Machine Translating From English to Chinese for E-Commerce Product Categorization[*]

Kitty Duong
University of Windsor
Windsor, Canada
duongy@uwindsor.ca

Miaomiao Zhang
University of Windsor
Windsor, Canada
zhang3s2@uwindsor.ca

ABSTRACT

This study explores the application of a machine translation model, Meta AI's NLLB-200[8], for the translation of Amazon product categories from English to Chinese. Given the critical role of accurate category translation in enhancing user experience and facilitating seamless e-commerce navigation, this research aims to evaluate the efficacy and accuracy of the NLLB-200 model against the existing Chinese categories on Amazon. Through a systematic translation of the sample data from Amazon and a comparison process between experiment results and the existing results, the study assesses the accuracy of NLLB-200 translations, identifying both the model's strengths and its limitations in handling e-commerce terminology. The findings indicate that while NLLB-200 shows promise in accurately translating a wide range of product categories, discrepancies in certain translations highlight areas for further refinement.[9] This research contributes to the ongoing discussion on improving machine translation for e-commerce applications and offers insights into the potential of NLLB-200 to support multilingual e-commerce platforms. The outcomes not only underscore the importance of leveraging advanced AI for localization efforts but also pave the way for future enhancements in machine translation technologies.

KEYWORDS

Multilingual NLP, machine translation, e-commerce product categories, e-commerce translation, translation accuracy, translation evaluation

1 PROBLEM DEFINITION

The global nature of e-commerce demands accurate and efficient localization of platform content, including product categories, to cater to diverse linguistic audiences. This localization is pivotal for user experience, searchable, and navigation efficiency on multinational platforms like Amazon. Traditional machine translation tools often struggle with maintaining accuracy, especially for languages with significant syntactic and semantic differences, such as English and Chinese.[4] Introducing advanced machine translation models like NLLB-200 offers a potential solution to these challenges by promising high-quality translations across a wide range of languages, including those less represented in digital resources.

However, the effectiveness of NLLB-200 in the specific context of e-commerce, particularly for the accurate translation of product categories from English to Chinese, remains an open question. Given the critical role of these categories in user interaction and the unique challenges posed by specialized e-commerce terminology, there is a pressing need to evaluate the performance of NLLB-200. This involves assessing its translation accuracy compared to manually curated categories on Amazon's Chinese platform and identifying any systematic discrepancies that could impact user experience.[5] Addressing this problem requires a detailed analysis of NLLB-200's translation outcomes and a comparison framework that considers both direct translation accuracy and the semantic integrity of category labels.

2 PROPOSED APPROACH

Our proposed machine translation evaluation method aims to fine-tune a pre-trained model and further train it for e-commerce product category translation from English to Chinese, then compare the accuracy of the results with the existing Chinese version on Amazon, as defined in the previous section. The approach works through two pipeline phases: the translation process and the comparison framework. In the following subsections, we describe the details of each step.

2.1 Translation Process

In our approach, the translation process for converting Amazon product categories from English to Chinese using the NLLB-200 model[2] entails preparing a standardized list of categories, setting up access to the NLLB-200 model through its API or SDK, and configuring translation parameters. Categories are then translated in batches, with errors and limitations managed appropriately. The resulting translations undergo initial review before being compared to Amazon's official Chinese categories to assess accuracy. The data of the experiment results would be stored in a file named "result.csv".

2.1.1 Data Preparation. In this experiment, we used the data from "Amazon_Ecommerce_Data_2020.csv"[12] as the dataset, which could provide a comprehensive list of Amazon product categories in English. Due to the size of the dataset, we selected around 90 pieces of data by random, which covers a wide range of product categories, so that we could test the translation capabilities of NLLB-200 across different terminologies and contexts.

2.1.2 Fine-tune NLLB-200 Model. We utilized the API provided by MetaAI to set up an environment with any necessary dependencies and authentication for accessing the NLLB-200 model. Besides, we followed the tutorial from Hugging Face to use PyTorch Trainer to finetune the pre-trained model[6]. Based on the configuration settings to specify the source (English) and target (Chinese) languages, we organized the product categories into batches for efficient processing, and then we sent batches to NLLB-200 for translation processing.

[*] https://github.com/duongy18418/Multilingual-NLP/tree/main/Code

Algorithm 1 Finetuning and comparing process

```
Inputs:
    dataset $\mathbb{D}$, Amazon Chinese categories $\mathbb{Z}$, sample data $\mathbb{S}$
Initialization:
    import transformer and datasets
    load NLLB-200 Model
    size of sample data = 90
    size of accurate data = 0
    $S \subset D, S = [d_1, d_2, d_3.....d_i]$
Output: 90 sample data in Chinese saved in result.csv
procedure FINETUNE AND COMPARE
    if $(D! = \varnothing) \wedge (S \subset D)$ then
        for all $d_i \in S$ do
            $d'_i \leftarrow finetune(d_i)$
            $d'_i \in S'$
    for all $d'_i \in \mathbb{S}'$ do
        $res \leftarrow ndiff(d_i, d'_i)$
        if $res == true$ then
            $acc \leftarrow acc + 1$
        else
            get_close_matches($d_i, d'_i$)
```

2.1.3 *Post-Translation Processing.*

As for the experiment results, we stored the translated categories after the translation processing, alongside their original English counterparts in a file named "result.csv" for further analysis. For the initial quality assurance, we collected the existing Chinese categories from the Amazon website and then conducted a preliminary review of the translation results to identify any glaring errors or inconsistencies that could indicate issues with the translation process.

