Tradução Automática: uma revisão sistemática

Larissa Kelmer de Menezes Silva[1], **Daniel Sabino**[1]

[1]Instituto Metrópole Digital – Universidade Federal do Rio Grande do Norte (UFRN)
Caixa Postal 1524 – 59.078-970 – Natal – RN – Brazil

Abstract. *The various personal assistants, online translators, speech recognition and speech transcription systems are just a few examples of the contribution of the Natural Language Processing (NLP) area. This strand is focused on the processing of multiple types of human languages by computers. Currently, this line of Computational Intelligence constitutes a field of great development, mainly due to the greater availability of data and greater computational processing found today. Part of this set of methods is automatic translation. The task of translation, which remains a current demand, has been highlighted by globalization and the consequent need for expanded communication. In this work, a systematic review is developed to identify which characteristics govern recent works in the area. Finally, the study manages to answer all the questions that govern it, from primary information, such as places and years of publication, to the main metric used in the studies.*

Resumo. *As diversas assistentes pessoais, tradutores online, sistemas de reconhecimento de voz e de transcrição de fala são apenas alguns exemplos da contribuição da área de Processamento de Linguagem Natural (PLN). Essa vertente é voltada para o processamento de múltiplos tipos de linguagens humanas por computadores. Atualmente, tal linha de IC se constitui como um campo de grande desenvolvimento, principalmente pela maior disponibilização de dados e maior processamento computacional encontrados na atualidade. Como parte desse conjunto de métodos está a tradução automática. A tarefa de tradução, que segue como uma demanda atual, tem seu destaque aumentado pela globalização e consequente necessidade de comunicação ampliada. Neste trabalho, é desenvolvida uma revisão sistemática a fim de identificar quais características regem trabalhos recentes na área. O estudo, por fim, consegue responder todas as perguntas que o regem, desde informações primárias, como locais e anos de publicação, à principal métrica usada nos estudos.*

1. Introdução

A maior capacidade de processamento de computadores e de suas encadeadas consequências, como a produção e o consumo exacerbado de dados, permitiu que técnicas computacionais pudessem ser melhor aplicadas e evoluídas. Não obstante, também permitiu o desenvolvimento de novos métodos de aprendizado automático. Quanto mais se dá esse desenvolvimento, mais intrínseco é o uso de tecnologias associadas ao dia-a-dia social e mais natural é esse processo, o que pode ser percebido pela utilização das técnicas em questão em outras áreas como a de Processamento de Imagens[1] e de Sons[8][7]; e no desenvolvimento da subárea do Processamento de Linguagem Natural[6].

Nesse sentido, e visando uma globalização contínua, as técnicas de *Machine Translation* (MT), ou Tradução Automática (TA), surgem como possibilitadorass: já que

a abundância em informação é tal que se torna impraticável depender apenas de tradutores humanos[10][20] e tampouco pode-se esperar que as pessoas sozinhas consigam quebrar a barreira idiomática. Assim, a demanda por *softwares* de tradução é crescente. A TA evoluiu desde sistemas baseados unicamente em dicionários e regras gramaticais até sistemas capazes de aprender profundamente e se adaptar a outras tarefas inclusive. Apesar disso, qualidade da tradução é uma preocupação constante, já que, no geral, a qualidade da tradução automática é menor do que a qualidade atingida em traduções manuais. Uma vez sendo um problema em aberto, diversos são os estudos que visam suprir essa demanda.

Considerando os pontos levantados, é nítida a necessidade de estudos propostos para identificar esses trabalhos de TA, avaliar seu rigor científico, e pontuar suas características e seus problemas. Esta revisão, então, pretende agir para esse fim e divide-se como se segue: a primeira seção será voltada a estabelecer a metodologia de pesquisa do presente trabalho. Nela, pretende-se determinar quais questões irão regê-lo, quais os métodos para fazer o levantamento de pesquisas e quais critérios devem ser atendidos para a inclusão das pesquisas levantadas nesta. Em seguida, apresentar-se-á os resultados do levantamento citado e, por fim, procurar-se-á responder as questões propostas inicialmente.

2. Metodologia

Este trabalho baseai-se no método de Kitchenham e pretende construir uma revisão sistemática, cujo objetivo é o levantamento do atual estado da arte em Traduções Automáticas. As seções seguintes dedicar-se-ão a estruturar tal revisão a partir de: levantamento de questões orientadoras, definição uma *string* de busca, definição criterios de inclusão e exclusão dos estudos. Após isso, exibir-se-á o produto da aplicação desses passos sobre os trabalhos selecionados.

2.1. Questões de Pesquisa

- QP1. O objetivo principal é uma tradução automática, um pós-processamento ou uma adaptação de um modelo já existente?
- QP2. Quais as principais técnicas utilizadas na tradução automática?
- QP3. Qual o tamanho e o tipo da base de dados? Há pré-processamento? Há pós-processamento?
- QP4. Quais métricas são usadas para avaliar o modelo?

O objetivo principal da primeira pergunta é mapear os trabalhos que se concentram em criar novos algoritmos e diferenciá-los daqueles cujo foco é melhorar (ou adequar) um modelo. Assim, é possível perceber se há modelos bem estabelecidos ou se os esforços futuros devem voltar-se a criar um modelo do zero. A segunda pergunta permite que se mapeie quais técnicas estão sendo mais utilizadas. Já a terceira pergunta permite que se levante o tamanho médio e o tipo de base de dados utilizada, se são textos inteiros, palavras individualmente, dicionários; se, caso sejam textos, são textos coloquiais, se são de cunho artístico ou acadêmico, se o corpus inclui textos mais de uma área de conhecimento, dentre outras possíveis delimitações.A última pergunta permite que se analise quais as métricas mais utilizadas e dentro de qual contexto ela são mais utilizadas.

