Olhares sobre a tradução automática: uma análise sobre o desempenho dos sistemas DeepL e Google Tradutor

Views on machine translation: An analysis on the performance of DeepL and Google Translate

Renata Ribeiro da Silva*

Thiago Blanch Pires**

* Bacharel em Línguas Estrangeiras Aplicadas ao Multilinguismo e à Sociedade da Informação (LEA-MSI) pela Universidade de Brasília (UnB). Email: renatarbrsilva@gmail.com.

** Professor adjunto vinculado ao Departamento de Línguas Estrangeiras e Tradução (LET) e ao Programa de Pós-Graduação em Linguística (PPGL) da Universidade de Brasília (UnB). Doutor em Ciência da Informação pela UnB e mestre em Letras-Inglês pela UFSC. E-mail: pirestb@unb.br.

Recebido em: 14-jun-2024
Aceito em: 23-jul.-2024

Resumo: Este trabalho avalia comparativamente o desempenho de dois sistemas de tradução automática, a saber, DeepL e Google Tradutor. O estudo parte de uma breve visão geral sobre traduções automáticas com a utilização de redes neurais, seguido de uma reflexão sobre os resultados da tradução automática ao utilizar a avaliação automática e humana. Subsequentemente, aborda a diferença de desempenho entre ambos os sistemas, além de identificar possíveis problemas linguísticos advindos dos resultados da tradução automática gerada por eles. Para a realização dessa comparação, o trabalho emprega um trecho da obra de Machado de Assis, Dom Casmurro, do capítulo "Olhos de Ressaca". O estudo utiliza as traduções geradas de forma automática para analisar criticamente qual sistema apresenta mais qualidade no resultado da tradução e examina comparativamente os erros linguísticos gerados pelos resultados do DeepL e Google Tradutor. Os critérios de avaliação incluem a precisão da tradução, a capacidade de manter o significado e a estrutura da frase original, a fluência e adequação da tradução resultante.

Palavras-chave: Tipologia de erros de resultados de tradução automática. Tradução automática neural. Avaliação de tradução automática.

Abstract: *This paper comparatively evaluates the performance of two machine translation systems, namely DeepL and Google Translator. It starts with a brief overview of machine translation using neural networks, followed by a reflection concerning the performance of machine translation output using human and automatic evaluation. After that, this study addresses the difference in performance between both systems and identifies possible linguistic problems, possibly generated from these machine translation output. To carry on such comparison, this study employs an excerpt from Machado de Assis's Dom Casmurro, taken form the chapter "Olhos de Ressaca". This investigation uses the automatically generated translation outputs to critically analyze which system conveys more quality in their translation output and to comparatively examine linguistic errors generated by DeepL e Google Translate results. The evaluation criteria include the accuracy of the translation, the ability to maintain the meaning and structure of the original sentence, as well as fluency and adequacy of the resulting translation.*

Palavras-chave: Machine translation output error typology. Neural machine translation. Machine translation evaluation.

Introdução

Em 2016, pesquisadores do Google Tradutor passaram a utilizar o que a empresa Google denominou de Google Neural Machine Translation (GNMT), trocando o método de estatística

que vinha sendo usado desde 2017 para aumentar a fluência e capacidade de suas traduções automáticas. No Google, a aprendizagem do algoritmo é realizada por meio do método em que se aprende por milhões de exemplos, ampliando o contexto e deduzindo a tradução com maior relevância (Schuster et al., 2016).

Por outro lado, de acordo com o DeepL Press Information, o DeepL, lançado em 2017, utiliza-se de uma geração nova de redes neurais e inteligência artificial, que aprende a captar os significados sutis das frases com o banco de dados Linguee. Diz-se que a tradução é gerada a partir do uso de um supercomputador que atinge 5,1 petaflops (medidas imensas de velocidade de processamento) e é operado na Islândia com energia hidrelétrica.

É nesse contexto em que o presente estudo se circunscreve. Utiliza-se das contribuições de Brita Banitz (2020) como principal aporte teórico-metodológico para observar a existência de diferenças entre sistemas de tradução automática. Os procedimentos que se busca empregar na presente investigação para adquirir os resultados na coleta de dados advém dos métodos Bilingual Evaluation Understudy Score (BLEU), Translation Error Rate (TER), e a avaliação humana, tendo como objeto a tradução (português-inglês) de um trecho escolhido do capítulo 32 “Olhos de Ressaca” da obra Dom Casmurro, de Machado de Assis.

Levando em consideração esses aspectos e propondo pesquisar sobre as áreas de Tradução e Tecnologia, as perguntas de pesquisa formuladas são as seguintes:

• Ao considerar os sistemas DeepL e Google Tradutor, podemos dizer, após análises comparativas, em termos de um trecho do capítulo “Olhos de Ressaca” da obra Dom Casmurro, de Machado de Assis, do português para inglês, que um dos sistemas apresenta resultados mais precisos em relação aos dados analisados?

• Quais métodos seriam eficazes na identificação dos erros linguísticos em resultados de comparação entre os sistemas DeepL e Google Tradutor?

