CloudSheep System for WMT24 Discourse-Level Literary Translation

Lisa Liu, Ryan Liu, Angela Tsai, Jingbo Shang
University of California, San Diego
{lil043, ryl001, cjt002, jshang}@ucsd.edu

Abstract

This paper describes the CloudSheep translation system for WMT24 Discourse-Level Literary Translation shared task. We participated in the Chinese-English direction on the unconstrained track. Our approach to the task used a pipeline of different tools in order to maximize the translation accuracy and flow of the text by combining the strengths of each tool. In particular, our focus was to translate names consistently and idioms correctly. To achieve consistent names throughout a text, a custom name dictionary was generated for each text, containing person and place names, along with their translations. A common honorific dictionary was applied for consistency with titles, especially in historical or cultivation novels. The names were found and translated with GPT 3.5-turbo. To achieve accurate and concise translations of idioms, which are often translated literally and verbosely, we integrated the CC-CEDICT library to provide official definitions. Then, we used GPT-4 to pick the best dictionary definition that fit the context and rephrase it to fit grammatically within a sentence. For the translation of non-name and non-idiom terms, we used Google Translate. We compared our approach's performance with Google Translate as a baseline using BLEU, chrF, and COMET, as well as A/B testing.

1 Introduction

Machine translation techniques customized for webnovels have been researched more during the past few years (Wang et al., 2023). With the widespread availability of commercial and open-source large language models, it has become easier to fine tune existing models for a specific kind of data. Many of the top translation solutions to last year's task approach the problem of webnovel translation from the fine tuning perspective, experimenting with combining and tuning different machine learning models to find the best method for translation (Lopez et al., 2023; An et al., 2023).

When scored by human annotators, each of last year's machine translation systems, without exception, had more errors in the categories of Accuracy and Fluency compared to the other categories of Style, Terminology, Localization, and Other (Wang et al., 2023). This may indicate that inconsistency and inaccuracy are still issues that need more attention.

With a background in reading and translating webnovels as human translators, specifically in the Chinese to English direction, we wanted to approach the machine translation problem from the human readability perspective. As a reader, one of the biggest qualities of a translation is consistency. When a character is referred to as A in one sentence and referred to as B in the next, it is very hard to follow the translation, even if the writing style and vocabulary choices are immaculate. On the other hand, even if the character is wrongly referred to as B the whole time, the consistency allows the reader to follow the translation and the events. At the time of our background research, the most up-to-date version of DeepL, a popular machine translation tool in the webnovel translation community, still had name translation inconsistencies even within the same sentence, as shown in Figure 1.

Another important aspect of a good translation from a reader's perspective is correctly translating Chinese phrases with an English equivalent that matches it in tone. As a translator, that means that we often aim to convey the figurative meaning, rather than the literal meaning. This is especially common for idioms, or "chengyu" in Chinese. These phrases often originated from ancient texts, and their meaning often comes from the myth, story, or historical event they were derived from, rather than the actual characters. Due to this, a literal translation fails to convey the meaning, and often is too formal for the modern settings where they are used as a casual part of speech. For example, the phrase "脑子进水" literally translates to

Proceedings of the Ninth Conference on Machine Translation, pages 960–966

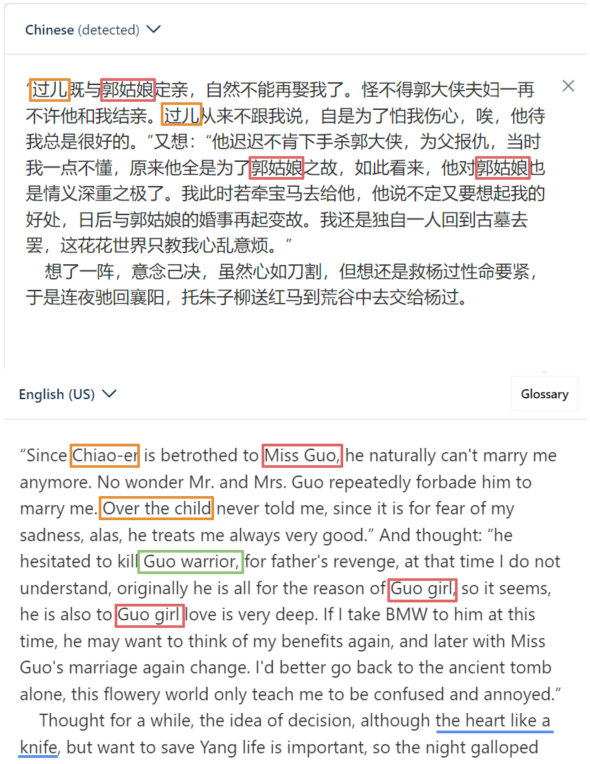

Figure 1: Name inconsistencies within DeepL translation for a single passage.