2.2. Processo de Pesquisa

Os bancos de dados usados para realizar as pesquisas foram o Scopus (Elsevier)[1], por manter a maior base de dados na área de computação. Inicialmente, usou-se uma *string* de teste, ((”machine” OR ”machine learning” OR ”natural language” OR ”automatic” OR ”NLP”) AND (”translation” OR ”translating” OR ”translation”)). A *string* de busca foi obtida a partir de termos relacionados à Tradução Automática e à Processamento de Linguagem Natural, em que se tentou mais de uma *string*, em ambos os bancos de dados, até que se obteve aquela com resultados mais relevantes (categorizados assim pela leitura dos título e resumo). Por fim, a *string* escolhida foi: ((”automatic translation”) OR (”machine translation”) OR (”neural machine translation”)). Além disso, durante o processo de busca, os artigo foram ordenados segundo relevância em todos os casos.

2.3. Critérios de inclusão e exclusão

Os trabalhos selecionados foram avaliados segundo dois conjuntos de critérios, sendo avaliados primeiro pelos de exclusão e, em sequência, pelos de inclusão. Caso fossem classificados em ambos, os trabalhos eram ou lidos a fundo ou, se já nessa etapa, eram excluídos.

Assim, começou-se a classificação pelo próprio filtro de busca, em que se determinou a data, apenas estudos de depois de 2017 foram considerados, se eram *peer-reviewed*, se o título se relacionava com o tema deste trabalho e se se classificavam como *research articles*. Ainda que outros documentos tenham sido usados como referência pra produzir o presente trabalho, apenas pesquisas com resultados práticos foram incluídas.

Assim, *reviews*, levantamentos do estado da arte e demais trabalhos de cunho teórico apenas foram desconsiderados. Além dos pontos citados, os critérios de exclusão também contavam com: resumos que não incluíam objetivos, métodos e resultados de forma nítida, trabalhos em que a tradução envolvia linguagens computacionais (como língua fonte ou alvo), trabalhos em outras línguas que não o inglês, resumos publicados como curtos ou como pôsteres.

2.4. Avaliação de Qualidade

Os estudos foram avaliados conforme os critérios de avaliação de qualidade definidos por Dyba [5]:

- QA1. Há uma declaração que indique os objetivos da pesquisa?
- QA2. Existe uma descrição adequada do contexto em que a pesquisa foi realizada?
- QA3. O estudo tem valor para pesquisas ou práticas?
- QA4. Há uma declaração sobre os achados da pesquisa?

2.5. Coleta de Dados

Os dados extraídos dos estudos selecionados foram: título, autores, ano, objetivo, tipo de tradução, tipo de algoritmo, se há a criação ou a melhora de algum modelo já existente, tipo de pré/pós-processamento (caso haja), tipo de base de dados utilizada e métricas utilizadas.

[1] https://www.scopus.com/home.uri

3. Resultados

Essa seção objetiva detalhar as etapas do processo a partir dos resultados obtidos, além de apresentar um resumo dos estudos classificados.

3.1. Resultados da Busca

A busca inicial na base Elsevier gerou 147 resultados, os quais foram filtrados e analisados segundo os critérios, resultando em 18 estudos. Por fim, os estudos restantes foram lidos na íntegra e, deles, 10 trabalhos se encaixavam perfeitamente nos requisitos.

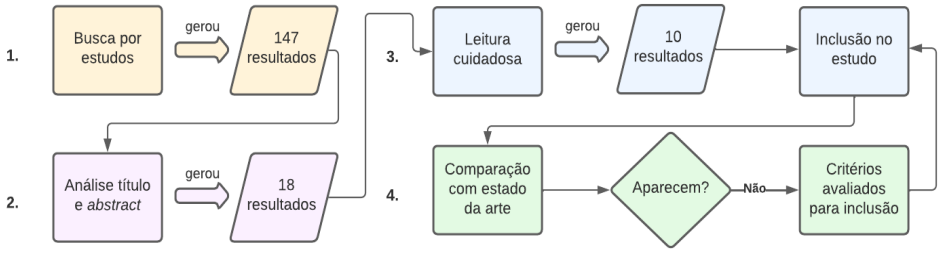

Figure 1. Etapas metodológicas.

Após a terceira etapa, que demandou a leitura completa dos artigos remanescentes, obteve-se os seguintes 10:

Table 1. 10 Estudos Submetidos à Leitura Cuidadosa.

Id	Author	Year	Venue
T1	Zhang Z., et al.	2021	Journal
T2	Su C., et al.	2020	Journal
T3	Pan B., et al.	2020	Journal
T4	Shi X.. et al.	2021	Journal
T5	Wang F., et al.	2019	Journal
T6	Yang Z., et al.	2018	Journal
T7	Su J., et al.	2019	Journal
T8	Tan Z., et al.	2018	Journal
T9	Liv Y., et al.	2021	Journal
T10	Yang B., et al.	2019	Journal

A mesma *string* foi utilizada em outras bases, como a arXiv[2] e a IEEE[3], com o objetivo de encontrarmos os principais trabalhos na área, também denominados *Foundation Models*[2], segundo quantidade de citações. Entretanto, uma vez que, numa análise exploratória desses, não foram encontrados tais trabalhos, optou-se por incluí-los à parte, como complemento ao estudo. A causa disso pode ser devido ao caráter multi-modal dos ”trabalhos-base”.

Considerando os *Foundation Models*[2], de aplicação à tradução automática e, como indica a figura 1, tendo os critérios de inclusão avaliados, os trabalhos também foram escolhidos segundo a quantidade de citações.

Título	Citações Semantic Scholar	Citações Google Scholar
Language Models are Few-Shot Learners	3524	3350
Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer	2902	2902
BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding	9999	36172
Attention is All You Need	31008	39332

Table 2. *Foundation Models*[2] por critérios de busca e Citações dos Semantic Scholar e Google Scholar.

