lidos exclusivamente por especialistas.

Textos de divulgação científica cumprem um papel importante de possibilitar um entendimento básico a respeito de temas especializados por meio de uma linguagem mais acessível. Entre outros efeitos positivos, esse entendimento possibilita um melhor exercício da cidadania, já que temas que têm potencial de afetar toda a população, como é o caso das mudanças climáticas ou dos avanços da inteligência artificial, podem ser debatidos e, por vezes, enfrentados, de modo mais democrático.

2.3 Imperialismo linguístico: inglês como *lingua franca*

Uma outra consequência do desenvolvimento acelerado da tecnologia e seus desdobramentos foi a disseminação mundial da língua inglesa e a incorporação de termos em inglês ao vocabulário técnico de diversos campos do saber em muitas línguas, particularmente nos meios digitais. Embora o empréstimo linguístico constitua um fenômeno natural e algumas vezes saudável na relação entre línguas, o caso do inglês nos domínios especializados do saber, em especial a tecnologia, tem peculiaridades que podem constituir o que Phillipson (1992) cunhou de "imperialismo linguístico", fenômeno pelo qual a língua inglesa domina, marginaliza e lentamente apaga outras línguas.

É importante destacar que tradutores se deparam frequentemente com diversos desafios terminológicos no exercício da tradução de textos especializados, como encontrar um equivalente, na língua de chegada, para um termo que pode ter sido cunhado no próprio texto que está sendo traduzido. Outras vezes, a equivalência entre termos em duas línguas é apenas parcial e insuficiente para o propósito daquela tradução ou contexto específicos; por vezes, é difícil determinar se há equivalente adequado na língua de chegada. Estes são desafios legítimos de Terminologia elencados por (CABRÉ, 1999) e amplamente reconhecidos.

Stolze (1999) propõe algumas soluções para problemas de não-equivalência, entre as quais se encontram:

- empréstimos
- decalques
- neologismos
- paráfrases
- explicações.

Os **empréstimos** são termos e expressões da língua de origem introduzidos na língua de chegada (BARBOSA, 2004), ou seja, são estrangeirismos utilizados sem tradução, como é o caso de "mouse". O **decalque**, por sua vez, é muito similar ao empréstimo, tendo em vista que ele é a acomodação do empréstimo na língua de chegada, sendo adaptado foneticamente ou, até mesmo, morfologicamente, tal como o aportuguesamento do verbo "clicar", do inglês *to click*. Os **neologismos** são expressões ou palavras novas para representar objetos sem denominação anterior na língua de chegada e, assim como o decalque, podem ser considerados parte do processo de adaptação de um estrangeirismo, tal como o termo "*pen drive*", um termo que não existe no inglês e que se refere ao dispositivo "*flash drive*".

Cabe ressaltar que o tradutor especializado não se confunde com o terminólogo, no sentido de ter autonomia, no exercício da tradução para criar terminologias sem o concurso de especialistas da área. Mas, como reconhecem Cabré (1999) e outros autores, o tradutor muitas vezes vai, sim, acabar desempenhando esse papel, e é importante que esteja bem preparado para tal. O tradutor pode se preparar familiarizando-se com uma visão geral de fundamentos teóricos de Terminologia, bem como seus conceitos básicos e a sua relação com outros conhecimentos, com destaque à Tradução. Em alguns cursos universitários, futuros tradutores podem aplicar esses conceitos em disciplinas dedicadas à prática de Terminologia Aplicada, elaborando pequenos glossários e dicionários com auxílio de princípios da Linguística de Corpus.

As paráfrases e explicações, por sua vez, geralmente são usadas quando o tradutor não pode recorrer a outros artifícios e por isso precisa estender o texto para acomodá-las. A **paráfrase** é uma espécie de explicação breve que é inserida no corpo do texto, enquanto a **explicação** pode ser uma nota de rodapé.

Com uma vasta gama de alternativas para solucionar esses problemas terminológicos, chama atenção a frequência com que textos técnicos em outras línguas, como é o caso do português brasileiro, adotam o empréstimo linguístico do inglês. Essa frequência pode ser fruto de um viés que, por sua vez, é reflexo da prevalência da língua inglesa na comunicação global e constitui uma faceta do imperialismo linguístico e, num contexto mais amplo, político e econômico. Os Estados Unidos da América se estabeleceram como força econômica, tecnológica e militar hegemônica com relação ao resto do mundo e, na posição de super potência hegemônica, são exportadores de ideologia, cultura e língua. O estabelecimento do inglês como *lingua franca* internacional é consequência dessa hegemonia.

Se, por um lado, a comunicação facilitada entre pessoas de diversas partes do mundo constitui uma possibilidade de intercâmbios culturais valiosos e de enriquecimento de todas as partes envolvidas, é importante refletir a respeito de alguns de seus aspectos problemáticos.

O primeiro desses aspectos é a uniformização da linguagem e a visão do inglês como língua "neutra" ou que sua imposição como *lingua franca* é "apolítica" e não uma representação da cultura dominante. Mais que isso, sua prevalência faz crer que essa imposição, na verdade, foi uma escolha da comunidade internacional. E essa problemática passa, certamente, pela (não)tradução.

Outro aspecto relevante é a observação de que a dominação pela língua e pela cultura são mecanismos de perpetuação de desigualdades econômicas e sociais, contanto que elas beneficiem grupos dominantes historicamente. Determinadas ideias e discursos acabam se propagando muito mais facilmente em virtude do controle exercido sobre meios de

comunicação por esses grupos, e a ciência, geralmente considerada um campo mais neutro de ideologias e preocupada com o coletivo do saber, também reproduz e perpetua esses ideais hegemônicos.

Além do mais, há outros impactos a serem discutidos quando se pensa em inglês como *lingua franca*, tendo em vista que o uso hegemônico de anglicismos leva à redução da diversidade linguística e da acessibilidade não apenas na língua oral como no discurso técnico e científico, e coloca-se, especialmente no caso da ciência, como uma barreira para o acesso ao conhecimento por parte da população geral de menor escolaridade. Em função dessa hegemonia, profissionais e estudantes enfrentam desafios adicionais para acessar e publicar conteúdos de divulgação científica, como a necessidade de aprender inglês para quaisquer profissões, ou de decorar termos em inglês de áreas especializadas que não têm tradução para o português brasileiro, por exemplo. A esse respeito, Olohan (2016) observa que, tanto quanto possível, falantes não-nativos do inglês optam por ler artigos em suas línguas maternas quando a escolha é possível. Ou seja, a tradução de textos de divulgação científica ajuda a cumprir o propósito de popularizar a ciência junto ao público que não domina o inglês.

Felizmente, existem estratégias de mitigação de impacto que podem ser aplicadas à elaboração de produções científicas e que também nortearam a tradução do capítulo 7, objeto deste trabalho. Essas estratégias incluem, especialmente, o incentivo à publicação de artigos e traduções que privilegiem escolhas tradutórias em português ou em outros idiomas que não o inglês, para contribuir com a difusão de termos da área de tecnologia que estejam traduzidos para o português, ainda que possa haver o uso de termos concorrentes em língua inglesa em textos especializados da área. Ademais, a tradução tem um papel crucial no fortalecimento de outras línguas além do português.

A seguir, descrevemos os passos seguidos em preparação à tarefa tradutória para, em seguida, apresentar um relatório da tradução do texto selecionado, ressaltando as características e fenômenos que mais nos chamaram a atenção nesse processo.

3 METODOLOGIA

Esta seção tem por objetivo apresentar e descrever as etapas do processo tradutório do capítulo 7 da obra em estudo à luz das referências teóricas apresentadas no capítulo anterior. Serão detalhados o preparo do documento, as ferramentas auxiliares utilizadas e as decisões tradutórias tomadas em pontos mais complexos do texto de partida.

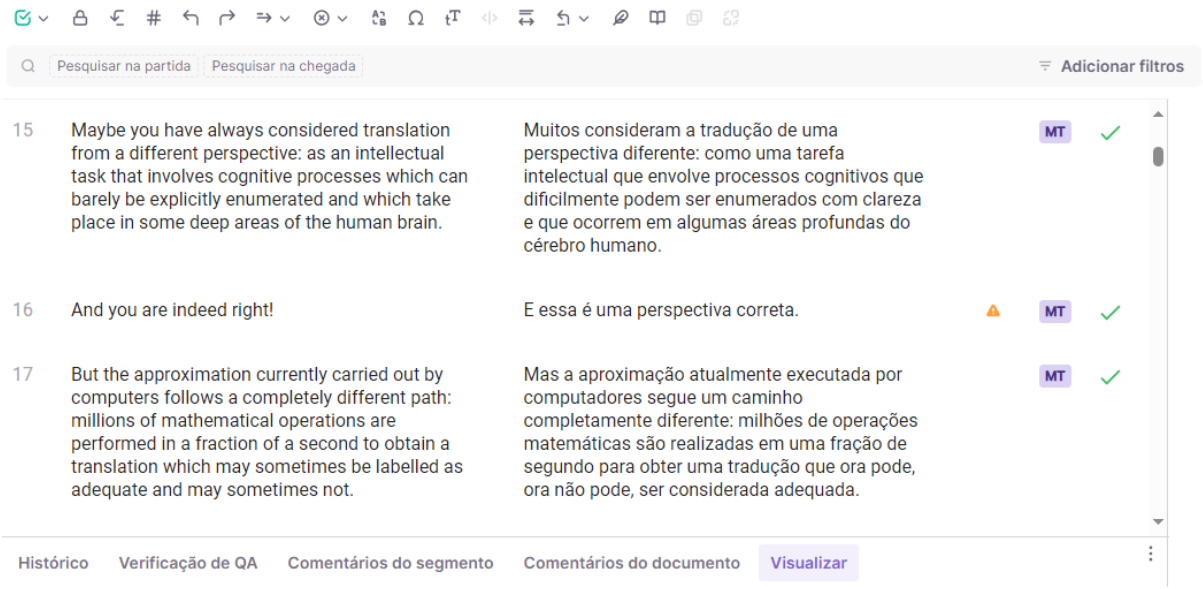
3.1 Preparo do documento

O material traduzido neste trabalho, o capítulo 7, *Como funciona a tradução automática neural*, do livro organizado por Dorothy Kenny (2022), está disponível gratuitamente na rede no formato .pdf[2]. Para utilizar a ferramenta automática de auxílio à tradução Smartcat[3], escolhida para o trabalho por familiaridade e praticidade, o arquivo .pdf precisou ser convertido para o formato de texto editável .docx, o que foi feito por meio do Google Docs. Em seguida, o arquivo convertido precisou passar por um processo de limpeza e formatação, no qual foi editado para que alguns caracteres fossem corrigidos, como hífens separando palavras e alguns caracteres Unicode renderizados incorretamente na conversão (por exemplo, o símbolo "€" no lugar de "ç").

O arquivo foi, enfim, carregado em um projeto de tradução no Smartcat, que o dividiu em segmentos, geralmente correspondentes a sentenças, numerados e dispostos na forma de uma tabela de duas colunas, de tal modo que segmentos em inglês ficassem lado a lado com suas respectivas traduções para o português, conforme ilustrado na Figura 1. Isso foi particularmente útil em virtude de facilitar o planejamento do trabalho.

[2] Disponível em: <https://langsci-press.org/catalog/book/342>. Acesso em outubro de 2023.
[3] Disponível em: <https://smartcat.com/>.

Figura 1: documento em processo de edição no Smartcat.



Fonte: autoria própria.

3.2 Processo tradutório

O processo de traduzir o texto de partida foi feito tirando proveito da interface de tradução, com todas as suas ferramentas e organização do texto na tela oferecidas pelo site do Smartcat. Como o texto é dividido em segmentos de uma sentença, a tradução foi feita segmento a segmento, nos quais várias alternativas de tradução foram comparadas. Em uma grande parte dos trechos, foram comparadas as sugestões de tradução fornecidas pela ferramenta de tradução automática padrão do próprio Smartcat, e as sugestões retiradas externamente pelo DeepL, que é uma ferramenta de tradução de desempenho superior ao do Google Translate (COLDEWAY, 2017) e que, apesar de ter sido usada como uma opção adicional de tradução automática, também pode ser integrada diretamente ao Smartcat, assim como outras ferramentas, substituindo a tradução automática fornecida como padrão por essa *CAT Tool*. Por vezes, a própria sugestão de tradução fornecida pelo Google Translate foi verificada, a fim de determinar se um termo específico poderia ser traduzido de outra forma. A

escolha por uma determinada versão dentre as oferecidas, pela combinação dessas versões ou pela tradução "manual" de cada segmento feita inteiramente pela autora deste trabalho, com base em seu discernimento a respeito da qualidade das sugestões produzidas pelos tradutores automáticos consultados.

Em trechos considerados simples, como aqueles de uma ou duas palavras, títulos de seções e afins, a tradução automática do Smartcat foi aceita na maior parte dos casos, a fim de ganhar tempo para dedicar a trechos mais complexos, de maior nuance ou que exigissem decisões tradutórias que afetariam uma porção grande do texto.

3.3 Cotejo e pós-edição

Depois de realizada a tradução, passou-se ao cotejo do bitexto, em que os segmentos traduzidos foram, um a um, comparados com seus correspondentes no texto de partida, a fim de verificar se nenhum trecho tinha sido deixado para trás, e também para avaliar diversos aspectos técnicos dos dois textos, como a manutenção do sentido original, adequação do registro, clareza e acurácia dos termos técnicos. Para isso, foi feita uma verificação terminológica utilizando a *web* como *corpus*.

O último aspecto mencionado merece uma explicação especial, uma vez que a pesquisa de termos existentes é um trabalho realizado mais frequentemente por terminólogos.

Ao fim deste processo, foi necessário realizar a pós-revisão e pós-edição do capítulo já traduzido. Aqui, procurou-se por possíveis erros de digitação e diagramação, que foram então corrigidos, e uma atenção especial foi dada à naturalidade e fluidez do texto, tendo em mente o público-alvo brasileiro. Por fim, foi feita uma conferência do leiaute do documento de saída produzido pelo Smartcat, a fim de certificar que o documento traduzido estivesse em conformidade visual com o original. Foi nesta etapa que se traduziram também as figuras do documento contendo texto em outra língua.

A seguir, detalhamos alguns aspectos recorrentes e relevantes do processo tradutório à luz das abordagens teóricas utilizadas.

4 RELATÓRIO DE TRADUÇÃO

Traduzir é, mesmo em condições favoráveis, uma tarefa desafiadora. Não apenas por causa do esforço subentendido à atividade, que passa por aprender não um, mas dois idiomas com um determinado grau de proficiência que seja suficiente para tornar possível o intermédio da comunicação entre dois indivíduos ou grupos, mas também porque requer do profissional equilibrar inúmeras condições por vezes conflitantes: a fidelidade ao material de origem, o comprometimento com o público-alvo, as idiossincrasias dos idiomas envolvidos, convenções de gênero, considerações de tamanho do texto. Também se deve levar em conta que o trabalho do tradutor é moldado pela língua de partida e da língua de chegada dos textos traduzidos, enquanto também ajuda a moldar esse discurso.

Esta seção trata dos principais desafios enfrentados no processo de traduzir o capítulo da obra em discussão. Cada uma das subseções dirá respeito a um tema e apresentará as dificuldades existentes, as soluções encontradas, os embasamentos das decisões tomadas, bem como as vantagens e desvantagens dessas soluções.

4.1 A prevalência de termos em inglês em textos de tecnologia

Traduzir textos a respeito de tecnologia da informação, suas áreas correlatas e suas sub-áreas, representa um desafio, entre outros motivos, em vista do fato de que muitos dos termos que descrevem equipamentos, técnicas e procedimentos são importados diretamente do inglês, sem adaptações. É o caso de dispositivos como *smartphones*, *notebooks* / *laptops*, *chips*, e *mouses*, de objetos virtuais como *dataframes*, *links*, *softwares*, de ações como *downloads*, *uploads*, *backups* e da própria *internet*. Áreas inteiras, muitas vezes, são conhecidas exclusivamente ou predominantemente em inglês, como *Machine Learning* e *Data Science*.

Ao se deparar com determinados termos no texto, a tradutora ou tradutor se vê diante de uma escolha difícil. Tomemos o exemplo da expressão *word embedding*, que é um dos

conceitos chave do texto e, portanto, aparece diversas vezes, representando uma decisão tradutória de extrema relevância. O próprio texto de partida, além de diversas outras fontes consultadas, explica que *word embedding* se refere ao ato ou efeito de se representar uma palavra como uma lista de números separados por vírgulas, chamada de "vetor". Isso seria vantajoso do ponto de vista computacional, pois permite, entre diversas outras operações matemáticas, transmitir matematicamente a ideia de semelhança semântica entre duas palavras, dado que é possível calcular, matematicamente, a distância entre dois vetores. Ou seja, transformar palavras em vetores tem a distinta vantagem de associar palavras de significado parecido a vetores próximos entre si. Como é possível calcular a distância, essas listas de números, que são os vetores, e como é possível transformar palavras em vetores, torna-se também possível calcular "distâncias semânticas" entre palavras por meio dos vetores, de modo que palavras com significados próximos sejam representadas por vetores separados entre si por distâncias pequenas. Isso permite que um programa de computador consiga "entender" a noção intuitiva a humanos de que a palavra "gatinho" é mais parecida, do ponto de vista semântico, com a palavra "gato" do que com a palavra "abacaxi". Com essa explicação em mente, é possível pensar em algumas alternativas para traduzir as ocorrências de *word embedding* (ou apenas *embedding*) no texto.

A primeira e mais simples dessas alternativas é realizar um empréstimo do inglês. É uma alternativa possível que é utilizada amplamente nos textos em línguas que não o inglês no que diz respeito à tradução automática, inclusive nos (relativamente poucos) textos em português encontrados em extensivas consultas feitas na internet para a elaboração deste trabalho. Além disso, conforme explicado nesta seção, é uma medida muito adotada em textos de tecnologia da informação traduzidos do inglês. Uma desvantagem desta alternativa, entretanto, é que ela tem o potencial de atrapalhar o entendimento do público que não fala inglês ou mesmo do público iniciante que esteja interessado no assunto, embora o uso de

anglicismos em textos de tecnologia possa ser comum e aceito, ainda existem pessoas a serem incluídas. A expressão *word embedding* em um texto em português pode ser de fácil compreensão para alguém que já esteja familiarizado com o tópico, mas não é imediata para quem está entrando em contato com o assunto pela primeira vez ou na própria leitura do texto de divulgação, ainda que essa pessoa tenha conhecimentos de inglês.

Uma segunda alternativa, também usada por algumas das fontes consultadas, consiste em uma mistura entre português e inglês, com a expressão "*embedding* de palavras". Ela pode parecer ligeiramente melhor para um brasileiro já familiarizado com o assunto, mas não chega a resolver o problema da não transparência de significado, ou a problemática da dominação linguística do inglês nas áreas da ciência, uma vez que é razoável supor que o principal impeditivo à compreensão de *word embedding* está no termo *embedding*, não no termo *word*.

Uma terceira alternativa é a de traduzir a expressão de modo literal. O verbo em inglês *to embed* costuma ser traduzido em português como *embutir*, *incorporar* ou, ainda, *imergir*. A expressão *word embedding* vem de uma operação matemática dentro do campo da topologia. Essa operação recebe o nome de *embedding* em inglês e de imersão em português (LIMA, 1970). Naturalmente, isso levaria à escolha de *imersão de palavras* ou *imersão lexical* para traduzir a expressão. Na busca por traduções existentes da expressão em português utilizando a *web* como *corpus*, não foram encontrados resultados para nenhuma dessas duas opções no contexto de tecnologia, o que pode ser frustrante para um indivíduo que, após a leitura do texto, procure se informar a respeito do assunto. É interessante constatar, entretanto, que algumas entradas da Wikipédia em outras línguas utilizam essa tradução de aspecto matemático, ou seja, com o termo equivalente a "imersão" nessas línguas, o que indica que, para elas, traduzir dessa forma pode ser uma opção viável (em francês, por exemplo, a entrada para *word embedding* na Wikipédia recebe o nome de *plongement lexical*).

Até agora, entre as opções de tradução apresentadas, as alternativas que parecem mais adequadas são reiterações dos termos existentes em inglês, como ocorreu na consagração de termos como *chip* e vários outros em português. Entretanto, é possível chegar a um resultado mais aceitável levando em conta a própria natureza da operação à qual se refere *word embedding*. Haja vista que o processo consiste em representar uma palavra como um vetor, algumas alternativas *prima facie* de tradução para a expressão surgem, todas elas dialogando entre si, com respaldo na literatura em português (CASELI et al, 2022), relativamente intercambiáveis e intuitivas para o público geral: “vetor lexical”, “vetor de palavra”, “vetor de sentido”, ou “vetor semântico”. Essas escolhas também tornam possível uma nuance de sentido que não existia na versão original e podem facilitar o entendimento do público-alvo: a distinção entre vetorização e vetor. Estes termos surgem como alternativas por conta de dizerem respeito ao processo de transformar em uma palavra, em língua natural, em um vetor levando em conta seu significado.

A vetorização lexical pode se referir ao processo de representar uma palavra como vetor, enquanto o vetor lexical é o resultado do processo da vetorização. No texto fonte, o termo *embedding* é utilizado tanto para se referir ao processo quanto ao seu resultado. Em virtude das vantagens apresentadas, o texto foi traduzido predominantemente com os termos “vetor semântico”, para se referir ao resultado, e “vetorização semântica” (ou apenas “vetorização"), para se referir ao processo.

Essa escolha se deve a um conjunto de fatores considerados relevantes. Inicialmente, é interessante destacar que, entre as quatro escolhas apresentadas, duas se referiam ao fato de o vetor ser a representação matemática de uma palavra (“vetor lexical”, “vetor de palavra”), enquanto as outras duas se referiam ao fato de essa representação matemática levar em conta o significado dessa palavra. Como a característica fundamental e a vantagem que possibilita a tradução por meio de redes neurais é justamente a possibilidade de transmitir ao vetor

elementos do significado da palavra que ele representa, optou-se por uma das alternativas que fosse condizente com essa propriedade. Entre um "vetor de sentido" e um "vetor semântico", optou-se pelo último em virtude da polissemia da palavra "sentido" e da tecnicalidade do adjetivo "semântico". Quanto à polissemia da palavra "sentido", é importante notar que vetores, na Física, correspondem a grandezas como forças e velocidades, e são objetos matemáticos especiais dotados de uma grandeza (uma velocidade de 100 m/s, por exemplo), uma direção e um sentido. A direção e o sentido, nesse caso, indicam, respectivamente, a trajetória que o objeto com aquela velocidade se desloca e a forma como ele percorre aquela trajetória. Por exemplo, um objeto pode estar, em relação a algum referencial, a 100 m/s (grandeza) no eixo Norte-Sul (direção), a caminho do Norte (sentido). A expressão "vetor de sentido" pode remeter a uma ideia diferente daquela que se deseja expressar em uma parte significativa do público-alvo. A expressão "vetor semântico", por sua vez, além de ser ligeiramente mais breve, induz o leitor a pensar em algo mais acurado, na medida em que a Semântica trata justamente da relação entre um significante e seu significado. Por esses motivos, seu uso foi considerado mais adequado.

A Tabela 1 apresenta alguns exemplos de segmentos retirados da tradução na plataforma Smartcat para ilustrar os contextos em que os termos "vetor semântico", "vetorização semântica" e "vetorização" foram utilizados.

№	Partida (EN-GB)	Chegada (PT-BR)
152	Each of these vectors represents a possible word embedding for these two words.	Cada um desses vetores representa uma possível vetorização semântica para essas duas palavras.
185	By conveniently using attention to concentrate on some words in the sentence, the **embedding vector** corresponding to the word season, for example, will differ between the sentences in examples 1 and 2 below:	Ao usar a atenção de forma conveniente para se concentrar em algumas palavras da frase, o **vetor semântico** correspondente à palavra temporada, por exemplo, será

		diferente nas frases dos exemplos 1 e 2 abaixo:
3	What makes these **embeddings** really useful is that those words with similar meanings or that usually co-occur in the same contexts end up having similar embeddings.	O que torna essas **vetorizações** realmente úteis é o fato de que as palavras com significados semelhantes ou que comumente ocorrem nos mesmos contextos acabam por ter vetorizações semelhantes.
4	In fact, both the weights and the **embeddings** are learned at the same time.	Na verdade, tanto os pesos de ponderação como as **vetorizações** são aprendidos ao mesmo tempo.

Tabela 1: alguns contextos do uso de "vetor semântico" e "vetorização semântica" como traduções de *word embedding*.

4.2 Sentenças usadas como exemplo no texto de partida

Para ilustrar o processo de tradução tal como ele seria feito por uma rede neural e como ela contornaria alguns obstáculos que poderiam ser difíceis para computadores, os autores utilizaram algumas palavras e sentenças como exemplo. O primeiro desses exemplos diz respeito à palavra *letter*, como ilustrado na Figura 2.

Figura 2: uso da palavra *letter* como exemplo de polissemia

Words do not always have the same meaning in every sentence. The embedding of the word *letter*, for example, should not be the same when the word refers to a character of an alphabet or when it refers to a document addressed to another person.

Fonte: (KENNY, 2022)

Como o significado de algumas palavras homônimas pode depender do contexto de uso (em português, podemos citar a palavra "banco", que pode ser uma peça do mobiliário usada para sentar ou uma instituição financeira), as redes neurais devem decidir, ao traduzir do inglês, se a palavra *letter* significa "letra" ou "carta", por meio de uma técnica, chamada

atenção[4], cuja funcionalidade é tomar decisões sobre quais contextos serão usados para traduzir uma palavra polissêmica; por meio dessa técnica, a rede neural recebe um vetor para a palavra do texto de partida com seu significado descontextualizado e em seguida produz um vetor contextualizado indicando qual dos significados ela julga ser o mais provável e adequado para aquele contexto. Traduzir o exemplo dessa palavra, especificamente, constituiu uma pequena dificuldade que acabou sendo uma decisão importante para a continuidade do texto.

