Proceedings of the 5th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/99/20251767

Diversified Research and Analysis of Machine Translation Models and the Development Prospects of Translation for Less-Commonly Spoken Language

Zijun Lan[1,a,*]

[1]*Samuel Ginn College of Engineering, Auburn University, AL, USA*
a. zijunlan59@gmail.com
**corresponding author*

Abstract: This paper begins by addressing issues in machine translation, reviewing the current state of research on translation problems, and summarizing various mainstream machine translation models. It also analyzes the structural characteristics and advantages of each model. Based on the current research landscape, mainstream machine translation models generally underperform in scenarios involving less-commonly spoken languages. Therefore, future research needs to focus on breakthroughs in such scenarios. Preliminary findings from recent studies indicate that advanced techniques such as data augmentation, reinforcement learning, and transfer learning have significantly contributed to addressing these issues. Future research could integrate these theoretical approaches to further enhance the quality of machine translation.

Keywords: Neural networks, language translation problems, machine translation models, transfer learning for less-commonly spoken languages.

1. Introduction

For translation problems, the first requirement is to establish a corpus, as it serves as the foundation for translation and is needed for both model training and testing. Next comes model training. The mainstream machine translation models currently include CNN and NMT. However, CNN has limited use, mostly applied to languages on the verge of extinction, while NMT is also utilized in emotional contexts and is further divided into several types, such as SMT and MNMT. In addition to these two models, new models like the BERT model have emerged.

As time goes on, translation issues are continuously being refined and enhanced, evolving from early machine learning translation to more sophisticated lexical data, and now focusing on the translation of less-commonly spoken languages. While translation problems are not an issue for most of us today, for some regions, language remains an integral part of their culture, making the translation of minority languages a primary challenge.

2. Literature Review

In certain specialized fields, people may need to read reports or articles from other countries, which requires translation. However, certain terms that are less frequently used can lead to translation errors. Antonio L. Lagard [1] proposed “Automatic Post-Editing in an Online Learning Framework,” which

DOI: 10.54254/2755-2721/99/20251767

involves receiving feedback from users after translation and updating the translation in real time. To better evaluate this technology, the authors used three corpora. For the EMEA and Xerox corpora, the "RBMT, Web, and SMT systems achieved improvements of up to 16, 11, and 7 BLEU points, respectively," while in the i3media corpus, which contains more n-gram syntax, the experiments demonstrated the importance of n-gram syntax for OL technology.

Over time, the vocabulary of some languages continues to expand. To prevent future translation issues due to insufficient lexical data, Grandee Lee [2] utilized the "Matrix Language Framework" theory to generate "synthetic code-switching" to convert data. The authors classified and translated texts from two languages using MLF, generating "synthetic code-switching data" during model training with this theory. Experiments showed that perplexity was reduced by 21%, and WER improved by 1.45% when this data was used. However, the experiment mostly focused on model training results and lacked substantial real-world examples for support.

In translation, semantic similarity impacts translation quality. While some mainstream languages have developed the ability to distinguish this, certain languages have not been sufficiently researched in this area due to their complexity. MD. Asif Iqbal [3] selected Bengali for research, a language with significant grammatical and morphological variations. The authors first developed a corpus to build an embedding model and then used embedding technology to compute semantic similarity, followed by a comparison of techniques. The results indicated that "the performance of tFastText embeddings using the CBOW (FT+CBOW) method (ρ=77.28%) outperformed other methods for the Bengali semantic similarity task (77.28%)" (Pearson's correlation (ρ)).

Neural machine translation (NMT) has made significant progress, but for some uncommon sentences, translation remains inadequate. To address this issue, Xuewen Shi [4] proposed an NMT method called "Learning Semantic Knowledge via Transfer of Bilingual Sentence Alignment." First, a recognizer was created to evaluate aligned sentences, where n-gram training was applied. Then, the method improved translation adequacy by incorporating an adversarial training framework and "alignment-aware decoding." The authors conducted experiments on translations of several common languages, and the analysis showed that this method outperformed traditional NMT translations.

Emotion is an essential aspect of translation, helping speakers of different languages understand each other's tone. However, due to differences in emotional distribution across languages, emotional alignment between two languages is necessary. Today, in the pursuit of translation efficiency, emotional translation has been neglected. To address this, Xun Zhu [5] introduced an innovative method called "Improved Code-Switching Emotion Detection." The authors first translated "code-switched text" into two languages and then aligned the two languages in parallel. An encoder was used for adversarial training to retrieve language features. Experimental results showed that the F1 score of translations in some languages improved by more than 1.5%.

Neural machine translation has certain limitations, one of which is the impact of sentence length on translation quality. This is the issue the authors aimed to solve. Yao Huang [6] proposed a method called "Decay Weight Function" to address this. The approach involves first selecting "the longest noun phrase in a sentence, retaining specific markers or core words, and forming a sentence framework with the remaining parts of the sentence." After this, the selected words are translated first, followed by progressively translating to lower levels. Combining this method with a "part-of-speech embedding-based encoding model" significantly improved the performance of both sentiment analysis and named entity recognition tasks. Although this method improved the BLEU score by 0.89 over the baseline, the overall translation quality did not improve much.

