Analysis of Transformer Performance in Chinese-to-English Translation Tasks

Haotian Xu
hax030@ucsd.edu

1 Introduction

The Transformer architecture introduced by (Vaswani et al., 2017) has brought significant improvements in the Natural Language Processing (NLP) field and serves as the foundation of many current language models. Machine translation, as one of the traditional but still evolving tasks in NLP systems, has also seen notable advancements, leading to more accurate and fluent translations between languages. However, the performance of machine translation systems can be influenced by the language models used, the specific language pairs, and various other factors.

In this analysis, we specifically focus on Simplified Chinese-to-English translation tasks. We evaluate a subset of the WMT19 zh-en dataset (Foundation, 2019) using the pretrained transformer model from (Tiedemann and Thottingal, 2020). We use SacreBLEU, introduced by (Post, 2018), as the evaluation metric for the generated translations. In the latter part of this report, we will refer to this metric simply as BLEU. We conduct a detailed analysis of the behavior of the translation transformer model through a customized project on the ZenoML platform (Cabrera et al., 2023), where we can slice the data into categorized subsets and observe the model's performance.

Our analysis reveals various translation errors, such as name and term inaccuracies, missing information, overly summarized translations, and issues stemming from the dataset quality. These errors highlight the limitations of the current model, the impact of noisy data, and the complexities of the Chinese language. We suggest that employing larger models, cleaner datasets, specialized tokenizers, and fine-tuning techniques could significantly enhance translation accuracy.

2 Dataset

The dataset we selected is the zh-en subset from the WMT19 translation dataset (Foundation, 2019). The original dataset contains approximately 26 million rows, which is computationally expensive to process, even for model inference. Due to our computational limitations and time constraints, we carefully processed the data and selected a subset with 450 rows for our analysis. The subset is arranged into 15 topics, each containing 30 samples. Each row or sample includes one or several sentences with a source entry in Chinese and a target entry in English. Below is the procedure we followed to process and sample the data.

2.1 Topic Modeling

We first chose the validation split of the zh-en translation dataset as our base dataset, which contains 3,981 rows. We used the topic modeling technique from (Grootendorst, 2022) to create topic clusters with a minimum size of 30 rows per topic. The topics were generated using the translation labels in English. The results are shown in Table 1. The topic number of -1 indicates outliers, while all others are valid topics with some representative words. We added a column to the dataset to map each row to its corresponding topic.

2.2 Sampling

Even though the validation subset is much smaller than the entire dataset, generating translations for these data is still costly on our machine. Therefore, we further sampled the dataset by randomly selecting 30 samples for each valid topic. Since we have 15 valid topics from the previous topic modeling, we obtained a subset of 450 rows. We show some example input/output pairs in Table 2.

Topic	Count	Name
-1	858	-1_frigates_statement_very_security
0	2295	0_cpc_congress_committee_cooperation
1	105	1_trump_donald_presidential_president
2	97	2_olympics_olympic_medals_pyeongchang
3	82	3_reconciliation_said_she_says
4	69	4_riodoce_riodoces_valdez_valdezs
5	63	5_neymar_neymars_uefa_madrid
6	59	6_activism_lynchings_black_ferrells
7	58	7_jokowi_paniai_indonesia_papua
8	56	8_explosion_exploded_incident_eruption
9	55	9_actors_imdb_dramas_film
10	46	10_fossil_dinosaur_dinosaurs_cretaceous
11	39	11_prophylaxis_hiv_aids_prevention
12	37	12_wounded_attackers_attacked_soldiers
13	32	13_colin_paul_lutis_brothers
14	30	14_bedroom_apartment_onebedroom_room

Table 1: Topic Modeling Results

3 Analysis Approach

We perform our analysis by first loading the pretrained model and using it to generate translations for each sample in the dataset. Then, we calculate the BLEU score between the generated translations and the reference labels. We also measure the length of each generated translation. Finally, we upload our project and data to ZenoML (Cabrera et al., 2023) to evaluate the behavior of the pretrained translation model.

3.1 Translation Generation

Our selected pretrained model is a transformer model trained on Chinese-to-English translations from (Tiedemann and Thottingal, 2020). The model is trained on the OPUS dataset, which is different from the WMT19 dataset (Foundation, 2019) that we sampled. We load the model and use it to generate translations for our dataset. The example translations are shown in Table 3.

