rice2024 Proceedings of the Ninth International Conference on Research in Intelligent Computing in Engineering pp. 1–10
DOI: 10.15439/2024R81
ISSN 2300-5963 ACSIS, Vol. 42

Seq2Seq Transformer-Based Model for Optimized Chinese-to-English Translation

Sumayya Afreen
Department of Computer Science and Engineering
Stanley College of Engineering and Technology for Women
Hyderabad, India
asumayya@stanley.edu.in

Nguyen Thi Dieu Linh
Department of Science and Technology
Hanoi University of Industry
Hanoi, Vietnam
nguyenlinh79.haui@gmail.com

Sritisha Kodur
Department of Computer Science and Engineering
Stanley College of Engineering and Technology for Women
Hyderabad, India
sritishakodur@gmail.com

Asma Begum
Department of Artificial Intelligence and Data Science & Computer Engineering
Stanley College of Engineering and Technology for Women
Hyderabad, India
basma@stanley.edu.in

***Abstract*—The use of transformer models for machine translation from Chinese to English is examined in this research. The transformer design, which is well-known for its self-attention mechanism, makes it possible to handle Chinese's intricate linguistic structures with efficiency. We assess the model's effectiveness using benchmark datasets, examine its translation correctness through cosine similarity scores, Rouge metric scores and draw attention to important issues including managing context and sentence structure inconsistencies. We also explore situations in which language complexity is observed to result in low accuracy, providing valuable information for enhancing future models. This paper highlights areas for optimization in practical situations and shows how transformers might improve translation quality.**

***Index Terms*—Cosine Similarity, Language complexity, Machine translation, Rouge Metrics, Transformer models.**

I. Introduction

RECENT advances in Natural Language Processing (NLP), especially regarding the transformer model, have had a significant impact on machine translation progress. Vaswani et al. (2017) created the transformer architecture, which has completely changed the way textual input is processed by removing the drawbacks of conventional Recurrent Neural Networks (RNNs) such parallelization problems and vanishing gradients. As a result, transformers have excelled at several NLP tasks, including machine translation, by exploiting self-attention mechanisms for better addressing long-range dependencies in text Transformer-based models, which exclusively rely on self-attention processes instead of RNNs' sequential processing approach, have completely changed the field of machine translation. With the help of this attention mechanism, the model can concentrate on several phrase components at once, better capturing long-range dependencies and enhancing translation accuracy. Unlike RNNs, which suffer from vanishing gradient difficulties and are hard to parallelize, transformers enable efficient parallel processing, which not only accelerates training but also enhances the model's capacity to handle complicated sentence structures. The better performance of the transformer over traditional models in a variety of language pairs, including Chinese-to-English, has been used to illustrate its usefulness in neural machine translation (NMT) tasks [1].

There are more than just linguistic differences in structure when translating from Chinese to English. The gap between literal and non-literal translation procedures has long been highlighted by translation theorists such as Newmark (1981) and Vinay & Darbelnet (1958). This distinction is particularly crucial when translating between languages that are as dissimilar as Chinese and English. In Chinese-to-English translation, nonliteral translation strategies are essential since many Chinese expressions—particularly idioms and colloquialisms—cannot be translated literally without losing their original meaning. To guarantee that the translated meaning stays faithful to the original, nonliteral expressions like "刑让我的辛苦白费了" (which means "Don't let my hard work be wasted") must be understood in context. Research has demonstrated that these kinds of nonliteral translations have historically proven difficult for machine translation models, especially RNN-based systems, often producing grammatically correct but semantically inaccurate translations [2].

The self-attention mechanism of the transformer model is crucial in this situation. Transformers are better able to handle translations including nonliteral phrases or complex sentence structures by letting the model determine the relative

value of various words in a sentence. Additionally, recent improvements in pre-trained language models such as BERT and GPT have further increased the capabilities of transformer models by including contextual awareness and semantic nuances [3]. For example, when applied to Chinese-English translation tasks, transformers have demonstrated considerable gains in BLEU scores, a metric typically used to evaluate the accuracy of machine translations. Transformer models have certain drawbacks even if they typically function incredibly well. One aspect that still requires work is their capacity to handle exceedingly long and contextually complex statements. Transformers can catch word dependencies thanks to the self-attention mechanism, but it can occasionally have trouble keeping long sentences coherent, particularly when translating materials that call for in-depth cultural knowledge or subject-specific expertise. According to studies by Chen et al. (2020) and Zhang et al. (2021), transformer models occasionally fall short of accurately capturing the context of colloquial terms or domain-specific jargon, which might result in less accurate translations in specialised sectors like legal or medical writings [4]

The linguistic and cultural disparities between Chinese and English have frequently presented difficulties for those translating public signage in China. Amenador and Wang (2020) draw attention to this problem by pointing out that a lot of translations fall short of the original meaning, which causes misunderstanding among audiences that speak English. Their work applies functional theory to the analysis of translation errors in public signs, highlighting the significance of tailoring translations to the target audience's communicative needs rather than following the source text exactly. The study proposes that translators can enhance the quality of translations and hence improve China's foreign image by adopting a purpose-driven approach [5]. To sum up, the transformer model, which offers a more reliable and scalable method than previous RNN-based systems, has greatly advanced the field of Chinese-to-English translation. Its self-attention mechanism makes it possible to handle literal and nonliteral translations more effectively, which makes it an effective tool for resolving the difficulties that come with translating between Chinese and English. To solve issues with context retention and the translation of texts that are extremely domain-specific, more study is still needed. Future advancements in transformer-based designs, such the incorporation of trained models like BERT or GPT, have the potential to significantly enhance translation quality in a variety of language contexts. The purpose of this study is to examine these developments and go over the issues still facing near-human translation accuracy [6].