Por sorte, ambas as possíveis traduções de *letter* têm, em português, significados dependentes do contexto, o que permite preservar em português a ambiguidade desejada pelos autores para ilustrar o conceito. "Letra" é um termo que poderia se referir a um dos elementos do alfabeto, ao texto de uma canção ou mesmo à caligrafia de um indivíduo. Analogamente, "carta" pode se referir a uma comunicação escrita entre duas pessoas ou dois grupos, a um dos elementos de um baralho, ao cardápio de um restaurante ou a um documento probatório de qualquer natureza (como em "carta de aceite"). Essa polissemia encontrada no exemplo do texto de partida impõe ao tradutor a escolha entre traduzir o termo *letter* como "letra", "carta" ou alguma outra palavra com a característica de ter múltiplos significados distintos dependentes do contexto. Um pouco adiante no texto, os autores mencionam a possibilidade de a palavra poder receber dois vetores diferentes, no caso de se interpretar o termo *letter* como "carta", a depender do conteúdo da carta: uma carta de amor pode ter um vetor contextual no caso de ser uma carta de amor ou outro, no caso de ser uma carta de reclamação. Essa menção serviu de incentivo para traduzir *letter* como "carta", mas é interessante notar que um termo como "banco" também poderia cumprir o mesmo papel, com pequenas modificações adicionais: o texto traduzido poderia dizer que o termo "banco" teria vetorizações diferentes no caso de ser um "banco comercial" ou um conjunto de dados, como um "banco de dados", por exemplo. Como foi possível utilizar uma tradução válida do termo *letter* de modo que o argumento

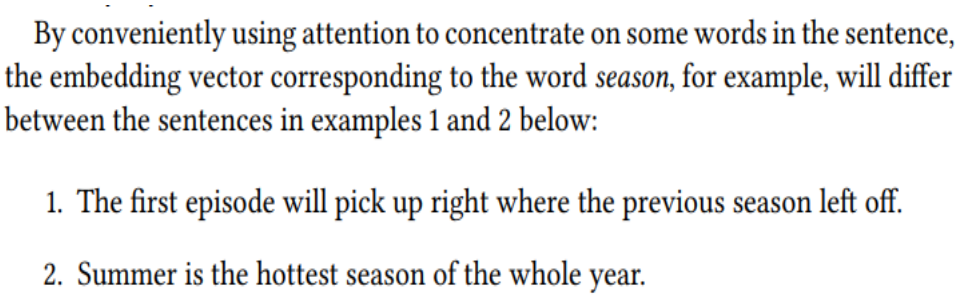
[4] Cf. <https://en.wikipedia.org/wiki/Attention_(machine_learning)>

central do texto ainda se sustentasse em português e como um termo alternativo como “banco” dificilmente traria alguma grande vantagem de modo imediato no entendimento do texto em português, optou-se pelo termo “carta” (em detrimento de “letra”).

Um outro exemplo de natureza similar, mas significativamente mais complexo, ocorreu quando os autores do texto de partida usaram outro exemplo, dessa vez aproveitando a polissemia da palavra *season* em inglês para exemplificar a abordagem e a técnica de atenção das redes neurais em casos de polissemia. Essa palavra pode ser traduzida principalmente como “estação (do ano)”, em um contexto meteorológico, ou como “temporada”, no contexto de uma série televisiva (em português, também é possível traduzir como “temporada” no contexto meteorológico, mas esse significado vem caindo em desuso).

Os autores usaram um par de frases para ilustrar a ambiguidade do termo *season* e, em seguida, se alongaram com a segunda frase para mostrar o processo tradutório de uma rede neural, como se observa na Figura 3.

Figura 3: Introdução do par de exemplos que ilustra a polissemia da palavra *season*



By conveniently using attention to concentrate on some words in the sentence, the embedding vector corresponding to the word *season*, for example, will differ between the sentences in examples 1 and 2 below:

1. The first episode will pick up right where the previous season left off.
2. Summer is the hottest season of the whole year.

Fonte: (KENNY, 2022)

A segunda frase em questão é “*Summer is the hottest season of the whole year*”. Em tese, é uma frase relativamente simples, que pode ser traduzida como “O verão é a estação mais quente do ano todo”. Entretanto, a análise feita em seguida pelos autores impediria essa tradução mais imediata, intuitiva e literal, pois eles decorrem longamente a respeito do papel que cada palavra cumpre na frase de origem e na sua relação com a palavra polissêmica em

estudo, *season*. Alguns dos elementos destacados pelos autores incluem a quantidade de palavras da frase, sua ordem (uma vez que as redes neurais apresentadas funcionam de modo a produzir o texto palavra a palavra, de modo análogo ao programa de previsão de textos de mensagens em aplicativos de email e celulares modernos) e suas classificações morfológicas. Eles ainda criaram um vetor de probabilidades com nove posições ocupadas por números para mostrar o papel desempenhado por cada palavra na frase (no exemplo dado, o terceiro elemento do vetor é 10%, indicando que a terceira palavra tem 10% de relevância na tradução da palavra *season*).

Em face disso, a frase foi traduzida por "O verão é a temporada mais quente do ano". Como em toda escolha tradutória, existem prós e contras. Entre as suas vantagens, é possível observar que ela conseguiu atender uma quantidade razoável de restrições impostas, listadas a seguir, pelo contexto de uso no texto de partida (que vai muito além do segmento onde ocorre pela primeira vez). Essas restrições têm o intuito de minimizar a quantidade de interferências necessárias no texto de origem, uma vez que os autores comentam longamente a respeito de elementos da estrutura da frase, como a quantidade de palavras na frase e a quantidade de palavras de cada classe gramatical, a ordem das palavras, entre outros. São elas:

- Preservar a ambiguidade do termo original (*season*);
- Preservar a quantidade de artigos da frase original, uma vez esta é mencionada diretamente em parte posterior do texto;
- Preservar a quantidade de palavras da frase, dada a importância que esse número tem na análise feita na sequência;
- Fazer com que a maior parte das palavras de classes gramaticais equivalentes ocupasse a mesma posição na frase;

Entre as desvantagens da tradução escolhida, pode-se destacar:

- O uso de “temporada” no lugar de “estação”, que seria uma escolha mais natural, tendo em vista que é a palavra mais usada como equivalente para esse sentido, em português brasileiro.
- Os artigos preservados não serem os mesmos (“o” e “a”, quando o original usava “the” duas vezes). O fato de a palavra *the* ser repetida no original tem importância na argumentação do texto de partida.
- A frase ter perdido parte da ênfase: o original destacava que a estação era a mais quente do ano *todo*, e não apenas do ano.
- Os vetores da palavra *season* precisaram sofrer uma leve alteração para continuarem ilustrando adequadamente o ponto que os autores pretendiam sustentar:
  - ✓ no original, o vetor era [25%, 8%, 10%, 15%, 25%, 8%, 2%, 0%, 7%];
  - ✓ na tradução, ficou [10%, 25%, 8%, 10%, 25%, 5%, 10%, 0%, 7%].

Algumas posições dos números precisaram ser alteradas para que os valores correspondessem ao ordenamento da frase original. Por exemplo, a palavra *summer* (25%) ocupa a primeira posição da frase de origem. Seu equivalente em português, “verão”, ocupa a segunda posição na frase-alvo. Para preservar a equivalência entre os números, o valor 25% foi posto na segunda posição da lista no vetor do texto em português. Além disso, a palavra *hottest*, que foi traduzida para “mais quente”, precisou ter seu valor de 15% decomposto em duas parcelas que também somassem 15%, de modo a refletir que as somas dos valores associados a “mais” e “quente” pudessem equivaler, de certa forma, ao original. Para tanto, escolheram-se os valores de 5% e 10%, respectivamente. Os números em si foram escolhidos por serem múltiplos de 5, já que a maioria dos números no vetor de origem também o eram, e com o intuito de dar um peso maior a “quente” do que a “mais” na tradução da palavra *hottest*. Observe-se que o impacto de sacrificar a ênfase dada pela palavra *whole* acabou não sendo

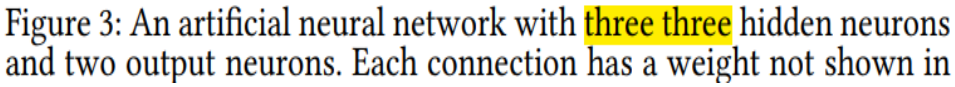
muito grande, dado que ela recebeu o valor de 0% de atenção no vetor usado como exemplo. Esta solução não é a mais rigorosa possível, o que corresponde a uma limitação genuína do método, uma vez que o processo da construção do vetor correspondente à frase em português usaria um outro corpus e muito provavelmente geraria outros números, mas foi considerada aceitável em virtude de sua didática e simplicidade. Ademais, redes neurais diferentes poderiam obter resultados diferentes no processo, o que é indicativo de que não há um resultado definitivo para se criar essa correspondência entre a posição e o valor dos números nos vetores do texto em inglês e português.

Outras traduções seriam possíveis e válidas, mas a escolhida foi considerada aceitável em virtude dos princípios exemplificados, que puderam ser respeitados com ela; outras alternativas poderiam funcionar melhor como tradução da frase fora de seu contexto como exemplo de uma explicação, mas provavelmente incorreriam em várias outras alterações em partes posteriores para justificá-las retroativamente, o que poderia ser uma interferência excessiva no material de origem. Em certo sentido, é possível traçar um paralelo entre a tradução dessa frase em seu contexto com a tradução de um poema no qual se desejasse prezar não apenas pelo significado das palavras na fonte, mas também por questões como métrica, rimas, estilo e ambientação histórica do poema.

4.3 Erros tipográficos e questões de estilo

O capítulo do livro em estudo ainda apresentou alguns problemas tipográficos que não deveriam ser carregados para a tradução, pois não contribuiriam com o entendimento do público acerca do assunto. A Figura 4 ilustra um dos erros encontrados, a saber, a repetição desnecessária da palavra *three*.

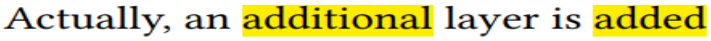
Figura 4: legenda com provável repetição acidental da palavra *three*.



Fonte: (KENNY, 2022)

Talvez mais relevantes que os poucos erros de tipografia do texto original tenham sido algumas questões estilísticas, que também têm baixo impacto direto no entendimento do texto, mas que podem ser relevantes na fluidez da leitura. Um exemplo simples deste ponto é o trecho da Figura 5, onde se lê que uma camada *adicional* é *adicionada*.

Figura 5: Repetição estilisticamente questionável de palavras derivadas de *add.*



Fonte: (KENNY, 2022)

É possível argumentar que não há nada de objetivamente errado com o trecho em seu contexto e que a repetição é justificada, estilisticamente, por uma questão de ênfase ao ato de adicionar uma camada, além de que a repetição no inglês ser muito menos mal vista do que no português, tendo em vistas os textos diversos com repetições abundantes que podem ser encontrados no inglês. Entretanto, é possível atingir o mesmo efeito em português acrescentando variedade ao vocabulário, o que costuma configurar como boa prática em manuais de estilo (WAHL, 2016). Por exemplo, seria possível expressar essa ideia como “uma camada extra é acrescentada”, “uma camada adicional é acrescentada”, “uma camada extra é adicionada” sem perda do senso enfático à ideia de adicionar e evitando a repetição sonora do par “adicional”/“adicionada”, que poderia causar estranhamento no leitor e possivelmente a impressão de redundância.

4.4 Graus de formalidade em textos de divulgação científica em inglês e em português

A linguagem acadêmica, embora compartilhe de elementos como a formalidade e a estruturação textual em diferentes idiomas, apresenta variações significativas entre o inglês e

o português. Em inglês, textos científicos tendem a favorecer uma estrutura mais direta e concisa, com menor uso de construções passivas e um estilo menos rebuscado, com o objetivo de facilitar a leitura e o entendimento (WILLIAMS; NADEL, 1989). Há uma preferência por frases curtas e parágrafos bem delimitados, valorizando a clareza da informação e a brevidade. Por outro lado, a linguagem acadêmica em português frequentemente se caracteriza por um estilo mais formal e rebuscado. Não é raro encontrar frases longas e complexas, com uso extensivo de construções elaboradas e uma maior variação de vocabulário, caracterizado por escolhas pouco comuns ao uso cotidiano da língua. Além disso, o uso da voz passiva é uma característica marcante.

O texto escolhido, apesar de ser um texto de divulgação voltado para o público geral, tem sua origem e potencial público-alvo na comunidade acadêmica. No trecho a seguir, é possível identificar uma transição na escala formalidade: de casual, em que os autores se referem ao leitor diretamente, ao utilizar a segunda pessoa do discurso, típico dos artigos de divulgação científica em inglês, para um tom mais formal e impessoal, que busca manter um distanciamento entre autor e leitor, um estilo de escrita mais bem aceito em textos acadêmicos de divulgação em português, como mostra o trecho a seguir:

227	At this point, you are hopefully in a good position to understand how NMT works, even if we describe its fundamentals in only a few sentences as we do next.	Neste ponto, o leitor deste artigo deve ter condições suficientes para entender como NMT funciona, mesmo que os seus fundamentos sejam descritos em poucas frases, como será feito a seguir.

4.5 Estratégias de tradução de unidades terminológicas

As principais soluções adotadas para traduzir unidades terminológicas encontradas no texto de partida, como apontado na fundamentação teórica, foram o empréstimo, a tradução literal, o decalque e a paráfrase. Como já elucidado anteriormente, a natureza do empréstimo e do decalque é similar, dado que o decalque, ou "aclimatação", é o "processo através do qual

os empréstimos são adaptados à língua que os toma" (BARBOSA, 2004, p.73), ou seja, a incorporação de uma estrutura ou significado da língua de partida; enquanto a paráfrase é, grosso modo, uma forma de explicação.

O decalque pode ser tanto morfológico quanto lexical. No primeiro caso, ele é uma adaptação da morfologia de um estrangeirismo ao português (BARBOSA, 2004). Um exemplo do fenômeno é ilustrado na palavra "deletar". No segundo caso, o decalque lexical pode ser classificado como um estrangeirismo ou empréstimo, que, segundo Barbosa (2004),

> "[...] consiste em transferir (transcrever ou copiar) para o TLT vocábulos ou expressões da LO que se refiram a um conceito, técnica ou objeto mencionado no TLO que seja desconhecido para os falantes da LT." (BARBOSA, 2004, pp. 71-72)

Aqui, TLT refere-se ao "Texto na Língua de Tradução", LO à "Língua Original" e TLO ao "Texto na Língua Original". Esse processo também pode envolver a transferência de sentido mantendo um sintagma já existente na língua de chegada, como "aplicar" (para uma vaga) para "apply", no sentido de "inscrever-se". No contexto de redes neurais, aprendizado de máquina e inteligência artificial, a palavra "*generalisation*" pode ser traduzida com uma abordagem de decalque lexical, como pode-se observar nos seguintes fragmentos da tradução:

Figura 6: Segmento onde a palavra "generalisation" foi traduzida por meio de decalque lexical

123	Similarly, generalisation happens when an organism which already responds to a certain stimulus in a particular way responds to similar stimuli in similar ways.	Analogamente, a generalização acontece quando um organismo que já reage a um determinado estímulo de uma maneira particular, responde de formas semelhantes a estímulos parecidos.

Fonte: autoria própria.

Apesar de a palavra "generalização" existir e ser usada no português com o sentido de aplicar a todos elementos de um grupo um conceito geral, no discurso, ela é sempre

acompanhada de um verbo, como “fazer generalizações” ou “aplicar generalizações”, podendo este ser considerado o uso mais natural e corrente do termo.

Um exemplo análogo à tradução de “*generalisation*” por “generalização” no contexto da tradução do capítulo 7 seria o de traduzir “*to overthink*” para “sobrepensar”, dado que a expressão considerada mais natural seria algo próximo da expressão “pensar demais”. No entanto, o decalque “sobrepensar” não seria tão bem aceito quanto “generalização”, conforme usada no presente contexto. O decalque lexical na referida tradução se dá pela aplicação do significado técnico que existe somente na palavra em inglês à palavra “generalização”, em português, e pelo uso da palavra sem estar acompanhada de um verbo, um uso típico do inglês.

Empréstimos linguísticos são fenômenos comuns em todas as línguas e ocorrem quando uma língua incorpora palavras ou expressões de outra, geralmente devido à influência cultural, tecnológica ou científica. Eles são caracterizados pela preservação morfológica e semântica de termos ao transferi-los para outro idioma e são bastante apropriados em caso de falta de equivalentes diretos ou em áreas amplamente globalizadas. Mesmo neste trabalho, cujo contexto também é resistência a anglicismos, o empréstimo pode ser justificado por razões práticas relacionadas à predominância do inglês como língua franca.

Um exemplo de estrangeirismo ou empréstimo na tradução do capítulo 7 é a manutenção da palavra “transformer” para se referir a um modelo de rede neural. A manutenção da palavra conforme está no inglês pode ser vista como uma escolha pragmática, posto que a palavra em inglês já está reconhecida e difundida na literatura científica nacional e internacional. Além disso, é importante levar em consideração a prevalência da globalização na área da tecnologia e a praticidade de se ater à terminologia que já está padronizada e que facilita a busca de artigos sobre o assunto.

A palavra “transformador” já tem um significado específico e bem estabelecido em português, principalmente no contexto da engenharia elétrica, referindo-se a um dispositivo

que transfere energia elétrica entre dois ou mais circuitos. Portanto, a utilização desse termo para se referir a um modelo de rede neural poderia causar confusão e imprecisão técnica. A manutenção de "*transformer*" evita essa ambiguidade, mantendo clara a distinção entre conceitos. Nesse sentido, a escolha de preservar a palavra "*transformer"* em textos traduzidos pode ser vista não como um abandono da língua portuguesa, mas como um reconhecimento prático das necessidades de um campo de estudo globalizado e como um meio de evitar ambiguidades. Também é relevante ressaltar que foram feitas buscas relativas ao empréstimo "transformer" usando a *web* como corpus, tendo sido encontrados números significativos de utilizações do termo.

Em um campo dominado por terminologias em inglês, priorizar termos que refletem mais precisamente o significado e uso na língua portuguesa pode ser visto como um passo positivo na preservação da identidade linguística e cultural. Ainda que estratégias de empréstimo tenham sido empregadas neste trabalho de tradução, como no caso da palavra "*transformer*", o objetivo ainda é reforçar a autonomia e a riqueza cultural do português. Desse modo, mesmo a tradução de "*natural language processing*" ter sido, até certo ponto, consagrada em português como "processamento de linguagem natural", optou-se, na tradução, pelo uso do termo "processamento de língua natural", visto que essa substituição reflete a distinção semântica presente na língua portuguesa entre "língua" e "linguagem". Embora ambas as opções, "processamento de linguagem natural" e "processamento de língua natural" sejam exemplos de tradução literal, o termo "linguagem" pode se referir a qualquer sistema de sinais ou símbolos usados para comunicação, como linguagem corporal ou linguagem de programação, enquanto a palavra "língua" é mais especificamente associada à fala e à escrita, que são capacidades exclusivamente humanas.

Um outro desafio importante da tradução do sintagma "*natural language processing*" é a semelhança com "*neural language processing*". A tradução de "*neural language processing*"

se distancia da tradução direta que deve ser usada para "*natural language processing*" (processamento de língua natural), visto que "*natural*", em "*natural language processing*" se refere à "*language*" (língua) e "*neural*", em "neural language processing" (processamento neural de língua) se refere a "processing" (processamento) e essa é uma característica que poderia confundir alguns tradutores que chegariam em traduções como "processamento de linguagem neural" em vez de "processamento neural de língua", um distanciamento de significado que não só vem da ordem das palavras como da similaridade entre os termos quando estão em inglês. As traduções que melhor reproduziriam os significados dos termos são "processamento de língua natural", para "natural language processing", e "processamento neural de língua" para "neural language processing", como no quadro a seguir:

132	In the field of **natural language processing**, and as indicated above, the information processed by neural networks is made up of words, and their representations within the network are usually referred to as embeddings (Mikolov et al. 2013).	No campo do **processamento de língua natural**, e conforme já foi mostrado, as informações processadas pelas redes neurais são compostas de palavras, e suas representações na rede costumam ser chamadas de vetores (*embeddings*) (Mikolov et al. 2013).
163	As sentence representations are obtained from word embeddings, we may conclude that representing similar words with similar numbers is a precondition for generalisation in **neural natural language processing**.	Como as representações de frases são obtidas a partir de vetorizações de palavras, pode-se concluir que representar palavras semelhantes com números semelhantes é uma condição prévia para a generalização no **processamento neural de língua natural**.

Além das outras estratégias, é essencial abordar como a paráfrase foi estrategicamente utilizada para adequar o registro do texto original, que apresentava um estilo informal em inglês, para um registro mais adequado ao público acadêmico que lê em português. A paráfrase, neste contexto, não se limita à tradução dos termos, mas envolve uma reestruturação das frases para alinhar o texto com as normas e expectativas formais da escrita acadêmica em português. Um exemplo claro dessa abordagem é a tradução do exclamativo "you are right!" para "E essa é uma perspectiva correta", como no segmento abaixo:

16	And you are indeed right!	E essa é uma perspectiva correta.

Aqui, a paráfrase suaviza o tom direto e informal do inglês e reformula a frase de modo a manter a essência do conteúdo, ao mesmo tempo em que confere um caráter mais formal e reflexivo, alinhado com as convenções da escrita acadêmica em português. Essa escolha reflete um esforço consciente para garantir que a tradução não apenas transmita a informação de forma precisa, mas também respeite as nuances culturais e linguísticas da língua de chegada.

4.6 Considerações finais sobre a tradução

As estratégias de tradução adotadas foram guiadas pelo princípio de clareza e acessibilidade para o público-alvo e foram embasadas por estudos teóricos da tradução especializada e da terminologia. Em casos como o termo *word embedding*, optamos por soluções que equilibrassem a fidelidade ao conceito original com a compreensibilidade para leitores em português. Essa escolha reflete um compromisso com a transmissão eficaz de conhecimento técnico, respeitando as características linguísticas e culturais do português brasileiro, bem como uma rejeição ao fenômeno de imperialismo linguístico do inglês em relação a outras línguas.

Nas considerações finais deste capítulo, destaca-se a escolha intencional de limitar os empréstimos linguísticos em favor da preservação da língua portuguesa. Esta abordagem, enfatizando técnicas como decalque e paráfrase, visa manter a autenticidade e riqueza do idioma. Um exemplo notável é a tradução de "natural language processing" para "processamento de língua natural" em vez do mais comum "processamento de linguagem natural". Esta escolha reflete não apenas a busca por precisão técnica, mas também o respeito pelas nuances semânticas específicas do português. Ao evitar anglicismos desnecessários, a tradução não só facilita a compreensão do leitor, mas também reforça o valor e a identidade linguística do português no contexto acadêmico e científico global.

5 CONCLUSÃO

No cenário contemporâneo, a tecnologia, especialmente a tradução automática, assume um papel crucial em diversas esferas da sociedade, transformando significativamente as formas de comunicação e o ato de traduzir. As ferramentas de tradução automática, que evoluem rapidamente com o desenvolvimento de redes neurais, não só aceleram o processo de tradução, permitindo aos tradutores trabalhar mais rapidamente e com maior qualidade, mas também democratizam o acesso à informação e, simultaneamente, desafiam os tradutores a aprimorar suas habilidades. A habilidade de traduzir com sensibilidade cultural e precisão técnica continua sendo um elemento chave, o que reafirma a relevância contínua dos profissionais da tradução. Contudo, para se manterem alinhados com esta evolução tecnológica, é vital que tanto os tradutores quanto o público em geral busquem um entendimento aprofundado sobre o funcionamento da tradução automática para que sejam capazes de empregar estas ferramentas de forma mais consciente dos benefícios e desvantagens.

Para a elaboração deste trabalho, foram discutidos os aspectos mais relevantes entre aqueles que levaram à escolha do texto a ser traduzido neste projeto. Foram discutidos o papel das traduções automáticas no desempenho profissional do tradutor e a prevalência de seu uso para indivíduos que precisam ou querem traduzir palavras ou textos sem ter nenhum conhecimento de outro idioma. Além disso, entender seu funcionamento é útil ao profissional do texto, pois permite que a incorpore de modo proveitoso em seu trabalho.

O tema do desenvolvimento de redes neurais para traduções automáticas também foi discutido em algum detalhe, não apenas por sua relevância histórica, mas também para destacar o quanto a área é recente e está em estado de constante transformação. A frequência com que esses avanços ocorrem faz aumentar a importância de várias categorias profissionais, entre as quais se encontram o divulgador científico e o tradutor especializado. É possível notar que ambas as profissões, em certo sentido, cumprem parcialmente o papel uma da outra: é possível

interpretar o trabalho do divulgador científico como o de alguém que traduz a linguagem acadêmica, muitas vezes considerada densa e obscura por quem não tem conhecimento especializado, para a linguagem popular, fazendo com que o grande público possa ter uma compreensão melhor do mundo a sua volta.

Analogamente, é possível entender o papel do tradutor especializado como o de divulgador científico, ao aumentar a exposição dos avanços científicos para o público de um determinado idioma. Por fim, o aumento de produtividade proporcionado por traduções automáticas, aliado à abrangência comunicativa permitida por avanços relativamente recentes como a própria internet, faz também aumentar a demanda por traduções especializadas. Atualmente, elas ocupam a maior parte do mercado editorial de tradução (BYRNE, 2014).

