UNIVERSIDADE DE BRASÍLIA
INSTITUTO DE LETRAS – IL
DEPARTAMENTO DE LÍNGUAS ESTRANGEIRAS APLICADAS E TRADUÇÃO – LET
CURSO DE LÍNGUAS ESTRANGEIRAS APLICADAS (LEA-MSI)

CLÓVIS HENRIQUE MARTINS PIMENTEL

TREINANDO UM MODELO DE TRADUÇÃO AUTOMÁTICA BASEADO EM *TRANSFORMERS*

BRASÍLIA, DF
2023

CLÓVIS HENRIQUE MARTINS PIMENTEL

TREINANDO UM MODELO DE TRADUÇÃO AUTOMÁTICA BASEADO EM *TRANSFORMERS*

Trabalho de Conclusão de Curso apresentado ao Departamento de Línguas Estrangeiras e Tradução como requisito parcial para a obtenção do título de Bacharel em Línguas Estrangeiras Aplicadas ao Multilinguismo e à Sociedade da Informação (LEA - MSI)

Orientador: Prof. Dr. Thiago Blanch Pires

BRASÍLIA, DF
2023

CLÓVIS HENRIQUE MARTINS PIMENTEL

TREINANDO UM MODELO DE TRADUÇÃO AUTOMÁTICA BASEADO EM *TRANSFORMERS*

Trabalho de Conclusão de Curso apresentado ao Departamento de Línguas Estrangeiras e Tradução como requisito parcial para a obtenção do título de Bacharel em Línguas Estrangeiras Aplicadas ao Multilinguismo e à Sociedade da Informação (LEA - MSI)

Orientador: Prof. Dr. Thiago Blanch Pires

Aprovado em: ____ / ____ / ____

Banca Examinadora

_____

Prof. Dr. Thiago Blanch Pires
Universidade de Brasília
Orientador

_____

Profa. Dra. Clarissa Prado Marini
Universidade de Brasília
Avaliadora

_____

Prof. Dr. Cláudio Corrêa e Castro Gonçalves
Universidade de Brasília
Avaliador

BRASÍLIA
2023

RESUMO

O presente trabalho possui como objetivo a análise dos modelos de tradução automática baseados em *Transformers*. Em específico, visa o teste da viabilidade do uso de modelos treinados a partir de corpus especializado. Para o treinamento do modelo foi construído corpus paralelo inglês-francês a partir de sete textos relativos à Convenção de 25 de outubro de 1980 sobre os Aspectos Civis do Rapto Internacional de Crianças e Adolescentes, posteriormente utilizado para treinar o modelo T5- *small*. Os resultados de tradução obtidos pelo modelo treinado foram comparados com resultados produzidos pelo Google Tradutor. Para a avaliação dos resultados foram utilizados os métodos de avaliação automática sacreBLEU e avaliação humana baseada na fluência, adequação e erros cometidos pelos modelos em suas traduções. Os resultados da avalição automática de frases produzidas pelo modelo treinado foram, em média, maiores que aquelas geradas pelo modelo não-treinado. A avaliação humana das frases revelou que houve erros de adequação no uso da linguagem específica à matéria da Convenção da Haia de 1980 tanto em frases geradas pelo modelo treinado, quanto em frases geradas pelo modelo do Google Tradutor.

Palavras-chave: Linguística Computacional. Tradutor automático. *Transformers*. Corpus Paralelo. Convenção Internacional.

ABSTRACT

The objective of this work is to analyze machine translation models based on Transformers. It aims to test the viability of using trained models based on a specific corpus. For the training of this model, a parallel English-French corpus was built with seven texts related to the Convention of 25 October 1980 on the Civil Aspects of International Child Abduction, later used to train the T5-*small* model. The translation results obtained by the trained model were compared with results produced by Google Translate. For the evaluation of the results, automatic evaluation methods sacreBLEU and human evaluation based on fluency, adequacy and errors committed by the models in their translations were used. The results of the automatic evaluation of sentences produced by the trained model were, on average, higher than those generated by the non-trained model. The human evaluation of the sentences revealed that there were adequacy errors in the use of language specific to the subject matter of the 1980 Hague Convention both in sentences generated by the trained model and in sentences generated by the Google Translate model.

Keywords: Computational Linguistics. Machine Translation. Transformers. Parallel Corpus. International Convention.

1 INTRODUÇÃO

Dentro da Linguística Computacional (LC), a tradução automática baseada em aprendizado de máquina[1] é uma área da inteligência artificial que se concentra em desenvolver sistemas e algoritmos capazes de simular o modo de aprendizagem humana e traduzir automaticamente textos ou fala de um idioma para outro. Essa técnica de tradução utiliza modelos de linguagem para treinar algoritmos a partir de dados de treinamento que normalmente consistem em pares de frases ou textos em ambos os idiomas para os quais se deseja realizar a tradução. O modelo é então treinado para aprender padrões e regularidades nessas correspondências linguísticas, permitindo assim a tradução automática.

Em meados de 2017, com a publicação do artigo *Attention is All You Need*, pela equipe da Google (Vaswani *et al*., 2017), iniciou-se a mudança de paradigma das técnicas utilizadas no processamento de dados sequenciais, especialmente no campo do Processamento de Linguagem Natural (PLN). Essa nova arquitetura de rede neural denominada de *Transformers* possibilitou um aprimoramento significativo na qualidade e precisão de tarefas como a tradução multilíngue, antes amplamente baseada em modelos de processamento de redes neurais recorrentes (RNNs).[2]

Até então, os tradutores automáticos baseados em RNNs eram dominantes, mas ainda apresentavam limitações, como a dificuldade em capturar dependências de longo prazo, resultando na perda de informações contextuais importantes (Iosifova, 2020). Com a chegada dos *Transformers*, a tradução automática passou a apresentar resultados significativamente melhores, possibilitando a criação de modelos com maior capacidade de lidar com contextos complexos e produzir traduções mais precisas. Essa mudança de paradigma inaugurou uma nova era na tradução automática, com modelos cada vez mais sofisticados e precisos.

Atualmente, é possível encontrar de forma bem acessível tradutores automáticos que produzem traduções multilíngues de altíssima qualidade. No entanto, em muitos casos, ainda é necessário que o texto dessas traduções geradas de forma automática passem por uma revisão humana que garanta uma adequação linguística multilíngue mais precisa. Isso se torna ainda

[1] Cf. https://www.ibm.com/topics/machine-learning. Acesso em: 12 jul. 2023.

[2] Tipo de arquitetura de rede neural projetada para lidar com dados sequenciais, onde a ordem e a dependência temporal são importantes. Ao contrário das redes neurais tradicionais, que processam os dados de forma independente, as RNNs possuem uma dimensão temporal e têm conexões que permitem que as informações sejam retransmitidas nesse aspecto, construindo algo semelhante a uma memória. Isso faz com que os dados de entrada processados sejam não tão somente os dados de um instante t mas também aqueles gerados em uma iteração t-1 (Iosifova, 2020).

mais eminente quando a tradução envolve a utilização de termos técnicos altamente especializados, provenientes de campos do conhecimento igualmente específicos.

Diante dessa situação, surge a indagação sobre como seria possível desenvolver um tradutor multilíngue especializado, baseado em uma arquitetura *Transformers* capaz de lidar com a tarefa de tradução automática voltada para áreas específicas do saber. Relacionado a isso, este trabalho explora a relação do desenvolvimento dessa técnica automática de tradução com o campo do Direito Internacional Privado, desenvolvendo um modelo de tradução especializado no léxico utilizado em torno da Convenção de 25 de outubro de 1980 sobre os Aspectos Civis do Rapto Internacional de Crianças e Adolescentes e das consequentes aplicações e políticas públicas internacionais que gravitam em torno dela.

Este artigo tem como objetivo geral a análise sobre os modelos de tradução automática baseados em transformadores. Partindo de premissas eminentes da Linguística Computacional, é de fundamental importância para este trabalho a busca por um amplo entendimento sobre a essa técnica criada para a tradução automática. Considerar atualmente os *Transformers* como uma das melhores arquiteturas para realizar tarefas de PLN implica na compreensão geral de sua estrutura de processamento de dados.

Um dos objetivos específicos do trabalho é, primeiramente, testar a viabilidade do uso de modelos de tradução automática baseados em *Transformers*. Aqui, será levado em consideração a viabilidade de implementação do modelo (programas e plataformas necessárias para a implementação do código, tutoriais disponíveis para a efetivação do processo, valores dispendidos para o treinamento do modelo), sua velocidade de processamento relativa ao corpus usado para o treinamento e o nível dos resultados gerados em comparação com todo o tempo despendido para a construção de um corpus específico relativo ao tema, sua implementação e seu treinamento.

Em segundo lugar, busca o desenvolvimento de um corpus específico relativo à Convenção de 25 de outubro de 1980 sobre os Aspectos Civis do Rapto Internacional de Crianças, que servirá de base para o treinamento de um modelo tradutor especializado para esta área do conhecimento, sendo assim capaz de traduzir textos de forma mais satisfatória (com maior precisão e adequação das palavras utilizadas) do que tradutores automáticos não especializados.