3.2 Evaluation Metrics

We use SacreBLEU (Post, 2018) to calculate the BLEU scores between the generated translations and the reference labels. This package provides standard tools for BLEU metrics, avoiding potential variations in scores due to different parameters and settings. Additionally, this package can directly use raw texts as inputs, facilitating our process by preventing us from generating tokens. We also measure the text length (in characters) of the generated outputs, as we hypothesize that text length may influence translation accuracy. Table 3 displays these results.

3.3 Behavioral Evaluation

We use ZenoML (Cabrera et al., 2023) for the behavioral evaluation of the model. This platform has many public projects that analyze model behavior on different datasets and metrics. However, we did not find one that fits our purposes, so we uploaded our dataset and results and created a customized project. The ZenoML platform allows us to easily slice data and observe how the model performs on these slices. We created slices for each topic and different tiers of performance based on the BLEU score and output length. This allowed us to see the performance on different topics and analyze commonalities in low-performance slices.

4 Errors and their Categorization

In this section, we will examine the translation model's behaviors and categorize the potential causes of low performance.

5 Errors and their Categorization

In this section, we will examine the translation model's behavior and categorize the potential causes of low performance.

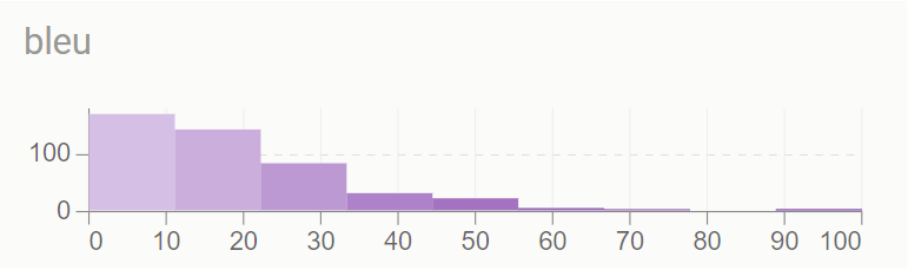

Figure 1: Distribution of the BLEU Score

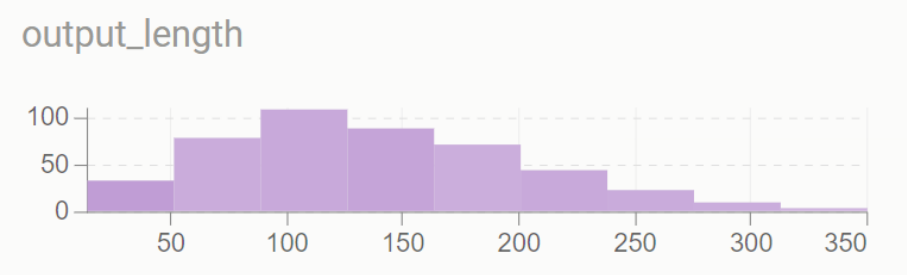

Figure 2: Distribution of the Output Length

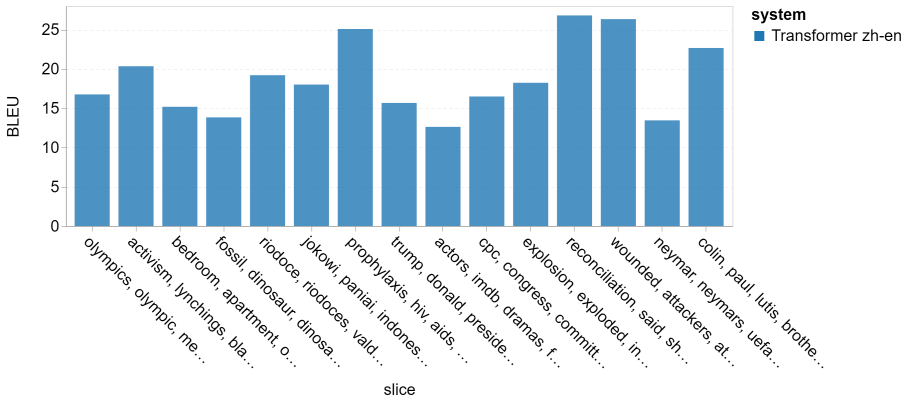

Figure 3: Topics vs. BLEU

id	Input (Chinese)	Output (English)	Topic
35	这部由布莱恩·盖特伍德和亚历山德罗·田中创作的喜剧，被导演里斯·托马斯推上了电视屏幕：这是一部难兄难弟喜剧片，剧情背景设置在80年代的罗马尼亚，讲述两名布加勒斯特官员为了真理、无神论和马克思列宁主义路线而奋斗的故事。	The comedy, created by Brian Gatewood and Alessandro Tanaka, with Rhys Thomas directing, purports to be found TV: a Romanian 1980s buddy-cop drama about two Bucharest officers fighting for truth, atheism and the Marxist-Leninist way.	actors, imdb, dramas, film, awards, emmy, audience, directing, drama, actress
182	有关鸟类起源问题，国际学术界已争论了100多年，但以往这些争论都缺少化石证据。	International academic circles have disputed issues pertaining to the origin of birds for more than a century. However, such past disputes lack fossil evidence.	fossil, dinosaur, dinosaurs, cretaceous, specimens, preserved, museum, specimen, preservation, creature
276	为确保里约奥运会的安全，巴西将启动有史以来最大规模的安保联动计划。	In order to ensure the safety of the Rio Olympics, Brazil will start the largest ever security linkage plan.	olympics, olympic, medals, pyeongchang, medal, korea, winter, athletes, athletics, competition

Table 2: Example Input/Output Pairs for Chinese-to-English Translation

id	Label	Prediction	Output Length	BLEU Score
272	”We’re putting this enormous load on them for three or four days and it’s very challenging,” Mora said.	”In three or four days, we will give them a huge amount of training, which is very challenging.” Mora said.	107	15.98
316	People with flu should stay home, and not go to work or school for a week.	Influenza patients should stay at home without going to work or school for a week.	82	43.76