[2]ArXiv não é um *journal*, é um repositório online e gratuito, que costuma abrigar pré-publicações, antes que sejam publicadas de fato. https://arxiv.org/

[3]https://ieeexplore.ieee.org/Xplore/

Por fim, os trabalhos incluídos foram:

Table 3. 10 Estudos Submetidos à Leitura Cuidadosa.

Id	Autor	Ano	Local de Publicação
T1	Zhang Z., et al.	2021	Journal
T2	Su, C., et al.	2020	Journal
T3	Pan, B., et al.	2020	Journal
T4	Shi, X.. et al.	2021	Journal
T5	Wang, F., et al.	2019	Journal
T6	Yang, Z., et al.	2018	Journal
T7	Su, J., et al.	2019	Journal
T8	Tan, Z., et al.	2018	Journal
T9	Li, Y., et al.	2021	Journal
T10	Yang, B., et al.	2019	Journal
T11	Brown, T., et al.	2020	Conferência
T12	Raffel, C., et al.	2019	ArXiv
T13	Devlin, J., et al.	2018	ArXiv
T14	Vaswani, A., et al.	2017	Conferência

3.2. Visão Geral dos Resultados

Assim, uma vez em posse dos trabalhos incluídos, pôde-se obter os seguintes dados sob uma primeira análise:

- I. Locais de publicação: dentre os trabalhos levantados, a maioria dos trabalhos era de publicação vinculada à instituições chinesas, exceto por 8 estadunidenses, sendo 1 publicação dividida com o Canadá. A quantidade de publicações podem estar diretamente ligada à quantidade de falantes de ambas[4] as línguas atrelado ao desenvolvimento tecnológico dos países de origem dessas pesquisas[5]. Além disso, todos os *Foundation Models* são de intituições estadunidenses, ainda que nem todos os autores o sejam.

Figure 2. Estudos selecionados por países dos autores.

[4]http://www.ethnologue.com/statistics/size
[5]https://www.wipo.int/pressroom/en/articles/2021/article_0008.html

- II. Anos de publicação: a partir da análise quanto à instituição de filiação dos autores, obteve-se os dados da figura 3. O fato de os estudos terem aumentado, ainda que mantenham uma certa constância, pode indicar um maior índice de publicações na área. Pode-se esperar, senão um aumento, uma constância no desenvolvimento acadêmico do tema.

Figure 3. Estudos selecionados por ano de publicação.

4. Discussão

Esta seção obejtiva responder as questões de pesquisa com base nos artigos selecionados. O objetivo é responder essas questões com base em comparações entre os trabalhos, de forma que, caso algum trabalho não seja incluído em um ou mais tópicos, será pela falta de informação ou de clareza no texto.

4.1. QP1. O objetivo principal é um novo algoritmo de tradução automática ou uma adaptação de um modelo já existente?

Em relação ao objetivo das pesquisas, observou-se que a maior parte dos trabalhos, T1-T5, T7-T10, não propunha métodos de tradução automática com grandes novidades, apoiando-se em técnicas de NMT já conhecidas e estabelecidas. Tal feito só foi realizado por um único trabalho, o T6, dentre os selecionados inicialmente. Entretanto, todos esses trabalhos propunham mudanças arquiteturais significativas o suficientes para representar uma melhoria no processo: nos trabalhos T4, T3, T6, T8, observou-se a suplantação do modelo-base em todas as categorias experimentias. Já nos trabalhos T1, T5, T7, T9, essa superação não foi generalizada para todos os testes, mas representou um estado competitivo com os modelos de base utilizados em cada caso.

Aqueles incluídos *a posteriori*, T11-T14, também apresentaram a característica de serem inovadores, ou *groundbreakers*, motivo pelo qual os tornou o estado da arte em MT e, por conseguinte, que os fez serem incluídos aqui. É de se destacar que esses últimos trabalhos focaram-se em tarefas gerais de Processamento de Linguagem Natural, podendo ser estendidos à tradução automática, exceto pelos Transformers, cujo objetivo inicial era a própria TA. Fazendo o caminho inverso dos demais, o T6 usou-se de uma técnica reconhecida em outras áreas, como a de Visão Computacional, a *Generative Adversarial Network*, para gerar traduções. Assim, seu objetivo era, de fato, a TA. Por fim, os autores do trabalho em questão deixaram como proposta a diversificação do fim do modelo.

Não obstante, apesar de alguns trabalhos terem feito uso de pré-processamento em suas bases de dados, esse aparato foi apenas complementar ao método principal proposto em questão. Quanto ao pós-processamento, nenhum dos estudos selecionados utilizou de tal técnica

4.2. QP2. Quais as principais técnicas utilizadas na tradução automática?

Os trabalhos selecionados a princípio basearem-se, invariavelmente, em técnicas de NMT e de atenção, de forma que as principais diferenciações giravam em torno do nível da tradução (a nível de caracter, de palavra, de sentença ou de arquivo), na técnica de aprendizado de máquina utilizada (*Tree-based* ou *Recurrent Neural Network* (RNN) ou se não discutiam esse nível de aprendizado), na quantidade (variando entre 1 e 2) e tipo (uni ou bidirecionais) de *encoders/decoders* e em mudanças arquiteturais ainda menores, por mais significatantes, como a dimensão do modelo ou a forma de seleção dos parâmetros. De tal forma que, para responder essa questão, QP2, optou-se por iniciar esta seção com um levantamento breve sobre os trabalhos, a fim de respaldar a resposta para essa. Como todos os trabalhos utilizam-se dos *Foundation Models* para construírem suas respectivas arquiteturas, primeiro desenvolver-se-á sobre esses modelos, para, só então, analisar os demais.

