Improving Machine Translation in the E-commerce Luxury Space. A case study

José-Manuel de-la-Torre-Vilariño
Acclaro
jvilarino@acclaro.com

Juan-Luis García-Mendoza
Université Sorbonne Paris Nord
garciamendoza@lipn.univ-paris13.fr

Alessia Petrucci
Acclaro
apetrucci@acclaro.com

Abstract

This case study presents a Multilingual e-commerce Project, which principal aim is to create an improved system that translates product titles and descriptions, plus other content in multiple languages. The project consisted of two main phases; a research-intensive solution using state-of-the-art Machine Translation systems and baseline language models for two language pairs, and the development of a Machine Translation system. The features implemented included Quality Estimation, model benchmarking, entity recognizers, and automatic domain detection. mBART model was used to create the system for the specific domain of e-commerce, for luxury items.

1 Introduction

Machine Translation (MT) is the automatic translation of text from one language to another without human intervention (Stahlberg, 2020). When this translation is performed using Deep Neural Networks (DNN), it is referred to as Neural Machine Translation (NMT) (Stahlberg, 2020). NMT technology has made significant progress in recent years, however, they have generally been trained on domain-general data, directly affecting domain-specific translations (Martins et al., 2022; Saunders, 2022). According to Martins et al. (2022), most methods for adapting MT systems to a specific domain focus on fine-tuning.

One of those specific domains is e-commerce. In today's globalized world, e-commerce has become a crucial part of the economy. With the rise of online shopping, businesses must be able to communicate with customers in their native languages to provide a seamless shopping experience. However, translating product titles and descriptions, reviews, and other content while maintaining formal and informal styles, and dealing with lengthy and very short sentences, can be a daunting task.

In this work, an e-commerce Multilingual Project aimed to improve machine translation quality is introduced. This improvement was carried out with mBART model (Liu et al., 2020), which is leading the way in Multilingual Translation. This model is designed to handle multiple languages simultaneously, making it ideal for e-commerce applications where content needs to be translated quickly and accurately. The project has been led by Acclaro[1], a leading company with extensive experience in professional translation services.

2 Arquitectures

In the state-of-the-art (SoTA) of NMT, the architectures proposed by Radford et al. (2019), Liu et al. (2020) and Tang et al. (2020) stand out. On the one hand, GPT-2 model proposed by Radford et al. (2019) is based on the architecture of large transformers (Vaswani et al., 2017). Besides, GPT-2 follows the details of the OpenAI GPT model proposed by Radford et al. (2018). On the other hand, according to Liu et al. (2020), mBART is "a multilingual sequence-to-sequence denoising auto-encoder" that uses BART (Lewis et al., 2020) large-scale monolingual corpora across many languages. This model was pre-trained using a subset

[1] https://www.acclaro.com/

of data in 25 languages extracted from Common Crawl[2]. Finally, Tang et al. (2020) add to mBART the ability to perform multilingual finetuning and extend it to 50 languages without training from scratch. This paper refers to these mBART-based models as mBART25 and mBART50, respectively.

The mBART-based models were selected for their ability to surpass the state-of-the-art results in the English-German and English-French language pairs. The evaluation was performed using the BLEU measure (Papineni et al., 2002), which compared the output of machine translation systems with human reference translations. These results are in agreement with those reported by Hendy et al. (2023). It should be noted that these architectures are implemented in NMT framework fairseq[3] (Ott et al., 2019).

Finally, other state-of-the-art architectures were taken into account, including M2M-100 (Fan et al., 2021), NLLB-200 (Costa-jussà et al., 2022), OpenNMT (Klein et al., 2017) and MarianNMT (Junczys-Dowmunt et al., 2018). However, it should be noted that these architectures were only analyzed and not implemented.

3 Methodology

Figure 1 represents the methodology applied in the luxury e-commerce Multilingual Project. Initially, a baseline was defined to set the minimum acceptable performance (see Section 3.2). Then, the sentence pairs to be processed and filtered within the e-commerce domain were established (see Section 3.3). The best quality pairs were used to train and finetune the models (see Section 3.4). The trained models obtained were evaluated and compared with the initial baseline or the baseline of the previous iteration (see Section 3.5). If the model performance improves the baseline, it is deployed using REST API services (see Section 3.6) and a new baseline is established.

Finally, the errors detected in the translations of these models were sent to expert linguists for examination, therefore improving the training pairs for the next iteration.

3.1 Data

The bilingual corpora utilized in this study is property of a luxury e-commerce company and consist of product information (titles and descriptions). At the outset, the initial sentence pairs were built by human translators and post-editors. The totals reached 244386 English-German pairs and 229709 English-French pairs. With the methodology proposed the dataset has since increased to 255643 and 242932 pairs respectively.