Por fim, procura avaliar em que medida a construção de um corpus específico e sua implementação em um modelo de tradução se torna mais satisfatória do que uma tradução feita em um modelo com léxico não especializado. Em outras palavras, testar se a qualidade dos resultados gerados por um modelo treinado consegue ser maior, gerando textos com vocábulos

mais precisos e adequados lexicalmente que os resultados de traduções feitas por outras plataformas não treinadas, como o Google tradutor.

Resumidamente, baseados em preceitos da LC, este trabalho explora as vantagens e desvantagens do uso de modelos de tradução automática baseados em *Transformers* e treinado a partir de corpora específicos para a tradução de textos técnicos e jurídicos, principalmente no contexto da Convenção sobre o Rapto Internacional de Crianças.

Em relação à metodologia utilizada, este trabalho busca estudar o modelo de tradução automática T5,[3] introduzido pelo Google *Research* em 2019 (Raffel, 2020) e disponibilizado pela plataforma *HuggingFace*, empresa especializada em processamento de linguagem natural (PLN) e aprendizado de máquina. Para a definição do corpus especializado sobre o tema "Convenção sobre os Aspectos Civis do Sequestro Internacional de Crianças", foi feita uma seleção de textos nas línguas Inglês e Francês – as duas línguas oficiais da Conferência da Haia de Direito Internacional Privado. Todo o conteúdo do corpus foi retirado de textos e traduções oficiais disponibilizados em seu site.

Após a construção do corpus, este foi aplicado ao modelo de tradução automática preexistente, treinando-o e o afinando[4] para a obtenção de traduções que utilizassem de forma semanticamente mais precisa expressões com maior recorrência no âmbito do léxico da convenção internacional aqui mencionada. Ao final, as traduções obtidas foram analisadas e comparadas aos resultados gerados pelo mecanismo de tradução automática multilíngue disponível online da Google.

Partindo de aplicações de técnicas desenvolvidas pela LC, este trabalho também se demonstra relevante para a área do Direito Internacional Privado, uma vez que aqui se desenvolve um modelo de tradutor altamente personalizado e relevante para tarefas de tradução multilíngue relativas à Convenção da Haia de 1980 e busca contribuir para o desenvolvimento de ferramentas e técnicas que possam melhorar a eficácia e precisão dos sistemas de tradução automática em campos especializados, ampliando as possibilidades de comunicação multilíngue em diferentes áreas do conhecimento.

---

[3] T5 é um modelo codificador-decodificador pré-treinado em uma mistura de tarefas não supervisionadas e supervisionadas, em que cada tarefa é convertida para um formato de *text-to-text*. Esse modelo performa muito bem em uma grande variedade de tarefas sem a necessidade de ajustes. Para isso é necessário, no entanto, adicionar um prefixo diferente à entrada correspondente a cada tarefa. Para que esse modelo realize a tarefa de tradução, por exemplo, é necessário inserir o comando: "traduzir do inglês para o francês: ". Cf. https://huggingface.co/docs/transformers/model_doc/t5. Acesso em: 12 jul. 2023.

[4] O afinamento, processo que integra o treinamento do modelo, ocorre após o pré-treinamento do modelo e envolve a melhoria adicional do desempenho do modelo, ajustando-o em um conjunto de dados mais específico e direcionado para a tarefa de tradução aqui em questão. Cf. https://huggingface.co/docs/transformers/main/training. Acesso em: 11 jul. 2023.

A seguir, no capítulo 2, serão discutidos alguns conceitos técnicos e linguísticos mais relevantes para este trabalho e será feita uma revisão de literatura, baseada em trabalhos prévios que abarcam o tema aqui tratado. Após isso, no capítulo 3, será narrada de forma mais extensa a metodologia utilizada para a produção dos resultados, que, por sua vez, serão analisados no capítulo 4. Por fim, as considerações finais explicitarão a qualidade os resultados obtidos e ressaltarão a relevância deste trabalho como um todo.

2 REVISÃO DE LITERATURA

Baseado no estado-da-arte da tecnologia de tradução automática, aqui neste artigo ilustrado pelas pesquisas de Lakew *et al*. (2018), Banitz (2020), Tian *et al*. (2020), Iosifova *et al*. (2020) e Kimera *et al*. (2023), para citar algumas, este trabalho opta pela utilização de um modelo pré-treinado baseado na arquitetura *Transformer* para o processamento de seus dados. Essa escolha se dá mediante extensiva literatura comprovando a eficácia desse modelo em detrimento do uso de RNNs (Vaswani, 2017; Raffel, 2020 e Wolf, 2019). Tal característica é confirmada por Lakew *et al*. (2018), que também atesta a superioridade que essa arquitetura possui quando performa tarefas de processamento de dados em modelos bilíngues, objeto aqui estudado.

Outro trabalho extremamente relevante é o modelo de tradução automática francês-inglês, baseado em uma arquitetura *Transformer*, desenvolvido por Tian *et al*. (2020). O artigo referente a esse trabalho afirma a superioridade do modelo na performance de tarefas de tradução automática e afirma que modelos baseados em RNNs possuem dois grandes defeitos. O primeiro deles seria a sua limitação no processamento, restritos ao processamento individual de palavras e ocasionando um atraso em relação a sua velocidade de treinamento. O segundo negativo mencionado em relação modelo que utiliza RNNs é a incapacidade dessa arquitetura de processar dados de forma acurada quando as frases se tornam demasiadamente longas.

Utilizando então um modelo baseado em *Transformers*, Tian *et al*. (2020), treinaram um modelo de tradução automática francês-inglês. Os resultados do treinamento puderam demonstrar uma acurácia de 80% em suas traduções, resultado informado como maior que aquele produzido por um modelo baseado em RNNs. O trabalho, no entanto, se limita a explicitar somente os dados obtidos a partir do processo de treinamento do modelo (perda de validação e de treinamento). O artigo não demonstra os valores de acurácia do modelo RNNs e tampouco procede à avaliação automática ou humana de nenhuma de nenhuma tradução produzida pelo modelo treinado baseado em *Transformers*.

Seguindo ainda o estudo de Iosifova *et al*. (2020), este trabalho também parte do pressuposto de que um modelo de tradução automática produz traduções mais adequadas semanticamente quando, após pré-treinado, é submetido a um processo de afinação com dados que possibilitem o treinamento do modelo para a melhora de uma tarefa específica. No caso deste trabalho, a utilização do corpus paralelo inglês-francês para a melhora na precisão do vocábulo utilizado em traduções sobre o tema da Convenção de Haia de 1980.

Sobre esse aspecto, podemos citar um modelo de tradução inglês-francês para textos de conteúdo jurídico, disponível na plataforma *HuggingFace*. O modelo "SEBIS/legal_t5_*small*_multitask_en_fr",[5] foi treinado a partir de um corpus paralelo de 9 milhões de pares de frases, 220 milhões de parâmetros, *batch size* de tamanho 4096, *sequence length* de 512, e corpus pré-processado a partir de 88 milhões de frases possui o *score* sacreBLEU de 38,063. Essas configurações e suas implicações serão explorados em momento mais adequado durante este trabalho, mas cabe ressaltar que os números aqui definidos indicam grande capacidade de processamento, muito superior ao que se é possível realizar em um computador de uso pessoal, por exemplo.

Tal modelo se aproxima da proposta estabelecida neste trabalho – treinar um modelo de tradução automática no par de língua inglês-francês –, porém se distancia drasticamente em relação às dimensões e capacidade de processamento no qual este trabalho se insere. Não obstante, serve como uma ótima referência para a confirmação de que o treinamento de modelos de tradução automática requerem ainda um enorme número de dados coletados e uma capacidade de processamento ainda fora do alcance de máquinas pessoais.

Um outro trabalho que se aproxima mais daquele aqui realizado foi desenvolvido por Kimera *et al*. (2023). Em seu artigo, os autores descrevem a construção de um corpus paralelo contendo 41.070 pares de frases em inglês e luganda, posteriormente usado para o treinamento de um modelo baseado em *Transformer*. Após o treinamento de vários modelos testando hiperparâmetros[6] diferentes, em um deles foi possível obter o *score* BLEU final de 17,47 para traduções de inglês para luganda. Apesar de ainda contar com um número muito maior de frases alinhadas e capacidade de processamento de dados igualmente superior, esse trabalho lança uma luz sobre a realidade dos materiais necessários para se produzir um modelo de tradução automática minimamente capaz de melhorar um modelo pré-treinado não especializado.

---

[5] Disponível em HuggingFace.co/SEBIS/legal_t5_*small*_multitask_en_fr. Acesso em: 12 jul. 2023.

[6] Variáveis de configuração externa, definidas manualmente, usadas para gerenciar o treinamento do modelo de aprendizado de máquina. Os hiperparâmetros determinam os recursos taxa de aprendizado e complexidade do modelo. Cf. https://aws.amazon.com/pt/what-is/hyperparameter-tuning. Acesso em: 17 jul. 2023.

