Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Melhorando a Qualidade da Maquina de Tradução Chinês para Português com RoBERTa

Guo Ruizhe

Dissertação apresentada como requisito parcial para
conclusão do Mestrado em Informática

Orientador
Prof. Dr. Li Weigang

Brasília
2023

Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Melhorando a Qualidade da Maquina de Tradução Chinês para Português com RoBERTa

Guo Ruizhe

Dissertação apresentada como requisito parcial para
conclusão do Mestrado em Informática

Prof. Dr. Li Weigang (Orientador)
CIC/UnB

Prof.a Dr.a Maristela Terto de Holanda
CIC/UNB

Prof. Dr. Thiago de Paulo Faleiros
CIC/UNB

Prof. Dr. Zhao Liang
FFCLRP, Universidade de São Paulo

Prof. Dr. Ricardo Pezzuol Jacobi
Coordenador do Programa de Pós-graduação em Informática

Brasília, 31 de Janeiro de 2023

Dedicatória

É por causa do constante apoio e incentivo de muitas pessoas que tenho a sorte de estudar na Universidade de Brasília. Sou grato aos professores, supervisores e colegas que apoiam meus estudos na Universidade de Brasília.

Agradecimentos

A presente dissertação de mestrado não poderia chegar a esse ponto sem o precioso apoio de várias pessoas. Quero agradecer a todas as pessoas que me apoiaram desde o momento que tomei a decisão de começar o trabalho de mestrado no Brasil. Sem o apoio deles, não haveria a possibilidade de conclusão deste trabalho.

Em primeiro lugar, não posso deixar de agradecer ao meu orientador, Professor Doutor Li Weigang, por toda a paciência, empenho e sentido prático com que sempre me orientou neste trabalho e em todos aqueles que realizei durante as aulas e os seminários do mestrado. Muito obrigado pelas correções quando foram necessárias e por sempre me motivar.

Agradeço a Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pela concessão da bolsa de estudos do meu curso de mestrado.

Desejo igualmente agradecer a todos os meus professores e colegas do mestrado, do Departamento de Ciência da Computação, que foram tão importantes na minha vida acadêmica e no desenvolvimento deste trabalho. Aos Professores Alba Cristina Magalhães Alves de Melo, Thiago de Paulo Faleiros, Maristela Terto de Holanda e Célia Ghedini Ralha, aos quais tive o prazer de assistir suas aulas. Agradeço também aos meus amigos e colegas do mestrado com quem compartilhei muito durante meus estudos, obrigado Zheng Jianya, Dennis, Harley, Marcos e José, cujo os seus apoio e motivação incondicional me ajudaram a tornar este trabalho uma válida e agradável experiência de aprendizagem.

Quero agradecer à minha esposa Guo Yue pelo seu amor, partilha, companheirismo e apoio incondicional. Agradeço a sua enorme compreensão, generosidade e alegria com que me brindou constantemente, contribuindo para chegar ao fim deste percurso.

Por fim, o meu profundo e sensível agradecimento a todas as pessoas que contribuíram para a concretização desta dissertação, estimulando-me intelectual e emocionalmente.

Resumo

As mudanças contínuas na época da informação promovem o desenvolvimento do campo da tradução e da tradução automática, acompanhadas pelo surgimento da inteligência artificial, mostrando uma tendência e apresentando próspero desenvolvimento. A tradução automática é um tópico importante no processamento de linguagem natural. A aplicação da tradução automática neural na tradução automática foi revivida e desenvolvida nos últimos anos. Com a introdução de algoritmos de excelência e a melhora da capacidade de processamento dos computadores, a tradução automática neural mostrou-se com grande potencial.

Existem grandes diferenças na forma e na expressão da linguagem entre o Português e o chinês. A comunicação entre o chinês e o Português está em fase de desenvolvimento, e os materiais básicos de tradução são muito escassos. O estudo da tradução automática entre chinês e Português não só servirá no auxílio às populações de língua chinesa e portuguesa, como também é tema de suma importância para a tradução entre idiomas onde os dados básicos são escassos.

Esta dissertação apresenta um estudo sobre Tradução Automática Neurais (Neural Machine Translation) para o par de línguas Português (PT) $\leftrightarrow$ Chinês (ZH) e adiciona as direções de tradução Chinês-Português (Brasil) e Português (Brasil)-Chinês. O objetivo é buscar um modelo mais adequado entre as línguas acima com algoritmos e arquiteturas avançadas, de forma a melhorar o nível atual de tradução chinês-Português, bem como o nível de tradução Chinês-Português (Brasil).

Modelos de tradução de última geração são utilizados na tradução automática chinês-Português. O modelo RoBERTa é o mais avançado e a estrutura de segmentação mistas de palavras é usado para pré-treinamento, e o BERT é usado para tradução seguinte. No corpus paralelo chinês-Português disponível e público, seleciona o Opensubtitles2016 que tem maior quantidade dos dados. E usa BLEU e Rouge-dois indicadores de avaliação que são mais versáteis na tradução automática.

No final, obtivemos os resultados dos impactos de fatores diferentes na tradução automática chinês-Português sob os recursos existentes e um modelo melhor de tradução automática chinês-Português, ao mesmo tempo, descobrindo alguns trabalhos efetivos

que devem ser feitos no campo da tradução automática chinês-Português no futuro.

Palavras-chave: Aprendizado Profundo, RoBERTa, Português, NMT, PTMs, Redes Neurais Artificiais

Abstract

The continuous changes in the information age have promoted the development of the translation field, and machine translation, accompanied by the rise of artificial intelligence, is showing a trend of prosperity and development. Machine translation is an important topic in natural language processing. The application of neural machine translation in machine translation has been revived and developed in recent years. With the introduction of excellent algorithms and the improvement of computer computing power, neural machine translation has shown great potential.

There are big differences in the form and expression of the language between Portuguese and Chinese.The communication between Chinese and Portuguese is in the development stage, and the basic translation materials are very scarce. The study of automatic translation between Chinese and Portuguese will not only help Chinese and Portuguese speaking populations, but it is also an important topic for translation between languages where basic data are scarce.

This dissertation presents a study on Neural Machine Translation (NMT) for the language pair Portuguese (PT) ↔Chinese (ZH) and adds the Chinese-Portuguese (Brazil) and Portuguese (Brazil)-Chinese translation directions. The objective is to seek a more suitable model among the above languages with advanced algorithms and architectures, in order to improve the current level of Chinese-Portuguese translation, as well as the level of Chinese-Portuguese (Brazil) translation.

State-of-the-art translation models are used in Chinese-Portuguese machine translation. The model RoBERTa is the most advanced, the mixed word segmentation framework is used for pre-training, and BERT is used for subsequent translation. In the available and public Chinese-Portuguese parallel corpus, select the Opensubtitles2016 that has the largest amount of data. And it uses BLEU and Rouge-two evaluation indicators that are more versatile in machine translation.

In the end, we got the results of the impacts of different factors on Chinese-Portuguese machine translation under existing resources and a better model of Chinese-Portuguese machine translation, at the same time discovering some effective works that should be done in the field of Chinese-Portuguese machine translation in the future.

Keywords: Deep Learning, RoBERTa, Portuguese, NMT, Chinese, PTMs, Artificial Neural Networks

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

Atualmente existem vários idiomas no mundo, e as traduções baseadas em idiomas diferentes também estão surgindo em um fluxo interminável. O nível de pesquisa de idiomas com base na tradução automática neural tem como maior fator de influência a necessidade de comunicação entre as línguas. Ao mesmo tempo, a tradução automática neural, com maturidade, também preencherá a lacuna na comunicação causada pelas diferenças linguísticas, aumentando assim a capacidade de comunicação e criando um círculo virtuoso sustentável.

O objetivo desta dissertação é o de estudar a tradução automática neural entre Chinês e Português usando algoritmos avançados para a realização de seu uso em larga escala nas duas línguas, uma vez que ainda carecem de pesquisa. O intuito é encontrar um modelo de tradução automático mais adequado entre as duas línguas através de um algoritmo e processo de operação, de modo a melhorar o nível atual de tradução Chinês-Português.

1.1 Motivação

A linguagem e a escrita são os meios de comunicação mais utilizados pelos seres humanos e são também a base para o desenvolvimento da civilização humana, surgindo e se desenvolvendo com o desenvolvimento da sociedade. Em contextos e ambientes históricos diferentes, pessoas de diferentes regiões e civilizações inventaram e derivaram idiomas diferentes através dos períodos. Cada idioma foi desenvolvido, absorvido, alterado e isolado ao longo do tempo, dando origem não apenas a idiomas novos, mas também a diferentes dialetos e sotaques em um mesmo idioma. É preciso muito tempo e esforço para aprender um idioma.

O Processamento de Linguagem Natural (NLP) é uma disciplina transversal e está relacionada a uma série de disciplinas como linguística, informática, matemática, psicologia, teoria da informação, acústica, etc.[9]. Nos últimos anos, com o aumento do

poder computacional e a quantidade crescente de pesquisas baseadas no Processamento de Linguagem Natural (NLP), a Neural Machine Translation (NMT) tem resolvido muitos problemas de tradução entre idiomas diferentes. Uma boa precisão e rápida velocidade de tradução podem ser obtidas entre vários idiomas, como Chinês-inglês, inglês-Português, e a maior parte da tradução é quase inteiramente feita por máquinas. Devido à existência de diversos idiomas no mundo, a tradução automática neural atualmente só pode concentrar os estudos nos idiomas usados por grande quantidade de pessoas e pela frequência do uso do idioma. Existem alguns idiomas, que devido à sua geografia, desenvolvimento social e outras questões, exigem mais meios de comunicação do que a quantidade atual. É o caso entre o Chinês e o Português: são poucas as pesquisas dedicadas ao Chinês-Português com tradução automática, até mesmo o corpus básico da tradução automática Chinês-Português é muito raro. É difícil desenvolver a tradução automática entre o Chinês e o Português.

Os países de língua chinesa ocupam uma posição importante no mundo, e os países de língua portuguesa também são poderosos, como é o caso do Brasil. Há uma demanda enorme de tradução entre ambas as línguas. De acordo com as suas características, embora existam muitos dialetos em Chinês e muitos sotaques em Português, em termos de escrita, os diferentes dialetos do Chinês são relativamente consistentes na escrita, existindo apenas diferenças de pronúncia; os diferentes sotaques que existem em Português não interferem na escrita, uma vez que a escrita é quase a mesma. Portanto, para resolver o problema da tradução dos dois idiomas, o processamento de tradução da escrita é uma maneira eficaz.

Esta dissertação trata de construir um modelo de tradução automática mais adequado para a tradução Chinês-Português, baseado em material Chinês-Português existente, para contribuir tanto para os países de língua portuguesa como para os países de língua chinesa.

1.2 A Língua Chinesa

O Chinês é uma língua mais falada no mundo e os falantes da língua chinesa constituem cerca de 19% da população mundial[10], e o país com o maior número de falantes é a China. De modo geral, acredita-se que o Chinês tenha uma história de 3.000 anos. Como o surgimento da linguagem é tão antigo quanto o da escrita, a história do Chinês como língua é, obviamente, muito maior do que 3.000 anos [11]. O idioma Chinês é um dos seis idiomas de trabalho prescritos pelas Nações Unidas. O Chinês é uma escrita ideográfica e não uma escrita fonética, o que é bem diferente de outras línguas como latim. Existem dialetos no Chinês e, embora a comunicação oral entre cada dialeto não seja completamente difícil, a linguagem escrita é quase igual.

Nos últimos anos, devido à longa política de amizade da China e ao desenvolvimento rápido de regiões de língua chinesa e da China, esta região tornou-se um parceiro muito importante para o intercâmbio econômico e cultural da maioria dos países no mundo e, já em 2012, a China se tornou o maior parceiro comercial de mais de 120 países do mundo, com um valor total de 39,1 trilhões de Renminbi de importações e exportações de mercadorias em 2021[12]. A China também é o maior parceiro comercial de Portugal da Ásia[13] e o maior parceiro comercial do Brasil por 13 anos consecutivos[14]. Devido às diferenças na história, política, leis, hábitos comerciais e métodos de comunicação entre os países, uma comunicação de sucesso é particularmente importante. Nos últimos anos, a China vem fortalecendo sua própria construção cultural e cada vez mais a cooperação estrangeira adota métodos de comunicação baseados no Chinês, no entanto, aprender Chinês é difícil, portanto, o processamento de linguagem natural tornou-se um tópico de pesquisa importante.

O processamento de linguagem natural (NLP) está se desenvolvendo rapidamente no mundo. Hoje os países e regiões de língua chinesa estão se desenvolvendo rapidamente, especialmente nos aspectos econômicos e culturais. A influência internacional do Chinês está aumentando dia a dia, e os países de língua chinesa também estão na posição de liderança no campo da pesquisa em NLP. Atualmente, os resultados da tradução entre as principais línguas do mundo têm sido excelentes, mas devido à falta de materiais básicos, de escassez de pesquisadores, de localização remota e de pouco apoio à pesquisa, a qualidade da tradução automática Chinês-Português não é excelente. Hoje a pesquisa de Chinês-Português no mundo vem principalmente de Portugal, do Brasil e da China. Entre eles, Macau, China, são as principais regiões de pesquisa Chinês-Português.

No caso da China, um país de língua chinesa com uma população de mais de 1,41[15] bilhões de habitantes, e do Brasil, um país de língua portuguesa com uma população de mais de 210 milhões[16], existe contato político e econômico próximo entre os dois países, mas com uma distância de mais de 16.632 km e uma diferença de tempo de 11 horas, o tempo de comunicação oportuna entre os povos do ambos países é muito precioso.

1.3 A Língua Portuguesa

O Português é a sexta língua mais falada no mundo com cerca de 280 milhões de falantes. Existem dois tipos de Português comum: um é com hábito da língua portuguesa e o outro é com hábito da língua brasileira. Há uma pequena diferença na pronúncia entre os dois, mas basicamente não tem barreira de comunicação no dia a dia, e a diferença no uso das palavras entre os dois é menor do que a diferença entre o inglês britânico e o inglês americano.

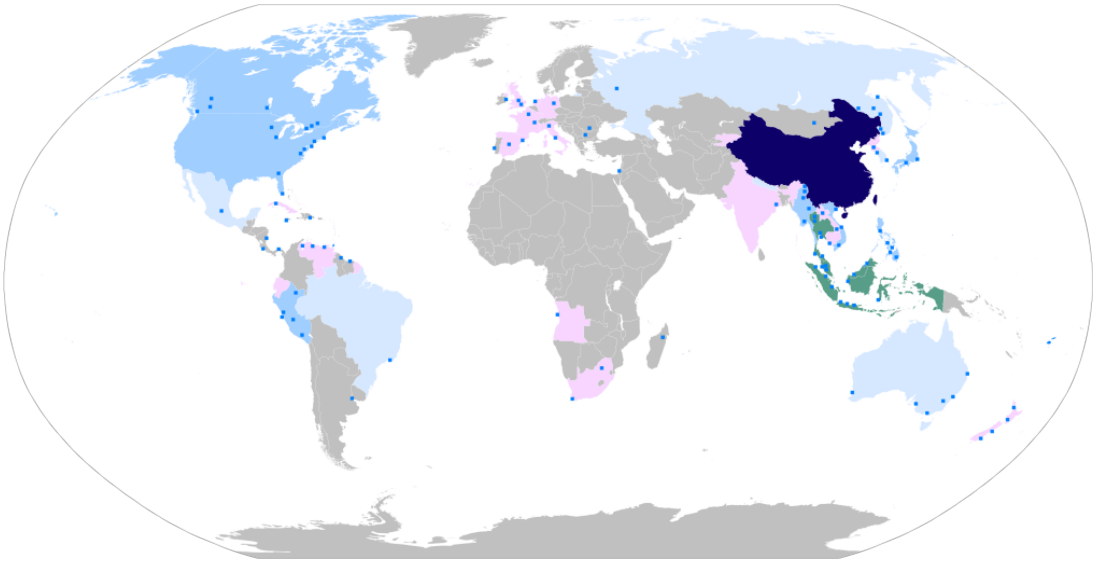

Figure 1.1: Países e regiões com maior número de falantes de Chinês do mundo (partes coloridas) da Wikipedia (2021)[1]

O Português é amplamente distribuído e falado em nove países e territórios no mundo: Portugal, Brasil, Angola, Moçambique, Cabo Verde, São Tomé e Príncipe, Guiné-Bissau, Timor-Leste, Macau e China. Têm falantes de Português na Europa, África, América, Ásia e uma pequena parte na Oceania. Atualmente, a pesquisa baseada na tradução entre Português e Chinês é mais sobre o Português de Portugal, e quase não há pesquisas sobre tradução automática entre Português brasileiro e Chinês. O Português é a língua mais falada no hemisfério sul, principalmente pelo fato de o Brasil ter mais de 210 milhões de pessoas e ser o país com o maior número de falantes de Português, representando cerca de 4/5 do total da população lusófona; o Brasil é um dos países mais importantes do mundo em termos de aspectos econômicos e culturais. Segundo estatísticas do Fundo Monetário Internacional em 2021, o PIB(Produto Interno Bruto) do Brasil ocupa a 12ª posição no mundo, e esse nível é ainda maior do que em anos anteriores.

O Português também é um importante ramo da família linguística latina, e a comunicação com outras famílias linguísticas latinas é muito frequente, havendo muitos estudos de tradução automática entre Português e outras línguas. No entanto, na tradução Português-Chinês, devido a fatores como cultura, história e distância geográfica, as pesquisas sobre tradução automática sempre foram muito raras. Existem alguns estudiosos portugueses que estão em posição de liderança na pesquisa de tradução automática Português-Chinês, mas em termos de número total de experimentos, eles não atingiram o nível em larga escala.

À medida em que há o desenvolvimento da globalização, devido ao impacto de covid-19 na economia global e no movimento populacional no ano de 2020, o desenvolvimento

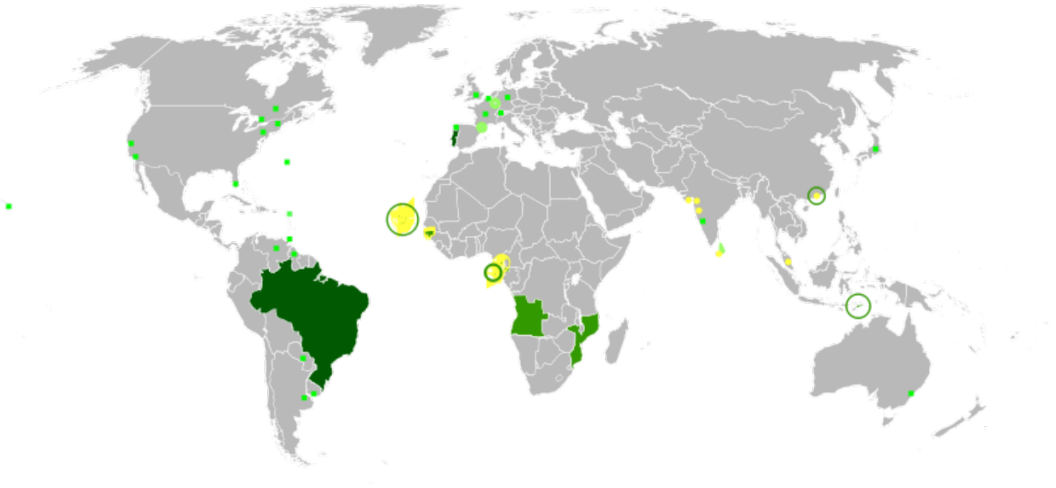

Figure 1.2: Países e territórios onde o Português é falado (incluindo segunda língua e minorias, crioulo com base no Português em amarelo) da Wikipedia[2]

econômico global e os intercâmbios culturais estão em situação de retrocesso. É um bom momento para recuperação global em 2022 e a bem-sucedida Cúpula BRICS também indica o aumento do intercâmbio e da cooperação entre a China e o Brasil. A comunicação entre a população de língua portuguesa e de língua chinesa tornou-se particularmente importante, e a tarefa de fortalecer a tradução automática Português-Chinês é iminente. Essa também é a intenção original da pesquisa sobre tradução Chinês-Português.

1.4 Português do Brasil

O Português brasileiro (pt-br) é uma variante do Português no Brasil, sendo também a língua portuguesa mais falada no mundo, com uma população de mais de 190 milhões. É falado no Brasil e nas comunidades de imigrantes brasileiros em todo o mundo.

Historicamente o Português brasileiro veio de Portugal e, após anos de desenvolvimento contínuo no Brasil, é uma língua que propriamente se autodenomina um sistema. No entanto, não foge do escopo do sistema Português, podendo ser convertido entre Português do Brasil e Português de Portugal apenas com ajustes pequenos.

O Português brasileiro não só tem algumas diferenças na pronúncia do Português de Portugal, mas também enriqueceu o conteúdo e simplificou muitas expressões incômodas no processo de evolução e desenvolvimento de longo prazo. No geral, embora a comunicação sem barreiras entre o Português brasileiro e o Português de Portugal exija apenas mudanças de sotaque, há tantas diferenças de expressão que tornam o Português brasileiro um sistema enorme e separado. Por exemplo, se o Brasil importar um filme de Portugal, normalmente, as legendas em Português serão traduzidas para o Português do Brasil antes da lançada.

Por causa do grande território e da enorme população do Brasil, o idioma oficial é baseado no sotaque de São Paulo como o sotaque padrão para unificar os sotaques de outras regiões. Até agora, regiões diferentes do Brasil ainda usam sotaques diferentes e estão em harmonia uns com os outros.

