Are these sentences really that difficlut? Translation of Chinese RPASS

Chudan Huang
huangchd6@mail2.sysu.edu.cn

Shurui Liu
liushr29@mail2.sysu.edu.cn

Pengfei Shen
shenpf3@mail2.sysu.edu.cn

Ying Shen*
shen76@mail.sysu.edu.cn

Abstract

The paper tackles the complexities of translating Repetition Polysemous Ambiguous Sentences (RPAS) in Chinese through the introduction of the RRR dataset and the assessment of various neural network models, including RNNs, Transformers, and LLMs. It highlights innovations like the Weighted Transformer and TTT layers, designed to boost translation accuracy and efficiency. Experimental results on the RRR dataset show significant improvements, especially with tailored prompts for the GPT-4o model.ur code was located in https://github.com/HowCCB/RAPS.

1. Introduction

With the development of large language models, the field of machine translation has achieved very fruitful results, most Chinese is expected to be well translated into other languages, but there is a special class of Chinese sentences, most of the current machine translation algorithms are not competent. For example, the following sentences. O

“我一把把车把把住了。” (1)

“用毒毒毒蛇毒蛇会不会被毒毒死。” (2)

“一行行，行行行。” (3)

“奏一乐乐一乐。” (4)

These sentences have the following characteristics.

Repetition Words This repetition is not just a simple repetition of syllables, but the multiple occurrences of words in a sentence to construct a specific expression.In the sentence(3), the word “行” appears many times.

Polysemous word Although words are repetitive, the context and meaning may be different each time they appear. In the sentence(4), the word “乐” can mean happiness or music or like as a verb.

Ambiguous meaning These sentences are generally vague and difficult in punctuation.

We named these sentences as **R**epetition **P**olysemous **A**mbiguous **S**entences. In our paper, we try to solve this challenging **RPAS** translation problem.

The main contributions of this article are as follows.

1. The polysemy dataset is automatically augmented to expand a large number of sentences containing polysemous words, which we call the **RRR dataset**, which can be used to train translation models.

2. We use RNNs, Transformer-based models, and LLMs to fit RPASS sentences, and evaluated in the **RRR dataset** we proposed in this paper.

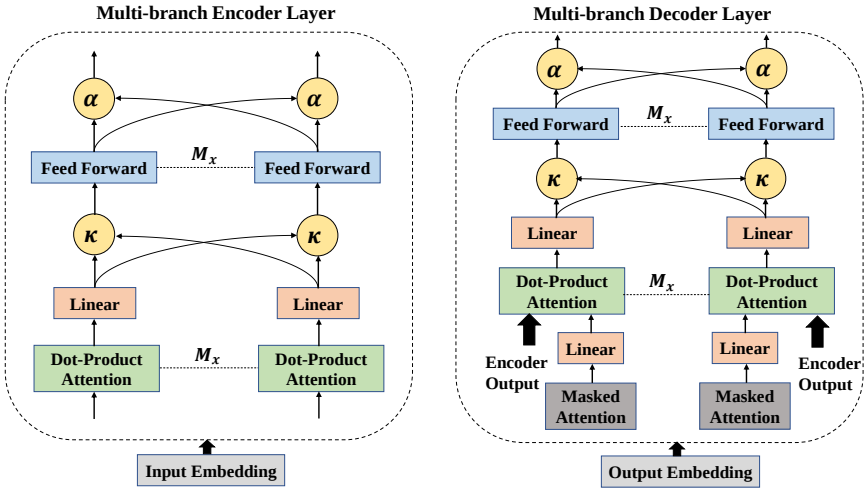

图 1. **The architecture of Weighted Transformer.**

3. We design prompts suitable for different LLMs to optimize the translation performance of RPAS sentences.

2. Related Works

2.1. RNN

The current reality of Recurrent Neural Networks (RNNs) presents a double-edged sword. On one hand, the main advantage of RNNs over Transformers lies in their linear (as opposed to quadratic) complexity. On the other hand, their inherent auto-regressive nature—requiring previous hidden states to be computed before the current time step—limits their potential to leverage parallelization, posing a significant constraint. Even when faced with sufficiently long contexts, existing RNN architectures struggle to effectively utilize the additional contextual information, highlighting their limitations in fully exploiting the available data.