4.2.1. Levantamento

Em sendo assim, T14[3], ou Transformer, é um dos trabalhos que mais revolucionou a área de NLP, ao escantear a recorrência e ser o primeiro modelo a utilizar apenas mecanismos de auto-atenção, *self-attention*. Utiliza de *multi-head attention*. Além disso, esse trabalho é construído sobre a estrutura *encoder-decoder*, ou codificador-decodificador, mas sem se utilizar de redes recorrentes ou convolucionais. Sua arquitetura conta com o uso de normalização (regularização), a função *softmax*, o *byte-pair encoder* (BPE), ambas as duas últimas tecnologias também são usadas em T11[3], e o otimizador Adam. O Transformer foi pensado para o uso em tarefas de traduções, mas considerou-se a extensão do modelo para outras tarefas.

O modelo de T11[3], o GPT-3, outro dos *Foundation Models*, tem sua arquitetura baseada em Transformer, tal qual T12[13] e T13[4]. Portanto, é um modelo codificador-decodificador, baseado em atenção. Esse modelo usa uma pré-normalização e "(...) alternating dense and locally banded sparse attention patterns in the layers of the Transformer (...)". O trabalho não faz maiores detalhamentos, exceto compará-lo ao seu antecessor, GPT-2. Um dos motivos para seu reconhecimento tão expressivo é seu tamanho: é considerado (até onde sabemos) o modelo treinado com maior número de parâmetros, motivo pelo qual não necessita de nenhum *fine-tuning* para realizar tarefas especídifcas de linguagem. Ao não objetivar uma só função, o modelo realiza tipos distintos de testes, que fogem ao escopo deste trabalho.

O modelo do trabalho T12[13], também conhecido por "T5", é um modelo voltado ao *transfer learning*. Assim como T11, esse também é voltado para *multi-task*s, como T13[4]. Ele também conta com a normalização de camadas e a pré-normalização. Dessa forma, o modelo também se baseia na arquitetura do BERT[18]. Além disso, aplica estratégias de *pre-training*, tal qual T13[4].

O trabalho T13[4] traz o codificador bidirecional como um de seus principais adendos ao Transformer, além de a *self-attention*, motivo pelo qual ganhou tanto destaque. Esse trabalho também conta com o uso de *softmax* como normalizador e o otimizador Adam.

O T1[24], por ter implementado uma variação do BERT[4], usou a mesma estrutura base, adicionando, à ela, dois módulos. Além de T1, apenas T9 menciona o BERT. O método utilizado por T1 neste sentido, é o dito "baseado em recursos", cujo modelo é do tipo pré-treinado, cujo *encoder* e o *decoder* possuem camadas com módulos de fusão, ou *fusion*, e de atenção conjunta, ou *joint-attention*, responsável por integrar camadas de atenção e realocá-las entre diferentes representações.

T2[16] incorpora Gumble-Tree-LSTM[6] em ambos os codificadores e decodificadores. Por um lado, ao usar RNN, o modelo se utiliza de LSTM no codificador e decodificador, além de contar com um codificador de árvore latente. Por outro, o modelo de auto-atenção tem seu primeiro codificador e seu decodificador replicados do Transformer, além do codificador de árvore latente.

[6]Gumble-Tree-LSTM é uma tecnologia de árvore não-supervisionada, usada para melhoria de modelos NMT. É utilizada para capturar e codificar o conhecimento de acerca de composição de estruturas de árvores - aqui, essa informação é aplicada ao contexto.

Outro trabalho que se utiliza de técnicas de aprendizadoo em árvore é o T10: um modelo NMT sequencial e de dependência, que usa a arquitetura atencional encoder-decoder, com o *encoder bidirecional*. Também se utiliza de RNN, especificamente GRUs *left-to-right* e *right-to-left*, além de a geração de frases ser representada como *bottom-up*. É um dos poucos trabalhos selecionados que não se utiliza de técnicas de *self-attention*.

Fazendo um paralelo tanto com T2 quanto com T10, T8[17] também se utiliza de técnicas de latência para a construção do seu modelo. Nesse trabalho, são usados dois codificadores baseados em latência, um com pré e outro com pós-composição. Para os codificadores são usadas RNNs bidirecionais, GRUs, e, para o codificador, é usada uma GRU com *deep-stacked* nas representações. O trabalho destaca a possibilidade de se usar um LSTM.

De forma semelhante, o T9 também se baseia em técnicas de latência para a construção de seu modelo. Ele se utiliza de um *latent feature encoder* (LFE), com o intuito de capturar as representações latentes de *features*, acoplado a um NMT *encoder*. Além desse, ele também usa de um *Transformer encoding block*, como complemento ao seu outro codificador. Assim como o codificador, o decodificador também é baseado em Transformer.

Em contrapartida, o T3 usa dois decodificadores e um codificador, além de um auto-codificador, ou *auto-encoder*, composto por um dos decodificadores. Além disso, T3 destaca dar uma menor importância à ordem das palavras. Esse modelo usa LSTM bidirecional pro encoder, que é divido. É um dos poucos trabalhos que menciona o Google's Neural Machine Translation (GNMT) System e também compara com o Transformer.

Já T4 é um dos poucos trabalhos que não é voltado ao desenvolvimento de um modelo dito *end-to-end*. Ele aplica suas propostas de melhorias sobre um modelo RNN-NMT e sobre o Transformer. Esse modelo RNN utiliza-se de um discriminador, formado por um codificador bidirecional, e um decodificador, também com RNN, mas sem maiores distinções. O T4 foca no alinhamento de sentença com discriminador - seu principal objetivo.

O T5[19] utiliza um segmentador de palavras, além da arquitetura padrão *encoder-decoder*. Seu encoder possui duas GRUs bidirecionais separadas, em que uma é compartilhada com o módulo de segmentação de palavras; da mesma forma, seu decodificador também é cosntruído com uma GRU.

O [22] destacou-se como um dos trabalhos com abordagem mais diferenciável. Ele usa da *generative adversarial net* com dois submódulos: um discriminador, pra discriminar entre as traduções geradas por máquinas e as geradas por humanos; e um modelo generativo, que traduz uma sentença fonte na sentença alvo. O gerador não assume uma arquitetura específica, mas baseia-se em atenção e cujo encoder usa uma GRU com uma camada bidirecional e um decoder com GRU e camada condicional. Já o discriminador usa uma CNN. Além disso, ele propõe uma alternativa ao próprio modelo, que segue o mesmo padrão do outro, porém com dois codificadores e dois decodificadores. Diz não assumir arquitetura NMT específica.

