Experimentação e análise comparativa dos modelos T5, mBart, MarianMT e MarianMT ROMANCE para tradução automática de reportagens científicas

Lucas M. D. Brum[1], Guilherme R. Rodrigues[2], Santiago Lühring[3]

[1]Instituto de Informática – Universidade Federal do Rio Grande do Sul (UFRGS)
Caixa Postal 15.064 – 91.501-970 – Porto Alegre – RS – Brazil

{lucas.brum,grrodrigues,slbluhring}@inf.ufrgs.br

Abstract. *This article details the experimentation and comparative performance analysis of the generative models T5, mBart, and MarianMT (two different models) for the automatic translation of sentences contained in research texts from the São Paulo Research Foundation's(FAPESP) magazine from English to Portuguese, using ROUGE, METEOR, BLEU metrics, and human evaluation.*

Resumo. *Este artigo detalha a experimentação e análise comparativa de desempenho entre os modelos generativos T5, mBart e MarianMT (dois modelos diferentes) para tradução automática de sentenças contidas em textos de pesquisa da revista da Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) do inglês para o português utilizando métricas ROUGE, METEOR, BLEU e avaliação humana.*

1. Introdução

A tradução automática (TA) é uma ferramenta muito eficiente para a transição entre idiomas, sejam elas parecidas ou não, envolvendo interpretação de uma língua de partida e a geração de uma língua de chegada. Para nós, seres humanos, certos elementos entre quaisquer duas linguagens utilizadas pelo mundo podem parecer muito semelhantes e fazer com que máquinas experimentem e executem isto pode ser algo bem interessante. Segundo [Martins 2008], a TA surgiu de forma concreta pela primeira vez em 1949, ainda que a mesma tenha sido discutida pelo menos desde o século XVII. Neste trabalho, incorpora-se o uso da tradução automática para averiguar a eficiência na tradução de textos de reportagens da Fundação de Amparo à Pesquisa do Estado de São Paulo[11] com quatro modelos, são eles, *mBart*, T5, e dois modelos *MarianMT* (Romance e Big), avaliando a precisão e a qualidade semântica das traduções geradas por esses modelos e também, com auxílio da avaliação humana para tais. A qualidade das traduções será medida pelas métricas BLEU, METEOR e ROUGE. A métrica BLEU foi usada para avaliar a correspondência exata entre os n-gramas da tradução gerada e da referência, enfatizando a precisão das traduções. Já ROUGE nos permitirá avaliar a

[1] Disponível em https://revistapesquisa.fapesp.br/

cobertura do conteúdo, ajudando a captar o quão bem as traduções preservam o sentido do texto original. Avaliamos também o tempo de execução de cada modelo na realização da tradução.

Nosso trabalho levou cerca de 50 minutos para processar um artigo de 143 sentenças e 38 minutos para outro de 107 sentenças. Assim, gerou texto através dos 4 modelos. Obtivemos resultados que se mostraram promissores, medianamente falando, porém para o mBart percebemos métricas mais modestas (quanto a BLEU principalmente), além de um tempo muito maior para processamento, tradução e cálculo de suas métricas.

2. Fundamentação Teórica

Do amplo campo de Processamento de Linguagem Natural, foram utilizados alguns conceitos importantes para fundamentar este trabalho. Conceitos técnicos como tokenização e a avaliação humana foram empregados para aferir o desempenho de cada modelo.

2.1. Conceitos técnicos/teóricos

A seguir, apresentamos os conceitos técnicos e teóricos que embasam e sustentam a proposta de nosso estudo. Explicamos os conceitos de transformer, dataset, sentença, tokenização, avaliação humana, e N-grama, que foram elementos essenciais para a execução do presente trabalho.

2.1.1. Sentença

A sentença é a unidade gramatical completa composta por uma ou mais palavras, expressando uma ideia. Comumente pontuada e estruturada compondo algo que faça sentido para o ser humano.

2.1.2. Transformer

O *Transformer* é uma arquitetura recente que visa resolver tarefas de forma sequencialmente enquanto lida com dependências de longo alcance com facilidade. Depende inteiramente de um mecanismo chamado autoatenção, que permite que o modelo avalie diferentes posições de uma mesma sequência ao calcular a representação de uma palavra. No contexto do Transformer, a autoatenção é utilizada nas camadas do codificador e do decodificador, possibilitando que o modelo identifique dependências e relações de longo alcance na sequência de entrada, utilizado para calcular

representações de sua entrada e saída, permitindo que capturem relações complexas em sequências de texto.