O presente estudo justifica-se ao propor uma inovação no campo da avaliação de tradução automática. Trata-se de uma tentativa de identificar questões linguísticas inéditas sob a ótica de um sistema de tradução já amplamente conhecido como o Google e outro que ainda está ganhando espaço no mercado, o DeepL. A diferença de desempenho entre os dois é analisada sob um novo ângulo. Tem como objetivo proporcionar uma nova abordagem ao tema a partir do estudo dos aspectos linguísticos da tradução, e determinar se existem diferenças entre o trecho original em português e o trecho traduzido automaticamente para o inglês.

Assim, este estudo não visa resolver o problema da tradução automática, mas se propõe a analisar se o resultado de tradução de um desses sistemas supera o outro no que diz respeito a potenciais erros linguísticos. Assim, os achados deste estudo final contribuem para a interface da Linguística Computacional e Estudos da Tradução.

Fundamentação teórica

Segundo Tan et al. (2020), a Tradução Automática (TA), em inglês Machine Translation, é um subcampo do processamento da linguagem natural, vertente da inteligência artificial que realiza o auxílio das máquinas/computadores a compreenderem a linguagem humana. Ainda seguindo o raciocínio da obra "Neural machine translation: a review of methods, resources, and tools" (Tan et al., 2020), a tradução automática se destina a traduzir frases em linguagem natural com o auxílio de computadores, sendo que os primeiros passos da tradução automática dependiam de regras de tradução manual e habilidades linguísticas. Devido à complexidade da linguagem natural, as regras de tradução manual eram limitadas e não conseguiam alcançar todas as irregularidades de linguagem. Por esta razão, com a disponibilidade de corpora (grandes coletâneas de textos em formato eletrônico), métodos baseados em dados (que aprendem informações linguísticas por meio de dados para reter e aprender informações de linguagem) estão cada vez mais sendo utilizados (Tan et al., 2020).

Redes neurais

Com os avanços contínuos da tecnologia, muitos sistemas de tradução automática estão aplicando conhecimentos e técnicas do campo da inteligência artificial e redes neurais, como, por exemplo, o Google Tradutor e o DeepL. As traduções automáticas que se utilizam de técnicas neurais conseguem alcançar resultados cada vez mais próximos de traduções realizadas por humanos, visto que o algoritmo utilizado aprende a cada consulta (Tan et al., 2020).

Porém, a melhora das traduções automáticas baseadas em redes neurais também faz com que inevitavelmente as métricas de avaliação automática percam sua eficácia (Isabelle et al., 2017). As avaliações automatizadas conseguem mostrar apenas percentagens da diferença das traduções automáticas ou editadas. Já nas avaliações humanas, é possível ter mais detalhes e especificidades da qualidade de tais resultados de tradução (Densmer, 2019).

Avaliação humana de TA

Segundo Banitz (2020), a avaliação humana da tradução automática é capaz de fornecer, de forma detalhada, uma análise da qualidade da tradução. O tradutor tem a capacidade de escolher a tradução com base em traduções de referência (traduções realizadas totalmente por um especialista ou pós-editadas por eles a partir de resultados de traduções automáticas). Neste trabalho, para realizar a avaliação humana, usaremos as tabelas de fluência e adequação de Callison-Burch et al. (2006), propostas por Banitz (2020).

Somos apontados para diversas problemáticas referente às avaliações das traduções automáticas, sejam automatizadas ou humanas. Isto se dá pelo fato de que é possível encontrar “falhas” em todo e qualquer avaliador, e a avaliação pode ser considerada subjetiva, por depender do julgamento e critério de cada tradutor (Banitz, 2020). Por esta razão, estudiosos passaram a desenvolver métodos-padrões em que a própria máquina avalia as traduções, porém, esses também acabam tendo suas próprias falhas.

Ainda que o foco do presente trabalho seja o processo tradutório da máquina e aquele do humano sejam distintos, a

comparação entre os resultados de cada um pode sugerir padrões que podem servir tanto como discernimento das escolhas tradutórias, dentro do contexto de treinamento de tradutores, quanto de melhoria dos sistemas de TA.

Avaliação automática de TA

Pode-se entender com os comentários de Banitz (2020) mencionando autores como Forcada (2010), que a avaliação automática de TA se tornou um procedimento padrão a ser seguido, visto que é mais ágil, rentável e objetivo, garantindo um elevado número de resultados com resposta imediata.

Tomando como exemplo o método BLEU, apesar de ser um dos mais conhecidos e tradicionais modelos de avaliação, é limitado a uma considerável sequência de palavras (Banitz, 2020). Ao utilizar este método com sistemas de tradução baseados em redes neurais, que estão mais próximos do desempenho humano do que sistemas baseados em estatística, não é possível notar diferenças mais sutis das traduções (Isabelle et al., 2017), porém, ainda pode ser considerado um bom parâmetro de utilização.

Conforme Forcada (2010), é necessário observar quando e o que esperar das traduções automáticas, para realmente poder usá-las de forma eficaz. Isto se dá pelo fato de que as traduções brutas, isto é, sem edições ou correções produzidas por sistemas de tradução automática se diferenciam das traduções produzidas por tradutores humanos profissionais. Essa diferença não significa que as traduções automáticas não sejam úteis, mas que ao depender da finalidade para qual será utilizada, pode ser necessário utilizar algum método para avaliá-la.