[15] se baseia em atenção e também se utiliza do modelo *encoder-decoder*, com uso de GRUs, ainda que seja aplicável a outras unidades de RNN. Aqui, como diferencial, ele propõe um *backward-decoder* e um *forward-decoder*, que também é bidirecional -

assim como seu codificador. Como base para construção do modelo, ele se utiliza da arquitetura do Transformer. Esse estudo também envolve pré e pós-tradução, algo quase não utilizado em outros estudos. Foca na apreensão do contexto.

4.2.2. Análises

Com base no levantamento, é possível, de fato, comparar quais tecnologias são mais utilizadas e, portanto, representam o a TA atualmente.

É unânime que todos os trabalhos reconhecem a efetividade da Atenção, porém, os mecanismos escolhidos e o nível de aplicação diferem entre si. A atenção conjunta baseia-se no mecanismo de atenção proposto por [18], em que todos os trabalhos levantados aqui também se baseiam, mas que apenas T1 se utiliza. A atenção conjunta é formada pela combinação da *self* e da *cross-attention*. Ainda que nenhum outro trabalho faça menção à *cross-attention*, T2-T4, T9, T12-T14 utilizam-se de *self-attention*, que, inclusive, foi introduzida pelo BERT[4]. utiliza-se de mecanismos de atenção, mas não faz menções diretas a *self* ou *cross-attention*. Outro tipo de atenção encontrada nos trabalhos é a *multi-head* attention, empregada pelo Transformer[18], que é usada em T4 e T7. Os trabalhos T5, T6, T8, T10 e T11 não usam diferenciações de atenção para além do mecanismo base.

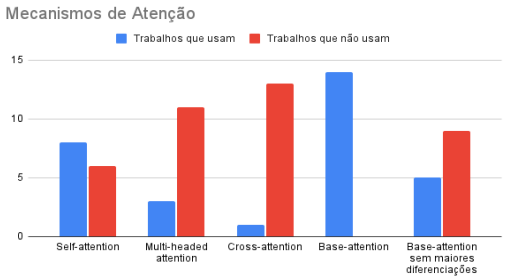

Figure 4. Trabalhos e Mecanismos de Atenção.

É interessante observar que alguns trabalhos, por mais que falem em atenção, estão se referindo a outros fins que não o mecanismo de atenção em si. Por exemplo, o T5 introduz a ”atenção” híbrida, ou *hybrid-attention*. Entretanto, a atenção híbrida diz respeito ao nível de unidade a que se presta atenção: caracter ou palavra (nesse caso). Outras distinções quanto a nível de unidade a que se foca serão levantados mais à frente. Além de T5, T7 utiliza ”atenção” ao se referir ao seu codificador-decodificador, o qual codifica ”contextos reversos”. Ou seja, estando no lado-destino, essa estrutura codifica contextos para a tradução da direita para a esquerda (*right-to-left*), o que é possível pelo caráter assíncrono do modelo. Outros modelos referem-se a mecanismos de função semelhante apenas definindo se o codificador/decodificador são bi ou unidirecionais.

Como complemento a mecanismos de atenção, todos os trabalhos utilizam-se da estrutura codificador-decodificador, mas apenas T6 não baseia seu modelo NMT nessa estrutura. As maiores variações dessa característica estão na quantidade de cada mecanismo desses e a direção de processamento. Observou-se, então, que T1-T8, T10 e T13 fazem uso de codificadores bidirecionais. Quanto a estruturas divergentes, T5 se utiliza de uma camada bidirecional decodificadora, além do codificador do modelo NMT, no seu

segmentador de palavras. T9 também se utiliza de uma dessas representações em uma camada, porém, neste caso, ele usa um codificador replicado de um modelo baseado em Transformer[7], além contar com um codificador NMT. T6 usa de um codificador e um decodificador dentro do seu gerador, mas não de um codificador-decodificador no modelo NMT, seguindo a distinção feita por T5. Ele, ainda, é o único trabalho a fazer uso de Redes Neurais Convolucionais, CNNS, e de Redes Adversariais Generativas, GANs. Em contrapartida, esse último trabalho faz uso de um gerador e de um discriminador. Desses, apenas T7 conta com dois decodificadores, um *right-to-left* e outro *left-to-right*, mas dito bidirecional. Em T7, "bidirecional" diz sobre a atenção do modelo, como já pontuado, e não sobre para qual lado o processamento ocorre. Assim, pode-se dizer que a estrutura de decodificação é bidirecional, porém feita separadamente. Além desses, T10 apresenta um codificador de dependência lexicalizada, variação pensada para estruturas de árvore - que esse trabalho segue. T4 divide-se em dois modelos: um baseado em NMT RNN, seguindo a arquitetura *encoder-decoder* padrão, mas com o codificador bidirecional. Pontua-se que o modelo pode ser construído tanto com LSTM quanto com GRU. O outro modelo de T4, baseia-se em Transformer. T11 não esmiuça sobre sua estrutura de codificação e decodificação. T12, ainda que tenha algumas pequenas mudanças nessa estrutura, se baseia em Transformer também. T14 é um modelo de *encoder-only*, por isso, não possui decodificadores.

Quanto à forma de se materializar os modelos, os principais tipo são: baseado em RNN (LSTM e GRU), em árvore ou apenas em atenção. Assim, 8 dos 14 trabalhos são baseados em RNN. Desses, T2-T4 são LSTMs; e T5-T8 e T10 são GRUs. Os trabalhos T3, T7 e T8 indicam a possibilidade de serem construídos com outro tipo de RNN. Já T2 e T10, ainda que se utilizem de RNN, têm seu enfoque principal nos modelos de árvores, além de, em ambos, o uso de mecanismos de latência ser uma das características mais relevantes - característica semelhante à vista em T8 e T9. Por conseguinte, sobre a T1, T5, T6, T9 e T11-T14 o uso de atenção apenas, computando 57% dos trabalhos.