This paper aims to develop and evaluate a robust Seq2Seq Transformer model tailored for accurate and efficient Chinese-to-English translation. The study leverages ROUGE metrics to assess the quality of generated translations and employs cosine similarity to measure semantic alignment with reference translations, ensuring both linguistic and contextual fidelity.

The paper is organized as follows: Section 2: Literature Review provides an overview of existing research on neural machine translation and the application of Seq2Seq Transformer models, highlighting gaps addressed in this work. Section 3: Methodology details the model architecture, data preprocessing steps, and the evaluation framework using ROUGE and cosine similarity metrics. Section 4: Results presents the experimental outcomes and compares the model's performance with existing benchmarks. Section 5: Future Work outlines potential enhancements, including broader language coverage and advanced optimization techniques. Finally, Section 6: Conclusion summarizes key findings and their implications for translation systems.

II. Related Work

Verb Semantics for English-Chinese Translation, published in 1995 by Martha Palmer and Zhibiao Wu, offers a thorough analysis of the difficulties associated with machine translation, with a particular emphasis on lexical differences between Chinese and English verbs. The authors stress that current NLP and machine translation (MT) systems frequently rely on precompiled lexicons that are unable to handle unforeseen word usage, and they frequently presume a fixed number of verb meanings. To overcome these shortcomings, this study suggests an improved method of verb semantics that takes lexical selection issues and verb sense expansions into consideration

Lexical Divergences between English and Chinese Verbs: The main problem noted in the paper is the substantial vocabulary differences between Chinese and English, especially in the structure of verbs and verb phrases. Verbs in English frequently combine action and consequence into a single lexical component. For example, in English, verbs like break might indicate a state or an activity without clearly stating the procedure or result. On the other hand, Chinese verbs often use compound formulations to clarify the action and the outcome. Due to this distinction, translating words between the two languages can be difficult since in Chinese, verb compounds like "da-sui" (hit-into-pieces) express both the action and the object's end condition. Given that verb translations must take into consideration both of these divergences, more sophisticated translation processes than those offered by conventional lexicons are needed.

Limitations of Existing Systems: Current MT and NLP systems are criticised by Palmer and Wu for their strict reliance on static verb sense lists, which makes them unsuitable for managing sense extensions or unforeseen verb usages. For example, English verbs like break may have numerous senses depending on context—whether referring to the physical state of an object, a break in continuity, or the malfunctioning of a device. The authors contend that the adaptability and inventiveness present in natural language are insufficiently addressed by systems built around a limited range of predetermined verb senses. This is especially evident when translating between Chinese and English, two languages with quite distinct vocabulary patterns. The tech-

nique of selectional limitations, which is often used in MT systems to limit verb arguments by specified categories, is criticised in this study. Although this approach can deal with senses of verbs, it is not very effective when dealing with more general linguistic events. This is particularly true when translating from English, which has more generalised verbs, to Chinese, which needs specificity in the description of actions and outcomes. The paper *Verb Semantics for English-Chinese Translation* by Martha Palmer and Zhibiao Wu (1995) provides an in-depth exploration of the complexities involved in machine translation, specifically focusing on lexical divergences between English and Chinese verbs. The authors emphasize that existing machine translation (MT) and natural language processing (NLP) systems often assume a fixed number of verb senses, relying on precompiled lexicons that lack the flexibility needed for dealing with unexpected word usages. This paper addresses these limitations by proposing an enhanced approach to verb semantics that accounts for both lexical selection problems and verb sense extensions.

Conceptual Lattices and UNICON: The authors suggest a novel approach based on conceptual lattices, a technique for encoding verb senses that enables more dynamic and adaptable mappings between verbs in various languages, to get around these restrictions. Verb meanings are arranged into a hierarchical structure by a conceptual lattice, which allows related senses to be placed together according to common semantic elements. With this method, the system can recognise verbal similarities between languages even in the absence of direct translations. By expanding verb senses beyond their preset definitions, the research prototype system UNICON puts this idea into practice and shows more accurate lexical selection.

One of the system's main advantages is its capacity to expand verb senses. The system can determine the closest related sense in the conceptual lattice when an unexpected verb usage happens and utilise this information to suggest a translation. This approach works especially well when there isn't a direct translation available in the target language. When an English verb, like break, is employed in an unusual way, for instance, the system can look up comparable verb senses in Chinese and choose the most appropriate one. UNICON can handle unexpected verb usages that would normally be outside the scope of typical lexicons by permitting sense extensions.

By emphasising the shortcomings of static verb sense lists and selectional limits in handling lexical divergences between English and Chinese, the paper significantly advances the field of machine translation. Palmer and Wu provide a more adaptable and context-sensitive method of verb semantics through their conceptual lattice framework, which enhances translation accuracy by taking into consideration the wider variety of verb usages and their potential extensions. The creation of UNICON, a useful tool for improving machine translation systems, proves the viability of this strategy [7].

Wazib Ansar, Saptarsi Goswami, and Amlan Chakrabarti's paper "A Survey on Transformers in NLP with Focus on Efficiency" delves deeply into the emergence of transformer-based models in NLP, emphasising the need to strike a balance between these models' performance and efficiency. The introduction of transformer models like BERT, GPT, and XLNet has transformed the way linguistic tasks are carried out as the area of NLP has developed. In tasks like text summarization, sentiment analysis, and machine translation, these models have proven to perform at the cutting edge. Transformers are resource-intensive and require a large amount of memory, processing power, and energy, thus this advancement is not without a price (2406.16893v1).