Table 3: Example Translation Results with BLEU Score

5.1 Overall Performance

The mean BLEU score is 18.70, which is much lower than the score of 36.1 in the model’s original test (Tiedemann and Thottingal, 2020). This gap is potentially due to the different patterns in the WMT19 and OPUS datasets. The distribution of the BLEU scores is shown in Figure 1. From the figure, we can see many samples receive relatively low scores, while samples with high scores are rare. We will particularly focus on these low-performance samples and analyze their errors in the following sections. The distribution of the output length is shown in Figure 2. We do not see a correlation between the output length and performance. However, these measures may help explain some extreme cases.

5.2 Topic

We will refer to each topic slice by the first word in the topic name. Figure 3 shows the BLEU scores of different topic slices. We can see the ”reconciliation” topic and ”wounded” topic have relatively good average BLEU scores. By checking through these two topic slices, we find some commonalities. They are mostly composed of common and easy words, and their text lengths are relatively short (the ”reconciliation” topic has an average output length of 64.30, and the ”wounded” class has an average output length of 117.60). The following examples show the above traits.

ID: 358
Input (Chinese): “把他从车里拖出来！”
Label (English): ”Get him out of the car!”
Output (English): ”Get him out of the car!”
BLEU: 100.00

Example 1: The ”reconciliation” Topic Example

ID: 440
Input (Chinese): 据人权观察组织称，五名年轻抗议者被杀，还有更多人受伤。
Label (English): ”According to Human Rights Watch, five young protesters were killed and many more injured.”
Output (English): ”According to HRW, five young protesters were killed and many more injured.”
BLEU: 70.14

Example 2: The ”wounded” Topic Example

On the other hand, the ”fossil” topic, ”actor” topic, and ”neymar” topic have relatively low performance. We find that these topics have many names and terms in the texts, leading to translation errors. We will delve into this factor in the next section. The following examples show the above traits.

ID: 202
Input (Chinese): 它们包括高棘龙，一种重达6吨、长38 英尺（11.5 米）的巨兽。
Label (English): They include Acrocanthosaurus, a 38ft (11.5m) long monster weighing six tonnes.
Output (English): They include high-crested dragons, a giant beast weighing 6 tons long and 38 feet long (11.5 metres).
BLEU: 6.29

Example 3: The ”fossil” Topic Example

ID: 40
Input (Chinese): 而且有众多名人配音演员- 其中包括珍妮·斯蕾特、尼克·奥弗曼、马赫莎拉哈什巴兹·阿里和科洛·塞维尼- 打造了一场精彩的声音盛会。
Label (English): And the roster of famous voice-over actors - among them Jenny Slate, Nick Offerman, Mahershala Ali and Chloë Sevigny - makes for a decent game of spot-the-voice.
Output (English): And there were a lot of celebrities – including Jenny Slater, Nick Ofman, Mahsala Hashbaz Ali and Kolo Sevini – who made a wonderful sound.
BLEU: 4.42

Example 4: The ”actor” Topic Example

5.3 Name and Term

Among the low-performance translations, we find errors in names and terms are frequent. Names could be a person’s name, location name, drama name, etc. In Example 4, the generated translation misspelled every actor’s name, leading to a low BLEU score. We provide another example of location name errors.