3.2 Baseline

The baseline was established using the BLEU evaluation measure on the output of the translation systems. Initially, the values obtained with Google Translate and DeepL were used. While, future iterations, were calculated based on the output of the systems trained on the specific domain. The evaluation period was quarterly.

3.3 Data preprocessing and filtering

In the preprocessing step, elements such as punctuation marks, form texts and Out-Of-Vocabulary (OOV) characters were standardized. Paired sentences of 50 words in length were also removed. In particular, for the English-German case, new orthographic conventions were introduced, plus the normalization of lexical redundancy with the help of Part-of-speech (POS) tagging. On top, tokenization was a key step, removing words with no semantic significance, and corpus markup, providing information about the text itself, by categories in the e-commerce space.

The quality of these bilingual pairs was evaluated using multilingual embedding comparisons, Quality Estimation (QE) models, POS tagging, Named Entity Recognition (NER) and domain classifiers. NMT models achieve good translation quality on domain-specific data via simple fine-tuning on representative training corpora. In addition, a manual evaluation was performed by expert linguists. Pairs with low quality were removed from the set. All experiments were conducted using the NMT framework fairseq. Subword segmentation was handled using SentencePiece (Kudo and Richardson, 2018).

3.4 Training

In this section, fine-tunings of existing pre-trained models is presented. Training from scratch powerful models like GPT-2 (Radford et al., 2019) or mBART (Liu et al., 2020) requires tens of GB of text, which is impossible and more so in the e-commerce space. Also, it's resource expensive, according to Liu et al. (2020), mBART trained for

[2] https://commoncrawl.org/

[3] https://github.com/facebookresearch/fairseq

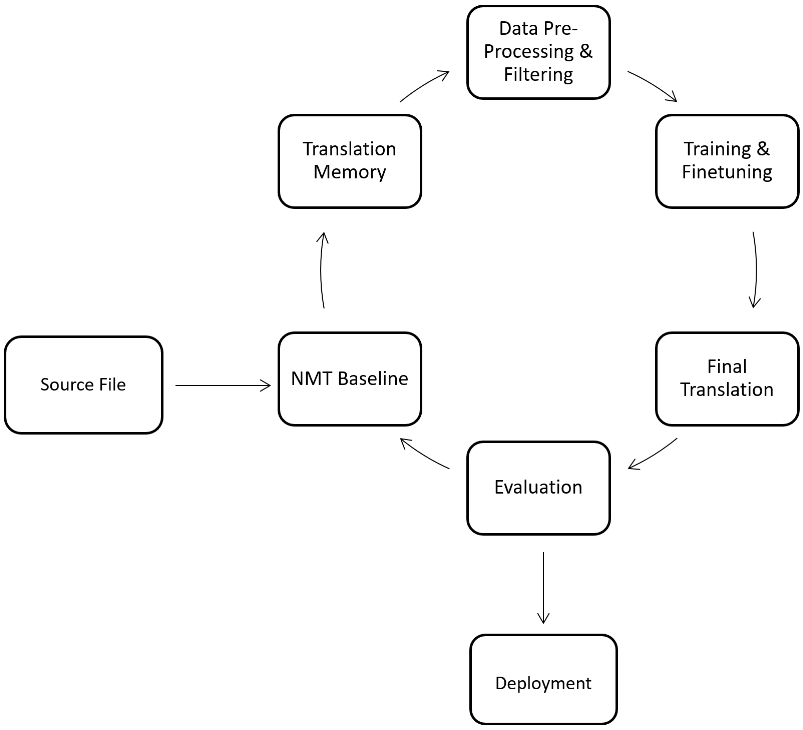

Figure 1: Project methodology.

2.5 weeks on 256 Nvidia V100 GPUs. For example:

- GPT-2, 1.5 billion parameters.
- GPT-3, 175 billion parameters.
- mBART25, 610 million parameters [4].
- mBART50, +610 million parameters.

As mentioned in Section 2, in this work the mBART-based models, mBART25 and mBART50 were used. These models were finetuned using the parameters suggested in the SoTA.

Later, variations in the parameters were made in both models according to the specific domain and data availability. The values of these parameters directly influence the quality of the translations. The best values for each parameter were highlighted.

- Learning Rate: The values $1e^{-3}$, $1e^{-4}$ and $\underline{5e^{-3}}$ with decay scheduled were used.
- Dropout: The values 0.0, 0.05, $\underline{0.1}$, 0.2 and 0.3 were used.
- Attention Dropout: The values 0.0, 0.05, $\underline{0.1}$, 0.2 and 0.3 were used.
- Embedding Layer Normalization: $\underline{\text{Yes}}$ and no.
- Optimizer: Adam.
- Temperature Sampling: The values 0.5, $\underline{1.0}$ and 1.5 were used.
- Beam Search: The values $\underline{5}$, 6, 7 and 9 were used.

Our full model is trained on 4Nvidia V100 Gpus (24GB) for 500K steps.