2.1.2. Dataset

O dataset é um conjunto de dados estruturados, organizado geralmente em tabelas ou vetores, usados para treino ou análise em modelos. Neste trabalho utilizamos um artigo estruturado em sentenças dispostas uma em cada linha, de forma que o pré-processamento seja eficiente em uma leitura de arquivo linha por linha por uma rotina definida.

2.1.3. Tokenização

A tokenização é um conceito de PLN dentro de um procedimento chamado pré-processamento que trata da separação de palavras através de separadores, comumente espaços em branco em palavras menores denominadas *tokens.* Apesar de não ser trivial a depender da linguagem, trata-se de uma etapa obrigatória para o pré-processamento, procedimento fundamental para a execução de testes de modelos preditivos.

2.1.4. Avaliação humana

A avaliação humana dos textos gerados por modelos é uma tática onde o usuário avalia um texto ou sentença gerado por um modelo qualquer comparando seu texto com um texto de referência e aferindo sua coerência e eficiência na geração.

2.1.5. N-Grama

Um N-Grama é uma sequência de elementos adjacentes em uma cadeia. Por exemplo, em uma frase, um bigrama (2-grama) considera pares de palavras consecutivas. Um modelo de n-grama é um tipo de modelo de linguagem probabilístico utilizado para prever o próximo elemento em uma sequência, baseando-se no princípio de um modelo de

Markov.

Estes modelos são amplamente empregados em diversas áreas, como probabilidade, teoria da comunicação, linguística computacional (por exemplo, no processamento estatístico de linguagem natural), biologia computacional (como na análise de sequências biológicas) e compressão de dados. Dois dos principais benefícios dos modelos de n-grama, e dos algoritmos que os utilizam, são sua simplicidade e escalabilidade. Com o aumento do valor de n, o modelo pode armazenar mais contexto, proporcionando um equilíbrio bem compreendido entre espaço e tempo, o que permite escalar experimentos menores de maneira eficiente.

2.2. Descrição dos modelos utilizados

Todos estes modelos foram obtidos no site do *Hugging Face*[2], que é uma plataforma de código aberto, que democratiza o acesso ao aprendizado de Inteligência Artificial e Aprendizado de Máquina, onde é possível compartilhar códigos e repositórios que consistam em projetos, bem como pesquisar outros projetos para se inspirar. Entre projetos gratuitos e outros os quais são necessários pagamento de uma API para uso, podemos encontrar modelos pré-treinados e alimentá-los com outros dados a fim de executar experimentos que aferem suas métricas de eficiência quanto ao conjunto de dados escolhidos. Para este trabalho, foram escolhidos os seguintes modelos.

2.2.1. mBart

O mBart utilizado é o *facebook/mbart-large-50-many-to-many-mmt*, uma fine-tuning de *facebook/mbart-large-50*. Este modelo pode traduzir diretamente entre quaisquer pares de 50 idiomas.

Foi introduzido em *Multilingual Translation with Extensible Multilingual Pretraining and Finetuning,2020*

mBart é uma extensão multilingue do BART e é ideal para cenários de tradução envolvendo múltiplos pares de línguas, especialmente em tarefas onde a qualidade da geração de texto e o contexto global são críticos.

2.2.2. T5

O modelo T5 utilizado é o *unicamp/translation-en-pt-t5*. Performa traduções entre inglês e português, neste sentido (EN-PT). É fornecido pela Universidade Estadual de Campinas (UNICAMP). Foi introduzido em *Lite Training Strategies for Portuguese-English and English- Portuguese Translation,2020.*

Este modelo demanda a introdução de “translate English to Portuguese:” no começo de cada sentença a ser traduzida.

2.2.3. MarianMT

OpusMTs são modelos neurais de Machine Translation desenvolvidas sobre MarianMT framework, uma ferramenta open source para sistemas NMT, foi criada pela Microsoft Research. O projeto OPUS é fruto da Universidade de Helsinque, com o modelo base treinado sobre corpora de mesmo nome. Aqui utilizamos dois destes modelos.

Marian MTs são modelos especializados, desenvolvidos para tradução direta e eficiente entre pares específicos de línguas.