O trabalho de Kimera *et al*. (2023), se demonstra extremamente relevante para a compreensão da relação entre o modo como o modelo é treinado e a aferição dos resultados produzidos pelo modelo. Durante o artigo são demonstrados alguns dados que compravam a relação direta entre os hiperparâmetros escolhidos pelos autores (como o tamanho do *batch*) e a influência que eles possuem na aferição do *score* BLEU do modelo. Alterando alguns hiperparâmetros era possível obter *scores* menores em relação ao *score* final e o artigo mostra, então, a definição de parâmetros não ótimos que leva ao *score* BLEU 13,96.

Semelhante ao que se propõe este trabalho, o artigo produzido por Kimera *et al*. (2023), busca não somente apresentar o *score* do modelo após seu treinamento, mas explora também o resultado de 4 frases traduzidas do inglês para luganda produzidas pelo seu modelo de tradução. Apesar disso, diferentemente do que se propõe no presente trabalho, o artigo de Kimera *et al*., 2023, não apresenta nenhuma avaliação humana consistente de suas traduções e nem pôde comparar, apesar de ter explicitado que esse era um dos objetivos futuros do trabalho, as traduções com resultados gerados pelo Google Tradutor, por ainda não existir, até a data da submissão do trabalho, a língua luganda disponível na plataforma online.

Cabe, por fim, mencionar o trabalho de Banitz (2020), que realiza um trabalho exímio na avaliação de traduções realizadas por dois modelos diferentes, um deles sendo o Google Tradutor e o outro o Systran. Em seu trabalho, a autora utiliza tanto métodos de avaliação automática (*Translation Error Rate*/TER *score*)[7] como métodos de avaliação humana para avaliar as primeiras 24 frases traduzidas a partir de seu corpus.

Em relação à avaliação automática, a autora (Banitz, 2020) esquematiza os *scores* TER obtidos dos resultados de tradução dos dois mecanismos automáticos em uma tabela e os compara explicitando que as traduções produzidas pelo Google Tradutor requerem, ao final, menos pós-edição, uma vez que apresentam uma taxa de erro menor que o outro modelo avaliado.

Já em relação à avaliação humana dos resultados, o trabalho de Banitz (2020) utiliza métricas bem delineadas de avalição de fluência e adequação das frases geradas (assunto mais bem explorado em subseção específica). Também em uma tabela comparativas dos resultados de tradução gerados pelos dois mecanismos de tradução automática, a autora compara os *scores* atribuídos às frases e aponta os resultados do Google Tradutor como mais satisfatórios. Todavia,

[7] Apesar de justificar a escolha desse método de avaliação automática como um método mais intuitivo do "quão bom" é a tradução, a autora explicita as limitações impostas por essa métrica de avaliação como: (i) o método não reflete necessariamente a adequação da tradução gerada e (ii) a métrica depende diretamente da qualidade da tradução-referência, uma vez que qualquer desvio da tradução humana será penalizada (Banitz, 2020).

a autora ainda explicita os desafios linguísticos que a tradução automática encontra em relação a erros semânticos, lexicais, sintáticos e morfológicos.

Apesar de não abordar o uso de modelos *Transformers* especificamente treinados para a melhora do resultado de traduções, o trabalho de Banitz (2020) abarca importantes discussões sobre os métodos de avaliação desenvolvidos e utilizados para a avaliação e comparação de traduções automaticamente geradas. É nesses aspectos que o trabalho da autora contribui para este artigo, fornecendo grande suporte científico para o método de comparação dos resultados aqui elaborado.

Para a melhor compressão dos parâmetros apresentados por Banitz (2020), é necessário entender que o erro de tradução, objeto de grande problematização dentro dos estudos de tradução, passa a ser utilizado, dentro do contexto da computação, com seu sentido matemático de "cômputo" (Pires, 2017). Como afirma o Professor Doutor Thiago Pires (2017, p. 18) em sua tese de doutorado: "Por essa razão é raramente conceitualizado, questionado ou teorizado quando o mesmo é relacionado a uma aplicação de tradução automática".

Seguindo ainda a linha apresentado pelo Professor Thiago Pires, os erros de tradução são aqui entendidos como "configurações de incompatibilidades linguísticas (lexical, semântica e sintática) entre o texto de entrada e o texto de saída gerado por uma tradução automática em um dado contexto de produção" (Pires, 2017, p. 18) e o "método de comparação entre o resultado da tradução gerado automaticamente e traduções consideradas "corretas", ainda que humano, é abstrato" (Vilar, 2006 apud Pires, 2017, p. 42). Pires se baseia em White (2003, p. 214), quando esse último afirma, sobre o problema da falta de verdade absoluta na tradução, que:

> MT evaluation is harder than this. Only people who know both languages can know just by looking whether it got a translation right. And as we noted above, there is great latitude for disagreement about what constitutes "exactly right" in translation. So we cannot take full advantage of the notion of "ground truth": the set of right answers that form a universally agreedKupon standard for comparison of evaluation results (e.g., the answer key of a school quiz, or the map of a minefield). Therefore we must somehow accommodate some highly subjective judgments about which translation might be better than which other translation.

Após a análise de alguns trabalhos relevantes na área, o presente estudo abordará no próximo capítulo, a metodologia desenvolvida neste artigo para treinar um modelo de tradução automática baseado em *Transformer*, levando em consideração o vocabulário específico da Convenção da Haia de 1980. Exploraremos técnicas específicas para o alcance de maior precisão, consistência e adaptação do léxico ao campo do conhecimento aqui em questão.

3 METODOLOGIA

Primeiramente, para a confecção do corpus especializado no tema da Convenção sobre os Aspectos Civis do Rapto Internacional de Crianças, optou-se somente pela seleção de textos e traduções que fossem reconhecidos como oficiais pela própria Conferência da Haia de Direito Internacional Privado. Essa instituição é uma organização intergovernamental referente à área do Direito Privado Internacional que administra diversas convenções internacionais, protocolos e instrumentos de *soft law* (regras de valor normativo limitado e que não possuem caráter jurídico obrigatório), com o objetivo de unificar progressivamente as normas na área do direito internacional privado.

Apesar de suas convenções internacionais não possuírem mandatoriamente o valor de lei para os países que assim não as convalidam, essas Convenções e instrumentos fornecem clareza e direção em relações transfronteiriças em relação a várias matérias de Direito Internacional Privado, dentre elas o Direito Internacional de Família e Proteção à Criança e Adolescente. É tarefa dessa Conferência uniformizar as normas aplicadas, zelar pelo seu adequado cumprimento e difundi-las entre os países. Com isso, a exclusiva seleção de texto oficiais produzidos pela própria Conferência da Haia atribui ao corpus uma maior segurança e uniformidade em seu léxico.[8]

Em relação à escolha das duas línguas selecionadas para integrar o corpus paralelo[9] criado, a opção pelas línguas Inglês e Francês se deu pela grande disponibilidade de material produzido e facilmente disponibilizado nesses idiomas, que, até o presente momento, são as duas únicas línguas oficiais dessa organização intergovernamental.[10] Assim, a maioria dos textos possui sua primeira redação em francês, com sua subsequente tradução oficial para o inglês e posterior disponibilização em diversas outros idiomas.[11]

---

[8] Cf. https://www.hcch.net/pt/about. Acesso em: 17 jul. 2023.

[9] Um corpus paralelo é uma coleção de textos em um ou mais idiomas juntamente com suas traduções para outros idiomas, armazenados em um formato legível por máquina. Ele se refere a textos originais e suas traduções, ou seja, os mesmos textos em diferentes idiomas, podendo também ser conhecido como “corpus de tradução” (Hallebeek, 2000). Também pode ser útil considerar o atributo "paralelo" como referindo-se a um tipo de arquitetura de corpus, em vez do status dos textos em relação à tradução. Os corpora paralelos podem ser entendidos como corpora nos quais dois ou mais componentes estão alinhados, ou seja, são subdivididos em unidades composicionais e sequenciais (Fantinuoli; Zanettin, 2015).

[10] Cabe destacar que apesar de o site já possuir versões completas em outras línguas como o alemão, português e espanhol, somente essa última introduzida ao rol de línguas oficiais dessa organização, em 1 de julho de 2024. Cf. https://www.hcch.net/pt/news-archive/details/?varevent=907. Acesso em: 13 jul. 2023.

[11] Para a construção do corpus foram utilizados, ao todo, 7 pares de textos, cada um com sua versão original em francês e versão em inglês: (i) texto oficial da Convenção sobre os Aspectos Civis do Rapto Internacional de Crianças; (ii) Relatório Explicativo de Eliza Pérez-Vera; e (iii) Guias de Boas Práticas sobre a Convenção da Haia de Subtração Internacional de Crianças e Adolescentes (Partes I a V). Textos disponíveis em https://www.hcch.net/pt/publications-and-studies/publications2. Acesso em: 12 jul. 2023.

Após a seleção dos textos relativos ao tema, foram necessárias a limpeza e formatação desses textos na plataforma Word, removendo quaisquer sinais gráficos ou de tabulação que dificultassem o posterior alinhamento das frases em uma tabela Excel. Uma vez inserido o conteúdo de todos os textos em suas duas versões na tabela Excel, procedeu-se ao alinhamento das frases, fazendo com que as frases de uma mesma linha na tabela tivessem a mesma correspondência semântica em suas respectivas línguas, conforme Quadro 1.