Os trabalhos T1-T4, T6, T7, T9-T12 explicitam o enfoque ao contexto, ainda que seja sabido que outros trabalhos também o considerem. Quanto ao nível a que se atentam, ou unidade de atenção, os modelos de T4, T5, T8 e T11 trabalham com o nível de caracter - exceto por T11, são trabalhos produzidos na China e com testes exaustivos envolvendo línguas asiáticas (principalmente o chinês). Teoriza-se, aqui, que o motivo para tanto é a quantidade de informação armazenada em um único símbolo básico da língua, o caracter. Os trabalhos T5 e T6 foram os únicos a, explicitamente, usarem o nível de palavras. De forma parecida, apenas T6, T7 e T9 usaram o nível de sentença; e T13 foi o único a destacar o nível de *token*.

A segmentação de palavras foi usada nos trabalhos T2, T4-T10 e T13, compondo 64% dos trabalhos. T1, T3-T5, T7-T9 e T10 fizeram uso de *byte pair encoding* (BPE), que é, basicamente, um compressor de palavras. Todos os trabalhos, exceto T5, T6, T9 e T11, utilizaram-se da função *softmax* para compor a predição de traduções.

Todos os trabalhos, exceto por T6 e T8 ou basearam-se (incluindo a réplica de partes do modelo) ou comparam-se ao Transformer, o que era de se esperar devido à repercussão de tal modelo e por ele ter sido, enquanto esses outros trabalhos eram desen-

[7] Vanilla Transformer encoder.

volvidos, o estado-da-arte em Tradução Automática. Ainda teoriza-se aqui que, o motivo pelo qual o BERT teve tão poucas citações é o fato de ter-se usado, nesta revisão, o ano de publicação dos trabalhos. Dessa forma, metade dos estudos foram publicados até 2019, mesmo ano de publicação do BERT. É razoável supôr que estudos práticos como esses são demoram alguns meses para serem completados. Dessa forma, estudos publicados em 2020, ano de mais publicações selecionadas, começaram ou de forma cocomitante ao BERT ou um pouco depois - possivelmente até antes de sua consolidação entre pares. Confirmando essas suposições, estudos que se basearam no BERT, T1 e T9, só foram aplicados em novembro de 2020. Outro estudo que citou o BERT foi o T12, mas, esse, apenas o usou para fins de comparação - ainda que também tenha sido aplicado e publicado em 2020.

Outras características encontradas com menor frequência são:

- Alinhamento de sentença: T2 e T4.
- Discriminador: T4 e T6.
- Compararam com e/ou se basearam em GNMT[21]: T3, T5 (usou apenas como *baseline*) e T14.
- Compararam com e/ou se basearam em DL4MT[8]: T5 (usou apenas como *baseline*), T6 e T10.
- Compararam e se basearam no BERT: T1, T9 e T12.

4.3. QP3. Qual o tamanho e o tipo da base de dados? Há pré-processamento? Há pós-processamento da base?

Na esmagadora maioria dos casos, o tipo da base de dados utilizada não importou. Isso é, os estudos não focavam-se em traduzir um tipo textual específico (como notícias ou gêneros literários). Eles utilizavam-se das mesmas bases que seus respectivos *base-lines*, a fim de comprovar sua eficácia, mas sem estender seus estudos às possíveis nuances de cada tipo textual.

É importante ressaltar que a maioria dos textos, inclusive os focados em traduções do/para chinês, utilizaram-se de bases de línguas ocidentais (ou orientais, caso o texto focasse em traduções ocidentais) para fins competitivos. A única exceção foi o T9, que só conduziu experimentos de traduções do inglês para o alemão e o T3, que, apesar de não ter conduzido experimentos para o chinês, conduziu para o vietnamita.

Ficou muito nítida, ainda, a repetição do uso das bases. Ainda que sirva de ponto para comparação, caso uma dessas bases seja enviezada ou não apresente as nuances da tradução, a não-expansão de bases pode ocasionar uma perpetuação de parâmetros que não são ótimos e que poderiam ser melhores desenvolvidos. Por esse motivo, destacam-se os trabalhos, T3, T7, T10 e T12, que, além de bases para fins de comparação, utilizaram a experimentação para testar seus modelos em línguas *low resources*. Portanto, línguas cujas bases são menores e menos utilizadas.

As bases de dados utilizadas giravam em torno de 1 milhão de pares de sentenças. Como era esperado pelos locais de publicação e pelo número de falantes da língua (além de a possibilidade de ser um boa métrica de avaliação dada a diferença do modelo de escrita ocidental), a maioria dos trabalhos (T1, T2,T4-T8 e T10) baseou-se, principalmente,

[8]https://github.com/nyu-dl/dl4mt-tutorial

em traduções do chinês para o inglês ou ao contrário. O T3, ainda que não tenha utilizado o chinês, usou o vietnamita.

As principais bases testadas e seus respectivos tamanhos podem ser observadas na tabela[4] e em seu complemento[5].

4.4. QP4. Quais métricas são usadas para avaliar o modelo?

Todos os trabalhos, T1-T12 e T14, utilizaram-se do BLEU[9] como métrica comparativa dos seus trabalhos ou de alguma variante do BLEU. Parte disso, muito possivelmente, deve-se ao parâmetro de comparação de trabalhos anteriores, que usaram essa mesma mética (por ser a mais estabelecida na área). Desses, apenas T2 e T7 contaram com outras métricas, de modo a complementar e não substituir o BLEU. T13 é o único trabalho que não faz qualquer menção à métrica em questão. Muito possivelmente isso se deve ao fato de ele exibir o treinamento de outras *tasks* que não a tradução automática.