Precisamos sempre lembrar que as avaliações de traduções automáticas precisam de um olhar atento e uma boa análise, de preferência com acompanhamento de um tradutor experiente, pois poderá ser um processo complexo. Aprofundando, de acordo com Kalyani et al. (2014, p. 54, apud Banitz, 2020, p. 58), "[não existe] nenhum padrão ouro [desempenho equivalente a um tradutor humano] no qual uma tradução poderia ser avaliada", ou seja, podemos presumir que se faz necessário um tradutor profissional que seja hábil para manusear e criar traduções de referência que sejam utilizadas conjuntamente com alguns métodos de avaliação automática, tais como o BLEU e o TER.

BLEU (Bilingual Evaluation Understudy Score) é um método de avaliação automática, com pontuação de avaliação bilíngue e uma das técnicas mais utilizadas nesse ramo, servindo como inspiração e influência para a criação de novas métricas. Essa métrica funciona através da comparação de traduções já existentes, sendo uma delas uma tradução de referência (por vezes uma tradução totalmente humana ou de pós-edição), e a tradução produzida pelo sistema de TA (Densmer, 2019).

A pontuação foi desenvolvida para avaliar as previsões feitas por sistemas de tradução automática, sendo a pontuação 1 para a tradução com correspondência perfeita, idêntica à tradução de referência utilizada, e a pontuação 0 para a tradução sem correspondência alguma. Estes critérios são realizados através da programação, em que se compara n-gramas da tradução produzida pela TA, e n-gramas da tradução de referência, contando os números que ambas possuem correspondentes. Nessa lógica considera-se como um n-grama cada token, normalmente uma palavra, e cada bigrama, um par

de palavras. Quanto mais combinações forem feitas, melhor será o resultado da tradução automática (Brownlee, 2019).

A pontuação Translation Error Rate, abreviada como TER, é a taxa de erros de tradução que faz a medição do número de edições que seriam necessárias para alcançar uma tradução realizada e/ou revisada por humanos especialistas (“padrão ouro”). Apesar de dar indícios do quão próximo a TA está da tradução humana, a pontuação TER não necessariamente mostra a adequação ou aceitação da tradução automática, e ela depende da qualidade da tradução de referência (Banitz, 2020).

Metodologia

A presente investigação busca realizar uma análise comparativa e de desempenho das traduções automáticas dos sistemas DeepL e Google Tradutor, além de utilizar uma pesquisa bibliográfica como base, conforme revisado na seção anterior.

A forma para analisar comparativamente ambos os sistemas de tradução foi por intermédio de um texto curto traduzido pelos dois sistemas de tradução. Neste estudo, consideramos a utilização dos sistemas DeepL e Google Tradutor entre as datas de novembro e dezembro de 2022.

Utiliza-se um trecho escolhido do capítulo “Olhos de Ressaca” da obra Dom Casmurro, de Machado de Assis, subsequentemente são realizadas as traduções português-inglês nos sistemas de tradução automática DeepL e Google Tradutor, além da pós-edição de cada uma das traduções para criar traduções de referência. O trecho de Dom Casmurro (Assis, 2019, p.51-52) escolhido está representado pelo Quadro 1 a seguir.

Quadro 1: Trecho do capítulo “Olhos de Ressaca” da obra Dom Camurro

Texto Original
“Tinham-me lembrado a definição que José Dias dera deles, ‘olhos de cigana oblíqua e dissimulada’. Eu não sabia o que era oblíqua, mas dissimulada sabia, e queria ver se podiam chamar assim. Capitu deixou-se fitar e examinar. Só me perguntava o que era, se nunca os vira; eu nada achei extraordinário; a cor e a doçura eram minhas conhecidas. A demora da contemplação creio que lhe deu outra ideia do meu intento; imaginou que era um pretexto para mirá-los mais de perto, com os meus olhos longos, constantes, enfiados neles, e a isto atribuo que entrassem a ficar crescidos, crescidos e sombrios, com tal expressão que... Retórica dos namorados, dá-me uma comparação exata e poética para dizer o que foram aqueles olhos de Capitu. Não me acode imagem capaz de dizer, sem quebra da dignidade do estilo, o que eles foram e me fizeram. Olhos de ressaca? Vá, de ressaca. É o que me dá ideia daquela feição nova. Traziam não sei que fluido misterioso e enérgico, uma força que arrastava para dentro, como a vaga que se retira da praia, nos dias de ressaca. Para não ser arrastado, agarrei-me às outras partes vizinhas, às orelhas, aos braços, aos cabelos espalhados pelos ombros; mas tão depressa buscava as pupilas, a onda que saía delas vinha crescendo, cava e escura, ameaçando envolver-me, puxar-me e tragar-me. Quantos minutos gastamos naquele jogo? Só os relógios do Céu terão marcado esse tempo infinito e breve. A eternidade tem as suas pêndulas; nem por não acabar nunca deixa de querer saber a duração das felicidades e dos suplícios. Há de dobrar o gozo aos bem-aventurados do Céu conhecer a soma dos tormentos que já terão padecido no inferno os seus inimigos; assim também a quantidade das delícias que terão gozado no Céu os seus desafetos aumentará as dores aos condenados do inferno. Este outro suplício escapou ao divino Dante; mas eu não estou aqui para emendar poetas. Estou para contar que, ao cabo de um tempo não marcado, agarrei-me definitivamente aos cabelos de Capitu, mas então com as mãos, e disse-lhe, — para dizer alguma coisa, — que era capaz de os pentear, se quisesse.”