Quadro 1- Frases em inglês e francês alinhadas de acordo com sua correspondência semântica

Texto em Inglês	Texto em Francês
Convention on the Civil Aspects of International Child Abduction	Convention Sur Les Aspects Civils De L'enlèvement International D'enfants
Concluded 25 October 1980	Conclue le 25 octobre 1980
The States signatory to the present Convention,	Les Etats signataires de la présente Convention,
Firmly convinced that the interests of children are of paramount importance in matters relating to their custody,	Profondément convaincus que l'intérêt de l'enfant est d'une importance primordiale pour toute question relative à sa garde,
Desiring to protect children internationally from the harmful effects of their wrongful removal or retention and to establish procedures to ensure their prompt return to the State of their habitual residence, as well as to secure protection for rights of access,	Désirant protéger l'enfant, sur le plan international, contre les effets nuisibles d'un déplacement ou d'un non-retour illicites et établir des procédures en vue de garantir le retour immédiat de l'enfant dans l'Etat de sa résidence habituelle, ainsi que d'assurer la protection du droit de visite,
Have resolved to conclude a Convention to this effect and have agreed upon the following provisions.	Ont résolu de conclure une Convention à cet effet, et sont convenus des dispositions suivantes.

Fonte: Autor

Este passo demonstrou-se o mais trabalhoso, sendo necessário a limpeza e o alinhamento de um total final de 5.494 frases pareadas em inglês e francês (EN-FR). Levando em conta todos os textos utilizados, é possível ainda documentar a quantidade de 134.943 *tokens* (total de ocorrências) e 6.453 *types* (vocábulos distintos) presentes nos textos em inglês e 143.285 *tokens* e 8.932 *types* relativos aos textos em francês.

O próximo passo foi decidir a plataforma que seria utilizada para o processamento desses dados e notou-se que a *HuggingFace* seria a melhor opção para isso. Ela é uma plataforma de desenvolvimento de código aberto projetada para trabalhar com modelos de inteligência artificial (IA), especialmente voltada para modelos de processamento de linguagem natural (PLN) e aprendizado de máquina. Ela oferece uma ampla gama de modelos pré-treinados e bibliotecas (inclusive a que disponibiliza o modelo *Transformers*) que facilitam o desenvolvimento, treinamento e implantação de modelos de tradução automática. Com todos esses recursos disponíveis, a própria plataforma ensina, por meio de um tutorial utilizando Python, a implementar a sua biblioteca *Transformers*.

No entanto, para que fosse possível carregar o corpus paralelo EN-FR criado para o treinamento do modelo especializado de tradução automática, foi necessário utilizar uma formatação especial para o corpus, criando-se, a partir de código escrito no aplicativo Bloco de Notas nativo do sistema Windows, um dicionário[12] em arquivo JSON[13] com os pares de frases indexados, conforme exemplo a seguir:

> [{"id": "0", "translation": {"en": "Convention on the Civil Aspects of International Child Abduction", "fr": "Convention Sur Les Aspects Civils De L'enlèvement International D'enfants"}}]

Uma vez feito o *upload* deste dicionário a uma página dedicada a este *dataset* criada em meu perfil na plataforma *HuggingFace*, procedeu-se a divisão do corpus em conjunto de teste e conjunto de treinamento. Enquanto este contém os dados, aqui pares de frases, que serão apresentados ao algoritmo encarregado de criar o modelo, aquele abarca os dados que serão apresentados ao modelo após a sua criação, simulando traduções que o modelo de tradução gerará, permitindo a verificação do desempenho do modelo treinado em um novo conjunto de dados. Esta prática muito comum em aprendizado de máquina[14] e PLN é feita para avaliar a capacidade do modelo de generalizar o conhecimento aprendido durante o treinamento e medir sua performance em dados não vistos anteriormente. Aqui, manteve-se o valor presente no código-exemplo utilizado pela plataforma, sendo 20% do corpus (1.099 frases) utilizado para teste e o restante (4.395 frases) utilizado para o treinamento do modelo.

Partindo-se então do código-exemplo[15] de implementação da biblioteca *Transformers* presente na plataforma *HuggingFace* e utilizando o arquivo JSON previamente mencionado, procedeu-se às adaptações necessárias do código, para que o corpus paralelo pudesse vir a ser processado e utilizado para o treinamento de um novo modelo especializado de tradução.

---

[12] Estrutura de dados que relaciona chaves e valores. Neste trabalho, as chaves continham o número de identificação do par de frases inglês-francês e os valores continham frases com o mesmo sentido semântico em inglês e francês. Cf. https://docs.python.org/3/tutorial/datastructures.html#dictionaries. Acesso em: 13 jul. 2023.

[13] Formato de arquivo amplamente utilizado na Linguística Computacional por apresentar um formato leve de armazenamento de informações estruturadas. Aqui utilizado por permitir a fácil interpretação e apresentação de dados alinhados, característica extremamente útil quando um dos objetivos deste trabalho envolve a construção de corpora paralelo. CF. https://www.w3schools.com/js/js_json_intro.asp. Acesso em: 13 jul. 2023.

[14] Aprendizado de máquina é um campo da inteligência artificial que se concentra no desenvolvimento de algoritmos e modelos capazes de aprender e tomar decisões ou realizar tarefas sem serem explicitamente programados. Em vez disso, os modelos de aprendizado de máquina são treinados em grandes conjuntos de dados para reconhecer padrões e fazer previsões com base neles. O aprendizado de máquina é baseado na ideia de que os modelos podem aprender com exemplos previamente processados para melhorar seu desempenho futuro (Han, 2022 e KOEHN, 2009)

[15] Cf. https://*HuggingFace*.co/docs/transformers/tasks/translation. Acesso em: 12 jul. 2023.

Em relação à escolha do modelo de *Transformers* a ser utilizado a partir da biblioteca disponibilizada pela *HuggingFace*, optou-se por um modelo pré-treinado e que não demandasse uma grande capacidade de processamento computacional para o seu treinamento a partir do corpus especializado. Assim, optou-se pela aplicação do modelo T5-*small*,[16] um modelo de rede neural pré-treinado para tarefas de *text-to-text*,[17] extremamente adequado para tarefas de tradução automática (Raffel, 2020).[18]

Inicialmente, foi utilizada a plataforma de processamento online Google Colaboratory para o processamento prévio de alguns códigos-teste que simulavam, em menor escala, a viabilidade da implementação de todas as bibliotecas necessárias para este projeto e treinava o corpus especializado para a tradução automática. No entanto, após várias tentativas, em decorrência do tamanho do corpus paralelo final, essa plataforma de processamento online não foi capaz de suportar o processamento de tamanha quantidade de dados, sempre quebrando a conexão depois de alguns minutos.

Em decorrência disso, foi necessária a nova adaptação do código em Python para que fosse criado um arquivo local contendo o dicionário com o corpus paralelo em formato JSON e *script* em Python que também fosse processado localmente. Para isso optou-se pelo Anaconda Prompt,[19] que pôde utilizar toda a capacidade de processamento do CPU, GPU e RAM da máquina na qual o corpus foi treinado. Isso fez com que o processamento ocorresse todo localmente e fosse dispensável a conexão com a internet ou o uso do Google Colaboratory.

A partir daí, apesar de não ser mais necessária a conexão com a internet e acesso a processadores remotos, a máquina também apresentava limitações, sendo necessária a realização de uma nova bateria de testes com corpora menores para melhor adequar a

[16] O T5-*small* é uma versão menor do modelo T5 em comparação com outras variantes, como T5-base e T5-large, por exemplo. Ele possui menos parâmetros e é mais leve em termos de recursos computacionais necessários para executar o modelo. Essa versão menor é mais rápida em relação ao tempo de inferência, o que a torna útil em cenários com restrições de recursos computacionais ou quando a velocidade de processamento é prioritária. Embora o T5-*small* possa ter um desempenho um pouco inferior em relação a versões maiores do T5 em algumas tarefas, ele ainda é capaz de realizar uma ampla gama de tarefas de PLN com bons resultados. A título de comparação, o modelo T5-*small* possui em torno de 60 milhões de parâmetros, ao passo que os modelos T5-base e T5-large possuem em torno de 220 milhões e 770 milhões de parâmetros, respectivamente (Raffel, 2020).

[17] Abordagem em que um modelo de PLN é treinado para converter um texto de entrada em outro texto de saída, independentemente da tarefa específica (Raffel, 2020).

[18] Esse modelo foi escolhido por sua capacidade de lidar com uma ampla gama de tarefas de PLN, inclusive traduções multilíngues, e, por já se encontrar em uma fase de pré-treinamento, ser necessário somente o treinamento desse modelo em relação ao léxico especializado da Convenção da Haia de 1980 (Raffel, 2020).