[9]https://en.wikipedia.org/wiki/BLEU

5. Conclusão e Trabalhos Futuros

Este trabalho apresentou uma revisão sistemática da literatura, acerca de modelos de Tradução Automática. Tinha, como fim, o levantamento e a análise de técnicas da TA, sos trabalhos que compõem o estado-da-arte atual, dos seus métodos de avaliação, dentre alguns outros aspectos. Para isso, delimitou-se um escopo de 10 trabalhos, segundo um rigor científico determinado na seção II, aos quais acresceu-se alguns dos trabalhos de NLP de maior repercussão e que compõem os ”modelo-base”, segundo número de citações. Por fim, obteve-se 14 trabalhos, os quais serviram como suporte para que se respondesse as perguntas orientadoras do estudo. Observou-se, ainda, uma tendência crescente no número de estudos - por mais que estudos no último ano não tenham tido maior frequência, o presente trabalho determinou incluir apenas trabalhos revisados. Para trabalhos futuros, considera-se a possibilidade de se investigar mais a fundo os métodos avaliativos e os modelos-base. Considera-se, ainda, o enfoque deste trabalho em pesquisas voltadas ao português e as ditas línguas *low-resources*.

ID	Treino				Validação		Teste	
	Base	Tamanho	Idiomas	Processamento	Base	Tamanho	Base	Tamanho
T1	IWSLT'14	0.160 mi	Inglês - Alemão	Lowercased Tokenização BPE	7k do set de treino	-	dev2010, dev2012,tst2010, tst2011, tst2012	-
	IWSLT'14	0.183 mi	Inglês - Espanhol	Tokenização BPE	dev2010,tst2010, tst2011, tst2012	-	tst2013, tst2014	-
	IWSLT'17	0.236 mi	Inglês - Francês	Tokenização BPE	tst2011, tst2012, tst2013, tst2014, tst2015	-	tst2016, tst2017	-
	IWSLT'17	0.235 mi	Inglês - Chinês	Tokenização BPE	tst2011, tst2012, tst2013, tst2014, tst2015	-	tst2016, tst2017	-
	WMT'14	4.5 mi	Inglês - Alemão	BPE Tokenização	newstest2012, newstest2013	-	newstest2014	-
T2	BOLT (LDC2013E80, LDC2013E81, LDC2013E83, LDC2013E85, LDC2013E118, LDC2013E125, LDC2013E132, LDC2014E08, LDC2014E69, LDC2013E119)	0.121 mi	Chinês - Inglês	-	BOLT	4.935	BOLT	4.977
	ASPEC (WAT'15)	1.5 mi	Inglês - Japonês	Segmentação (japonês) Filtragem por tamanho (50 tokens)	ASPEC	1.790	ASPEC	1.812
	LDC (LDC2002E18, LDC2003E07, LDC2003E14, LDC2004T07, Hansardsportion of LDC2004T08, e LDC2005T06)	1.678 mi	Chinês - Inglês	Segmentação (chinês) Alinhamento de sentença	NIST02 (MT02) testset	878	NIST03/4/5/6/8/ 12General /08-12Progress (MT03/4/5/6/8/12/08-12)	8.809
	WMT14	4.52 mi	Inglês - Alemão	-	newstest2013	3.000	newstest2014	2.737
T3	WMT14	1.9 mi	Inglês - Alemão	BPE	-	-	newstest2014	*
	WMT14	2.0 mi	Inglês - Francês	BPE	-	-	-	*
	IWSLT 2015	0.133 mi	Inglês - Vietnamita	BPE	-	-	IWSLT 2015	1.2 k
T4	LDC (LDC2005T10, LDC2003E14, LDC2004T08 e LDC2002E18)	1.3 mi	Chinês - Inglês	Tokenização Filtragem por tamanho (100 tokens) BPE Segmentação (chinês)	NIST05	-	NIST02/3/4	-
	Uyghur to Chinese News Translation Task CCMT2019	0.17 mi	Uigur - Chinês	Tokenização Filtragem por tamanho (100 tokens) BPE Segmentação (chinês)	CWMT2018-uc-news-test	-	CCMT2019_UC_test	-
	WMT14	4.5 mi	Inglês - Alemão	Tokenização Filtragem por tamanho (100 tokens) BPE	newstest2013	-	newstest2014	-
T5	LDC	1.5 mi	Chinês - Inglês	Tokenização (inglês) Segmentação (chinês) Filtragem por tamanho (100 tokens source, 50 tokens target)	NIST02	-	NIST03/4/5	-
T6	LDC (LDC2002L27, LDC2002T01, LDC2002E18, LDC2003E07,LDC2004T08, LDC2004E12,LDC2005T10)	1.25 mi	Chinês - Inglês	Filtragem por tamanho (50 tokens) Segmentação (chinês)	NIST02	-	NIST03/4/5	-
	WMT14	4.5 mi	Inglês - Alemão	Filtragem por tamanho (50 tokens) BPE	newstest2013	-	newstest2014	-

Table 4. Tabela de detalhamento de bases por trabalho.