Ao longo do texto, fez-se necessário considerar aspectos técnicos, estilísticos e situacionais para garantir um resultado satisfatório. Considerações técnicas foram necessárias, naturalmente, para que o texto produzido fosse adequado terminologicamente, mas elas também precisaram ser conciliadas com questões estilísticas e situacionais, a fim de que a tradução soasse idiomática, fluida e agradável ao público brasileiro do século XXI.

Questões filosóficas também vieram à tona em várias decisões tradutórias, pois todas elas foram norteadas por tentativas de respostas a perguntas que nem sempre são fáceis, como: qual o papel da divulgação científica? Qual o papel da tradução, no geral, e dentro desse gênero textual? O que o público-alvo entenderá do texto? Uma outra consideração relevante que apareceu em vários casos, foi a do impacto específico de uma tradução na língua portuguesa dentro de um tema que não possui muito material nessa língua. Quando há pouco material disponível acerca de um determinado tema em uma língua, pode ocorrer de as primeiras traduções de determinados termos serem as que, por terem surgido primeiro, alcançarem maior abrangência e se tornarem as versões consagradas desses termos. É interessante, portanto, que

essas traduções sejam tão precisas quanto possível, a fim de não fossilizar anglicismos ou imprecisões na língua de chegada.

Com este trabalho, espera-se ter contribuído positivamente com a literatura em português brasileiro a respeito de traduções especializadas e tradução automática, seja fornecendo a profissionais da tradução soluções linguísticas em português para a tradução de certos termos ou oferecendo ao público brasileiro um texto informativo a respeito de uma ferramenta cada vez mais prevalente no cotidiano de todas as pessoas.

REFERÊNCIAS BIBLIOGRÁFICAS

BAHDANAU, Dzmitry; CHO, Kyunghyun; BENGIO, Yoshua. **Neural machine translation by jointly learning to align and translate**. arXiv preprint arXiv:1409.0473, 2014.

BARBOSA, H. **Procedimentos Técnicos da Tradução: uma nova proposta**. Campinas: Editora Pontes, 2004.

BEVILACQUA, Cleci Regina; KILIAN, Cristiane Krause. **Tradução e terminologia: relações necessárias e a formação do tradutor**. Domínios de Lingu@ gem, v. 11, n. 5, p. 1707-1726, 2017.

BOJAR, Ondřey et. al. **Findings of the 2016 Conference on Machine Translation (WMT16)**. Disponível em: <https://web.archive.org/web/20180127202851/https://cris.fbk.eu/retrieve/handle/11582/307240/14326/W16-2301.pdf>

BYRNE, JODY (Ed.). **Technical translation**. Dordrecht: Springer Netherlands, 2006.

CABRÉ, Maria Teresa. **Terminology: Theory, methods, and applications**. John Benjamins Publishing, 1999.

CASELI, Helena; FREITAS, Cláudia; VIOLA, Roberta. Processamento de Linguagem Natural. **Sociedade Brasileira de Computação**, 2022.

COLDEWAY, Devin; LARDINOIS, Frederic. **DeepL schools other online translators with clever machine learning**. Disponível em: <https://techcrunch.com/2017/08/29/deepl-schools-other-online-translators-with-clever-machine-learning>

GU, Xinxing. **New features make Translate more accessible for its 1 billion users**. Disponível em: <https://blog.google/products/translate/new-features-make-translate-more-accessible-for-its-1-billion-users/>

JENKINS, Jennifer. **English as a lingua franca in the international university: The politics of academic English language policy**. Routledge, 2013.

KENNY, Dorothy. **Machine translation for everyone: Empowering users in the age of artificial intelligence**. Language Science Press, 2022.

LIMA, Elon Lages. **Elementos de topologia geral**. Ao Livro Técnico, Editora da Universidade de São Paulo, 1970.

OLOHAN, M. **Scientific and Technical Translation**.New York: Routledge, 2016.

PHILLIPSON, Robert. **Linguistic imperialism**. Oxford University Press, 1992.

ROGERS, Margaret; ROGERS, Margaret. **Specialised Translation: An Orientation**. Specialised Translation: Shedding the 'Non-Literary'Tag, p. 20-42, 2015.

RUSH HOVDE, Marjorie. **Creating procedural discourse and knowledge for software users: Beyond translation and transmission**. Journal of Business and Technical Communication, v. 24, n. 2, p. 164-205, 2010.

TUROVSKY, Barak. **Found in translation: More accurate, fluent sentences in Google Translate**. Google, 2016. Disponível em: <https://blog.google/products/translate/found-translation-more-accurate-fluent-sentences-google-translate/>

SHANKLAND, Stephen. **Google Translate now serves 200 million people daily**. Disponível em: <https://www.cnet.com/tech/services-and-software/google-translate-now-serves-200-million-people-daily/>. Acesso em: 30 out. 2023.

STOLZE, R**. Die Fachübersetzung: eine Einführung.** Tübingen: Narr, 1999.

WAHL, Daniel. **The Sense of Style: The Thinking Person's Guide to Writing in the 21st Century**, by Steven Pinker. The Objective Standard, v. 11, n. 1, 2016.

WILLIAMS, Joseph M.; NADEL, Ira Bruce. **Style: Ten lessons in clarity and grace**. Glenview, IL: Scott, Foresman, 1989.

APÊNDICE I

Tradução do capítulo 7 de *Machine Translation for Everyone*

№	Partida (EN-GB)	Chegada (PT-BR)
1	Chapter 7	Capítulo 7
2	How neural machine translation works	Como funciona a tradução automática neural
3	Juan Antonio Pérez-Ortiz	Juan Antonio Pérez-Ortiz
4	Universitat d'Alacant, Spain	Universidade de Alicante, Espanha
5	Mikel L. Forcada	Mikel L. Forcada
6	Universitat d'Alacant, Spain	Universidade de Alicante, Espanha
7	Felipe Sánchez-Martínez	Felipe Sánchez-Martínez
8	Universitat d'Alacant, Spain	Universidade de Alicante, Espanha
9	This chapter presents the main principles behind neural machine translation systems.	Este capítulo apresenta os princípios fundamentais subjacentes aos sistemas de tradução automática neural.
10	We introduce, one by one, key concepts used to describe these systems, so that the reader achieves a comprehensive view of their inner workings and possibilities.	São introduzidos, um a um, conceitos-chave utilizados para descrever esses sistemas, para que o leitor obtenha uma visão abrangente de seu funcionamento interno e possibilidades.
11	These concepts include: neural networks, learning algorithms, word embeddings, attention, and the encoder–decoder architecture.	Esses conceitos incluem: redes neurais, algoritmos de aprendizagem, vetorização de palavras, mecanismos de atenção e arquitetura codificador–decodificador.
12	1 Introduction	1 Introdução
13	The first thing you should know about neural machine translation (NMT) is that it considers translation as a task involving operations on numbers performed by mathematical systems called artificial neural networks: these systems take a sentence and transform it into a series of numbers.	A primeira coisa que se deve entender em relação à tradução automática neural (NMT, do inglês Neural Machine Translation), é que ela considera a tradução como uma tarefa que envolve operações em números realizadas por sistemas matemáticos chamados redes neurais artificiais: esses sistemas tomam uma sentença e a transformam em uma série de números.
14	They add some more numbers here (usually, thousands or millions of them), multiply by other numbers there, perform a few additional, relatively simple, mathematical operations, and eventually output a translation of the original sentence into another language.	Adicionam mais alguns números aqui (geralmente, milhares ou milhões deles), multiplicam por outros números ali, realizam algumas operações matemáticas adicionais relativamente simples e, por fim, produzem uma tradução da sentença original para outro idioma.
15	Maybe you have always considered translation from a different perspective: as an intellectual task that involves cognitive processes which can barely be explicitly enumerated and which take place in some deep areas of the human brain.	Muitos enxergam a tradução de uma perspectiva diferente: como uma tarefa intelectual que envolve processos cognitivos que dificilmente podem ser enumerados com clareza e que ocorrem em algumas áreas profundas do cérebro humano.
16	And you are indeed right!	E essa é uma perspectiva correta.
17	But the approximation currently carried out by computers follows a completely different path: millions of mathematical operations are performed in a fraction of a second to obtain a translation which may sometimes be labelled as adequate and may sometimes not.	Mas a abordagem executada atualmente por computadores segue um caminho completamente diferente: milhões de operações matemáticas são realizadas em uma fração de segundo para obter uma tradução que ora pode, ora não pode ser considerada adequada.
18	And it turns out that the percentage of times they happen to be adequate has dramatically in the last few years.	E o fato é que a percentagem de vezes em que a tradução é adequada aumentou dramaticamente nos últimos anos.

19	But, historically, artificial neural networks were devised as a simplified model of how natural neural networks such as our brains work, and the cognitive processes carried out in it are also the result of distributed neural computation processes which are not that different from the mathematical operations mentioned above.	Mas, historicamente, as redes neurais artificiais foram concebidas como um modelo simplificado de como funcionam as redes neurais naturais, como o cérebro humano, e os processos cognitivos realizados nelas são também o resultado de processos difusos de computação neural que não são tão diferentes das operações matemáticas mencionadas acima.
20	This chapter will teach you the key elements of NMT technology.	Este capítulo explicará em detalhe elementos-chave da tecnologia NMT.
21	We will start off by pointing out the connection between how translation could be carried out in a human brain and how an NMT system undertakes it.	A começar por destacar a conexão entre as formas possíveis de traduzir de um cérebro humano e de um sistema NMT.
22	This will help us to introduce the basic concepts needed to get a comprehensive overview of the principles of machine learning and artificial neural networks, which constitute two of the cornerstones of NMT.	Explicar essa conexão ajudará na apresentação dos conceitos básicos necessários para obter uma visão abrangente dos princípios de aprendizado de máquina e redes neurais artificiais, que constituem dois dos pilares da NMT.
23	After that, we will discuss the essential principles of non-contextual word embeddings, a computerised representation of words with many interesting properties that, when combined through a mechanism known as attention, will produce the so-called contextual word embeddings, a key factor in the realisation of NMT.	Em seguida, serão discutidos os princípios essenciais de vetores não contextuais (non-contextual word embeddings), uma representação computadorizada de palavras com diversas propriedades interessantes que, quando combinadas através de um mecanismo conhecido como "atenção", produz os chamados vetores de palavras contextuais, um fator-chave no entendimento da NMT.
24	All these ingredients will allow us to present an overall picture of the inner workings of the two most used NMT models, namely, the transformer and the recurrent models.	Todos esses ingredientes permitirão apresentar um quadro geral do funcionamento interno dos dois modelos de NMT mais utilizados, a saber, o transformer e os modelos de redes neurais recorrentes.
25	The chapter wraps up by introducing a series of secondary themes that will improve your knowledge on how these systems run behind the scenes.	O capítulo termina com a apresentação de uma série de temas secundários para expandir o conhecimento público sobre a forma com que estes sistemas funcionam por trás das cortinas.
26	2 An imperfect analogy between human translation and NMT	2 Uma analogia imperfeita entre tradução humana e NMT
27	To simplify the discussion a bit, let us make the radical approximation that translating a text is equivalent to translating each of its sentences independently of each other.	Para simplificar um pouco a discussão, vamos imaginar que traduzir um texto equivale a, grosso modo, traduzir cada uma das suas sentenças de forma independente.
28	Let us now assume for a minute that translating a sentence is a two-step process: the translator first determines the interpretation or meaning of the whole source sentence and then produces in one go a sentence that allows more or less the same interpretation, but is now written in the target language.	Suponha, por um momento, que a tradução de uma sentença é um processo de duas etapas: o tradutor primeiro interpreta ou determina o significado de toda a sentença de origem e, em seguida, produz de uma só vez uma sentença que permite mais ou menos a mesma interpretação, mas que agora está escrita na língua de chegada.
29	But every day translators encounter sentences that they have never seen before, such as "The pencil slipped from my hand, stood up, and started talking to me", and can still translate them: how is that possible?	Mas, todos os dias, tradutores encontram sentenças que nunca viram antes, como "o lápis escorregou da minha mão, levantou-se e começou a falar comigo", e ainda conseguem traduzi-las: como isso é possível?
30	Linguistics has formulated the answer to this question as a principle, the principle of semantic compositionality: we build the interpretation of each sentence by combining	A linguística formulou a resposta a essa questão como o princípio da composicionalidade semântica: os seres humanos constroem a interpretação de cada sentença combinando as interpretações individuais

	the individual interpretations of its component words, and the order in which they are combined is dictated by the syntactic structure of the sentence in which words form phrases, phrases form larger phrases, until one gets to the whole sentence.	de suas palavras e a ordem em que são combinadas é ditada pela estrutura sintática da sentença em que as palavras formam orações, orações formam orações maiores, até chegar a toda a sentença.
31	A translator would then analyse this interpretation and perform the inverse procedure, but in the target language.	Um tradutor então analisa essa interpretação e executa o procedimento inverso, mas na língua de chegada.
32	Of course, translators do not always process sentences as a whole, particularly when they are long, and they may take shortcuts to avoid building interpretations of whole sentences, but let us stick to this simplification for a while.	É evidente que os tradutores nem sempre interpretam as sentenças como um todo, especialmente ao se depararem com sentenças longas, podendo recorrer a atalhos para evitar interpretações de sentenças inteiras logo de início. Contudo, vamos manter essa simplificação, por enquanto.
33	NMT works in a similar way.	NMTs funcionam de forma semelhante.
34	When translating a sentence, during the encoding phase, the system assigns a neural representation, or embedding, to each source-text word in isolation.	Ao traduzir uma sentença, durante a sentença de codificação, o sistema atribui uma representação, ou vetorização, para cada palavra do texto-fonte isoladamente.
35	These neural representations are then combined to produce a similar representation, but this time at sentence level.	Essas representações neurais são então combinadas para produzir uma representação semelhante, mas desta vez no nível da sentença.
36	As they are combined, individual representations are also modified according to their context; one could consider this a contextualised representation of interpretation or meaning.	À medida que são combinadas, as representações individuais também são modificadas de acordo com o seu contexto; pode-se considerar isso uma representação contextualizada de interpretação ou significado.
37	Then, in the decoding phase, the sentence-level representations are unravelled step by step to predict, one by one, the words in the target sentence.	Em seguida, na fase de decodificação, as representações das sentenças são desvendadas passo a passo para prever, uma a uma, as palavras na sentença-alvo.
38	The encoder and the decoder performing these two phases are artificial neural networks interconnected into a single composite neural network.	O codificador e o decodificador que executam essas duas fases são redes neurais artificiais interconectadas que formam uma única rede neural composta.
39	As in the case of translators, current neural architectures do not really work by considering the whole source sentence when producing each target word, but rather have learned to pay attention to the relevant source words and the target words already produced when they do so.	Como no caso dos tradutores, as arquiteturas neurais atuais não funcionam, de fato, considerando toda a sentença de origem ao produzir cada palavra-alvo, mas aprenderam a prestar atenção nas palavras-fonte relevantes e nas palavras-alvo já produzidas quando fazem isso.
40	In the remaining sections of this chapter we will describe in more detail the nature of these representations, the structure of the artificial neural networks (which we may simply call ”neural networks” from now on) that build and transform them by selectively paying attention to what is important, and the ways in which these artificial neural networks can be trained to do this task using translation examples.	Nas seções restantes deste capítulo, o leitor será apresentado a uma descrição mais detalhada da natureza destas representações, à estrutura das redes neurais artificiais (que podemos simplesmente chamar de "redes neurais", a partir de agora) que as constroem e transformam, prestando atenção seletivamente ao que é importante, e às formas como estas redes neurais artificiais podem ser treinadas para realizar tal tarefa utilizando exemplos de tradução.
41	3 Artificial neural networks	3 Redes Neurais Artificiais
42	To make sense of NMT, one needs to consider in more detail the artificial neural networks (Goodfellow et al. 2016) that perform it: what they are made of, how they work and how they are trained.	Para entender o conceito de NMT, é preciso considerar com mais detalhes as redes neurais artificiais (Goodfellow et al. 2016) que a realizam: do que são feitas, como funcionam e como são treinadas.

43	The name neural clearly invokes neurons and the way in which the nervous systems of animals, and particularly people's brains, work.	O adjetivo "neural" remete diretamente a neurônios e à maneira como funcionam os sistemas nervosos de animais e, principalmente, do cérebro humano.
44	Artificial neural networks are indeed made up of thousands or millions of artificial units that resemble neurons whose activation (that is, how excited or inhibited they are) depends on the signals they receive from other neurons and the strength of the connections carrying these signals.	As redes neurais artificiais são, de fato, constituídas por milhares ou milhões de unidades artificiais que se assemelham a neurônios cuja ativação (ou seja, o quanto estão excitados ou inibidos) depende dos sinais que recebem de outros neurônios e da força das conexões que transmitem esses sinais.
45	3.1 Artificial neurons	3.1 Neurônios artificiais
46	Artificial neurons are the main building blocks of artificial neural networks.	Os neurônios artificiais são os principais blocos de construção das redes neurais artificiais.
47	These artificial neurons (we will simply call them neurons from now on) may be seen as operating in two steps when updating their state or activation.	A operação desses neurônios artificiais (que, a partir de agora, chamaremos simplesmente de neurônios) de atualizar seu estado ou ativação pode ser analisada em duas etapas.
48	Let us imagine the simple situation in Figure 1 in which we study how the activation of neuron □4 is updated in response to stimuli received from neurons □1, □2, and □3.	Apresenta-se, na Figura 1, uma situação simplificada, na qual se observa como o grau de ativação do neurônio4 é atualizado em decorrência dos estímulos recebidos pelos neurônios 1, □2, e S3.
49	Figure 1: Updating the state □4 of artificial neuron 4 in	Figura 1: atualização do estado S4 do neurônio 4 em
50	response to stimuli received from neurons 1, 2 and 3.	resposta aos estímulos recebidos dos neurônios 1, 2 e 3.
51	In the first step, the activations of neurons □1, □2 and □3, all of them connected to neuron □4, are added, but first each one is multiplied by a weight (□1, □2 and □3) representing the strength of their connections; these weights determine how their activations are turned into actual stimuli for neuron □4.	Na primeira etapa, os graus de ativação dos neurônios S1, □2 e □3, todos eles ligados ao neurônio4, são somados, mas primeiro cada um é multiplicado por um peso (□1, □2 e □3) representando a força de suas conexões; esses pesos determinam como sua ativação é transformada em estímulos reais para o neurônio S4.
52	Weights may be positive or negative.	Os pesos podem ser positivos ou negativos.
53	For instance, if weight □2 is positive and the activation of □2 is high, it will contribute to exciting neuron □4 (a positive stimulus); if, however, □2 is negative, it will contribute to inhibiting neuron □4 (a negative stimulus).	Por exemplo, se o peso w2 é positivo e o grau de ativação de S2 é elevado, ele contribuirá para ativar o neurônio S4 (um estímulo positivo); no entanto, se w2 é negativo, contribuirá para inibir o neurônio S4 (um estímulo negativo).
54	In general terms, neurons connected through positive weights tend to be simultaneously excited or inhibited, while neurons connected through negative weights tend to be in opposite states.	Em termos gerais, os neurônios conectados por pesos positivos tendem a ser ativados ou inibidos simultaneamente, enquanto os neurônios conectados por pesos negativos tendem a estar em estados opostos.
55	Coming back to neuron □4, if we add the stimuli coming from each neuron, we get a net stimulus:	Retornando à análise do neurônio S4, ao se adicionar os estímulos provenientes de cada neurônio, obtém-se um saldo líquido de estímulo:
56	(1)	(1)
57	The net stimulus □ can take any possible value, negative or positive, but it is not the activation of neuron □4 yet.	O estímulo líquido pode assumir qualquer valor possível, negativo ou positivo, mas ainda não é a ativação do neurônio S4 .
58	In the second step, neuron □4 reacts to this stimulus.	Na segunda etapa, o neurônio S4 reage a este estímulo.
59	In the example, when the stimulus is intermediate, that is, not too positive or too negative, the neuron □4 is very sensitive to it.	No exemplo, quando o estímulo é intermediário, ou seja, não muito positivo ou muito negativo, o neurônio S4 é muito sensível a ele.
60	However, when stimuli get large (no matter if positive or negative), changes in their values	No entanto, quando os estímulos ficam grandes (não importa se positivos ou negativos), as

	have a lesser impact on the output, as the neuron is respectively largely inhibited or largely excited.	mudanças em seus valores têm um impacto menor na produção, pois o neurônio é, respectivamente, amplamente inibido ou amplamente ativado.
61	In the example, neuron □4 is such that its activation is bound between -1 and +1.	No exemplo, o neurônio S4 é tal que o seu grau de ativação está confinado entre -1 e +1.
62	Figure 2 represents how neuron □4 reacts to the stimulus in equation 1.	A Figura 2 representa o modo como o neurônio S4 reage ao estímulo da equação 1.
63	The reaction is represented with a function □(…), called the activation function, which is applied to the stimulus; the result is the activation of □4:	A reação é representada por uma função F(...), denominada função de ativação, que é aplicada ao estímulo; o resultado é o grau de ativação de S4:
64	(2)	(2)
65	Figure 2: How a neuron reacts to the total stimulus received	Figura 2: como um neurônio reage ao estímulo total recebido
66	As can be seen, for values around 0 in the horizontal axis the reaction is proportional to the stimulus, but for large positive or negative stimuli, when the neuron is very inhibited or very excited, the reaction is much smaller.	Como pode ser visto, para valores em torno de 0 no eixo horizontal, a reação é proporcional ao estímulo, mas, para grandes estímulos positivos ou negativos, quando o neurônio está muito inibido ou muito excitado, a reação é bem menor.
67	For this kind of neuron, the actual extreme values of -1 and +1 are never reached, no matter how strong the total stimulus is.	Para este tipo de neurônio, os valores extremos reais de -1 e +1 nunca são alcançados, não importa o quanto o estímulo total seja forte.
68	As said above, neuron □4 in our example is a specific type of neuron with an activation that varies between -1 and +1.	Como dito acima, o neurônio S4 , no exemplo em questão, é um tipo específico de neurônio com um grau de ativação que varia entre -1 e +1.
69	There are other kinds of activation functions with different ranges, but exploring them is out of the scope of this chapter.	Existem outros tipos de funções de ativação com diferentes intervalos, mas explorá-las está fora do âmbito de aplicação deste capítulo.
70	3.2 From neurons to networks	3.2 De neurônios até redes.
71	Neurons like the one discussed in the previous section may be connected to form an artificial neural network that performs a specific computational task, to solve a specific problem.	Neurônios como os discutidos na seção anterior podem ser conectados para formar uma rede neural artificial que executa uma tarefa computacional específica para resolver um problema específico.
72	In a network, some neurons receive external stimuli which act as inputs to the network (much as our eyes are connected to our brain and feed it with images) and represent an instance of the problem to be solved; some neurons, known as hidden neurons, receive stimuli only from other neurons; and finally, some neurons, known as output neurons represent the solution to the problem (a bit like the signals sent to the muscles of one of your hands to move it in a specific way).	Em uma rede, alguns neurônios recebem estímulos externos que atuam como entradas para a rede (assim como os olhos estão ligados ao cérebro humano cérebro e o alimentam com imagens) e representam uma instância do problema a ser resolvido; alguns neurônios, conhecidos como neurônios ocultos, recebem estímulos apenas de outros neurônios; e, finalmente, alguns neurônios, conhecidos como neurônios de saída, representam a solução para o problema (de maneira similar aos sinais enviados aos músculos de uma das mãos para movê-la de uma forma específica).
73	Figure 3 shows an example of such a neural network with five neurons; the network takes three inputs, which are fed to three hidden neurons, which in turn stimulate two output neurons.	A Figura 3 mostra um exemplo de uma rede neural com cinco neurônios; a rede recebe três entradas, que são alimentadas a três neurônios ocultos, que, por sua vez, estimulam dois neurônios de saída.
74	Figure 3: An artificial neural network with three three hidden neurons and two output neurons.	Figura 3: uma rede neural artificial com três neurônios ocultos e dois neurônios de saída.
75	Each connection has a weight not shown in the diagram.	Cada conexão tem um peso não indicado no diagrama.
76	The three input neurons on the left are represented by smaller circles to emphasise the idea that they directly emit the values of	Os três neurônios de entrada à esquerda são representados por círculos menores para enfatizar a ideia de que eles emitem diretamente os valores da

	the external input, but, unlike regular neurons, they do not compute a stimulus or react to it via an activation function.	entrada externa, mas, ao contrário dos neurônios regulares, eles não calculam um estímulo ou reagem a ele por meio de uma função de ativação.
77	When building a neural network to solve a specific problem, one first needs to determine its architecture: how many neurons it has, how they are connected, which neurons receive external inputs and which neurons are designated as output neurons; but the actual computation performed depends on the weights of all of the connections in the network.	Ao construir uma rede neural para resolver um problema específico, primeiro é necessário determinar sua arquitetura: quantos neurônios ela têm, como eles estão conectados, quais recebem entradas externas e quais são designados como neurônios de saída; mas o cálculo efetivamente realizado depende dos pesos de ponderação de todas as ligações na rede.
78	How these weights are arrived at is explained in Section 3.5.	A forma como estes pesos são calculados é explicada na Seção 3.5.
79	Suffice it to say here, that one nice feature of artificial neural networks is that they may be trained to perform a task from examples, that is, their weights may be set to specific values by observing a set of solved examples, each one made up of the values of input signals representing the problems, and the values of the desired output activations representing the solutions.	Por ora, basta dizer que uma característica interessante das redes neurais artificiais é que elas podem ser treinadas para executar uma tarefa a partir de exemplos. Ou seja, os seus pesos podem ser definidos para valores específicos, observando um conjunto de exemplos resolvidos, cada um composto pelos valores dos sinais de entrada que representam os problemas e os valores dos graus de ativação de saída desejados, que representam as soluções.
80	3.3 Layers of neurons	3.3 Camadas de neurônios
81	Imagine that you are an absolute beginner and want to learn some basic techniques to paint landscapes in oils.	Supunha um completo iniciante que queira aprender algumas técnicas básicas para pintar paisagens em óleo sobre tela.
82	A manual might teach you a step-by-step over-simplified method with, for example, these four stages: drawing (a rough composition is sketched in), colour distribution, drawing refinement, and finish (when the final touches are made).	Um manual pode ensinar um método passo-a-passo simplificado com, por exemplo, estas quatro etapas: desenho (uma composição aproximada é esboçada), distribuição de cores, refinamento do desenho e acabamento (quando os retoques finais são feitos).
83	The point here is not the number of stages or the particular characteristics of each of them, but the fact that the whole process flows in an incremental manner in such a way that the output of one step becomes the input to the next one.	O ponto aqui não é o número de passos ou as características particulares de cada uma deles, mas o fato de que todo o processo flui de forma incremental de tal forma que a saída de uma etapa se torna a entrada para a próxima.
84	Each step refines the previous outcome: the outcome of the second step (colour distribution) is more of an actual landscape painting than the outcome of the first one (drawing) and, similarly, the outcome of the fourth stage (finish) can be conceptually considered as a better painting than those resulting from any of the previous steps.	Cada etapa refina o resultado anterior: o resultado da segunda etapa (distribuição de cores) está mais próxima de uma pintura de paisagem real do que o resultado da primeira (desenho) e, da mesma forma, o resultado da quarta etapa (acabamento) pode ser conceitualmente considerado como uma pintura melhor do que as resultantes de qualquer uma das etapas anteriores.
85	It turns out that neural computation benefits from a similar step-by-step incremental process.	Ocorre que a computação neural se beneficia de um processo incremental passo-a-passo semelhante.
86	Back in the sixties, researchers discovered that by including multiple layers of neurons more complex tasks could be tackled.	Nos anos sessenta, pesquisadores descobriram que, ao incluir várias camadas de neurônios, tarefas mais complexas poderiam ser abordadas.
87	Each layer in a multilayer neural network refines the output of the previous layer and takes a bigger or smaller step towards the ultimate solution.	Cada camada em uma rede neural multicamada refina a saída da camada anterior e dá um passo maior ou menor em direção à solução final.