Os modelos OPUS, fornecidos pela Universidade de Helsinque, são uma série de modelos neurais de MT e parte de um projeto cujo objetivo é torná-los mais acessíveis e amplamente disponíveis para muitas línguas.

Helsinki-NLP/opus-mt-en-ROMANCE

Este é um OpusMT que permite a tradução de inglês para línguas românicas, aquelas descendentes do latim (*e.g. italiano, espanhol, português europeu e brasileiro*). O modelo em tese se utiliza da relação entre elas para maximizar a transferência de conhecimento e melhorar a qualidade das traduções.

Este modelo demanda a inserção de um token (>>pt_BR, por exemplo) no começo de cada sentença a ser traduzida.

Helsinki-NLP/opus-mt-tc-big-en-pt

O tc-big é um modelo otimizado para tradução direta entre inglês e português, possuindo arquitetura maior e capaz de capturar mais contexto do que a variante ROMANCE. Em contrapartida, o modelo tem maior demanda computacional.

Introduzido em *OPUS-MT-Building open translation services for the World,2020.*

2.3. Descrição das métricas utilizadas

Nesta seção exploraremos as métricas que ajudaram a elucidar a eficiência dos modelos descritos anteriormente; são elas: ROUGE, BLEU e METEOR.

2.3.1. ROUGE

ROUGE (*Recall-Oriented Understudy for Gisting Evaluation*) é um conjunto de métricas úteis para tradução automática e para sumarização. As métricas trabalham comparando resumo ou tradução produzido automaticamente com um resumo ou tradução de referência (de alta qualidade). ROUGE compara a tradução ou o resumo gerado automaticamente com um ou mais resumos e/ou traduções de referência, buscando medir a sobreposição de n-gramas entre eles.

Neste trabalho, utilizamos o ROUGE-L, pois este avalia a LCS (*Longest Common Sentenc*e), a maior sequência em comum entre a frase de referência e a candidata para tradução, capturando a estrutura e coerência gramatical da tradução, sendo uma boa métrica para avaliar a sintaxe da predição gerada.

2.3.2. BLEU

BLEU (*Bilingual Evaluation Understudy*), introduzido em 2002 no artigo *BLEU: A method for automatic evaluation of machine translation*. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (*Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J.)* é uma métrica que compara uma tradução candidata a uma ou mais traduções de referência, medindo a sobreposição de n-gramas entre elas. Embora desenvolvido para tradução, BLEU também serve para avaliar o texto gerado para várias outras tarefas de Processamento de Linguagem Natural, como paráfrase e sumarização. A métrica não é perfeita, enfrentando dificuldades para avaliar traduções que utilizam estrutura e vocabulários diferentes do texto referência. Além disso, o BLEU pode não refletir com precisão aspectos semânticos e fluência das traduções, pois se concentra exclusivamente em n-gramas exatos. Mesmo assim, o BLEU se correlaciona bastante com avaliações humanas, o que o torna útil para a avaliação automatizada de grandes volumes de dados. Através de um cálculo um tanto complicado e em etapas, BLEU elabora mapeamentos de unigrama detalhando como alinhar uma frase traduzida (candidata) com uma frase de referência, calcular métricas de precisão e recall unigrâmico, ajustar os resultados considerando n-gramas maiores e aplicar uma penalidade baseada em descontinuidades no alinhamento

2.3.3. METEOR

METEOR (Metric for Evaluation of Translation with Explicit Ordering) é uma métrica desenvolvida para superar as limitações do BLEU, especialmente no que diz respeito à incorporação de variabilidade linguística, como sinonímia, lematização e reordenação das palavras. A principal diferença do METEOR em relação ao BLEU é que, enquanto o BLEU se concentra em n-gramas exatos, o METEOR busca uma correspondência mais flexível e humana, levando em consideração não apenas a precisão e o recall de unigrama, mas também a semelhança semântica entre palavras, além de penalidades por reordenação de palavras. A abordagem do METEOR é valiosa pois avalia não somente a semântica da tradução automática mas também sua estrutura gramatical. A contrapartida é que METEOR é mais complexo que outras métricas e por consequência computacionalmente mais caro. Introduzido em *METEOR: An automatic metric for MT evaluation with improved correlation with human judgments (Banerjee, S., & Lavie, A. ,2005).*

3. Trabalhos Relacionados

Nesta seção apresentamos a seguir alguns trabalhos relacionados que são relevantes e ofereceram contribuições importantes para o presente estudo.

