COSC 5P84 Final Project Report

Madeline Janecek
mj17th@brocku.ca

Abstract

In their recent publication, *CipherDAug: Ciphertext based Data Augmentation for Neural Machine Translation*, authors Kambhatla et al. introduce CipherDAug, a multi-view Neural Machine Translation (NMT) model that leverages encrypted versions of the original input to improve performance. Although the authors demonstrate how CipherDAug outperforms other state-of-the-art NMT models, they do not provide a thorough analysis of their data augmentation technique's contribution to the observed improvement. To address this gap, this paper presents a focused investigation of CipherDAug's translation capabilities, where the encrypted text is replaced by a wider range of supplementary data sources. This testing is done with an challenging machine translation dataset to test the model's generalizablity. The results of this study indicate that CipherDAug's success is, in part, attributable to the encryption-based data augmentation approach. However, using other redundant representations of the input text are found to produce similar or improved results.

1 Introduction

Neural Machine Translation (NMT) is the current state-of-the-art machine translation (MT) method. Broadly speaking, NMT utilizes artificial neural networks (ANNs) to convert text from a source language into a target language (Tan et al., 2020). One of the primary drawbacks of NMT models is that their efficacy is heavily dependent on the availability of large-scale high-quality training data. This poses a significant challenge, especially for low-resource language pairs where such data is often not readily available (Koehn and Knowles, 2017; Tan et al., 2020). To address this limitation, researchers have explored various data augmentation techniques to fully exploit the available training resources (Ranathunga et al., 2023).

Despite the widespread use of data augmentation techniques in NMT, the underlying reasons why they improve the performance of NMT models are not yet well understood. Instead, the evaluation of data augmentation approaches for NMT are typically based solely on manual or automated analyses of translation results. To address this issue, this project entailed going beyond these surface level evaluations.

Specifically, several tests were conducted to take a closer look at CipherDAug, a NMT model that produces high quality translations using an encryption-based data augmentation technique (Kambhatla et al., 2022). As a multi-view learning model, CipherDAug examines the input data simultaneously with encrypted versions of the input. For this project, several versions of CipherDAug were trained using different redundant representations of the data, each of which were chosen to investigate the following research questions:

- How much of the CipherDAug's success can be attributed to the model itself, as opposed to the data augmentation technique?
- Is the authors' claim that the supplementary input should be non-overlapping and lexically diverse justified?
- Can the model's translation capabilities be improved using an alternative supplementary data source?
- Would the CipherDAug model be as successful with a more challenging dataset?

The rest of this paper is organized as follows: Section 2 provides an overview of related work, Section 3 describes the project's methodology, Section 4 presents the experimental results that are discussed in Section 5, and finally Section 6 concludes the paper.

2 Background and Related Work

2.1 Neural Machine Translation

In comparison to earlier Statistical Machine Translation (SMT) approaches, which employ feature engineering to model parallel corpora as a probabilistic framework (Osborne, 2010), NMT models have demonstrated higher performance in terms of accuracy, adequacy, and fluency (Stasimioti et al., 2020). However, despite being the current state-of-the-art approach, there still remains a lack of understanding as to how and why NMT models produce their superior translation results (Tan et al., 2020).

To enhance the performance of NMT models some approaches utilize multi-view learning, which involves leveraging redundant views of the input data to enhance machine learning models (Xu et al., 2013). For instance, some works treat different layers in their transformer architecture as distinct views, which are subsequently merged to generate the final translation (Wang et al., 2020; Yang et al., 2023). CipherDAug is another example of an NMT model that exploits multi-view learning, as it treats enciphered versions of the input as different views. The goal of this project is to examine the impact these views have on the overall quality of the translation.

2.2 CipherDAug

CipherDAug is a multi-view NMT approach that was designed to leverage a novel encryption-based data augmentation technique (Kambhatla et al., 2022). It starts by creating alternate views of the source language through ROT-k encryption. Although ROT-k ciphers are not considered secure due to their simplicity, they are well-suited for generating consistent alternative representations of the input data. The resulting ciphertext (c) are given to the CipherDAug model in tandem with the original input (i). The model is then trained to generate an appropriate target for i as well as c by minimizing a combined loss function.

The overall CipherDAug approach has several benefits, including that it does not require any external data, preserves the distributional features of the original input, and is shown to improve translation results. Kambhatla et al. provide a more detailed overview of CipherDAug's design and results in the original CipherDAug paper (Kambhatla et al., 2022).