2.2. LLMs

Large Language Models (LLMs) such as GPT-4 have demonstrated strong translation capabilities. However, there remains a need for research on enhancing the in-context learning abilities of existing open-source models through fine-tuning for real-time adaptive machine translation (MT), and comparing these enhancements to current approaches. These models can be fine-tuned to perform better in in-context learning scenarios, using specialized prompt templates that incorporate domain-specific sentences, phrases, or terminology. This direction holds promise for improving both translation quality and efficiency.

3. Dataset

In the construction of the original dataset, there is a significant limitation in that its content is primarily focused on everyday language, comprising a total of 20,403 records. While this size is sufficient to support preliminary training and evaluation of language models, the homogeneity of the data restricts its generalization capability and the breadth of its application scenarios. We add several key linguistic elements. The dataset collection process includes automated and manual checks to ensure translation quality.

In addition to the original data, the RRR dataset we constructed was also enriched with 5,000 training texts containing 15 polysemy words.

3.1. Idioms

We have collected 4310 unique Chinese idioms and their corresponding English translations from A dictio-

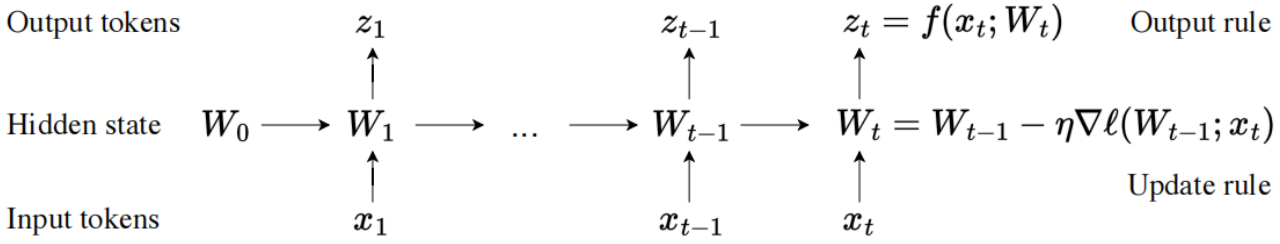

图 2. **Architecture of Test-Time Training (TTT) process.**A generic sequence modeling layer expressed as a hidden state that transitions according to an update rule. All sequence modeling layers can be viewed as different instantiations of three components in this figure: the initial state, update rule and output rule.

nary of Chinese idioms with English translation.

3.2. Polyphonic characters

We carefully collected 200 sentences from the Internet that specifically involve the complex linguistic phenomenon of polyphonic characters. Then, with the help of the advanced machine translation tool DeepL and manual calibration, we produced high-quality translations of these sentences.

4. Methodology

4.1. Weighted Transformer

The traditional Transformer network uses multi-head attention mechanisms which processes the input sequence through multiple independent attention heads, concatenates the outputs of these heads, and then applies a linear transformation to get the final attention representation.

$$\text{headi} = \text{Attention}(QWQ_i, KWK_i, VWV_i) \quad (1)$$
$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{headi})W^O \quad (2)$$

The Weighted Transformer Network [1] introduces branch self-attention layers to replace the multi-head attention mechanism in the original Transformer. And the contribution of each branch is learned during the training process.

Linear transformation and weighting are shown in Equation 3.

$$\overline{\text{headi}} = \text{headi}W^{Oi} \times \kappa_i \quad (3)$$

Weighted sum is shown in Equation 4.

$$\text{BranchedAttention}(Q, K, V) = \sum_{i=1}^{M} \alpha_i \text{FFN}(\overline{\text{headi}}) \quad (4)$$

where κ_i are the learned branch weights and α_i are the learned weighting parameters, ensuring $\sum \kappa_i = 1$ and $\sum \alpha_i = 1$.

By learning the κ_i and α_i parameters, the model can adaptively adjust the importance of each branch, enhancing the expressive power of the attention mechanism and the training efficiency of the model. The network architecture is shown in Figure 1.

4.2. Test-Time Training (TTT) layers

TTT layer [2] proposes a new class of sequence modeling layers that combine the linear complexity of RNNs with a more expressive hidden state. In a TTT layer, the hidden state is itself a machine learning model, and the update rule is a step of self-supervised learning. This means that the hidden state is updated by training even on test sequences, hence the name Test-Time Training (TTT). Instantiations of TTT Layers include two TTT-Linear and TTT-MLP.