Fonte: Assis, 2019

A escolha desse trecho justifica-se pelo teor de criatividade da escrita literária, o que poderia trazer certa variação nos resultados dos sistemas de tradução automática em questão. Além disso, leva-se em consideração o fato de a obra de Machado de Assis estar em domínio público, portanto, sem necessidade de autorização dos detentores dos direitos autorais. Por isso, é um facilitador na divulgação dos achados deste estudo.

O mesmo trecho foi traduzido automaticamente para o inglês utilizando os sistemas de DeepL e Google Tradutor, respectivamente representados pelo Quadro 2 e Quadro 3 a seguir.

Quadro 2: Tradução do DeepL para o inglês

Tradução DeepL
“They had reminded me of Jose Dias definition of them, ‘the eyes of an oblique and disguised gypsy’. I didn’t know what oblique was, but dissimulated I did, and I wanted to see if they could be called that. Capitu let herself stare and examine. She only wondered what it was, if she had never seen them; I found nothing extraordinary; the color and sweetness were familiar to me. The delay in her contemplation, I believe, gave her another idea of my intentions; she imagined it was a pretext to look at them more closely, with my long, constant eyes stuck in them, and to this I attribute that they began to grow bigger and bigger and darker, with such an expression that... Lovers’ rhetoric, give me an exact and poetic comparison to say what those eyes of Capitu’s were. I have no image that can tell, without breaking the dignity of style, what they were and what they did to me. Hangover eyes? Come on, hungover. That’s what gives me an idea of that new feature. They brought I don’t know what mysterious and energetic fluid, a force that dragged me in, like the wave that retreats from the beach on hangover days. To avoid being dragged away, I clung to the other neighboring parts, to the ears, to the arms, to the hair spread across the shoulders; but as soon as I reached for the pupils, the wave that came out of them grew bigger, hollow and dark, threatening to envelop me, to pull me in and swallow me. How many minutes did we spend in that game? Only the clocks in heaven will have marked that infinite and brief time. Eternity has its pendulums; it never stops wanting to know the duration of happiness and suffering. It will double the joy of the blessed in Heaven to know the sum of the torments that their enemies will have already suffered in Hell; and the amount of the delights that their enemies will have enjoyed in Heaven will increase the pains of the damned in Hell. This other torment escaped the divine Dante; but I am not here to amend poets. I’m here to tell you that, after an unmarked time, I definitely grabbed Capitu’s hair, but then with my hands, and told her, - to say something, - that I was capable of combing it, if I wanted to.”

Fonte: Tradução automática do DeepL.

Quadro 3: Tradução do Google Tradutor para o inglês

Tradução Google Tradutor
“They had reminded me of the definition that José Dias had given them, ‘oblique and concealed gypsy eyes’. I didn’t know what oblique was, but covertly I did, and I wanted to see if they could call it that. Capitu allowed himself to be stared at and examined. He only asked me what it was, if he had never seen them; I found nothing extraordinary; the color and sweetness were my acquaintances. The delay of contemplation I think gave him another idea of my intention; he imagined that it was a pretext to look at them more closely, with my long, steady eyes fixed on them, and to this I attribute that they began to grow larger, larger and somber, with such an expression that... give me an exact and poetic comparison to say what those eyes of Capitu were. I can’t think of an image capable of saying, without breaking the dignity of the style, what they were and what they did to me. Hangover eyes? Come on, hangover. It’s what gives me the idea of that new feature. They brought some mysterious and energetic fluid, a force that dragged them inwards, like the wave that leaves the beach, on days of undertow. In order not to be dragged away, I clung to the other neighboring parts, to the ears, to the arms, to the hair spread over the shoulders; but as quickly as I sought the pupils, the wave that came out of them was growing, hollow and dark, threatening to envelop me, pull me in and swallow me. How many minutes did we spend on that game? Only the clocks of Heaven will have marked this infinite and brief time. Eternity has its pendulums; not even because it never ends, he never ceases to want to know how long the happiness and torments will last. It will double the joy of the blessed of Heaven to know the sum of the torments that their enemies will have already suffered in hell; so, also the amount of delights that his enemies will have enjoyed in Heaven will increase the pains of the damned in hell. This other torment escaped the divine Dante; but I’m not here to amend poets. I’m about to tell you that, at the end of an unmarked time, I grabbed Capitu’s hair definitively, but then with my hands, and told her, — to say something — that I was capable of combing them if I wanted to.”

Fonte: Tradução automática do Google Tradutor.

Para realizar as análises das avaliações das traduções, três diferentes métodos foram escolhidos, a saber: i) BLEU, ii) TER e iii) a avaliação humana realizada por meio da tabela de fluência e adequação. Para utilizar os dois métodos de avaliação automática foi necessário criar duas traduções de referência, sendo uma para a tradução do Google Tradutor (Quadro 4 a seguir) e outra para a tradução do DeepL (Quadro 5 subsequente).