The Evolution of NLP and the Rise of Transformers: The article begins with a succinct overview of NLP's past, showing how computational techniques have changed over time, moving from rule-based systems to machine learning techniques. The capacity to handle complex linguistic patterns was restricted by earlier techniques like rule-based and classic machine learning approaches, such Support Vector Machines (SVM) and Naive Bayes, which also required intensive feature engineering. Recurrent Neural Networks (RNNs) were a breakthrough in deep learning; nonetheless, it was hampered by issues such as vanishing gradients and managing long-range dependencies. With their self-attention mechanism, transformer models—which were first shown by Vaswani et al. (2017)—addressed these issues and enabled improved parallelization and long-range context capture (2406.16893v1).

As the study discusses, transformers have a few benefits over earlier systems. These models can provide varying weights to distinct words in a sentence according on how relevant they are to the context according to the self-attention mechanism, which improves understanding of linguistic subtleties. Because of this, transformers are especially useful for jobs like translation and summarization that call for an awareness of the larger context. Models such as GPT (Generative Pretrained Transformer) and BERT (Bidirectional Encoder Representations from Transformers) established new performance standards for NLP tasks, resulting in their broad use in research and industry (2406.16893v1).

The authors provide research showing how training big NLP models has an impact on the environment. For instance, Strubell et al. (2019) discovered that the energy required for several transatlantic trips can be like that required for training a sizable NLP model. There is increasing worry about how these models affect the environment, which is driving scientists to look into more accurate but more efficient alternatives. This is especially pertinent in light of sustainability and the rising demand for environmentally friendly artificial intelligence solutions (2406.16893v1).

Strategies for Improving Transformer Efficiency the study examines several techniques for improving transformer models' operating efficiency to address efficiency concerns without sacrificing performance. These include model com-

pression methods that have been used to decrease the size and computational load of these models, such as pruning, quantization, and knowledge distillation (Gordon et al., 2020; Hinton et al., 2015). Through pruning, a trained model's less significant weights are removed, creating a more efficient version that uses less memory and processing resources. Conversely, quantization lowers the weights in the model's numerical precision, which can help minimise resource requirements while keeping Through concentrating on condensing the information included in big models into more manageable and effective models, scientists can capitalise on the advantages of transformer architectures while lessening their impact on the environment and operational expenses [8].

Sparse Attention Mechanisms: Another unique way to boosting transformer efficiency mentioned in the research is the use of sparse attention mechanisms. For lengthy sequences, traditional transformers can become computationally expensive as they must compute attention scores for every pair of input tokens. The computational load is decreased by sparse attention techniques, which restrict the attention computations to a subset of pertinent tokens (Child et al., 2019). Models like Long former and Reformer indicate that by leveraging sparse attention patterns, transformers can effectively manage longer sequences without a substantial reduction in performance. These models are appropriate for a range of NLP applications because they not only increase efficiency but also maintain the contextual knowledge that transformers are recognized for.

Hardware's Place in Efficiency: The significance of hardware developments for transformer model optimization is also highlighted in the article. The speed and effectiveness of training and inference procedures can be greatly increased by using custom hardware solutions, such as Field Programmable Gate Arrays (FPGAs) and Tensor Processing Units (TPUs) (Jouppi et al., 2017). Researchers can reduce energy usage and enhance the performance of these models by customizing hardware for transformer computations. The efficiency issues raised by large models can be effectively resolved by combining transformer topologies with optimized hardware.

A. Research Directions for the Future

The necessity for continued research to investigate novel approaches for raising the efficiency of transformer models is emphasised in the paper's conclusion. As natural language processing (NLP) advances, it is critical to create models that are both highly effective and ecologically sound. The authors support a comprehensive strategy that integrates developments in model architecture, hardware, and algorithms to produce transformer systems that are more effective. Subsequent investigations ought to focus on discovering novel approaches to optimise the advantages of transformers while reducing their ecological footprint and resource needs. [8]

The development, applications, and prospects of Transformer-based models in natural language processing (NLP) are reviewed in detail in the paper "End-to-End Transformer-Based Models in Textual-Based NLP" by Abir Rahali and Moulay A. Akhloufi. The architecture, training modalities, and particular applications of Transformer-based models are highlighted by the authors, who also emphasise how these models have a transformative effect on NLP tasks.

The paper starts out by highlighting the profound change in natural language processing (NLP) that deep learning (DL) architectures—specifically, Recurrent Neural Networks (RNNs)—have brought about. It then shifts to the enhanced powers of Transformer models, which Vaswani et al. The self-attention mechanism, which allows the model to take into account long-range dependencies between tokens in a sequence, is the main innovation that distinguishes Transformers from RNNs and Convolutional Neural Networks (CNNs). This allows for more effective and scalable processing of textual data.

The review by the authors is structured around the development of Transformer-based (TB) models, beginning with the basic Transformer model and moving on to several adaptations that tackle architectural and performance issues. Despite their prior effectiveness in sequential data processing, the research demonstrates that classic RNN-based models were constrained by problems including disappearing gradients and the incapacity to manage long-term relationships efficiently. Transformer models tackle these challenges by leveraging self-attention techniques, allowing parallelization, and making them more efficient for tackling large-scale NLP jobs.

The review claims that the flexibility of Transformer-based models is their main advantage. The paper offers thorough insights into several important Transformer variations, each intended for a particular NLP purpose. Examples of models that use the Transformer architecture for different purposes are discussed in detail, such as BERT (Bidirectional Encoder Representations from Transformers) and GPT (Generative Pre-trained Transformer), where BERT is used for natural language understanding and GPT is used for text generation.

The literature study also focuses on newer models like RoBERTa, which improves on BERT by modifying training processes and using larger datasets. XLNet, which aims to combine the benefits of auto-regressive and auto-encoding models, and T5 (Text-to-Text Transfer Transformer), which reformulates all NLP tasks into a text-to-text format, are two other noteworthy models that were explored.

The writers argue that Transformer models are useful for a variety of natural language processing (NLP) tasks, such as sentiment analysis, question-answering, summarization, and machine translation. Because these models are better at learning contextual representations of text, they have shown themselves to be more effective than earlier architectures at handling such tasks. Although the enormous processing demands for training these models are acknowledged in the review as a potential drawback, pre-training and transfer learning advancements have somewhat alleviated these difficulties.