2.2 Comparison Framework

In this experiment, the comparison work would focus on semantic equivalence and terminology appropriateness between the experiment results and the existing Amazon Chinese version. Furthermore, discrepancies are documented, and a statistical analysis is conducted to evaluate NLLB-200's overall performance, aiming to highlight the model's strengths and areas for improvement in e-commerce localization.

For each translated category, we attempt to find an exact or closely matching category within Amazon's existing Chinese version. As for evaluating the accuracy of translation results, we need to consider factors such as semantic equivalence, appropriateness of terminology, and consistency with Amazon's category naming conventions.

As for the 90 pieces of testing data in this experiment, we did a manual review for the initial assessment of translation accuracy. The experiment result shows around 80% of accurate translation compared with the Chinese version on the Amazon website. There would be some inappropriate translations concerning semantic analysis.[11] For instance, the translation of the 7th testing data in the result.csv ("Window Treatments") is "窗口治疗", which translated the whole vocabulary into two vocabularies separately. "窗口治疗" means "window" plus "treatments" in Chinese, which is coming with no actual meanings. "Window Treatments" should be translated into "窗上用品" or something similar, which means a set of decorations for a window, often including curtains or blinds.[3] Another example is the 9th testing data, the translation of "Baby&Toddler toys". The translation result is "宝宝" for both "Baby" and "Toddler". There is a distinct difference between these two vocabularies, as "Baby" means kids at the stage from zero to twelve months while "Toddler" means kids at the stage from one to three years.[10] Furthermore, there would be duplication errors for some testing data, the result is repeating the same characters.[7] As we could find from the comparison outcome in the experiment of the 90 pieces of testing data, the NLLB-200 model is unable to make appropriate translations for some data.

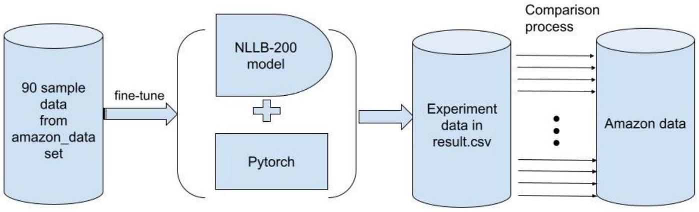
90 sample data from amazon_data set
fine-tune
NLLB-200 model
Pytorch
Experiment data in result.csv
Comparison process
Amazon data

Figure 1: Finetune process architecture.

As for the further analysis for semantic equivalence, appropriateness of terminology, and consistency with Amazon's category naming conventions, we would continue the evaluation process in the next stage, to provide a more specific and detailed analysis, which is expected to be implemented in Python via methods provided by Python API, such as "ndiff" method for detailed comparison and "get_close_matches" method for finding similar lines, [1]in which way we can acquire the completely same and similar results of all the pieces of data in the Amazon English categories.

REFERENCES

[1] [n. d.]. *difflib — Helpers for computing deltas.* https://docs.python.org/3/library/difflib.html

[2] 2024. *facebook/nllb-200-distilled-600M · Hugging Face.* https://huggingface.co/facebook/nllb-200-distilled-600M

[3] 2024. *Window-treatment Definition & Meaning | YourDictionary.* https://www.yourdictionary.com/window-treatment

[4] Shihua Brazill, Michael Masters, and Pat Munday. 2017. *Digital Commons @ Montana Tech ANALYSIS OF HUMAN VERSUS MACHINE TRANSLATION ACCURACY.* https://digitalcommons.mtech.edu/cgi/viewcontent.cgi?article=1226&context=grad_rsch

[5] Li-Ching Chang. 2022. Chinese language learners evaluating machine translation accuracy. *The JALT CALL Journal* 18, 1 (Apr 2022), 110–136. https://doi.org/10.29140/jaltcall.v18n1.592

[6] Hugging Face. [n. d.]. *Fine-tune a pretrained model.* https://huggingface.co/docs/transformers/en/training

[7] Maarit Koponen. 2010. *Assessing Machine Translation Quality with Error Analysis.* https://www.sktl.fi/@Bin/40701/Koponen_MikaEL2010.pdf

[8] Meta. [n. d.]. *No Language Left Behind.* https://ai.meta.com/research/no-language-left-behind/

[9] Yasmin Moslem. 2023. *Fine-tuning Large Language Models for Adaptive Machine Translation.* https://arxiv.org/pdf/2312.12740.pdf

[10] American Academy of Pediatrics. 2019. *Ages & Stages.* https://www.healthychildren.org/english/ages-stages/pages/default.aspx

[11] Martha Palmer and Zhibiao Wu. 1995. Verb semantics for English-Chinese translation. *Machine Translation* 10, 1-2 (Mar 1995), 59–92. https://doi.org/10.1007/bf00997232

[12] PromptCloud. 2020. *Amazon Product Dataset 2020.* https://www.kaggle.com/datasets/promptcloud/amazon-product-dataset-2020