ID	Treino				Validação		Teste	
	Base	Tamanho	Idiomas	Processamento	Base	Tamanho	Base	Tamanho
T7	LDC (nível de sentença: LDC2002E18, LDC2003E07, LDC2003E14, partes do LDC2004T08 e nível de documento: LDC2002T01, LDC2004T07, LDC2005T06, LDC2005T10, LDC2009T02, LDC2009T15, LDC2010T03)	2 mi	Chinês - Inglês	BPE	NIST06	-	NIST02/3/4/5	-
	WMT15	4.46 mi	Inglês - Alemão	BPE	newstest2013	-	newstest2014	-
	WMT15	1.93 mi	Finlandês - Inglês	BPE	newsdev2015	-	newstest2015	
T8	CTB, PKU e MSR	2.5 mi	Chinês - Inglês	Segmentação (chinês) Filtragem por tamanho (70 tokens source, 50 tokens target)	NIST05	-	NIST02/3/4/5/6/8	-
	ASPEC	75.6 mi	Japonês - Inglês	Segmentação (japonês) Filtragem por tamanho (70 tokens source, 50 tokens target)	WAT	-	WAT	-
T9	WMT17	20.5 mi	Chinês - Inglês	Tokenização BPE Filtragem por tamanho (50 tokens)	newsdev2017	-	newstest2017	-
	WMT15	4.58 mi	Inglês - Alemão	Tokenização BPE Filtragem por tamanho (60 tokens)	newstest2013	-	newstest2014	-
T10	WMT14	4.39 mi	Inglês - Alemão	Parser (inglês) Filtragem por tamanho (50 tokens)	newstest2013	0.003.000 mi	newstest2014	0.00300.3 mi
T11	Common Crawl WebText2 Books1 Books2 Wikipedia	410 bi 19 bi 12 bi 55 bi 3 bi	Inglês - Francês Inglês - Alemão Inglês - Romeno	Pré-treino Common Crawl (treino) XLM tokenização (teste)	-	-	WMT14 e WMT16	-
T12	Crawled Corpusc	364 mi	Inglês - Alemão Inglês - Francês Inglês - Romeno		Transformer newstest2013 newstest2014 newstest2015 newstest2016	-	newstest2014 newstest2015 newstest2016	-
T13	BooksCorpus English Wikipedia	800 mi 2.5 mi		Pré-treino				
T14	WMT14	4.5 mi 0.036 mi	Inglês - Alemão Inglês - Francês	BPE Tokenização Otimização Regularização	-	-	newstest2014	-

Table 5. Tabela de detalhamento de bases por trabalho.

References

[1] ARVELOS, César Augusto et al. **Buzz trap: Identificação de abelhas usando características acústicas e inteligência artificial**. 2021.

[2] BOMMASANI, Rishi et al. **On the opportunities and risks of foundation models**. arXiv preprint arXiv:2108.07258, 2021.

[3] BROWN, Tom et al. **Language models are few-shot learners. Advances in neural information processing systems**, v. 33, p. 1877-1901, 2020.

[4] DEVLIN, Jacob et al. **Bert: Pre-training of deep bidirectional transformers for language understanding**. arXiv preprint arXiv:1810.04805, 2018.

[5] DYBÅ, Tore; DINGSØYR, Torgeir. **Empirical studies of agile software development: A systematic review**. Information and software technology, v. 50, n. 9-10, p. 833-859, 2008.

[6] HERCHONVICZ, Andrey L.; FRANCO, Cristiano R.; JASINSKI, Marcio G. **A comparison of cloud-based speech recognition engines**. Anais do Computer on the Beach, p. 366-375, 2019.

[7] HU, Ning; DANNENBERG, Roger B.; TZANETAKIS, George. **Polyphonic audio matching and alignment for music retrieval**. In: 2003 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (IEEE Cat. No. 03TH8684). IEEE, 2003. p. 185-188.

[8] KAWAKITA, Satoshi; ICHIKAWA, Kotaro. **Automated classification of bees and hornet using acoustic analysis of their flight sounds**. Apidologie, v. 50, n. 1, p. 71-79, 2019.

[9] KITCHENHAM, Barbara et al. **Guidelines for performing systematic literature reviews in software engineering version 2.3**. Engineering, v. 45, n. 4ve, p. 1051, 2007.

[10] LAGARDA, Antonio L. et al. **Translating without in-domain corpus: Machine translation post-editing with online learning techniques**. Computer Speech Language, v. 32, n. 1, p. 109-134, 2015.

[11] LI, Yachao; LI, Junhui; ZHANG, Min. **Improving neural machine translation with latent features fee dback**. Neurocomputing, v. 463, p. 368-378, 2021.

[12] PAN, Boyuan et al. **Bi-decoder augmented network for neural machine translation**. Neurocomputing, v. 387, p. 188-194, 2020.

[13] RAFFEL, Colin et al. **Exploring the limits of transfer learning with a unified text-to-text transformer**. arXiv preprint arXiv:1910.10683, 2019.

[14] SHI, Xuewen et al. **Improving neural machine translation with sentence alignment learning**. Neurocomputing, v. 420, p. 15-26, 2021.

[15] SU, Jinsong et al. **Exploiting reverse target-side contexts for neural machine translation via asynchronous bidirectional decoding**. Artificial Intelligence, v. 277, p. 103168, 2019.

[16] SU, Chao et al. **Neural machine translation with Gumbel Tree-LSTM based encoder**. Journal of Visual Communication and Image Representation, v. 71, p. 102811, 2020.

References

[17] TAN, Zhixing et al. **Lattice-to-sequence attentional Neural Machine Translation models**. Neurocomputing, v. 284, p. 138-147, 2018.

[18] VASWANI, Ashish et al. **Attention is all you need.** Advances in neural information processing systems, v. 30, 2017.

[19] WANG, Feng et al. **Hybrid attention for Chinese character-level neural machine translation**. Neurocomputing, v. 358, p. 44-52, 2019.

[20] WAY, Andy. Quality expectations of machine translation. In: Translation quality assessment. Springer, Cham, 2018. p. 159-178.

[21] WU, Yonghui et al. **Google's neural machine translation system: Bridging the gap between human and machine translation**. arXiv preprint arXiv:1609.08144, 2016.

[22] YANG, Zhen et al. **Generative adversarial training for neural machine translation**. Neurocomputing, v. 321, p. 146-155, 2018.

[23] YANG, Baosong et al. **Improving tree-based neural machine translation with dynamic lexicalized dependency encoding**. Knowledge-Based Systems, v. 188, p. 105042, 2020.

[24] ZHANG, Zhebin et al. **BERT-JAM: Maximizing the utilization of BERT for neural machine translation**. Neurocomputing, v. 460, p. 84-94, 2021.