As shown in Figure 2, all sequence modeling layers can be expressed as a hidden state that transitions according to an update rule. The key idea is to make the hidden state itself a model f with weights W, and the update rule a gradient step on the self-supervised loss ℓ. Therefore, updating the hidden state on a test

sequence is equivalent to training the model f at test time. This process, known as test-time training (TTT), is programmed into TTT layers.

4.3. Fine-tuning of LLMs

ChatGLM-6B is an open-source bilingual conversational language model supporting both Chinese and English, based on the General Language Model (GLM) architecture, with 6.2 billion parameters. Utilizing model quantization techniques, users can deploy it locally on consumer-grade graphics cards (requiring as little as 6GB of VRAM at the INT4 quantization level). ChatGLM-6B employs similar techniques to ChatGPT and is optimized for Chinese Q&A and dialogue. With training on approximately 1 trillion bilingual tokens, enhanced by supervised fine-tuning, self-help feedback, and reinforcement learning from human feedback, the 6.2 billion parameter ChatGLM-6B is capable of generating responses that align well with human preferences.

Existing pretrained models mainly fall into three categories: autoregressive models (e.g., GPT), autoencoding models (e.g., BERT), and encoder-decoder models (e.g., T5). These models perform differently across various tasks, with no single framework excelling in all tasks. The GLM architecture used in this model improves blank-filling pretraining by incorporating 2D position encodings and allowing for arbitrary order prediction spans. GLM pretrains different types of tasks by varying the number and length of blanks (as shown in the Figure 3). Across a wide range of tasks, GLM outperforms BERT, T5, and GPT under the same model size and data conditions, achieving the best performance in a single pretrained model and demonstrating its versatility across different downstream tasks.

ChatGLM-6B has been specifically optimized for Chinese and English, pretrained on approximately 1 trillion Chinese and English tokens, and incorporates techniques such as supervised fine-tuning, self-help feedback, and reinforcement learning from human feedback. This enables it to perform exceptionally well in handling Chinese-English translation tasks.

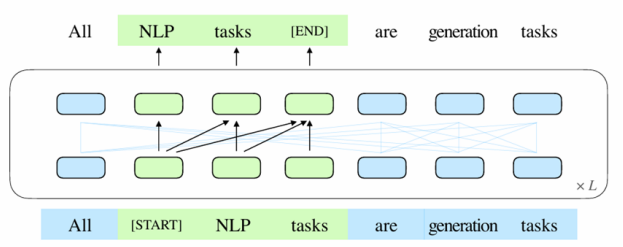

图 3. **Illustration of GLM. GLM blank out text spans (green part) and generate them autoregressively.**

We have conducted a certain degree of fine-tuning on ChatGLM-6B to specialize it further in the field of Chinese-English translation, and found that it exhibits excellent translation performance.

4.4. Prompt Engineering

We designed a series of prompts for different LLMs, evaluated them on the RRR dataset, and finally gave suggestions for the design of prompts for RPAS sentences.

5. Experiments

5.1. Comparison on RRR

We trained the RNN network and Weighted-Transformer network on the RRR dataset, fine-tuned ChatGLM, fine-tuned TTT, and used prompt engineering for prompting. At the same time, some very challenging sentences such as " 一行行，行行行" were used to test, We write the results of the test to whether it was successful or not in the RPAS column, and the results were like table 1 shown.

5.2. Prompt Design

We use GPT-4o as an example to design a prompt that is best suited for translating RPAS sentences.We design prompts from four perspectives: role, task, tone, and examples. It was designed as Table 2 shown.

Method	Accuracy(%)	RPAS Test
RNN-based	40.24	✗
Self-Attention	90.39	✗
Weighted Transformer	96.20	✗
Fine-tuned Chat-GLMs	94.66	✓
Fine-tuned TTT	96.42	✓
Prompt Engineering	**99.25**	✓

表 1. **Comparison of experimental results of different methods.**Fine-tuned Chat-GLM partially passed the RPAS test, Fine-tuned TTT and Prompt Engineering passed all the RPAS test.