The use of pre-training and fine-tuning strategies in Transformer-based models is covered in a noteworthy section of the article. These models can be fine-tuned on smaller, task-specific datasets after gaining a generalised grasp of language through pre-training on large-scale, unlabeled datasets. It is now commonplace in NLP to use this transfer learning strategy to attain excellent performance on tasks with little labelled data. A discussion of the future directions for Transformer-based model research rounds off the overview. Enhancing these models' scalability and efficiency remains an open problem, especially in low-resource languages and data-poor areas. In order to further improve performance, the authors also point out the possibility of hybrid models, which integrate the advantages of several neural network architectures.

Overall, the study presents a complete overview of the state of Transformer-based models in NLP, highlighting their evolution, applicability, and ongoing research concerns [9]. Natural language processing (NLP) can be used to analyze and compare various English translations of The Analects, a classic work of Confucian philosophy. This is explored in the paper "A Semantic Similarity Analysis of Multiple English Translations of The Analects: Based on a Natural Language Processing Algorithm" by Liwei Yang and Guijun Zhou. To better understand how translation choices and styles affect the overall semantic integrity of the original text, the study compares and calculates the semantic similarities across 15 English translations of The Analects using a variety of NLP techniques.

The first section of the paper discusses The Analects' cultural relevance and its enduring influence on Chinese and international intellectual traditions. The Analects, which has been translated into English multiple times since the 17th century, is an important resource for international dialogue as well as Chinese cultural and philosophical study. However, the availability of different translations causes issues for readers, especially those unfamiliar with the historical and cultural background of the original text. The need to give readers a better grasp of the semantic divergence and convergence of various translations is the driving force behind this work.

The writers begin by reviewing The Analects' historical English translations, pointing out the variety of approaches taken by various translators. While some translators prioritised keeping the text as faithful to the source Chinese as possible, others focused more on making it readable and accessible to English-speaking audiences. The translators are divided into three categories in the paper: native English speakers, Chinese translators with Western education, and Chinese translators with traditional education. This classification is predicated on the linguistic and cultural backgrounds of the translators, which the authors speculate may impact translation decisions and, in turn, alter the semantic output of the translations.

The authors use five distinct natural language processing (NLP) techniques, namely TF-IDF (Term Frequency-Inverse Document Frequency), Word2Vec, GloVe, BERT, and SimHash, to evaluate the semantic similarity between translations. These algorithms provide various methods for analysing the semantics of text. Word2Vec and GloVe, for instance, build vector representations of words to capture semantic associations, while TF-IDF estimates the relevance of words in a text based on their frequency and how rare they are across documents. BERT, a transformer-based approach, leverages deep learning to capture increasingly complicated semantic patterns. Conversely, SimHash provides a less sophisticated but simpler comparison by gauging the similarity between texts based on their binary hash representations.

The study focuses on fifteen popular English versions of The Analects that have attracted a lot of interest from users on sites like Google Scholar, Goodreads, and Amazon. The authors determine pairwise similarities across translations using various NLP methods, and they provide the results as quantitative data. Their results show that most translations have a high degree of semantic similarity, especially those done by well-known translators like James Legge and D.C. Lau, although there are also notable discrepancies. The translators' employment of different Chinese annotations and interpretive frameworks, rather than their origins or the historical era in which they worked, is primarily responsible for these disparities, the authors contend.

One of the paper's main findings is that the selection of Chinese annotations is a significant factor in figuring out how similar various translations are semantically. For instance, translations that rely significantly on Zhu Xi's annotation, a Song Dynasty scholar, show higher levels of semantic coherence. Semantic divergence is higher in translations that attempt to rethink the original meaning of the text or include more contemporary remarks.[10].

The study "Advances in Chinese Natural Language Processing and Language Resources" by Jianhua Tao, Fang Zheng, Aijun Li, and Ya Li, offers an overview of recent developments in Chinese Natural Language Processing (CNLP), highlighting the construction and exploitation of linguistic resources and consortiums. The paper underlines the relevance of data-driven techniques and linguistic resources in NLP research, focusing on how the availability and sharing of corpora have contributed to major gains in both text and speech processing in Chinese.

Key Themes and Scope:The authors start out by stressing how important well-constructed corpora and linguistic data are to the advancement of CNLP research. For accuracy and usefulness, modern NLP techniques—especially statistical approaches—heavily rely on real-world data. Due to the complexity of Chinese and its distinct linguistic properties, it is essential to have access to large and reliable corpora in order to make significant progress in natural language processing (NLP) applications such as text categorisation, machine translation, speech recognition, and more. The wide variety of NLP tasks, such as machine translation, syntactic parsing, part-of-speech (POS) tagging, and word segmentation, are

also covered in the study. Notably, with accuracy rates of 98% and 95% in Chinese, word segmentation and POS tagging have attained notable success. But more difficult jobs like syntactic and semantic parsing continue to be difficult, partly because of the ambiguity and complexity inherent in natural language

B. Contributions to Chinese NLP

Lexicons: From generic word segmentation to specialized lexicons for Chinese geographic names, proper nouns, and organizations, the paper examines a variety of lexicons developed for diverse reasons. Notably, a large amount of word frequency data that is essential for many NLP applications can be found in the Chinese Web 5-gram Corpus.

POS-Tagged Corpora: POS-tagged resources, such as the People's Daily Corpus, offer crucial training data for machine learning models in applications like information retrieval and text categorization.

