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

In this study, we evaluate the performance of two machine translation models, NLLB-200[3] and Google Translator[4], in translating Amazon product categories[15] from English to Chinese. The dataset consists of a collection of product categories extracted from the Amazon platform. We employ BLEU (Bilingual Evaluation Understudy)[1] as the evaluation metric to compare the translations produced by the two models. Google Translator serves as the baseline system due to its widespread use and availability. The experimental results reveal the effectiveness of NLLB-200[10] in capturing the nuances of the product categories, outperforming Google Translator in terms of BLEU scores. Through a comprehensive analysis of the results, we discuss the strengths and weaknesses of both translation models[8], highlighting potential areas for improvement. Our findings contribute to the understanding of machine translation performance[7] in the context of e-commerce product categorization and provide insights for further research in this domain.

KEYWORDS

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

1 EXPERIMENTAL SETUP AND EVALUATION

1.1 Dataset

In our experiments, we use a publicly available Amazon e-commerce product category dataset [1] collected and published by PromptCloud [2]. It consists of approximately 10,000 entries based on 22 diverse main categories. Each entry of the data would include the specific and detailed subcategories under the main category, i.e., Sports Outdoors Sports Fitness Team Sports, in which "Sports Outdoors" is the main category. At the same time "Team Sports" is the second-level subcategory of the subcategory "Sports Fitness".

1.2 Setup

Our proposed approach consists of two phases to set up the environment for e-commerce product category translation from English to Chinese; translating the original context via a pre-trained model NLLB-200 and preparing the train data and test data. Here, we provide the implementation details and the setup of our approach in each of these phases.

1.2.1 Pre-trained model translation. The translation process for converting Amazon product categories from English to Chinese using the NLLB-200 model entails preparing a standardized list of categories, setting up access to the NLLB-200 model through its API or SDK, and configuring translation parameters. Besides, we followed the tutorial from Hugging Face[6] to use PyTorch Trainer to finetune the pre-trained model[12]. We stored this model as "E-commerce_Translation_Model"[2], which is posted in Hugging face. 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. Categories are then translated in batches, with errors and limitations managed appropriately.

1.2.2 Preparing training and testing data. In the data preparation phase, we curated two distinct datasets as the training dataset $\mathbb{D}_i$ and testing dataset $\mathbb{D}_j$, $\forall \mathbb{D}_i, \mathbb{D}_{j\neq i} : \mathbb{D}_i \cap \mathbb{D}_j = \varnothing$. The training data[13], derived from Google-Product, was selected due to its suitability for training machine translation models and its availability. Before training, the Google Product dataset underwent preprocessing steps, including data cleaning and tokenization, to ensure consistency and quality. Conversely, the testing data[5] consisted of Amazon e-commerce product categories, chosen to represent real-world translation scenarios relevant to our task. Preprocessing of the testing data involved similar steps to align it with the format and requirements of the translation models. Throughout this process, careful attention was paid to maintaining the integrity and representativeness of both datasets to facilitate an accurate evaluation of the translation models' performance.

1.3 Metrics

As for the evaluation metric, we utilized BLEU (Bilingual Evaluation Understudy)[14]. BLEU is a widely used metric for evaluating the quality of machine-translated text by comparing it to one or more reference translations. It quantifies the similarity between the machine-generated translation and the reference translations based on n-gram overlap. In the context of our project, BLEU provides a standardized measure to assess the accuracy and fluency of the translated Amazon e-commerce product categories produced by the NLLB-200 and Google Translator models. A higher BLEU score[11] indicates a closer resemblance between the generated translations and the reference translations, thus reflecting a higher quality of translation. By employing BLEU as the evaluation metric, we aim to

*https://github.com/duongy18418/Multilingual-NLP/tree/main/Code
[1]https://www.kaggle.com/datasets/promptcloud/amazon-product-dataset-2020/

[2]https://huggingface.co/duongy18418/E-commerce_Translation_Model/tree/main

Conference'17, July 2017, Washington, DC, USA
2024. ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

objectively compare the performance of the two translation models across different categories of e-commerce products and provide insights into their relative strengths and weaknesses. In the result analysis part, we will display the comparison diagram based on the BLEU scores of the two models.

1.4 Baselines

In this report, the baseline system utilized for comparison is Google Translator. Google Translator serves as a benchmark against which the performance of the NLLB-200 model in translating Amazon e-commerce product categories from English to Chinese is evaluated. Google Translator[16] is chosen as the baseline due to its widespread use and availability as a popular machine translation tool. By comparing the translations generated by NLLB-200 with those produced by Google Translator, we can assess the relative effectiveness and accuracy of the NLLB-200 model in capturing the nuances of the product categories. The comparison with Google Translator provides a practical reference point for evaluating the performance of NLLB-200 and gaining insights into its strengths and limitations. Additionally, the baseline comparison helps contextualize the results obtained from NLLB-200, offering valuable benchmarks for assessing the progress and advancements in machine translation technology.

1.5 Evaluation results

In the result section of this report, we will provide a comprehensive analysis of the performance of both the baseline model (Google Translator) and the proposed model (E-commerce-Translation-Model) based on the BLEU scores obtained. The result analysis is conducted from perspectives of BLEU scores and runtime comparison.