"water entered the brain," but the meaning is "lost one's mind" or "gone crazy." Making the appropriate choice between them depends on the sentence's tone and context.

With these two aspects in mind, our translation system aims to target inconsistencies in name translation and inaccuracies in idiom translation. We accomplished the former through generating a dictionary of the names found in the Chinese raws along with their English translations. We accomplished the latter through finding the figurative meaning of idioms from an open-source dictionary and using GPT-4 to rephrase the best definition to fit the sentence.

2 Data and Tools

We primarily used the GuoFeng Webnovel Corpus provided by the organizers (WMT23, 2023) (Wang et al., 2024). The data we used for self-evaluation came from the test data in last year's dataset, because of the relatively short lengths of the texts provided per novel and the reference English translation provided as well. We also looked for short excerpts of novels through publicly available translations (found through NovelUpdates) and their original Chinese texts (found through JJWXC) to test our system's ability to translate idioms and names.

We also used public blog posts to compile a dictionary of honorific translations, in order to maintain consistent translations across novels and texts. We used open-source dictionaries like CC-CEDICT to obtain the most accurate translation for idioms. Finally, we used prompt engineering and GPT models to tie together the different translation tools we used to create a comprehensive translation.

3 Translation Pipeline

3.1 Text Segmentation (Name Translation)

We wanted to find a way to reliably build a name dictionary that would get a majority of the names without incurring too much cost. The first place submission in last year's task's unconstrained task, DUTNLP (Zhao et al., 2023), used Jieba, a segmentation tool for Chinese. Text segmentation is the process of dividing text into meaningful words or phrases. Different segmentation granularities can significantly impact translation performance, especially for languages like Chinese (Zhao et al., 2013). In Chinese, spaces are not used to separate words, which can be made up of multiple characters, making good text segmentation very important for determining which words are present in a sentence.

We tested Jieba in our own system, aiming to use its ability to identify proper nouns to form a basic dictionary of names in the text. Specifically, we filtered for phrases tagged "nr" (person name) and "ns" (place name) (Jieba, 2020). Unfortunately, Jieba had a high false positive rate, and often split up phrases or names, which made it unsuitable to form our name dictionary. For example, if the name contained a common noun that could be part of many phrases, replacing that part of the name with the English meaning would be very unhelpful and create a weird-sounding name. However, although Jieba was not suitable for identifying proper nouns, it was still useful for determining a phrase's part of speech.

A name dictionary's main purpose is to translate names consistently, and is more useful when it contains names that appear often. If a character or place appears only once within the story, readers do not need a consistent translation across mentions to recognize it. Additionally, it is likely to be insignificant to the story, so even if the translation is not the best, it is unlikely to affect enjoyment much.

Names commonly occurring in the text are likely to be re-occurring characters, such as the main protagonist or important supporting characters.

We decided to try to feed a percentage of lines to GPT-3.5 for name identification. GPT-3.5 was good at identifying names, rarely returning false positives. However, we didn't need to get every single name from a text, just the re-occurring ones. This meant that GPT was sifting through a large number of duplicates, and incurring extra cost through the API.

We manually identified the names within the sample dataset from last year's task, and for each text, we calculated the total number of unique names, the total percentage of characters within the text that belonged to a name, and the average number of lines that would contain a name. We found that the character percentages ranged from 5% to 10%, and the line percentages ranged from 11% to 22%.

By only giving GPT a certain percentage of lines that were randomly selected from the text, we introduced an element of chance into our pipeline that meant GPT may not be able to see all the names from the text it is given. We selected 20% as a number on the higher end of the range we found, so it was likelier that GPT would be given a majority of the names.