Multilingual Corpora: As machine translation has grown in popularity, the use of multilingual corpora—which pair Chinese with languages like English and Japanese—has become crucial for enhancing translation precision. Resources such as the Tsinghua Chinese Treebank are helpful in the development of tools like syntactic parsers and event detection systems, which are essential for higher-level NLP tasks. Another important aspect of CNLP that is covered in the study is speech processing. Specialized speech corpora have proven extremely beneficial for tasks such as speaker identification, synthesis, and speech recognition. As an illustration, the CASIA Mandarin Corpus is cited as a crucial tool for enhancing Mandarin voice synthesis and recognition. The study also highlights the significance of emotional speech data and regional dialects, both of which are utilized to enhance speech recognition systems' performance in a wider range of scenarios.

The importance of resource-sharing programs such as the Chinese Linguistic Data Consortium (CLDC) and the Chinese Corpus Consortium (CCC) is emphasized in the paper. The creation, gathering, and distribution of linguistic resources for use in scholarly and industrial applications is greatly aided by these consortiums. By making high-quality data available to researchers, they enable the development of better-performing CNLP systems and facilitate the replication of results across the field [11].

III. Methodology

A. Transformer model

The Transformer model, which broke with the conventional sequential data processing techniques of Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks, was first presented by Vaswani et al. in their seminal paper "Attention Is All You Need" (2017). This change fundamentally altered the field of Natural Language Processing (NLP). The self-attention mechanism, which is the foundation of the Transformer model, enables it to assess the connections between various input sequence elements regardless of how far apart they are from one another. The Transformer processes all tokens at once, in contrast to RNNs, which process tokens in a sequential manner. This makes the Transformer more computationally efficient and better suited for capturing long-range dependencies.

The encoder and the decoder are the two primary parts of the Transformer model's architecture. After processing the input sequence, the encoder creates hidden representations of the input. Then, using the previously generated output and these hidden representations, the decoder creates the output sequence. A multi-head self-attention mechanism and a position-wise feedforward network are the two sub-layers that make up each of the several layers that comprise the encoder and decoder. The model may focus on pertinent tokens and their context by attending to diverse areas of the input concurrently thanks to the multi-head self-attention method.

The application of positional encoding is another significant feature of the Transformer concept. The model misses the natural order of words in a series since it does not process input tokens sequentially. To ensure word order is taken into account during training, positional encodings are added to each token to provide the model with information about its location in the sequence.

NLP tasks including text summarization, language modelling, and machine translation have been transformed by transformers, which greatly outperform earlier models in terms of accuracy and speed. They are especially useful for activities like machine translation and question-answering systems that need to comprehend intricate relationships between tokens because of their capacity to manage long-range dependencies and parallelize computations.

B. seq 2 seq model

A particular use of the Transformer design for sequence-to-sequence operations is the Seq2Seq Transformer model, which entails producing an output sequence based on a supplied input sequence. This architecture is extensively used for jobs where one sequence needs to be converted into another, such as speech recognition, text synthesis, and machine translation.

Under the Seq2Seq Transformer model, the encoder generates a context-aware representation of the complete input by processing the input sequence first. Every word in the input sequence is examined in connection to every other word in the phrase, not only in isolation. This enables the model to comprehend the context-specific meaning of every phrase. In machine translation, for instance, the term "bank" in the sentence "I went to the bank" can mean different things depending on whether the context indicates that it's a riverbank or a financial institution.

The next token in the output sequence is then generated by the decoder using these context-aware representations from the encoder and the previously generated output tokens. In order to ensure coherence and consistency throughout the translation or sequence production, the decoder also uses self-attention mechanisms to examine all of the previously generated tokens. For example, when translating from

English to Chinese, the decoder creates a natural translation by utilising the context of the source English sentence in addition to the Chinese words that it has already generated.

The Seq2Seq Transformer model's superiority over conventional RNN-based Seq2Seq models lies in its capacity to manage intricate dependencies and distant interactions in sequences. When processing lengthy sequences, RNNs' sequential data processing frequently results in issues like disappearing gradients, where the model finds it difficult to remember information from previous tokens. Transformers are able to get over these restrictions and outperform other models on tasks that require lengthy sequences of events by processing all tokens concurrently and utilising self-attention processes.

Furthermore, the Transformer is far faster than RNNs or LSTMs due to its capacity for parallel processing, particularly when working with big datasets or lengthy sequences. Because of this, it is now the preferred model for a wide range of complex natural language processing tasks, especially in real-time applications such as live translation or automated speech recognition.

To sum up, the Seq2Seq Transformer is an effective model that performs well in tasks involving the conversion of one sequence into another. It does this by utilising parallel processing and the attention mechanism to increase accuracy and speed. Its design has served as the foundation for numerous cutting-edge NLP models, including Google's T5 and OpenAI's GPT, proving its adaptability and efficacy in a variety of contexts.

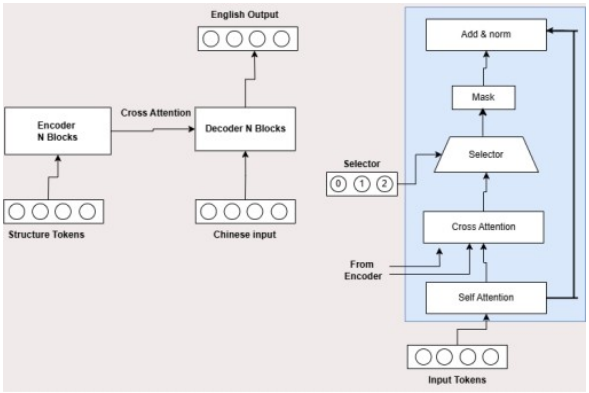

Fig. 1: Architecture Diagram

C. Data Set Used

A training set and a test set are the two halves of the dataset. There are one million sentence pairings in the training set and one thousand in the test set. The machine translation model is trained using the training set, and its performance is assessed using the test set.