No ensino de Português em alguns países não lusófonos, o Português do Brasil é frequentemente usado como o conteúdo principal de ensino em vez do Português de Portugal. No entanto, o Português brasileiro e o Português de Portugal ainda são as mesmas línguas.

O Português do Brasil é a língua mais falada nos países lusófonos, e tem um papel importante na comunicação entre a China e os países da língua portuguesa, sendo uma língua que vale a pena estudar. Na pesquisa linguística atual, incluindo a tradução automática, o Português brasileiro é usado como objeto de pesquisa. O Português brasileiro também é uma língua importante no mundo.

1.5 Características de Chinês e de Português

O Chinês é um ideograma pictográfico e o Português é uma escrita fonética. O Chinês é muito difícil de aprender e, portanto, no processo de desenvolvimento do Chinês, ele foi simplificado continuamente. Desde 1955, a China começou a simplificar os caracteres chineses. Em 1976, a lista de caracteres chineses simplificados emitida em Cingapura era exatamente a mesma emitida na China. Assim, os caracteres simplificados também começaram a se espalhar. Em 1977, a China também tentou publicar versões simplificadas de caracteres simplificados, que foram abolidos em 1986 por vários motivos. Hoje em dia, existem cerca de 2.500 caracteres comumente usados após a simplificação e 1.000 caracteres subcomuns.

Devido à dificuldade da compreensão do Chinês, um grupo de professores chineses na China começou a estudar a expressão fonética do Chinês em 1950, que foi chamada de pinyin e foi popularizada e usada em 1956. Atualmente usada no ensino oficial na China, o pinyin se tornou uma ferramenta importante para crianças aprenderem Chinês; e em medida certa, o pinyin também pode se tornar um substituto para o Chinês, mas ainda há um longo caminho até a completa substituição. Devido à particularidade do idioma, o pinyin só pode ajudar o Chinês a se expressar em alguns casos e não pode substituir ou transmitir completamente o significado. Muitos estudos de processamento de linguagem natural em Chinês usaram pinyin, mas os resultados não são satisfatórios ainda.

Os caracteres chineses são baseados na representação gráfica e o pinyin é baseado na representação fonética. Os dois não são complementares, e o pinyin é somente usado para auxiliar os caracteres chineses.

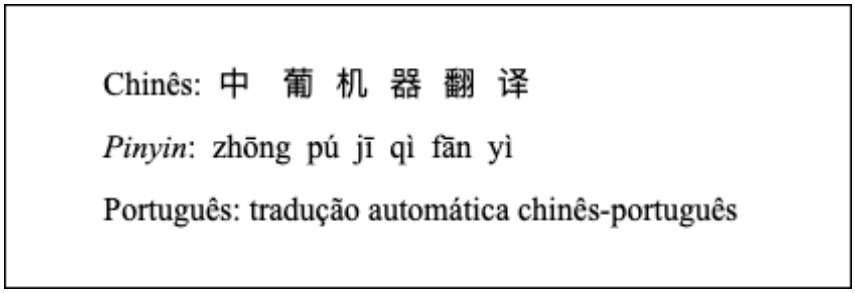

Figure 1.3: Notação pinyin chinesa

O Português é uma língua especial; por razões históricas, o Português está amplamente distribuído no mundo. Entretanto, devido a dificuldades de comunicação e razões geográficas, a comunicação ainda é insuficiente. Os falantes de língua portuguesa distinguem essa diferença como sotaque diferente, não como dialeto diferente. Depois de alguns estudiosos terem argumentado que o Português europeu foi o que mais mudou na evolução da história. A maioria dos países e regiões lusófonos na África utilizam o Português europeu, sendo que o Português no mundo tem atualmente quatro sotaques mais influentes, nomeadamente Coimbra, Lisboa, Rio de Janeiro e São Paulo.

Em termos de escrita, a diferença entre os dialetos do Chinês ou a diferença entre os sotaques do Português é pequena e pode até ser ignorada. Portanto, a dificuldade de pesquisa sobre a escrita é bem menor do que a dificuldade de pesquisa oral nas duas línguas.

A maior diferença entre o Chinês e o Português é que o Chinês é uma língua expressa graficamente, enquanto o Português é uma língua expressa foneticamente. No passado, muitos estudiosos tentaram usar o pinyin para se conectar com a linguagem dos caracteres fonéticos a fim de obter um bom resultado de pesquisa, mas nenhum deles obteve resultados experimentais satisfatórios.

Os estudos de tradução atuais entre o Chinês e a língua fonética são mais convertidos para a tradução entre caracteres chineses e o idioma de destino, ao invés de pinyin ou usando pinyin como ajuda para fazê-lo.

Há pouca diferença entre o Chinês e o Português em termos de expressão linguística, estrutura gramatical das frases e conjugação de pessoas, e as expressões semânticas entre as duas línguas são muito próximas.

Esta dissertaçõ abreviará alguns idiomas, como Chinês-(zh), inglês-(en), Português-(pt) e Português brasileir -(pt br). Na tarefa de tradução, o primeiro idioma é o idioma de origem e o segundo é o idioma de destino da tradução, por exemplo, zh-pt significa que usa Chinês como idioma de origem e Português como idioma de destino da tradução. Não será descrito especialmente a seguir.

1.6 O Estado da Arte da Tradução Automática

Foi realizada uma introdução aos dois idiomas anteriormente e, devido à diversidade e complexidade dos idiomas naturais, a tradução perfeita de um idioma para outro continua sendo uma tarefa difícil. Para o Chinês-Português, dois idiomas que não são da mesma família, realizar tradução automática seria mais representativo e poderia resolver mais tarefas de tradução similares à tradução Chinês-Português.

Atualmente, com a melhora contínua do poder de computação e o acúmulo contínuo de materiais básicos de linguagem, a tradução automática (MT) tem mostrado um grande potencial. Entre elas, a tradução automática estatística (SMT) e a tradução automática neural (NMT) foram desenvolvidas nos últimos anos e resolveram parte dos problemas de tradução. A tradução automática neural (NMT) até desenvolveu uma nova abordagem para a tradução automática. Este método só precisa de corpus paralelo bilíngue e os equipamentos de poder de computação para treinar modelos de tradução em larga escala. A qualidade das tarefas de tradução está positivamente correlacionada com o número de sessões de treinamento. Na situação atual de interação gradual do algorítmo, ele não apenas tem um valor alto de pesquisa, como ao mesmo tempo, possui fortes capacidades de industrialização. O uso da tradução automática (MT) para obter a conversão entre diferentes idiomas tornou-se um importante campo de pesquisa em processamento de linguagem natural e inteligência artificial.

Muitas empresas de Internet, como Google Tradutor, Baidu Tradutor, Microsoft Bing Tradutor, etc., realizaram tradução on-line individualizada entre diferentes idiomas. A equipe do Google desenvolveu algoritmos convencionais, como Transformer e BERT, que melhoraram muito a qualidade da tradução automática. Embora ainda não seja possível que a tradução automática substitua completamente o trabalho humano, mais pesquisadores têm se dedicado a esse tema, tornando abrangente a capacidade da perfeita e aprimorada tradução automática ano a ano.

1.7 Contexto

Esta dissetação foi pesquisada quando eu estava cursando mestrado em ciência da computação na Universidade de Brasília. Atualmente, há poucos pesquisadores estudando tradução automática Chinês-Português no mundo, a maioria estão em Macau e Portugal, e quase nenhum pesquisador do Brasil tem feito isso. Como sou Chinês e estudante no Brasil vou ao encontro de que o intercâmbio entre a China e o Brasil é muito próximo e a relação bilateral é muito importante, e que os talentos que conseguem se comunicar estão longe de serem suficientes para que os dois países cooperem completamente, e até

muitas vezes, devido à incapacidade de tradução reversa, tem que recorrer a um terceiro idioma como o inglês para comunicar, o que torna a expressão pouco clara e aumenta a dificuldade de compreensão. E é necessária a tradução totalmente manual nos documentos oficiais ou pesquisas acadêmicas. Então acho que é minha responsabilidade começar e continuar a estudar tradução automática Chinês-Português.

Pessoalmente, quero fazer algumas melhorias na tradução Chinês-Português, o que vai ajudar muita gente, seja empresa ou pessoa física, seja empresário ou estudante; desde que haja comunicação entre a população chinesa e a população portuguesa, é inevitável que eles precisem de uma ferramenta. Assim, tentei iniciar minha pesquisa e tentei trazer alguns métodos relativamente novos para a tradução automática Chinês-Português dos últimos anos.

Esta dissertação é minha primeira pesquisa sobre tradução automática Chinês-Português. No processo de tradução automática, a tradução Chinês-Português e a tradução Português-Chinês precisam ser repetidas duas vezes, pois há poucos estudos sobre tradução automática Chinês-Português na história, e os dados básicos são relativamente escassos. Selecionei alguns corpora mais adequados para pesquisa de modelos de tradução automática, a fim de obter um modelo melhor de tradução entre o Chinês e o Português.

1.8 Organização do Trabalho

O objetivo desta dissetação é tentar usar um modelo NMT para combinar modelos de tradução Chinês-Português e de tradução Português-Chinês, no caso de falta de recursos de corpus, após o treinamento do corpus, para melhorar o tempo e a precisão da tradução.

Conduziremos experimentos utilizando modelos NMT pré-treinados e treinados adequados à situação existente do corpus Chinês-Português, a fim de obter um melhor resultado experimental.

Esta dissertação está organizado da seguinte maneira.

No Capítulo 1, serão apresentados o contexto e o objetivo da o pesquisa.

No Capítulo 2, serão apresentados a motivação e o próximo plano de pesquisa.

No Capítulo 3, serão apresentados na pesquisa atual de NLP, o progresso de NMT e o modelo que escolhi.

No Capítulo 4, serão apresentados a estrutura de pesquisa e o trabalho que realizamos, bem como a fluência e operação específica do experimento, incluindo a avaliação e comparação do experimento.

No Capítulo 5, serão apresentados o processo de projeto experimental e implementação experimental.

No Capítulo 6, serão apresentados os métodos específicos de avaliação e os resultados da experiência, assim como uma análise dos resultados.

No Capítulo 7, serão apresentados a análise dos resultados experimentais e planejamento para os trabalhos futuros.

A apêndice fornece informações adicionais.

Capítulo 2

Objetivos e Planejamento da Pesquisa

Realizar a tradução automática Chinês-Português é uma tarefa demorada e trabalhosa. Se estabelecem metas e planos para melhorar o trabalho. No Capítulo 2, serão apresentados o objetivo e o processo de criação da dissertação. Na Seção 2.1, serão apresentados o objetivo do trabalho e os meios para se para atingir o objetivo. Na Seção 2.2 e Seção 2.3, serão apresentados o plano de trabalho e o tempo real utilizado para sua realização.

2.1 Objetivo

O objetivo é realizar um estudo de tradução de idiomas correspondente, uma vez que é necessário para um grande número de pessoas e têm poucos pesquisadores na área. Esta dissertação espera realizar a tradução correspondente na ausência de corpus básico, falta de pesquisa de outras pessoas e em duas línguas muito diferentes, usando apenas recursos e algoritmos disponíveis gratuitamente, e alcançar certos resultados. Será utilizado o atual modelo NMT de última geração e o modo de pré-treinamento NMT para melhorar a qualidade da tradução Chinês-Português e a qualidade da tradução Português-Chinês.

O processamento de linguagem natural é um importante tópico de pesquisa e leva muito tempo para entendê-lo e se familiarizar. Os algoritmos e modelos mais avançados são usados para selecionar os dados adequados para a pesquisa e, finalmente, aplicá-los.

E o maior fator da pesquisa é o de encontrar os dados básicos apropriados – o corpus paralelo em Chinês e Português, o que é fundamental para a pesquisa. Um corpus excelente deve ser de alta qualidade, e devido à particularidade de direção da pesquisa, a pesquisa requer um corpus com quantidade moderada, não muito pequeno, qualidade moderada e sem erros grandes. E na pesquisa, não é necessário obter o melhor corpus, pois por melhor que seja o corpus, os recursos do corpus são escassos e insuficientes. Assim, a pesquisa requer recursos de compromisso para métodos de pesquisa mais práticos.

Depois disso, deve-se encontrar um método adequado para calcular a conexão entre os dois idiomas para obter um efeito de tradução mais preciso. Deve-se dar mais prioridade aos prós e contras do método principal e, em seguida, Deve-se selecionar os dados básicos de acordo com o método principal que escolho e outros métodos auxiliares, para tentar garantir que o método obtido seja o melhor.

Para os dados, é necessário encontrar um método de segmentação de palavras adequado e, em seguida, escolher um dispositivo de segmentação de palavras mais adequado. Como a segmentação de palavras em Chinês tenha relativamente mais dificuldade, o dispositivo de segmentação de palavras deve ser mais adequado para o Chinês. E é preciso separar as frases complexas de Chinês e Português do dicionário sem demarcação.

Por último, use uma excelente métrica de avaliação para avaliar se o trabalho que estou fazendo está obtendo os resultados desejados.

Levando em consideração o que foi dito anteriormente, foram estabelecidos os seguintes objetivos potenciais para trabalho levando a esta dissertação:

- Adquirir conhecimentos sobre Processamento de Linguagem Natural(NLP), em geral, e os modelos relacionados à NMT em particular;
- Conhecer o processo que outros pesquisadores estão trabalhando e identificar áreas para melhoria;
- Coletar corpora paralelos mais usados e escolher o mais adequado;
- Buscar um tokenizer adequado a separar frases para construir dicionários e modelos de entrada;
- Encontra o caminho certo para se adaptar ao ambiente Chinês-Português,Português-Chinês existente;
- Encontra os modelos, especialmente modelos relacionados pré-treinados, e contexto adequado de treinamento;
- Treina e melhorar o modelo para obter um resultado mais satisfatório;
- Encontra as métricas adequadas de avaliação e comparações para verificar o desempenho do modelo;
- Com base nos resultados da avaliação, resumir e desenvolver um sistema de tradução entre o Chinês e o Português.

2.2 Plano

Para atingir os objetivos acima mencionados, um conjunto de diretrizes foi definido previamente ao início dos trabalhos que conduziram a esta dissertação. O plano inicialmente proposto era:

- Adquirir conhecimento sobre os fundamentos do NLP, bem como sua evolução.

Adquirir conhecimento sobre os fundamentos do NMT, bem como os métodos atuais de pesquisa convencionais.

• Adquirir conhecimento sobre os atuais métodos comuns de tradução automática Chinês-Português e levar em consideração a existência de outros métodos de excelência.

• Coletar os corpora paralelos Chinês-Português atualmente disponíveis.

• Definir em qual equipamento posso executar os experimentos e escolhar o mais apropriado em custo para aprender a usá-lo.

• Começar a experimentar e ajustar modelos, corrigir bugs.

• Avaliar, modificar e melhorar o modelo e comparar o desempenho do modelo.

• Analisar modelos e escrever dissertação.

2.3 Execução do Plano

Para implementar o plano, no início, verifica literatura.[17, 6, 18, 19, 20] Com a introdução e conhecimento do ramo, aprende alguns conceitos superficiais e depois busca alguns livros relevantes para os fundamentos e princípios da NLP, o que levou mais tempo do que esperava. Passe 1 mês para concluir a pesquisa sobre NLP.

Depois disso, investica a NMT da mesma forma que investica a NLP, que também é ponto de conhecimento chave na pesquisa. Também leva um mês para adquirir o conhecimento e os algoritmos de ponta da NMT. Começa a seguir alguns cases e tenta fazer alguns procedimentos simples.

Quando pesquisa a tradução automática Chinês-Português, pois atualmente são poucos os estudos e o método de pesquisa principal é usar uma terceira língua como meio para obter melhores resultados do que a tradução direta. A literatura relevante e decide fazer uma tradução direta do Chinês para o Português.

Depois disso, colota alguns corpora paralelos Chinês-Português. O corpus Chinês-Português que pode ser obtido gratuitamente é muito raro e a escala da coleção também é relativamente pequena. Existem alguns estudos sobre corpora paralelos Chinês-Português, mas eles não são totalmente abertos, então só obtive alguns corpora paralelos públicos, o que por acaso é suficiente para sustentar pesquisa.

Então escolher e investigar em que tipo de dispositivo a pesquisa deveria funcionar, e como o aprendizado profundo tem altas exigências de hardware, considera que algumas das tecnologias mais avançadas seriam usadas. Depois de experimentar dispositivos GPU Windows, dispositivos de chip IOS M1, Plataforma Google Cloud, Plataforma Amazon Cloud, Plataforma JiuTian Cloud e Plataforma AutoDL, escolhe JiuTian e AutoDL Plataform como plataforma de execução experimental. Houve muito trabalho interativo realizado aqui e levou mais de 2 meses.

Depois disso, começa a coletar dados, testar o programa, executá-lo e ajustá-lo, melhorar o modelo e corrigir os erros do programa. Isso levou 4 meses. Ao mesmo tempo, começa a trabalhar em dissertação.

Depois de obter os dados experimentais, para testar o bom resultado dos experimentos, traz os dados gerados por modelos diferentes e conjuntos de testes diferentes para o modelo de avaliação separadamente e obte as notas de avaliação, que em vez disso funcionaram mais rápido do que o planejado e leva apenas 1 mês.

A lista fornece uma visão geral do tempo necessário para a realização das diversas entradas do planejamento. Embora o tempo necessário para algumas tarefas tenha mudado em relação ao plano, o plano foi concluído sem nenhuma omissão.

Capítulo 3

Trabalhos Relacionados

Ao conduzir a pesquisa, investiga-se primeiramente o status atual da tradução automática Chinês-Português. Neste capítulo, a pesquisa existente é elaborada. Na Seção 3.1 desta dissertação é apresentado o progresso da pesquisa da tradução de linguagem natural e é elaborado o desenvolvimento da tradução automática neural e os algoritmos de última geração atuais. Na seção 3.2, é apresentado o estado atual de desenvolvimento do corpus, dentre os quais o corpus paralelo Chinês-Português é um dado básico importante para apoiar a pesquisa. Na seção 3.3, é apresentado o progresso da pesquisa e os métodos de pesquisa da tradução automática Chinês-Português, bem como os métodos adotados pelos pesquisadores.

3.1 Tradução de Linguagem Natural

Atualmente,a tradução automática (MT) tornou-se o principal método para tradução no mundo. Com a melhora contínua da capacidade de processamento pelos computadores,a realização do Deep Learning tornou-se um tópico de pesquisa prático. Nos últimos anos, a tradução automática neural (NMT) foi muito interada. Esta dissertação utiliza o modelo NMT de última geração para realizar a tradução entre o Chinês e o Português.

Este capítulo vai apresentar o desenvolvimento e a situação atual da tradução automática e gradualmente descrever o processo de evolução e os motivos do algoritmo. Na seção 3.1.1 descreve brevemente o processo de desenvolvimento da MT e, em seguida, durante as seções 3.1.2-3.1.7 serão apresentados o processo de desenvolvimento e os principais nós da evolução da NMT.

3.1.1 Tradução Automática (MT) e Tradução Automática Estatística (SMT)

O conceito de tradução automática tem uma longa história e pode ser rastreado até o conceito de línguas internacionais. Naquela época, as pessoas esperavam que o significado principal de várias línguas pudesse ser convertido indiretamente através de uma linguagem semântica intermediária para alcançar o efeito de tradução mútua. Em 1954, um experimento do IBM [21] traduziu com sucesso uma pequena quantidade de palavras em russo para o inglês, realizando a tradução automática real pela primeira vez. Durante a pesquisa e contínuo desenvolvimento da tradução automática, a capacidade de processamento dos computadores foi um dos principais fatores que promoveram o progresso da pesquisa. Foi somente por volta de 1980 que os computadores começaram a ser amplamente utilizados e a tradução automática começou a entrar em alto nível.

Baseado no desenvolvimento do ramo da estatística, a tradução automática estatística (SMT) tornou-se a principal força na tradução automática por alguns anos. Em 1949, W. Weaver [22] propôs a ideia básica da tradução automática estatística. O IBM propôs 5 modelos de tradução palavra-a-palavra no tempo de pesquisa subsequente, o que em muito contribuiu para o desenvolvimento da SMT.

A Neural Machine Translation (NMT) pode ser rastreada até o algoritmo Perceptron proposto por Rosenblatt em 1957, que é a rede neural mais simples. No entanto, devido à estrutura simples do Perceptron, ele é incapaz de resolver alguns problemas complexos, como o problema da inseparabilidade linear, de modo que a pesquisa não alcançou o desenvolvimento suficiente. Até que o algoritmo de retropropagação (Backpropagation, BP) e o multilayer perceptron (Multilayer Perceptron, MLP) fossem utilizados, os pesquisadores o chamaram de Feedforward Neural Network, FNN. Com o desenvolvimento de GPU (Graphics Processing Unit), alguns algoritmos que exigem mais cálculos puderam ser realizados, e o NMT gradualmente começou a substituir o SMT como o método principal de tradução automática (MT).

Após o surgimento do BLEU[6] (índice de avaliação), a pesquisa do SMT passou a ter uma referência para comparação horizontal, evitando muito esforço e gasto de comparação manual. Gradualmente, a pesquisa do SMT foi compelida para uma altura mais alta. O SMT conta a frequência de palavras do corpus de acordo com a probabilidade e no processo de tradução, a palavra com a maior frequência de uso é usada para tradução mútua. Embora existam alguns problemas, a capacidade de tradução é excelente, portanto, antes do surgimento de neuro tradução automática, o SMT vem ocupando quase todos os mercados da tradução automática. Depois de 2016 a maioria dos pesquisadores e empresas comerciais começaram a recorrer à tradução automática neural (NMT).