3.1. *Tradução Automática de Abstracts: avaliação do potencial e das limitações de três ferramentas da web*

Neste trabalho, o autor José Paulo de Araújo pesquisa três ferramentas da internet para executar tradução automática de *abstracts* e avaliar sua utilidade para tal feito. O processo de seleção de bibliografia para trabalhos acadêmicos é dificultado pela barreira linguística, especialmente para estudantes que não dominam o inglês. A solução tradicional de tradução humana é onerosa, e o trabalho propõe investigar ferramentas gratuitas de tradução automática como uma alternativa prática para lidar com abstracts em língua estrangeira.

Essa abordagem é relevante para a democratização do acesso à produção acadêmica e pode reduzir desigualdades em contextos onde o inglês é um requisito essencial, mas nem sempre acessível para todos.

Segundo Araújo, o processo não é preciso ainda e frequentemente o texto traduzido, o texto-alvo, precisa ser revisado por um tradutor ou pelo menos por um usuário nativo ou fluente na língua-alvo. A TA pode ser especialmente útil quando se deseja um resultado que necessite de pouca ou nenhuma revisão posterior por um ser humano.

3.2. *Treinamento e análise de um modelo de tradução automática baseado em Transformer*

Neste artigo de autoria de Clovis Pimentel e Thiago Blanch Pires, um modelo de tradução baseados em Transformer foi treinado com um corpus especializado, composto de sete textos paralelos inglês-francês da Convenção de 25 de outubro de 1980 sobre os Aspectos Civis do Rapto Internacional de Crianças. Seu intuito foi a comparação da tradução de textos traduzidos neste modelo com aquela realizada pelo Google Translate. A avaliação foi feita utilizando métodos *sacreBLEU* e também avaliação humana. Este trabalho se assemelha ao nosso por utilizar da tradução automática e o baseando em Transformer assim como o fizemos.

4. Base de dados

A base de dados escolhida para o trabalho foi um *corpus* da ”Revista Pesquisa Fapesp Parallel Corpora”, advinda de *Fully Automatic Compilation of a Portuguese-English Parallel Corpus for Statistical Machine Translation.* O *corpus* consiste em quatro textos, duas reportagens da revista em inglês e estas mesmas reportagens, mas em português.

As versões em inglês são usadas para a tradução e geram, após a realização desta, um conjunto de sentenças que são comparadas com a versão em português e as métricas calculadas.

Uma reportagem é ”*The Bill on Biosafety is going to boost Brazilian research, which was already strong in the area*”de Marcos Piveta e ela possui 143 sentenças.

Outra é *“The university reform proposal changes bureaucratic structures without defining what the country wants from these institutions”* de Fabrício Marques e possui 107 sentenças.

5. Metodologia do trabalho

5.1. Entendimento do problema

Com o fascínio por tradução automática como das áreas mais importantes de PLN, decidimos aplicar a uma base de dados que relata pesquisas da FAPESP para analisar como modelos trabalham com diversos outros *corpus*, nunca experimentados antes e aferir uso de métricas diferentes e importantes para a execução de trabalhos da área.

5.2. Objetivo

O objetivo do presente trabalho é aferir a eficiência de modelos de tradução automática que sejam capazes de executar tradução automática testando outros *corpus*. Os modelos *mBart*, *MarianMT Romance* e *MarianMT Big* e T5 foram escolhidos por possuírem renome na literatura referente a Processamento de Linguagem Natural.

5.3. Organização dos dados

Em um arquivo com várias reportagens, escolhemos duas (descritas em 4) com menos de 150 sentenças, uma por linha.

Estas sentenças foram separadas em blocos de 20 para serem processadas. A execução das métricas em blocos menores, ao invés de executar no texto como um todo, se deu para melhorar, manter e poder perceber a coesão entre elementos mais próximos.

5.4. Pré-processamento

Na etapa de pré-processamento, colhemos dois dos arquivos dispostos em formato .txt e o 'tokenizamos' para separá-los em sentenças e executar os modelos pré-treinados nos mesmos. A tokenização funciona usando uma estratégia complexa, pois cada modelo possui uma metodologia diferente. Os tokenizadores *Marian*, T5 e *mBART* usam vocabulário específico do modelo e esquemas de tokenização de subpalavras, como Byte Pair Encoding (BPE) ou *SentencePiece*, otimizados para seus respectivos dados e arquiteturas de treinamento. Alguns modelos, como o mBART, inserem tokens específicos do idioma, enquanto o T5 usa formatação de entrada baseada em prefixo ()."translate English to Portuguese: " + sentence. Tokenização em nível de frase para métricas como BLEU e METEOR utiliza nltk.word_tokenize, que é independente de modelos.