ID: 152
Input (Chinese): 新京报记者从新乡县委宣传部获悉，爆炸共致1人受伤。
Label (English): The Beijing News reporter learned from the Xinxiang County Party Committee Propaganda Department that one person was injured in the explosion.
Output (English): The news reporter from the New Towns Committee’s Advocacy Department learned that the explosion had injured a total of one person.
BLEU: 7.26

Example 5: The Location Name Error Example

The above example mistranslated a specific location in China from ”Xinxiang” to ”New Towns”. Specific location names are typically translated by pronunciation, but the translation model translates them by the meaning of each Chinese character. This example contains another obvious error as it translated ”Beijing News” directly into ”news”, which is a kind of missing information that we will analyze later.

Scientific terms can also be confusing for the translation model. In Example 3, the model trans-

lated the term ”Acrocanthosaurus” into ”high-crested dragons”. Another example below also shows this kind of error.

ID: 180
Input (Chinese): 最新研究发现出现在《当代生物学》杂志上的科学家们相信，结节龙死后被冲到海里，之后在泥土中形成了木乃伊。
Label (English): The scientists, whose latest findings appear in the journal Current Biology, believe Borealopelta was washed out to sea after it died and mummified in mud.
Output (English): Recent studies have found that scientists in Contemporary Biology magazine believe that nodes were washed into the sea after their death and that mummies were formed in the soil.
BLEU: 4.52

Example 6: Scientific Term Error Example

In this example, the dinosaur ”Borealopelta” was translated into ”nodes”, which is far from its original meaning. These examples show that it is difficult for the model to match scientific terms in two languages.

5.4 Missing Information

Another common error is missing information. The model may miss some information in the input Chinese and does not contain them in the output translations. The following example shows this situation.

ID: 121
Input (Chinese): 活动围绕“科技、安全、环保”主题展开，为期2天。届时，全国烟花爆竹行业中最先进的花炮机械、最安全环保的花炮原辅材料和新产品将齐聚浏阳。
Label (English): The two-day event will be focused on the theme “Technology, Safety, Environmental Protection”. Then, the most advanced fireworks and firecracker machinery, the safest and most environmental friendly fireworks and firecracker raw and auxiliary materials as well as new products in China’s fireworks and firecracker industry will gather at Liuyang.
Output (English): The event will focus on the theme “Technology, safety, and environmental protection” for two days.
BLEU: 3.39

Example 7: Missing Information Example

The error in this example is clear. The output translation only contains the first sentences and ignores the next one. This kind of error is frequent in our samples, resulting in low performance.

5.5 Overly Summarization

The error of overly summarization is similar to the missing information error but has some variations. The model sometimes summarizes the meaning of inputs excessively, potentially leading to the loss of original nuance or information. The following example clearly demonstrates this kind of error.

ID: 129
Input (Chinese): 要树立正确的事业观、权力观、地位观，树牢备战打仗意识。
Label (English): It is necessary to establish proper outlook on career, outlook on power and outlook on position and strengthen their awareness in war preparation and war fighting.
Output (English): To build the right vision of business, of power, of status, and to be war-aware.
BLEU: 1.89

Example 8: Overly Summarization Example

The output translation in this example is much more concise than the label, as it summarizes the content. It is difficult to say the translation is wrong in meaning, but it may lose the original tone. The original input in Chinese and the label in English have a sense of seriousness and formality. However, the generated translation is overly summarized and loses such a feeling.

5.6 Dataset Issues

There are some other errors that are strange and may be related to the training dataset of the model or the evaluation dataset that we selected. These datasets may not be clean and may contain some noise or incorrect labels. These noises lead to inconsistencies between the label and the translation. We illustrate such situations in the following examples.

ID: 124
Input (Chinese): 做好宣讲，让全社会形成共识，形成更好的思想基础
Label (English): Good presentation helps the whole society to reach a consensus and form a better ideological foundation.
Output (English): (c) Make a good case for a common understanding of society and a better basis for thinking.
BLEU: 5.04

Example 9: Dataset Problem Example

In this example, the "(c)" at the beginning of the translation is odd, as it is not present anywhere in the input. It is possible that the training dataset contains data in multiple-choice format and does not wrap the choice properly.

ID: 325
Input (Chinese): 所收费用均捐献给前列腺癌慈善组织Prostate Cancer UK。
Label (English): All proceeds go to Prostate Cancer UK.
Output (English): The fees collected were donated to Prostate Chancer UK, a prostate cancer charity.
BLEU: 6.25

Example 10: Dataset Problem Example

This example shows an error in the evaluation label. The generated translation correctly captures the meaning of the original input sentence. However, the label assumes there is preceding context and ignores information about fees and the introduction of Prostate Cancer UK. This error results in a low BLEU score for this translation, even though the translation is not wrong.