88	The resulting architecture would be similar to that in Figure 3 but with a number of additional hidden layers.	A arquitetura resultante seria semelhante à da Figura 3, mas com várias outras camadas ocultas.
89	One can clearly see this layered structure in the simple network in Figure 3: computation, performed by two layers, takes place in two steps.	Pode-se ver claramente esta estrutura em camadas na rede simples na Figura 3: a computação, realizada por duas camadas, ocorre em duas etapas.
90	A model made of neurons organised in layers is referred to as a layered neural network.	Um modelo feito de neurônios organizados em camadas recebe o nome de rede neural em camadas.
91	In spite of theoretical results proving that a two-layer network has enough computational power to perform virtually any task (Hornik 1991), in the real world, the computational power of neural networks appears to be correlated with the number of layers; models with more than a few layers are often labelled as deep neural networks and the corresponding training algorithms are known as deep-learning algorithms.	Apesar de os resultados teóricos provarem que uma rede de duas camadas tem poder computacional suficiente para realizar praticamente qualquer tarefa (Hornik 1991), no mundo real, o poder computacional das redes neurais parece estar correlacionado com o número de camadas; modelos com mais do que algumas poucas camadas são frequentemente rotulados como redes neurais profundas e os algoritmos de treinamento correspondentes são conhecidos como algoritmos de aprendizagem profunda.
92	“OpenAI’s GPT-3 language model: A technical overview” (2020).	"Modelo de linguagem GPT-3 da OpenAI: uma revisão técnica" (2020).
93	Retrieved from https:// lambdalabs.com/blog/demystifying-gpt-3.	Obtido em https:// lambdalabs.com/blog/demystifying-gpt-3.
94	As an example of the complexity that these deep models may reach, GPT- 3 (Brown et al. 2020), one of the largest neural networks released in 2020 in the field of natural language generation, has 96 layers with tens of thousands of neurons each, which results in around 175,000 million weights to be learned by the training algorithm.	Como exemplo da complexidade que esses modelos profundos podem atingir, o modelo GPT-3 (Brown et al. 2020), uma das maiores redes neurais lançadas em 2020 no campo da geração de língua natural, possui 96 camadas com dezenas de milhares de neurônios cada, o que resulta em cerca de 175 bilhões de pesos a serem aprendidos pelo algoritmo de treinamento.
95	Supercomputers were used to train the GPT-3 system, a process that can take several weeks or even months, but it has been estimated that learning the weights for such a model with a single powerful gaming desktop personal computer would have taken more than 350 years1.	Para treinar o sistema GPT-3, foram utilizados supercomputadores, em um processo que pode demorar várias semanas ou mesmo meses, mas estima-se que a aprendizagem dos pesos para um modelo desse tipo com um único computador pessoal para jogos poderoso teria levado mais de 350 anos.
96	3.4 Neural machine translation	3.4 Tradução automática neural
97	If we manage to represent a source sentence as a set of inputs to a neural network, and we can interpret the neural network’s outputs as a target sentence, we have a neural machine translation (NMT) system.	Se uma sentença de origem for representada como um conjunto de entradas para uma rede neural, e se for possível interpretar as saídas da rede neural como uma sentença de destino, um sistema de tradução automática neural (NMT) é estabelecido.
98	NMT first processes the words in the source sentence.	A NMT primeiro processa as palavras na sentençade origem.
99	Each time a source word is ingested by the encoder part of the neural network, the activations of sets of specific neurons in the network change.	Cada vez que uma palavra-fonte é ingerida pela parte codificadora da rede neural, os graus de ativação de conjuntos específicos de neurônios na rede mudam.
100	When the whole source sentence has been processed, the decoder part of the network starts its work.	Quando toda a sentença de origem tiver sido processada, a parte decodificadora da rede inicia o seu trabalho.
101	It has been trained to provide, step by step, a probability score for each possible target word in the translation, given the target words it has already output.	Ela foi treinada para fornecer, passo a passo, um valor de probabilidade para cada palavra-alvo possível na tradução, dadas as palavras-alvo que já produziu.

102	This is similar to how predictive keyboards in contemporary smartphones work, but, as we will see, word predictions in NMT also depend on the source sentence, as they are meant to be a translation of it.	Esse mecanismo é semelhante à forma como os teclados preditivos em smartphones contemporâneos funcionam, mas, como será elucidado, as previsões de palavras de uma NMT também dependem da sentença de origem, pois devem ser uma tradução dela.
103	NMT systems are deep neural networks with architectures that will be discussed later in section 6.	Os sistemas NMT são redes neurais profundas com arquiteturas que serão discutidas mais adiante, na seção 6.
104	They have thousands of neurons and millions of weights (or many more) which have to be trained by providing examples taken from a parallel corpus containing millions of source sentences and their translations.	Eles têm milhares de neurônios e milhões de pesos (ou muitos mais) que precisam ser treinados fornecendo exemplos retirados de um corpus paralelo contendo milhões de sentenças-fonte e suas traduções.
105	Mathematical representations of the words in a given sentence in the source language are fed as inputs to the neural network and the words in the corresponding target-language sentence are used to represent the desired output.	Representações matemáticas das palavras de uma determinada sentença na língua de origem são alimentadas à rede neural como entradas e as palavras da sentença correspondente na língua de destino são usadas para representar a saída desejada.
106	As you might expect, training a large network in reasonable time is computationally demanding: one needs very powerful, specialised number-crunching hardware to train the network by showing it the examples over and over again.	Como se pode imaginar, treinar uma grande rede em tempo razoável é computacionalmente intensivo: é necessário um sistema de processamento muito poderoso e especializado para treinar a rede, mostrando os exemplos repetidamente.
107	On each iteration, small changes are made to the weights in the network to improve its prediction of target words.	A cada iteração, pequenas alterações são feitas nos pesos na rede para melhorar sua previsão das palavras-alvo.
108	3.5 Training neural networks	3.5 Treinando redes neurais
109	Training a neural network is the process of determining the weight of the connections between its neurons so that, given a training set of input–output examples, it produces an actual output which is as close as possible to that in the relevant example.	Treinar uma rede neural é o processo de determinar o peso das conexões entre seus neurônios de modo que, dado um conjunto de treinamento de exemplos de entrada e suas saídas respectivas, produz uma saída real o mais próxima possível da do exemplo em estudo.
110	Some of you may recognise here the mathematical concept of derivative of a function.	Alguns dos leitores poderão reconhecer aqui o conceito matemático de derivada de uma função.
111	Training starts with a set of random weights or with weights taken from a neural network solving a similar task.	O treinamento começa com um conjunto de pesos aleatórios ou com pesos retirados de uma rede neural resolvendo uma tarefa semelhante.
112	During training weights are modified in such a way that the value of an error function (also known as a loss function), which measures how much actual outputs deviate from the desired outputs, is made as small as possible.	Durante o treinamento, os pesos são modificados de tal forma que o valor de uma função de erro (também conhecida como função de perda), que mede o quanto as saídas reais se desviam das saídas desejadas, seja o menor possível.
113	Training algorithms (also called learning algorithms) repeatedly compute small corrections (updates) to weights until the error function is minimal or small enough for all examples in the training set, or a certain performance is observed in a different development set, which has been reserved or ”held out” for this purpose (see Section 7.2).	Algoritmos de treinamento (também chamados de algoritmos de aprendizagem) realizam repetidamente pequenas correções (atualizações) nos pesos até que a função de erro seja mínima (ou pequena o suficiente) para todos os exemplos no conjunto de treinamento, ou até que um determinado desempenho seja observado em um conjunto de desenvolvimento, que é reservado para esse propósito (ver seção 7.2).
114	The technical details of the training algorithm are beyond the scope of this chapter; let us just say that it is usually based on computing how much the error function varies when	Os detalhes técnicos do algoritmo de treinamento excedem o escopo deste capítulo; mencione-se apenas que ele se fundamenta, em geral, no cálculo do quanto a função de erro varia quando cada peso

	each weight is varied by a fixed but very small amount (the gradient of the error function), and then varying each weight a bit in the direction in which it reduces the error function.2 This type of training is called gradient descent; it is not guaranteed to find the very best weights, but it is likely that good candidates will be found.	é variado por uma quantidade fixa, mas muito pequena (o gradiente da função de erro) e, em seguida, cada peso é ligeiramente ajustado na direção que reduz a função de erro.2 Este tipo de treinamento chama-se gradiente descendente; não é garantia de que ele vai encontrar os melhores pesos, mas é provável que encontre bons candidatos.
115	The intensity of these weight variations is regulated by a parameter called the learning rate; this learning rate is usually higher in the first steps of the training algorithm, but its magnitude is made progressively smaller as the weights get closer to their final values.	A intensidade destas variações de peso é regulada por um parâmetro denominado taxa de aprendizagem; esta taxa de aprendizagem é geralmente maior nas primeiras etapas do algoritmo de treinamento, mas sua magnitude diminui progressivamente à medida que os pesos se aproximam de seus valores finais.
116	Note that training neural networks is quite laborious: many examples are necessary and they need to be presented many times to learn.	Observe que o treinamento de redes neurais é bastante trabalhoso: muitos exemplos são necessários e precisam ser apresentados muitas vezes para que elas aprendam.
117	This is often due to limitations of the training algorithms, however, rather than to the lack of capacity of a specific neural network to represent the solution to a problem.	No entanto, isto se deve frequentemente às limitações dos algoritmos de treinamento, e não à falta de capacidade de uma rede neural específica para representar a solução de um problema.
118	Once the weights are determined, training stops (see Section 7.2) and the neural network can be used to obtain the outputs for new inputs which are not included among the examples used during training.	Uma vez determinados os pesos, o treinamento para (ver seção 7.2) e a rede neural pode ser utilizada para obter as saídas para novas entradas que não estão incluídas entre os exemplos utilizados durante o treino.
119	3.6 Generalisation in neural networks	3.6 Generalização em redes neurais
120	Generalisation is a fundamental cognitive process for humans and animals.	Generalização é um processo cognitivo fundamental para humanos e animais.
121	It allows us to use what we learned in the past in new situations which can be regarded as similar but not identical to the situation in which learning originally took place.	Permite-se, por meio dele, a utilização do que foi aprendido no passado em novas situações que podem ser consideradas semelhantes, mas não idênticas à situação em que a aprendizagem ocorreu originalmente.
122	A person does not need to relearn how to drive when entering a new street or driving a new car.	Uma pessoa não precisa reaprender a dirigir ao entrar em uma nova rua ou dirigir um carro novo.
123	Similarly, generalisation happens when an organism which already responds to a certain stimulus in a particular way responds to similar stimuli in similar ways.	Analogamente, a generalização acontece quando um organismo que já reage a um determinado estímulo de uma maneira particular, responde de formas semelhantes a estímulos parecidos.
124	Generalisation is also key to language learning: young children soon learn to say sentences they have never heard before.	Aplicar generalizações também é fundamental para a aprendizagem de línguas: as crianças aprendem rapidamente a pronunciar sentenças que nunca ouviram antes.
125	Neural networks may ideally generalise in the context of machine translation by producing similar outputs when fed with similar inputs, independently of whether they were included in the training set or not.	As redes neurais podem, idealmente, aplicar generalizações no contexto da tradução automática, produzindo resultados semelhantes quando alimentadas com entradas semelhantes, independentemente de terem sido ou não incluídas no conjunto de treinamento.
126	One feature of neural networks is the smoothness of the computations, meaning that if the input values are slightly changed, the result of the formulas will not vary significantly.	Uma característica das redes neurais é a suavidade dos cálculos, o que significa que, se os valores de entrada forem ligeiramente alterados, o resultado das fórmulas não irá variar significativamente.
127	In a broad sense, in order to achieve generalisation, similar sentences should get	Em sentido amplo, para aplicar uma generalização, sentenças semelhantes devem obter representações

	similar representations, and as sentence representations will be obtained from word representations, we may conclude that representing similar words with similar numbers is a precondition for generalisation in neural language processing.	parecidas e, como as representações de sentenças serão obtidas a partir de representações de palavras, conclui-se que representar palavras similares e com números semelhantes é uma pré-condição para a generalização no processamento neural de língua.
128	The next section will delve into how we can end up with a convenient list of neural representations for the words in a sentence that benefits from the smoothness of neural networks so that, after training, the system is able to generalise properly to sentences it has not seen before.	Na próxima seção, será aprofundada a maneira com que se obtém uma lista de representações neurais para palavras de uma sentença que se beneficiaria da fluidez das redes neurais de forma que, após o treinamento, o sistema seria capaz de fazer generalizações adequadas de sentenças que nunca viu antes.
129	4 Word embeddings as vector representation of words	4 Vetores como representações de palavras
130	In the previous section we noted that neurons are usually arranged in layers in such a way that the output of the neurons of one layer becomes the input to the neurons of the following one.	Na seção anterior, foi observado que os neurônios geralmente são organizados em camadas de maneira que a saída dos neurônios de uma camada se transforma na entrada para os neurônios da camada seguinte.
131	Interestingly, the output of the set of neurons in a given layer constitutes a representation of the information they are processing at that stage.	Curiosamente, o produto do conjunto de neurônios em uma determinada camada consiste numa representação das informações que eles processam naquele estágio.
132	In the field of natural language processing, and as indicated above, the information processed by neural networks is made up of words, and their representations within the network are usually referred to as embeddings (Mikolov et al. 2013).	No campo do processamento de língua natural, e conforme já foi mostrado, as informações processadas pelas redes neurais são compostas de palavras, e suas representações na rede costumam ser chamadas de vetores (embeddings) (Mikolov et al. 2013).
133	What makes these embeddings really useful is that those words with similar meanings or that usually co-occur in the same contexts end up having similar embeddings.	O que torna essas vetorizações realmente úteis é o fato de que as palavras com significados semelhantes ou que comumente ocorrem nos mesmos contextos acabam por ter vetorizações semelhantes.
134	In order to better understand this, take a piece of paper and draw a square with sides of about 10 centimetres.	Para uma melhor compreensão, use um pedaço de papel para desenhar um quadrado com lados de aproximadamente dez centímetros.
135	Now, take the words in the following list and put them all on the square by following a criterion that places words which are closer in meaning nearer each other than words with less related meanings.	A seguir, reúna as palavras e organize-as dentro do quadrado, observando o critério de que palavras com significados mais similares estejam posicionadas mais próximas entre si em comparação às palavras com significados mais distintos.
136	If this concept of meaning closeness seems imprecise to you, you may place the words based on their frequency of co-occurrence in sentences or paragraphs.	Se o conceito de proximidade de sentido se mostrar inexato, é possível organizar as palavras com base na frequência de ocorrência conjunta em sentenças ou parágrafos.
137	The words are: restaurant, red, garden, fountain, flower, tomato, balloon, waiters, knife, flowers, menu, cooked, chromosome and consistently.	As palavras são: restaurante, vermelho, jardim, fonte, flor, tomate, balão, garçons, faca, flores, cardápio, cozido, cromossomo e consistentemente.
138	Do this before reading on.	Faça isso antes de dar continuidade à leitura.
139	The restriction imposed by means of the criterion of word meaning proximity implies that you have not been able to freely distribute the words on the square.	A restrição imposta por meio do critério de proximidade semântica significa que não existe a possibilidade de distribuir livremente as palavras no quadrado.
140	Probably, you have decided to group words such as restaurant, menu and waiters, on the	Um agrupamento mais intuitivo poderia ser composto de palavras como restaurante, cardápio e

	one hand, and words such as garden, flower and fountain, on the other hand.	garçons, de um lado, e palavras como jardim, flor e fonte, de outro.
141	There are, however, some doubtful cases: red is clearly a neighbour of tomato, but it should be close to flower as well; a compromise solution would be to put it somewhere in between, a little bit closer to tomato than to flower if we acknowledge that red is not as essential to flowers as it is to tomatoes.	Há, no entanto, alguns casos questionáveis: vermelho é claramente próximo de tomate, mas também deve ficar próximo a flor; uma solução conciliatória poderia ser colocá-lo em um ponto intermediário, um pouco mais perto de tomate do que de flor, levando em consideração que a cor vermelha não é tão essencial às flores quanto é aos tomates.
142	We have deliberately placed Figure 4 a few pages on, so that you do not see it before you attempt the exercise.	Propositalmente, colocamos a Figura 4 algumas páginas mais adiante, para que o leitor não a veja antes de tentar fazer o exercício.
143	You may have noticed some clusters in your design: an island representing the semantic field of restaurants and related things, and another island around the idea of gardens and orchards.	Alguns subgrupos são perceptíveis neste agrupamento: uma ilha representando o campo semântico de restaurantes e coisas relacionadas, e outra ilha relacionada a jardins e pomares.
144	There are some outliers on the list, especially the word consistently, which seems in principle disconnected from the rest of words, forcing us to put it as far as possible from all of them.	Há algumas exceções na lista, especialmente a palavra consistentemente, que parece, por princípio, desvinculada das outras palavras, impondo a necessidade de colocá-la o mais distante possível de todas as outras.
145	Chromosome is another isolated word, but as flowers and waiters use chromosomes to carry their genetic information, it may be put somewhere in the middle of the line between these words but at the same time not very close to red.	Cromossomo é outra palavra isolada, mas, como as flores e os garçons, têm cromossomos para transportar as próprias informações genéticas, a palavra pode ser colocada em algum lugar em entre essas palavras, porém, ao mesmo tempo, um pouco longe da palavra vermelho.
146	See Figure 4 for a possible solution that may not match yours exactly.	Consulte a Figura 4 e observe uma solução possível e que talvez não corresponda exatamente à solução típica elaborada pelo leitor.
147	In order to assign mathematical codes to the words in our list, let's assign coordinates to each word to reflect its position on the square.	Com o propósito de atribuir códigos matemáticos às palavras da lista em questão, serão designadas coordenadas para cada palavra, a fim de refletir suas respectivas posições no quadrado
148	As we are in a two-dimensional space, we need two coordinates for each word: the first coordinate is a number that represents the distance to the left vertical side of the square; the second coordinate is a number that represents the distance to the bottom horizontal side of the square.	Considerando que o espaço em questão é bidimensional, duas coordenadas são necessárias para cada palavra: a primeira coordenada é um número que representa a distância horizontal até o canto inferior esquerdo do quadrado; a segunda coordenada é um número que representa a distância vertical até o mesmo ponto.
149	The word restaurant could be assigned, for example, the two numbers 0.25 and 1.1, and the word menu the numbers 0.6 and 1.3, close to restaurant as seen in Figure 4.	A palavra restaurante poderia ser atribuída, por exemplo, aos dois números 0,25 e 1,1, e a palavra menu aos números 0,6 e 1,3, próximo a restaurante, como visto na Figura 4.
150	These coordinate values can be represented using vector notation, which simply consists of writing the numbers as a comma-separated list of values between brackets.	Essas coordenadas podem ser representadas usando a notação vetorial, que consiste simplesmente em escrever os números como uma lista de valores separados por vírgulas entre colchetes.
151	The vectors corresponding to restaurant and menu would therefore be [0.25, 1.1] and [0.6, 1.3], respectively.	Os vetores correspondentes a restaurante e cardápio seriam, portanto, [0,25, 1,1] e [0,6, 1,3], respectivamente.
152	Each of these vectors represents a possible word embedding for these two words.	Cada um desses vetores representa uma possível vetorização semântica para essas duas palavras.
153	Although it may not be completely obvious, considering embeddings made up of two numbers instead of a single number boosts the possibilities of solving the problem of placing	Embora não seja completamente óbvio, avaliar vetorizações compostas por dois números em vez de um único número aumenta as possibilidades de uma solução para o problema de posicionar

	words closer or farther apart as we have more freedom to satisfy all the restrictions.	palavras mais perto ou mais longe, pois mais liberdade é proporcionada para satisfazer todas as restrições.
154	In fact, moving from two dimensions to a higher number of dimensions increases these possibilities even more.	De fato, a mudança de duas dimensões para um número maior de dimensões aumenta ainda mais essas possibilidades.
155	A five-dimensional representation of a word could be, for example, [2.34, 1.67, 4.81, 3.01, 5.61].	Uma representação de cinco dimensões de uma palavra poderia ser, por exemplo, [2,34, 1,67, 4,81, 3,01, 5,61].
156	NMT systems consider embeddings with hundreds of dimensions, and the input sentence to be translated is represented by a collection of these vast word embeddings.	Os sistemas de NMT analisam as vetorizações com centenas de dimensões, e a sentença de entrada a ser traduzida é representada por uma coletânea de vastas vetorizações semânticas.
157	Word embeddings are learned using the very same algorithm used to learn the weights of the neural network presented in Section 3.5.	A vetorização semântica é aprendida com o mesmo algoritmo usado para aprender os pesos da rede neural apresentada na Seção 3.5
158	In fact, both the weights and the embeddings are learned at the same time.	Na verdade, tanto os pesos como as vetorizações são aprendidos ao mesmo tempo.
159	Bearing in mind that the input layer of a neural network involved in NMT usually consists of the embeddings of the words in the input sentence, there is no need to limit ourselves to fixed vectors.	Ao levar em conta que a camada de entrada de uma rede neural utilizada na NMT é geralmente composta pelas vetorizações das palavras na sentença de entrada, não há necessidade de limitar-se a vetores fixos.
160	Instead, their values can be repeatedly updated during training in such a way that the value of the error function is minimised.	Em vez disso, seus valores podem ser atualizados repetidamente durante o treinamento, de modo que o valor da função de erro seja minimizado.
161	4.1 Generalisation	4.1 Generalização
162	As already discussed, for the network to be able to properly generalise, that is, to be able to learn to translate and be capable of translating sentences never seen before, similar sentences should get similar representations.	Como já discutido, para que a rede possa fazer generalizações adequadamente, ou seja, para aprender a traduzir e ser capaz de traduzir sentenças nunca vistas antes, sentenças semelhantes devem receber representações semelhantes.
163	As sentence representations are obtained from word embeddings, we may conclude that representing similar words with similar numbers is a precondition for generalisation in neural natural language processing.	Como as representações de sentenças são obtidas a partir de vetorizações de palavras, pode-se concluir que representar palavras semelhantes com números semelhantes é uma condição prévia para a generalização no processamento neural de língua natural.
164	Following our example, words such as poured, rained, pouring or raining should ideally share similar embeddings as all of them are semantically similar; the codes for pouring and raining should also be closer to words such as driving since the three of them are gerunds and may appear in similar contexts; poured and rained should be neighbours as well because both of them are past tenses.	No exemplo proposto, palavras como alagou, choveu, alagando ou chovendo devem, idealmente, têm vetorizações semelhantes, uma vez que todas são semanticamente semelhantes; os códigos para alagando e chovendo também devem estar mais perto de palavras como dirigindo, uma vez que os três são gerúndios e podem aparecer em contextos semelhantes; alagou e choveu devem ficar próximos também porque ambos são verbos no pretérito.
165	This is why we usually need many dimensions: we want words to be close to each other in different ways or for different reasons, simultaneously.	É por essa razão que a necessidade de muitas dimensões é frequentemente observada: busca-se que as palavras estejam, simultaneamente, próximas umas das outras de maneiras diferentes ou por motivos diferentes.
166	4.2 Geometric properties of word embeddings	4.2 Propriedades geométricas de vetorização semântica.
167	Word embeddings exhibit interesting properties that demonstrate that they represent	As vetorizações têm propriedades interessantes que representam características semânticas (ou relacionadas à semântica) de palavras.