Quadro 4: Tradução de referência a partir de resultado do Google Tradutor

Tradução de referência – via Google Tradutor
“They had reminded me of the definition that José Dias had given them, ‘oblique and disguised gypsy eyes’. I didn’t know what oblique was, but disguised I did, and I wanted to see if they could call it that. Capitu allowed herself to be stared at and examined. I only asked myself what it was; if I had never seen them. I found nothing extraordinary; the color and sweetness were my acquaintances. The delay of contemplation, I think, gave her another idea of my intention; she imagined that it was a pretext to look at them more closely, with my long, steady eyes fixed on them. And, to this, I attribute that they began to grow larger and darker, with such an expression that gives me an exact and poetic comparison to say what those eyes of Capitu were. I can’t think of an image capable of saying, without breaking the dignity of the style, what they were and what they did to me. Undertow eyes? Come on, undertow. It’s what gives me the idea of that new face. They brought some mysterious and energetic fluid, a force that dragged them inwards, like the wave that retreats from the shore. In order not to be dragged away, I clung to the other neighboring parts, to the ears, to the arms, to the hair spread over the shoulders; but as quickly as I sought the pupils, the wave that came out of them was growing, hollow and dark, threatening to envelop me, pull me in and swallow me. How many minutes did we spend on that game? Only the clocks of Heaven will have marked this infinite and brief time. Eternity has its pendulums; not even because it never ends, he never ceases to know how long the happiness and torments will last. It will double the joy of the blessed of Heaven if they know the sum of torments that their enemies will have already suffered in hell. Also, the number of delights his enemies will have enjoyed in Heaven will increase the pains of the damned in hell. This other torment escaped the divine Dante, but I’m not here to amend poets. I’m about to tell you that, at the end of an unmarked time, I grabbed Capitu’s hair definitively, but then with my hands, and told her — to say something — that I was capable of combing them if I wanted to.”

Fonte: Tradução de referência a partir do Google Tradutor, realizada pela autora.

Quadro 5: Tradução de referência a partir de resultado do DeepL

Tradução de referência – via DeepL
“They had reminded me of Jose Dias definition of them, ‘oblique and dissimulated gypsy eyes’. I didn’t know what oblique was but dissimulated I did, and I wanted to see if they could be called that. Capitu let herself be stared at and examined. I only wondered what it was; if I had never seen them. I found nothing extraordinary; the color and sweetness were familiar to me. The delay in her contemplation, I believe, gave her another idea of my intentions; she imagined it was a pretext to look at them more closely, with my long, constant eyes stuck in them. And, to this, I attribute that they began to grow bigger and darker, with such an expression that... Lovers’ rhetoric gives me an exact and poetic comparison to say what those eyes of Capitu’s were. I have no image that can tell without breaking the dignity of style what they were and what they did to me. Undertow eyes? Come on, undertow. That’s what gives me an idea of that new face. They brought a mysterious and energetic fluid, a force that dragged me in like the wave that retreats from the beach. To avoid being dragged away, I clung to the other neighboring parts, to the ears, to the arms, to the hair spread across the shoulders; but as soon as I reached for the pupils, the wave that came out of them grew bigger, hollow and dark, threatening to envelop me, pull me in and swallow me. How many minutes did we spend in that game? Only the clocks in heaven will have marked that infinite and brief time. Eternity has pendulums; it never stops wanting to know the duration of happiness and suffering. It will double the joy of the blessed in Heaven to know the sum of torments their enemies will have already suffered in Hell. And the amount of delights their enemies will have enjoyed in Heaven will increase the pains of the damned in Hell. This other torment escaped the divine Dante, but I am not here to amend poets. I’m here to tell you that, after an unmarked time, I definitely grabbed Capitu’s hair, but then with my hands, and told her - to say something - that I was capable of combing it if I wanted to.”

Fonte: Tradução de referência a partir do DeepL, realizada pela autora.

Para a criação das traduções de referência, o presente estudo conduziu uma edição em cada uma das traduções, utilizando os parâmetros de avaliação humana escolhidos, buscando adequar cada referência a cada tradução automática para que a avaliação pudesse aferir o desempenho de ambos os sistemas.

A pontuação BLEU foi escolhida por se tratar de um dos grandes e renomados métodos no mundo da tradução automática ao se tratar de avaliações de TA, sendo até mesmo considerada como um padrão de avaliação. Assim, sua utilização foi essencial para este trabalho, avaliando não somente o trecho em si traduzido automaticamente, mas também mostrando a pontuação dos dois sistemas de TA baseados em redes neurais. Avaliou-se, com a pontuação BLEU, palavra por palavra da tradução. Para acrescentar à avaliação BLEU, escolheu-se o método TER, para dar uma medição precisa de quantas alterações seriam necessárias para se alcançar a tradução de referência.

E, por fim, as escalas de Fluência e de Adequação propostas por Callison-Burch et al. (2006) foram escolhidas como avaliação humana, para agregar ao estudo um parâmetro humano acerca das traduções realizadas pelos sistemas. Nessas escalas temos não somente a questão da fluência de cada sentença, mas também o quanto do significado expresso da tradução de referência é também expresso na tradução literal. Assim, os três métodos que foram escolhidos se utilizam de traduções de referência para a realização das análises.