[19] Anaconda é uma plataforma de ciência de dados e distribuição de software que simplifica o gerenciamento de pacotes e ambientes de programação para Python e R. Ela inclui um Ambiente de Desenvolvimento Integrado (IDE) chamado Anaconda *Navigator*, bem como uma ferramenta de linha de comando chamada Anaconda Prompt. O Anaconda Prompt, por sua vez, é uma interface de texto que permite executar comandos e *script*s python. Cf. https://www.anaconda.com/open-source. Acesso em: 13 jul. 2023

capacidade de processamento da máquina ao tamanho do corpus a ser treinado (levando em consideração o desejo por um tempo relativamente rápido para a realização de todo processo).

Em relação à melhor adequabilidade entre o tempo de processamento e qualidade do modelo a ser gerado, dois são os hiperparâmetros que mais influenciam nesses resultados, a saber: o número de *epochs* e o número de *batches*. O primeiro define o número de vezes que todo o conjunto de treinamento é percorrido durante o treinamento, sendo que um número muito baixo pode levar a um *underfitting*[20] do modelo, enquanto um número muito alto pode levar ao *overfitting*.[21]

De forma geral, não há um número específico pré-determinado de *epochs* a ser utilizado em cada modelo; isso varia de acordo com o tamanho do corpus a ser treinado e a complexidade de seus dados. Idealmente, deve-se estabelecer o número de *epochs* baseado no cruzamento de dados de aprendizado do modelo: validação e treinamento.[22] Enquanto as duas curvas estiverem apresentando uma queda em seus valores, deve-se manter o treinamento do modelo, estendendo o número de *epochs* de forma indefinida (Jabbar; Khan, 2015).

No entanto, para este trabalho, levou-se primordialmente em consideração o fator tempo na definição do hiperparâmetro *epoch* deste modelo. O número de *epochs* sugerido pela plataforma *HuggingFace* para a implementação do código era de dois, mas verificou-se que depois de cinco *epochs* o *score* da avaliação automática não passava a ter um considerável aumento em seu valor (menos de 0,06), não sendo muito mais rentável (visando uma otimização do tempo) continuar treinando o modelo. Fazer isso não resultaria em uma alteração muito mais significativa no valor de sua avaliação automática quando comparado com a diferença dos *scores* obtida entre os quatro primeiros *epochs*, mas aumentaria consideravelmente o tempo de treinamento do modelo (um modelo com 10 *epochs* levaria em torno de 127 horas para ser

[20] O *underfitting* ocorre quando um modelo não é capaz de capturar os padrões e relações complexas nos dados de treinamento. Isso resulta em um desempenho insuficiente tanto nos dados de treinamento quanto nos dados de teste. Um modelo *underfit* pode ser muito simples ou ter uma capacidade limitada de representar a complexidade dos dados, não sendo capaz de aprender os padrões e nuances necessários para gerar traduções precisas e coerentes. Sinais de *underfitting* incluem um desempenho ruim nos dados de treinamento e um desempenho igualmente ruim nos dados de teste (Jabbar; Khan, 2015).

[21] O *overfitting* ocorre quando um modelo se ajusta muito bem aos dados de treinamento, mas não é capaz de generalizar corretamente para dados não vistos, como o conjunto de teste. Isso acontece quando o modelo "memoriza" os padrões e detalhes específicos dos dados de treinamento em vez de aprender os padrões mais gerais e relevantes que podem ser aplicados a novos dados. Sinais de *overfitting* incluem um desempenho muito bom nos dados de treinamento, mas um desempenho pior nos dados de teste. O modelo pode se tornar muito complexo, capturando o ruído nos dados de treinamento e se tornando excessivamente especializado neles. Isso pode resultar em traduções excessivamente precisas para os exemplos de treinamento, mas menos precisas ou até mesmo inconsistentes em exemplos diferentes (Jabbar; Khan, 2015).

[22] A curva de aprendizado do treinamento é calculada a partir do conjunto de dados de treinamento que fornece uma ideia de quão bem o modelo está aprendendo. Por sua vez, a curva de aprendizado de validação é calculada a partir de um conjunto de dados de validação separado, que fornece uma ideia de quão bem o modelo está generalizando (Jabbar; Khan, 2015).

treinado). Assim, chegou-se à conclusão de que cinco *epochs* seriam ideais para os objetivos aqui visados, alcançando um bom balanço entre o *score* obtido a partir do método de avaliação automática e o tempo de treinamento despendido; número final adotado.

Já o número de *batches* especifica a quantidade de blocos de pares de frases que serão analisados pelo modelo cada vez que ele processar os dados do conjunto de treinamento (realizar uma iteração). Um tamanho de lote maior pode acelerar o treinamento, mas requer mais memória, enquanto um tamanho de lote menor pode levar a uma convergência mais suave, mas com um treinamento mais lento (Popel, 2018). Em relação ao número de *batches*, após a aferição do tempo despendido para o treino de modelos-teste, optou-se por manter o valor de 16 *batches* já presente no código, visto que eminentemente ótimo para esse trabalho.[23]

Todos esses ajustes no código possibilitaram, após 3 horas 46 minutos e 36 segundos, a obtenção do treinamento bem-sucedido de modelo de tradução especializado no vocabulário referente à Convenção sobre os Aspectos Civis do Rapto Internacional de Crianças.

Após o treinamento do modelo, os resultados obtidos foram comparados com traduções realizadas pelo Google Tradutor, que produz seus resultados a partir de uma arquitetura híbrida de RNNs e *Transformers*, o *Google Neural Machine Translation*/GNMT (WU *et al*., 2016). Buscou-se aqui avaliar a eficácia do modelo de tradução treinado a partir de um corpus especializado em relação aos resultados produzidos pelo tradutor online. Com essa metodologia, foi possível aferir a capacidade do modelo criado a partir de um corpus especializado em produzir traduções linguisticamente precisas e contrastar seus resultados com aqueles produzidos pelo GNMT.

3.1 Avaliação Automática

Para a aferição dos resultados optou-se pela utilização de dois métodos: (i) métrica de comparação automática e (ii) avaliação humana, esta última abordada na próxima subseção deste trabalho. Em relação à métrica de comparação automática, foi utilizado a ferramenta de código aberto sacreBleu, que desempenha o cálculo do *score* BLEU para avaliar a qualidade das traduções automáticas em comparação com as referências humanas (Papieni *et al*., 2002; Post, 2018).

[23] Para se obter o número de vezes que o modelo percorrerá os dados – realizará as iterações – é necessário dividir o número de pares de frases do conjunto de treinamento (aqui 4395 pares de frases), pelo tamanho definido de um *batch*, aqui 16. O modelo teve então o valor inteiro de 275 iterações por *epoch*.

O método sacreBLEU é uma implementação específica do cálculo do *score* BLEU (*Bilingual Evaluation Understudy*)[24] para a avaliação automática de traduções. Ele se baseia nos princípios daquele, automatizando todo o processo de avaliação e fornecendo uma ferramenta de código aberto para facilitar a medição da qualidade das traduções automáticas em relação às referências humanas (Papieni *et al*., 2002; Post, 2018).[25]

O método mencionado segue uma série de etapas para realizar o cálculo preciso do *score* BLEU e então gerar sua própria avaliação e pontuação, que varia de 0 a 100, sendo 100 uma tradução ótima. Ao receber as traduções automáticas e as referências correspondentes, o *script* do sacreBLEU realiza automaticamente o pré-processamento dos dados a serem avaliados utilizando sua métrica interna (Papieni *et al*., 2002; Post, 2018).[26]

Essa metodologia permite uma avaliação objetiva e sistemática da qualidade das traduções automáticas, comparando-as com referências humanas e levando em consideração a precisão dos n-gramas. Contudo, assim como o *score* BLEU, o sacreBLEU também possui limitações que devem ser consideradas ao interpretar os resultados aqui obtidos, como a sua falta de compreensão semântica e as limitações impostas pelas referências que ele utiliza para avaliar a tradução.[27]

Em relação à falta de compreensão semântica, o BLEU não avalia a precisão das traduções, não sendo capaz capturar as nuances semânticas cruciais para traduções que requerem o uso de termos especializados. É possível que uma tradução tenha o mesmo significado que uma referência utilizada pela própria métrica, mas receba um *score* baixo. Da mesma forma, é possível que uma tradução tenha altas pontuações de BLEU, mas transmita um significado diferente ou incorreto em comparação com as referências (caso ilustrado pelo Tabela 1, no capítulo 4 deste trabalho).

Já tratando-se das limitações impostas pelas referências que ele utiliza para avaliar a tradução, esse método de avaliação depende de frases-referência definidas por humanos para a avaliação. A escolha das referências pode ser subjetiva e pode não capturar toda a variedade de traduções aceitáveis. Além disso, os resultados BLEU nem sempre melhoram com o aumento

[24] O BLEU é um método de avaliação automática amplamente utilizado na área de processamento de linguagem natural para medir a qualidade de traduções automáticas em relação a referências humanas (Papieni *et al*., 2002).
[25] O BLEU e o sacreBLEU compartilham a mesma abordagem de avaliação baseada em n-gramas. Ambos consideram a correspondência de n-gramas presentes nas traduções automáticas em relação às referências humanas. No entanto, o sacreBLEU é uma implementação específica que oferece uma ferramenta pronta para uso, tornando o processo de avaliação automática mais acessível e conveniente para seus usuários em geral (Papieni *et al*., 2002; Post, 2018).
[26] Cf. https://github.com/mjpost/sacrebleu. Acesso em: 12 jul. 2023.
[27] O *score* também apresenta outras limitações (insensibilidade à ordem das palavras e sintaxe, ênfase excessiva na correspondência do número de n-gramas, favorecimento de traduções mais curtas insensibilidade a sinônimos e paráfrases), essas não tão relevantes para a análise dos resultados desse trabalho (Post, 2018).

do número de frases- referência, e estudos recentes comprovam melhor desempenho do *score* quando somente uma frase referência foi utilizada (Freitag *et al*., 2020).