	semantic characteristics (or something related to semantics) of words.	
168	As already explained, the embedding of a word consists of several real numbers, usually hundreds or thousands of them, and each of these numbers seems to capture a certain aspect of the meaning of a word.	Como já explicado, a vetorização de uma palavra consiste em vários números reais, geralmente centenas ou milhares deles, e cada um desses valores captura um determinado aspecto do significado de uma palavra.
169	For example, the word embedding for Dublin should capture several semantic-related aspects of it: a city, the capital of Ireland, the place for the headquarters in Europe of several multinational companies, etc.	Por exemplo, a vetorização semântica de palavras para Dublin deve reunir vários aspectos relacionados a ela: uma cidade, a capital da Irlanda, o local da sede na Europa de várias empresas multinacionais etc.
170	Thanks to this specialisation of the different dimensions of the embeddings, we can perform some arithmetic operations with the embeddings and obtain meaningful results.	Graças a essa particularização das diferentes dimensões dos vetores semânticos, é possível executar algumas operações aritméticas com as incorporações e obter resultados expressivos.
171	These operations are simply additions and subtractions that are straightforward to compute.	Essas operações são simplesmente adições e subtrações que são fáceis de calcular.
172	Adding (or subtracting) two embeddings simply consists of adding (or subtracting) the components of the vectors one by one; for example, [1.24, 2.56, 5.23]+[0.12, 1.12, 0.01] = [1.36, 3.68, 5.24].	A adição (ou subtração) de dois vetores consiste simplesmente em adicionar (ou subtrair) seus componentes um a um; por exemplo, [1.24, 2.56, 5.23] + [0.12, 1.12, 0.01] = [1.36, 3.68, 5.24].
173	Below are two examples of arithmetic operations with meaningful results performed on embeddings that NMT systems usually learn:	Abaixo estão dois exemplos de operações aritméticas com resultados significativos realizados em incorporações que os sistemas NMT normalmente aprendem:
174	where the square brackets refer to the embedding of a word, and with ≃ we mean that the resulting embedding after the operation is close to the embedding of the word on the right-hand side of the example.	onde os colchetes referem-se às vetorizações de uma palavra e, com o uso do símbolo ≃, indica-se que a vetorização resultante após a operação é aproximada da vetorização da palavra ao lado direito da "equação" do exemplo.
175	This can be interpreted as indicating that king is to man what queen is to woman, a male or female monarch; and Dublin is to Ireland what Paris is to France, the capital of a country.	Isso pode ser interpretado como uma indicação de que o rei é para o homem o que a rainha é para a mulher, um monarca masculino ou feminino; e Dublin é para a Irlanda o que Paris é para a França, a capital de um país.
176	Figure 4: Placement of words in a two-dimensional area in such a way that related words are positioned close to each other, but far from words they have less in common with.	Colocação de palavras em uma área bidimensional de forma que as palavras relacionadas sejam posicionadas próximas umas das outras, mas longe das palavras com as quais têm menos em comum
177	5 Contextual word embeddings through attention	5 Vetores contextuais por meio de atenção
178	Words do not always have the same meaning in every sentence.	As palavras nem sempre têm o mesmo significado em todas as sentenças.
179	The embedding of the word letter, for example, should not be the same when the word refers to a character of an alphabet or when it refers to a document addressed to another person.	A vetorização da palavra carta, por exemplo, não deve ser a mesma quando a palavra se refere a um elemento do baralho ou quando se refere a um documento destinado a outra pessoa.
180	In fact, it may even be interesting for an NMT system to represent the word with different embeddings depending on whether it refers to a love letter or a complaint letter.	Na verdade, pode até ser interessante para um sistema NMT representar a palavra com diferentes vetorizações, a depender de o termo se referir a uma carta de amor ou a uma carta de reclamação.
181	The embeddings we introduced before are non-contextual: they were computed by considering words that usually co-occur in	As vetorizações previamente apresentadas são não-contextuais: elas foram calculadas levando em consideração palavras que geralmente ocorrem

	sentences but without taking into consideration the different meanings words may have.	juntas em sentenças, mas sem levar em conta os diferentes significados que as palavras podem ter.
182	In the NMT arena, attention plays an important role as it allows the neural network to compute contextual word embeddings, that is, vector representations of the words in a sentence computed in such a way that the representation obtained for a word is adapted to its meaning in each particular sentence.	No contexto da NMT, a atenção desempenha um papel importante, pois permite que a rede neural compute vetores contextuais, ou seja, representações vetoriais das palavras em uma sentença computadas de tal forma que a representação obtida para uma palavra seja adaptada ao seu significado em cada sentença específica.
183	Attention is, once again, a concept which is implemented by means of mathematical operations conveniently learned by a training algorithm.	A atenção é, mais uma vez, um conceito que é implementado por meio de operações matemáticas aprendidas de forma conveniente por um algoritmo de treinamento.
184	In our context, attention is similar, to the situation in which we pay attention to something or someone in our everyday lives.	No contexto em discussão, a atenção assemelha-se à situação na qual a atenção é dispensada a algo ou a alguém na vida cotidiana.
185	By conveniently using attention to concentrate on some words in the sentence, the embedding vector corresponding to the word season, for example, will differ between the sentences in examples 1 and 2 below:	Ao usar a atenção de forma conveniente para se concentrar em algumas palavras da sentença, o vetor semântico correspondente à palavra temporada, por exemplo, será diferente nas sentenças dos exemplos 1 e 2 abaixo:
186	The first episode will pick up right where the previous season left off	O primeiro episódio vai continuar exatamente de onde a temporada anterior parou.
187	Summer is the hottest season of the whole year.	O verão é a temporada mais quente do ano.
188	In principle, it may sound as if the purpose of contextual word embeddings is that the different meanings of a word get different representations, but, while this will be usually true, the idea goes beyond this.	Em princípio, pode parecer que o propósito dos vetores contextuais de palavras fosse o de atribuir representações diferentes a significados diferentes de uma palavra, mas, embora isso geralmente seja verdade, a ideia vai além disso.
189	The contextual word embeddings for season in the sentences “Winter is the coldest season of the year in polar and temperate zones”, “Summer is the hottest season of the whole year” and even “Of the whole year, summer is the hottest season” will all be different, although presumably closer to each other than the representation of season in “The first episode will pick up right where the previous season left off”.	Os vetores contextuais da palavra para temporada nas sentenças "o inverno é a temporada mais fria do ano nas zonas polares e temperadas", "o verão é a temporada mais quente de todo o ano" e até mesmo "de todo o ano, a temporada de verão é a mais quente" serão todos diferentes, embora presumivelmente mais próximos uns dos outros do que a representação de temporada em "o primeiro episódio vai continuar exatamente de onde a temporada anterior parou".
190	These divergences result from the fact that the words in the sentences or the order in which they are placed differ.	Estas divergências resultam do fato de as palavras das sentenças ou a ordem em que são colocadas serem diferentes.
191	Remarkably, the two instances of the in each of our examples will get two different contextual vectors because the context of each instance is also different.	É possível notar que, em cada um dos exemplos propostos, as instâncias repetidas de um mesmo artigo apresentarão dois vetores contextuais distintos, uma vez que o contexto de cada instância varia.
192	How are contextual embeddings mathematically computed through attention?	Como os vetores são computados matematicamente através da atenção?
193	Given the sentence in example 2 above (”Summer is the hottest season of the whole year.”), the procedure starts by obtaining the non-contextual word embeddings that were introduced in Section 4.	Dada a sentença no exemplo 2 acima ("o verão é a temporada mais quente do ano."), o procedimento é iniciado pela obtenção dos vetores não-contextuais de palavras apresentados na seção 4.
194	As the sentence has nine words, the result is a collection of nine vectors which are the ingredients for the next step.	Dado que a sentença contém nove palavras, o resultado é uma coleção de nove vetores, que são os elementos fundamentais para a próxima etapa.

195	Now, in order to compute the contextual word embedding for the word season in the sentence, an attention vector is mathematically produced by the neural network.	Agora, para calcular o vetor contextual da palavra temporada na sentença, um vetor de atenção é produzido matematicamente pela rede neural.
196	This attention vector will have nine percentages representing the degree of attention that needs to be paid to each of the words in the sentence in order to obtain the representation of the word season.	Este vetor de atenção terá nove valores percentuais que representam o grau de atenção que deve ser prestado a cada uma das palavras da sentença para obter a representação da palavra temporada.
197	The element at a certain position in the vector corresponds to the attention to the word at that position in the sentence.	O elemento em uma determinada posição no vetor corresponde à atenção para a palavra nessa posição na sentença.
198	For example, an attention vector [25%, 8%, 10%, 15%, 25%, 8%, 2%, 0%, 7%] would indicate that in order to compute a contextual vector representation of the word season in the running sentence, the word embeddings for summer and season will be equally highly relevant (together, they receive fifty percent of the total attention), which makes sense as they are semantically connected to the concept of a meteorological season.	Por exemplo, um vetor de atenção [10%, 25%, 8%, 10%, 25%, 5%, 10%, 0%, 7%] indicaria que, para calcular uma representação vetorial contextual da palavra temporada na sentença sendo traduzida, as vetorizações para verão e temporada serão igualmente relevantes (em conjunto, recebem cinquenta por cento da atenção total), o que faz sentido, uma vez que estão semanticamente ligados ao conceito de estação meteorológica.
199	Notice that the preceding determiner gets some attention too (10%), which may be explained by the fact that it helps to label season as a noun.	Observe que o determinante anterior também recebe alguma atenção (10%), o que pode ser explicado pelo fato de ajudar a rotular temporada como um substantivo.
200	The contribution of the verb (8%) to the contextual embedding may also be described in terms of its contribution to marking the number of season as singular.	A contribuição do verbo (8%) para a incorporação contextual pode também ser descrita em termos da sua contribuição para a marcação do número de estação como singular.
201	Note that the percentages always add up to 100%.	Note que as percentagens somam sempre 100%.
202	Determining how the attention vector is used in order to obtain a new embedding that combines the original non-contextual embeddings to get a new embedding is beyond the scope of this chapter.	A determinação de como o vetor de atenção é utilizado para obter uma nova vetorização que combine os vetores não-contextuais originais para obter um novo vetor está fora do âmbito do presente capítulo.
203	Suffice to say that the procedure involves a specific sequence of mathematical operations and that the resulting embedding will be located somewhere in between the original embeddings.	Basta dizer que o procedimento envolve uma sequência específica de operações matemáticas e que o vetor resultante estará em algum lugar entre os vetores originais.
204	Following our running example, nine different attention vectors will be computed for this sentence (one for each word) and then applied to the original noncontextual embeddings in order to obtain a collection of nine new embeddings, each one corresponding to a different word in the sentence.	No exemplo em análise, nove vetores de atenção distintos serão calculados para a sentença em questão (um para cada palavra) e, em seguida, aplicados aos vetores não-contextuais originais, a fim de obter uma coleção de nove novos vetores, cada um correspondendo a uma palavra diferente na sentença.
205	These new embeddings may be considered as contextual embeddings as they are influenced to different degrees by the rest of the words in the sentence.	Esses novos vetores podem ser considerados como vetores contextuais, pois são influenciados em diferentes níveis pelo resto das palavras da sentença.
206	5.1 Many attention layers, better than one	5.1 Várias camadas de atenção são melhores que uma
207	“Turing-NLG: A 17-billion-parameter language model by Microsoft”, 2020.	"Turing-NLG: um modelo de linguagem de 17 bilhões de parâmetros da Microsoft", 2020.

208	Retrieved from https://www.microsoft.com/en-us/research/blog/turing-nlg-a-17-billion-parameterlanguage-model-by-microsoft/	Disponível em: https://www.microsoft.com/en-us/research/blog/turing-nlg-a-17-billion-parameterlanguage-model-by-microsoft/
209	Previously, in Section 3.3 of this chapter, we discussed the benefits of successively refining neural computations by exploiting models with different layers.	Anteriormente, na seção 3.3 deste capítulo, foram discutidos os benefícios de refinar sucessivamente os cálculos neurais por meio da exploração de modelos com diferentes camadas.
210	Consequently, it will come as no great surprise that in order to obtain more precise representations, the contextual embeddings just obtained may be combined with new attention vectors to obtain yet another new embedding for each word.	Consequentemente, não surpreende descobrir que, para obter representações mais precisas, os vetores contextuais recentemente obtidos possam ser combinados com novos vetores de atenção para obter mais um novo vetor para cada palavra.
211	As a real-life example, Turing Natural Language Generation (T-NLG), another of the largest language models published in 2020, has 78 attention layers that successively polish embeddings of 4,256 dimensions.	Como exemplo da vida real, o Turing Natural Language Generation (T-NLG), outro dos maiores modelos de linguagem publicados em 2020, tem 78 camadas de atenção que refinam sucessivamente vetores de 4.256 dimensões.
212	Recall that these representations, which are learned by applying many consecutive layers, are known as deep representations.	É importante lembrar de que essas representações, aprendidas pela aplicação de muitas camadas consecutivas, são conhecidas como representações profundas.
213	5.2 Many heads, better than one	5.2 Muitas cabeças pensam melhor que uma
214	There is no reason to restrict ourselves to a single attention vector for each word in each layer.	Não há motivos para nos restringirmos a um único vetor de atenção para cada palavra em cada camada.
215	For example, given the sentence “My grandpa baked bread in his oven daily”, it could be interesting to have an embedding for oven which has the flavour of grandpa to reflect that this oven belongs to an older person, and a different embedding for oven with the flavour of bread to reflect what has been cooked in it.	Por exemplo, dada a sentença "Meu avô assava pão em seu forno diariamente", pode ser interessante ter uma vetorização par forno que tenha o aspecto de avô para refletir que o forno em questão pertence a alguém mais velho, e uma vetorização diferente para forno com o aspecto de pão para refletir o que foi assado dentro dele.
216	A single attention vector would have to mix both flavours in a single embedding containing too much heterogeneous information that could affect negatively the search for a translation for the word represented by the embedding.	Um único vetor de atenção teria de misturar ambos os aspectos em uma única vetorização contendo informações demasiadamente heterogêneas que poderiam interferir negativamente na busca por uma tradução da palavra representada no vetor.
217	For this reason, some NMT systems obtain different attentions for each word in each layer and use them to compute a number of different embeddings for each word.	Por esse motivo, alguns sistemas de NMT obtêm diferentes vetores de atenção para cada palavra em cada camada e depois os utilizam para calcular vários vetores diferentes para cada palavra.
218	Each of these embeddings is said to be computed by a different head.	Cada um desses vetores foi calculado por uma cabeça diferente.
219	T-NLG has 28 attention heads in each layer.	T-NLG tem 28 cabeças de atenção em cada camada.
220	Therefore, its last layer produces 28 different 4,256-dimensional embeddings for each word.	Portanto, sua última camada produz 28 vetores de 4256 dimensões para cada palavra.
221	5.3 Contextual word embeddings in natural language processing	5.3 Vetores contextuais em processamento de língua natural
222	Embeddings are the cornerstone of NMT but they have also proved to be useful in many other natural language processing applications such as sentiment analysis and automatic summarisation.	Vetores são os pilares da NMT, mas eles também se mostraram úteis em muitas outras aplicações de processamento de língua natural, como análise de sentimento e resumo automático.

223	As an illustration, systems that automatically classify as positive or negative the sentences in a text containing a product review may work by first computing a collection of deep contextual embeddings for each word in the sentence and then feeding these embeddings to a much simpler neural network that will compute a number between 0 and 1 indicating the degree of positiveness of the sentence (for example, 0.95 will indicate a decidedly positive sentence, 0.2 a negative sentence, and 0.51 a neutral sentence).	Para ilustrar, sistemas que classificam automaticamente como positivas ou negativas as sentenças da avaliação de um produto podem calcular primeiro uma coleção de vetores contextuais longos para cada palavra da sentença e depois usá-los para alimentar uma rede neural muito mais simples, que vai calcular um número entre 0 e 1 indicando o grau de positividade da sentença (por exemplo, 0,95 indica uma sentença definitivamente positiva; 0,2, uma sentença negativa e 0,51, uma sentença neutra).
224	These systems are usually trained with a corpus of sentences manually tagged by humans.	Esses sistemas normalmente são treinados com um corpus de classificadas manualmente por seres humanos.
225	The part of the model that computes the embeddings is not necessarily trained for a particular corpus as pre-trained models already trained with millions of sentences are freely available for many languages.	A parte do modelo que calcula os vetores não necessariamente é treinada com um corpus em particular, dado que modelos pré-treinados com milhões de sentenças já existem de modo disponível gratuitamente em muitas línguas.
226	6 Neural machine translation, at last	6 Por fim, a tradução automática neural
227	At this point, you are hopefully in a good position to understand how NMT works, even if we describe its fundamentals in only a few sentences as we do next.	Neste ponto, o leitor deste artigo deve ter condições suficientes para entender como NMT funciona, mesmo que os seus fundamentos sejam descritos em poucas sentenças, como será feito a seguir.
228	We will focus on two architectures: those of so-called transformer and recurrent neural networks.	Vamos nos concentrar em duas arquiteturas específicas: os assim chamados transformers e as redes neurais recorrentes.
229	6.1 Transformer: Attention-based encoder–decoder	6.1 Transformer: um par codificador-decodificador baseado em atenção
230	Put simply, a transformer NMT system is composed of a module that computes contextual word embeddings for each word in the source input sentence and a second module which successively predicts each word in the target sentence.	Em termos simplificados, um sistema transformer de NMT é composto por um módulo que calcula vetores contextuais para cada palavra na sentença de entrada e um segundo módulo que prevê cada palavra na sentença-alvo.
231	The former module is called an encoder and the latter module is known as a decoder.	O primeiro módulo recebe o nome de codificador e o segundo, de decodificador.
232	For predicting the words in the target language, the decoder pays attention to the embeddings of all the words in the source sentence as well as to the embeddings of the target words already generated.	Para prever as palavras da sentença-alvo, o decodificador presta atenção à vetorização de todas as palavras da sentença-origem, bem como à vetorização das palavras já produzidas na sentença-alvo.
233	The whole architecture is called a transformer (Vaswani et al. 2017).	A arquitetura completa é chamada de transformer (Vaswani et al., 2017)
234	Figure 5 shows an example of a three-layered encoder and the degrees of attention considered in order to compute an embedding in the second layer and in the third one.	A Figura 5 mostra um exemplo de um codificador de três camadas e os níveis de atenção considerados ao calcular as vetorizações nas segunda e terceira camadas.
235	Figure 6 depicts this encoder in an extended diagram that also includes the decoder so that it represents the whole transformer architecture.	A Figura 6 mostra este codificador em um diagrama expandido que também inclui o decodificador, de modo a representar toda a arquitetura do transformer.
236	Figure 5: The encoder of a transformer-based neural machine translation system.	Figura 5: O codificador de um sistema de tradução automática neural de modelo transformer.
237	The symbol start is usually prefixed to explicitly mark the beginning of the sentence.	O símbolo de partida geralmente é prefixado de modo explícito para marcar o começo da sentença.
238	The diagram also shows that first-layer embeddings for brown and fox contribute to	O diagrama também mostra que a vetorização da primeira camada para as palavras "brown"

	different degrees to obtain the embedding for fox in the second layer; similarly, the embedding for brown in the last layer integrates information from all the embeddings in the second layer using different degrees of attention.	(marrom) e "fox" (raposa) dão contribuições de níveis diferentes para obter a vetorização de "fox" na segunda camada; analogamente, a vetorização de "brown" na última camada integra informações de todas as vetorizações da segunda camada usando diferentes níveis de atenção.
239	A parallel corpus is used by the learning algorithm to obtain a set of weights, embeddings and attention vectors for the transformer such that the training data can be reproduced up to a certain degree and the system is able to generalise beyond the sentences in the training set.	Um corpus paralelo é usado pelo algoritmo de aprendizagem para obter um conjunto de pesos, vetores lexicais e de atenção para o transformer, de tal forma que os dados de treinamento possam ser reproduzidos até um certo grau de acurácia e o sistema seja capaz de fazer generalizações para além das sentenças do conjunto de treinamento.
240	Figure 6: A complete transformer-based neural machine translation system translating a sentence.	Figura 6: Um sistema completo de tradução automática neural, baseado no modelo transformer, em processo de tradução de uma sentença.
241	An enlarged version of the encoder can be seen in Figure 5.	Uma versão aumentada do codificador pode ser vista na Figura 5.
242	Note how the prediction of zorro is obtained by paying attention to the embeddings of the previous target words but also to the embeddings corresponding to some of the input words coming from the last layer of the encoder.	Observe como a previsão "zorro" é obtida por meio da atenção prestada à vetorização das palavras-alvo anteriores, mas também à vetorização correspondendo a algumas das palavras de entrada que vêm da última camada do codificador.
243	For example, assume that a transformer with one single head per layer is used to translate the sentence “My grandpa baked bread in his oven daily” into Spanish.	Por exemplo, suponha que um transformer com uma única cabeça por camada seja utilizado para traduzir a sentença "Meu avô assava pão em seu forno diariamente" para o espanhol.
244	The encoder first produces a collection of eight embedding vectors.	O codificador primeiramente produz uma coleção de oito vetores lexicais.
245	The decoder then computes an 8-dimensional attention vector such as [60%, 10%, 0%, 0%, 0%, 30%, 0%, 0%] and uses it to obtain a flavour of the source sentence that allows it to obtain an embedding for the first word in the target sentence.	O decodificador então calcula um vetor de atenção de oito dimensões, como [60%, 10%, 0%, 0%, 0%, 30%, 0%, 0%] e o usa para obter um aspecto da sentença de origem que o permita obter uma vetorização da primeira palavra na sentença-alvo.
246	Let us assume that the system correctly generates the Spanish word mi.	Suponha que o sistema produza corretamente a palavra espanhola "mi".
247	The decoder will then compute a 9-dimensional attention vector such as [50%, 10%, 0%, 0%, 0%, 20%, 0%, 0%, 20%] (the last percentage corresponds to the attention paid to the first word in the target sentence) and use it to obtain an embedding for the second word in the target sentence.	O decodificador então irá calcular um vetor de nove dimensões como [50%, 10%, 0%, 0%, 0%, 20%, 0%, 0%, 20%] (o último valor percentual corresponde ao grau de atenção prestado à primeira palavra na sentença-alvo) e usá-lo para obter uma vetorização da segunda palavra na sentença-alvo.
248	The procedure will continue until the decoder generates a special token that marks the end of the sentence.	O procedimento vai continuar até o decodificador gerar um símbolo especial que marca o fim da sentença.
249	The output of the decoder at each step is not exactly an estimation of the embedding of the next word.	A saída do decodificador a cada iteração não é exatamente uma estimativa da vetorização da próxima palavra.
250	Actually, an additional layer is added at the end of the decoder to compute a vector of probabilities or likelihoods for each word in the target-language vocabulary.	Na verdade, uma camada adicional é sobreposta ao fim do decodificador para calcular um vetor de probabilidades ou verossimilhanças para cada palavra no vocabulário da língua de chegada.
251	Section 7.3 will discuss how these probabilities can be used in order to obtain the sequence of words that result in the target-language sentence.	A seção 7.3 irá discutir como essas probabilidades podem ser usadas para se obter uma sequência de palavras que resultam na sentença da língua de chegada.

252	6.2 Recurrent architectures	6.2 Arquiteturas recorrentes
253	The transformer, as presented in the previous section, is the model used in most current commercial NMT systems, but alternative neural models exist.	O transformador, conforme apresentado na seção anterior, é o modelo adotado na maioria dos sistemas comerciais de NMT, embora existam modelos neurais alternativos.
254	Another top model is the recurrent encoder–decoder model (Bahdanau et al. 2015).	Outro modelo muito utilizado é o de codificador-decodificador (Bahdanau et al. 2015).
255	Similarly to transformer-based models, there is an encoder that produces a collection of embeddings for the words in the input sentence and a decoder that uses attention to compute embeddings for each target word by integrating the information from the input words and the already generated target words.	De forma parecida com os modelos baseados em transformers, há um codificador que produz uma coleção de vetores para as palavras na sentença de entrada e um decodificador que usa atenção para calcular vetores de cada palavra-alvo integrando a informação das palavras de entrada e as palavras-alvo já produzidas.
256	The encoder and decoder in the recurrent model, however, compute the contextual word embeddings in a local manner in such a way that the embeddings for the fifth encoded word, for example, are based on the embeddings of the four first words, on the one hand, and the embeddings of the next words, on the other hand.	Entretanto, o codificador e o decodificador, no modelo recorrente, calculam os vetores contextuais de forma local de tal modo que os vetores da quinta palavra, por exemplo, sejam baseados nos vetores das quatro primeiras palavras, por um lado, e nos vetores das palavras seguintes, por outro.
257	This is achieved by traversing the input sentence from left to right and from right to left; see Figure 7 for a diagram of this model showing only left-to-right processing.	Esse procedimento é possível porque a sentença de entrada é lida tanto da esquerda para a direita quanto da direita para a esquerda; a Figura 7 traz um diagrama deste modelo mostrando apenas o processamento da esquerda para a direita.
258	Figure 7: Left-to-right submodel of the encoder of a recurrent neural machine translation system, just after processed “<start> The brown” and when about to processs “fox”.	Figura 7: submodelo da esquerda para a direita do codificador de um sistema de tradução automática, logo após o processamento de "<início> A raposa" e logo antes de processar "marrom".
259	It is worth noting that the mathematical model used imposes some restrictions on the relevance given to the words around the word for which the contextual word embeddings are computed (in our example the fifth one), resulting in a mechanism that specially focuses on the nearest words and tends to ignore the representations of distant words.	É relevante observar que o modelo matemático empregado impõe certas restrições ao grau de relevância atribuído às palavras circundantes àquela para a qual os vetores estão sendo calculados (no exemplo em questão, a quinta palavra), culminando em um mecanismo que confere um enfoque especial às palavras mais próximas e tende a menosprezar as representações das palavras mais distantes.
260	Similarly to the transformer, a final layer at the end of the decoder computes a vector that gives the probability of each target-language word being the word at the corresponding position in the output sentence.	De forma semelhante ao transformer, uma camada final, ao fim do decodificador, calcula um vetor que dá a probabilidade de que cada palavra traduzida ocupe a posição correspondente na sentença de resultado.
261	Forcada (2017) describes in more detail the recurrent encoder–decoder model and also discusses the kind of outputs that NMT produces.	Forcada (2017) descreve em maior detalhe o modelo recorrente de codificador-decodificador e também discute os tipos de saída que a NMT produz.
262	7 Additional settings	7 Parâmetros adicionais
263	7.1 Words and sub-words	7.1 Palavras e sub-palavras
264	According to what has been presented in this chapter, independently of whether a transformer or a recurrent model is used, an embedding is obtained for each word after training.	De acordo com o que foi apresentado neste capítulo, independentemente de se usar um modelo de transformer ou recorrente, obtém-se uma vetorização para cada palavra após o treinamento.
265	Does this mean that we end up having an embedding for every possible word in the language?	Isso significa que ao fim do processo é obtida uma vetorização correspondente a todas as palavras possíveis de uma língua?