Para comparar as características da tradução automática estatística e da tradução automática neural, se obteve a seguinte comparação, que será elaborada na seguinte apresentação:

Métricas de avaliação	NMT	SMT
Método de exibição	Continuamente	Disperso
Modelo	Não linear	Log linear
Tamanho do Modelo	Pequeno	Grande
Tempo de treino	Muito tempo	Tempo curto
Interpretabilidade do modelo	Poderoso	Fraco
Pegada de memória	Pequena	Grande
GPU	Necessário	Não é necessário
Modo de treinamento incremental	Suporte	Não suporta

Table 3.1: As diferenças entre NMT e SMT[7]

3.1.2 Modelo de Linguagem Probabilística de Rede Neural (NLPM)

O modelo de linguagem probabilística de rede neural foi proposto por Bengio [23] e seu colega em 2003. Ao contrário da SMT tradicional, ele propôs um importante conceito de "vetor de palavras", que trouxe a tradução automática para um novo nível. A ideia principal do modelo de linguagem tradicional é encontrar a frequência de ocorrência de palavras e frases primeiro no conjunto de treinamento e, em seguida, usar o método da teoria da probabilidade para calcular a probabilidade. Embora esse método tenha se mostrado eficaz e tende a alcançar um bom efeito, têm alguns problemas que não podem ser resolvidos atualmente. O exemplo mais típico são palavras de baixa frequência. Supondo que existam as seguintes frases em um grande conjunto de treinamento:

Canecas de vidro são frágeis.

Canecas de cerâmica são frágeis.

No processo de coleta de frequência de palavras na maioria dos corpora, assume-se que "canecas de vidro" aparece 500 vezes, enquanto "canecas de cerâmica" aparece apenas 5 vezes. Nesse caso, no processo de cálculo do modelo de linguagem tradicional, "canecas de cerâmica" " será coberto por "canecas de vidro", e nos resultados, apenas "canecas de vidro" aparecerá. Sob o conceito de vetor de palavras proposto por Bengio et al., os dois obterão vetores de palavras semelhantes mas diferentes, em diferentes cenários de uso de frequência alta e média, que obterão um método para resolver palavras de frequência baixa, ampliando o escopo de uso. Sua pesquisa não apenas melhorou o problema da escassez de dados em SMT, mas também lançou as bases para o uso de redes neurais no campo da tradução automática.

3.1.3 Tradução Automática Estatística(SMT) e Tradução Automática Neural (NMT)

A Tradução Automática Estatística(SMT) e Tradução Automática Neural (NMT) adotam a ideia de solução probabilística ao resolver problemas de tradução, e a diferença está principalmente no método de implementação. SMT expande a probabilidade estatística de acordo com o princípio bayesiano para obter a seguinte fórmula[24]:

$$p(\boldsymbol{t} \mid \boldsymbol{s}) = \frac{p(\boldsymbol{t})p(\boldsymbol{s} \mid \boldsymbol{t})}{p(\boldsymbol{s})} \tag{3.1}$$

O $p(s)$ na fórmula (3.1) representa a probabilidade da sentença do idioma de origem,o $p(t)$ representa a probabilidade a priori,t e s representam condição e previsão, respectivamente. Na tarefa de treinamento do modelo de tradução, $p(s)$ é um valor fixo. A solução geral é maximizar o produto à direita para obter $p(t|s)$ um valor máximo ou seja, resolver:

$$\hat{t} = \arg\max_{t} p(\boldsymbol{t})p(\boldsymbol{s} \mid \boldsymbol{t}) \tag{3.2}$$

respectively. $\hat{t}$ denotes the translation output with the highest translation probability. $p(s|t)$ is usually decomposed using the log-linear model:

$$\hat{\mathbf{t}} = \arg\max_{\mathbf{t}} \frac{\exp\left(\sum_{i=1}^{I} \lambda_i h_i(\mathbf{s}, \mathbf{t})\right)}{\sum_{\mathbf{t}'} \exp\left(\sum_{i=1}^{I} \lambda_i h_i\left(\mathbf{s}, \mathbf{t}'\right)\right)} \tag{3.3}$$

$h_i(.)$ indica o recurso de translação e λ_i é seu peso ideal correspondente, que é aprendido maximizando com um conjunto de desenvolvimento. I indica o número de função total.

Em tarefas estatísticas de tradução automática, o problema geralmente é decomposto em vários submódulos, como modelo de linguagem, modelo de tradução, modelo de sequenciamento etc., e depois combinado com um modelo log-linear para finalmente obter o resultado da tradução.

O NMT é um método para gerar mapeamentos entre linguagens naturais usando redes neurais de aprendizado profundo. Ao contrário do método de representação discreta da tradução automática estatística, a tradução automática neural (NMT) usa representação de espaço contínuo (Continuous Space Representation) para representar palavras, frases e sentenças. Em termos de modelagem de tradução, não há necessidade das etapas de tradução automática estatística, como alinhamento de palavras e extração de regras de tradução, e a rede neural é totalmente usada para concluir o mapeamento do idioma

de origem para o idioma de destino. O modelo NMT é otimizado com base no modelo SMT, o que é orientado por uma estrutura de codificador-decodificador. O Codificador-Decodificador é uma estrutura de modelo comum em aprendizado profundo. O texto é convertido em um vetor fixo pelo codificador, o texto de entrada é convertido em uma representação vetorial e, em seguida, o vetor é gradualmente decodificado pelo decodificador.

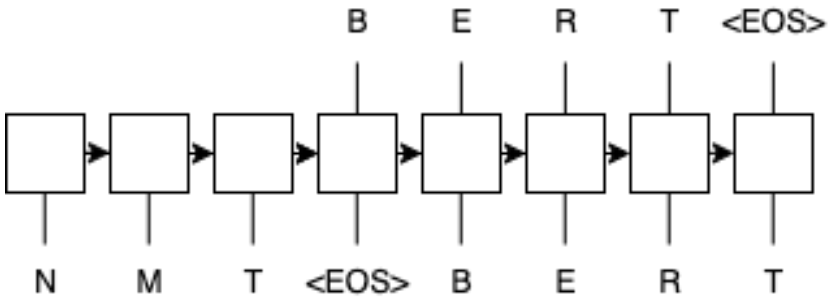

Figure 3.1: Modelo Ponta a Ponta

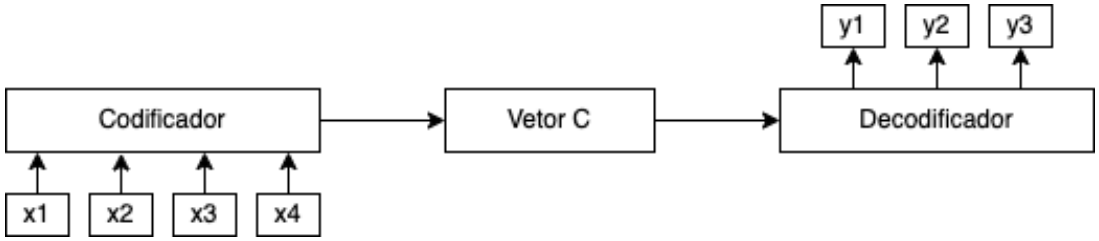

Figure 3.2: Arquitetura do Codificador-Decodificador

Devido às limitações da capacidade de processamento computacional e dados de linguagem, as primeiras redes neurais foram usadas principalmente em modelos de linguagem de tradução automática estatística (SMT), alinhamento de palavras, extração de regra de tradução, etc.[25] e em 2013, uma nova arquitetura foi proposta por Nal Kalchbrenner e Phil Blunsom [26], que implementa a tradução automática por meio de codificação de rede neural convolucional (CNN) e decodificação de Recurrent neural network(RNN) de ponta a ponta. Sua pesquisa é vista como o início da NMT.

Em 2014, Sutskever et al.[18] e Cho et al.[19][20] desenvolveram uma abordagem sequência a sequência (seq2seq), que difere daquela proposta por Nal Kalchbrenner (2013). o RNN é usado na codificação e decodificação, ao mesmo tempo em que a Long short-term memory (LSTM) é introduzida. O LSTM é uma variante do RNN e o problema da explosão de gradiente é resolvido melhor pelo LSTM. Comparado com o NMT com arquitetura RNN comum, o NMTs podem ter um melhor desempenho em sequências

mais longas com a adição de LSTM. E a introdução do LSTM aumentará incerteza e complexidade no caso de sentenças muito longas.

3.1.4 Atenção

O Encoder-Decoder é um framework muito bom, mas tem uma grande limitação. Por exemplo, existe apenas um vetor semântico de comprimento fixo entre o codificador e o decodificador, e o vetor semântico não pode conter todas as informações. Ao mesmo tempo, quanto maior o comprimento da frase, mais o vetor semântico é coberto, resultando em informações insuficientes de entrada o que, por consequência, deixa a decodificação imperfeita.

O que realmente torna a NMT o mainstream da tradução automática é a introdução de um mecanismo de atenção no modelo NMT. O mecanismo de atenção não foi originalmente usado no campo da NLP, mas foi proposto pela DeepMind na resolução de problemas de imagem. Bahdanau [3] e outros introduziram o mecanismo de atenção no campo da PNL em 2014 pela primeira vez. Em 2017, o Google propôs o mecanismo de autoatenção para tornar o mecanismo de atenção mais perfeito.

A adição do mecanismo de atenção altera a conexão do colar de comprimento fixo entre o Codificador-Decodificador em um único vetor e o passa para o decodificador, permitindo que o decodificador acesse o estado de cada codificador, obtendo assim informações válidas e reduzindo a semântica.

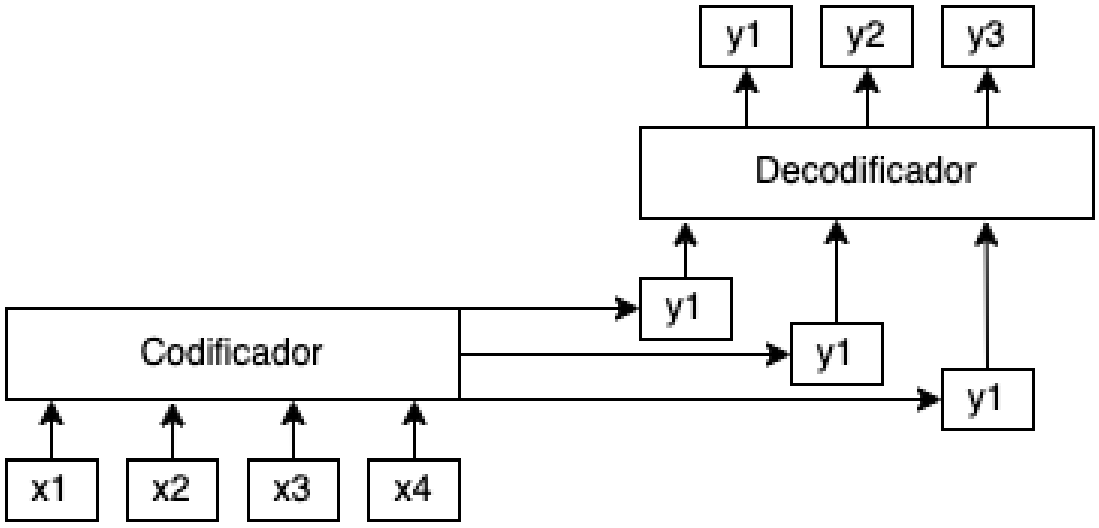

Figure 3.3: Arquitetura do Atenção

O mecanismo de Atenção é mostrado na figura (3.3). A característica do modelo de Atenção é que o Encoder não mais codifica toda a sequência de entrada em um "vetor intermediário "C" de comprimento fixo, mas o codifica em uma sequência de vetores.

Desta forma, quando cada saída é gerada, a informação transportada pela sequência de entrada pode ser totalmente utilizada. E este método alcançou resultados muito bons em tarefas de tradução.

A correspondência de vocabulário bilíngue alcançada pelo mecanismo de Attetion é chamada de alinhamento suave (Soft-alignment). Comparado com o método de alinhamento rígido da tradução automática estatística, esse método não limita o comprimento do alinhamento das palavras do idioma de destino e das palavras do idioma de origem, o que pode evitar problema no método de alinhamento rígido.

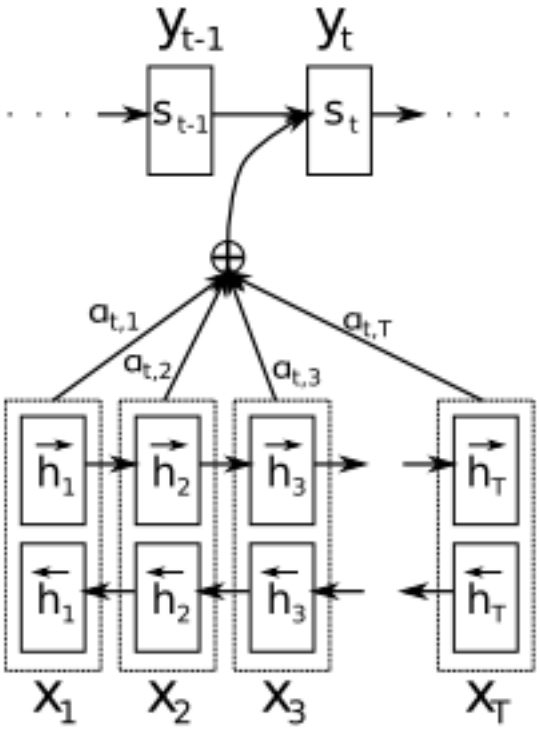

Figure 3.4: O modelo de tradução automática proposto por Bahdanau et al.[3]

A Figura(3.4) mostra o modelo de tradução automática proposto por Bahdanau et al.[3], no qual o codificador utiliza uma rede neural recorrente bidirecional. Eles são a linguagem fonte de entrada direta e reversa respectivamente. No RNN direto, os dados são inseridos em ordem, de modo que o j-ésimo estado da camada oculta pode carregar apenas $h_j^{\rightarrow}$ palavras e algumas informações anteriores dela. Na RNN reversa, os dados são inseridos na ordem inversa, que $h_j^{\leftarrow}$ contém as informações da j-ésima palavra e seguintes. Se esses dois estados de camada oculta forem combinados, $h_j = \left[h_j^{\rightarrow}, h_j^{\leftarrow}\right]$conterá as informações antes e depois da entrada j-ésima.

O desenvolvimento do mecanismo de atenção também está sendo aprimorado nos últimos anos. A primeira introdução do mecanismo de atenção, as palavras-alvo, são

independentes e não podem memorizar as partes "traduzidas" e "não traduzidas". Além disso, o mecanismo de atenção tem uma quantidade relativamente grande de computação e falta de informações de restrição. Os mecanismos de Atenção, por exemplo, Atenção Global, Atenção Local e alguns de Atenção que integram informações de restrição, estão constantemente aprimorando o próprio mecanismo de atenção.

No processo de decodificação, primeiro é necessário calcular o grau de correlação entre cada estado da camada oculta do codificador $h_1 \sim h_T$ e o estado da camada oculta do decodificador s_{t-1} e executar uma operação de normalização softmax para obter o peso a_{ij} de cada camada vetorial oculta.Calculado da seguinte forma:

$$\begin{aligned} e_{ij} &= a\left(s_{i-1}, h_j\right) \\ &= v_a^T \tanh\left(W_a s_{i-1} + U_a h_j\right) \\ a_{ij} &= \frac{\exp\left(e_{ij}\right)}{\sum_{k=1}^{T} \exp\left(e_{ik}\right)} \end{aligned} \tag{3.4}$$

Em equação(3.4) e_{ij} representa a correlação entre a i-ésima saída anterior do estado da camada oculta s_{i-1} e o j-ésimo vetor da camada oculta de entrada h_j.O valor de peso normalizado a_{ij} pode ser obtido passando e_{ij} para a função cálculo softmax.

Em seguida, a soma ponderada é realizada em $h_1 \sim h_T$ para obter o vetor de codificação c_i da sequência de entrada $(x_1, x_2, x_3 ... x_T)$ correspondente a esta decodificação. A fórmula de cálculo é a seguinte:

$$c_i = \sum_{j=1}^{T} a_{ij} h_j \tag{3.5}$$

Finalmente, decodifique de acordo com o vetor de codificação c_i, primeiro calcule o estado da camada oculta s_i no momento do decodificador i e, em seguida, calcule a saída y_i do decodificador no momento i. A forma de cálculo é a seguinte:

$$y_i = g\left(y_{i-1}, s_i, c_i\right) \tag{3.6}$$

A etapa mais importante do mecanismo de Atenção é gerar vetores diferentes de codificação de linguagem a cada momento, indicando quais partes da sequência de entrada focar na próxima saída e, em seguida, gerar a próxima saída de acordo com a área de interesse.

O mecanismo de Atenção é um mecanismo que busca o aprendizado da atenção humana. Quando os seres humanos prestam atenção à área alvo, haverá um foco e uma área de atenção para obter informações mais detalhadas e, ao mesmo tempo, reduzir e

evitar a obter informações desnecessárias e inúteis. Esta é a maneira como os humanos obtêm informações, o que melhora muito a precisão e a velocidade do acesso abrangente à informação. O mecanismo de atenção é ponderar cada parte da sentença fonte em cada processo de decodificação e, em seguida, determinar a saída do decodificador por meio de diferentes pesos.

O desenvolvimento do mecanismo de Atenção também está se aprimorando nos últimos anos. A primeira introdução do mecanismo de Atenção, as palavras-alvo, são independentes e não podem memorizar as partes "traduzidas" e "não traduzidas". Além disso, o mecanismo tem uma quantidade relativamente grande de computação e falta de informações de restrição. Os mecanismos de Atenção, por exemplo Atenção Global, Atenção Local e alguns de Atenção que integram informações de restrição, estão constantemente aprimorando o próprio mecanismo de Atenção.

No mecanismo de Atenção foi desenvolvido o mecanismo de Auto Atenção por meio de ajustes e modificações contínuas. O mecanismo comum de Atenção ocorre na semelhança entre uma palavra na sentença Alvo de saída e cada palavra na sentença Fonte de entrada. Como o nome sugere, Auto Atenção não se refere ao mecanismo de Atenção entre Alvo e Fonte, mas ao mecanismo de Atenção que ocorre entre os elementos internos da Alvo ou entre os elementos internos do Fonte.

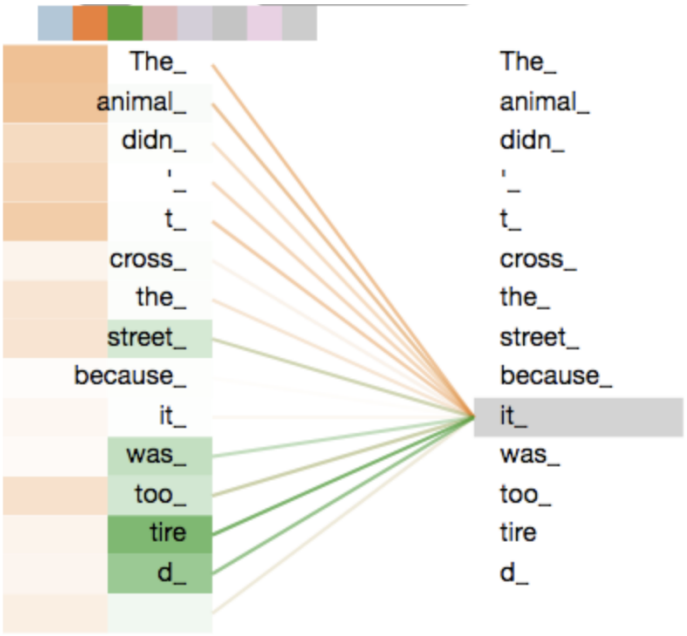

Figure 3.5: Auto Atenção[4]

A figura(3.5) mostra que Auto Atenção capta os traços semânticos entre as palavras de uma mesma frase, e o objeto de referência de "it" é "the animal".

Durante o processo de cálculo, Auto Atenção conectará diretamente a conexão entre quaisquer duas palavras na frase por meio de um resultado de cálculo, de modo que a distância entre os recursos dependentes de longa distância seja bastante reduzida, o que é propício para o uso eficaz desses recursos. Além disso, Auto Atenção também ajuda diretamente a aumentar o paralelismo dos cálculos. Portanto, esta também é a principal razão pela qual Auto Atenção está gradualmente sendo amplamente utilizada.

3.1.5 *Transformer*

Em 2017, a equipe do Google propôs o modelo Transformer, que substituiu gradualmente os modelos RNN, como o LSTM, e se tornou o método preferido no campo da NLP. No Transformer, a CNN e a RNN no conceito tradicional são abandonadas, e toda a estrutura da rede é completamente composta por Auto Atenção e Rede Neural Feed Forward.