In order to pick lines more likely to contain names, we used Jieba to identify the number of nouns within a line. We theorized that it was unlikely for lines to contain no names, so if Jieba didn't identify any nouns at all within a sentence, it likely missed a name, which may be a combination of characters that are verbs or adjectives on their own. We first ran Jieba's segmentation on the text, and then selected only from a pool of longer sentences without any nouns identified.

We also theorized that for characters such as these, their introductions are more likely to be concentrated within the beginning or middle of the text, rather than the end. As we only need to get one occurrence of each name, we decided to weigh sentences earlier as more likely to be selected. We picked 15% of the lines from the first $\frac{3}{4}$ of the text, and 5% of the lines from the last $\frac{1}{4}$ of the text.[1]

[1] "Lines" in the text file are sometimes multiple sentences in the Chinese raws; so if Jieba identifies 0 nouns in a "line", that can equate to 0 nouns in a paragraph.

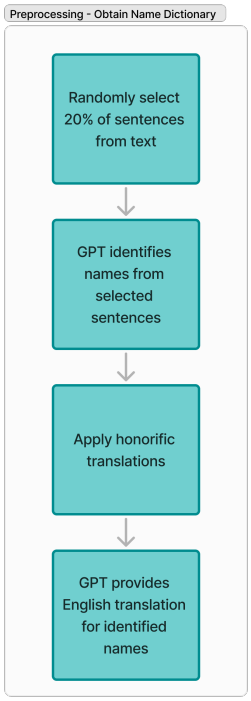

Figure 2: Flow chart describing name translation process.

3.2 Honorifics (Name Translation)

We compiled a list of honorifics, ranging from common honorifics such as "哥哥" (brother) to martial art novel honorifics such as "师爷" (grandmaster) (Mountain, 2017) to historical novel honorifics such as "公公" (eunuch) (Wyhcwe, 2022). We acknowledge that the translations of such terms can sometimes vary across different translations, but we wanted to make a standard translation across all of our translations.

The names identified by Jieba and GPT-3.5 in the previous subsection include these honorifics, so by first applying the honorific translation and only asking GPT to translate the remaining characters left behind as the name, it can standardize the name translations and also ensure the honorifics aren't translated as pinyin directly. For example, our code would go through these steps to translate a name: 韩少爷→ Young Master 韩→ Young Master Han. We used the prompt: "Translate this name to English: [name]. Only list the English name."

3.3 CC-CEDICT (Idiom Translation)

We used the Chinese-English dictionary, CC-CEDICT. It is an free online dictionary that is regularly updated through crowdsourcing, and every contribution is verified regularly and added to

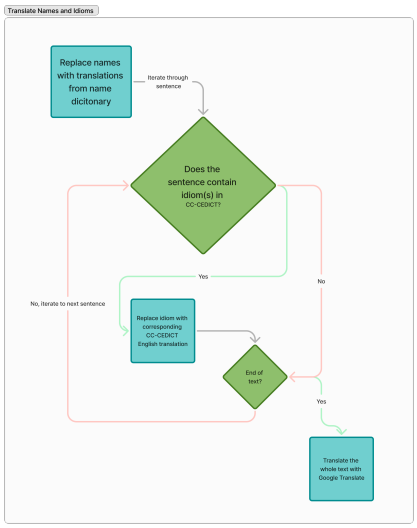

Figure 3: Flow chart describing idiom replacement process.

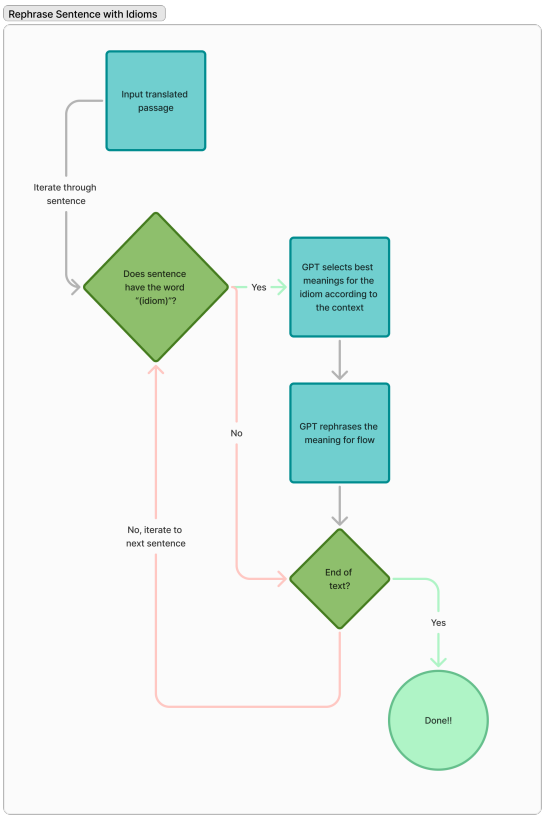

Figure 4: Flow chart describing final rephrasing process.