The dataset's sentences are pre-processed to eliminate extraneous characters and symbols before being encoded in Unicode format. Tokenization is the process of breaking the sentences up into separate words or phrases and giving each one a unique ID. The words in the input and output sequences of the machine translation model are represented by the IDs.

Link:https://www.kaggle.com/datasets/qianhuan/translation

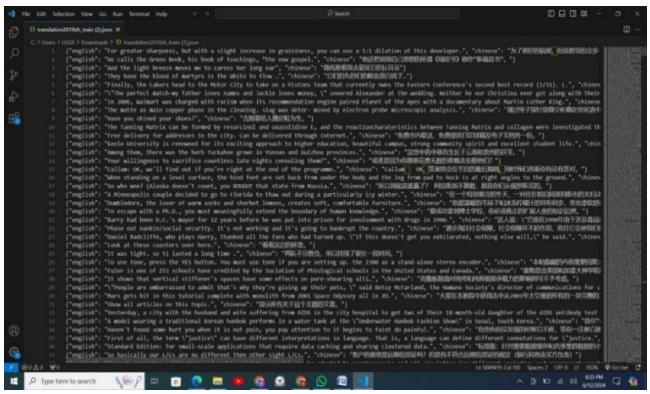

Fig. 2: Data Set Example

D. Technique used

Self-Attention Mechanism: The multi-head self-attention mechanism, which calculates relationships between every token in a sequence simultaneously, is the central component of the Transformer. Because of this, the model can concentrate on pertinent words regardless of where they are in the text, which makes it very useful for translation tasks where context is crucial.

Structure of the Encoder-Decoder: The encoder in the model is designed to handle the Mandarin input, while the decoder produces the English output. The input sequence is transformed into context-aware representations by the encoder, which the decoder then combines with previously produced output to predict the next word.

Positional Encoding: Word order is not taken into account by self-attention by itself, so positional encodings are added to the token embeddings to help the model comprehend the structure of the sequence.

Layers that feed forward: Following attention, the attended representations are processed by a feedforward neural network in each layer of the encoder and decoder, honing them before sending them to the subsequent layer.

Personalised Learning Rate Calendar: A Custom Schedule is used to regulate the learning rate. It does this by modifying the rate according to the number of warm-up steps and the model's dimensionality (d_model). This facilitates a more effective convergence of the model during training.

Functions of Accuracy and Loss: To stop the network from learning inaccurate predictions for padding tokens, a masked loss function is only used to compute loss for non-padding tokens.

Masked accuracy makes sure that padded values are ignored and that the accuracy computation only considers real tokens.

E. Pseudocode: Chinese-to-English Transformer Model

Step 1: Import Libraries

Import TensorFlow, libraries for data handling, file management (e.g., os, urllib).

Step 2: Download and Extract Data

Download dataset (Chinese-English pairs), check file type (ZIP/TAR), and extract to a target directory.

Step 3: Preprocess Dataset

Tokenize Chinese/English text, add <SOS>, <EOS>, <PAD>.

Convert tokens to numerical IDs, pad sequences to uniform lengths.

Step 4: Positional Encoding

Define positional encoding function using sine and cosine functions to represent token positions.

Step 5: Attention Mechanism

Implement scaled dot-product attention to calculate token relationships using softmax for weights.

Step 6: Transformer Encoder

Apply multi-head attention and feed-forward layers with residual connections and normalization.

Step 7: Transformer Decoder

Process encoder output and previous tokens with self-attention and encoder-decoder attention.

Add feed-forward layers, residuals, and normalization.

Step 8: Build Transformer Model

Stack encoder/decoder layers.

Add a final linear layer with softmax for token prediction.

Step 9: Training Setup

Use cross-entropy loss (ignore <PAD>), Adam optimizer, and learning rate scheduler.

Initialize weights.

Step 10: Train Model

For each epoch and batch, pass inputs through encoder and decoder, compute loss, and update weights.

Step 11: Evaluate Model

Translate Chinese inputs with encoder and autoregressive decoding in the decoder.

Use BLEU or accuracy for evaluation.

Step 12: Save Model

Save model weights, architecture, and checkpoints for reuse or fine-tuning.

Step 13: Post-Training Evaluation

Evaluate test data with BLEU score or similar metrics.

Optionally visualize attention maps for insight.

Output: High-quality English translations with performance metrics.

IV. Results

A. Metrics Used

Cosine similarity: The similarity between two vectors in an inner product space is measured by cosine similarity. It determines whether two vectors are pointing in about the same direction and is calculated by taking the cosine of the angle between them. In text analysis, it is frequently used to gauge document similarity.

Thousands of attributes, each documenting the frequency of a specific word (such a keyword) or phrase in the document, might be used to represent a document. As a result, every document is an object that is represented by a term-frequency vector. For instance, Table 2.5 shows that the word "team" appears five times in Document1, yet "hockey" appears three times. A count value of 0 indicates that the word "coach" is not present throughout the document.

$$SC(x, y) = x.\ y\ /\ ||x|| \times\times ||y||, \quad (1)$$

where the product of the vectors "x" and "y" is x. The length (magnitude) of the two vectors "x" and "y" is equal to ||x|| and ||y||. The regular product of the two vectors "x" and "y" is ||x|| × × ||y||.

Rouge Metrics: Recall serves as the primary basis for the ROUGE ratings, which were really created with text summary in mind, where the model-generated text is typically shorter than the reference text. In essence, ROUGE contrasts the reference and candidate summaries in terms of n-grams, word pairings, and word sequences.

Important ROUGE Measures

1. The n-gram overlap between the reference text and the generated text is measured by ROUGE-N.
2. To capture structural similarity, ROUGE-L uses the longest common subsequences (LCS).
3. Contiguous matches that weigh more than other n-grams are weighed by ROUGE-W.