"Cuneiform writing is one of the earliest recorded writing systems in human history," but it has been lost over time. To help people understand texts written in cuneiform, Gai Gutherz [7] used convolutional neural networks (CNN) in combination with a neural machine translation (NMT) model. In the NMT model, the methods "Cuneiform to English Task (C2E)" and "Transliteration to

DOI: 10.54254/2755-2721/99/20251767

English Task (T2E)" were used. The results showed that both methods performed similarly but significantly outperformed the baseline model of the translation memory library, with T2E performing even better. The comparable performance of both methods indicates that skipping the transliteration step and translating directly yields higher quality.

Neural machine translation is closely linked to statistical machine translation (SMT), but SMT is gradually falling behind. To improve SMT translation quality, Debajyoty Banik [8] focused on enhancing SMT's emotional context. The authors first trained a neural network to learn and model emotions, then incorporated contextual understanding into the translation process, which increased emotional depth. The study primarily experimented with English and Hindi, and the results showed improved translation accuracy ("4.66 BLEU points, 4.09 LeBLEU points, 4.67 NIST points, 5.71 RIBES points") and a "7.79% retention of emotions."

Multilingual neural machine translation (MNMT) has made significant advances, but it faces challenges with less-commonly spoken languages due to the need for parallel corpora. To address this issue, Yingli Shen [9] explored the use of monolingual data. The authors proposed training the MNMT model (XLM-DM) "in an unsupervised manner" using a "cross-lingual encoder." During the experiments, the MNMT model was initialized with a pre-trained cross-lingual encoder, and two levels of alignment were used to further align the representation space in the MNMT model. The results showed that XLM-DM exhibited strong language and domain transfer capabilities with fewer samples and did not affect the performance of the core tasks during operation.

This article focuses on the study of translating Khasi, a language from the Austroasiatic language family, into English. Due to the limited resources available for this language, research has been minimal. Given the scarcity of resources, a new corpus was first established before translation. The authors, Aiusha Vellintihun Hujon [10], based their research on LSTM, GRU, and transformer models, and divided the experimental data into "tokenized," "non-tokenized," and "subword BPE" categories. The results of the experiment showed that the transformer model performed better in both qualitative and quantitative evaluations and analyses. Additionally, the article applied a "transfer learning method," achieving a BLEU score of up to 58.1 in similar domains and 17.7 in general domains.

Today, the number of users on social platforms continues to grow, with people from various countries, which inevitably leads to conflicts due to differing interests, including language-based attacks. In response, some social platforms are actively addressing these issues. Aya Mousa [11] selected Arabic to detect and classify offensive language. The authors introduced a new BERT model and combined it with several deep learning models and traditional classifiers to form a cascading model. Traditional classifiers were also used to assess performance. The experimental results showed that the cascading model outperformed traditional models in terms of precision, accuracy, recall, and F1 scores, achieving 98.4%, 98.2%, 92.8%, and 98.4%, respectively.

Research on Tamil is relatively scarce globally, but it remains an important language. To recognize its handwritten form, Jayasree Ravi [12] used convolutional neural networks (CNN) for image recognition. To find the optimal method, the authors developed three CNN models and compared them with two popular pre-trained models, VGG-11 and VGG-16. The method used in the experiment involved first classifying handwritten letters, adjusting the images to make them clearer, and finally outputting the results via CNN for better model comparison. The experimental results indicated that the CNN models outperformed the VGG models in terms of accuracy. Among the CNN models, those with fewer "conv2d" layers had higher accuracy, while models with more "epochs" were prone to overfitting.

The authors of this article, Harshita Samota [13], aimed to create a system that could translate Punjabi text into Bharati Braille. To achieve this, they used several neural machine translation (NMT) systems and trained the system on six lac parallel sentences. In these sentences, the system captured and analyzed grammatical structures. To improve translation quality, both grammar-enhanced and

DOI: 10.54254/2755-2721/99/20251767

baseline systems were used. The results showed that the grammar-enhanced MT system outperformed the baseline system in terms of BLEU scores, indicating that the grammar-enhanced system was better. However, due to the limitations of the corpus, the current results remain as such.

3. Discussion

Neural Machine Translation (NMT) is currently the dominant model for machine translation, subdivided into various categories, such as SMT, MNMT, and RNN. Statistical Machine Translation (SMT) conducts statistical analysis on parallel corpora, which requires a large volume of samples to support it. This enables SMT to achieve better accuracy and richer emotional content than other models, but at the cost of lower efficiency. This means that SMT is suited for addressing emotional gaps in translation and can be integrated with other models. Multilingual Neural Machine Translation (MNMT) translates one language into multiple languages. Since it requires establishing one-to-one models to achieve this, its computational efficiency is relatively low. However, as it does not require extensive text data, it is more suitable for smaller languages, and transfer learning is also applicable to this model.