2.3 NMT Data Augmentation

Other data augmentation techniques for NMT can be generally divided into three categories (Ranathunga et al., 2023). The first category consists of methods that generate new training examples by substituting words or phrases with semantically similar alternatives (Sennrich et al., 2016; Wu et al., 2021). The second category, known as back-translation, involves adding NMT-translated sentences into the training data (Xia et al., 2019). Finally, parallel data mining techniques involve extracting translation equivalents from comparable corpora (Ranathunga et al., 2023).

While the goal of data augmentation techniques is to improve the quality and diversity of training data, these methods may unintentionally degrade its semantic quality, thereby rendering it unsuitable for NMT purposes (Kambhatla et al., 2022). For instance, word replacement approaches may misinterpret context, leading to the production of fragmented training data. Similarly, using back-translation with inadequate models may yield semantically poor results that impact subsequent training. CipherDAug's use of redundancy eliminates the need for generating new data, so it does not face the same issues as those seen with the aforementioned methods.

3 Methodology

To evaluate the impact that the ciphertexts had on the overall performance, several CipherDAug models were trained and assessed using a broader range of supplementary data sources. The CipherDAug model architecture was kept constant[1], with only the supplementary data generation method being modified. Each model was given two supplementary data views, as this was shown to lead to optimal results in the original work (Kambhatla et al., 2022).

The data types given to the models include the ciphertext, duplicate, reverse, shuffle, random, blank, and random data types. These options were specifically selected to test some aspect of the CipherDaug's authors' claim that their success is due to the ciphertexts' semantically similar yet lexically diverse representation of the input. Figure 1 presents a visual overview of each alternative data type, while more detailed descriptions are included in the subsequent subsections.

[1]CipherDAug's implementation has been made publicly available https://github.com/protonish/cipherdaug-nmt

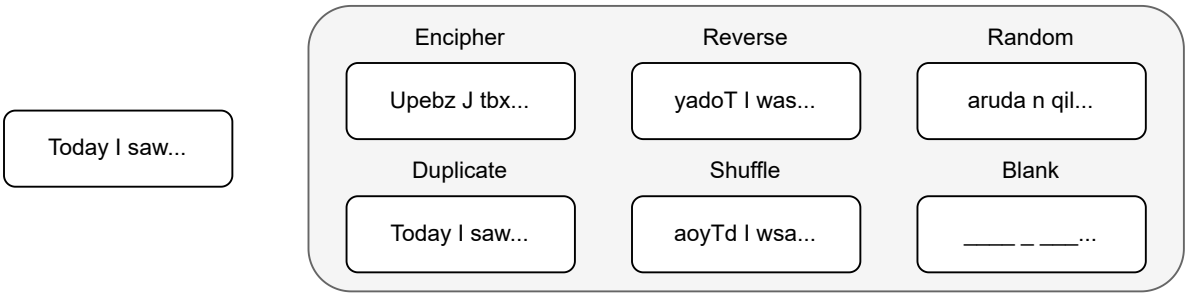

Figure 1: A visual overview of the six different supplementary data types examined in this work.

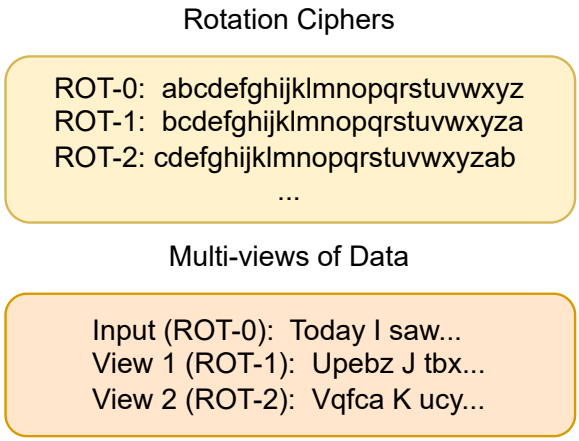

Figure 2: How ROT-k encipherment is used to create two redundant views of the input data.

3.1 Ciphertext

For this data type, the supplementary input data is encrypted using a ROT-k cipher, exactly as the original CipherDAug approach. ROT-k ciphers work by substituting each letter in a text with the letter k positions down the alphabet (see Figure 2). For the purposes of this study, the ciphertext approach serves as a baseline for evaluating the effectiveness of the other supplementary data generation techniques.