Utilizando a avaliação automática também para avaliar não só o modelo, mas frases produzidas por ele, foram realizadas avaliações de frases individuais produzidas tanto pelo Google Tradutor quanto pelo modelo treinado. Seguindo técnica utilizada por Freitag *et al*. (2020), para cada tradução comparada, este trabalho escolheu a frase correspondente do texto oficial em francês como parâmetro de referência utilizado pelo sacreBLEU para a avaliação automática e geração de *scores* individuais. Os resultados foram posteriormente comparados seguindo modelo de formatação utilizado por Banitz (2020), oportunamente apresentado no capítulo dedicado à discussão dos resultados.

Optando então pela utilização de uma métrica de avaliação automática de fácil implementação e amplamente difundida, este trabalho utiliza o sacreBLEU como parâmetro de avaliação automática. Necessário, entretanto, lembrar que cada métrica de avaliação encontra algum tipo de limitação que deve ser considerada quando utilizada na avaliação automática de modelos de tradução de mesma natureza. Em decorrência disso, este trabalho preza também pelo desenvolvimento da avaliação humana, que se torna imprescindível para a obtenção de uma compreensão mais abrangente da qualidade da tradução gerada pelo modelo treinado, assunto tratado a seguir.

3.2 Avaliação Humana

Para a avalição humana, este trabalho tem seus alicerces em parâmetros estabelecidos principalmente em dois artigos científicos: o primeiro, de Banitz (2020), brevemente abordado no capítulo 2 e o segundo, elaborado por Vilar *et al*. (2006). O primeiro utiliza os parâmetros de fluência e adequabilidade para aferir a qualidade das traduções, enquanto o segundo apresenta a classificação (com respectivas subclasses) de uma série de erros cometidos pela máquina ao realizar uma tradução.

Primeiramente, em relação aos termos utilizados pela autora Banitz (2020), ela explica que a fluência de uma tradução gerada também pode ser entendida como o nível de sua inteligibilidade, compreendendo tanto a correção gramatical quando a escolha de palavras utilizadas na tradução (Douglas, 1994; Kalyani *et al*., 2014). Por sua vez, a adequação pode ser também entendida como a acurácia ou fidelidade da tradução produzida, e se relaciona com o grau no qual a tradução consegue representar o significado original da frase traduzida (Douglas, 1994 e Kalyani *et al*. 2014).

Para a metrificação de cada um dos dois parâmetros, autora descreve a atribuição de um valor que varia de 1 a 5, cabendo a um humano avaliar a frase traduzida com base nesses valores. Para a fluência, a atribuição do valor 1 corresponde a incompreensibilidade da frase e o valor 5 a uma frase perfeitamente inteligível em determinada língua. Em relação à adequação, atribuir o valor 1 a uma frase implica em dizer que o significado expressado nela não se aproxima de forma alguma do significado expressado na frase que se pretendia traduzir. Já o valor 5 expressa que não houve perda de nenhum significado na frase traduzida.

Já em relação ao trabalho produzido por Vilar *et al*. (2006), esse apresenta uma série de categorias e subcategorias de erros passíveis de identificação quando da análise de uma tradução gerada automaticamente. Este trabalho, no entanto, opta for focar em uma subcategoria específica apresentada pelos autores: o erro aferido a partir da utilização errônea de palavras na tradução, especialmente quando levado em consideração o sentido da palavra presente o texto gerado. Isso pode ocorrer, de acordo com os autores, por uma escolha lexical errada ou por uma desambiguação incorreta. O presente trabalho foca principalmente no fator relacionado à uma escolha lexical errada para comparar as traduções geradas pelo modelo treinado e aquelas geradas pelo Google Tradutor, acreditando residir aqui a maior diferença dentre esses resultados.

Utilizando esses métodos de avaliação humana aqui expostos e métricas de avaliação automática anteriormente discutidos, este trabalho passa a apresentar e discutir, em capítulo próximo, os resultados deste estudo.

4 RESULTADOS

Para o treinamento do modelo de tradução automática, o corpus paralelo construído foi composto por quatorze textos, sendo sete pares de textos em inglês-francês. Os textos selecionados foram todos retirados do site oficial da Conferência da Haia sobre Direito Internacional Privado. Um desses pares de textos representa o texto da própria Convenção de 25 de outubro de 1980 sobre os Aspectos Civis do Rapto Internacional de Crianças e Adolescentes. Os outros pares de textos selecionados possuem relação direta com essa Convenção, sendo um relatório explicativo sobre a própria convenção e outros cinco guias de boas práticas também relacionadas à implementação dessa convenção internacional.

Após a coleta dos textos, foram computados 134.943 *tokens* e 6.453 *types* presentes nos textos em inglês e 143.285 *tokens* e 8.932 *types* relativos aos textos em francês. Ao final foram alinhadas 5.494 frases em inglês-francês. Tais quantidades se mostram muito aquém daquilo

que se encontra em trabalhos de semelhantes aqui já citados, que podem chegar a contar com nove milhões de pares de frases alinhadas. No entanto, o objetivo desse trabalho sempre foi testar a viabilidade do treinamento de um modelo simples, utilizando pouquíssimos recursos de tempo e acesso limitado à capacidade de processamento, para então aferir os resultados gerados a partir disso.

No presente artigo, o modelo aqui treinado recebeu o *score* sacreBLEU de 7,6467. Quando comparado com modelos treinados com técnicas semelhantes, levando em consideração o tamanho extremamente pequeno do corpus usado para treino neste modelo (5.494 pares de frases), e os hiperparâmetros ajustados de forma a ser possível o processamento em uma máquina com configurações feitas primariamente para o uso doméstico, vemos que o valor do resultado da avaliação automática encontra-se alinhado quando comparado com os estudos previamente mencionados que seguem a mesma linha metodológica (Lakew *et al*., 2018; Banitz, 2020; Tian *et al*., 2020 e Kimera *et al*., 2023).

Apesar de apresentar um *score* de aproximadamente 7,65, o modelo tende a performar melhor sua tarefa de acordo com aquilo para que foi treinado. Assim, deve-se também avaliar individualmente a pontuação de alguns de seus resultados de frases traduzidas, para que seja possível compreender o que esse valor atribuído ao modelo, apesar de baixo, realmente representa quando da realização de tarefas de tradução automática sobre o tema aqui exposto.

Para a análise dos resultados da avalição automática, será utilizado o formato similar àquele apresentado por Banitz (2020), identificando 20 frases em inglês (retiradas do par inglês-francês presente no corpus) alimentadas aos dois modelos de tradução por meio do número de sua chave correspondente no corpus. Os *scores* sacreBLEU gerados para cada tradução em francês proveniente de cada modelo seguem logo depois. Para a obtenção dos *scores*, cada tradução gerada, tanto pelo modelo treinado quanto pelo Google Tradutor, foi comparada com a frase referência em francês presente no corpus. Foi então feita uma acareação dos *scores* individuais dessas frases traduzidas a partir do corpus construído, conforme Tabela 1.

Tabela 1 – Comparação dos scores sacreBLEU atribuídos às frases geradas pelo modelo treinado e pelo Google Tradutor

Chave	Score sacreBLEU Modelo Treinado	Score sacreBLEU Google Tradutor
316	32,5	32,6
384	29,5	58,1
635	71,9	50,7
796	22,8	23,2
852	8,3	6,3
950	20,6	12,6
965	21	63,9
1013	82,4	91,2
1166	7,3	7,1
1377	22	26,3
1390	33	7,8
1399	49,2	11,4
1411	49,8	36,1
1418	23,4	10
1437	5,7	14,1
1451	14,4	9,4
1455	29	24,6
1471	37,5	33,5
1486	41,4	34,8
1520	31.9	27,2
Média	31,6684211	29,045

Fonte: Autor

A partir dos resultados obtidos pela avaliação automática, pode-se perceber que o modelo treinado ainda é capaz de obter *scores* maiores em relação ao modelo de tradução disponibilizado pela Google. O resultado da média dos *scores* obtidos dentre as traduções geradas pelo modelo treinado foi de aproximadamente 31,66. Em contrapartida, o resultado da média dos *scores* obtidos dentre as traduções geradas pelo Google Tradutor foi de 29,045. A diferença não é grande, mas dificilmente seria, levando em consideração o tamanho do corpus paralelo utilizado para o treinamento do modelo de tradução automática. No entanto, essa pequena diferença já indica um passo no caminho certo para o desenvolvimento de um trabalho que realmente será capaz de oferecer melhores pontuações.