266	Not really.	Na verdade, não.
267	Languages, specially those which are highly inflected or agglutinative, may easily have hundreds of thousands or even millions of different word forms.	Línguas, especialmente aquelas que são altamente flexionais ou aglutinantes, podem facilmente atingir centenas de milhares ou até milhões de formas diferentes das palavras.
268	In order to understand why this poses a challenge for NMT systems you should know that the number of word embeddings (which is referred as the vocabulary) conditions the number of weights in the neural network and that large neural networks often struggle to generalise to unseen data.	Para entender como esse cenário pode ser desafiador para NMTs, tenha em mente que o número de vetorizações lexicais (que é chamado de vocabulário) condiciona o número de pesos na rede neural e que redes neurais grandes muitas vezes têm dificuldade para generalizar resultados para dados não vistos.
269	The size of the vocabulary could be reduced by considering only those word forms present in the training corpus but this usually still implies considering a substantial number of words and raises a new issue: when training is finished and the NMT system undertakes the translation of new sentences containing words not in the training set, these unseen words will make the model perform clumsily and lose accuracy as every unknown word is assigned a single non-contextual embedding reserved for this situation.	O tamanho do vocabulário pode ser reduzido se considerarmos apenas as formas morfológicas presentes no corpus de treinamento, mas, normalmente, isso ainda implica utilizar um grande número de palavras e ainda provoca outro problema: quando o treinamento termina e o sistema NMT se propõe a traduzir novas sentenças contendo palavras fora do conjunto de treinamento, essas palavras não vistas vão fazer o modelo atuar de maneira desajeitada e perder acurácia, uma vez que cada palavra desconhecida vai receber um vetor não-contextual reservado para aquela situação.
270	The solution engineers came up with is to split words into so-called sub-word units.	A solução que os engenheiros bolaram foi de dividir palavras nas chamadas unidades sublexicais.
271	Ideally, these units should make linguistic sense and carry some components of meaning; for instance, splitting demystifying as de- + -myst- + -ify- + -ing surely makes more linguistic sense (and is therefore likely to be more helpful when it comes to performing machine translation) than splitting it as dem- + -ystif- + -yi- + -ng.	Em um mundo ideal, essas palavras devem fazer sentido linguisticamente e ter componentes com significado: por exemplo, dividir a palavra "desmistificando" como des + misti + fic + ando certamente faz mais sentido linguístico do que dividi-la como desm + istif + ican + do.
272	But performing a linguistically sound splitting requires the existence of a set of splitting rules and procedures for the language in question, a resource that may not be available for many languages.	Mas fazer uma divisão lexical adequada linguisticamente requer a existência de um conjunto de regras de divisão e procedimentos para a língua em questão, o que é um recurso que, para muitas línguas, pode não estar disponível.
273	There are more advanced methods such as SentencePiece (Kudo & Richardson 2018), which treats the whole text as a sequence of characters and performs word division (tokenization) and sub-word division in one fell swoop.	Existem métodos mais avançados, como o SentencePiece (Kudo & Richardson 2018), que trata o texto todo como uma sequência de caracteres e realiza a divisão do texto em palavras (tokenização) e as divisões sublexicais em uma tacada só.
274	Byte-pair encoding was originally a text compression algorithm: frequent letter (byte) sequences would be stored once and replaced by short codes to reduce the total storage needed.	A codificação de pares de bytes era originalmente um algoritmo de compressão: letras (bytes) frequentes são armazenados uma única vez e substituídos por códigos mais curtos para reduzir o armazenamento necessário.
275	A commonly-used workaround is to automatically learn splitting rules by inspecting large texts, such as one containing all the source or all the target sentences in the training set.	Uma forma comum de contornar o problema consiste em aprender automaticamente as regras de divisão sublexical ao inspecionar textos longos, como um que contenha todas as sentenças de origem ou todas as sentenças-alvo no conjunto de treinamento.
276	A popular approach is called byte-pair encoding (BPE) (Sennrich et al. 2016), and starts with letter-sized units which are joined	Uma abordagem popular é a chamada de codificação de pares de bytes (BPE, byte-pair encoding) (Sennrich et al., 2016), que começa com unidades do tamanho de uma letra que são unidas

	into two-letter, three-letter, etc. units when they appear frequently in the corpus.	para formar unidades de duas letras (ou três etc) quando elas aparecem frequentemente no corpus.
277	Byte-pair encoding would probably identify a frequent -ing suffix in many verb forms (marching, considering) and chop it off, even for unseen forms (such as bartsimpsoning); -ing would then be turned into a contextual embedding carrying its atomic meaning.	A codificação por pares de bytes provavelmente identificaria um sufixo -ndo frequente em várias formas verbais (andando, pensando) e o cortaria, mesmo para formas verbais não vistas (como bartsimpsonando); o sufixo -ndo seria então transformado em um vetor contextual que traria consigo seu significado atômico.
278	7.2 Stopping criteria and metrics	7.2 Critérios de parada e métricas
279	As mentioned in section 3.5, in addition to a large training corpus, a small development corpus is usually held out and not used for training.	Conforme mencionado na seção 3.5, além de um corpus de treinamento grande, um corpus de desenvolvimento pequeno normalmente é reservado e não é utilizado para treinamento.
280	The purpose of this corpus is to monitor the performance of the NMT system while it is being trained, to decide, for instance, when training should stop.	O propósito desse corpus é monitorar o desempenho de um sistema NMT enquanto ele está sendo treinado, para decidir quando o treinamento deve parar.
281	Training tries to minimise an error function (or, in NMT, actually maximise the probability of the target sentences in the training corpus).	O treinamento busca minimizar uma função de erro (ou, no caso da NMT, maximizar a probabilidade das sentenças-alvo no corpus de treinamento).
282	One possible problem that may occur is that training too deep on the training corpus hurts generalisation as the neural network ends up memorising the example translations too much.	Um possível problema que pode surgir é que treinar demais no corpus de treinamento prejudica as generalizações, pois a rede neural acaba memorizando excessivamente os exemplos de treinamento.
283	This is where the development corpus comes into play: after a certain number of iterations or steps of the training algorithm, the source sentences in the development corpus are translated with the neural network and the output is automatically compared to the desired target sentences in the corpus using simple approximate automatic evaluation metrics (see Rossi & Carré 2022 [this volume]), the most common of which is BLEU (Papineni et al. 2002).	É aí que entra o corpus de desenvolvimento: depois de um certo número de iterações (ou passos) do algoritmo de treinamento, as sentenças de origem do corpus de desenvolvimento são traduzidas com a rede neural e a saída é comparada automaticamente com as sentenças-alvo desejadas do corpus por meio de métricas simples de avaliação automática aproximada (vide Rossi & Carré 2022 [volume atual]), dentre as quais a mais comum é o BLEU (Papineni et al., 2002).
284	BLEU measures how many one-word, twoword, three-word and four-word sequences in the output are found in the reference, and computes a score that varies from 0 (no match) to 100% (all stretches found).	O BLEU conta quantas sequências de uma, duas, três e quatro palavras na saída são encontradas na referência, e calcula uma nota que varia de 0 (nenhuma correspondência encontrada) até 100% (todas correspondências encontradas).
285	If, during training, BLEU on the development set starts to signal a degradation of performance, training may be stopped, or the current set of weights may be stored and training then continued for a while to see if BLEU improves again.	Se, durante o treinamento, o BLEU encontrar uma deterioração do desempenho, o treinamento pode ser interrompido, ou o conjunto atual de pesos pode ser congelado e o treinamento continuar por algum tempo para ver se o BLEU aumenta novamente.
286	Of course, there are many other automatic evaluation metrics which can take the place of BLEU in this process.	Naturalmente, existem muitas outras métricas automáticas de avaliação que podem tomar o lugar do BLEU no processo.
287	7.3 Beam search	7.3 Busca por feixe
288	The decoder in NMT systems produces the output sentence sequentially, one target word at a time, as explained in Sections 6.1 and 6.2.	O decodificador em sistemas NMT produz a sentença de saída sequencialmente, uma palavra-alvo por vez, como explicado nas seções 6.1 e 6.2.
289	At each time step, the neural network produces a probability or likelihood (a value	A cada iteração, a rede neural produz uma probabilidade ou verossimilhança (um valor entre 0

	between 0 and 100%) for every single word in the target vocabulary.	e 100%) a cada uma das palavras do vocabulário-alvo.
290	One way of using this information is to pick the most likely target word and output it, ignoring other possibilities.	Uma forma de usar essa informação é escolher a palavra-alvo mais provável e reproduzi-la, ignorando outras possibilidades.
291	It is worthwhile noting that, in doing so, we are completely determining the ensuing steps taken by the NMT system as the current prediction is given as input to the decoder in the next step (see, for example, the word zorro in Figure 6).	É interessante notar que, ao proceder desta maneira, os próximos passos dados pelo sistema NMT são completamente determinados, já que a previsão atual é usada para alimentar o decodificador do passo seguinte (vide, por exemplo, a palavra zorro, na Figura 6).
292	One possible way to explore more possibilities is to consider, for instance, the three most likely words, and clone the system into three systems, each of which would be determined respectively by each of the three choices, and see how they fare.	Uma forma de explorar mais possibilidades é considerar, por exemplo, as três palavras mais prováveis e fazer três cópias do sistema, cada uma das quais determinada respectivamente pelas três palavras escolhidas, e avaliar seu desempenho.
293	But one cannot do this indefinitely, as one would triplicate the number of systems translating the sentence at each step, and their number would grow exponentially.	Entretanto, não é possível fazer isso indefinidamente, pois o número de sistemas traduzindo a sentença triplicaria a cada passo, aumentando exponencialmente.
294	To avoid that, only a certain number of systems are allowed to survive, namely those obtaining the best value in an approximate calculation of the probability of the full sentence that would be produced.	Para evitar isso, apenas um certo número de sistemas sobreviveria, a saber, aqueles que obtivessem a maior nota em um cálculo da probabilidade da sentença a ser produzida.
295	This is usually called beam search and is a common approximation in other probabilistic models of human language processing such as speech recognition.	Esse método normalmente recebe o nome de busca por feixe e é uma aproximação comum em outros modelos probabilísticos de processamento de língua natural, como reconhecimento de fala.
296	8 Conclusions	8 Conclusões
297	A multilingual model is a single neural network that is trained to translate between many different language pairs so that knowledge from well-resourced languages may be transferred to low-resourced ones.	Um modelo multilíngue é uma rede neural única que é treinada para traduzir entre vários pares de línguas diferentes, de modo que o conhecimento a respeito de línguas de corpora ricos possa ser transferido para línguas de corpora pobres.
298	Interestingly, multilingual models bring the possibility of zero-shot translation (Ko et al. 2021) in which a system may be able to translate with reasonable quality, for example, between Spanish and Upper Sorbian using a multilingual model trained on German–Upper Sorbian and Spanish–German corpora, even when no Spanish–Upper Sorbian parallel corpus is available.	Modelos multilíngues acarretam a possibilidade da tradução de zero paralelismo (Ko et al., 2021), no qual um sistema pode traduzir com razoável qualidade, por exemplo, entre espanhol e alto sorábio usando modelos treinados com corpora dos pares alemão-alto sorábio e espanhol-alemão, mesmo sem um corpus paralelo de espanhol e alto-sorábio.
299	Unsupervised NMT goes a step further by learning NMT systems from monolingual corpora only.	NMTs não supervisionadas vão um passo além, aprendendo apenas com sistemas de NMT de corpora monolinguais.
300	To train an NMT system, one needs thousands or even millions of examples of source sentence–target sentence pairs.	Para treinar um sistema NMT, são necessários milhares ou até milhões de exemplos de pares de sentenças de origem e sentenças-alvo.
301	For many language pairs, many domains and many text genres, such resources do not exist, which constrains many specific applications, but for well-resourced languages, general-purpose NMT is a reality and is very widely used, not only by translators.	No caso de muitos pares de línguas, muitos campos e muitos gêneros textuais, tais recursos não existem. Isso impõe restrições a muitas aplicações específicas, mas, para línguas abundantes em recursos, a NMT faz-tudo já é uma realidade e muito utilizada não apenas por tradutores.
302	Moreover, scientific advances in approaches such as multilingual models or unsupervised	Além disso, avanços científicos em abordagens como modelos multilíngues ou NMTs não supervisionadas recentemente começaram a

	NMT have recently started to produce promising results in low-resource scenarios.	produzir resultados promissores em cenários de escassez de recursos.
303	This chapter has introduced – and provided technical details of – the key elements in NMT systems, and explored how they interact in the two currently most popular architectures, namely transformer-based and recurrent neural networks.	Este capítulo apresentou e forneceu detalhes técnicos dos elementos chaves de sistemas de NMT e explorou como eles interagem nas duas arquiteturas mais populares, a saber, as baseadas em transformers e as baseadas em redes neurais recorrentes.
304	Research activity in the area is so intense at the time of writing that proposals for new models arise almost every month.	A pesquisa na área é tão intensa que, ainda no período de escrita deste artigo, propostas de novos modelos surgem quase todo mês.
305	Transformers are currently the paradigm of choice if enough parallel corpora are available for training, because they require shorter training times and allow subtle quality improvements in comparison to recurrent neural networks, but the picture may change dramatically at any time.	Atualmente, transformers constituem o paradigma escolhido no caso de haver um número suficiente de corpora paralelos disponíveis para treinamento, porque eles requerem menor tempo de treinamento e permitem melhoras sutis de qualidade em relação às redes neurais recorrentes, mas esse cenário pode mudar drasticamente a qualquer momento.
306	References	Referências
307	Bahdanau, Dzmitry, Kyunghyun Cho & Yoshua Bengio. 2015.	Bahdanau, Dzmitry, Kyunghyun Cho & Yoshua Bengio. 2015.
308	Neural machine translation by jointly learning to align and translate.	Neural machine translation by jointly learning to align and translate.
309	In Yoshua Bengio & Yann LeCun (eds.), 3rd International Conference on Learning Representations, ICLR 2015.	In Yoshua Bengio & Yann LeCun (eds.), 3rd International Conference on Learning Representations, ICLR 2015.
310	DOI: 10.48550/arXiv.1409.0473.	DOI: 10.48550/arXiv.1409.0473.
311	Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Dario Amodei, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever & Dario Amodei. 2020.	Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Dario Amodei, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever & Dario Amodei. 2020.
312	Language models are few-shot learners.	Language models are few-shot learners.
313	CoRR abs/2005.14165. https://arxiv.org/abs/2005.14165.	CoRR abs/2005.14165. https://arxiv.org/abs/2005.14165.
314	Forcada, Mikel. 2017.	Forcada, Mikel. 2017.
315	Making sense of neural machine translation.	Making sense of neural machine translation.
316	Translation Spaces 6(2). 291–309.	Translation Spaces 6(2). 291–309.
317	Goodfellow, Ian, Yoshua Bengio & Aaron Courville. 2016.	Goodfellow, Ian, Yoshua Bengio & Aaron Courville. 2016.
318	Deep learning.	Deep learning.
319	Cambridge, MA: MIT Press.	Cambridge, MA: MIT Press.
320	Hornik, Kurt. 1991.	Hornik, Kurt. 1991.
321	Approximation capabilities of multilayer feedforward networks.	Approximation capabilities of multilayer feedforward networks.
322	Neural Networks 4(2). 251–257.	Neural Networks 4(2). 251–257.
323	Ko, Wei-Jen, Ahmed El-Kishky, Adithya Renduchintala, Vishrav Chaudhary, Naman Goyal, Francisco Guzmán, Pascale Fung, Philipp Koehn & Mona Diab. 2021.	Ko, Wei-Jen, Ahmed El-Kishky, Adithya Renduchintala, Vishrav Chaudhary, Naman Goyal, Francisco Guzmán, Pascale Fung, Philipp Koehn & Mona Diab. 2021.

324	Adapting high-resource NMT models to translate low-resource related languages without parallel data.	Adapting high-resource NMT models to translate low-resource related languages without parallel data.
325	In Proceedings of the 59th annual meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, 802–812.	In Proceedings of the 59th annual meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, 802–812.
326	Kudo, Taku & John Richardson. 2018.	Kudo, Taku & John Richardson. 2018.
327	SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing.	SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing.
328	In	In
329	Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 66–71.	Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 66–71.
330	Brussels, Belgium: Association for Computational Linguistics.	Brussels, Belgium: Association for Computational Linguistics.
331	Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg Corrado & Jeffrey Dean. 2013.	Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg Corrado & Jeffrey Dean. 2013.
332	Distributed representations of words and phrases and their compositionality.	Distributed representations of words and phrases and their compositionality.
333	In Advances in Neural Information Processing Systems 30, 3111–3119.	In Advances in Neural Information Processing Systems 30, 3111–3119.
334	Papineni, Kishore, Salim Roukos, Todd Ward & Wei-Jing Zhu. 2002.	Papineni, Kishore, Salim Roukos, Todd Ward & Wei-Jing Zhu. 2002.
335	BLEU: A method for automatic evaluation of machine translation.	BLEU: A method for automatic evaluation of machine translation.
336	In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, 311–318.	In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, 311–318.
337	Philadelphia, Pennsylvania, USA: Association for Computational Linguistics.	Philadelphia, Pennsylvania, USA: Association for Computational Linguistics.
338	DOI: 10.3115/1073083.1073135.	DOI: 10.3115/1073083.1073135.
339	Rossi, Caroline & Alice Carré. 2022.	Rossi, Caroline & Alice Carré. 2022.
340	How to choose a suitable neural machine translation solution: Evaluation of MT quality.	How to choose a suitable neural machine translation solution: Evaluation of MT quality.
341	In Dorothy Kenny (ed.), Machine translation for everyone: Empowering users in the age of artificial intelligence, 51–79.	In Dorothy Kenny (ed.), Machine translation for everyone: Empowering users in the age of artificial intelligence, 51–79.
342	Berlin: Language Science Press.	Berlin: Language Science Press.
343	DOI: 10.5281/zenodo.6759978.	DOI: 10.5281/zenodo.6759978.
344	Sennrich, Rico, Barry Haddow & Alexandra Birch. 2016.	Sennrich, Rico, Barry Haddow & Alexandra Birch. 2016.
345	Neural Machine translation of rare words with subword units.	Neural Machine translation of rare words with subword units.
346	In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, 1715–1725.	In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, 1715–1725.
347	Berlin: Association for Computational Linguistics.	Berlin: Association for Computational Linguistics.
348	Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser & Illia Polosukhin. 2017.	Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser & Illia Polosukhin. 2017.
349	Attention is all you need.	Attention is all you need.
350	In Advances in Neural Information Processing Systems 30, 5998–6008.	In Advances in Neural Information Processing Systems 30, 5998–6008.

APÊNDICE II

Capítulo 7

Como funciona a tradução automática neural

Juan Antonio Pérez-Ortiz
Universidade de Alicante, Espanha

Mikel L. Forcada
Universidade de Alicante, Espanha

Felipe Sánchez-Martínez
Universidade de Alicante, Espanha

Este capítulo apresenta os princípios fundamentais subjacentes aos sistemas de tradução automática neural. São introduzidos, um a um, conceitos-chave utilizados para descrever esses sistemas, para que o leitor obtenha uma visão abrangente de seu funcionamento interno e possibilidades. Esses conceitos incluem: redes neurais, algoritmos de aprendizagem, vetorização de palavras, mecanismos de atenção e arquitetura codificador–decodificador.

1 Introdução

A primeira coisa que se deve entender em relação à tradução automática neural (NMT, do inglês Neural Machine Translation), é que ela considera a tradução como uma tarefa que envolve operações em números realizadas por sistemas matemáticos chamados *redes neurais artificiais*: esses sistemas tomam uma sentença e a transformam em uma série de números. Adicionam mais alguns números aqui (geralmente, milhares ou milhões deles), multiplicam por outros números ali, realizam algumas operações matemáticas adicionais relativamente simples e, por fim, produzem uma tradução da sentença original para outro idioma.

Muitos enxergam a tradução de uma perspectiva diferente: como uma tarefa intelectual que envolve processos cognitivos que dificilmente podem ser enumerados com clareza e que ocorrem em algumas áreas profundas do cérebro humano.

E essa é uma perspectiva correta. Mas a abordagem executada atualmente por computadores segue um caminho completamente diferente: milhões de operações matemáticas são realizadas em uma fração de segundo para obter uma tradução que ora pode, ora não pode ser considerada adequada. E o fato é que a percentagem de vezes em que a tradução é adequada aumentou dramaticamente nos últimos anos. Mas, historicamente, as redes neurais artificiais foram concebidas como um modelo simplificado de como funcionam as *redes neurais naturais*, como o cérebro humano, e os processos cognitivos realizados nelas são também o resultado de processos difusos de computação neural que não são tão diferentes das operações matemáticas mencionadas acima.

Este capítulo explicará em detalhe elementos-chave da tecnologia NMT. A começar por destacar a conexão entre as formas possíveis de traduzir de um cérebro humano e de um sistema NMT. Explicar essa conexão ajudará na apresentação dos conceitos básicos necessários para obter uma visão abrangente dos princípios de *aprendizado de máquina* e *redes neurais artificiais*, que constituem dois dos pilares da NMT. Em seguida, serão discutidos os princípios essenciais de *vetores não contextuais (non-contextual word embeddings)*, uma representação computadorizada de palavras com diversas propriedades interessantes que, quando combinadas através de um mecanismo conhecido como *"atenção"*, produz os chamados *vetores de palavras contextuais*, um fator-chave no entendimento da NMT. Todos esses ingredientes permitirão apresentar um quadro geral do funcionamento interno dos dois modelos de NMT mais utilizados, a saber, o *transformer* e os *modelos de redes neurais recorrentes*. O capítulo termina com a apresentação de uma série de temas secundários para expandir o conhecimento público sobre a forma com que estes sistemas funcionam por trás das cortinas.

2 Uma analogia imperfeita entre tradução humana e NMT

Para simplificar um pouco a discussão, vamos imaginar que traduzir um texto equivale a, grosso modo, traduzir cada uma das suas sentenças de forma independente. Suponha, por um momento, que a tradução de uma sentença é um processo de duas etapas: o tradutor primeiro *interpreta* ou determina o *significado* de toda a sentença de origem e, em seguida, produz de uma só vez uma sentença que permite mais ou menos a mesma interpretação, mas que agora está escrita na língua de chegada. Mas, todos os dias, tradutores encontram sentenças que nunca viram antes,

como "o lápis escorregou da minha mão, levantou-se e começou a falar comigo", e ainda conseguem traduzi-las: como isso é possível? A linguística formulou a resposta a essa questão como o *princípio da composicionalidade semântica*: os seres humanos *constroem* a interpretação de cada sentença combinando as interpretações individuais de suas palavras e a ordem em que são combinadas é ditada pela estrutura sintática da sentença em que as palavras formam orações, orações formam orações maiores, até chegar a toda a sentença. Um tradutor então analisa essa interpretação e executa o procedimento inverso, mas na língua de chegada. É evidente que os tradutores nem sempre interpretam as sentenças como um todo, especialmente ao se depararem com sentenças longas, podendo recorrer a atalhos para evitar interpretações de sentenças inteiras logo de início. Contudo, vamos manter essa simplificação, por enquanto.

NMTs funcionam de forma semelhante. Ao traduzir uma sentença, durante a sentença de codificação, o sistema atribui uma *representação*, ou *vetorização*, para cada palavra do texto-fonte isoladamente. Essas representações neurais são então combinadas para produzir uma representação semelhante, mas desta vez no nível da sentença. À medida que são combinadas, as representações individuais também são modificadas de acordo com o seu contexto; pode-se considerar isso uma representação contextualizada de interpretação ou significado. Em seguida, na fase de decodificação, as representações das sentenças são desvendadas passo a passo para prever, uma a uma, as palavras na sentença-alvo. O *codificador* e o *decodificador* que executam essas duas fases são redes neurais artificiais interconectadas que formam uma única rede neural composta.

Como no caso dos tradutores, as arquiteturas neurais atuais não funcionam, de fato, considerando toda a sentença de origem ao produzir cada palavra-alvo, mas aprenderam a prestar *atenção* nas palavras-fonte relevantes e nas palavras-alvo já produzidas quando fazem isso.

Nas seções restantes deste capítulo, o leitor será apresentado a uma descrição mais detalhada da natureza destas representações, à estrutura das redes neurais artificiais (que podemos simplesmente chamar de "redes neurais", a partir de agora) que as constroem e transformam, prestando atenção seletivamente ao que é importante, e às formas como estas redes neurais artificiais podem ser treinadas para realizar tal tarefa utilizando exemplos de tradução.

3 Redes Neurais Artificiais

Para entender o conceito de NMT, é preciso considerar com mais detalhes as redes neurais artificiais (Goodfellow et al. 2016) que a realizam: do que são feitas, como funcionam e como são treinadas.

