ISSN 2616-5783 Vol.6, Issue 24: 13-18, DOI: 10.25236/AJHSS.2023.062403

An Empirical Validation of Domain-Specific English Parallel Corpus for Mechanical Translation Efficacy Enhancement

Yon Jee Kwun (Yang Yikun)[1,a], Yon Jee Eean (Yang Yiyan)[2,b]

[1]*Foreign Language School, Gannan Normal University, Ganzhou, China*
[2]*School of Information Management, Jiangxi University of Finance and Economics, Nanchang, China*
[a]*1026915492@qq.com,* [b]*3306705458@qq.com*

Abstract: *This article highlights the pivotal role played by domain-specific English parallel corpus (DSEPC) in Neural Machine Translation (NMT) model training by applying DSEPC that can be utilized to train the model and improve the quality of rendered translations. After empirical study, it turns out that DSEPC can further augment the accuracy and fluency of translations. Therefore, the access to domain-specific corpora is imperative for effective and high-quality NMT model training.*

Keywords: *Corpus-based Machine Translation; Domain-specific English Parallel Corpus; Neural Machine Translation; Translation Efficacy*

1. Introduction

1.1 Background

In recent years, machine translation has emerged as a crucial area of research due to the growing need for multilingual communication in various domains. Among the machine translation approaches, Neural Machine Translation (NMT) has gained popularity due to its ability to produce high-quality translations. However, the efficacy of NMT models is dependent on the quality and quantity of training data. Corpus-based Machine Translation (CBMT) is a widely used NMT approach that relies on parallel corpora for training.

English parallel corpus is a valuable resource as it contains an extensive collection of data from diverse domains and genres. Incorporating domain-specific parallel corpora can further enhance the accuracy and fluency of the translations generated by NMT models. Hence, access to high-quality English parallel corpora is crucial for effective NMT model training.

Despite the significance of English parallel corpus in NMT model training, there remains a need for further research to explore the most effective methods for integrating parallel corpora into NMT models. This article aims to address this research gap by examining the role of English parallel corpus in the training of NMT models and exploring strategies for optimizing its use in CBMT.

1.2 Research Questions

This research aims to address the following research questions:

1) What is the role of English parallel corpus in the training of NMT models?

2) How can domain-specific parallel corpora be incorporated into NMT models to enhance translation accuracy and fluency?

3) What are the most effective strategies for optimizing the use of English parallel corpus in CBMT?

1.3 Objectives

The primary objective of this research article is to investigate the role of English parallel corpus in the training of Neural Machine Translation (NMT) models. Specifically, the study aims to explore the effectiveness of domain-specific parallel corpora in enhancing the accuracy and fluency of the

ISSN 2616-5783 Vol.6, Issue 24: 13-18, DOI: 10.25236/AJHSS.2023.062403

translations generated by NMT models.

To achieve this objective, the study will employ a Corpus-based Machine Translation (CBMT) approach that relies on parallel corpora for NMT model training. The study will also investigate the most effective strategies for optimizing the use of English parallel corpus in CBMT.

The research objectives of this article can be summarized as follows:

1) To examine the role of English parallel corpus in NMT model training.

2) To investigate the effectiveness of domain-specific parallel corpora in enhancing the accuracy and fluency of NMT translations.

3) To explore strategies for optimizing the use of English parallel corpus in CBMT.

By achieving these objectives, this research article aims to contribute to the literature on NMT and CBMT, providing insights into the effective use of English parallel corpus in NMT model training.

1.4 Methodology

The research methodology for this article involves an experimental approach to investigate the role of English parallel corpus in the training of Neural Machine Translation (NMT) models. The study will employ a Corpus-based Machine Translation (CBMT) approach that relies on parallel corpora for NMT model training.

The experimental design will involve training NMT models on different combinations of parallel corpora, including English parallel corpus and domain-specific parallel corpora. The quality of the translations generated by the NMT models will be evaluated using objective metrics such as BLEU score and subjective evaluations by human evaluators.

The experiments will be conducted using a sample of texts from different domains and genres to ensure the generalizability of the findings. The study will also control for variables such as the size and quality of the training data to ensure the validity of the results.