Análise

Esta seção, dando importância a todas as informações adquiridas desde a seção introdutória, analisa as traduções automáticas dos tradutores DeepL e Google Tradutor. À vista disso, encontra-se neste tópico os resultados do estudo: o diferencial entre os sistemas de TA considerando que ambos se baseiam na utilização de redes neurais.

Análise dos resultados avaliados pelo BLEU

O resultado obtido das avaliações do BLEU foi de dois decimais iguais, ou seja, ambos sistemas tiraram a mesma nota de 0,82. Porém, a terceira casa decimal mostrou que, apesar dos dois sistemas estarem igualados em questões de pontuação, o Google Tradutor se sobressaiu minimamente ao DeepL, conforme vemos na Tabela 1 a seguir.

Tabela 1: Comparação das pontuações BLEU dos sistemas de tradução automática Google Tradutor e DeepL

Sistema de Tradução Automática	**BLEU**
Google Tradutor	0,82876
DeepL	0,82576

Fonte: Autores.

A pontuação BLEU varia de 0 a 1. Em poucos casos a pontuação chega a 1, a menos que a tradução seja idêntica à tradução de referência, ou seja, a tradução editada pelos autores e considerada neste estudo como padrão ou “ouro”. Assim, quanto mais traduções de referências houver, maior será a pontuação BLEU, pois a tradução terá ainda mais chances de ter correspondentes. Por esta razão, utilizou-se apenas uma tradução de referência para cada um dos tradutores, tornando a comparação justa em termos de adequação.

Mesmo com as duas pontuações tendo o número tão próximo, ainda assim foi possível reparar que o Google Tradutor teve um melhor desempenho e foi mais preciso por 0,003. Podemos identificar que esse fato ocorreu, pois a pontuação BLEU se baseia na contagem de n-gramas em comum entre a

tradução-alvo e a tradução de referência, mostrando que o Google conseguiu se aproximar mais da tradução de referência do que o DeepL. Portanto, por mais que não seja tão nítido a diferença de ambos os sistemas de tradução automática entre um e outro, podemos observar a similitude entre ambos. As TAs que utilizam redes neurais se saíram bem neste caso em específico, com a pontuação de 0,82, quase se aproximando à pontuação 1.

Análise dos resultados avaliados pelo TER

Na análise TER, é utilizada a tradução de referência em comparação com a tradução-alvo, indicando quantas exclusões, alterações e substituições seriam necessárias para se alcançar a tradução de referência. A tradução do Google Tradutor resultou numa pontuação média TER de 6,29, enquanto a pontuação média TER do DeepL foi de 5,17, indicando que a tradução do DeepL se sobressaiu na pontuação, precisando realizar menos edições, alterações ou exclusões para alcançar a tradução de referência. Este método de avaliação nos indica que, no geral, o DeepL necessita de menos edições para se aproximar à referência. Ao contrário, referente à tradução do Google Tradutor, das 17 sentenças alvo, somente as sentenças 3, 6, 10, 11 e 17 alcançaram uma pontuação TER inferior.

Ao considerar as pontuações TER inferiores do Google Tradutor, vê-se como exemplo na sentença 3, referente à tradução do Google Tradutor, trocou-se apenas o pronome “himself” para “herself”, ao se tratar da personagem Capitu. Em contrapartida, na tradução do DeepL, o sentido completo da frase teve de ser alterado, tendo como tradução de referência a ideia central de que Capitu deixou que o outro personagem da obra a admirasse,

e não ela como a admiradora. Uma das razões da tradução Google ter alcançado uma pontuação mais elevada, se refere ao fato de que a tradução alvo colocou o gênero da personagem Capitu no masculino, que foi alterado para o feminino em todas as sentenças na tradução de referência, causando um aumento na quantidade de edições necessárias.

Análise dos resultados avaliados pela avaliação humana

Para realizar a avaliação humana das traduções automáticas, as escalas de fluência e de adequação propostas por Callison-Burch et al. (2006) foram utilizadas. Nessa análise, o presente trabalho avaliou a tradução humana por meio da fluência, dando uma pontuação de 1 a 5, sendo de incompreensível à um inglês impecável respectivamente, e por meio da adequação, analisando quanto do sentido que está na tradução de referência também está na tradução de cada um dos sistemas.

Ao considerar a pontuação de adequação, nota-se que o DeepL foi melhor classificado, ficando com 4,29 de média, em comparação com a média do Google Tradutor de 4,17. Dessa forma, a tradução que mais alcançou a adequação conforme a pergunta “Quanto do significado expresso na tradução de referência é também expresso na tradução literal?”, foi a do DeepL, sendo uma representação mais próxima do texto em português.

Nota-se que é dada a adequação “razoável” em 4 sentenças do Google Tradutor, e a mesma em somente 2 sentenças do DeepL. A pontuação “majoritariamente”, que trata do significado ser o mesmo da tradução de referência, foi dado em 6 sentenças

do Google e em 8 do DeepL. Por último, a pontuação de “100%” estava presente tanto em 7 sentenças do Google Tradutor quanto também em sentenças do DeepL. Portanto, após os dados apresentados, é possível inferir que o fator decisivo na pontuação de adequação foi as sentenças consideradas com aproximação razoável e majoritária, visto que a pontuação 100% ocorreu em 7 sentenças igualmente em cada um dos dois sistemas.