O processo da tokenização em nível de modelo depende do tokenizador específico da biblioteca *HuggingFace* transformers, variando com base na arquitetura e método de treinamento do modelo. Essa separação garante flexibilidade, mas também significa que a tokenização pode impactar as métricas dependendo do alinhamento entre o método de tokenização e o treinamento do modelo.

Para melhor aferição das médias de métricas escolhidas, os dados foram divididos em blocos de 20 sentenças, totalizando, em uma reportagem, 7 blocos de 20 sentenças e 1 de 3 sentenças e, noutra reportagem, 5 blocos 20 de sentenças e 1 de 7 sentenças.

5.5. Avaliação Humana

A avaliação humana foi essencial para complementar as métricas automáticas utilizadas. Os critérios estabelecidos para avaliar os modelos foram: Fluência, isto é, quão natural e gramaticalmente correta é a tradução; Adequação, quanto do significado do texto foi preservado na tradução; e Consistência, se a tradução preserva o uso de termos e estilo ao longo do texto.

Avaliamos as traduções dos modelos buscando detectar erros e comparando as traduções par a par, escolhendo a(s) melhor(es) dentre as quatro de cada "par".

6. Experimentos e resultados

Nesta seção exploraremos como foi feito todo o processo de experimentação e aferição dos resultados das traduções preditas pelos modelos, bem como breves discussões sobre os mesmos

6.1. Softwares utilizados e recursos

Nesta seção apresentamos os programas que utilizamos para executar as análises e o ambiente o qual todo o procedimento foi de fato executado, incluindo linguagem, dinamicidade do projeto quanto aos envolvidos, pacotes importantes e relevantes que viabilizaram a execução do trabalho.

6.1.1. Linguagem Python

Utilizamos a linguagem Python[1], que é referência para aprendizado de máquina e também para processamento de linguagem natural pelo alto aparato que oferece para as duas áreas.

Dentro da linguagem Python, diversas bibliotecas e diretivas foram utilizadas entre elas NTLK, ROUGE, transformers.

6.1.2. Controle de versionamento com GitHub

Para melhor assessoramento ao código, utilizamos do GitHub[2], que permite que haja mobilidade do projeto em qualquer dispositivo que se queira trabalhar. Favorável também para que outros integrantes do projeto também possam trabalhar simultaneamente no mesmo.

6.1.3. Microsoft VS Code

O programa utilizado para executar os procedimentos foi o Microsoft Visual Studio Code[3], pela facilidade, versatilidade e leveza para execução de qualquer linguagem e projeto que se queira trabalhar.

---

[2] Disponível em https://github.com/psicosleazy-dev/pln
[3] Disponível em https://code.visualstudio.com/

6.2. Análise quantitativa

Nesta seção, estão apresentados os resultados obtidos com a metodologia apresentada anteriormente, bem como dados quantitativos e análises.

A seguir, estão apresentadas as tabelas que correspondem às métricas escolhidas para avaliar a eficiência dos modelos, trazendo os *scores* obtidos. Na tabela 1, vemos que o METEOR atingiu valores relativamente altos para para ambas as reportagens e todos os modelos, vemos isso pela coluna final da tabela que traz uma boa média para ambos. O menor valor deles ocorreu no Marian Romance para a reportagem de autoria de Marques (0,689).

Tabela 1. METEOR

	Marian Romance	**Marian Big**	**T5**	**mBart**	**Média**
Pivetta	0.74596	0.77254	0.73991	0.64907	0.72687
Marques	0.689	0.70539	0.71697	0.62901	0.68509

Na tabela 2 abaixo vemos os *scores* de BLEU, nota-se que para a reportagem de Marques os scores foram bem baixos, notavelmente para mBART (0,38447).