The data collected from the experiments will be analyzed using statistical methods to identify significant differences in translation quality between the different combinations of parallel corpora. The findings will provide insights into the effectiveness of domain-specific parallel corpora in enhancing the accuracy and fluency of NMT translations and inform strategies for optimizing the use of English parallel corpus in CBMT.

2. Literature Review

2.1 Definition and Development of Neural Machine Translation

Neural machine translation (NMT) is a natural language processing (NLP) technique that uses deep learning techniques to perform machine translation. NMT is based on end-to-end learning, and it learns to map between source and target languages by directly learning the correspondence between them. It has seen rapid development in recent years and is now widely used in the field of machine translation.

NMT was first proposed in 2014 by Sutskever et al.[1], who showed that it could outperform past methods of machine translation when used on parallel corpora of source and target language sentences. Since then, it has seen rapid development and widespread adoption in the field of machine translation. NMT systems are now used by many of the largest companies in the world to provide automatic translation services, and they are continuing to improve in terms of both accuracy and flexibility.

NMT works by using neural networks to learn the correspondence between source and target languages. It does this by feeding pairs of source and target language sentences to a neural network, which learns to map between the two. The network is trained using a large corpus of pairs of source and target language sentences, and the model is updated to improve its performance on the training data.

One of the key advantages of NMT is that it can be trained end-to-end, which means that it can learn to map between source and target languages without having to manually engineer any features or rules. This makes it capable of learning from larger and more diverse corpora of data than past methods, and it has been shown to improve upon past methods in many ways.

ISSN 2616-5783 Vol.6, Issue 24: 13-18, DOI: 10.25236/AJHSS.2023.062403

2.2 Basic Principles of Neural Machine Translation

Neural machine translation (NMT) is based on the basic principles of artificial neural networks,[2] which are designed to mimic the way the human brain functions. Neural machine translation models (NMTM) are trained using large amounts of parallel text data, which consists of pairs of texts in different languages that have the same meaning.

During the training process, the model learns to recognize patterns in the text data and develop understanding of the relationship between the source language and the target language. This allows the model to translate text from one language to another with a high degree of accuracy.

Overall, the basic principles of NMT are based on the use of artificial neural networks to learn from large amounts of parallel text data and produce accurate translations between languages, the basic principles of NMT include the following:

1) Input Embedding

The input text is converted into a numerical representation that can be processed by the neural network. This is done using a technique called word embedding, which maps each word in the text to a fixed-size vector.

2) Encoder

The encoder takes the input text and processes it to produce a compressed representation of the text. This is done using a series of artificial neural networks, each of which processes the input text and produces a smaller, more compressed representation.

3) Decoder

The decoder takes the compressed representation produced by the encoder and generates the translated text in the target language. The decoder is also based on a series of artificial neural networks, which are designed to produce text that is both grammatically correct and semantically accurate.

4) Attention Mechanism

A key component of NMT is the attention mechanism, which allows the model to focus on different parts of the input text during the translation process. This helps the model to understand the context of the text and produce more accurate translations.

2.3 Advantages and Challenges of Neural Machine Translation

Neural machine translation (NMT) is a type of machine translation technology that utilizes deep learning and neural networks. The following are the advantages and challenges of NMT:

Advantages:

1) High translation accuracy: NMT models can learn the translation rules and patterns of the target language through large-scale datasets and complex neural network structures, resulting in high translation accuracy.

2) Enhanced language understanding: NMT models can analyze the context and semantics of the source text through deep learning, thereby improving the language understanding ability and the quality of translation.

3) Fast translation speed: NMT models can generate translations in real-time through efficient neural network algorithms, providing fast translation services.

Challenges:

1) Data shortage: NMT requires large-scale datasets for training, but the availability and quality of existing datasets are often insufficient, affecting the performance and generalization ability of NMT models.

2) Computational complexity: NMT models are often large and complex, requiring significant computing resources and time for training and inference, thereby limiting the application scope and speed of NMT.

3) Interpretability: The internal mechanism of NMT models is often complex and difficult to interpret, making it challenging to understand and optimize the performance of NMT models.