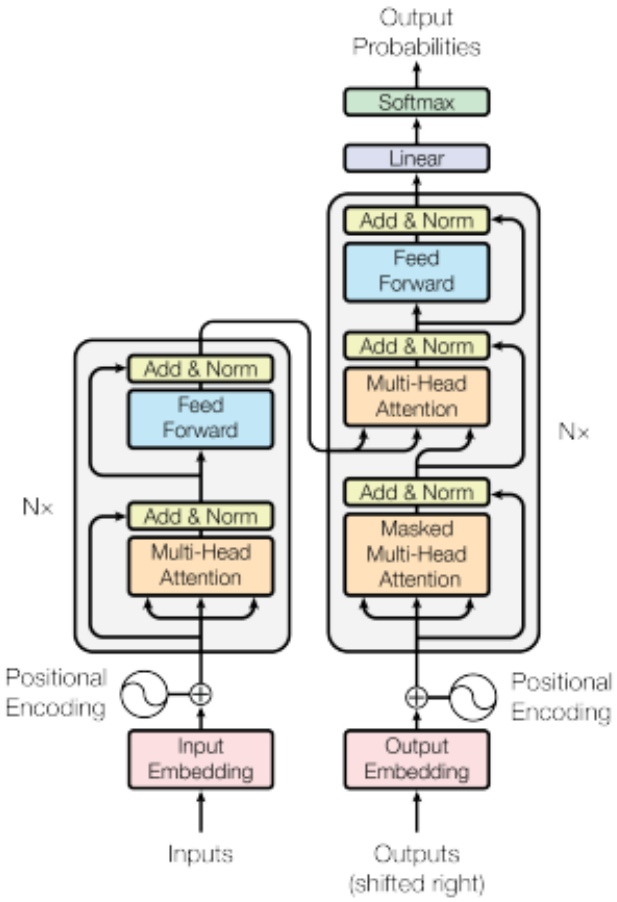

Figure 3.6: *Transformer*[5]

A essência do *Transformer* é uma estrutura Encoder-Decoder. As partes do Codificador e do Decodificador usam camadas de autoatenção empilhadas e totalmente conec-

tadas por pontos. No *Transformer*, a unidade básica é a unidade de atenção do produto escalar. Existem várias cabeças de atenção em cada camada, e cada cabeça de atenção representa a atenção de diferentes tags. A potência e a atenção das cabeças longas podem calcular pesos diferentes para diferentes "correlações".

Dois mecanismos diferentes de Atenção são mostradas na figura(3.7). A função de Atenção em um conjunto de consultas simultaneamente, agrupadas em uma matriz Q. As chaves e valores também são agrupados nas matrizes K e V.

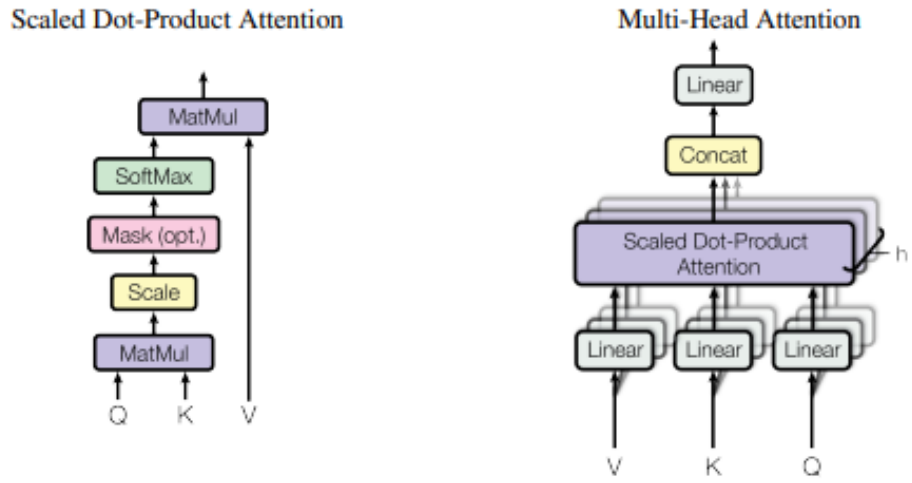

Figure 3.7: Duas atenções diferentes[5]

Existem 2 tipos de métodos de cálculo dos Mecanismos de Atenção.

$$\begin{aligned} \text{MultiHead}(Q,K,V) &= \text{Concat}\left(\text{ head }_1,\ldots,\text{ head }_h\right)W^O \\ \text{head }_i &= \text{ Attention }\left(QW_i^Q, KW_i^K, VW_i^V\right) \end{aligned} \tag{3.7}$$

O método utilizado pela Auto Atenção é o produto escalar escalado, então:

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{3.8}$$

d_k é o número de matrizes Q, K, ou seja, dimensões vetoriais. As unidades de Auto Atenção são processadas em paralelo através da unidade de atenção multicabeça.

Na etapa Encoder de Transformer, o número passará pelo módulo de Auto Atenção para obter um vetor ponderado e, em seguida, enviará o vetor ponderado obtido para a Rede Neural Feed Forward, repetirá o processo por 6 vezes e, por fim, obterá a saída de toda a parte de codificação. Na etapa Decoder, primeiro é calculado o score de Au-

toatenção para a saída. Diferentemente da parte do Encoder, após a realização da Autoatenção, a saída do Autoatenção e a saída do módulo Decodificadores são calculadas uma vez na pontuação do mecanismo de atenção e, em seguida, inserida a Rede Neural Feed Forward.

O modelo Transformer melhora a pontuação BLEU (3,8) do NMT. A chave para a melhoria de desempenho do projeto Transformer é definir a distância entre duas palavras como 1, o que é eficaz para resolver o espinhoso problema de dependência de longo prazo em NLP. Com base na ideia do Transformer, muitos modelos poderosos como BERT e GPT-2 foram produzidos posteriormente.

3.1.6 BERT (*Bidirectional Encoder Representation From Transformers*)

O BERT é gerado com base no Transformer. Com base no Transformer, o método Pretrain é adicionado, e o *Masked Language Model(MLM)* e a *Next Sentence Prediction(NSP)* são usados para obter palavras-alvo e frases precisas por meio do pré-treinamento.

Durante o pré-treinamento, o BERT utiliza o método *Masked Language Model*, que é chamado de tarefa Cloze (Taylor, 1953) para fazer Mask aleatoriamente, 15% das palavras do corpus de treinamento. Nas palavras selecionadas, 80% adicionam [MASK], 10% de palavras de substituição e 10% sem modificação; os métodos diferentes são usados para realizar operações de MASK nas palavras selecionadas. As operações de MASK são as seguintes:

- Use [MASK] para uma parte das palavras selecionadas

  Eu sou estudante da UNB → Eu sou [MASK] da UNB

  essa parte das palavras corresponde a 80% das palavras selecionadas;

- Substitua parte da palavra selecionada:

  Eu sou estudante da UNB → Ela sou estudante da UNB

  esta parte da palavra corresponde a 10% da palavra selecionada;

- Uma parte das palavras selecionadas permanece inalterada:

  Eu sou estudante da UNB → Eu sou estudante da UNB

  esta parte das palavras representa 10% das palavras selecionadas.

O BERT usa isso para obter resultados de treinamento mais precisos, mas também devido à operação de MASK de 15%, requer treinamento repetido para concluir cada frase.

Em seguida, será realizada a *Next Sentence Prediction*. Há muitas questões importantes que requerem uma compreensão da relação entre os contextos. A tarefa de pré-treinamento gerará uma sentença de teste, combinada com a operação MASK, para fornecer treinamento para o modelo, enquanto a frase de teste tem uma probabilidade geral de estar correta, a outra metade da probabilidade é o texto incorreto, por exemplo:

- Input = [CLS] Há uma montanha [MASK] longe[SEP]

  Há um rio sob [MASK] montanha [SEP]

  Label = IsNext

- Input = [CLS] Há [MASK] montanha ao longe [SEP]

  suco de maçã [MASK] doce [SEP]

  Label = NotNext

Como é muito treinamento para a capacidade de processamento de computação do computador, para acelerar o treinamento, 90% dos comprimentos de treinamento são 128 e os 10% restantes usarão comprimentos de 512. O BERT estabeleceu um novo recorde em 11 tarefas de NLP, que é uma das razões pelas quais o NMT substituiu completamente o SMT como a escolha primeira para o MT. As características do BERT são tendenciosas para dados grandes, modelos grandes e sobrecarga computacional grande.

O BERT é um algoritmo excelente, mas precisa de dados enormes como suporte e, também, requer um poder de computação extremamente forte e um tempo muito longo para o processamento. Grandes projetos usam o BERT como algoritmo principal e seu desempenho é excelente.

3.1.7 RoBERTa*(A Robustly Optimized BERT)*

Em julho de 2019, RoBERTa[27], um modelo novo baseado em BERT, foi publicado pelo Facebook e pela Universidade de Washington, que obteve melhora de resultados em alguns rankings de idiomas.

RoBERTa é um novo modelo ajustado baseado no algoritmo BERT por Liu Y, Ott M, Goyal N, et al (2019) [27] . Com base no BERT, RoBERTa propõe um método de treinamento e pré-treinamento mais avançado.

Em termos de dados de treinamento, o RoBERTa usa material de treinamento com uma escala maior do que o BERT. Em termos de parâmetros de treinamento, RoBERTa adiciona o batch-size de treinamento e altera Adam (Kingma e Ba, 2015)[28] de 0,999 para 0,98, adicionando a etapa de treinamento. RoBERTa usa uma mask dinâmica na tarefa de pré-treinamento, para que o material de treinamento inicie a operação de mask

aleatória antes de entrar no modelo, cancelando a tarefa *Next Sentence Prediction (NSP)* e adicionando o mecanismo *Full-Sentences*. Em termos de codificação de texto, *Byte-Pair Encoding(BPE)* é uma mistura de representações em nível de caractere e em nível de palavra, que suporta o processamento de palavras muito comuns em corpora de linguagem natural. Os pesquisadores usaram um vocabulário BPE de nível de byte maior do que o algoritmo BERT para treinar o BERT. Esse vocabulário contém 50 mil unidades de subpalavra sem nenhum pré-processamento adicional ou segmentação de palavra da entrada.

O RoBERTa obteve pontuações mais altas do que o BERT na avaliação geral de compreensão da linguagem GLUE, no conjunto de dados de respostas a perguntas SQuAD Stanford e na re-compreensão do exame RACE.

O método principal desta dissertação é o RoBERTa, que é o melhor algoritmo em tradução automática atualmente. E o processo detalhado será apresentado principalmente no Capítulo 4.

3.2 Corpus

Corpus é um material importante e básico para tradução automática. A tradução automática precisa ser baseada em corpus. Um excelente corpus pode melhorar a qualidade da tradução para tradução automática. Nesta seção, o corpus será descrito em detalhes. Na seção de 3.2.1, será apresentado o corpus e seu desenvolvimento. Na seção de 3.2.2 e 3.2.3, serão apresentados respectivamente o estado atual do corpus paralelo e do corpus paralelo bilíngue Chinês-Português.

3.2.1 Desenvolvimento de Corpus

Diferentes estudiosos têm opiniões diferentes sobre a definição de corpus, mas seja baseado em linguagem falada ou texto escrito, deve ser uma linguagem natural, não derivada de outros meios.

Já em 1991, David Crystal [29] deu a definição como:

Uma coleção de dados linguísticos, sejam textos escritos ou uma transcrição de fala gravada, que pode ser usada como ponto de partida da descrição linguística ou como meio de verificação de hipóteses sobre uma língua. [30] Essa definição não apenas aponta que o corpus é uma transcrição da linguagem escrita ou falada, mas também teve um papel orientador no estudo do corpus.

No mesmo ano, John Sinclair [31] definiu a palavra Corpus como:

Uma coleção de linguagem txt que ocorre naturalmente, escolhida para caracterizar um estado ou variedade de uma linguagem [31].

Explica o que o corpus deve conter e por que deve ser usado.

Em 2006, McEnery, Xiao & Tono et al.[32], generalizaram as características de corpus, e a definição de corpus se tornou mais precisa:

(1)É texto eletrônico legível por máquina;

(2)É linguagem real (falada ou escrita);

(3)É uma amostra de idioma obtida por meio de amostragem estrita (em vez de coletada aleatoriamente);

(4)É destina-se a representar um idioma ou variante de idioma.

Corpus não é apenas um trabalho de pesquisa complexo, mas também é a base de suporte da NLP de processamento de linguagem natural. O desenvolvimento de corpus afeta diretamente o auge da pesquisa de NLP.

3.2.2 Corpus Paralelo

Houve desacordo sobre a definição de corpus paralelo. Mona Baker (1995) [33] considerou que os textos incluídos no corpus paralelo são os textos na língua A e suas traduções na língua B. Stig Johansson (1998) [34] pensava que um corpus paralelo é um corpus que contém dois textos linguísticos com relação comparável. A visão de Mona Baker agora é geralmente aceita.

Entre todos os corpus paralelos, de acordo com as relações diferentes de quantitativas correspondentes, o corpus paralelo pode ser dividido em corpus paralelo bilíngue e corpus paralelo multilíngue. O corpus paralelo bilíngue é composto por textos de tradução um para um, e o corpus paralelo multilíngue é composto por um idioma, correspondente a várias cópias dos textos de tradução um para um.

De acordo com a direção correspondente do corpus paralelo, ele pode ser dividido em corpus paralelo unidirecional, corpus paralelo bidirecional e corpus paralelo multidirecional. Um corpus paralelo unidirecional é um idioma e sua tradução para outro idioma (por exemplo, Chinês e sua tradução para inglês). Um corpus paralelo bidirecional consiste em um idioma e sua tradução para outro, e contém o texto traduzido deste último (por exemplo, Chinês com sua tradução para o inglês e inglês com sua tradução para o Chinês). O corpus paralelo multidirecional é o texto fonte em um idioma e o texto traduzido em vários idiomas (por exemplo, Chinês para inglês e Chinês para francês).

O corpus paralelo é o dado básico da tradução automática NMT, e é uma condição necessária para a tradução e treinamento. Atualmente, a pesquisa sobre corpus paralelo no mundo ainda está em fase de desenvolvimento, o que é complementar à tradução automática.

3.2.3 Corpus Paralelo Bilíngue Chinês-Português

Atualmente, existem poucos corpora Chinês-Português no mundo, e um corpus paralelo Chinês-Português que está disponível publico e gratuitamente é ainda mais raro.

Liu, Siyou, Longyue Wang e Chao-Hong Liu (2018)[35] dedicaram-se à construção de um corpus paralelo sino-Português, mencionando que há pouquíssimos estudos sobre corpora sino-Português no mundo, e apenas alguns vários corpora paralelos estão disponíveis: OPUS(Tiedemann,2012)[36] ;OpenSubtitles2016(Lison and Tiedemann, 2016)[8];News-Commentary11(Tiedemann,2012);Tanzil (Tiedemann, 2012).

OpenSubtitles2016 (Lison e Tiedemann, 2016)[8] tem 6,7 milhões de pares de frases Chinês-Português no corpus paralelo acima mencionado. A maioria desse corpus paralelo vem de filmes, principalmente peças curtas, embora não sejam a primeira escolha para máquina básica. Mas é verdade que no corpus paralelo Chinês-Português atual , o corpus com mais pares de idiomas paralelos também é a melhor escolha no momento. Além disso, News-Commentary11 (Tiedemann, 2012) [36], em que os pares de corpus são, em sua maioria, de notícias. Este é um corpus de preferência de qualidade no corpus Chinês-Português, mas sua escala não é grande e suficiente, com apenas 10873 pares de corpus. O material de treinamento é insuficiente para suportar todo o processo de tradução automática, e este corpus pode ser o corpus de teste. E Tanzil (Tiedemann, 2012)[36] é uma coleção de textos religiosos do Alcorão, também é um corpus de preferência de qualidade, mas tem duas deficiências: uma é que a linguagem religiosa é muito obscura e difícil de entender, que não é o melhor para NMT, e a outra é que é muito pequeno em escala, com apenas 1,2 mil pares de corpus, o que não é suficiente para suportar todo o processo de treinamento de tradução automática.

No mesmo ano, a Universidade de Macau construiu um corpus paralelo Chinês-Português UM-PCorpus, que tem cerca de 6 milhões de sentenças, (Chao et al., 2018)[37], na pesquisa de Santos et al. (2019)[9] , usando 1 milhão de sentenças paralelas $PT \leftrightarrow ZH$ fornecidas pelo UM-PCorpus, juntamente com 5.000 sentenças adicionais para fins de teste. No artigo, UM-PCorpus é considerado um corpus paralelo adequado para pesquisa, mas infelizmente, as sentenças paralelas $PT \leftrightarrow ZH$ do UM-PCorpus não foram obtidas com sucesso.

Após comparação, decidimos usar OpenSubtitles2016 (Lison e Tiedemann, 2016) [8] como material de corpus para a pesquisa.

De acordo com as características do corpus paralelo do OpenSubtitles2016, não é adequado usar outros conjuntos de dados para teste. Para fazer a comparação experimental, vamos separar os últimos 10% do corpus no OpenSubtitles2016 como um conjunto de teste.

3.3 Tradução Automática Chinês-Português

Há pouca pesquisa sobre tradução automática Chinês-Português. Os pesquisadores da Universidade de Macau e da Universidade de Lisboa fizeram a maior parte do trabalho. Como é uma região com duas línguas oficiais, o Chinês e o Português, Macau tem contribuído muito para o desenvolvimento da comunicação sino-portuguesa.

Fai Wong e Sam Chao da Universidade de Macau(2010)[38] publicaram artigo sobre o tema. Neste artigo, as ferramentas de tradução automática (MT) que foram desenvolvidas são utilizadas para ajudar as pessoas de ambas as línguas. Essas ferramentas variam de dicionários bilíngues a modelos baseados em regras.

Nos anos subquentes, não haviam muitos estudos sobre tradução automática Chinês-Português até que Chao et al. (2018) criaram um corpus paralelo de 6 milhões de frases, o corpus é obtido atravé do site governamental de Macau para $PT \leftrightarrow ZH$. Em sua pesquisa, eles fizeram grandes contribuições para o desenvolvimento do corpus sino-Português, e disseram que publicariam uma parte do corpus. Na pesquisa posterior de Santos et al.[9], usaram com sucesso o corpus construído por eles e fizeram pesquisas. Mas, infelizmente, não consegui obter os dados do corpus.

Santos et al. (2019) [9] utilizaram o modelo Transformer para estudar a tradução Chinês-Português usando Different Model for Each Direction (Direct), Pivot Language(Pivot) e Single Model for All Pairs(Many-to-Many). Em Direct, foram criados modelos para zh-pt e pt-zh para testar o efeito.

Em Pivot, o autor faz 4 modelos: os dois primeiros traduzem zh-en e depois en-pt respectivamente. Os dois últimos traduzem pt-en e en-zh, respectivamente. Nesse método, zh-pt não traduz diretamente, mas transmite a tradução por meio do idioma do hub.

Em Many-to-Many, zh-en e en-pt são colocados juntos no mesmo modelo após marcação especial. E através do método central, zh-pt é traduzido indiretamente.

Após comparação, Santos acredita que o corpus criado por Chao et al.(2018) é o mais adequado para pesquisas em tradução automática. E obteve 1 milhão de materiais de treinamento do corpus. Infelizmente, não conseguimos obter o mesmo corpus paralelo.

Embora Santos e outros pesquisadores também tenham tentado usar o método da linguagem central para realizar muitas pesquisas, apontaram que, no caso de dados de pesquisa insuficientes, a tradução direta não é necessariamente a tradução ideal. Em termos de pesquisa sino-Português, a maioria dos excelentes dados são também obtidos através do método da língua central. Mas não importa como o idioma central é usado para pesquisa, é impossível ir além da limitação do idioma central para melhorar a qualidade da tradução de zh-pt.

Capítulo 4

Metodologia

O capítulo apresenta materiais e métodos de pesquisa. Este capítulo, primeiramente, apresenta o propósito, as ideias e os problemas que precisam ser resolvidos, materiais adquiridos e núcleo de questões de pesquisa respectivamente. Finalmente, faz o modelo necessário para resolver o problema.

4.1 Resumo do Problema

Melhorar a qualidade da tradução automática Chinês-Português é uma questão complexa. Nesta fase, a tradução Chinês-Português já começou, mas há pouco trabalho de pesquisa. Para melhorar a qualidade da tradução Chinês-Português realmente, deve-se realizar experimentos em combinação com características do idioma, dados básicos, excelentes algoritmos e avaliação. No que diz respeito à tradução automática Chinês-Português, existem duas grandes dificuldades no desenvolvimento: uma é a falta de dados básicos e de pesquisas relacionadas, e a outra é a combinação de modelos avançados de tradução automática e tradução Chinês-Português.

Para obter um excelente modelo de tradução automática Chinês-Português é necessário fazer o trabalho nos três aspectos: dados, estrutura e algoritmo. No campo da tradução automática, o BERT e outros algoritmos derivados baseados no BERT são atualmente os melhores . Para um modelo de tradução, a quantidade de dados e o número de tempo de treinamentos estão em proporção direta com a excelência dos resultados obtidos. Embora atualmente o corpus paralelo Chinês-Português seja muito escasso, é necessário escolher um que possa obter bons resultados em dados de pequena escala e se adaptar às mudanças nos dados de escala maior provocadas por pesquisas e desenvolvimentos no futuro.