O adjetivo *"neural"* remete diretamente a neurônios e à maneira como funcionam os sistemas nervosos de animais e, principalmente, do cérebro humano. As redes neurais artificiais são, de fato, constituídas por milhares ou milhões de unidades artificiais que se assemelham a neurônios cuja *ativação* (ou seja, o quanto estão *excitados* ou *inibidos*) depende dos sinais que recebem de outros neurônios e da força das conexões que transmitem esses sinais.

3.1 Neurônios artificiais

Os neurônios artificiais são os principais blocos de construção das redes neurais artificiais. A operação desses neurônios artificiais (que, a partir de agora, chamaremos simplesmente de neurônios) de atualizar seu estado ou ativação pode ser analisada em duas etapas. Apresenta-se, na Figura 1, uma situação simplificada, na qual se observa como o grau de ativação do neurônio4 é atualizado em decorrência dos estímulos recebidos pelos neurônios 1, □2, e S3.

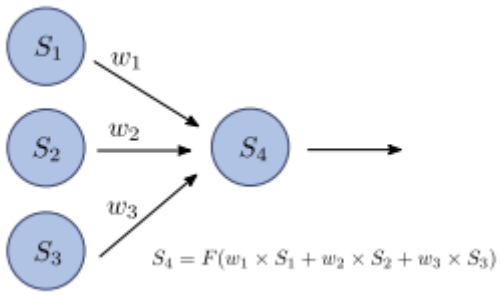


Figura 1: atualização do estado S4 do neurônio 4 em resposta aos estímulos recebidos dos neurônios 1, 2 e 3.

Na primeira etapa, os graus de ativação dos neurônios S1, □2 e □3, todos eles ligados ao neurônio4, são somados, mas primeiro cada um é multiplicado por um *peso* (□1, □2 e □3) representando a força de suas conexões; esses pesos determinam como sua ativação é transformada em estímulos reais para o neurônio S4. Os pesos podem ser positivos ou negativos. Por exemplo, se o peso w2 é positivo e o grau de ativação de S2 é elevado, ele contribuirá para ativar o neurônio S4 (um

estímulo positivo); no entanto, se w2 é negativo, contribuirá para inibir o neurônio S4 (um estímulo negativo). Em termos gerais, os neurônios conectados por pesos positivos tendem a ser ativados ou inibidos simultaneamente, enquanto os neurônios conectados por pesos negativos tendem a estar em estados opostos. Retornando à análise do neurônio S4, ao se adicionar os estímulos provenientes de cada neurônio, obtém-se um *saldo líquido de estímulo*:

$$□ = □_1 □ □_1 + □_2 □ □_2 + □_3 □ □_3 \ (1)$$

O estímulo líquido pode assumir qualquer valor possível, negativo ou positivo, mas ainda não é a ativação do neurônio S4 . Na segunda etapa, o neurônio S4 *reage* a este estímulo. No exemplo, quando o estímulo é intermediário, ou seja, não muito positivo ou muito negativo, o neurônio S4 é muito sensível a ele. No entanto, quando os estímulos ficam grandes (não importa se positivos ou negativos), as mudanças em seus valores têm um impacto menor na produção, pois o neurônio é, respectivamente, amplamente inibido ou amplamente ativado.

No exemplo, o neurônio S4 é tal que o seu grau de ativação está confinado entre -1 e +1. A Figura 2 representa o modo como o neurônio S4 reage ao estímulo da equação 1. A reação é representada por uma função F(...), denominada *função de ativação*, que é aplicada ao estímulo; o resultado é o grau de ativação de S4:

$$□_4 = □(□) = □(□_1 □ □_1 + □_2 □ □_2 + □_3 □ □_3) \ (2)$$

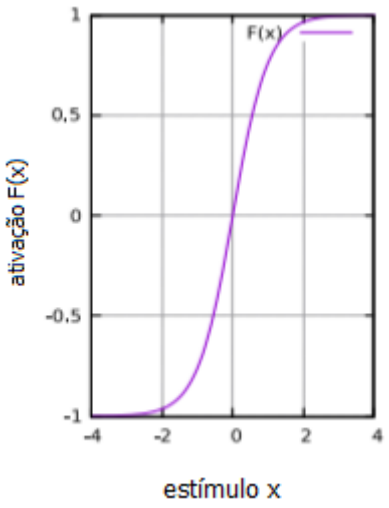


Figura 2: como um neurônio reage ao estímulo total recebido

Como pode ser visto, para valores em torno de 0 no eixo horizontal, a reação é proporcional ao estímulo, mas, para grandes estímulos positivos ou negativos, quando o neurônio está muito

inibido ou muito excitado, a reação é bem menor. Para este tipo de neurônio, os valores extremos reais de -1 e +1 nunca são alcançados, não importa o quanto o estímulo total seja forte. Como dito acima, o neurônio S_4, no exemplo em questão, é um tipo específico de neurônio com um grau de ativação que varia entre -1 e +1. Existem outros tipos de funções de ativação com diferentes intervalos, mas explorá-las está fora do âmbito de aplicação deste capítulo.

3.2 De neurônios até redes.

Neurônios como os discutidos na seção anterior podem ser conectados para formar uma rede neural artificial que executa uma tarefa computacional específica para resolver um problema específico. Em uma rede, alguns neurônios recebem estímulos externos que atuam como *entradas* para a rede (assim como os olhos estão ligados ao cérebro humano cérebro e o alimentam com imagens) e representam uma instância do problema a ser resolvido; alguns neurônios, conhecidos como *neurônios ocultos*, recebem estímulos apenas de outros neurônios; e, finalmente, alguns neurônios, conhecidos como neurônios de saída, representam a solução para o problema (de maneira similar aos sinais enviados aos músculos de uma das mãos para movê-la de uma forma específica). A Figura 3 mostra um exemplo de uma rede neural com cinco neurônios; a rede recebe três entradas, que são alimentadas a três neurônios ocultos, que, por sua vez, estimulam dois neurônios de saída.

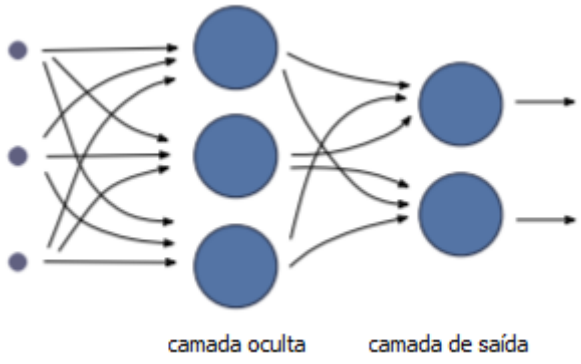


Figura 3: uma rede neural artificial com três neurônios ocultos e dois neurônios de saída. Cada conexão tem um peso não indicado no diagrama. Os três neurônios de entrada à esquerda são representados por círculos menores para enfatizar a ideia de que eles emitem

> diretamente os valores da entrada externa, mas, ao contrário dos neurônios regulares, eles não calculam um estímulo ou reagem a ele por meio de uma função de ativação.

Ao construir uma rede neural para resolver um problema específico, primeiro é necessário determinar sua *arquitetura*: quantos neurônios ela têm, como eles estão conectados, quais recebem entradas externas e quais são designados como neurônios de saída; mas o cálculo efetivamente realizado depende dos pesos de ponderação de todas as ligações na rede. A forma como estes pesos são calculados é explicada na Seção 3.5. Por ora, basta dizer que uma característica interessante das redes neurais artificiais é que elas podem ser *treinadas* para executar uma tarefa a partir de exemplos. Ou seja, os seus pesos podem ser definidos para valores específicos, observando um conjunto de exemplos resolvidos, cada um composto pelos valores dos sinais de entrada que representam os problemas e os valores dos graus de ativação de saída desejados, que representam as soluções.

3.3 Camadas de neurônios

Supunha um completo iniciante que queira aprender algumas técnicas básicas para pintar paisagens em óleo sobre tela. Um manual pode ensinar um método passo-a-passo simplificado com, por exemplo, estas quatro etapas: desenho (uma composição aproximada é esboçada), distribuição de cores, refinamento do desenho e acabamento (quando os retoques finais são feitos). O ponto aqui não é o número de passos ou as características particulares de cada uma deles, mas o fato de que todo o processo flui de forma incremental de tal forma que a saída de uma etapa se torna a entrada para a próxima. Cada etapa refina o resultado anterior: o resultado da segunda etapa (distribuição de cores) está mais próxima de uma pintura de paisagem real do que o resultado da primeira (desenho) e, da mesma forma, o resultado da quarta etapa (acabamento) pode ser conceitualmente considerado como uma pintura melhor do que as resultantes de qualquer uma das etapas anteriores.

Ocorre que a computação neural se beneficia de um processo incremental passo-a-passo semelhante. Nos anos sessenta, pesquisadores descobriram que, ao incluir várias camadas de neurônios, tarefas mais complexas poderiam ser abordadas. Cada camada em uma rede neural multicamada refina a saída da camada anterior e dá um passo maior ou menor em direção à solução final. A arquitetura resultante seria semelhante à da Figura 3, mas com várias outras

camadas ocultas. Pode-se ver claramente esta estrutura em camadas na rede simples na Figura 3: a computação, realizada por duas camadas, ocorre em duas etapas.

Um modelo feito de neurônios organizados em camadas recebe o nome de *rede neural em camadas*. Apesar de os resultados teóricos provarem que uma rede de duas camadas tem poder computacional suficiente para realizar praticamente qualquer tarefa (Hornik 1991), no mundo real, o poder computacional das redes neurais parece estar correlacionado com o número de camadas; modelos com mais do que algumas poucas camadas são frequentemente rotulados como *redes neurais profundas* e os algoritmos de treinamento correspondentes são conhecidos como *algoritmos de aprendizagem profunda*.

Como exemplo da complexidade que esses modelos profundos podem atingir, o modelo GPT-3 (Brown et al. 2020), uma das maiores redes neurais lançadas em 2020 no campo da geração de língua natural, possui 96 camadas com dezenas de milhares de neurônios cada, o que resulta em cerca de 175 bilhões de pesos a serem aprendidos pelo algoritmo de treinamento. Para treinar o sistema GPT-3, foram utilizados supercomputadores, em um processo que pode demorar várias semanas ou mesmo meses, mas estima-se que a aprendizagem dos pesos para um modelo desse tipo com um único computador pessoal para jogos poderoso teria levado mais de 350 anos[5].

3.4 Tradução automática neural

Se uma sentença de origem for representada como um conjunto de entradas para uma rede neural, e se for possível interpretar as saídas da rede neural como uma sentença de destino, um sistema de *tradução automática neural* (NMT) é estabelecido. A NMT primeiro processa as palavras na sentençade origem. Cada vez que uma palavra-fonte é ingerida pela parte codificadora da rede neural, os graus de ativação de conjuntos específicos de neurônios na rede mudam. Quando toda a sentença de origem tiver sido processada, a parte decodificadora da rede inicia o seu trabalho. Ela foi treinada para fornecer, passo a passo, um valor de probabilidade para cada palavra-alvo possível na tradução, dadas as palavras-alvo que já produziu. Esse mecanismo é semelhante à forma como os teclados preditivos em smartphones contemporâneos funcionam, mas, como será

[5] "Modelo de linguagem GPT-3 da OpenAI: uma revisão técnica" (2020). Obtido em https://lambdalabs.com/blog/demystifying-gpt-3.

elucidado, as previsões de palavras de uma NMT também dependem da sentença de origem, pois devem ser uma tradução dela.

Os sistemas NMT são redes neurais profundas com arquiteturas que serão discutidas mais adiante, na seção 6. Eles têm milhares de neurônios e milhões de pesos (ou muitos mais) que precisam ser treinados fornecendo exemplos retirados de um corpus paralelo contendo milhões de sentenças-fonte e suas traduções. Representações matemáticas das palavras de uma determinada sentença na língua de origem são alimentadas à rede neural como entradas e as palavras da sentença correspondente na língua de destino são usadas para representar a saída desejada. Como se pode imaginar, treinar uma grande rede em tempo razoável é computacionalmente intensivo: é necessário um sistema de processamento muito poderoso e especializado para treinar a rede, mostrando os exemplos repetidamente. A cada iteração, pequenas alterações são feitas nos pesos na rede para melhorar sua previsão das palavras-alvo.

3.5 Treinando redes neurais

Treinar uma rede neural é o processo de determinar o peso das conexões entre seus neurônios de modo que, dado um *conjunto de treinamento* de exemplos de entrada e suas saídas respectivas, produz uma saída real o mais próxima possível da do exemplo em estudo.

O treinamento começa com um conjunto de pesos aleatórios ou com pesos retirados de uma rede neural resolvendo uma tarefa semelhante. Durante o treinamento, os pesos são modificados de tal forma que o valor de uma função de erro (também conhecida como *função de perda*), que mede o quanto as saídas reais se desviam das saídas desejadas, seja o menor possível. *Algoritmos de treinamento* (também chamados de algoritmos de aprendizagem) realizam repetidamente pequenas correções (atualizações) nos pesos até que a função de erro seja mínima (ou pequena o suficiente) para todos os exemplos no conjunto de treinamento, ou até que um determinado desempenho seja observado em um *conjunto de desenvolvimento*, que é reservado para esse propósito (ver seção 7.2). Os detalhes técnicos do algoritmo de treinamento excedem o escopo deste capítulo; mencione-se apenas que ele se fundamenta, em geral, no cálculo do quanto a função de erro varia quando cada peso é variado por uma quantidade fixa, mas muito pequena (o *gradiente* da função de erro) e, em seguida, cada peso é ligeiramente ajustado na direção que reduz

a função de erro.[6] Este tipo de treinamento chama-se *gradiente descendente*; não é garantia de que ele vai encontrar os melhores pesos, mas é provável que encontre bons candidatos. A intensidade destas variações de peso é regulada por um parâmetro denominado *taxa de aprendizagem*; esta taxa de aprendizagem é geralmente maior nas primeiras etapas do algoritmo de treinamento, mas sua magnitude diminui progressivamente à medida que os pesos se aproximam de seus valores finais. Observe que o treinamento de redes neurais é bastante trabalhoso: muitos exemplos são necessários e precisam ser apresentados muitas vezes para que elas aprendam. No entanto, isto se deve frequentemente às limitações dos algoritmos de treinamento, e não à falta de capacidade de uma rede neural específica para representar a solução de um problema.

Uma vez determinados os pesos, o treinamento para (ver seção 7.2) e a rede neural pode ser utilizada para obter as saídas para novas entradas que não estão incluídas entre os exemplos utilizados durante o treino.

3.6 Generalização em redes neurais

Generalização é um processo cognitivo fundamental para humanos e animais. Permite-se, por meio dele, a utilização do que foi aprendido no passado em novas situações que podem ser consideradas semelhantes, mas não idênticas à situação em que a aprendizagem ocorreu originalmente. Uma pessoa não precisa reaprender a dirigir ao entrar em uma nova rua ou dirigir um carro novo. Analogamente, a generalização acontece quando um organismo que já reage a um determinado estímulo de uma maneira particular, responde de formas semelhantes a estímulos parecidos. Aplicar generalizações também é fundamental para a aprendizagem de línguas: as crianças aprendem rapidamente a pronunciar sentenças que nunca ouviram antes.

As redes neurais podem, idealmente, aplicar generalizações no contexto da tradução automática, produzindo resultados semelhantes quando alimentadas com entradas semelhantes, independentemente de terem sido ou não incluídas no conjunto de treinamento. Uma característica das redes neurais é a *suavidade* dos cálculos, o que significa que, se os valores de entrada forem ligeiramente alterados, o resultado das fórmulas não irá variar significativamente.

[6] Alguns dos leitores poderão reconhecer aqui o conceito matemático de *derivada de uma função*.

Em sentido amplo, para aplicar uma generalização, sentenças semelhantes devem obter representações parecidas e, como as representações de sentenças serão obtidas a partir de representações de palavras, conclui-se que representar palavras similares e com números semelhantes é uma pré-condição para a generalização no processamento neural de língua.

Na próxima seção, será aprofundada a maneira com que se obtém uma lista de representações neurais para palavras de uma sentença que se beneficiaria da fluidez das redes neurais de forma que, após o treinamento, o sistema seria capaz de fazer generalizações adequadas de sentenças que nunca viu antes.

4 Vetores como representações de palavras

Na seção anterior, foi observado que os neurônios geralmente são organizados em camadas de maneira que a saída dos neurônios de uma camada se transforma na entrada para os neurônios da camada seguinte. Curiosamente, o produto do conjunto de neurônios em uma determinada camada consiste numa representação das informações que eles processam naquele estágio.

No campo do processamento de língua natural, e conforme já foi mostrado, as informações processadas pelas redes neurais são compostas de palavras, e suas representações na rede costumam ser chamadas de *vetores* (embeddings) (Mikolov et al. 2013). O que torna essas vetorizações realmente úteis é o fato de que as palavras com significados semelhantes ou que comumente ocorrem nos mesmos contextos acabam por ter vetorizações semelhantes. Para uma melhor compreensão, use um pedaço de papel para desenhar um quadrado com lados de aproximadamente dez centímetros. A seguir, reúna as palavras e organize-as dentro do quadrado, observando o critério de que palavras com significados mais similares estejam posicionadas mais próximas entre si em comparação às palavras com significados mais distintos. Se o conceito de proximidade de sentido se mostrar inexato, é possível organizar as palavras com base na frequência de ocorrência conjunta em sentenças ou parágrafos. As palavras são: *restaurante, vermelho, jardim, fonte, flor, tomate, balão, garçons, faca, flores, cardápio, cozido, cromossomo* e *consistentemente*. Faça isso antes de dar continuidade à leitura.

A restrição imposta por meio do critério de proximidade semântica significa que não existe a possibilidade de distribuir livremente as palavras no quadrado. Um agrupamento mais intuitivo

poderia ser composto de palavras como *restaurante, cardápio* e *garçons*, de um lado, e palavras como *jardim, flor* e *fonte*, de outro. Há, no entanto, alguns casos questionáveis: *vermelho* é claramente próximo de *tomate*, mas também deve ficar próximo a *flor*; uma solução conciliatória poderia ser colocá-lo em um ponto intermediário, um pouco mais perto de *tomate* do que de *flor*, levando em consideração que a cor vermelha não é tão essencial às flores quanto é aos tomates.

Alguns subgrupos são perceptíveis neste agrupamento: uma ilha representando o campo semântico de restaurantes e coisas relacionadas, e outra ilha relacionada a jardins e pomares. Há algumas exceções na lista, especialmente a palavra *consistentemente*, que parece, por princípio, desvinculada das outras palavras, impondo a necessidade de colocá-la o mais distante possível de todas as outras. *Cromossomo* é outra palavra isolada, mas, como as flores e os garçons, têm cromossomos para transportar as próprias informações genéticas, a palavra pode ser colocada em algum lugar em entre essas palavras, porém, ao mesmo tempo, um pouco longe da palavra *vermelho*. Consulte a Figura 4 e observe uma solução possível e que talvez não corresponda exatamente à solução típica elaborada pelo leitor.[7]

Com o propósito de atribuir códigos matemáticos às palavras da lista em questão, serão designadas coordenadas para cada palavra, a fim de refletir suas respectivas posições no quadrado Considerando que o espaço em questão é bidimensional, duas coordenadas são necessárias para cada palavra: a primeira coordenada é um número que representa a distância horizontal até o canto inferior esquerdo do quadrado; a segunda coordenada é um número que representa a distância vertical até o mesmo ponto. A palavra *restaurante* poderia ser atribuída, por exemplo, aos dois números 0,25 e 1,1, e a palavra *menu* aos números 0,6 e 1,3, próximo a *restaurante*, como visto na Figura 4. Essas coordenadas podem ser representadas usando a *notação vetorial*, que consiste simplesmente em escrever os números como uma lista de valores separados por vírgulas entre colchetes. Os vetores correspondentes a *restaurante* e *cardápio* seriam, portanto, [0,25, 1,1] e [0,6, 1,3], respectivamente. Cada um desses vetores representa uma possível vetorização semântica para essas duas palavras.

Embora não seja completamente óbvio, avaliar vetorizações compostas por dois números em vez de um único número aumenta as possibilidades de uma solução para o problema de posicionar

[7] Propositalmente, colocamos a Figura 4 algumas páginas mais adiante, para que o leitor não a veja antes de tentar fazer o exercício.

palavras mais perto ou mais longe, pois mais liberdade é proporcionada para satisfazer todas as restrições. De fato, a mudança de duas dimensões para um número maior de dimensões aumenta ainda mais essas possibilidades. Uma representação de cinco dimensões de uma palavra poderia ser, por exemplo, [2,34, 1,67, 4,81, 3,01, 5,61]. Os sistemas de NMT analisam as vetorizações com centenas de dimensões, e a sentença de entrada a ser traduzida é representada por uma coletânea de vastas vetorizações semânticas.

A vetorização semântica é aprendida com o mesmo algoritmo usado para aprender os pesos da rede neural apresentada na Seção 3.5 Na verdade, tanto os pesos como as vetorizações são aprendidos ao mesmo tempo. Ao levar em conta que a camada de entrada de uma rede neural utilizada na NMT é geralmente composta pelas vetorizações das palavras na sentença de entrada, não há necessidade de limitar-se a vetores fixos. Em vez disso, seus valores podem ser atualizados repetidamente durante o treinamento, de modo que o valor da função de erro seja minimizado.

4.1 Generalização

Como já discutido, para que a rede possa fazer *generalizações* adequadamente, ou seja, para aprender a traduzir e ser capaz de traduzir sentenças nunca vistas antes, sentenças semelhantes devem receber representações semelhantes. Como as representações de sentenças são obtidas a partir de vetorizações de palavras, pode-se concluir que representar palavras semelhantes com números semelhantes é uma condição prévia para a generalização no processamento neural de língua natural. No exemplo proposto, palavras como *alagou, choveu, alagando* ou *chovendo* devem, idealmente, têm vetorizações semelhantes, uma vez que todas são semanticamente semelhantes; os códigos para *alagando* e *chovendo* também devem estar mais perto de palavras como *dirigindo*, uma vez que os três são gerúndios e podem aparecer em contextos semelhantes; *alagou* e *choveu* devem ficar próximos também porque ambos são verbos no pretérito. É por essa razão que a necessidade de muitas dimensões é frequentemente observada: busca-se que as palavras estejam, simultaneamente, próximas umas das outras de maneiras diferentes ou por motivos diferentes.

4.2 Propriedades geométricas de vetorização semântica.

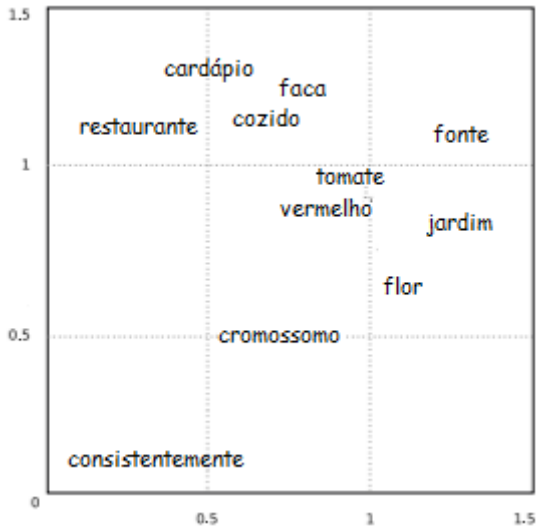
As vetorizações têm propriedades interessantes que representam características semânticas (ou relacionadas à semântica) de palavras. Como já explicado, a vetorização de uma palavra consiste em vários números reais, geralmente centenas ou milhares deles, e cada um desses valores captura um determinado aspecto do significado de uma palavra. Por exemplo, a vetorização semântica de palavras para *Dublin* deve reunir vários aspectos relacionados a ela: uma cidade, a capital da Irlanda, o local da sede na Europa de várias empresas multinacionais etc.

Graças a essa particularização das diferentes dimensões dos vetores semânticos, é possível executar algumas operações aritméticas com as incorporações e obter resultados expressivos. Essas operações são simplesmente adições e subtrações que são fáceis de calcular. A adição (ou subtração) de dois vetores consiste simplesmente em adicionar (ou subtrair) seus componentes um a um; por exemplo, [1.24, 2.56, 5.23] + [0.12, 1.12, 0.01] = [1.36, 3.68, 5.24]. Abaixo estão dois exemplos de operações aritméticas com resultados significativos realizados em incorporações que os sistemas NMT normalmente aprendem:

$$[\square\square\square] - [h\square\square\square\square] + [\square\square\square h\square\square] \simeq [\square\square\square\square h\square]$$

$$[\square\square\square\square\square\square] - [\square\square\square\square\square\square\square] + [\square\square\square\square ç\square] \simeq [\square\square\square\square\square]$$

onde os colchetes referem-se às vetorizações de uma palavra e, com o uso do símbolo ≃, indica-se que a vetorização resultante após a operação é aproximada da vetorização da palavra ao lado direito da "equação" do exemplo. Isso pode ser interpretado como uma indicação de que o *rei* é para o *homem* o que a *rainha* é para a *mulher*, um monarca masculino ou feminino; e *Dublin* é para a *Irlanda* o que *Paris* é para a *França*, a capital de um país.



Colocação de palavras em uma área bidimensional de forma que as palavras relacionadas sejam posicionadas próximas umas das outras, mas longe das palavras com as quais têm menos em comum

5 Vetores contextuais por meio de atenção

As palavras nem sempre têm o mesmo significado em todas as sentenças. A vetorização da palavra *carta*, por exemplo, não deve ser a mesma quando a palavra se refere a um elemento do baralho ou quando se refere a um documento destinado a outra pessoa. Na verdade, pode até ser interessante para um sistema NMT representar a palavra com diferentes vetorizações, a depender de o termo se referir a uma carta de amor ou a uma carta de reclamação. As vetorizações previamente apresentadas são *não-contextuais*: elas foram calculadas levando em consideração palavras que geralmente ocorrem juntas em sentenças, mas sem levar em conta os diferentes significados que as palavras podem ter.