3.2 Duplicate

With the duplicate data type, the supplementary views are exact copies of the original input. This approach was specifically selected to evaluate the claim that redundant views must be lexically diverse in order to improve translation quality. Furthermore, as the duplicate data does not introduce any new information, it can help estimate the extent to which CipherDAug's success can be attributed to the data augmentation method versus the model's architecture.

3.3 Reverse

The reverse data approach involves flipping the order of the characters within each word in the input text. This method differs from the duplicate data method, as it does modify the original input text. However, all supplementary views generated with the reverse data approach are identical to each other. This allows for further testing of whether the model can effectively utilize diverse views in the learning process.

3.4 Shuffle

For this supplementary data representation method, the characters of each word in the input text are randomly rearranged. Unlike the previous methods, this approach results in a different representation of the input each time it is applied. This method is meant to test whether a consistent alternative representation of the input data is essential to the model's effectiveness, or if slight variations in the data would still produce high quality translations.

3.5 Blank

For this data generation method, each character in the input text is replaced with an underscore ('_'). The intention of this approach is to provide the NMT model with as little additional information as possible. The reasons for picking the underscore are twofold. Firstly, the supplementary data had to be the same length as the original input, so discarding the text all together was not an option. Secondly. while whitespace would be the closest approximation to a true blank character, it would be automatically filtered out by the NMT model. Using the underscore fulfilled the requirement of being a recognized character, while still conveying as little meaning as possible.

Supplementary Data	BLEU Score
Ciphertext	59.4401
Duplicate	60.4400
Reverse	**60.5201**
Shuffle	60.0900
Blank	11.5400
Random	10.5500

Table 1: The BLEU score of each model determined using the CoGnition dataset.

3.6 Random

For this case, each character in the input was replaced with a randomly selected character. The goal was to provide the model with completely inconsistent and uninformative data, with the expectation being that its performance would be worse than what was observed with the blank data. Any success seen with the random data can be attributed to the model itself, not the data augmentation approach.

4 Results

Each of the experiments were conducted within an Ubuntu 22.04 environment with CUDA 11.7, Python 3.8.10, and PyTorch 2.0.0. All of the data and code needed to reproduce the experimental results has been made publicly available[2].

4.1 Data

All of the models were trained and tested using the CoGnition dataset[3], which is an English-Chinese parallel corpus designed to test a NMT model's robustness and compositional generalization (Li et al., 2021). This dataset has been shown to challenge models that excel with the standard IWSLT benchmark datasets, which were the primary datasets Kambhatla et al. used to initially evaluate CipherDAug (Kambhatla et al., 2022).

4.2 BLEU Score

Table 1 displays the BLEU scores (Papineni et al., 2002) of each CipherDAug model, which were utilized to compare their performance. Notably, the reverse data type model achieved the highest BLEU score of 60.5201, although it is worth mentioning that the duplicate, shuffle, and ciphertext data types led to comparable results.

[2]https://github.com/janecekm/CipherDAugTesting
[3]https://github.com/yafuly/CoGnition

Supplementary Data	BLEU Score
Duplicate	**35.1001**
Reverse	33.8500
Shuffle	32.1200
Random	6.6901

Table 2: The BLEU score of each model determined using the IWSLT14 De↔En dataset.

To further investigate the effectiveness of different data representation types, some of the experiments were repeated using the larger IWSLT14 De↔En dataset (Cettolo et al., 2014). The BLEU scores of these models are shown in Table 2. Among these models, the one employing the duplicate data type achieved the highest BLEU score of 35.1001.

5 Discussion

The experimental results can be considered from two distinct perspectives. Firstly, the performance of the model when presented with adversarial data, such as the blank and random data types (Section 5.1). Secondly, the model's response to legitimate alternative views of the original input, including the ciphertext, duplicate, reverse, and shuffle data types (Section 5.2).

5.1 Detrimental Data

The results of this study demonstrate that the performance of a CipherDAug model is significantly impaired by the blank and random data types, thus confirming that the data augmentation approach plays a crucial role in generating high quality translations. Notably, the random data type had a more pronounced adverse effect on the model's performance than the blank data type. These results support the notion that the inclusion of nonsensical information can be more detrimental to the model's performance than having no information at all. This highlights the importance of ensuring the quality of the supplementary data to achieve optimal translation results.