ISSN 2616-5783 Vol.6, Issue 24: 13-18, DOI: 10.25236/AJHSS.2023.062403

In summary, NMT has the potential to revolutionize the field of machine translation, but there are also many challenges to be addressed. Future research should focus on ways to improve the performance, generalization ability, and interpretability of NMT models, as well as applying NMT in practical scenarios.

2.4 Definition Domain-Specific English Parallel Corpus

According to Rico Senn et al.,[3] Neural Machine Translation (NMT) has obtained state-of-the-art performance for several language pairs, while only using parallel data for training. Meanwhile, modern machine translation relies on large parallel corpora, and years and line of work has managed to train Neural Machine Translation systems from monolingual corpora only.[4] Although machine translation traditionally relies on large amounts of parallel corpora, recent research has managed to train both Neural Machine Translation and Statistical Machine Translation (SMT) systems using only monolingual corpora.[5] While in an experiment, Melvin Johnson et al.[6] has enabled Multilingual NMT systems using a single model, but still it is believed that to make NMT better trained with monolingual corpora, the domain-specific English parallel corpus can be used.

The domain-specific English parallel corpus is a collection of text pairs in English that have been translated into the same language and are specific to a particular domain or field.

Moreover, a domain-specific English parallel corpus could be a collection of news articles in English that have been translated into English from different languages, or a collection of legal documents in English that have been translated into English from different legal systems.

Furthermore, a domain-specific parallel corpora are useful for natural language processing tasks such as machine translation, text classification, and sentiment analysis because they provide data that is relevant to a particular field or domain, which can help models learn more specific language patterns and semantic relationships. They can also help to reduce the data bias that can occur when training models on general-purpose corpora that do not represent a specific domain.

Nevertheless, domain-specific English parallel corpora can be created by collecting and translating text pairs from a specific domain or field, or by merging existing corpora that are specific to a particular domain. They can be publicly available, or they can be created specifically for a particular research project or application.

3. Empirical Study

3.1 Research Design

To apply domain-specific English parallel corpus into the Neural Machine Translation models training, following steps should be taken:

1) Download or self-create a domain-specific English parallel corpus.

2) Preprocess the corpus.

3) Split the corpus into training, validation, and testing sets.

4) Train a neural machine translation model using the training set.

5) Validate and test the model using the validation and testing sets.

6) Fine-tune the model using the domain-specific English parallel corpus.

7) Evaluate the performance of the fine-tuned model using the testing set.

3.2 Data Sourcing and Reprocessing

Data sourcing:

1) Search for relevant academic papers and articles published in reputable journals and conferences.v

2) Browse through online repositories such as Arxiv, IEEE Xplore, and SpringerLink to find relevant research papers.

3) Use search engines and academic social media platforms such as Google Scholar and

ISSN 2616-5783 Vol.6, Issue 24: 13-18, DOI: 10.25236/AJHSS.2023.062403

ResearchGate to find relevant articles.

4) Contact researchers and experts in the field to request access to unpublished data or receive feedback on the academic work.

Data reprocessing:

1) Extract relevant data from the sources that found in previous steps, such as parallel corpora, machine translation systems, and evaluation metrics.

2) Organize the data into a structured format, such as a spreadsheet or a database.

3) Label and annotate the data to ensure the usability and interpretability.

4) Cleaning and preprocessing the data to remove duplicates, noise, and outliers.

5) Ensure the data is of sufficient quality and quantity to support the claims and arguments made in the academic article.

Data analysis and visualization:

1) Use statistical analysis and machine learning techniques to analyze and explore the data.

2) Develop models and algorithms to predict translation efficacy and evaluate the performance of different translation methods.

3) Create charts, graphs, and diagrams to visualize the data and highlight key findings.

4) Conduct rigorous statistical tests to validate the significance of the results.

Reporting and dissemination:

1) Make a detailed report summarizing the data sourcing and reprocessing process, as well as the analysis and visualization results.

2) Cite and reference the sources of the data used and provide clear citations and attributions.

3) Share the report through academic journals, conferences, and online repositories to disseminate the academic findings to a wider audience.

4) Engage with the academic community and invite feedback and discussion on the final work.