Em relação à pontuação de fluência, o Google Tradutor ficou com a pontuação média de 3,94, enquanto o DeepL apresentou uma pontuação média 4. Nota-se como exemplo que, na tradução do Google Tradutor, os pronomes de Capitu são trocados do feminino para o masculino no decorrer do trecho, mudando significativamente o contexto da obra literária e também a caracterização da personagem.

É importante salientar que entre os resultados de ambos os sistemas, o adjetivo “ressaca”, notadamente atribuído pelo narrador ao olhar de Capitu, apresentaram a versão “hangover”, ao invés de “undertow” ou “whirlpool". Assim, embora esse resultado apresente uma das possibilidades de correspondente léxico-semântico, a escolha do mesmo pelos sistemas reconfigura a caracterização da personagem dentro do contexto da obra na versão em inglês.

Considerando a tradução humana realizada pelo método escolhido, a tradução automática do DeepL recebeu mais pontuações tanto de fluência quanto de adequação em relação à tradução automática do Google Tradutor, com a diferença de 0,06 na fluência, e de 0,12 na adequação. A diferença de pontuação entre os sistemas de TA não foi discrepante, porém, mostra que

mesmo ambos os sistemas usem redes neurais, cada um possui um algoritmo e uma forma de lidar com as traduções, podendo gerar sentenças com estruturas mais bem definidas levando em consideração o contexto original.

Considerações Finais

Perante o exposto, este trabalho apresentou um panorama geral sobre a tradução automática e redes neurais, além de comentar brevemente sobre os métodos de avaliação das traduções, e uma análise das avaliações das traduções do trecho do texto original propostas pelo DeepL e pelo Google Tradutor. O principal objetivo foi identificar a existência de algum diferencial entre os sistemas considerando que ambos se baseiam na utilização de redes neurais. Os dados levantados neste estudo mostram que, apesar da expectativa inicial de que os sistemas teriam um desempenho significativamente diferente um do outro, na realidade eles não estão distantes.

Com o presente trabalho, retomamos à pergunta de pesquisa: “Ao considerar os sistemas DeepL e Google Tradutor, podemos dizer, após análises comparativas em termos de um trecho do capítulo ‘Olhos de Ressaca’ da obra Dom Casmurro, de Machado de Assis, do português para inglês, que um dos sistemas apresenta resultados mais precisos em relação aos dados analisados?”. Levando em consideração tal pergunta, podemos dizer que em relação às avaliações automáticas, na pontuação BLEU, o Google Tradutor obteve um melhor resultado, já na pontuação TER, o DeepL se sobressaiu. E considerando a avaliação humana, usando como parâmetro a tabela de fluência e adequação, o DeepL apresentou uma melhor pontuação. Assim,

ao considerar as análises realizadas, apesar de ambos os sistemas estarem próximos em questão de qualidade, e que as diferenças de pontuação não sejam tão discrepantes, podemos concluir que o sistema DeepL ainda assim se sobressaiu, estando mais próximo majoritariamente de sua tradução de referência, apresentando resultados mais precisos em relação aos dados analisados.

Em relação à segunda pergunta de pesquisa: "Quais métodos seriam eficazes na identificação dos erros linguísticos em resultados de comparação entre os sistemas DeepL e Google Tradutor?", observa-se que os métodos utilizados não seriam tão eficazes para a identificação dos erros linguísticos, nem mesmo para uma análise aprofundada de cada resultado de TA.

Na pontuação BLEU, somos limitados às traduções de referência, não havendo uma solução para dar uma pontuação para uma palavra que fosse correta e semelhante, nos levando à resposta de que não poderíamos confiar totalmente nessa pontuação, pois uma palavra correta e também similar poderia ser considerada como um erro. Na pontuação TER, temos o mesmo problema, pois para a avaliação também seria necessária a utilização de uma tradução de referência, e a pontuação seria então referente a quantas edições, correções fossem necessárias para alcançá-la. Na avaliação humana, utilizando os critérios de adequação e fluência, temos a questão da subjetividade antes mencionada na fundamentação teórica, visto que o próprio tradutor, baseado em seus estudos e experiência profissional/pessoal, irá pontuar cada sentença. Como percebemos nas análises, os métodos utilizados nos levaram a determinadas métricas que não eram esperadas. Ao decorrer da investigação, os dados apontavam para situações em que eles não

abrangiam totalmente as traduções realizadas por sistemas neurais.

O presente estudo entende relevante retomar os fundamentos, e os pensamentos de Forcada (2010) sobre como ainda é preciso estar atento ao uso das traduções automáticas, pois elas ainda se diferenciam de traduções produzidas por especialistas. Nessa análise, foi utilizado um texto literário, e apesar de ter sido apenas um curto trecho, observou-se como a tradução automática do Google Tradutor não soube identificar corretamente o contexto que caracteriza a personagem Capitu, causando uma diferença substancial de tal trecho da obra de Machado de Assis vertida automaticamente para o inglês. Esse fato, considerando a análise realizada, nos leva à conclusão de que traduções automáticas ainda não podem ser utilizadas sem uma devida correção em textos literários.