O atual modelo de tradução seq2seq ainda é o framework mais utilizado. Em termos de algoritmo, opta pelo algoritmo RoBERTa, mais avançado, que não apenas usa MLM dinâmico para aumentar o efeito do treinamento, mas também pode se adaptar a treina-

mentos maiores e com mais etapas do que o BERT. Para usar o algoritmo e o modelo de pré-treinamento no modelo misto de palavras e usar o Unilm para tarefas de tradução downstream, a estrutura geral do processo é a seguinte:

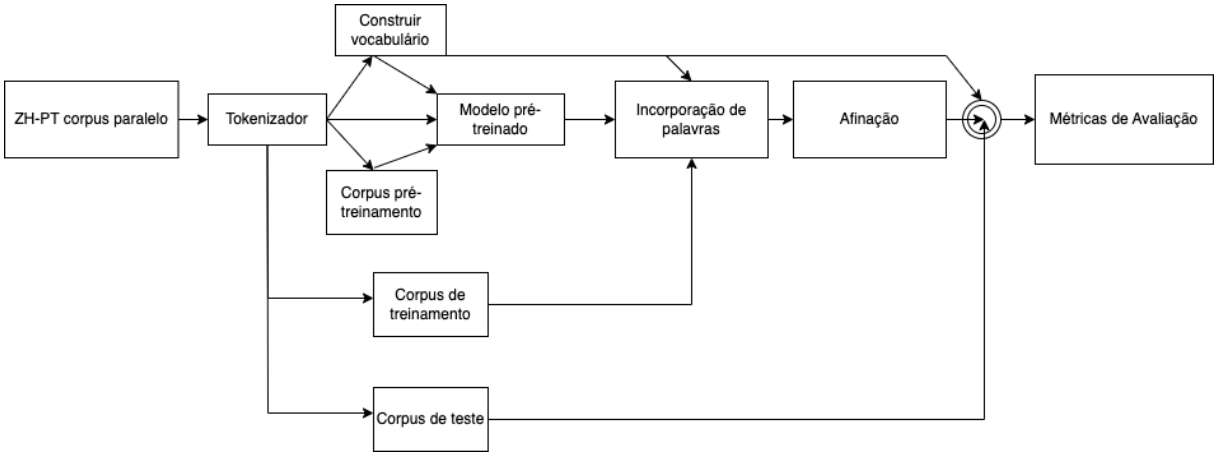

Figure 4.1: Arquitetura geral do modelo

Será feita uma série de experimentos para verificar e encontrar fatores que afetam o modelo e para se obter uma excelente pontuação de revisão. Considerando que a pesquisa sobre segmentação de palavras em tradução automática Chinês-Português deve focar na parte chinesa, utilizamos o BPE para a segmentação de palavras em Português; experimentos também serão feitos para encontrar o segmentador de palavras mais adequado para a parte chinesa.

4.2 Modelos Clássicos de Tradução Automática Neural

O modelo clássico de tradução automática neural é *end-to-end* ou modelo codificador-decodificador (*Encoder-Decoder Model*) [41]. O problema de tradução é o mesmo que resolver o problema de probabilidade. Ambos recebem o idioma de origem e encontram a probabilidade condicional alvo da linguagem.

O que se tem a fazer é pré-treinar a partir do corpus paralelo Chinês-Português, aprender esses parâmetros e maximizar a probabilidade condicional para obter o resultado ideal do modelo de tradução.

4.2.1 Os Modelos Pré-treinados (PTMs)

Nos últimos anos, com o avanço contínuo do aprendizado profundo, a fim de treinar totalmente os parâmetros do modelo profundo e evitar *overfitting*, geralmente é necessário

um treinamento de dados mais rotulados. No campo da NLP, os dados rotulados são um recurso caro. Os PTMs são pré-treinados a partir de uma grande quantidade de dados não rotulados, o que pode melhorar significativamente o desempenho de muitas tarefas de NLP enquanto reduz os custos de recursos. Em geral, os PTMs têm as seguintes vantagens:

• O pré-treinamento em grandes dados não rotulados pode obter uma representação de linguagem mais geral e beneficiar tarefas de downstream;

• Os PTMs fornecem um melhor parâmetro de inicialização para o modelo, têm melhor desempenho de generalização na tarefa de destino e aceleram a convergência;

• Os PTMs podem evitar o overfitting em pequenos conjuntos de dados até certo ponto;

• O modelo de pré-treinamento em tradução automática é treinar o modelo com o primeiro corpus e, em seguida, prosseguir com a tradução automática com base nesse modelo treinado preliminarmente. O treino e a utilização do modelo pré-treinado correspondem a duas fases: a fase de pré-treinamento e a fase de fune-tuning.

4.2.2 Algoritmos na Etapa de Pré-treinamento

No modelo de pré-treinamento, o algoritmo BERT é mais amplamente utilizado e usa MLM e NSP para desempenhar sua tarefa. E Robustly Optimized BERT Pretraining (RoBERTa) usa MLM dinâmico melhor. O BERT usa O MLM traditional, o que foi apresentado na seção de 3.2.6. No RoBERTa, o maior ajuste é mudar o MLM estático para um MLM dinâmico, e a operação de máscara aleatória começa antes que o material de treinamento entre no modelo. A operação em MASK é a seguinte:

MASK estática original: ao preparar dados de treinamento no BERT, cada amostra executará apenas uma MASK aleatória uma vez (portanto, cada época é repetida) e cada etapa de treinamento subsequente usa a mesma máscara,

Pré-processamento - adicionar operação MASK - salvar dados - modelo de entrada - executar cálculo.

MASK Dinâmica: No RoBERTa, a máscara não é executada durante o pré-processamento, mas sim gerada dinamicamente cada vez que a entrada é fornecida ao modelo, portanto, muda o tempo todo. Desta forma, no processo de entrada contínua de uma grande quantidade de dados, o modelo vai gradualmente se adaptando a diferentes estratégias de mascaramento e aprendendo diferentes representações de linguagem.

Pré-processar dados - salvar dados - executar operação MASK - inserir modelo - executar cálculo.

Comparado com o treinamento NSP do BERT, o RoBERTa usa frases completas sem NSP. No BERT, dois segmentos são unidos como um modelo de entrada de sequência e,

em seguida, a tarefa NSP é usada para prever se os dois segmentos têm uma relação contextual, mas a sequência e o comprimento total é inferior a 512. Descobriu-se por meio de experimentos no RoBERTa que FRASES COMPLETAS têm um efeito melhor e significa que as frases são continuamente extraídas de um artigo ou de vários artigos e preenchidas na sequência de entrada do modelo. Em outras palavras, uma sequência de entrada pode abranger vários limites de artigo. Especificamente, ele extrairá continuamente frases de um artigo para preencher a sequência de entrada, mas se chegar ao final desse artigo, continuará extraindo frases do próximo artigo para preencher a sequência, e o conteúdo em diferentes artigos ainda está em acordo com o delimitador SEP para dividir.

4.3 Modelo de Mistura de Letras e Palavras

O estudo da tradução automática requer dois idiomas diferentes. Em nossa pesquisa, no processamento de linguagem da tradução automática Chinês-Português, o Português é composto de letras, seja ele dividido por "espaço" ou processado por segmentação de palavras BPE, que obterá um valor muito bom em resultado de segmentação de palavras. A parte chinesa precisa ser focada e, atualmente, existem muitos dispositivos de segmentação de palavras que podem segmentar palavras chinesas, e suas características também são diferentes. Apresenta as características de cada dispositivo de segmentação de palavras e nossa escolha final na seção de 4.3.1.

4.3.1 Tokenizador

A segmentação de palavras é dividir uma frase em palavras-chave, palavras, conjuntos de palavras, etc. A segmentação de palavras é uma etapa básica do NLP e também é um processo importante para os computadores pré-processarem as sentenças. Tokenizadores diferentes têm métodos diferentes de segmentação para diferentes sentenças.

Existem métodos diferentes de segmentação de palavras para idiomas diferentes, e a escolha de um segmentador adequado vai ajudar a obter resultados melhores. Seleção de alguns tokenizers disponíveis para Chinês e Português atualmente:

a)Jieba, a atualização mais recente foi em 2020; atualmente é um segmentador de palavras chinesas convencional que suporta, relativamente, 4 modos: modo preciso, modo completo, modo de mecanismo de pesquisa e modo paddle, é um segmentador de palavras relativamente flexível e o código do segmentador de palavras é compatível. O desempenho dele é relativamente bom e suporta dicionários personalizados. Parece ser o separador de palavras mais usado para a língua chinesa.

b)O SnowNLP[39] é um tokenizador inspirado pelo TextBlob. Foi atualizado em 2017 e ofereceu segmentação de palavras chinesas, marcação de característica da palavras e

análise de sentimento. É um tokenizador excelente que pode marcar características das palavras, mas o processo de operação é relativamente lento.

c)O Stanford CoreNLP é um tokenizador excelente, atualizado em 2020, escrito por JAVA, suporta tagger part-of-speech (POS), named entity recognizer (NER), sentiment analysis e outras funções.

d)Lexical Analysis of Chinese (LAC), atualizada em 2021, é uma ferramenta que suporta funções como segmentação de palavras em Chinês, marcação de característica da palavra e reconhecimento de nomes próprios. O LAC foi desenvolvido pela equipe do Baidu. Em sua publicidade, propõe-se que ele tenha as características de boa efetividade, eficiência alta, personalização, chamada conveniente, suporte a terminal móvel e compatibilidade com python. É a segunda escolha para parte chinesa.

e)Byte Pair Encoding (BPE) é uma forma simples de compactação de dados. Primeiro, o texto é dividido nas unidades mais básicas, após segmentação e marcação adequadas. Em seguida, é contado o número de cada par de símbolos em todas as palavras e selecionado o par de símbolos com a frequência mais alta. A seguir, o par de símbolos com frequência maior é considerado como um símbolo novo, e a frequência de todos os pares de símbolos é recontada para todos os pares de símbolos na unidade constituinte mais básica, e a execução é repetida até que o número de símbolos combinados atenda ao requisito do número de vocabulário ou a uma condição de parada. BPE é um método de segmentação de palavras adequado para idiomas alfabéticos. O princípio é simples e fácil de entender, e a operação é altamente ajustável. O BPE também pode decodificar reversamente de acordo com o processo de codificação, o que também reduz bastante o custo computacional. BPE é usado como tokenizador para a parte portuguesa.

f)Space, como o Português é um idioma muito organizado, cada palavra será separada por espaços ou sinais de pontuação; portanto, em alguns casos, como na etapa de verificação de pontuação, o uso da segmentação de palavras espaciais mais simples para processar o Português serão considerados, para que não altere a frase.

4.3.2 Construção do Vocabulário

No processo de construção do vocabulário, utiliza a segmentação de palavras BPE para a parte de Português. A construção do vocabulário é o núcleo do BPE. Primeiro, prepare o corpus de treinamento do modelo e predefina um tamanho de vocabulário desejado. Em seguida, sepere todas as palavras do corpus de treinamento em sequências de caracteres e construa palavras iniciais a partir dessas sequências de caracteres divididas. Após a construção do vocabulário inicial, realize treinamento estatístico para calcular a frequência de cada par de bytes consecutivos no corpus; selecione o par de bytes com maior frequência para mesclar em uma nova subpalavra, atualizando o vocabulário e repetindo essa

operação até que o tamanho do vocabulário atinja o tamanho desejado ou a frequência dos pares de bytes restantes, que é no máximo 1.

Corpus de treinamento	Vocabulário
l o w </w> l o w e r <w> l o w e r <w> n e w e r <w>	('l','o','w','e','r','<w>','n')
lo w </w> lo w e r <w> lo w e r <w> n e w e r <w>	('w','e','r','<w>','n','lo')
low </w> low e r <w> low e r <w> n e w e r <w>	('w','e','r','<w>','n','low')
low </w> low er <w> low er <w> n e w er <w>	('w','e','<w>','n','low','er')

Table 4.1: O processo de construção de um vocabulário

Em Tabela 4.1, podemos ver o processo de construção do vocabulário. Na primeira etapa, obtivemos todas as letras e identificadores, incluindo 'l','o','w','e','r',' <w>','n'. Na segunda etapa, descobrimos que todos os 'l' e 'o' não apareciam sozinhos, mas sim combinados em 'lo', desta vez deletamos o 'l's e 'o's que não apareciam sozinhos, e adicionamos 'lo' como um novo membro do vocabulário. Na terceira etapa, descobrimos que todos os 'lo's e 'w's formam 'baixo', mas 'w' aparece sozinho, então mantemos 'w', excluímos 'lo' e 'baixo' adicionados ao vocabulário. Na quarta etapa, descobrimos que não há mais combinação de 'low', todos os 'r's são combinados com 'e' para formar 'er's, mas 'e' ainda existe sozinho, excluímos 'r' e mantemos 'e ', adicionamos 'er' ao vocabulário. Por fim, após não haver mais combinações, a construção do vocabulário é finalizada.

É usado o tokenizador jieba para construir o vocabulário Chinês, que é semelhante ao princípio do BPE. Comparado com o método de segmentação de palavras BPE, o jieba pode identificar o Chinês com mais precisão e realizar segmentação e rotulagem de palavras para Chinês.

4.3.3 Modelo Prioritário para Modelo de Mistura de Letras e Palavras

Na tradução automática, o uso de diferentes corpora produzirá frequências diferentes de letras e de palavras. Como cada palavra não é distribuída uniformemente, são produzidas algumas letras e palavras de frequência alta e baixa.

Na pesquisa, sob o mecanismo do Masked Language Model(MLM), as palavras com maior frequência são mais fáceis de serem calculadas pela tradução automática quando a segmentação da palavra é mais fragmentada. Ao contrário, letras e palavras com menor frequência precisam de mais cálculos no MLM com boas convergências. Portanto, no processo de uso real, utiliza o modelo letra-primeira e o modelo de palavra-primeira para tentar traduzir.

modelo letra-primeira:

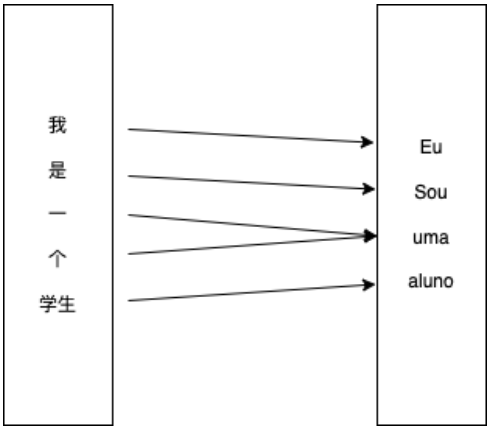

Figure 4.2: Modelo letra-primeira

Modelo palavra-primeira:

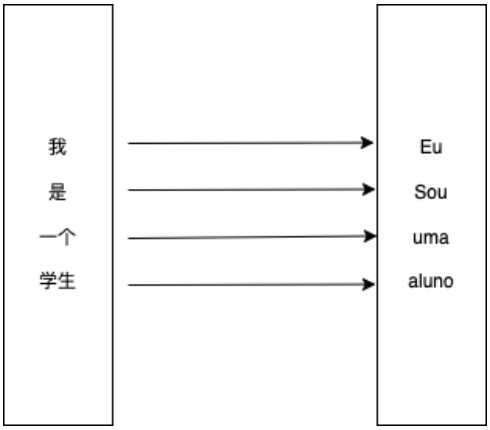

Figure 4.3: Modelo palavra-primeira

Conforme mostrado na Figura(4.2), no modelo letra-primeira, dois caracteres que podem formar uma palavra são separados, o que não corresponde bem a cada palavra em Português. Na Figura(4.3), o modelo palavra-primeira lida muito bem com esse problema, e tem uma boa correspondência entre Chinês e Português.

Em ambientes diferentes de linguagem, o modelo letra-primeira e o modelo palavra-primeira têm suas vantagens e desvantagens próprias.

4.3.4 *Unigram Language Model (ULM)*

O ULM é uma ferramenta para tentar resolver o problema de segmentação de palavras na tradução automática. O autor usa um método chamado marginalized likelihood para modelar o problema de tradução. Considerando o impacto de diferentes resultados de segmentação de palavras no resultado final da tradução, a probabilidade de segmentação de palavras é introduzida.

O ULM utiliza um modelo de linguagem para selecionar subpalavras e usa um método de redução, ou seja, um grande vocabulário é inicializado primeiramente e o vocabulário é continuamente descartado de acordo com os critérios de avaliação até que as condições limitantes sejam atendidas. O algoritmo ULM leva em consideração as diferentes possibilidades de segmentação de palavras da frase, de modo que pode gerar vários segmentos de subpalavras com probabilidades.

A frase S, $\vec{x} = (x_1, x_2, \ldots, x_m)$ é um resultado de segmentação de palavras, que consiste em m subpalavras. Portanto, o valor de verossimilhança da sentença S sob a segmentação de palavra atual pode ser expresso como:

$$P(\vec{x}) = \prod_{i=1}^{m} P\left(x_i\right) \tag{4.1}$$

Para a frase S, selecione aquela com o maior valor de verossimilhança como resultado da segmentação de palavras, que pode ser expressa como:

$$x^* = \text{argmax}_{x \in U(x)} P(\vec{x}) \tag{4.2}$$

Entre eles, $U(x)$ contém todos os resultados de segmentação de palavras da frase. Em aplicações práticas, o tamanho do vocabulário é de dezenas de milhares e não é prático listar diretamente todas as combinações de segmentação de palavras. Para resolver este problema, x^* pode ser obtido através do algoritmo de Viterbi.

Resolver $P(x_i)$ Em seguida, use o algoritmo EM, assumindo o vocabulário atual V, |D| é o número de dados no corpus,o objeto da maximização do passo M é a seguinte função de verossimilhança:

$$L = \sum_{s=1}^{|D|} \log\left(P\left(X^{(s)}\right)\right) = \sum_{s=1}^{|D|} \log\left(\sum_{x \in U\left(X^{(s)}\right)} P(x)\right) \tag{4.3}$$

(1)Inicialmente, construa um vocabulário suficientemente grande. Geralmente, todos os caracteres no corpus mais substrings comuns podem ser usados para inicializar o vocabulário e também podem ser inicializados pelo algoritmo BPE.

(2)Para o vocabulário atual, use o algoritmo Expectation Maximization Algorithm (EM) para resolver a probabilidade de cada subpalavra no corpus.

(3)Para cada subpalavra, calcule quanto a perda total diminui quando a subpalavra é removida do vocabulário e registre-a como loss da subpalavra.

(4)Classifique as subpalavras de acordo com o tamanho de loss, descarte uma certa proporção de subpalavras com a menor loss(como 20%) e gere um novo vocabulário para as subpalavras restantes. Deve-se notar aqui que caracteres únicos não podem ser descartados, isso é para evitar situações OOV.

(5)Repita as etapas (2) a (4) até que o tamanho do vocabulário seja reduzido ao intervalo definido.

Capítulo 5

Implementação

O objetivo desta dissertação de mestrado é realizar a tradução automática Chinês-Português com base no uso de dados e ferramentas livres, tanto quanto possível. Para atingir esse objetivo, apresentarei a principal tarefa neste capítulo:

Na seção 5.1, são apresentados os diferentes métodos que foram utilizados, são eles: (i)modelos de tradução automática que utilizam tokenizers diferentes para processar a parte chinesa; (ii)modelo Letra-primeira e do Modelo Palavra-primeira nos Resultados(iii)modelos que utilizam conjuntos de dados diferentes para implementar a tradução automática Chinês-Português; (iv)são utilizados os mesmos conjuntos de dados de diferentes escalas para implementar a máquina Chinês-Português.

Na seção 5.2, são descritos os diferentes conjuntos de dados e ferramentas fornecidos para cada método, bem como os dados usados como conjuntos de teste.

Na seção 5.3, apresenta a arquitetura geral de trazer dados para o RoBERTa para execução e o tempo gasto por cada método.

5.1 Arquitetura ao Treinamento

A principal dificuldade da pesquisa é utilizar recursos limitados para alcançar os melhores resultados. Portanto, projeta três grupos de experimentos e modelos que serão apresentados em detalhes a seguir.

5.1.1 O Impacto de Tokenizadores Diferentes no Modelo

A dificuldade de tradução entre Chinês e Português é a forma de entender o Chinês em grande parte. Como as palavras chinesas têm características de várias combinações, primeiro temos que experimentar os dados do corpus que foram utilizados, que tipo de impacto isso terá nos resultados da tarefa de tradução sob o processamento de diferentes

tokenizers. É um passo importante e um fator que vai diretamente afetar a qualidade da tradução.

No entanto, seu impacto no modelo não pode ser testado separadamente e, como têm fatores do custo e tempo de treinamento, escolhar dois tokenizers mais adequados para os dados que foram utilizados como ferramentas para experimentação.

Para realizar esta pesquisa, projeta o experimento (i) nas tarefas de tradução zh-pt de dois tokenizers diferentes e suas tarefas de tradução pt-zh, mantendo os mesmos dados, o modelo e a plataforma de treinamento, com um total de 2 modelos. O conteúdo experimental e a seleção serão 6.4.1 apresentados.

5.1.2 O Impato do Modelo Letra-primeira e do Modelo Palavra-primeira nos Resultados

No processo de tradução automática, o uso preferencial do vocabulário também terá um impacto na tradução. Atualmente, as formas mais comuns de usar o vocabulário são o modelo letra e o modelo de palavra-primeira. No modelo letra, quando encontra palavras de baixa frequência, dará prioridade a utilização de uma palavra única para prosseguir. No modelo de palavra-primeira, é reduzida a taxa de utilização de palavras de alta frequência e é aumentada a taxa de utilização de palavras de baixa frequência, de modelo a reduzir o impacto das palavras de alta frequência nos resultados da tradução.