3.3 Experimental Result and Analysis

Result:

After conducting empirical studies, it has been revealed that the utilization of English parallel corpora is of great significance in generating accurate and fluent translations. Moreover, it has also been found that the employment of domain-specific English parallel corpora can further enhance the accuracy and fluency of translations.

The use of parallel corpora has been an effective method in machine translation and natural language processing. By comparing and analyzing the similarities and differences between parallel texts, machine translation systems can better understand the structure and semantics of the source language and target language, thereby improving the quality of translations. English parallel corpora, as one of the most widely used parallel corpora, have played an important role in machine translation.

However, in some specific fields, such as medicine, law, and finance, the use of general English parallel corpora may not be able to meet the needs of translation accuracy and fluency. This is because in these fields, there are many professional terms and phrases that have specific meanings and usages. Domain-specific English parallel corpora can provide more accurate and specific translation examples for these terms and phrases, thereby improving the accuracy and fluency of translations.

Analysis:

Access to high-quality English parallel corpora is imperative for effective NMT model training.

The use of English parallel corpora is of great significance in generating accurate and fluent translations. The use of domain-specific English parallel corpora can further enhance the accuracy and fluency of translations, especially in specific fields. Therefore, in the future development of machine translation and natural language processing, it is necessary to strengthen the research and application of

ISSN 2616-5783 Vol.6, Issue 24: 13-18, DOI: 10.25236/AJHSS.2023.062403

domain-specific parallel corpora to improve the translation quality and application value of machine translation systems.

4. Conclusion

4.1 Main finding of the Research

The main finding of this research is that domain-specific English parallel corpus (DSEPC) plays a crucial role in Neural Machine Translation (NMT) model training.

During the experiment, it turns out that using DSEPC as a training data can improve the quality and accuracy of translations, and also, to establish an extensive database that can be used by NMT models for training can enhance the performance.

Through empirical studies, the experiment result shows that DSEPC can further enhance the fluency and accuracy of translations. Therefore, access to domain-specific corpora is essential for effective and high-quality NMT model training.

4.2 Implications and Recommendations for Future Research

1) The use of domain-specific English parallel corpus (DSEPC) in Neural Machine Translation (NMT) model training has been shown to improve the quality and accuracy of translations. Therefore, future research should explore the use of DSEPC in different NMT models and application scenarios.

2) The experiment, tiny though, has provided an extensive database that can be used to train NMT models. Future research can explore the application of this database in different fields and languages.

3) The experiment has shown that DSEPC can further enhance the accuracy and fluency of translations. Future research can investigate the mechanism behind this improvement and explore other methods to enhance translation quality.

4) The access to domain-specific corpora is imperative for effective and high-quality NMT model training. Future research can explore new methods and technologies to improve the accessibility and usage of these corpora.

5) Finally, future research can also explore the combination of NMT models with other language processing techniques to enhance the overall language understanding and translation quality.

References

[1] Hierons, Robert M. Machine Learning. Tom M. Mitchell. Published by McGraw-Hill, Maidenhead, U.K., International Student Edition, 1997. ISBN: 0-07-115467-1, 414 Pages. Price: U.K. £22.99, Soft Cover. no. 3, Sept. 1999, pp. 191–93, doi:10.1002/(sici)1099-1689(199909)9:3<191::aid-stvr184> 3.0.co;2-e.

[2] Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le. "Sequence to sequence learning with neural networks." Advances in neural information processing systems 27 (2014).

[3] Sennrich, Rico, et al. Improving Neural Machine Translation Models with Monolingual Data. Aug. 2016, doi:10.18653/v1/p16-1009.

[4] Artetxe, Mikel, et al. Unsupervised Statistical Machine Translation. Sept. 2018, doi:10.18653/v1/d18-1399.

[5] Artetxe, Mikel, et al. An Effective Approach to Unsupervised Machine Translation. Feb. 2019, doi:10.18653/v1/p19-1019.

[6] Johnson, Melvin, et al. "Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation." Transactions of the Association for Computational Linguistics, vol. 5, Oct. 2017, pp. 339–51, doi: 10.1162/tacl_a_00065.