6 Discussion

The errors we observed in our analysis may be related to the model, the dataset, and the language features.

6.1 Model

The naive implementation of the transformer model has limited tokens and parameters. It may be difficult for the model to capture and parameterize all information, leading to missing information errors. The model sometimes summarizes the meaning of inputs excessively, potentially leading to the loss of original nuance or information. Additionally, the model struggles with translating names correctly, particularly when it comes to distinguishing between pronunciation and meaning. These errors may result from the limited size of the model, which does not have enough parameters to learn these aspects correctly.

6.2 Dataset

As discussed in Section 5.6, the quality of datasets can influence translation accuracy. Inaccurate labels, as shown in Example 10, may lead to unfair evaluations. If the model is trained on these inconsistent labels, it may produce incorrect results. The oddity in Example 9 may also be caused by unclean datasets.

6.3 Language Features

The complexity of the Chinese language, with its numerous characters and nuances, presents unique challenges for translation models. The large number of characters requires larger token sizes, making it difficult for small models to learn them effectively. Additionally, the Chinese language has no spaces between words, making it challenging for translation models to distinguish individual characters and words. Errors in names and terms may be due to this feature. The model tends to translate the meaning of individual characters and combine them, instead of recognizing these characters as a word and translating them into correct names or terms.

7 The Way Forward

To improve translation accuracy, we propose several actions.

7.1 Large Model

A larger and more complex language model may effectively improve translation performance, especially for the Chinese language, which has numerous unique characters. Larger models can tokenize and learn Chinese characters more efficiently, thereby producing better translations. Currently, there are many language models that are significantly more complex than the original transformer model. Implementing a larger model for Chinese-to-English translation tasks is plausible and could lead to substantial improvements in accuracy and fluency.

7.2 Large and Clean Dataset

A high-quality dataset with a sufficient number of samples is also crucial for improving the performance of translation models. However, achieving such a dataset can be challenging. It requires considerable labor to find and translate high-quality texts in both languages. Efforts should be made to ensure that the dataset is free of noise and inconsistencies, as these can negatively impact the model's performance. Collaborations with linguistic experts and utilizing crowdsourcing platforms could be effective strategies to build and maintain such datasets.

7.3 Chinese Tokenizer

Since the Chinese language differs significantly from Latin-based languages, developing a tokenizer specifically for Chinese is essential. A well-designed Chinese tokenizer can handle the nuances of the language, such as the lack of spaces between words and the large number of characters. Investing in a robust tokenizer that can accurately segment Chinese text will greatly enhance the preprocessing step and, consequently, the overall translation quality.

7.4 Fine-Tuning

To improve translation accuracy for scientific terms or specific names, fine-tuning the model on a domain-specific dataset can be highly beneficial. Fine-tuning allows the model to adapt to the terminology and context of a particular field, resulting in more accurate translations. For instance, creating a specialized corpus for medical or archaeological translations and using it to fine-tune the model could significantly enhance its performance in those areas. This targeted approach ensures that the model is well-versed in the specific vocabulary and nuances of different domains.

8 Conclusion

In this analysis, we examined the performance of a pretrained transformer model on Chinese-to-English translation. We identified several types of errors, including name and term errors, missing information errors, overly summarized translations, and dataset issues. These errors can be attributed to the model's limitations, the noisy dataset, and inherent language features.

To improve the accuracy of Chinese-to-English translation tasks, we propose using a larger language model, a clean dataset, and a robust Chinese tokenizer. Additionally, fine-tuning techniques can be employed to enhance translation accuracy in specific domains.

9 Additional Information

Our ZenoML project link:
https://hub.zenoml.com/project/ac445ff2-d77b-4c4e-b31c-251e437bd9ce/zh-en%20Translation

References

Cabrera, A., Fu, E., Bertucci, D., Holstein, K., Talwalkar, A., Hong, J. I., and Perer, A. (2023). Zeno: An interactive framework for behavioral evaluation of machine learning. In *CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA. Association for Computing Machinery.

Foundation, W. (2019). Acl 2019 fourth conference on machine translation (wmt19), shared task: Machine translation of news. http://www.statmt.org/wmt19/translation-task.html.

Grootendorst, M. (2022). Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Post, M. (2018). A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Tiedemann, J. and Thottingal, S. (2020). OPUS-MT — Building open translation services for the World. In *Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT)*, Lisbon, Portugal.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.