Os modelos diferentes de prioridade de vocabulário são adequados para idiomas diferentes, algoritmos diferentes e modelos diferentes. Na tradução automática Chinês-Português, acredita-se que o uso do Chinês terá um impacto certo no resultado final do modelo, por isso decide por usar dois experimentos com diferentes modelos.

Para conduzir esta pesquisa, projeta o experimento (ii), utilizar os mesmos dados, modelo, plataforma de treinamento e tokenizer, e utilizar o tokenizer que teve um bom desempenho no experimento (i) para processar a parte chinesa. No experimento,com um total de 2 modelos. Como o treinamento do modelo de palavra-primeira já foi realizado em (i), pega os dados experimentais no experimento (i) e os novos dados experimentais no experimento (ii) para a comparação, que são os resultados da tradução do modelo de palavra-primeira zh-pt e pt-zh e os resultados da tradução do modelo letra-primeira, a fim de comparar e obter um excelente modelo para experimentos subsequentes.O conteúdo experimental e a seleção serão 6.4.2 apresentados.

5.1.3 O Impato do Banco Diferente dos Dados

Os recursos de corpus paralelos entre Chinês e Português são muito escassos, o que foi explicado na seção 3.2. Para experimentar o impacto de banco diferente dos dados na

tradução automática Chinês-Português, pois a diferença entre pt e pt(br) é muito pequena, dois bancos de dados diferentes são utilizados: zh-pt e zh-pt(br) em Opensubtitles2016[8]. Esses dois conjuntos de dados são de obras de cinema e televisão, e as expressões são mais parecidas com filmes e o comprimento das frases são relativamente os mesmos. A escolha desses dois conjuntos de dados minimiza o impacto no modelo devido a diferentes fontes de coleta de dados e diferentes expressões de linguagem utilizadas nos dados.

Para tanto, projeta o experimento (iii), a utilização do modelo com a melhor pontuação em experimento (i) como modelo experimental e realiza experimentos de teste em dois conjuntos de dados, a fim de entender o impacto de diferentes dados no modelo. 2 modelos no experimento(iii) , e 4 tarefas de tradução diferentes, zh-pt, pt-zh, zh-pt(br) e pt(br)-zh, foram concluídas e comparadas.O conteúdo experimental e a seleção serão 6.5.1 apresentados.

5.1.4 O Impacto de Tamanhos Diferentes no Modelo

Antes de desenharmos este estudo, buscou-se todos os corpora paralelos Chinês-Português abertos e livres. Como o tamanho dos dados obtidos era menor que o Opensubtitles2016 [8], considera experimentar todos os dados e reduzir uma parte dos dados a confirmar se os resultados do modelo têm algum efeito. Como o volume de treinamento é muito baixo, levará a resultados incertos; então, define o volume de dados da parte reduzida para 5 milhões de pares de frases, para o qual projeta o experimento (iV); no total têm 2 modelos. São eles zh-pt e pt-zh sob os dados completos e zh-pt e pt-zh sob os 5 milhões de pares de dados de sentença. No experimento (i), treinamos mais de 7 milhões de pares de dados completos, portanto, no experimento (iv), é necessário utilizar apenas os dados reduzidos para treinamento e 5 milhões de pares de dados de sentença zh-pt e pt-zh, e comparados com os resultados do modelo no experimento (i). O conteúdo experimental e a seleção serão 6.5.2 apresentados.

5.2 Dados e Ferramentas Usadas

5.2.1 Corpus Paralelo Chinês-Português

Como o corpus público de zh-pt é muito escasso, as características são analisadas em Seção 3.2.3 Antes do pré-treinamento, a informação básica mais necessária é o corpus paralelo Chinês-Português, que também são os dados básicos para tradução automática entre os dois idiomas. Deve-se escolher dados em quantidade suficientemente grande. Embora a escolha da qualidade dos dados não seja a melhor, o único corpus que atende às nossas necessidades é o Opensubtitles2016[8], onde existem dois corpora paralelos em Chinês e

Português, ambos de obras cinematográficas e televisivas, um é zh- pt que corresponde ao Chinês e Português, outro é o corpus paralelo de zh-pt (br) que é correspondente ao Português brasileiro e Chinês. Após um extenso trabalho de triagem e comparação, opta o corpus zh-pt e o corpus zh-pt(br) do Opensubtitles2016 [8] como dados de pesquisa.

Embora não sejam os dados mais perfeitos, depois de compará-los com outros corpora paralelos sino-portugueses que podemos obter, esta é a melhor escolha. Algumas correções no corpus serão feitas para que atenda às nossas necessidades.

Como as frases no Opensubtitles2016[8] vêm de obras de cinema e televisão, a desvantagem é que a semântica não é suficientemente popular e o alinhamento não é completo, mas a vantagem é que existem muitas frases traduzidas excelentemente. Os conteúdos específicos do nosso corpus selecionado são os seguintes:

Linguagem	alignment	all pairs	“alignment”words	“all pairs”words
zh-pt	6639359	7188706	56.05M	57.84M
zh-pt(br)	8583997	9295045	71.20M	73.55M

Table 5.1: Detalhes de Opensubtitles2016[8]

O “alinhamento” contém apenas unidades de tradução exclusiva, e “all pairs” incluem todas as unidades de alinhamento não vazias, incluindo duplicatas.

Todo o corpus contém muitas sentenças com boa qualidade de tradução e sentenças de compreensão difícil , o que pode ter algum impacto nos resultados de tradução do modelo. Por isso, cria um experimento para comparar o impacto de diferentes corpora nos resultados da tarefa de tradução.

Comparado com outros corpora que podem ser encontrados, o Opensubtitles2016[8] é o maior e o mais adequado para treinamento. Nos dados, há alguns problemas que precisam ser ajustados, como sinais de pontuação, placeholders, frases desalinhadas, etc. Para o treinamento, algumas correções são necessárias, algumas podem ser mantidos e detalhadas no processamento dos dados em experimentos subsequentes.

Para encontrar um conjunto de teste adequado, é avaliado manualmente o conteúdo do conjunto de dados, considerando que suas principais fontes são obras de cinema e televisão, e a maioria delas são frases curtas, que não são adequadas para notícias (News-Commentary11[36]), religião (Tanzil [36])) etc., como um conjunto de teste. Embora nosso modelo não se limite à tradução de obras de cinema e televisão, de acordo com os materiais de treinamento existentes, conjuntos de testes adequados para pontuação ainda devem ser encontrados em obras de cinema e televisão.

Após uma consideração abrangente, nós tentamos usar News-Commentary11 zh-pt(Ao escolher um corpus para este artigo, inclua o corpus reverso) e copiado os últimos 10%

e 5% do Opensubtitles2016 zh-pt e zh-pt(br)como um conjunto de teste para avaliar o modelo, de modelo que tenha um número apropriado para avaliação e corresponda ao tipo do nosso conjunto de treinamento.

5.2.2 Processamento de Dados

Em termos de qualidade das frases, o corpus paralelo Chinês-Português do Opensubtitles2016[8] não é excelente. Em sua maioria são frases curtas, e a maioria de suas palavras e estruturas são provenientes de filmes; muitas frases que não foram processadas e caracteres ilegíveis e palavras em inglês, e as frases não estão completamente alinhadas. O conjunto de dados possui conteúdo que pode afetar a tarefa de tradução, é processado uma parte dele e opta por manter a outra parte do alinhamento. Os problemas e soluções existentes são os seguintes:

1)Existem alguns sinais de pontuação que não correspondem, o que pode afetar a tradução.

zh	pt
喂	Estou?
是我	Sou eu.
起来	Levanta-te.

Figure 5.1: Exemplos de diferenças na pontuação e tradução nos dados

Na Figura 5.1 pode-se observar alguns problemas como pontuação; quase não há ponto "." em Chinês, enquanto há muitos pontos "." em Português. As frases são muito curtas, de modelo que é difícil obter a maioria do significado no contexto, que tem ambiguidade nas mesmas palavras.

2)Existem algumas frases no filme que não estão alinhadas e podem ter um impacto pequeno na tradução:

zh	pt
但我真的放不下	Mas importa
太好了	Sim
加了橄榄汁的马天尼	Um dirty martini?
他本来要将七大王国付之一炬...	Teria queimado os Sete Reinos...

Figure 5.2: Exemplos de viés de tradução devido ao desalinhamento de contexto

Na Figura 5.2 há algum conteúdo nos dados que é principalmente irrelevante para o texto original, mas do contexto, e as frases aqui não têm um impacto decisivo na tradução automática, por isso elas são mantidas.

3)Existem alguns idiomas que não são totalmente processados

	PT	ZH
Mais conteúdo	É um Van Gogh, o que tu pensas?	那是梵高 或者你认为它是什么？ 是啊。
Conteúdo não totalmente traduzido	Ohhh sim...	0H，是的。
Conteúdo não adequado para uso	gravação de Patterson e Gimlin..	Gimlin？

Figure 5.3: Frases a serem melhoradas na qualidade do corpus

Na Figura 5.3 não é difícil ver que a maioria dos pares de frases é produzida de acordo com o contexto, e têm algumas exclusões, ambiguidades e traduções modificadas, o que não tem um grande impacto no modelo de tradução, e a seleção do conjunto de teste para o modelo de tradução é relativamente importante. Para nossa linguagem cotidiana, esses materiais de treinamento não são um material excelente, mas para trabalhos de cinema e televisão, este material de treinamento é muito bom.

Pontuação Removida	,	。	!	……
Pontuação Deixados	?	“ ”	-	

Figure 5.4: Tratamento de sinais de pontuação

Para resolver o problema de pontuação mostrado na Figure 5.4, remova a seguinte pontuação e deixe algumas para distinguir o significado das palavras.

Use espaços para preencher onde a pontuação foi removida. Isso garante que a relação de alinhamento entre os pares de frases antes e depois da modificação não seja afetada e também é conveniente para nosso trabalho subsequente de segmentação de palavras.

Além disso, algumas frases têm espaços no final da frase, o que pode ter impacto em nosso modelo de tradução, porque existem algumas frases com espaço duplo nos dados e o problema é o espaço no final da frase; remova todos os 3 ou mais espaços consecutivos na frase. Embora ainda haja alguns espaços reservados de 2 ou menos espaços consecutivos no final da frase, a maioria das partes que afetam a tradução foram removidas.

Para os dados da Figure (5.3) que não são adequados para tradução, não há uma boa solução, pois não é comum e, ao mesmo tempo, difícil de encontrar; julga que seu impacto na tarefa de tradução não é grande, por isso não trata isso.

No experimento comparativo (iv), para evitar o desvio causado pelo conjunto não-treinador não incluindo o conjunto de teste, no modelo E (que será descrito no Capítulo 6), parte do experimento, o conjunto de treinamento é de 5 milhões de dados em ordem inversa.

5.2.3 Tokenizer e Modelo de Algoritmo.

Já apresentei na seção anterior a situação relevante do tokenizer. Para escolher um tokenizer que seja melhor para os resultados de tradução, é necessário rodar o modelo na íntegra para compará-lo; devido ao custo econômico e de tempo de treinamento, considerando a praticidade, decida escolher dois tokenizers adequados para tradução automática Chinês-Português à comparação experimental.

Considerando que utiliza principalmente a linguagem python, o tokenizer é utilizado apenas para lidar com Chinês. Como a publicidade é gratuita e o tempo de atualização é relativamente novo, escolha o tokenizer jieba e o tokenizer lac para segmentação de palavras. A situação experimental relevante será realizada no experimento (i). Nos experimentos de acompanhamento, utilize o tokenizer que apresenta melhor desempenho na tradução automática Chinês-Português.

Use dois tokenizers para pré-processar o Chinês e construir um vocabulário para utilizar um modelo de mistura, reduzir a ocorrência de palavras de frequência alta e afetar a tarefa de tradução; a tarefa de tradução utiliza preferencialmente as palavras do dicionário. As palavras são usadas para realizar operações.

5.3 Configuração de Execução do Modelo

Como o NMT precisa usar Graphics Processing Unit (GPU) para acelerar os cálculos e, como algumas GPUs avançadas são muito caras, selecione alguns equipamentos experimentais na medida de custo mais baixo possível para oferecer o suporte suficiente de computação.

Tenho um computador IOS com um chip M1. Embora o desempenho do chip M1 seja bom, não é a melhor escolha devido à configuração de ambiente complicada e ao fato de que a maioria dos modelos de computação baseados em RoBERTa não podem se adaptar perfeitamente ao chip M1. Portanto, procurei algumas plataformas de aluguel de nuvem de GPU para comparação e selecionei a plataforma mais adequada para nosso experimento. Algumas das melhores plataformas são as seguintes:

A plataforma Jiutian Bisheng é construída pela equipe Jiutian do Instituto de Pesquisa da China Mobile Communications Co., Ltd.. Ela fornece poder de computação GPU abundante, dados ricos e recursos de aprendizado prático para atender a todos os cenários do processo, como aprendizado de curso, classificações de competição e trabalho Hunting. Também é fornecido para faculdades e universidade para oferecer uma solução completa para ensino online, pesquisa científica e desenvolvimento. Jiutian Bisheng é uma plataforma de nuvem de GPU gratuita muito boa, e seu principal método de cálculo é convidar novos usuários para obter "feijões de computação" gratuitos. O efeito geral é muito bom. Desde o momento em que escolhemos a plataforma de poder de computação, não havia outra maneira de obter horas de GPU além de convidar novos usuários.

GCP (Google Cloud Platform) é um conjunto de serviços públicos de computação em nuvem fornecidos pelo Google. A plataforma inclui uma série de serviços gerenciados para computação, armazenamento e desenvolvimento de aplicativos em execução no hardware do Google. Desenvolvedores de software, administradores de nuvem e outros profissionais de TI corporativos podem acessar os serviços do Google Cloud Platform pela Internet pública ou conexões de rede privada. Comparado com Azure e AWS (Amazon), o GCP tem métodos de cobrança diferentes. Em alguns estudos, o custo de usar o GCP de pesquisa é relativamente barato e há mais garantias para a segurança dos dados.

Amazon Web Services (AWS) é a plataforma de nuvem mais abrangente e amplamente adotada, oferecendo mais de 200 serviços completos de data centers em todo o mundo. Milhões de clientes, incluindo as startups de crescimento mais rápido, as maiores empresas e as principais agências governamentais, usam a AWS para reduzir custos, aumentar a agilidade e acelerar a inovação [40]. AWS é uma plataforma segura e conveniente, muitos modelos de configuração podem economizar muitos problemas de configuração.

O AutoDL é uma plataforma de aluguel de GPU relativamente econômica. O preço é menor do que AWS e GCP. Devido ao número atual de usuários, a GPU também está em um estado relativamente suficiente e haverá um ambiente bem configurado , o que economiza muito tempo no início do programa e em problemas de configuração. A plataforma limpará os dados do usuário a cada 30 dias quando for desativada, o que não tem impacto substancial nos pesquisadores alugados a longo prazo.

Existem também algumas plataformas muito boas, como: Microsoft Azure, nuvem

Ali, nuvem Tencent, etc. Devido a fatores como preço, configuração e links de rede, não escolhi entre essas plataformas.

Jiutian Bisheng interromperá a operação do programa devido à sua manutenção regular, e o processo de obtenção de poder de computação é um pouco complicado. Então, no final, não escolhemos esta plataforma. Quanto ao GCP, devido ao grande número de pesquisadores, é difícil obter GPUs e outras configurações econômicas, e não é nossa melhor escolha. Comparado com outras plataformas, o AWS é mais caro. Esperamos realizar o maior número possível de pesquisas experimentais com o mesmo custo. Portanto, após triagem e julgamento, a plataforma AutoDL foi finalmente selecionada.

Capítulo 6

Experimento

6.1 Planejamento do Estudo de Caso

Por fim, são selecionados 2 conjuntos de dados: Opensubtitles2016 e News-Commentary11. Em Opensubtitles2016 usamos zh-pt e zh-pt(br), em News-Commentary11 usamos zh-pt. Em cada modelo, usaremos 3 conjuntos de dados para realizar experimentos:

Primeiro utilizamos todos os dados dos dois conjuntos de dados como conjunto de treinamento. Ao mesmo tempo, os últimos 10% dos pares para testar se o modelo pode funcionar corretamente.Usaremos o número 1 após os resultados do modelo para marcar.

Segundo utilizados os primeiros 95% os dados dos dois conjuntos de dados como conjunto de treinamento. Os últimos 5% dos pares de frases são usados como conjunto de teste. Usaremos o número 2 após os resultados do modelo para marcar.

Terceiro utilizamos todos os dados dos dois conjuntos de dados como conjunto de treinamento. E utilizamos News-Commentary11, zh-pt como corpus de teste. Usaremos o número 3 após os resultados do modelo para marcar.

Existem 7188706 pares de frases em Opensubtitles2016 zh-pt e 9295045 pares de frases em Opensubtitles2016 zh-pt(br), e existem 10873 pares de frases em News-Commentary11 zh-pt.

6.2 Métricas de Avaliação

Atualmente, o melhor índice de avaliação ainda usa a avaliação manual, o que é muito caro, demorado e trabalhoso. Pesquisadores inventaram alguns indicadores de avaliação para substituir ao máximo a avaliação manual. Embora não exista um sistema de índice de avaliação perfeito, essas têm sido algumas das maneiras de ajudar as pessoas a quantificar a precisão da tradução automática.

Entre os indicadores de avaliação da tradução automática, os indicadores de avaliação mais usados são Bilingual Evaluation Understudy (Bleu) [6] e Recall-oriented Understanding for gisting assessment (Rouge) [41]. O BLEU mede a qualidade da tradução de acordo com a taxa de precisão. O Rouge, por outro lado, mede a qualidade das traduções com base na recordação.

6.2.1 BLEU

O BLEU foi proposto pela IBM em 2002 para a avaliação de tarefas de tradução automática. Durante o processo de avaliação, a sentença gerada pela rede neural é candidata, a tradução padrão dada é referência e n-grama refere-se ao número de palavras consecutivas n. E o método de cálculo do BLEU é o seguinte:

$$BLEUn = \frac{\sum_{c \in \text{candidates}} \sum_{n-gram \in c} \text{Count}_{\text{clip}} \left(n - \text{gram }\right)}{\sum_{c' \in \text{candidates}} \sum_{n-\text{gram }' \in c'} \text{Count} \left(n - \text{gram }'\right)} \tag{6.1}$$

No equação 6.1, primeiro, o BLEU calcula as contagens de n-gramas cortadas para todas as frases candidatas e divide pelo número de n-gramas candidatos no corpus de teste para calcular uma pontuação de precisão modificada, p_n, para todo o corpus de teste.

Em seguida, o BLEU calcula um fator de Penalidade de Brevidade (BP) para evitar pontuações excessivas quando o comprimento da sentença traduzida é menor que o da sentença traduzida.

$$BP = \begin{cases} 1 & \text{if} \quad l_c > ls \\ e^{1-\frac{l_s}{l_c}} & \text{if} \quad l_c <= l_s \end{cases} \tag{6.2}$$

A fórmula de cálculo de BP é equação (6.2). l_c representa o comprimento da tradução da tradução automática, l_s representa o comprimento efetivo da resposta de referência e, quando houver várias traduções de referência, selecione o comprimento mais próximo da tradução. Quando o comprimento da tradução for maior que o comprimento da tradução de referência, o fator de penalidade é 1, o que significa que não há penalidade, e o fator de penalidade será calculado somente se o comprimento da tradução automática for menor que a resposta de referência.

Finalmente, a fórmula de cálculo do BLEU é a seguinte.

$$BLEU = BP \times \exp\left(\sum_{n=1}^{N} \boldsymbol{W}_n \log P_n\right) \tag{6.3}$$

O resultado da avaliação BLEU é um número entre 0 e 1. Para facilitar a comparação, o valor BLEU na parte posterior desta dissertação amplia esse número em 100 vezes e o exibe como um valor entre 0 e 100. Os indicadores comuns incluem BLEU-1, BLEU-2, BLEU-3 e BLEU-4.Uniform weights $W_n = 1/\mathrm{N}$. Entre eles, BLEU-4 é o indicador de avaliação mais usado. Todos os indicadores BLEU usam BLEU -4.

6.2.2 Rouge*(Recall-Oriented Understudy for Gisting Evaluation)*

O surgimento do Rouge é para resolver o problema do lançamento perdido (baixa taxa de recuperação) do NMT. Portanto, Rouge é adequado apenas para avaliar NMT, e não para SMT. O Rouge não se importa se de que tradução candidata seja fluente ou não. Seu método de cálculo é semelhante ao BLEU, com a exceção de que o BLEU é baseado em precisão, enquanto Rouge é baseado em recall.

$$Rouge-N = \frac{\sum_{S \in ReferenceSummaries} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in ReferenceSummaries} \sum_{gram_n \in S} Count(gram_n)}$$

Figure 6.1: Fórmula de cálculo Rouge [6]

A maior vantagem do método de avaliação automática Rouge é que não depende de ferramentas de processamento de linguagem; a desvantagem é que é rígido, pouco flexível e não considera a correspondência no nível semântico. Normalmente, os pesquisadores usam as pontuações de Rouge-1, Rouge-2 e Rouge-L para se referir à qualidade da tradução do modelo, onde Rouge-1 representa a precisão e a recuperação da sobreposição de combinação de uma única palavra entre o resumo do sistema e o resumo de referência. Rouge-2 representa a precisão e a recuperação da sobreposição de duas palavras entre o resumo do sistema e o resumo de referência, e Rouge-L é a auto-sequência comum mais longa. Como os nossos dados são principalmente frases curtas, não usamos Rouge-L.