The main challenges we face in machine translation today are related to lesser-used languages. Due to the smaller number of speakers worldwide, research on these languages is significantly more difficult than that for commonly used languages. However, through some research, there have been initial solutions to these challenges. First, regarding the issue of sample size, we can use data augmentation techniques during corpus collection to diversify the data. This involves replacing words and grammar in sentences with synonyms or adjusting the word order, thus enriching the corpus. Additionally, we can introduce more optimized reinforcement learning techniques, such as sentence expansion and abbreviation, to make the translated content and emotions more abundant. Second, transfer learning can be applied. This involves transferring the solution to one problem to a similar problem. For lesser-used languages, we can employ various transfer learning methods. First, we explore how early solutions for translating major languages were transferred to other major languages and then apply this method to lesser-used languages. Finally, we can use samples from languages similar to the lesser-used ones to fill in the gaps. This approach addresses the shortage of sample data for lesser-used languages and reduces the time cost of large-scale training.

4. Conclusion

This article focuses on translation challenges and examines the current state of machine translation technology. It finds that the focus of translation problems has shifted from improving translation quality and efficiency to addressing and supplementing translation for lesser-used languages. We primarily rely on NMT models to address these issues, as they cover a broad range and align with current approaches to solving translation problems. However, even with these advancements, we still encounter difficulties when dealing with lesser-used languages, such as the issue of limited samples—an issue not seen in previous translation challenges. This points to a future direction for development.

Another issue is the insufficient ability to process multimodal data. In real-world translation scenarios, there are often more complex and specific contexts, such as movie dialogues, where the translation depends on the movie's plot at that moment. Therefore, when addressing machine translation problems, it is necessary to consider not only textual data but also to handle multimodal data structures holistically, incorporating information from context, setting, plot, and character development to achieve a translation that is accurate, communicative, and elegant. However, most current mainstream machine translation research focuses only on textual data structures, lacking compatibility with multimodal data structures. This represents a significant area for further development and a feasible direction for improving the effectiveness of machine translation.

DOI: 10.54254/2755-2721/99/20251767

References

[1] *Lagarda, A. L., Ortiz-Martínez, D., Alabau, V., & Casacuberta, F. (2015). Translating without in-domain corpus: Machine translation post-editing with online learning techniques. Computer Speech & Language, 32(1), 109-134.*

[2] *Lee, G., Yue, X., & Li, H. (2019). Linguistically Motivated Parallel Data Augmentation for Code-Switch Language Modeling. In Interspeech (pp. 3730-3734).*

[3] *Iqbal, M. A., Sharif, O., Hoque, M. M., & Sarker, I. H. (2021). Word embedding based textual semantic similarity measure in Bengali. Procedia Computer Science, 193, 92-101.*

[4] *Shi, X., Huang, H., Jian, P., & Tang, Y. K. (2021). Improving neural machine translation with sentence alignment learning. Neurocomputing, 420, 15-26.*

[5] *Zhu, X., Lou, Y., Deng, H., & Ji, D. (2022). Leveraging bilingual-view parallel translation for code-switched emotion detection with adversarial dual-channel encoder. Knowledge-based systems, 235, 107436.*

[6] *Huang, Y., & Xin, Y. (2022). [Retracted] Deep Learning‐Based English‐Chinese Translation Research. Advances in Meteorology, 2022(1), 3208167.*

[7] *Gutherz, G., Gordin, S., Sáenz, L., Levy, O., & Berant, J. (2023). Translating Akkadian to English with neural machine translation. PNAS nexus, 2(5), pgad096.*

[8] *Banik, D. (2023). Sentiment induced phrase-based machine translation: Robustness analysis of PBSMT with senti-module. Engineering Applications of Artificial Intelligence, 126, 106977.*

[9] *Shen, Y., Bao, W., Gao, G., Zhou, M., & Zhao, X. (2024). Unsupervised multilingual machine translation with pretrained cross-lingual encoders. Knowledge-Based Systems, 284, 111304.*

[10] *Hujon, A. V., Singh, T. D., & Amitab, K. (2024). Neural machine translation systems for English to Khasi: A case study of an Austroasiatic language. Expert Systems with Applications, 238, 121813.*

[11] *Mousa, A., Shahin, I., Nassif, A. B., & Elnagar, A. (2024). Detection of Arabic offensive language in social media using machine learning models. Intelligent Systems with Applications, 22, 200376.*

[12] *Ravi, J. (2024). Handwritten alphabet classification in Tamil language using convolution neural network. International Journal of Cognitive Computing in Engineering, 5, 132-139.*

[13] *Samota, H., & Joshi, N. (2024). Improving the Punjabi-Hindi Braille Neural Machine Translation through Syntax Augmentation. Procedia Computer Science, 235, 1489-1497.*