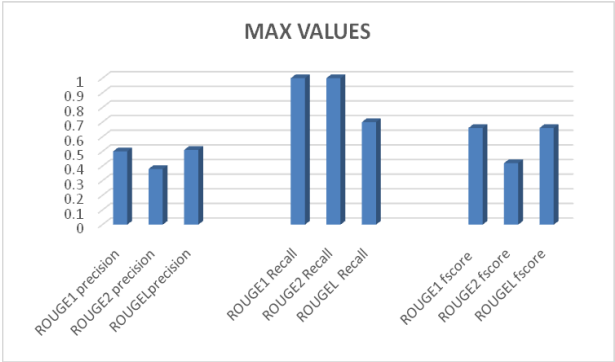

Fig. 3: Maximum Values of Precision, Recall and F score.

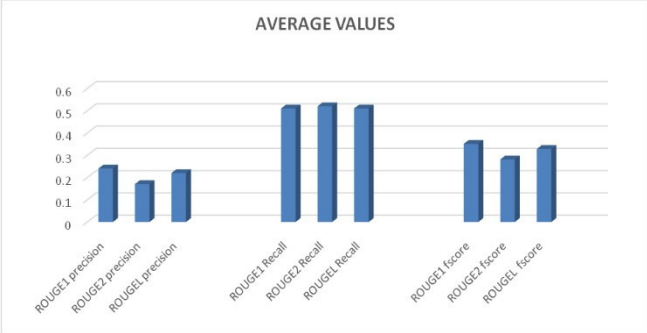

Fig. 4: Average Values of Precision, Recall and F score.

4. ROUGE-S is a measure of skip-bigram overlap, which considers two words that might not be next to each other.

ROUGE-N = {Number of matching n-grams} \ {Total n-grams in the reference}. (2)

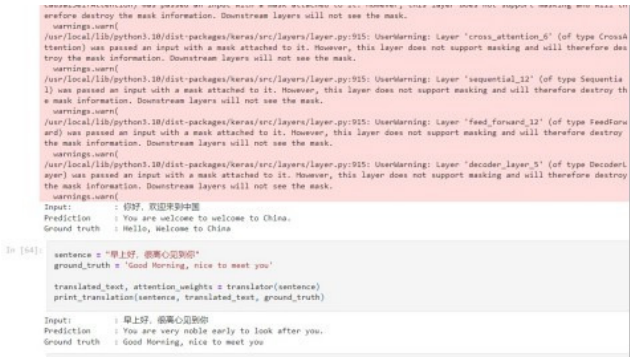

Fig. 5: Screenshot of executed Translations.

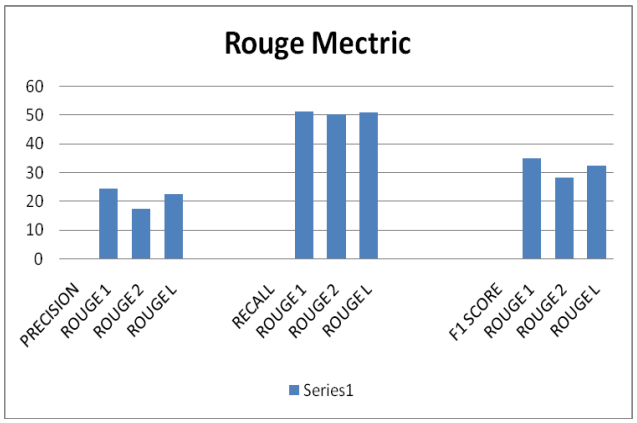

Fig. 6: Rouge metric Values of precision, Recall and F Score

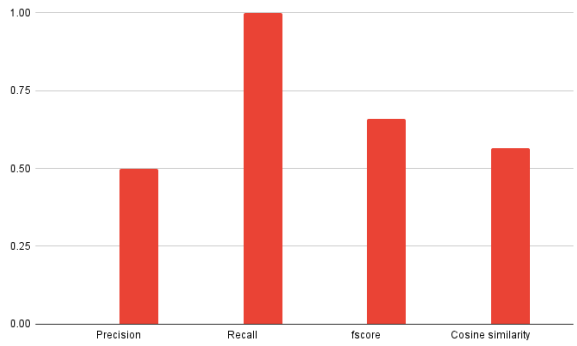

Fig. 7: Rouge metric Values along with Cosine Similarity.

V. Future Work

Model Optimisation and Compression: Given the computational complexity of transformer models, one potential direction for future research is to optimise the model for reduced resource usage. Model pruning, quantization, and knowledge distillation are a few methods that can be used to minimise the computing burden and size of the model without compromising accuracy. This is particularly crucial if the model is meant to be used with low-resource devices, like cell phones or edge devices. By reducing the environmental impact of operating large-scale transformers, optimising the model would also address concerns about energy use.

Table 1: Average Rouge metric values

ROUGE Metric	Precision	Recall	F Score
ROUGE-1	0.24	0.51	0.367
ROUGE-2	0.17	0.52	0.323
ROUGE-L	0.22	0.51	0.353

Including Attention Visualisation: Including attention visualisation tools would be a worthwhile and fascinating addition. This would make it easier for consumers or researchers to comprehend the translational focus of the model. Debugging and optimising translation outcomes, particularly in instances where errors arise, may be facilitated by visualising the attention weights between source and target tokens. Researchers can further refine the model to increase its accuracy by identifying regions where the model may be focused improperly.

Including Contextual Understanding: Existing transformer models frequently handle sentences as stand-alone entities without taking larger documents or conversations into account. In the future, the model could incorporate a context-aware translation process that considers earlier sentences when translating a given sentence. This would be very helpful when translating dialogue or longer texts where the meaning is spread out over several sentences.