É interessante realçar que o intuito deste trabalho foi o de contribuir com os estudos referente a sistemas de TA, e identificar se entre o Google Tradutor e o DeepL algum se sobressairia na tradução automática, além de identificar os métodos que melhor avaliariam o desempenho dos sistemas de TA. Em resumo, pode-se considerar que foi possível identificar e responder todos os objetivos propostos neste estudo, de maneira satisfatória.

Em especial, este estudo agrega à formação de estudantes de Letras no sentido de oferecer metodologias que adicionam não só o aspecto linguístico e tradutório, como também métricas e noções gerais de programação que podem ser incorporadas à formação, especificamente, por exemplo, de estudantes de Línguas Estrangeiras Aplicadas ao Multilinguismo e à Sociedade

da Informação (LEA-MSI), como é o caso da autora do presente trabalho. Mais especialistas nessa área podem contribuir para o avanço desses estudos dentro da perspectiva dos cursos de Letras, em especial, de LEA-MSI (Pimentel e Pires, 2024; Pires, 2020).

Visando pesquisas futuras que possam continuar a contribuição para a interface do campo de pesquisa dos Estudos de Tradução e Linguística Computacional, incluindo especificamente a Avaliação de Tradução Automática, seria interessante a realização de novas análises utilizando textos de outras áreas do conhecimento, além do campo literário, na intenção de descobrir o desempenho dos dois sistemas de TA ao considerar contextos e falas diferentes. Outra questão a ser aprofundada é a replicação dos métodos deste trabalho com o emprego de outros pares linguísticos, propondo-se a identificação do comportamento dos sistemas de TA, considerando a similaridade da pontuação em relação ao presente trabalho.

Referências

ASSIS, M. de. **Dom Casmurro**. Rio de Janeiro: Edições Câmara, 2019.

BANITZ, B. Machine translation: a critical look at the performance of rule-based and statistical machine translation. **Cadernos de Tradução**, 40(1), 2020. Disponível em: https://periodicos.ufsc.br/index.php/traducao/article/view/2175-7968.2020v40n1p54/42358 . Acesso em 17 de novembro de 2023.

BROWNLEE, J. **A Gentle Introduction to Calculating the BLEU Score for Text in Python.** 2019.https://machinelearningmastery.com/calculate-bleu-score-for-text-python/ > Acesso em 17 de novembro de 2023.

CALLISON-BURCH, C., OSBORNE, M., & KOEHN, P. **Re-evaluating the Role of BLEU in Machine Translation Research**, 2006. https://aclanthology.org/E06-1032.pdf . Acesso em 17 de novembro de 2023.

DEEPL PRESS INFORMATION. (n.d.). https://www.deepl.com/en/press.html. Acesso em 17 de novembro de 2023.

DENSMER, L.. **Interview with an Expert: How Do You Measure MT?** RWS. 2019. https://www.rws.com/blog/interview-with-an-expert-how-do-you-measure-mt/ . Acesso em 17 de novembro de 2023

FORCADA, M. Machine translation today. In: GAMBIER, Y.; DOORSLAER, L. V. (Orgs.). **Handbook of Translation Studies** (pp. 215–223). John Benjamins Publishing Company, 2010.

GOOGLE CLOUD. **Como avaliar modelos**. 2022. https://cloud.google.com/translate/automl/docs/evaluate?hl=pt-br. Acesso em 17 de novembro de 2023.

HUTCHINS, W. J., & SOMERS, H. L. **An introduction to machine translation** (pp. 215-223). Academic Press, 1992.

ISABELLE, P., CHERRY, C., & FOSTER, G. A challenge set approach to evaluating machine translation. **Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing**, 2486–2496, 2017. https://aclanthology.org/D17-1263.pdf . Acesso em 17 de novembro de 2023.

PIMENTEL, C. H. M.; PIRES, T. B. Treinamento e análise de um modelo de tradução automática baseado em Transformer. **Texto Livre,** Belo Horizonte-MG, v. 17, p. e49118, 2024. DOI: 10.1590/1983-3652.2024.49118. Disponível em: https://periodicos.ufmg.br/index.php/textolivre/article/view/49118 . Acesso em: 15 jul. 2024.

PIRES, T. B. A avaliação de tradução automática na atuação do bacharel em LEA-MSI. In: PEREIRA, Fernanda Alencar (Org.). **Línguas Estrangeiras Aplicadas**: trajetórias e possibilidades. Campinas: Pontes Editores, 2020. p. 61–75.

SABELLE, P., & KUHN, R. **A Challenge Set for French -> English Machine Translation.** 2018. https://arxiv.org/pdf/1806.02725.pdf. Acesso em 17 de novembro de 2023.

MATTHES, E. **Curso Intensivo de Python.** Novatec Editora. 2016.

PLANETCALC. **Levenshtein distance**.2022. http://planetcalc.com/1721. Acesso em 17 de novembro de 2023.

SCHUSTER, M., JOHNSON, M., & THORAT, N. **Zero-Shot Translation with Google's Multilingual Neural Machine Translation System**. 2016. https://ai.googleblog.com/2016/11/zero-shot-translation-with-googles.html . Acesso em 17 de novembro de 2023.

TAN, Z., WANG, S., YANG, Z., CHEN, G., HUANG, X., SUN, M., & LIU, Y. Neural machine translation: a review of methods, resources, and tools. **AI Open**, 5–21, 2020.