the database (CC-CEDICT, 2020). Due to continual updates by the owners, CC-CEDICT is a good choice for getting the most updated figurative meanings of idioms, slang, and other culturally specific terms. For example, the phrase "脑子进水" from the introduction section has the CC-CEDICT entry "to have lost one's mind crazy soft in the head."

We searched through all entries labelled as "idioms" within the CC-CEDICT dictionary. If any such idioms were found within a line of the original Chinese text, the Chinese idiom would be replaced directly with its corresponding CC-CEDICT English entry. Other Chinese text in the line not identified as idioms would not be translated at this step. The raw dictionary replacements did not account for grammatical context surrounding the idioms, and some entries contained more than one English translation phrase per Chinese idiom. Furthermore, in their raw formatting these entries were surrounded by brackets and contained a text flag "(idiom)". We kept the raw replacement formatting as-is, which we then processed further after translating the rest of the text.

3.4 Overall Translation

We were left with text that was primarily still in the original Chinese, but with names and idioms programmatically replaced with English translations. We experimented with two different translation engines, DeepL and Google Translate, to translate the remainder of the text. These engines are the two most mentioned translation engines amongst online webnovel forums before ChatGPT. The engines translated the remaining Chinese text without any modification to the already-present English idiom replacements, thus requiring a step to smooth out the sentences containing idioms.

3.5 GPT Rephrasing (Idiom Translation)

We decided to use GPT-4 and LangChain (LangChain, 2024) to replace every line that contained a raw idiom definition, as identified by their surrounding brackets and accompanying "(idiom)" flag, with a grammatically correct rephrasing. Langchain is an open-source framework that makes it easier to develop using GPT's API. We found that GPT-4 was better than GPT-3.5 at rephrasing only the sentences with idiom definitions within a given line. Because any output from this section would be inserted directly into the text as the final step, we decided to switch to GPT-4 for this step for better quality. To minimize the API cost in the rephrasing phase, we recorded the line numbers for the lines modified in the CC-CEDICT step, and only gave those lines to Langchain. Only about 2%-11% of lines across the samples we encountered contained idioms, so GPT-4 was only used on a small percentage of the text.

We used the prompt: "Please pick the idiom definition that best fits the context for the following sentences and rephrase only the part of the sentence with the idiom grammatically. Only output the new translation. Don't change the sentences without idioms. Favor the more concise meaning and find an English equivalent if possible." We added many instructions to our prompt as a result of experimentation; not asking for the "more concise meaning" or "English equivalent" often resulted in translations that were complicated amalgamations of every definition provided by the dictionary entry; not asking for "don't change the sentences without idioms" often resulted in sentences without idioms being changed and other content given being cut out.

Once the GPT-4 rephrasing was complete, the text translation was considered to be finished.

3.6 Evaluation

We used three metrics for automatically evaluating machine-translated text: BLEU, chrF, and COMET. BLEU evaluates word-level n-grams, calculating the precision between the machine translation and the reference, weighted by a brevity penalty (Papineni et al., 2002). ChrF evaluates character-level n-grams, scoring the overlap of short sequences of characters between the machine translation and the reference (Popović, 2015). COMET is a fine-tuned neural framework that takes in sentence embeddings from the source text, translation, and reference (Rei et al., 2020). We used these because last year's conference proceedings summary paper used them for the automatic evaluation (Wang et al., 2023).