Nesta dissertação é utilizado o Rouge-1 e o Rouge-2 com BLEU para avaliar nosso modelo.

6.3 Baseline

A fim de encontrar uma linha de base que seja conveniente para comparar, procure NMT zh-pt existentes que possam ser comparados, e dois deles podem ser usados como nossa

linha de base. Um é um modelo de tradução de Santos, Rodrigo Soares dos [17], e o outro é o Google Translate.

O modelo de Santos é baseado no Transformer, e eles fornecem uma página da web para tradução em pt-zh.

https://portulanclarin.net/workbench/lx-translator/

Considerando que seu método principal é o Transformer, e o BERT [42] é aprimorado pelo Transformer, e o RoBERTa [27] é ajustado do BERT, embora possa ser usado como nossa linha de base, devido a atualizações do algoritmo, uma comparação não é necessária. Após tentar, infelizmente, não consegue obter os dados de treinamento de seu modelo; embora seu site de tradução possa realizar a tradução Chinês-Português, possui um mecanismo de verificação, o que não é adequado para um grande número de trabalhos de amostragem de tradução.

Portanto, usa o Google Tradutor para comparação de linha de base.

O Google é uma empresa global super grande com uma ampla gama de negócios e é uma empresa líder no setor da Internet. O Google Translate é o site de tradução líder mundial, abrangendo não apenas uma grande variedade de idiomas, mas tambe dispondo uma precisão muito alta. Excelentes algoritmos, incluindo Transformer e BERT, também são da equipe do Google, e muitos pesquisadores pretendem estar próximos do nível de tradução do Google.

O Google tem um grande número de especialistas e técnicos, bem como um grande número de amostras de dados, execelentes algoritmos e poder de computação. É muito desafiador usar o Google Tradutor como baseline.

O Google Translate só permite traduções com menos de 5.000 caracteres por vez, porque há um problema de que "espaços" não podem ser totalmente refletidos, especialmente em Chinês. A situação é a seguinte.

	Exemplo	Hipótese	BLEU
Com espaço	Sou aluno da unb	Sou aluno da unb	100
Sem espaço	Soualunodaunb	Soualunodaunb	100
Com espaço	Sou aluno da unb	Sou aluno da unc	75
Sem espaço	Soualunodaunb	Soualunodaunc	0

Table 6.1: Impacto do espaço no BLEU-1

A tabela 6.1 mostra o impacto das sentenças em diferentes situações na pontuação BLEU. Quando as sentenças são exatamente iguais, a pontuação BLEU obtida é uma pontuação completa. Quando a estrutura de segmentação de palavras das sentenças é a mesma, há palavras como "unb " e "unc", em que se pode ver que a pontuação do BLEU vai variar muito. Quando a frase não possui uma estrutura de segmentação de palavras, o

BLEU tratará todo o conteúdo como uma palavra para processamento. Neste momento, pela diferença sutil entre "unb" e "unc", a dedução será ampliada.

Outro problema que precisa ser resolvido é a segmentação de palavras em Chinês.

O original	我是巴西利亚大学的学生
Jieba	我 是 巴西利亚大学 的 学生
LAC	我 是 巴西利亚 大学 的 学生
Espaço	我 是 巴 西 利 亚 大 学 的 学 生

Figure 6.2: Segmentação de palavras diferentes do Chinês por tokenizers diferentes

Como segmentações diferentes de palavras têm efeitos diferentes na pontuação de sentenças, opte por usar o tokenizer Jieba para segmentação de palavras na parte chinesa da pontuação BLEU.

Portanto, para garantir a integridade das frases, rastreie o conjunto de teste frase por frase e traduza no Google Translate. Por fim, os resultados são contados.

Após a estatística, realize a segmentação de palavras e pontue de acordo com os exemplos de frases antes da tradução e o texto obtido após a tradução. O resultado é o seguinte.

Conjunto de dados zh-pt:

	zh-pt(10%)	pt-zh(10%)	zh-pt(5%)	pt-zh(5%)	zh-pt(NC)	pt-zh(NC)
Rouge-1	32,57	36,56	33,01	37,12	40,12	41,31
Rouge-2	13,21	14,56	14,02	14,95	19,21	20,03
BLEU	5,15	4,51	5,35	4,72	11,15	13,50

Table 6.2: Pontuação BLEU do conjunto de dados zh-pt após a tradução do Google Translate

Conjunto de dados zh-pt (br):

Escolhe um para observar que as traduções do Google Translate são de alta qualidade.

Chinês: 你/不是/把/卡刷/爆/了/吧

Português: Não estás a atingir o limite dos teus cartões, estás?

pt-zh: 你/没有/达到/你/的/卡/限额/，/是/吗?

zh-pt: Você não explodiu seu cartão?

Figure 6.3: Selecione aleatoriamente um resultado de tradução

	zh-pt(br)10%	pt(br)-zh 10%	zh-pt(br)5%	pt-zh(br)5%
Rouge-1	31,44	34,12	32,21	34,45
Rouge-2	14,11	15,26	15,03	15,62
BLEU	5,02	4,73	5,41	4,96

Table 6.3: Pontuação BLEU do conjunto de dados zh-pt(br) após a tradução do Google Translate

6.4 Resultados dos Modelos Diferentes

6.4.1 Modelos de Tokenizers Diferentes

Se você deseja fazer uma tradução entre Chinês e Português, além de usar um excelente modelo de tradução, principalmente no processamento da parte chinesa, para encontrar um método de processamento mais adequado, é necessário projetar dois modelos com tokenizers diferentes: o Modelo A e o Modelo B. Durante o processo de treinamento utilize o mesmo conjunto de treinameno, o Opensubtitles2016 (zh-pt). Na parte do Português, ambos os conjuntos de modelos utilizam segmentação de palavras BPE. Para o processamento da parte chinesa, o Modelo A usa o tokenizer LAC e o Modelo B usa o tokenizer Jieba. Coletam-se as primeiras 25.000 unidades mínimas processadas pelos dois modelos, como listas de palavras chinesas, e elas são mescladas com a lista de palavras em Português processada pela segmentação de palavras BPE após duplicação. O Modelo A1 obteve uma lista de palavras com 40.901 unidades e o Modelo B1 obteve uma lista de palavra com 41.885 unidades.O Modelo A2 obteve uma lista de palavras com 40.887 unidades e o Modelo B2 obteve uma lista de palavra com 41.843 unidades.

No vocabulário, existem alguns fatores instáveis, pois são provenientes do próprio corpus. No entanto, se for processado, afeta artificialmente na avaliação do Modelo E impacta no treinamento, faz com que OOV adicione o efeito corpus de treinamento (palavra desconhecida), então mantamos esses fatores desestabilizadores.

Símbolo	七	ز	リ	ح	۵
Número	3000	41	1cH00FF00	700	99

Figure 6.4: Alguns exemplos de fatores instáveis

Os fatores instáveis dos símbolos vêm principalmente dos pares de sentenças não processados no corpus, e os elementos de números vêm principalmente das instâncias de sentenças. Como a frequência de ocorrência dos números é diferente, não existe um sistema completo de todos os números, mas existem os números sozinhos que aparecem nos pares de sentenças.

	Rouge-1	Rouge-2	BLEU
Modelo A1 zh-pt	22,25	14,47	11,87
Modelo B1 zh-pt	25,21	21,48	13,71
Modelo A2 zh-pt	22,01	14,13	11,05
Modelo B2 zh-pt	24,95	21,13	12,95
Modelo A3 zh-pt	20,95	12,21	7,54
Modelo B3 zh-pt	21,01	13,15	7,76
Modelo A1 pt-zh	22,64	15,31	12,21
Modelo B1 pt-zh	24,53	19,98	12,95
Modelo A2 pt-zh	22,07	14,85	12,06
Modelo B2 pt-zh	24,01	19,45	12,18
Modelo A3 pt-zh	21,16	12,65	7,85
Modelo B3 pt-zh	21,67	20,21	7,96

Table 6.4: Resultados dos Modelo A e Modelo B

Ao executar os cálculos, existem algumas plataformas gratuitas disponíveis em razão da necessidade de usar GPUs estáveis. Porém, por conta de interrupções eventuais no sistema de gerenciamento, é utilizada uma plataforma AutoDL estável e é alugada um GPU de V40-48GB e AMD EPYC 7543 32-Core Processor na plataforma para o nosso treinamento. Vamos substituir o vocabulário treinado no modelo de pré-treinamento RoBERTa para pré-treinamento, e observar manualmente a função de perda. Após a 11ª época do treinamento, a função de perda basicamente não muda, interrompe a operação e salva o vetor de peso para realizar tarefas downstream. Para tornar os níveis de treinamento dos dois modelos semelhantes, definimos as rodadas de treinamento do Modelo A e do Modelo

B para 7. O Modelo A1 usou total de 610 horas e o Modelo B1 usou 616 horas.O Modelo A2 usou total de 601 horas e o Modelo B1 usou 609 horas.

A distribuição de peso e o vocabulário treinados são trazidos para o nosso modelo de tradução, traduzidos através do conjunto de teste proposto Seção 6.1. Os resultados obtidos são avaliados diretamente para BLEU e Rouge. Os resultados obtidos da avaliação na Tabela 6.4.

6.4.2 Modelos Diferentes de Letra-primeira e Palavra-primeira

Para comparar modelos de letra-primeira e palavra-primeira. Projeta o experimento (ii), seleciona o Modelo B com melhores resultados no experimento (i) e o compara com o Modelo F. No Modelo F, usa o tokenizador mesmo, conjunto de dados mesmo, modelo de algoritmo mesmo do Modelo B. No modelo de prioridade de vocabulário, use o modelo de letra-primeira.

Para isso, realiza o treinamento. Vamos fazer o Modelo F1 e F2 como no Modelo B1 e B2, e o treinamento foi interrompido após 7 épocas e avaliado. O Modelo F1 após 615 horas de treinamento e o Modelo F2 após 607 horas de treinamento, foram obtidos os seguintes resultados(Na tabela 6.5):

	Rouge-1	Rouge-2	BLEU
Modelo F1 zh-pt	25,01	21,07	13,61
Modelo B1 zh-pt	25,21	21,48	13,71
Modelo F2 zh-pt	24,65	20,83	12,80
Modelo B2 zh-pt	24,95	21,13	12,95
Modelo F3 zh-pt	19,95	12,65	7,20
Modelo B3 zh-pt	21,01	13,15	7,76
Modelo F1 pt-zh	23,95	19,30	12,65
Modelo B1 pt-zh	24,53	19,98	12,95
Modelo F2 pt-zh	23,02	18,96	12,01
Modelo B2 pt-zh	24,01	19,45	12,18
Modelo F3 pt-zh	20,87	19,44	7,65
Modelo B3 pt-zh	21,67	20,21	7,96

Table 6.5: Resultados dos Modelo F e Modelo B

6.5 Resultados dos Dados Diferentes

6.5.1 Resultados dos Tipo de Idioma Diferentes

Após treinamento com modelos diferentes, descobrimos que os resultados de tradução do modelo usando segmentação de palavras jiaba foram ligeiramente melhores do que aqueles usando segmentação de palavras LAC. Ao testar diferentes Dataset, projete o Modelo C e Modelo D para manter a consistência, ambos usam o tokenizer Jieba para construir o vocabulário e a segmentação de palavras.

Use dois conjuntos de dados diferentes. Como pt e pt(br) são basicamente os mesmos em termos de expressões textuais, usae ambos para teste de dados. Os conjuntos de dados são do Opensubtitles2016. Usando o conjunto de dados zh-pt respectivamente, há 7.188.706 pares de sentenças e o conjunto de dados zh-pt(br) possui 9.295.045 pares de sentenças. Fiz uma comparação simples entre os dois conjuntos de dados, e o comprimento da frase, a qualidade da frase e o grau de processamento são basicamente os mesmos. Ao mesmo tempo, tente comparar o impacto de tamanhos diferentes de conjuntos de treinamento nos resultados do modelo. Portanto, realize 2 grupos em um total de 4 vezes de treinamento para obter resultados intuitivos. Com a finalidade de construir um conjunto de teste de tamanho adequado, extraímos os últimos 400.000 pares de sentenças dos dois corpora paralelos e os combinamos em ordem aleatória como um conjunto de teste, com um total de 800.000 pares de sentenças.

Dois conjuntos de experimentos foram realizados separadamente. A plataforma de treinamento é igual com o Modelo A. Para o primeiro conjunto de Modelo C e Modelo D, usa 6,5 milhões de pares de sentenças para treinamento, e o Modelo C usou os primeiros 6,5 milhões de pares de sentenças no zh-pt; o Modelo D usou os primeiros 6,5 milhões de pares de sentenças no zh-pt(br). Devido aos conjuntos diferentes de treinamento, o tamanho do vocabulário do Modelo C é 41.820, e o tamanho do vocabulário do Modelo D é 42.210. Na fase de pré-treinamento do Modelo C, a função de perda do Modelo C é basicamente inalterada após 10 épocas, então levam 10 épocas no Modelo C e no Modelo D. Depois que o Modelo C1 for executado por 554 horas e o Modelo D1 por 556 horas, o pré-treinamento é interrompido e é salva a distribuição de peso. Depois disso, o vocabulário e a distribuição de pesos são trazidos para o modelo de tradução e os indicadores são avaliados diretamente. Não é necessário fazer o Modelo C e o Modelo D a realizar o segundo conjunto de experimentos em seção 6.1, e o Modelo D não pode conduzir experimentos D3. Os resultados experimentais estão em tabela 6.6.

	Rouge-1	Rouge-2	BLEU
Modelo C1 zh-pt	24,46	18,95	12,65
Modelo D1 zh-pt(br)	24,56	17,05	12,44
Modelo C3 zh-pt	20,01	14,14	6,95
Modelo C1 pt-zh	22,52	18,45	12,92
Modelo D1 pt(br)-zh	24,01	19,21	12,76
Modelo C3 pt-zh	19,57	14,11	6,86

Table 6.6: Resultados dos Modelo C e Modelo D

6.5.2 Resultados dos Tamanhos Diferentes dos Bancos de Dados

Para os Modelo E e Modelo B, utilize o corpus paralelo zh-pt para os experimentos. O Modelo E utiliza 5 milhões de pares de sentenças, e o Modelo B utiliza todas as sentenças, ou seja, 7.188.706 pares de sentenças para pré-treinamento. Por conta dos parâmetros do Modelo B serem os mesmos do Modelo B na Seção 6.4.1 desta dissertação, então só é necessário experimentar o Modelo E. Além dos dados de treinamento, o tamanho do vocabulário de treinamento é 41.020, e o restante da configuração é igual ao Modelo B. Após 7 época e 420 horas, a função de perda do Modelo E permanece basicamente inalterada, podendo parar de treinar e salvar a distribuição de peso. Depois disso, o vocabulário e a distribuição de pesos são trazidos para o modelo de tradução e os indicadores são avaliados diretamente. O Modelo E não requer o segundo conjunto de experimentos em seção 6.1 Os resultados são os seguintes:

	Rouge-1	Rouge-2	BLEU
Modelo E1 zh-pt	19,64	12,46	9,55
Modelo B1 zh-pt	25,21	21,48	13,71
Modelo E3 zh-pt	16,54	10,38	5,75
Modelo B3 zh-pt	21,01	13,15	7,76
Modelo E1 pt-zh	21,20	13,35	10,12
Modelo B1 pt-zh	24,53	19,98	12,95
Modelo E3 pt-zh	17,12	13,55	6,02
Modelo B3 pt-zh	21,67	20,21	7,96

Table 6.7: Resultados dos Modelo E e Modelo B

6.6 Análise Completa

Conta todos os modelos e resultados experimentais, e os resultados estatísticos são os seguintes(tabela 6.8):

	Tokenizer	Modelo de Prioridade	Idioma	Tamanhos dos Dados	Tamanho do Vocabulário	Número de Épocas
Modelo A1	LAC	Palavra	zh-pt	7188706	40901	7
Modelo B1	Jieba	Palavra	zh-pt	7188706	41885	7
Modelo C1	Jieba	Palavra	zh-pt	6500000	41820	10
Modelo D1	Jieba	Palavra	zh-pt(br)	6500000	42210	10
Modelo E1	Jieba	Palavra	zh-pt	5000000	41020	7
Modelo F1	Jieba	Letra	zh-pt	7188706	41885	7
Modelo A2	LAC	Palavra	zh-pt	6829270	40887	7
Modelo B2	Jieba	Palavra	zh-pt	6829270	41843	7
Modelo F2	Jieba	Letra	zh-pt	6829270	41843	7
Modelo A3	LAC	Palavra	zh-pt	6829270	40887	7
Modelo B3	Jieba	Palavra	zh-pt	6829270	41843	7
Modelo C3	Jieba	Palavra	zh-pt	6500000	41820	10
Modelo E3	Jieba	Palavra	zh-pt	5000000	41020	7
Modelo F3	Jieba	Letra	zh-pt	6829270	41843	7

Table 6.8: Estatísticas de detalhes dos modelos

Este capítulo conduz um experimento de avaliação passo a passo, avaliando as pontuações de cada modelo Rouge-1, Rouge-2 e BLEU, incluindo: Google translate como a pontuação de baseline, usando os tokenizers Jieba e LAC e usando conjuntos diferentes de dados. No caso de quantidades diferentes de dados, avaliações diferentes são obtidas.

No geral, os modelos de pontuação mais alta foram aqueles usando o tokenizer Jieba e modelos usando conjuntos de dados maiores. Em termos de indicadores de avaliação, o Chinês utiliza particípios Jieba, o Português utiliza particípios BPE, e o BLEU usa 4 gramas.

	Rouge-1	Rouge-2	BLEU
Modelo A1 zh-pt	22,25	14,47	11,87
Modelo B1 zh-pt	25,21	21,48	13,71
Modelo C1 zh-pt	24,46	18,95	12,65
Modelo D1 zh-pt(br)	24,56	17,05	12,44
Modelo E1 zh-pt	19,64	12,46	9,55
Modelo F1 zh-pt	25,01	21,07	13,61
Modelo A2 zh-pt	22,01	14,13	11,05
Modelo B2 zh-pt	24,95	21,13	12,95
Modelo F2 zh-pt	24,65	20,83	12,80
Modelo A3 zh-pt	20,95	12,21	7,54
Modelo B3 zh-pt	21,01	13,15	7,76
Modelo C3 zh-pt	20,01	14,14	6,95
Modelo E3 zh-pt	16,54	10,38	5,75
Modelo F3 zh-pt	19,95	12,65	7,20
Modelo A1 pt-zh	22,64	15,31	12,21
Modelo B1 pt-zh	24,53	19,98	12,95
Modelo C1 pt-zh	22,52	18,45	12,92
Modelo D1 pt(br)-zh	24,01	19,21	12,76
Modelo E1 pt-zh	21,20	13,35	10,12
Modelo F1 pt-zh	23,95	19,30	12,65
Modelo A2 pt-zh	22,07	14,85	12,06
Modelo B2 pt-zh	24,01	19,45	12,18
Modelo F2 pt-zh	23,02	18,96	12,01
Modelo A3 pt-zh	21,16	12,65	7,85
Modelo B3 pt-zh	21,67	20,21	7,96
Modelo C3 pt-zh	19,57	14,11	6,86
Modelo E3 pt-zh	17,12	13,55	6,02
Modelo F3 pt-zh	20,87	19,44	7,65

Table 6.9: Resultados dos Modelos

6.7 Análise de Conclusão

A tabela(6.8) e tabela(6.9) apresenta os resultados experimentais, fornecendo informações claras. Usam-se diferentes modelos e materiais de treinamento para realizar experimentos comparativos com o intuito de obter um modelo melhor para a tradução automática Chinês-Português. Comparam-se quatro conjuntos de experimentos, que são respetivamente: Experimento(i), a comparação de treinamento entre um modelo que usa o tokenizer Jieba para processar o modo de mix-word Chinês e um modelo que usa o tokenizer LAC. Experimento (ii) modelo Letra-primeira e do Modelo Palavra-primeira nos Resultados (iii), a comparação entre o modelo de tradução automática com zh-pt como conjunto

de treinamento e o modelo de tradução automática com zh-pt (br) como conjunto de treinamento sob o mesmo tamanho de corpus. Experimento(iv), a comparação de treinamento entre o Modelo Com 5 milhões de conjuntos de treinamento e o Modelo Com mais de 6,5 milhões de conjuntos de treinamento. Além disso, também é testado o efeito de tradução de nosso conjunto de testes no Google translate, que é usado como uma comparação de baseline.

Ao realizar experimentos (i) comparativos, descobre-se que tokenizers diferentes têm impacto no modelo de tradução automática. Devido às características de nosso corpus com muitas frases curtas, processamento imperfeito e expressão mais tendenciosa para obras de cinema e televisão, entre a segmentação de palavras em Jieba e LAC, na tarefa de tradução zh-pt, o tokenizer Jieba obtém a pontução mais alta. O Modelo B1 usando o tokenizer Jieba tem uma pontuação BLEU de 13,71 pontos, que é maior do que o Modelo A1 usando o tokenizer LAC, que obteve 11,87 pontos. Também o Modelo B2 tem uma pontuação BLEU de 12,95 pontos, que é maior do que o Modelo A2, que obteve 11,05 pontos. E o Modelo B3 tem uma pontuação BLEU de 7,76 pontos, que é maior do que o Modelo A3, que obteve 7,54 pontos. E na tarefa de tradução pt-zh, a pontuação BLEU do Modelo B1 também foi superior em 12,21 pontos do Modelo A1 em 12,95. O Modelo B2 tem uma pontuação BLEU de 12,18 pontos, que é maior do que o Modelo A2, que obteve 12,06 pontos. E o Modelo B3 tem uma pontuação BLEU de 7,96 pontos, que é maior do que o Modelo A3, que obteve 7,85 pontos. Após a análise do corpus, acredita que linguagens, padrões de sentenças e comprimentos diferentes requerem o uso de tokenizers diferentes. Para outros experimentos desta dissertação, é selecionado o tokenizer Jieba.