Apesar de o foco desta pesquisa ser o treinamento de um modelo de tradutor automático baseado em Transformers, uma breve análise das traduções geradas demonstra-se relevante para a compreensão geral do *score* sacreBLEU do modelo gerado. Em relação à análise humana também dessas 20 traduções realizadas, este trabalho opta por seguir os parâmetros de adequação e fluência estabelecidos por Banitz (2020) e as categorias de erro identificadas em

traduções automáticas delineadas por Vilar *et al*. (2006). Dentro dessas categorias estipuladas pelo trabalho dos últimos autores, o único erro verificado nas traduções geradas automaticamente, tanto pelo modelo treinado quanto pelo Google Tradutor, foi o erro relativo à geração de traduções que contivessem palavras incorretas, alterando o sentido da frase por meio de uma escolha incorreta de léxico (Vilar *et al*.,2006).

Após a análise humana das traduções geradas baseada nos parâmetros descritos no parágrafo anterior, pode-se afirmar que, apenas duas traduções geradas apresentaram erros, sendo uma delas gerada pelo modelo treinado e a outra pelo Google Tradutor. Todas as outras não apresentam erros relevantes que influenciem na adequação e fluência da frase, podendo serem consideradas boas traduções.

Focando na análise então das duas frases que realmente não atingira um nível de adequação terminológico satisfatório, trataremos da primeira. Observando o Quadro 3 abaixo, pode-se perceber que, a partir da frase em inglês, a tradução gerada pelo modelo *Transformer* treinado, foi altamente adequada e fluente. Houve sim um erro (Vilar *et al*., 2006) relativo à ausência do artigo indefinido *une*, porém que não afeta a fluência e adequação do restante da frase (Banitz, 2020). Essa foi a única diferença que impediu que a frase traduzida ficasse idêntica à frase francesa oficial usada como referência (presente no corpus paralelo). Essa frase recebeu a pontuação de 32,5.

Quadro 3 – Resultado das traduções do modelo treinado e do Google Tradutor

Frase EN para tradução	Frase FR Referência	Resultado Tradução Modelo	Resultado Tradução Google
Convention of cooperation among authorities	Une convention de coopération entre autorités	Convention de coopération entre les autorités	Convention de coopération entre collectivités

Fonte: Autor

Por outro lado, a tradução do Google Tradutor gerou um resultado considerado fluente, porém inadequado para o contexto, uma vez que utiliza o termo "*collectivités*" para se referir aos órgãos federais de cooperação internacional relativa à Convenção Internacional da Haia de 1980 – as Autoridades Centrais. No entanto, o termo "*collectivité*" faz menção a departamentos administrativas reconhecidas na França, consideradas entidades territoriais coletivas, com poder de governo estabelecido. Essa terminologia engloba qualquer área que possua uma forma de governo local eleito e autoridade reguladora local e não faz referência à Autoridade Central, órgão incumbido da adoção de medidas para o cumprimento das obrigações impostas pela Convenção de Haia de 1980.

Utilizando a classificação de Vilar *et al*. (2020), este erro cometido pelo modelo de tradução do Google é claramente uma escolha lexical errada, afetando toda a adequação da tradução (Banitz, 2020). Ainda assim, essa tradução gerada pelo Google Tradutor recebeu uma pontuação de 32,6, infimamente maior que a tradução mais correta gerada pelo modelo treinado.

No entanto, o modelo treinado com base em *corpus* especializado também não foi sem erros e a segunda tradução que contém um erro, dentre as 20 anteriormente mencionadas, foi produto de seu processamento. Como demonstrado no Quadro 4, o modelo traduziu erroneamente a palavra "*Requesting*" para "*requises*", trocando totalmente o sentido da frase em inglês.[28] Assim, apesar de mantida a fluência da frase, não há adequação do termo utilizado (Banitz, 2020) e a escolha por esse léxico errôneo distorce o sentido da frase no idioma de entrada, podendo ser caracterizado como um erro proveniente de uma escolha errada de léxico (Vilar *et al*. 2006). O *score* sacreBLEU obtido para essa tradução foi de 5,7.

Quadro 4 - Resultado das traduções do modelo treinado e do Google Tradutor

Frase EN para tradução	Frase FR Referência	Resultado Tradução Modelo	Resultado Tradução Google
Requesting Central Authorities are often under pressure from applicants (usually left-behind parents) to provide daily reports of progress	Les demandeurs (généralement les parents privés de leur enfant) exercent souvent des pressions sur les Autorités centrales pour qu'elles leur fournissent des rapports de suivi journaliers	Les Autorités centrales requises sont souvent soumises à des pressions de la part des demandeurs (habituellement des parents laissés derrière eux) pour fournir des rapports quotidiens sur les progrès accomplis	Les Autorités centrales requérantes subissent souvent des pressions de la part des demandeurs (généralement des parents délaissés) pour qu'elles fournissent des rapports quotidiens sur les progrès

Fonte: Autor

Ainda olhando para o Quadro 4, em relação a tradução do Google dessa mesma frase, essa foi igualmente fluente, mas realmente mais adequada semanticamente ao contexto. Ela utilizou corretamente o termo "*requérantes*" para se referir às Autoridades Centrais Requerentes (*Requesting*); termo também utilizado na frase-referência oficial em francês. O *score* acompanhou essa lógica correta e atribuiu, à tradução automática do google, 14,1 pontos.

Após a análise de todos esses resultados gerados, é possível perceber que um modelo treinado com apenas 5.494 pares de frases já consegue desempenhar tarefas de tradução automática baseada em vocábulo especializado com um nível um pouco melhor que tradutores

[28] Importante aqui ressaltar que a Autoridade Central pode figurar tanto como Requerente (*Requérant*/*Requesting*) como Requerida (*Requise*/*Requested*). Esse título é importante para identificar onde o processo se inicia e qual o papel dessas Autoridades para a recuperação da criança ou adolescente.

automáticos disponíveis, quando levado em consideração a análise de resultados de avaliação automática.

Em relação aos erros apontados na avaliação humana, ambos os modelos de tradução, seja ele treinado ou não, estão suscetíveis a falhas e cabe melhorar o modelo *Transformer* treinado, onde der, para que seus resultados sejam ainda melhores. A revisão do *corpus*, o aumento no número de pares de frases alinhadas, o investimento em máquinas e técnicas de processamento de dados mais avançadas, a dedicação de maior tempo para o treinamento do modelo e o teste de mais alguns hiperparâmetros ajustáveis no momento de treinar o modelo são só alguns dos fatores concretos apontados por esse trabalho que garantem uma melhora no modelo e, consequentemente, nas traduções geradas.

5 CONSIDERAÇÕES FINAIS

Este artigo teve como objetivo geral a análise sobre os modelos de tradução automática baseados em transformadores. Os resultados aqui obtidos demonstram que essa análise, em geral, foi realizada a contento, sendo ainda possível, por meio deste trabalho, aplicar métodos e técnicas desenvolvidos ao longo de toda a graduação do curso de Línguas Estrangeiras Aplicadas ao Multilinguismo e Sociedade da Informação da Universidade de Brasília.

Em um de seus objetivos específicos, o trabalho buscou o teste da viabilidade do uso de modelos de tradução automática baseados em *Transformers*. A partir disso, foi possível perceber que a implementação do modelo *Transformer*, foi bastante facilitada pelo uso de código aberto e de bibliotecas disponibilizadas de forma gratuita pela plataforma *HuggingFace*, fatores que contribuíram imensamente para a etapa inicial de escrita do código utilizado.

No entanto, em etapas mais avançadas da implementação do modelo, este trabalho encontrou dificuldades na tarefa específica de treinamento do modelo, muito disso devido aos limites computacionais de capacidade de processamento de dados pela máquina utilizada. Como demonstrado, modelos de tradução automática baseados em *Transformers* que contam com bons *scores*, apesar de melhor aproveitarem a capacidade de processamento da máquina, ainda necessitam de uma quantidade bem grande de dados para que possam ser treinandos e, consequentemente, de máquinas que possam lidar de forma eficiente com esses dados. Com menor capacidade de processamento, o modelo pré-treinado utilizado T5-*small*, os hiperparâmetros utilizados e o tempo necessário para o treinamento se tornaram grandes limitadores para o desenvolvimento de um melhor modelo de tradução automática.

Não obstante, este trabalho iniciou o desenvolvimento de um corpus paralelo inglês-francês específico relativo à Convenção da Haia de 1980, sendo desenvolvido um conjunto de dados eminentemente promissor que serve de referência para pesquisas futuras e aplicações em trabalhos que busquem desenvolver técnicas semelhantes. O trabalho de organização, limpeza e anotação dos textos selecionados aqui realizado e documentado demonstram o rigor metodológico empreendido neste trabalho que, por sua vez, fornece os alicerces necessários para a revisão e expansão do corpus relativo à Convenção da Haia de 1980.

O trabalho ainda serviu como forma de avaliação da medida na qual a construção de um corpus específico e sua implementação em um modelo de tradução se torna mais satisfatória do que uma tradução feita em um modelo com léxico não especializado. Aqui é importante ressaltar o fator das dimensões envolvidas no trabalho. Como demonstrado, quanto maior a quantidade de texto disponível para alimentar o treinamento modelo, maior a qualidade dos resultados esperada. Dessa forma, apesar de ter seu treinamento baseado em um corpus que conta com somente 5.494 pares de palavras, o modelo de tradução desenvolvida se demonstra extremamente eficiente e performa, na maioria das vezes, melhor do que um modelo de tradução não treinado.