Tabela 2. BLEU

	Marian Romance	**Marian Big**	**T5**	**mBart**	**Média**
Pivetta	0.53313	0.59253	0.55979	0.41810	0.52589
Marques	0.46979	0.49487	0.49689	0.38447	0.46151

Na tabela 3 estão as pontuações para ROUGE, mais precisamente para ROUGE-L, que foram bastante favoráveis ao modelo, os maiores *scores* foram de recall e f1-macro para Pivetta. Em Marques os maiores *scores* também foram os de recall e f1-macro.

Tabela 3. ROUGE

	Métricas ROUGE-L	**Marian Romance**	**Marian Big**	**T5**	**mBart**	**Médias**
Pivetta	**recall**	0.72988	0.75586	0.72119	0.64896	0.71397
	precision	0.70603	0.74169	0.71742	0.63303	0.69954
	f1-macro	0.71762	0.74852	0.71901	0.64057	0.70643
Marques	**recall**	0.67903	0.70084	0.67749	0.62238	0.66994
	precision	0.64573	0.67965	0.65113	0.59139	0.64198
	f1-macro	0.66191	0.68969	0.66402	0.60647	0.65552

Como vemos na tabela 4, os tempos de execução foram bem variados, porém mBART chama mais a atenção pois foi o que levou mais tempo em ambas as reportagens (735.66 segundos para Pivetta e 736.87 para Marques), depois vemos que Marian Romance leva mais tempo para gerar texto para Pivetta e em seguida Marian Big para Marques.

Tabela 4. Tempo de execução (segundos)

	Marian Romance	Marian Big	T5	mBart
Pivetta	284,47	230,47	190,22	735,66
Marques	159,82	222,22	139,18	736,87

6.3. Resultados e Discussões

A análise dos resultados obtidos pelos modelos T5, mBART, Marian Romance e Marian Big evidencia diferenças importantes no desempenho em termos de precisão, fluência e conservação semântica nas traduções automáticas.

6.3.1. Desempenho dos Modelos

O modelo Marian Romance demonstrou alto grau de consistência gramatical e manutenção semântica. As predições preservaram, de maneira geral, a ideia central dos textos-alvo. Contudo, foram detectados alguns erros de tradução literal ou na falsa identificação de entidades nomeadas, como na Rodada 31 do conjunto de sentenças Pivetta onde ‘miracle’ e ‘Saint’ não foram traduzidos para ‘milagre’ e ‘Santo’, o que não reflete o contexto da sentença.

O modelo mBART utilizado consistentemente foi o mais demorado em tempo de execução e o com piores scores nas métricas, analisando manualmente o porquê, percebeu-se que apesar de construir estruturas sintáticas por vezes melhores do que outros modelos este introduziu diversas palavras em espanhol ao invés de suas equivalentes em português. O modelo também apresentou dificuldades em distinguir português de Portugal e do Brasil.

O modelo T5 apresentou desempenho intermediário em relação aos demais, com bons resultados em sentenças mais curtas e de menor complexidade. Porém para traduções de sentenças longas e complexas teve sua fluência e coerência global afetada, como na Rodada 40 do conjunto Pivetta, onde trocou sua tradução, que vinha sendo consistente em sentenças anteriores, de ‘células tronco’ para ‘células estaminais’.

O modelo MarianMT Big teve o melhor desempenho tanto em métricas quanto durante a avaliação humana, consistentemente tendo a tradução mais adequada e natural para um leitor de português.

Todos os modelos, à exceção do MarianMT Big e, por vezes o T5, enfrentaram dificuldades com o uso de aspas.

6.3.2. Tendências Identificadas nos Erros

Foram identificados três tipo recorrentes de erros: traduções literais, escolhas lexicais duvidosas e erros de semântica, ou erros contextuais.

Os modelos frequentemente optaram por traduções literais ao invés de equivalentes mais idiossincráticos, o que por vezes compromete a naturalidade do texto em português.

Porém em certos casos o oposto acontece, como na rodada 33 do conjunto Pivetta:

- Original:[...]red blood cells, (erythrocytes), the white blood cells[...]
- Romance:[...]glóbulos vermelhos, (eritrócitos), os glóbulos brancos[...]
- mBart:[...]células sanguíneas vermelhas (erytrocitos), células sanguíneas brancas[...]
- T5:[...]hemácias, (eritrócitos), as hemácias brancas[...]
- Big:[...]glóbulos vermelhos, (eritrócitos), os glóbulos brancos[...]

Neste exemplo podemos ver que os modelos que optaram por traduções menos literais se distanciaram do que seria mais natural em português, com o modelo T5 alucinando a existência de hemácias brancas.