No contexto da NMT, a *atenção* desempenha um papel importante, pois permite que a rede neural compute *vetores contextuais*, ou seja, representações vetoriais das palavras em uma sentença computadas de tal forma que a representação obtida para uma palavra seja adaptada ao seu significado em cada sentença específica. A atenção é, mais uma vez, um conceito que é implementado por meio de operações matemáticas aprendidas de forma conveniente por um algoritmo de treinamento. No contexto em discussão, a atenção assemelha-se à situação na qual a atenção é dispensada a algo ou a alguém na vida cotidiana.

Ao usar a atenção de forma conveniente para se concentrar em algumas palavras da sentença, o vetor semântico correspondente à palavra temporada, por exemplo, será diferente nas sentenças dos exemplos 1 e 2 abaixo:

1. O primeiro episódio vai continuar exatamente de onde a temporada anterior parou.
2. O verão é a temporada mais quente do ano.

Em princípio, pode parecer que o propósito dos vetores contextuais de palavras fosse o de atribuir representações diferentes a significados diferentes de uma palavra, mas, embora isso geralmente seja verdade, a ideia vai além disso. Os vetores contextuais da palavra para *temporada* nas sentenças "o inverno é a temporada mais fria do ano nas zonas polares e temperadas", "o verão é a temporada mais quente de todo o ano" e até mesmo "de todo o ano, a temporada de verão é a mais quente" serão todos diferentes, embora presumivelmente mais próximos uns dos outros do que a representação de *temporada* em "o primeiro episódio vai continuar exatamente de onde a temporada anterior parou". Estas divergências resultam do fato de as palavras das sentenças ou a ordem em que são colocadas serem diferentes. É possível notar que, em cada um dos exemplos propostos, as instâncias repetidas de um mesmo artigo apresentarão dois vetores contextuais distintos, uma vez que o contexto de cada instância varia.

Como os vetores são computados matematicamente através da atenção? Dada a sentença no exemplo 2 acima ("o verão é a temporada mais quente do ano."), o procedimento é iniciado pela obtenção dos vetores não-contextuais de palavras apresentados na seção 4. Dado que a sentença contém nove palavras, o resultado é uma coleção de nove vetores, que são os elementos fundamentais para a próxima etapa. Agora, para calcular o vetor contextual da palavra *temporada* na sentença, um vetor de atenção é produzido matematicamente pela rede neural. Este vetor de atenção terá nove valores percentuais que representam o grau de atenção que deve ser prestado a cada uma das palavras da sentença para obter a representação da palavra *temporada*. O elemento em uma determinada posição no vetor corresponde à atenção para a palavra nessa posição na sentença. Por exemplo, um vetor de atenção [10%, 25%, 8%, 10%, 25%, 5%, 10%, 0%, 7%] indicaria que, para calcular uma representação vetorial contextual da palavra *temporada* na sentença sendo traduzida, as vetorizações para *verão* e *temporada* serão igualmente relevantes (em conjunto,

recebem cinquenta por cento da atenção total), o que faz sentido, uma vez que estão semanticamente ligados ao conceito de estação meteorológica. Observe que o determinante anterior também recebe alguma atenção (10%), o que pode ser explicado pelo fato de ajudar a rotular *temporada* como um substantivo. A contribuição do verbo (8%) para a incorporação contextual pode também ser descrita em termos da sua contribuição para a marcação do número de *estação* como singular. Note que as percentagens somam sempre 100%.

A determinação de como o vetor de atenção é utilizado para obter uma nova vetorização que combine os vetores não-contextuais originais para obter um novo vetor está fora do âmbito do presente capítulo. Basta dizer que o procedimento envolve uma sequência específica de operações matemáticas e que o vetor resultante estará em algum lugar entre os vetores originais.

No exemplo em análise, nove vetores de atenção distintos serão calculados para a sentença em questão (um para cada palavra) e, em seguida, aplicados aos vetores não-contextuais originais, a fim de obter uma coleção de nove novos vetores, cada um correspondendo a uma palavra diferente na sentença. Esses novos vetores podem ser considerados como vetores contextuais, pois são influenciados em diferentes níveis pelo resto das palavras da sentença.

5.1 Várias camadas de atenção são melhores que uma

Anteriormente, na seção 3.3 deste capítulo, foram discutidos os benefícios de refinar sucessivamente os cálculos neurais por meio da exploração de modelos com diferentes camadas. Consequentemente, não surpreende descobrir que, para obter representações mais precisas, os vetores contextuais recentemente obtidos possam ser combinados com novos vetores de atenção para obter mais um novo vetor para cada palavra. Como exemplo da vida real, o Turing Natural Language Generation (T-NLG), outro dos maiores modelos de linguagem publicados em 2020, tem 78 camadas de atenção que refinam sucessivamente vetores de 4.256 dimensões[8]. É importante lembrar de que essas representações, aprendidas pela aplicação de muitas camadas consecutivas, são conhecidas como representações *profundas*.

[8] "Turing-NLG: um modelo de linguagem de 17 bilhões de parâmetros da Microsoft", 2020. Disponível em: https://www.microsoft.com/en-us/research/blog/turing-nlg-a-17-billion-parameterlanguage-model-by-microsoft/

5.2 Muitas cabeças pensam melhor que uma

Não há motivos para nos restringirmos a um único vetor de atenção para cada palavra em cada camada. Por exemplo, dada a sentença "Meu avô assava pão em seu forno diariamente", pode ser interessante ter uma vetorização par *forno* que tenha o aspecto de *avô* para refletir que o forno em questão pertence a alguém mais velho, e uma vetorização diferente para forno com o aspecto de pão para refletir o que foi assado dentro dele. Um único vetor de atenção teria de misturar ambos os aspectos em uma única vetorização contendo informações demasiadamente heterogêneas que poderiam interferir negativamente na busca por uma tradução da palavra representada no vetor. Por esse motivo, alguns sistemas de NMT obtêm diferentes vetores de atenção para cada palavra em cada camada e depois os utilizam para calcular vários vetores diferentes para cada palavra. Cada um desses vetores foi calculado por uma *cabeça* diferente. T-NLG tem 28 cabeças de atenção em cada camada. Portanto, sua última camada produz 28 vetores de 4256 dimensões para cada palavra.

5.3 Vetores contextuais em processamento de língua natural

Vetores são os pilares da NMT, mas eles também se mostraram úteis em muitas outras aplicações de processamento de língua natural, como análise de sentimento e resumo automático. Para ilustrar, sistemas que classificam automaticamente como positivas ou negativas as sentenças da avaliação de um produto podem calcular primeiro uma coleção de vetores contextuais longos para cada palavra da sentença e depois usá-los para alimentar uma rede neural muito mais simples, que vai calcular um número entre 0 e 1 indicando o grau de positividade da sentença (por exemplo, 0,95 indica uma sentença definitivamente positiva; 0,2, uma sentença negativa e 0,51, uma sentença neutra). Esses sistemas normalmente são treinados com um corpus de classificadas manualmente por seres humanos. A parte do modelo que calcula os vetores não necessariamente é treinada com um corpus em particular, dado que modelos *pré-treinados* com milhões de sentenças já existem de modo disponível gratuitamente em muitas línguas.

6 Por fim, a tradução automática neural

Neste ponto, o leitor deste artigo deve ter condições suficientes para entender como NMT funciona, mesmo que os seus fundamentos sejam descritos em poucas sentenças, como será feito a seguir. Vamos nos concentrar em duas arquiteturas específicas: os assim chamados transformers e as redes neurais recorrentes.

6.1 Transformer: um par codificador-decodificador baseado em atenção

Em termos simplificados, um sistema transformer de NMT é composto por um módulo que calcula vetores contextuais para cada palavra na sentença de entrada e um segundo módulo que prevê cada palavra na sentença-alvo. O primeiro módulo recebe o nome de codificador e o segundo, de decodificador. Para prever as palavras da sentença-alvo, o decodificador presta atenção à vetorização de todas as palavras da sentença-origem, bem como à vetorização das palavras já produzidas na sentença-alvo. A arquitetura completa é chamada de transformer (Vaswani et al., 2017) A Figura 5 mostra um exemplo de um codificador de três camadas e os níveis de atenção considerados ao calcular as vetorizações nas segunda e terceira camadas. A Figura 6 mostra este codificador em um diagrama expandido que também inclui o decodificador, de modo a representar toda a arquitetura do transformer.

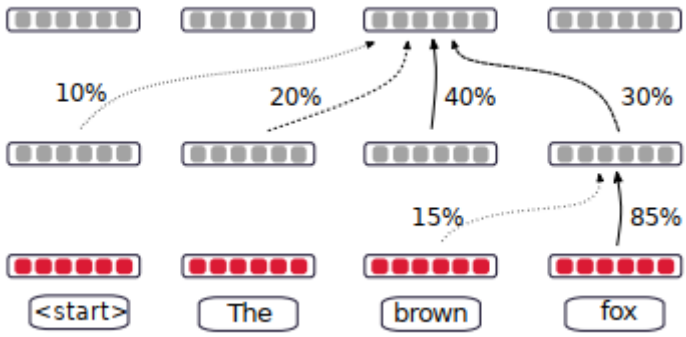


Figura 5: O codificador de um sistema de tradução automática neural de modelo transformer. O símbolo de partida geralmente é prefixado de modo explícito para marcar o começo da sentença. O diagrama também mostra que a vetorização da primeira camada para as palavras "brown" (marrom) e "fox" (raposa) dão contribuições de níveis diferentes para obter a vetorização de "fox" na segunda camada; analogamente, a vetorização de "brown" na última camada integra informações de todas as vetorizações da segunda camada usando diferentes níveis de atenção.

Um corpus paralelo é usado pelo algoritmo de aprendizagem para obter um conjunto de pesos, vetores lexicais e de atenção para o transformer, de tal forma que os dados de treinamento possam ser reproduzidos até um certo grau de acurácia e o sistema seja capaz de fazer generalizações para além das sentenças do conjunto de treinamento.

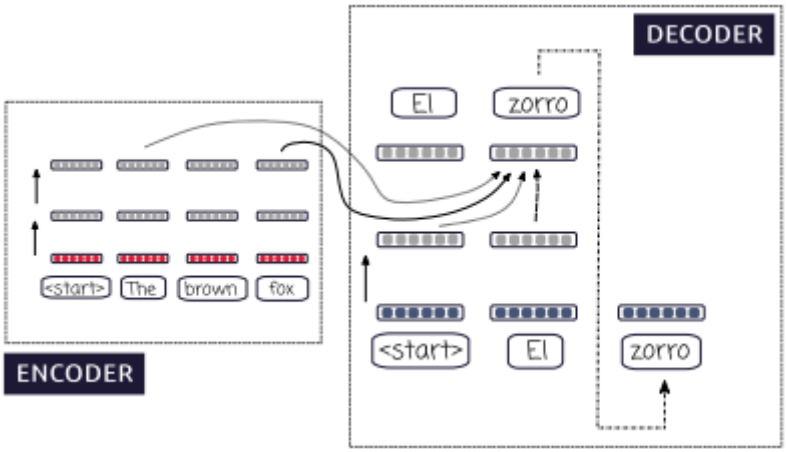


Figura 6: Um sistema completo de tradução automática neural, baseado no modelo transformer, em processo de tradução de uma sentença. Uma versão aumentada do codificador pode ser vista na Figura 5. Observe como a previsão "zorro" é obtida por meio da atenção prestada à vetorização das palavras-alvo anteriores, mas também à vetorização correspondendo a algumas das palavras de entrada que vêm da última camada do codificador.

Por exemplo, suponha que um transformer com uma única cabeça por camada seja utilizado para traduzir a sentença "Meu avô assava pão em seu forno diariamente" para o espanhol. O codificador primeiramente produz uma coleção de oito vetores lexicais. O decodificador então calcula um vetor de atenção de oito dimensões, como [60%, 10%, 0%, 0%, 0%, 30%, 0%, 0%] e o usa para obter um aspecto da sentença de origem que o permita obter uma vetorização da primeira palavra na sentença-alvo. Suponha que o sistema produza corretamente a palavra espanhola "mi". O decodificador então irá calcular um vetor de nove dimensões como [50%, 10%, 0%, 0%, 0%, 20%, 0%, 0%, 20%] (o último valor percentual corresponde ao grau de atenção prestado à primeira palavra na sentença-alvo) e usá-lo para obter uma vetorização da segunda palavra na sentença-

alvo. O procedimento vai continuar até o decodificador gerar um símbolo especial que marca o fim da sentença.

A saída do decodificador a cada iteração não é exatamente uma estimativa da vetorização da próxima palavra. Na verdade, uma camada adicional é sobreposta ao fim do decodificador para calcular um vetor de probabilidades ou verossimilhanças para cada palavra no vocabulário da língua de chegada. A seção 7.3 irá discutir como essas probabilidades podem ser usadas para se obter uma sequência de palavras que resultam na sentença da língua de chegada.

6.2 Arquiteturas recorrentes

O transformador, conforme apresentado na seção anterior, é o modelo adotado na maioria dos sistemas comerciais de NMT, embora existam modelos neurais alternativos. Outro modelo muito utilizado é o de codificador-decodificador (Bahdanau et al. 2015). De forma parecida com os modelos baseados em transformers, há um codificador que produz uma coleção de vetores para as palavras na sentença de entrada e um decodificador que usa atenção para calcular vetores de cada palavra-alvo integrando a informação das palavras de entrada e as palavras-alvo já produzidas. Entretanto, o codificador e o decodificador, no modelo recorrente, calculam os vetores contextuais de forma local de tal modo que os vetores da quinta palavra, por exemplo, sejam baseados nos vetores das quatro primeiras palavras, por um lado, e nos vetores das palavras seguintes, por outro. Esse procedimento é possível porque a sentença de entrada é lida tanto da esquerda para a direita quanto da direita para a esquerda; a Figura 7 traz um diagrama deste modelo mostrando apenas o processamento da esquerda para a direita.

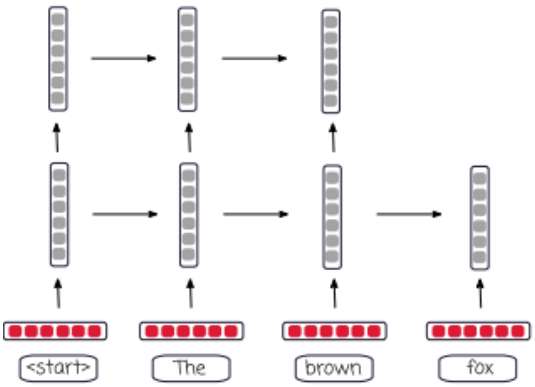


Figura 7: submodelo da esquerda para a direita do codificador de um sistema de tradução automática, logo após o processamento de "<início> A raposa" e logo antes de processar "marrom".

É relevante observar que o modelo matemático empregado impõe certas restrições ao grau de relevância atribuído às palavras circundantes àquela para a qual os vetores estão sendo calculados (no exemplo em questão, a quinta palavra), culminando em um mecanismo que confere um enfoque especial às palavras mais próximas e tende a menosprezar as representações das palavras mais distantes. De forma semelhante ao transformer, uma camada final, ao fim do decodificador, calcula um vetor que dá a probabilidade de que cada palavra traduzida ocupe a posição correspondente na sentença de resultado. Forcada (2017) descreve em maior detalhe o modelo recorrente de codificador-decodificador e também discute os tipos de saída que a NMT produz.

7 Parâmetros adicionais

7.1 Palavras e sub-palavras

De acordo com o que foi apresentado neste capítulo, independentemente de se usar um modelo de transformer ou recorrente, obtém-se uma vetorização para cada palavra após o treinamento. Isso significa que ao fim do processo é obtida uma vetorização correspondente a todas as palavras possíveis de uma língua? Na verdade, não. Línguas, especialmente aquelas que são altamente flexionais ou aglutinantes, podem facilmente atingir centenas de milhares ou até milhões de formas diferentes das palavras. Para entender como esse cenário pode ser desafiador para NMTs, tenha em mente que o número de vetorizações lexicais (que é chamado de vocabulário) condiciona o número de pesos na rede neural e que redes neurais grandes muitas vezes têm dificuldade para generalizar resultados para dados não vistos. O tamanho do vocabulário pode ser reduzido se considerarmos apenas as formas morfológicas presentes no corpus de treinamento, mas, normalmente, isso ainda implica utilizar um grande número de palavras e ainda provoca outro problema: quando o treinamento termina e o sistema NMT se propõe a traduzir novas sentenças contendo palavras fora do conjunto de treinamento, essas palavras não vistas vão fazer o modelo atuar de maneira desajeitada e perder acurácia, uma vez que cada palavra desconhecida vai receber um vetor não-contextual reservado para aquela situação.

A solução que os engenheiros bolaram foi de dividir palavras nas chamadas unidades sublexicais. Em um mundo ideal, essas palavras devem fazer sentido linguisticamente e ter componentes com significado: por exemplo, dividir a palavra "desmistificando" como des + misti + fic + ando certamente faz mais sentido linguístico do que dividi-la como desm + istif + ican + do. Mas fazer uma divisão lexical adequada linguisticamente requer a existência de um conjunto de regras de divisão e procedimentos para a língua em questão, o que é um recurso que, para muitas línguas, pode não estar disponível.

Uma forma comum de contornar o problema consiste em aprender automaticamente as regras de divisão sublexical ao inspecionar textos longos, como um que contenha todas as sentenças de origem ou todas as sentenças-alvo no conjunto de treinamento. Uma abordagem popular[9] é a chamada de codificação de pares de bytes (BPE, byte-pair encoding) (Sennrich et al., 2016), que começa com unidades do tamanho de uma letra que são unidas para formar unidades de duas letras (ou três etc) quando elas aparecem frequentemente no corpus.[10] A codificação por pares de bytes provavelmente identificaria um sufixo -ndo frequente em várias formas verbais (andando, pensando) e o cortaria, mesmo para formas verbais não vistas (como bartsimpsonando); o sufixo -ndo seria então transformado em um vetor contextual que traria consigo seu significado atômico.

7.2 Critérios de parada e métricas

Conforme mencionado na seção 3.5, além de um corpus de treinamento grande, um corpus de desenvolvimento pequeno normalmente é reservado e não é utilizado para treinamento. O propósito desse corpus é monitorar o desempenho de um sistema NMT enquanto ele está sendo treinado, para decidir quando o treinamento deve parar. O treinamento busca minimizar uma função de erro (ou, no caso da NMT, maximizar a probabilidade das sentenças-alvo no corpus de treinamento). Um possível problema que pode surgir é que treinar demais no corpus de

[9] Existem métodos mais avançados, como o SentencePiece (Kudo & Richardson 2018), que trata o texto todo como uma sequência de caracteres e realiza a divisão do texto em palavras (tokenização) e as divisões sublexicais em uma tacada só.

[10] A codificação de pares de bytes era originalmente um algoritmo de compressão: letras (bytes) frequentes são armazenados uma única vez e substituídos por códigos mais curtos para reduzir o armazenamento necessário.

treinamento prejudica as generalizações, pois a rede neural acaba memorizando excessivamente os exemplos de treinamento. É aí que entra o corpus de desenvolvimento: depois de um certo número de iterações (ou passos) do algoritmo de treinamento, as sentenças de origem do corpus de desenvolvimento são traduzidas com a rede neural e a saída é comparada automaticamente com as sentenças-alvo desejadas do corpus por meio de métricas simples de avaliação automática aproximada (vide Rossi & Carré 2022 [volume atual]), dentre as quais a mais comum é o BLEU (Papineni et al., 2002). O BLEU conta quantas sequências de uma, duas, três e quatro palavras na saída são encontradas na referência, e calcula uma nota que varia de 0 (nenhuma correspondência encontrada) até 100% (todas correspondências encontradas). Se, durante o treinamento, o BLEU encontrar uma deterioração do desempenho, o treinamento pode ser interrompido, ou o conjunto atual de pesos pode ser congelado e o treinamento continuar por algum tempo para ver se o BLEU aumenta novamente. Naturalmente, existem muitas outras métricas automáticas de avaliação que podem tomar o lugar do BLEU no processo.

7.3 Busca por feixe

O decodificador em sistemas NMT produz a sentença de saída sequencialmente, uma palavra-alvo por vez, como explicado nas seções 6.1 e 6.2. A cada iteração, a rede neural produz uma probabilidade ou verossimilhança (um valor entre 0 e 100%) a cada uma das palavras do vocabulário-alvo. Uma forma de usar essa informação é escolher a palavra-alvo mais provável e reproduzi-la, ignorando outras possibilidades. É interessante notar que, ao proceder desta maneira, os próximos passos dados pelo sistema NMT são completamente determinados, já que a previsão atual é usada para alimentar o decodificador do passo seguinte (vide, por exemplo, a palavra zorro, na Figura 6). Uma forma de explorar mais possibilidades é considerar, por exemplo, as três palavras mais prováveis e fazer três cópias do sistema, cada uma das quais determinada respectivamente pelas três palavras escolhidas, e avaliar seu desempenho. Entretanto, não é possível fazer isso indefinidamente, pois o número de sistemas traduzindo a sentença triplicaria a cada passo, aumentando exponencialmente. Para evitar isso, apenas um certo número de sistemas sobreviveria, a saber, aqueles que obtivessem a maior nota em um cálculo da probabilidade da sentença a ser produzida. Esse método normalmente recebe o nome de busca por feixe e é uma

aproximação comum em outros modelos probabilísticos de processamento de língua natural, como reconhecimento de fala.

8 Conclusões

Para treinar um sistema NMT, são necessários milhares ou até milhões de exemplos de pares de sentenças de origem e sentenças-alvo. No caso de muitos pares de línguas, muitos campos e muitos gêneros textuais, tais recursos não existem. Isso impõe restrições a muitas aplicações específicas, mas, para línguas abundantes em recursos, a NMT faz-tudo já é uma realidade e muito utilizada não apenas por tradutores. Além disso, avanços científicos em abordagens como modelos multilíngues ou NMTs não supervisionadas recentemente começaram a produzir resultados promissores em cenários de escassez de recursos.[11]

Este capítulo apresentou e forneceu detalhes técnicos dos elementos chaves de sistemas de NMT e explorou como eles interagem nas duas arquiteturas mais populares, a saber, as baseadas em transformers e as baseadas em redes neurais recorrentes. A pesquisa na área é tão intensa que, ainda no período de escrita deste artigo, propostas de novos modelos surgem quase todo mês. Atualmente, transformers constituem o paradigma escolhido no caso de haver um número suficiente de corpora paralelos disponíveis para treinamento, porque eles requerem menor tempo de treinamento e permitem melhoras sutis de qualidade em relação às redes neurais recorrentes, mas esse cenário pode mudar drasticamente a qualquer momento.

Referências

Bahdanau, Dzmitry, Kyunghyun Cho & Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In Yoshua Bengio & Yann LeCun (eds.), 3rd International Conference on Learning Representations, ICLR 2015. DOI: 10.48550/arXiv.1409.0473.

[11] Um modelo multilíngue é uma rede neural única que é treinada para traduzir entre vários pares de línguas diferentes, de modo que o conhecimento a respeito de línguas de corpora ricos possa ser transferido para línguas de corpora pobres. Modelos multilíngues acarretam a possibilidade da tradução de zero paralelismo (Ko et al., 2021), no qual um sistema pode traduzir com razoável qualidade, por exemplo, entre espanhol e alto sorábio usando modelos treinados com corpora dos pares alemão-alto sorábio e espanhol-alemão, mesmo sem um corpus paralelo de espanhol e alto-sorábio. NMTs não supervisionadas vão um passo além, aprendendo apenas com sistemas de NMT de corpora monolinguais.

Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Dario Amodei, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever & Dario Amodei. 2020. Language models are few-shot learners. CoRR abs/2005.14165. https://arxiv.org/ abs/2005.14165.

Forcada, Mikel. 2017. Making sense of neural machine translation. Translation Spaces 6(2). 291–309.

Goodfellow, Ian, Yoshua Bengio & Aaron Courville. 2016. Deep learning. Cambridge, MA: MIT Press.

Hornik, Kurt. 1991. Approximation capabilities of multilayer feedforward networks. Neural Networks 4(2). 251–257.

Ko, Wei-Jen, Ahmed El-Kishky, Adithya Renduchintala, Vishrav Chaudhary, Naman Goyal, Francisco Guzmán, Pascale Fung, Philipp Koehn & Mona Diab. 2021. Adapting high-resource NMT models to translate low-resource related languages without parallel data. In Proceedings of the 59th annual meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, 802–812.

Kudo, Taku & John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In

Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 66–71. Brussels, Belgium: Association for Computational Linguistics.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg Corrado & Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In Advances in Neural Information Processing Systems 30, 3111–3119.

Papineni, Kishore, Salim Roukos, Todd Ward & Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, 311–318. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics. DOI: 10.3115/1073083.1073135.

Rossi, Caroline & Alice Carré. 2022. How to choose a suitable neural machine translation solution: Evaluation of MT quality. In Dorothy Kenny (ed.), Machine translation for everyone: Empowering users in the age of artificial intelligence, 51–79. Berlin: Language Science Press. DOI: 10.5281/zenodo.6759978.

Sennrich, Rico, Barry Haddow & Alexandra Birch. 2016. Neural Machine translation of rare words with subword units. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, 1715–1725. Berlin: Association for Computational Linguistics.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser & Illia Polosukhin. 2017. Attention is all you need. In Advances in Neural Information Processing Systems 30, 5998–6008.