Na comparação de (ii), o impacto de diferentes modos de prioridade nos resultados não são grande, mas a partir dos dados gerais, os resultados experimentais no modelo de palavra-primeira são geralmente melhor do que os resultados da tradução no modelo de letra-primeira. Na pontuação BLEU, zh-pt na tarefa de tradução, o Modelo B1 de palavra-primeira com 13,71, seja maior do que o Modelo F1 de letra-primeira com 13,61, o Modelo B2 tem uma pontuação BLEU de 12,95 pontos, que é maior do que o Modelo F2, que obteve 12,80 pontos. E o Modelo B3 tem uma pontuação BLEU de 7,76 pontos, que é maior do que o Modelo F3, que obteve 7,20 pontos. E na tarefa de tradução pt-zh, o Modelo B1 de palavra-primeira com 12,95, seja maior outra vez do que o Modelo F1 de letra-primeira com 12,65. Portanto, o Modelo B2 tem uma pontuação BLEU de 12,18 pontos, que é maior do que o Modelo F2, que obteve 12,01 pontos. E o Modelo B3 tem uma pontuação BLEU de 7,96 pontos, que é maior do que o Modelo F3, que obteve 7,65 pontos. Na tarefa de tradução, a tradução automática entre o Chinês e Português será mais adequada com o modelo de palavra-primeira.

Na comparação de (iii), conjuntos diferentes de dados não trouxeram grandes diferenças. Na tarefa de tradução zh-pt, a pontuação de BLEU do Modelo C usando o corpus zh-pt em 12,65 foi um pouco maior do que o Modelo D usando zh-pt (br) em 12,44. Na tarefa de tradução pt-zh, o pontução de BLEU do Modelo C em 12,92 também é ligeiramente superior ao Modelo D em 12,76. A diferença da pontuação Rouge dos 2 modelos também não é grande. Corpora diferentes não trouxeram resultados diferentes, o que não é apenas um bom resultado para a estabilidade do modelo, mas também uma maior probabilidade de uso do Modelo Em mais corpora com características diferentes.

Então usa-se o Modelo B de experimento(i) e o Modelo E do pequeno conjunto de dados após (iv) comparação. Descobrimos que sob nosso modelo de treinamento entre o Modelo B, com o maior conjunto de treinamento e o Modelo E, com menor conjunto de treinamento, o Modelo B obteve resultados melhores. Entre os três escores de Rouge-1, Rouge-2 e BLEU, na tradução de zh-pt, a pontução de BLEU do Modelo B1 foi de 13,71, superando os 9,55 do Modelo E1, o Modelo B3 tem uma pontuação BLEU de 7,76 pontos, que é maior do que o Modelo E3, que obteve 5,75 pontos. Enquanto na tradução de pt-zh, a pontuação de BLEU do Modelo B1 em 12,95 também é maior do que do Modelo E1 em 10,12, o Modelo B3 tem uma pontuação BLEU de 7,96 pontos, que é maior do que o Modelo E3, que obteve 6,02 pontos. Os resultados acima são muito intuitivos e consistentes com o princípio do algoritmo principal. Um conjunto de treinamento maior pode trazer treinamento mais suficiente para obter melhores resultados de treinamento e de pontuações do modelo.

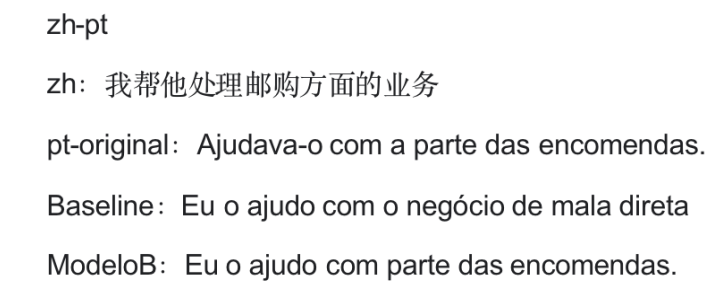

Figure 6.5: Um exemplo de resultados de zh-pt tradução

Comparei nossos resultados com Google translate como baseline e descobri que as pontuações de Rouge-1, Rouge-2 e BLEU foram maiores do que Google translate. Comparado o Modelo B que usa o mesmo conjunto de teste com Baseline no experimento(i), todos os dados são significativamente mais altos do que Baseline. A pontuação de avaliação BLEU da tarefa de tradução zh-pt em Modelo B1 13,71 é maior que a pontuação de Baseline em 5,15, e em Modelo B2 12,95 é maior que a pontuação de Baseline em 5,35. E a pontuação de avaliação BLEU da tarefa de tradução pt-zh em Modelo B1 12,95 é superior ao Baseline

em 4,51 e em Modelo B2 12,18 é superior ao Baseline em 4,72 na tabela 6.3. Por causa da grande diferença, pega um resultado de tradução para referência:

A figura(6.5) As sentenças do Baseline são mais rotineiras, e as sentenças do Modelo B estão mais próximas do texto original, razão pela qual a pontuação BLEU é muito baixa para o Baseline.

pt-zh

pt：Fecharam o negócio?

zh original：交易做完了吗

Baseline：他们完成交易了吗?

ModeloB：交易完成了吗?

Figure 6.6: Um exemplo de resultados de pt-zh tradução

A figura(6.6) Baseline novamente fornece uma boa tradução, e o uso de palavras e a ordem das palavras leva a uma grande diferença com o texto original do conjunto de teste, e o Modelo B obtém uma pontuação melhor.

Mas em zh-pt, o Modelo B3 tem uma pontuação BLEU de 7,76, não é tão bom quanto o 11,15 da baseline, em pt-zh, o Modelo B3 é 7,96, também não é tão bom quanto o 13,50 da baseline. No teste do conjunto de dados News-Commentary11, o google obteve melhores resultados.

Por fim, devido a certas diferenças no conjunto de teste, zh-pt(br) e pt(br)-zh não são precisos em comparados precisos com a baseline. No entanto, devido às partes sobrepostas do conjunto de teste, você pode tentar fazer uma comparação aproximada e não pode tirar uma conclusão. Em zh-pt(br) experimento (iii) Modelo D1 tem uma pontuação BLEU de 12,44, é maior que a pontuação de Baseline em 5,02, e em pt(br)-zh, Modelo D1 tem uma pontuação BLEU de 12,76, é maior que a pontuação de Baseline em 4,73.

Capítulo 7

Conclusão e Trabalho Futuro

Esta dissertação utiliza o algoritmo avançado de tradução automática para tradução Chinês-Português, utiliza o modelo de mistura de palavras e RoBERTa para pré-treinamento e, em seguida, utiliza BERT para tradução, a fim de obter melhor qualidade de tradução Chinês-Português por meio deste método. As principais contribuições desta dissertação estão listadas na seção 7.1, seguidas de uma análise de trabalhos futuros na seção 7.2.

7.1 Contribuição

As principais contribuições do trabalho realizado nesta dissertação são:

A viabilidade e necessidade da tradução automática Chinês-Português são analisadas. O mercado de aplicação de tradução automática Chinês-Português é muito grande, e a conexão entre os países de língua chinesa e os países de língua portuguesa está cada vez mais próxima. Atualmente, faltam pesquisas da tradução automática Chinês-Português no mundo. A pesquisa em tradução automática Chinês-Português é adequada e viável e os pesquisadores também são limitados, portanto é muito necessário fazer essa pesquisa.

Modelos de tradução de última geração são utilizados na tradução automática Chinês-Português. O modelo RoBERTa é o mais avançado e a estrutura de segmentação mistas de palavras são usados para pré-treinamento, e o BERT é usado para tradução. No corpus paralelo Chinês-Português disponível publicamente, seleciona o Opensubtitles2016 que tem maior quantidade dados. E usa os indicadores de avaliação BLEU e Rouge que são mais versáteis na tradução automática.

São projetados os modelos avançados para tradução automática Chinês-Português, e são realizados vários conjuntos de experimentos. Foram realizados experimentos de tradução automática Chinês-Português usando modelos avançados, e os resultados foram obtidos em diferentes situações. Experimentos de tradução automática foram realizados nos dados livres disponíveis e quatro grupos de experimentos foram desenhados, seja (i),

o efeito de tokenizers diferentes nos resultados do modelo de tradução Chinês-Português, (ii)modelo Letra-primeira e do Modelo Palavra-primeira nos Resultados (iii), os efeitos diferentes corpora no mesmo modelo, pt e pt(br) foram selecionados para experimentos respectivamente e, (iv) O impacto de tamanhos diferentes de conjuntos de treinamento no modelo de tradução. E compara com o Baseline fora dos três grupos de experimentos.

Resolveu o processo geral de tradução automática Chinês-Português, e analisou os problemas existentes na tradução automática Chinês-Português. Atualmente, os excelentes resultados da tradução automática Chinês-Português no mundo vêm do método de usar uma terceira língua como centro. Existem poucos estudos sobre tradução direta Chinês-Português. Adota-se o modelo avançado de uma tradução direta Chinês-Português, e a capacidade de tradução é limitada pelos corpora paralelos. Na pesquisa de acompanhamento, o foco da pesquisa deve ser na direção do corpus paralelo Chinês-Português.

7.2 Trabalho Futuro

7.2.1 Construção de Corpus Paralelo

No processo de tradução automática Chinês-Português, a importância dos dados básicos é profunda. E com falta de corpora paralelos na tradução automática, nossa pesquisa foi passiva.

Acredita-se que a tradução automática deva ser um tema de pesquisa impulsionado pela demanda do mercado e desenvolvido com a avanço de algoritmos e poder de computação. No entanto, no caso da tradução automática Chinês-Português, onde os recursos do corpus são escassos, onde só pode escolher os melhores do corpus paralelo existente, isso limita muito a margem da pesquisa, e a qualidade do corpus determina a qualidade da tradução em sua extensão.

Nas pesquisas futuras da tradução Chinês-Português, vamos nos concentrar na construção de um corpus paralelo. Atualmente, os pesquisadores obtiveram corpora paralelos de sites do governo, notícias, obras de cinema e televisão, etc. para construir corpora. Podemos seguir os seus métodos e tentar melhorá-los para construir um corpus maior, mais preciso e mais amigável à tradução automática.

Em nossa pesquisa, antes de entrar no modelo de tradução, foi submetida a um processamento tedioso em corpus e há alguns problemas que não são adequados para o processamento. No processo de construção do corpus, deve haver um pensamento inverso, partindo do uso real e processando os dados o máximo possível de uma só vez, de modo a fornecer um melhor suporte para a tradução automática.

Alguns pesquisadores tentam usar o método da língua central para a tradução automática Chinês-Português. Embora seja um bom método, ainda é prejudicado pela qualidade da língua central. No entanto, no processo de construção do corpus, a linguagem central pode se tornar uma boa direção para desenhar materiais. A escala do corpus paralelo pode ser expandida tanto quanto possível, e um bom ponto de equilíbrio deve ser atingido em termos de escala, qualidade e custo para melhorar realmente os dados básicos da tradução automática de idioma Chinês-Português.

7.2.2 Processamento de Corpus Paralelo

No processo de nossa pesquisa, o maior constrangimento é a falta de corpus básico. A qualidade da tradução é diretamente afetada pela qualidade do corpus de treinamento. Embora tenha feito muito processamento de dados, ainda não é suficiente para adaptar o corpus em um material de base adaptado para tradução automática.

Em trabalhos futuros, podemos tentar modelar e treinar o corpus paralelo Chinês-Português separadamente, tentando usar a rede Gan [43] para extrair sentenças alinhadas e sentenças de alta qualidade no corpus, e usando a automação para lidar com pontuação ns frases. Também é possível expandir o corpus em direções diferentes combinando frases no corpus em diferentes comprimentos e sequências.

Em trabalhos futuros, quero tentar usar a Back-translation [44] para criar ou expandir o corpus, e no estudo, usar o treinamento adversário [45] sobre os resultados da tradução de codificação, para aumentar o volume de treinamento e obter resultados mais precisos.

7.2.3 Métricas de Avaliação e Visualização

Na tradução automática, o índice de avaliação é um importante método de referência, mas por causa e ambientes linguísticos diferentes e difentes tipos de linguagem, atualmente não existe um índice de avaliação perfeito. Nos últimos anos também não houve um índice de avaliação perfeito, por isso é particularmente importante visualizar o modelo de tradução.

A pesquisa sobre tradução automática Chinês-Português de Santos et al.e de muitas empresas grandes visualizaram a tradução automática Chinês-Português. O que tornará mais fácil obter subjetivamente resultados de tradução de idiomas de tipos diferentes, campos diferentes, frequências diferentes e cenários diferentes.

A pesquisa sobre visualização não tem nada a ver com a qualidade da tradução automática em si, mas também é uma parte indispensável do conteúdo da pesquisa no caminho para o desenvolvimento da tradução automática.

Referências

[1] Wikimedia: *File:New-Map-Sinophone World.svg*. https://commons.wikimedia.org/wiki/File:New-Map-Sinophone_World.svg#mw-jump-to-license, 16.10.2022. xiii, 4

[2] Wikimedia: *Official Portuguese language in the World*. https://commons.wikimedia.org/wiki/File:Official_Portuguese_language_in_the_World.svg, 15.9.2022. xiii, 5

[3] Bahdanau, Dzmitry, Kyunghyun Cho e Yoshua Bengio: *Neural machine translation by jointly learning to align and translate*. arXiv preprint arXiv:1409.0473, 2014. xiii, 20, 21

[4] Alammar, J: *The Illustrated Transformer [Blog post]*. https://jalammar.github.io/illustrated-transformer/, 2018. xiii, 23

[5] Vaswani A, Shazeer N, Parmar N et al.: *Attention is all you need*. Advances in neural information processing systems, 30, 2017. xiii, 24, 25

[6] Papineni, Kishore, Salim Roukos, Todd Ward e Wei Jing Zhu: *Bleu: a method for automatic evaluation of machine translation*. Em *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, páginas 311–318, 2002. xiii, 13, 16, 51, 52

[7] Yachao, Li, Xiong Yimin e Zhang Min: *uma pesquisa sobre tradução automática neural*. Revista de Ciência da Computação, 41(12):2734–2755, 2018. xiv, 17

[8] Lison, Pierre e Jörg Tiedemann: *Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles*. 2016. xiv, 30, 43, 44, 45

[9] Xiaolong, Wang e Guan Yi: *processamento de linguagem natural do computador*. Tsinghua University Press Co., Ltd., 2005. 1, 30, 31

[10] Santos, Rodrigo Soares dos: *Portuguese-Chinese neural machine translation*. Tese de Doutoramento, 2019. 2

[11] Jianming, Lu e Shen Yang: *Quinze Palestras sobre Língua Chinesa e Estudos Chineses*. BEIJING BOOK CO. INC., 2016. 2

[12] GOV, China: *Em 2021, a escala de importação e exportação ultrapassará 6 trilhões de dólares americanos pela primeira vez*. http://www.gov.cn/xinwen/2022-01/15/content_5668288.htm, 15.1.2022. 3

[13] Bentang, Zhao: *Embaixador da China em Portugal, entrevista ao "New Weekly"de Portugal*. http://pt.china-embassy.gov.cn/sgxw/202202/t20220208_10639939.htm, 8.02.2022. 3

[14] São Paulo, Consulado Geral da China em: *China é o maior parceiro comercial do Brasil pelo 13º ano consecutivo*. http://stpaul.mofcom.gov.cn/article/jmxw/202202/20220203280694.shtml, 17.02.2022. 3

[15] China, Escritório Nacional de Estatísticas da: *Chinese population*. http://www.gov.cn/shuju/hgjjyxqk/xiangqing/np.html, 2022. 3

[16] IBGE: *Coordenação de População e Indicadores Sociais - COPIS*. https://ftp.ibge.gov.br/Estimativas_de_Populacao/Estimativas_2021/estimativa_dou_2021.pdf, 2021. 3

[17] Santos, Rodrigo Soares dos: *Portuguese-Chinese neural machine translation*. Tese de Doutoramento, 2019. 13, 53

[18] Sutskever, Ilya, Oriol Vinyals e Quoc V Le: *Sequence to sequence learning with neural networks*. Advances in neural information processing systems, 27, 2014. 13, 19

[19] Cho, Kyunghyun, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk e Yoshua Bengio: *Learning phrase representations using RNN encoder-decoder for statistical machine translation*. arXiv preprint arXiv:1406.1078, 2014. 13, 19

[20] Cho, Kyunghyun, Bart Van Merriënboer, Dzmitry Bahdanau e Yoshua Bengio: *On the properties of neural machine translation: Encoder-decoder approaches*. arXiv preprint arXiv:1409.1259, 2014. 13, 19

[21] Macdonald, Neil: *Language translation by machine-a report of the first successful trial*. Computers and automation, 3(2):6–10, 1954. 16

[22] Weaver, Warren: *The mathematics of communication*. Scientific American, 181(1):11–15, 1949. 16

[23] Bengio, Yoshua, Réjean Ducharme e Pascal Vincent: *A neural probabilistic language model*. Advances in neural information processing systems, 13, 2000. 17

[24] Smets, Philippe: *Belief functions: The disjunctive rule of combination and the generalized Bayesian theorem*. International Journal of Approximate Reasoning, 9(1):1–35, 1993, ISSN 0888-613X. https://www.sciencedirect.com/science/article/pii/0888613X9390005X. 18

[25] Zhang, Jiajun, Chengqing Zong *et al.*: *Deep Neural Networks in Machine Translation: An Overview*. IEEE Intell. Syst., 30(5):16–25, 2015. 19

[26] Kalchbrenner, Nal e Phil Blunsom: *Recurrent continuous translation models*. Em *Proceedings of the 2013 conference on empirical methods in natural language processing*, páginas 1700–1709, 2013. 19

[27] Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer e Veselin Stoyanov: *Roberta: A robustly optimized bert pretraining approach*. arXiv preprint arXiv:1907.11692, 2019. 27, 53

[28] Kingma, Diederik P e Jimmy Ba: *Adam: A Method for Stochastic Optimization. international conference on learning representations (2015)*. San Diego, California, 2015. 27

[29] Crystal, David: *Profiling linguistic disability*. 1992. 28

[30] Crystal, David: *A dictionary of linguistics and phonetics*. John Wiley & Sons, 2011. 28

[31] Sinclair, John e Les Sinclair: *Corpus, concordance, collocation*. 1991. 28

[32] McEnery, Tony, Richard Xiao e Yukio Tono: *Corpus-based language studies: An advanced resource book*. Taylor & Francis, 2006. 29

[33] Baker, Mona: *Corpora in translation studies: An overview and some suggestions for future research*. Target. International Journal of Translation Studies, 7(2):223–243, 1995. 29

[34] Johansson, Stig: *On the role of corpora in cross-linguistic research*. Language And Computers, 24:3–24, 1998. 29

[35] Liu, Siyou, Longyue Wang e Chao Hong Liu: *Chinese-Portuguese machine translation: a study on building parallel corpora from comparable texts*. arXiv preprint arXiv:1804.01768, 2018. 30

[36] Tiedemann, Jörg: *Parallel data, tools and interfaces in OPUS.* Em *Lrec*, volume 2012, páginas 2214–2218, 2012. 30, 44

[37] Chao, Lidia S, Derek F Wong, Chi Hong Ao e Ana Luìsa Leal: *UM-PCorpus: a large Portuguese-Chinese parallel corpus*. Em *Proceedings of the LREC 2018 Workshop "Belt & Road: Language Resources and Evaluation*, páginas 38–43, 2018. 30

[38] Wong, Fai e Sam Chao: *PCT: Portuguese-Chinese machine translation systems*. Journal of translation studies, 13(1-2):181–196, 2010. 31

[39] SnowNLP: *Simplified Chinese Text Processing*. `https://github.com/isnowfy/snownlp`, 1801.2020. 35

[40] [AWS]Amazon Web Services, Inc: *Serviços de computação em nuvem*. `https://aws.amazon.com/`, 2022. 48

[41] Lin, Chin Yew: *Rouge: A package for automatic evaluation of summaries*. Em *Text summarization branches out*, páginas 74–81, 2004. 51

[42] Devlin, Jacob, Ming Wei Chang, Kenton Lee e Kristina Toutanova: *Bert: Pre-training of deep bidirectional transformers for language understanding*. arXiv preprint arXiv:1810.04805, 2018. 53

[43] Mogren, Olof.: *C-RNN-GAN: Continuous recurrent neural networks with adversarial training.* arXiv preprint, 30, 2016. 67

[44] Edunov, Sergey, Myle Ott, Michael Auli e David Grangier: *Understanding back-translation at scale.* arXiv preprint arXiv:1808.09381, 2018. 67

[45] Mogren, Olof: *C-RNN-GAN: Continuous recurrent neural networks with adversarial training.* arXiv preprint arXiv:1611.09904, 2016. 67