Em especial notou-se que os modelos multilinguais, MarianMT Romance e mBart, por vezes utilizavam palavras em desacordo com a variante brasileira do português. O erro mais recorrente foi o uso de ‘células estaminais’, cuja ocorrência, à exceção de uma ocasião com o modelo T5, se restringiu a esses modelos. É possível que tenha ocorrido contaminação durante o treinamento e, pelas características multilingues dos modelos, tenham sido feitas inferências linguísticas, dada a similaridade entre os idiomas.

Erros Contextuais, ou de semântica, foram apresentados por todos os modelos, sendo o mais comum o de concordância nominal. Podemos ver isso claramente na tradução da Rodada 133 do conjunto Pivetta com o modelo mBart: “[..] Presidente George Bush *reserva apenas fundos* do orçamento federal para estudos [..]”, onde o texto original dizia “[..]only sets aside federal budgetary funds for studies carried out with[..]”. A tradução do modelo dá a entender que os estudos podem receber apenas fundos advindos do orçamento federal, uma tradução mais correta seria a observada no modelo Marian Big: “[...]*apenas reserva fundos* orçamentários federais para estudos realizados[...]”.

7. Conclusão

Com isso, concluímos que os modelos mBART, MarianMT Romance e MarianMT Big e T5 são de fato ótimos modelos e referências para tradução automática, salvo algumas questões importantes: mBART foi o mais penalizado principalmente utilizando o *dataset* proposto; o modelo foi o que mais levou tempo e o que mais sofreu com erros de tradução, e também por fazer uma tradução mais literal, isto é, menos livre do que as referências definidas pelo *dataset*.

De um modo geral, resumindo, Marian Romance demonstrou boa consistência gramatical e semântica, mas cometeu erros de tradução literal e na identificação de entidades. O mBART, embora mais lento, apresentou as piores métricas e incluiu palavras em espanhol, além de confundir variantes do português. O T5 teve desempenho intermediário, com melhores resultados em sentenças curtas e dificuldades em frases mais complexas. O Marian Big obteve o melhor desempenho geral, sendo o mais adequado para leitores de português. Aspas foram problemáticas para todos, exceto para o Marian Big e, ocasionalmente, o T5.

Erros comuns incluíram traduções literais, escolhas lexicais inadequadas e erros contextuais, como a tradução imprecisa de "células sanguíneas vermelhas" ou problemas de concordância nominal, exemplificados na tradução da frase sobre George Bush pelo mBART (conforme seção 6.3.2). Os modelos multilinguais, como Marian Romance e mBART, frequentemente introduziram termos em espanhol e variantes do português europeu, sugerindo contaminação nos dados de treinamento e inferências linguísticas equivocadas. Em resumo, o Marian Big demonstrou ser o modelo mais consistente, enquanto os multilinguais enfrentaram vários desafios relacionados à poluição linguística e inferências incorretas.

8. Referências

Slides do prof. Dennis Balreira

Alura. Hugging Face [Internet]. Available from: https://www.alura.com.br/artigos/hugging-face

Wikipedia. METEOR [Internet]. Available from: https://en.wikipedia.org/wiki/METEOR

Towards Data Science. Transformers [Internet]. Available from: https://towardsdatascience.com/transformers-89034557de14

NLPlanet. (2022) Two Minutes NLP: Learn the ROUGE Metric by Examples. Disponível em: https://medium.com/nlplanet/two-minutes-nlp-learn-the-rouge-metric-by-examples-f179cc285499. Acesso em: 2 jan. 2025.

Hugging Face. (n.d.) T5 Model Documentation. Disponível em: https://huggingface.co/docs/transformers/model_doc/t5. Acesso em: 2 jan. 2025.

Chiusano, F. (2023). *Two minutes NLP — Learn the BLEU metric by examples*. Disponível em: https://medium.com/nlplanet/two-minutes-nlp-learn-the-bleu-metric-by-examples-df015ca73a86. Acesso em: 2 jan. 2025.

GeeksforGeeks. (2023). *Transformer Attention Mechanism in NLP*. Disponível em: https://www.geeksforgeeks.org/transformer-attention-mechanism-in-nlp/#3-selfattention. Acesso em: 6 jan. 2025.

Martins, R. T. (2008). *Tradução automática*. Todas as Letras-Revista de Língua e Literatura, 10(2).