ChatGPT-4o	**Role**	"You are a Chinese language worker, and you are very familiar with the expression and sentence structure in Chinese, and you are also very familiar with English translation."
	Task	"Now please translate into English the passage I have given you, please note that there are many polysemous words in this passage that you have been given, and that these polysemous words will be repeated"
	Tone	"Please translate in as plain a tone as possible"
	Example	"Now to give you an example, the 'row' in 'row row, row row row' can represent both sufficient ability and industry. So it needs to be translated as 'if you are prominent in one area, then you can easily do work in other areas'"

表 2. **Prompt designed for GPT-4o.**We designed the prompt from 4 view: Role, Task, Tone and Example

6. Conclusion

The paper tackles the complexities of translating Repetition Polysemous Ambiguous Sentences (RPAS) in Chinese through the introduction of the RRR dataset and the assessment of various neural network models, including RNNs, Transformers, and LLMs. It highlights innovations like the Weighted Transformer and TTT layers, designed to boost translation accuracy and efficiency. Experimental results on the RRR dataset show significant improvements, especially with tailored prompts for the GPT-4o model.

参考文献

[1] Karim Ahmed, Nitish Shirish Keskar, and Richard Socher. Weighted transformer network for machine translation. *ArXiv*, abs/1711.02132, 2017. 3

[2] Yu Sun, Xinhao Li, Karan Dalal, Jiarui Xu, Arjun Vikram, Genghan Zhang, Yann Dubois, Xinlei Chen, Xiaolong Wang, Sanmi Koyejo, Tatsunori Hashimoto, and Carlos Guestrin. Learning to (learn at test time): Rnns with expressive hidden states. 2024. 3

Appendix

A. Training Process

During the training process of the Weighted Transformer, we conducted a series of experiments and iterations to determine the optimal parameter configuration as follows:

Parameter	Value
Batch size	64
Buffer size	1000
Epochs	50
Num layers	4
dff	1024
Dropout rate	0.3
Num heads	8

表 3. **Model Training Parameters.**

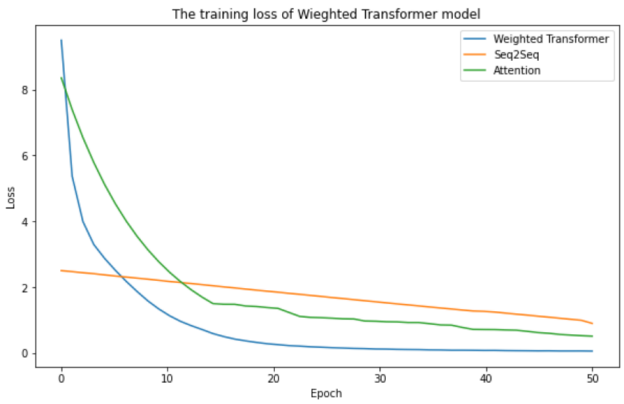

图 4. **The training loss of Weighted Transformer and two midterm models.**

We compare Weighted Transforemr with the two models from the midterm assignment. In our midterm assignment, we have found the optimal parameters of two models. Compared with Weighted Transformer, they require more epochs and more time to reach their optimal accuracy and loss values. Figure 4 shows a comparison of the loss values when they were also trained to 50 epochs. During this period, the two models operating during this period have not yet converged to the optimal value.

B. Simple User Interaction Design

In the translation system based on the ChatGLM model, we have developed three modes in the user interface of the ChatGLM-6B model: Chinese-to-English translation, English-to-Chinese translation, and free chat mode(as shown in the Figure 5). Users only need to enter the number corresponding to the different modes to switch between them.For each mode, we designed a set of prompt instructions to produce results that better meet the requirements in different system modes.

In the Chinese-to-English and English-to-Chinese translation modes,

Name	No.	Contribution
黄楚丹	21312024	Dataset & Weighted Transformer & Writing
刘书睿	21312295	TTT & Prompt Engineering & Writing
沈鹏飞	21312188	Chat-GLM & Simple User Interaction Design & Writing

表 4. **Labor Contribution.**

users only need to input the sentence they want to translate to get a concise and accurate translation. Additionally, in the translation modes, the system will automatically clear the chat history to avoid interfering with subsequent translation functions. In the free chat mode, the ChatGLM system works consistently with the original model's chat system. Users can ask ChatGLM questions according to their needs, and ChatGLM will provide corresponding answers. In this mode, the model will retain memory of the previous chat content to better meet the user's needs and preferences.

图 5. **Translation system interface display.**

C. Labor Contribution

Our labor contribution are shown on table 4.