Real-time Translation and Deployment: Further development of the model to facilitate real-time translation, which would enable speech-to-text or chat translation services, may also be part of future work. Low-latency translations are critical in real-time scenarios, and the algorithm might be modified to function well in such a circumstance. The model's use cases might be increased and made more widely available by integrating it into chat apps or deploying it as a service via APIs

VI. Conclusion

In conclusion, the creation of Transformer-based Chinese-to-English translator shows how effective contemporary Natural Language Processing (NLP) methods are in solving challenging linguistic problems. We achieved a considerable improvement in translation quality and accuracy by utilising the self-attention mechanism of the Transformer design. This allowed us to properly capture the subtle differences and contextual dependencies between the two languages. We ensured that the model had access to a rich and diverse dataset for learning, which led to more natural and contextu-

ally appropriate translations. The method used is based on a solid data curation and preprocessing pipeline.

The model's fine-tuning, together with the help of a dynamic learning rate scheduler and assessment techniques like cosine similarity scores, made sure that the translations retained both high accuracy and fluency in the original language. This was crucial since typical machine learning models frequently falter when dealing with non-literal translations and long-range relationships. The solution outperformed previous machine learning models like RNNs and LSTMs thanks to the Transformer model's architecture, which can process tokens in parallel and attend to multiple sections of a sentence at the same time.

In the future, we hope to further optimise the model by adding strategies like quantization and model compression to lessen computing burden and enhance real-time translation capabilities. Adding other languages to the system is a top goal as it will establish the model as a flexible instrument for multilingual translation. Pre-trained models like BERT or GPT could be incorporated to reduce training time and increase translation accuracy.

This study emphasises the broader implications of AI in bridging linguistic barriers in addition to the technical improvements in NLP. The goal is to promote greater cross-cultural communication and understanding through the increased accessibility of high-quality translation tools, thereby advancing international cooperation and respect. We are hopeful about the model's future as it develops and adjusts to feedback from the real world.

References

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin, arXiv:1706.03762 [cs.CL](2017)

[2] Chu, Y.J. (2020) On English Translation of Chinese Original Picture Books from the Perspective of Multimodality. Open Access Library Journal, 7: e6208.

[3] Jonathan Mallinson, Rico Sennrich, and Mirella Lapata. 2017. Paraphrasing Revisited with Neural Machine Translation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 881–893, Valencia, Spain. Association for Computational Linguistics.

[4] JacobDevlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova,arXiv:1810.04805 [cs.CL], (2019)

[5] Yinghui Li, Haojing Huang, Shirong Ma, Yong Jiang, Yangning Li, Feng Zhou, Hai-Tao Zheng, Qingyu Zhou, arXiv:2307.09007 [cs.CL], (2023)

[6] Amenador, Kate & Wang, Zhiwei. (2020). Analysis of the Chinese-English Translation of Public Signs: A Functional Theory Perspective. International Journal of Linguistics, Literature and Translation. 3. 176-188. 10.32996/ijllt.2020.3.7.20.

[7] Palmer, M., Wu, Z. Verb semantics for English-Chinese translation. *Mach Translat*10, 59–92 (1995). https://doi.org/10.1007/BF00997232

[8] Ansar, Wazib & Goswami, Saptarsi & Chakrabarti, Amlan. (2024). A Survey on Transformers in NLP with Focus on Efficiency. 10.48550/arXiv.2406.16893.

[9] Rahali, A.; Akhloufi, M.A. End-to-End Transformer-Based Models in Textual-Based NLP. AI 2023, 4, 54–110. https://doi.org/ 10.3390/ai4010004

[10] Yang L and Zhou G (2022) A semantic similarity analysis of multiple English translations of The Analects: Based on a natural language processing algorithm. Front. Psychol. 13:992890. doi: 10.3389/fpsyg.2022.992890

[11] Tao, Jianhua & Zheng, Fang & Li, Aijun & Li, Ya. (2009). Advances in Chinese Natural Language Processing and Language resources. 10.1109/ICSDA.2009.5278384.

[12] Haoxiang Shi, Cen Wang, and Tetsuya Sakai. 2020. A Siamese CNN Architecture for Learning Chinese Sentence Similarity. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 24–29, Suzhou, China. Association for Computational Linguistics.

[13] Bei Li, Ziyang Wang, Hui Liu, Quan Du, Tong Xiao, Chunliang Zhang, Jingbo Zhu, arXiv:2012.13866 [cs.CL],(2020)

[14] Xiong, Wen & Jin, Yaohong. (2011). A new Chinese-English machine translation method based on rule for claims sentence of Chinese patent. 378-381. 10.1109/NLPKE.2011.6138228.

[15] Li, Jason & Ng, Young & Wu, Ruixue. (2022). Strategies and problems in geotourism interpretation: A comprehensive literature review of an interdisciplinary chinese to english translation. International Journal of Geoheritage and Parks. 10. 10.1016/j.ijgeop.2022.02.001.

[16] Xiang'e Zhang, 2021. A Study of Cultural Context in Chinese-English Translation. Region - Educational Research and Reviews, 3(2), pp.11-14.

[17] Chen, Jiangping. (2006). A lexical knowledge base approach for English–Chinese cross-language information retrieval. JASIST. 57. 233-243. 10.1002/asi.20273.

[18] Khurana, D., Koli, A., Khatter, K. *et al.* Natural language processing: state of the art, current trends and challenges. *Multimed Tools Appl***82**, 3713–3744 (2023). https://doi.org/10.1007/s11042-022-13428-4

[19] Gillioz, Anthony & Casas, Jacky & Mugellini, Elena & Abou Khaled, Omar. (2020). Overview of the Transformer-based Models for NLP Tasks. 179-183. 10.15439/2020F20.

[20] Wen, Y., van Heuven, W.J.B. Chinese translation norms for 1,429 English words. *Behav Res* **49**, 1006–1019 (2017). https://doi.org/10.3758/s13428-016-0761-x