A shortcoming of automatic metrics such as BLEU is that they lack the ability to evaluate based upon semantics, instead favoring direct word-to-word matches between a translation and reference (Callison-Burch et al., 2006). This means a translation that achieves high grammatical quality but uses different words than a provided reference could potentially score poorly. As such, we also surveyed human readers to compare the quality of our system's translations. Participants were given 4 separate translations of a text sample ranging from 200-300 English words, each generated using a different method: one generated by our translation system using Google Translate ("pipeline Google Translate"), one generated by our translation system using DeepL ("pipeline DeepL"), one generated using only Google Translate ("pure Google Translate"),

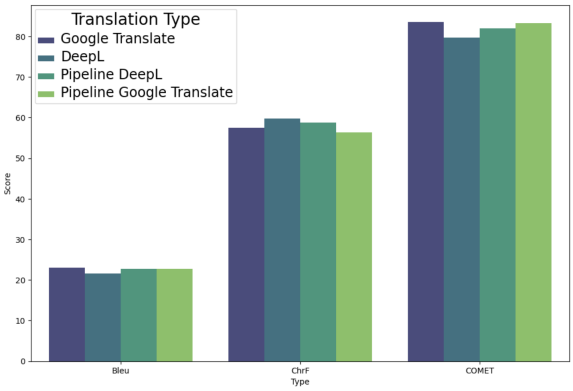

Figure 5: random sample 1, video games (20%)

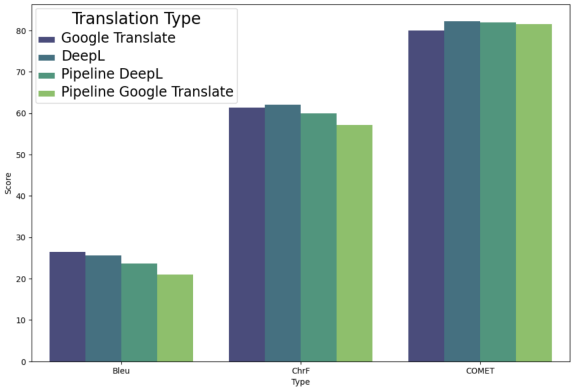

Figure 6: random sample 2, science fiction (23%)

and one using only DeepL ("pure DeepL"). These translations were given in a random order, and participants were not informed of which translation came from which source. After reading the translations, participants were asked to rank them from best to worst based on how readable they found the translations. This process was repeated over several different samples, and the rankings were recorded for each sample.

4 Results

We used the automated metrics to evaluate the results of the four techniques: pure Google Translate, pipeline Google Translate, pure DeepL, pipeline DeepL. To decide between Google Translate and DeepL for our final submission, we decided to compare the pure Google Translate and pure DeepL results. In this paper, we show the results for three random samples selected across the dataset, shown in Figures 5, 6, and 7. The genre of the sample is labelled, along with their distribution percentage in the training set (Wang et al., 2023). Google Translate and DeepL performed about the same for the first two samples, but Google Translate was signifi-

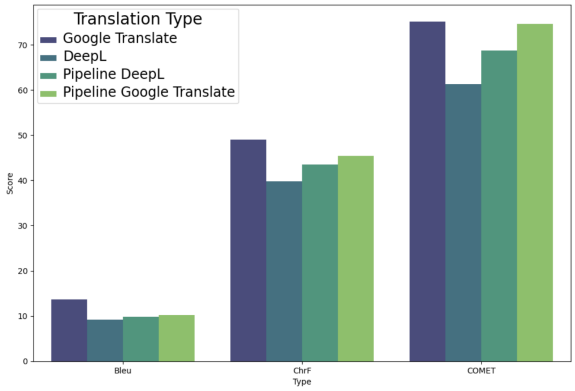

Figure 7: random sample 3, martial arts (2%)

cantly better than DeepL in the third sample, which was a classical martial arts novel. Although martial arts novels only make up a small percentage of the dataset, because idioms originate from classical Chinese literature, we decided to employ the translation pipeline with Google Translate ("pipeline Google Translate") for our final conference submission.

In our A/B testing, across all the samples, we found that participants ranked the pipeline Google Translate output the highest most, and the pure DeepL the lowest, as shown in Table 1. However, the distribution was mostly even, and about half the time, participants reported that the difference between translations was slight, which could be due to the limitations mentioned in the following limitations section.

Technique	1st	2nd	3rd	4th
pure Google (2)	2	2	1	2
pipeline Google (1)	3	1	3	0
pure DeepL (4)	1	3	0	3
pipeline DeepL (3)	2	0	4	1

Table 1: Times each technique was ranked 1st, 2nd, 3rd, or 4th across 7 samples. Ties were allowed.