3.5 Evaluation

Our ongoing evaluation systems described the hybrid approach of automatic metrics, plus a human-in-the-loop method in a Sentence-Level approach. The proprietary QE algorithms in conjunction with the BLEU measure, covered a wide range of the QA process, reducing the post-editors workload through a ranking of sentences on which direct assessment and editing were performed.

The evaluation effort feeds an adaptive neural network that is able to ingest new information and update the production instances. Acclaro linguistic specialist feedback enriches the NMT, and ensures the best possible output.

3.6 Deployment

The translation service is enabled for the client using Kubernetes[5] and REST API services. These

[4] https://github.com/facebookresearch/fairseq/tree/main/examples/mbart

[5] https://kubernetes.io/

services were implemented using the Django[6] framework and use the best models obtained. Besides, its behavior and performance was tested with JMeter[7] The main functionalities of these services are:

- Translate one or more sentences in the English-German or English-French directions
- Integration with Computer-Assisted Translation (CAT) Tools such as XTM[8]. This includes XLIFF format processing, job status management and batch translation.
- Storage of low-quality sentence pairs for future review by linguists. These sentence pairs are used to improve the models in the next iteration.
- Statistics of translations performed at several intervals (current day and year, last 7 and 30 days, last month, etc.)

In addition, a Telegram bot[9] was added to these services and performs the following operations:

- Select sentences with poor quality and send them to expert linguists.
- Translate one or more sentences sent from the Telegram application.
- Obtain the current status of the services.

4 Results

The Tables 1 and 2 show the values of the BLEU measure obtained on pairs of product titles and descriptions. These results are shown concerning to the quarters of the year 2022. The first three quarters were evaluated with mBART25 while the last one with mBART50. The initial baseline was the BLEU obtained by DeepL.

Table 1: BLEU scores for products titles using the model mBART25.

Pair	Google	DeepL	Q1	Q2	Q3	Q4*
English-German	0.667	0.671	0.688	0.700	0.706	0.733
English-French	0.669	0.674	0.691	0.710	0.720	0.729

* The mBART50 multilingual model was used.

Table 2: BLEU scores for product descriptions using the model mBART25.

Pair	Google	DeepL	Q1	Q2	Q3	Q4*
English-German	0.789	0.809	0.811	0.816	0.813	0.821
English-French	0.632	0.640	0.641	0.642	0.640	0.651

* The mBART50 multilingual model was used.

Acknowledgements: The present work was supported by Acclaro.

References

[Costa-jussà et al.2022] Costa-jussà, Marta R, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. arXiv preprint arXiv:2207.04672.

[Fan et al.2021] Fan, Angela, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. 2021. Beyond english-centric multilingual machine translation. The Journal of Machine Learning Research, 22(1):4839–4886.

[Hendy et al.2023] Hendy, Amr, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are gpt models at machine translation? a comprehensive evaluation. arXiv preprint arXiv:2302.09210.

[Junczys-Dowmunt et al.2018] Junczys-Dowmunt, Marcin, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In Proceedings of ACL 2018, System Demonstrations, pages 116–121, Melbourne, Australia, July. Association for Computational Linguistics.

[Klein et al.2017] Klein, Guillaume, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In Proceedings of ACL 2017, System Demonstrations, pages 67–72, Vancouver, Canada, July. Association for Computational Linguistics.

[Kudo and Richardson2018] Kudo, Taku and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 66–71.

[Lewis et al.2020] Lewis, Mike, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7871–7880.

[6] https://docs.djangoproject.com/

[7] https://jmeter.apache.org/

[8] https://xtm.cloud/

[9] https://core.telegram.org/bots

[Liu et al.2020] Liu, Yinhan, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. Transactions of the Association for Computational Linguistics, 8:726–742.

[Martins et al.2022] Martins, Pedro, Zita Marinho, and André FT Martins. 2022. Efficient machine translation domain adaptation. In Proceedings of the 1st Workshop on Semiparametric Methods in NLP: Decoupling Logic from Knowledge, pages 23–29.

[Ott et al.2019] Ott, Myle, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations), pages 48–53.

[Papineni et al.2002] Papineni, Kishore, 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.

[Radford et al.2018] Radford, Alec, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

[Radford et al.2019] Radford, Alec, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

[Saunders2022] Saunders, Danielle. 2022. Domain adaptation and multi-domain adaptation for neural machine translation: A survey. Journal of Artificial Intelligence Research, 75:351–424.

[Stahlberg2020] Stahlberg, Felix. 2020. Neural machine translation: A review. Journal of Artificial Intelligence Research, 69:343–418.

[Tang et al.2020] Tang, Yuqing, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. arXiv preprint arXiv:2008.00401.

[Vaswani et al.2017] Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, and Lukasz Kaiser. 2017. Attention Is All You Need. In 31st Conference on Neural Information Processing Systems (NIPS 2017), pages 5998–6008, Long Beach, CA, USA.