Por fim, considera-se que os resultados aqui relatados satisfatórios se encontram solidamente alinhados com aquele produzidos por pesquisas semelhantes, com real possibilidade de melhora a partir da revisão de fatores especificamente apontados ao longo do trabalho.

BIBLIOGRAFIA

BANITZ, Brita. Machine translation: a critical look at the performance of rule-based and statistical machine translation. **Cadernos de Tradução**, [S.L.], v. 40, n. 1, p. 54-71, 22 jan. 2020. Universidade Federal de Santa Catarina (UFSC). Disponível em: http://dx.doi.org/10.5007/2175-7968.2020v40n1p54. Acesso em: 12 jul. 2023.

DOUGLAS, Arnold *et al*. **Machine translation**: An introductory guide. Oxford: Blackwell, 1994.

FANTINUOLI, Claudio; ZANETTIN, Federico. Creating and using multilingual corpora in translation studies. **New directions in corpus-based translation studies**, p. 1-10, 2015. Disponível em: https://langsci-press.org/catalog/view/76/67/277-1. Acesso em: 13 jul. 2023.

HALLEBEEK, Jos. English parallel corpora and applications. **Cuadernos de Filologia Inglesa**, v. 9, n. 1, 2000. Disponível em: https://digitum.um.es/xmlui/bitstream/10201/1714/1/112499.pdf. Acesso em: 13 jul. 2023.

HAN, Lifeng. **An overview on machine translation evaluation**. arXiv preprint arXiv:2202.11027, 2022. Disponível em: https://arxiv.org/pdf/2202.11027. Acesso em: 13 jul. 2023.

HCCH. Conferência da Haia de Direito Internacional Privado. Publications. 28: **Convention of 25 October 1980 on the Civil Aspects of International Child Abduction**. Disponível em: http://www.hcch.net/index_en.php?act=conventions.publications&dtid=1&cid=24. Acesso em: 12 jul. 2023.

HCCH. Conferência da Haia de Direito Internacional Privado. Publications. **The Guide to Good Practice under the Hague Convention of 25 October 1980 on the Civil Aspects of International Child Abduction**: Part I – Central Authority Practice. Disponível em: https://www.hcch.net/en/instruments/conventions/publications1/?dtid=3&cid=24. Acesso em: 12 jul. 2023.

HCCH. Conferência da Haia de Direito Internacional Privado. Publications. **The Guide to Good Practice under the Hague Convention of 25 October 1980 on the Civil Aspects of International Child Abduction**: Part II – Implementing Measures. Disponível em: https://www.hcch.net/en/instruments/conventions/publications1/?dtid=3&cid=24. Acesso em: 12 jul. 2023.

HCCH. Conferência da Haia de Direito Internacional Privado. Publications. **The Guide to Good Practice under the Hague Convention of 25 October 1980 on the Civil Aspects of International Child Abduction**: Part III - Preventive Measures. Disponível em: https://www.hcch.net/en/instruments/conventions/publications1/?dtid=3&cid=24. Acesso em: 12 jul. 2023.

HCCH. Conferência da Haia de Direito Internacional Privado. Publications. **The Guide to Good Practice under the Hague Convention of 25 October 1980 on the Civil Aspects of International Child Abduction**: Part IV - Enforcement. Disponível em: https://www.hcch.net/en/instruments/conventions/publications1/?dtid=3&cid=24. Acesso em: 12 jul. 2023.

HCCH. Conferência da Haia de Direito Internacional Privado. Publications. **The Guide to Good Practice under the Hague Convention of 25 October 1980 on the Civil Aspects of International Child Abduction**: Part V - Mediation. Disponível em: https://www.hcch.net/en/instruments/conventions/publications1/?dtid=3&cid=24. Acesso em: 12 jul. 2023.

IOSIFOVA, Olena *et al*. *Techniques comparison for natural language processing*. **MoMLeT&DS**, v. 2631, n. I, p. 57-67, 2020. Disponível em: https://core.ac.uk/reader/328802590. Acesso em: 12 jul. 2023.

JABBAR, H.; KHAN, Rafiqul Zaman. Methods to avoid over-fitting and under-fitting in supervised machine learning (comparative study). **Computer Science, Communication and Instrumentation Devices**, v. 70, n. 10.3850, p. 978-981, 2015. Disponível em: https://www.researchgate.net/publication/295198699. Acesso em: 12 jul. 2023.

KALYANI, Aditi *et al*. Evaluation and ranking of machine translation output in Hindi language using precision and recall oriented metrics. **International Journal of Advanced**

Computer Research, 4.14. 2014. p. 54-59. Disponível em: https://arxiv.org/pdf/1404.1847. Acesso em: 12 jul. 2023.

KIMERA, Richard *et al*. **Building a Parallel Corpus and Training Translation Models Between Luganda and English**. arXiv preprint arXiv:2301.02773, 2023. Disponível em: https://arxiv.org/pdf/2301.02773. Acesso em: 12 jul. 2023.

KOEHN, Philipp. Statistical machine translation. **Cambridge University Press**, 2009.

LAKEW, Surafel M. *et al*. **A comparison of transformer and recurrent neural networks on multilingual neural machine translation**. arXiv preprint arXiv:1806.06957, 2018. Disponível em: https://arxiv.org/pdf/1806.06957. Acesso em: 12 jul. 2023.

PAPINENI, Kishore *et al*. Bleu: a method for automatic evaluation of machine translation. In: **Proceedings of the 40th annual meeting of the Association for Computational Linguistics**. 2002. p. 311-318. Disponível em: https://aclanthology.org/P02-1040.pdf. Acesso em: 13 jul. 2023.

PÉREZ-VERA, Elisa. Explanatory Report on the 1980 Hague Child Abduction Convention. In: **Acts and Documents of the Fourteenth Session** (1980), tome III, Child abduction. 1982. Hague: HCCH Publications, 1981. Disponível em: https://www.hcch.net/en/publications-and-studies/details4/?pid=2779. Acesso em: 12 jul. 2023.

PIRES, Thiago Blanch. **Ampliando olhares sobre a tradução automática online**: um estudo exploratório de categorias de erros de máquina de tradução gerados em documentos multimodais. Tese de doutorado apresentada à Universidade de Brasília. Brasília, 2017. Disponível em: https://repositorio.unb.br/bitstream/10482/23727/1/2017_ThiagoBlanchPires.pdf. Acesso em: 27 jul. 2023

POPEL, Martin; BOJAR, Ondřej. **Training tips for the transformer model**. arXiv preprint arXiv:1804.00247, 2018. Disponível em: https://arxiv.org/pdf/1804.00247. Acesso em: 13 jul. 2023.

POST, Matt. **A call for clarity in reporting BLEU scores**. arXiv preprint arXiv:1804.08771, 2018. Disponível em: https://arxiv.org/pdf/1804.08771. Acesso em: 12 jul. 2023.

RAFFEL, Colin *et al*. Exploring the limits of transfer learning with a unified text-to-text transformer. **The Journal of Machine Learning Research**, v. 21, n. 1, p. 5485-5551, 2020. Disponível em: https://arxiv.org/pdf/1910.10683. Acesso em: 12 jul. 2023.

TIAN, Taoling *et al*. A French-to-English machine translation model using transformer network. **Procedia Computer Science**, v. 199, p. 1438-1443, 2022. Disponível em: https://www.sciencedirect.com/science/article/pii/S1877050922001831/pdf?md5=de6f0f9663197d32d3160118cd31791c&pid=1-s2.0-S1877050922001831-main.pdf. Acesso em: 12 jul. 2023.

VASWANI, Ashish *et al*. Attention is All you Need. In: **Advances in Neural Information Processing Systems**. [S.l.]: Curran Associates, Inc., 2017. v.30. Disponível em: https://arxiv.org/abs/1706.03762. Acesso em: 12 jul. 2023

VILAR, D. *et al*. Error analysis of statistical machine translation output. In: **LREC**, 2006, Genoa. Proceedings of LREC. Genoa, 2006. Disponível em: http://www.lrec-conf.org/proceedings/lrec2006/pdf/413_pdf.pdf. Acesso em: 12 jul. 2023.

WHITE, J.S. How to evaluate machine translation. In: **Computers and Translation**: A translator's guide. Amsterdam/Philadelphia: John Benjamins Publishing, 2003. v. 35p. 211–244.

WOLF, Thomas *et al*. **Huggingface's transformers**: State-of-the-art natural language processing. arXiv preprint arXiv:1910.03771, 2019. Disponível em: https://arxiv.org/pdf/1910.03771. Acesso em: 12 jul. 2023.

WU, Yonghui *et al*. **Google's neural machine translation system**: Bridging the gap between human and machine translation. arXiv preprint arXiv:1609.08144, 2016. Disponível em: https://arxiv.org/pdf/1609.08144. Acesso em: 12 jul. 2023.