5 Conclusion

We created a machine translation system that creates consistent translations for names and accurate translations for idioms, both of which enhance human readability despite making up a small ratio of the overall text. Even though our pipeline did not see any major improvements in the automated evaluation metrics, the positive reception among human survey participants points to the potential value that our process provides.

6 Limitations

When providing lines of text for ChatGPT to identify names, we randomly selected a certain percentage of lines to use in order to reduce API usage costs. Though the selected lines were weighted based on factors such as whether or not Jieba found any proper nouns in a line, there is nonetheless a slight element of randomness that is introduced during our process. One limitation that could be further explored is how consistently our pipeline performs over multiple runs on the same input.

Another limitation in our results lay in our use of human evaluators. Participants were asked to rank translations that used our system against translations that did not. Though they were not informed which translations did or did not use our system, they also were not given any specific metrics to quantify their decisions. Participants also sometimes reported that the passages provided were too long to quickly judge the difference, and that reading four passages in a row that described the same content made it hard to evaluate the difference without an earnest effort to study the differences within the text. In the future, our team could work on developing a more robust approach to the human side of evaluations that addresses these limitations.

Acknowledgements

This research project was fully funded by the Halıcıoğlu Data Science Institute at UC San Diego, through their HDSI Undergraduate Scholarship Program. We would like to thank them for their support throughout the year. We would also like to thank our mentor, Professor Jingbo Shang, for guiding us through the project, consistently checking in with us on our progress, and supporting us in publishing our first paper.

References

Li An, Linghao Jin, and Xuezhe Ma. 2023. Max-isi system at wmt23 discourse-level literary translation task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 282–286.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of bleu in machine translation research. In *11th conference of the european chapter of the association for computational linguistics*, pages 249–256.

CC-CEDICT. 2020. Chinese-English Dictionary. https://cc-cedict.org/wiki/. [Online; accessed 28-July-2024].

Jieba. 2020. Jieba. https://github.com/fxsjy/jieba. [Online; accessed 30-July-2024].

LangChain. 2024. LangChain. https://www.langchain.com/langchain. [Online; accessed 30-July-2024].

Fabien Lopez, Gabriela González-Sáez, Damien Hansen, Mariam Nakhlé, Behnoosh Namdarzadeh, Marco Dinarelli, Emmanuelle Esperança-Rodier, Sui He, Sadaf Mohseni, Caroline Rossi, et al. 2023. The make-nmtviz system description for the wmt23 literary task.

Immortal Mountain. 2017. WuXia Terms of Address. https://immortalmountain.wordpress.com/glossary/terms-of-address. [Online; accessed 28-July-2024].

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Maja Popović. 2015. chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the tenth workshop on statistical machine translation*, pages 392–395.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. Comet: A neural framework for mt evaluation. *arXiv preprint arXiv:2009.09025*.

Longyue Wang, Siyou Liu, Minghao Wu, Wenxiang Jiao, Xing Wang, Jiahao Xu, Zhaopeng Tu, Liting Zhou, Yan Gu, Weiyu Chen, Philipp Koehn, Andy Way, and Yulin Yuan. 2024. Findings of the wmt 2024 shared task on discourse-level literary translation. proceedings of the ninth conference on machine translation (wmt).

Longyue Wang, Zhaopeng Tu, Yan Gu, Siyou Liu, Dian Yu, Qingsong Ma, Chenyang Lyu, Liting Zhou, Chao-Hong Liu, Yufeng Ma, et al. 2023. Findings of the wmt 2023 shared task on discourse-level literary translation: A fresh orb in the cosmos of llms. *arXiv preprint arXiv:2311.03127*.

WMT23. 2023. WMT23 Task Link. http://www2.statmt.org/wmt23/literary-translation-task.html. [Online; accessed 23-July-2024].

Wyhcwe. 2022. Historical Terms of Address. https://dreamsofjianghu.ca/%e5%85%ab%e5%ae%9d%e5%a6%86-eight-treasures-trousseau/glossary/. [Online; accessed 28-July-2024].

Anqi Zhao, Kaiyu Huang, Hao Yu, and Degen Huang. 2023. DUTNLP system for the WMT2023 discourse-level literary translation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 296–301.

Hai Zhao, Masao Utiyama, Eiichiro Sumita, and Bao-Liang Lu. 2013. An empirical study on word segmentation for chinese machine translation. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 248–263. Springer.