1.5.1 BLEU scores. The BLEU scores obtained for the E-commerce-Translation-Model demonstrate a slightly higher and more consistent performance than the baseline model, Google Translator. Across various categories of e-commerce products, the E-commerce-Translation-Model consistently yielded BLEU scores that were marginally superior to those of the baseline model. This suggests that the proposed model may better capture the nuances of product descriptions and translate them accurately from English to Chinese. Despite the modest improvements observed in BLEU scores[11], it's important to note that both models' translations fall significantly below the ideal score of 1, indicating that there is still room for improvement in terms of translation quality.

1.5.2 Runtime comparison. In terms of runtime, the proposed model E-commerce-Translation-Model exhibited a substantially longer processing time compared to the baseline model. The E-commerce-Translation Model took approximately 287.72 seconds to complete the translation task, whereas the baseline model only required 22.18 seconds. This significant disparity in runtime raises concerns about the scalability and efficiency of the E-commerce-Translation-Model, particularly in real-world applications where speed and responsiveness are critical factors. Further investigation is warranted to identify the factors contributing to the increased computational complexity[9] of the E-commerce-Translation-Model and explore strategies for optimizing its performance without compromising translation quality.

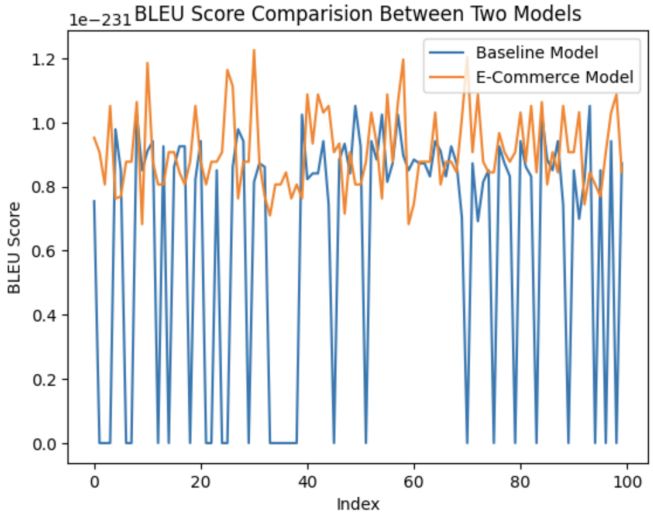

Figure 1: Comparative performance on Baseline model and E-commerce model.

REFERENCES

[1] [n. d.]. https://www.nltk.org/_modules/nltk/align/bleu_score.html
[2] 2021. https://www.promptcloud.com/
[3] 2024. *facebook/nllb-200-distilled-600M · Hugging Face.* https://huggingface.co/facebook/nllb-200-distilled-600M
[4] Nidhal Baccouri. [n. d.]. deep-translator: A flexible free and unlimited python tool to translate between different languages in a simple way using multiple translators. https://pypi.org/project/deep-translator/#google-translate-1
[5] R.A. DeMillo, R.J. Lipton, and F.G. Sayward. 1978. Hints on Test Data Selection: Help for the Practicing Programmer. *Computer* 11, 4 (Apr 1978), 34–41. https://doi.org/10.1109/c-m.1978.218136
[6] Hugging Face. [n. d.]. *Fine-tune a pretrained model.* https://huggingface.co/docs/transformers/en/training
[7] Peter Flach. 2019. Performance Evaluation in Machine Learning: The Good, the Bad, the Ugly, and the Way Forward. *Proceedings of the AAAI Conference on Artificial Intelligence* 33 (Jul 2019), 9808–9814. https://doi.org/10.1609/aaai.v33i01.33019808
[8] Hugh G. Gauch, J. T. Gene Hwang, and Gary W. Fick. 2003. Model Evaluation by Comparison of Model-Based Predictions and Measured Values. *Agronomy Journal* 95, 6 (Nov 2003), 1442–1446. https://doi.org/10.2134/agronj2003.1442
[9] Oded Goldreich. 2008. Computational complexity. *ACM SIGACT News* 39, 3 (Sep 2008), 35. https://doi.org/10.1145/1412700.1412710
[10] Yeskendir Koishekenov, Alexandre Berard, and Vassilina Nikoulina. 2023. Memory-efficient NLLB-200: Language-specific Expert Pruning of a Massively Multilingual Machine Translation Model. https://doi.org/10.48550/arXiv.2212.09811
[11] Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2020. Tangled up in BLEU: Reevaluating the Evaluation of Automatic Machine Translation Evaluation Metrics. *arXiv (Cornell University)* (Jun 2020). https://doi.org/10.48550/arxiv.2006.06264
[12] Bonan Min, Hayley Ross, Elior Sulem, Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. 2023. Recent Advances in Natural Language Processing via Large Pre-Trained Language Models: A Survey. *Comput. Surveys* 56, 2 (Jun 2023). https://doi.org/10.1145/3605943
[13] Robert Moore and William Lewis. 2010. *Intelligent Selection of Language Model Training Data.* 11–16 pages. https://aclanthology.org/P10-2041.pdf
[14] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. *BLEU: a Method for Automatic Evaluation of Machine Translation.* https://aclanthology.org/P02-1040.pdf
[15] PromptCloud. 2020. *Amazon Product Dataset 2020.* https://www.kaggle.com/datasets/promptcloud/amazon-product-dataset-2020
[16] Adi Sutrisno. 2020. The Accuracy and Shortcomings of Google Translate Translating English Sentences to Indonesia. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3747888