5.2 Meaningful Data

Upon examining the results of the CoGnition dataset, there was no significant variation in BLEU scores between the ciphertext, duplicate, reverse, and shuffle data types. With the IWSLT14 dataset there was some more variability, however the results were still very close. These findings suggest

that the effectiveness of the models was not reliant on the particular format of the supplementary data views but rather on the uniformity of word representation. The shuffle data type, which introduced some randomness into the data views, consistently yielded the weakest results out of these four data generation techniques, further substantiating this conclusion.

It is worth noting that the duplicate data type, which required no modification of the input data, yielded some of the best results. This indicates that even straightforward, cost-effective, and computationally efficient techniques such as duplicating the input data can result in substantial improvements in NMT model performance. In short, the encryption process itself does not appear to have a substantial role in CipherDAug's success.

6 Conclusion

In conclusion, this study evaluated the impact of various supplementary data types on the performance of the CipherDAug model. The results show that the quality of the supplementary data has a significant impact on the model's translation results, with the blank and random data types having a substantial negative impact. Moreover, the results indicate that the reverse data type resulted in the highest BLEU score with the CoGnition dataset, while the duplicate data type was the most effective with the larger IWSLT14 De↔En dataset. In general, it is recommended that this type of testing be more widely adopted when developing NMT data augmentation approaches to better understand their contributions to model performance.

This study presents several potential directions for future research. For example, the promising results obtained with the duplicate data type suggest that a multi-view NMT model utilizing copies of the input data could be developed. The experiments in this study utilized configurations optimized for the ciphertext data type, a tailored model leveraging the duplicate data type has the potential of improving upon CipherDAug in terms of both performance and complexity.

References

Mauro Cettolo, Jan Niehues, Sebastian Stüker, Luisa Bentivogli, and Marcello Federico. 2014. Report on the 11th IWSLT evaluation campaign. In *Proceedings of the 11th International Workshop on Spoken Language Translation: Evaluation Campaign*, pages 2–17, Lake Tahoe, California.

Nishant Kambhatla, Logan Born, and Anoop Sarkar. 2022. CipherDAug: Ciphertext based data augmentation for neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 201–218, Dublin, Ireland. Association for Computational Linguistics.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. Association for Computational Linguistics.

Yafu Li, Yongjing Yin, Yulong Chen, and Yue Zhang. 2021. On compositional generalization of neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4767–4780, Online. Association for Computational Linguistics.

Miles Osborne. 2010. *Statistical Machine Translation*, pages 912–915. Springer US, Boston, MA.

Kishore Papineni, 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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Surangika Ranathunga, En-Shiun Annie Lee, Marjana Prifti Skenduli, Ravi Shekhar, Mehreen Alam, and Rishemjit Kaur. 2023. Neural machine translation for low-resource languages: A survey. *ACM Comput. Surv.*, 55(11).

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Maria Stasimioti, Vilelmini Sosoni, Katia Kermanidis, and Despoina Mouratidis. 2020. Machine translation quality: A comparative evaluation of SMT, NMT and tailored-NMT outputs. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 441–450, Lisboa, Portugal. European Association for Machine Translation.

Zhixing Tan, Shuo Wang, Zonghan Yang, Gang Chen, Xuancheng Huang, Maosong Sun, and Yang Liu. 2020. Neural machine translation: A review of methods, resources, and tools. *AI Open*, 1:5–21.

Qiang Wang, Changliang Li, Yue Zhang, Tong Xiao, and Jingbo Zhu. 2020. Layer-wise multi-view learning for neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4275–4286, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Xueqing Wu, Yingce Xia, Jinhua Zhu, Lijun Wu, Shufang Xie, Yang Fan, and Tao Qin. 2021. mixSeq: A simple data augmentation methodfor neural machine translation. In *Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT 2021)*, pages 192–197, Bangkok, Thailand (online). Association for Computational Linguistics.

Mengzhou Xia, Xiang Kong, Antonios Anastasopoulos, and Graham Neubig. 2019. Generalized data augmentation for low-resource translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5786–5796, Florence, Italy. Association for Computational Linguistics.

Chang Xu, Dacheng Tao, and Chao Xu. 2013. A survey on multi-view learning.

Jian Yang, Yuwei Yin, Liqun Yang, Shuming Ma, Haoyang Huang, Dongdong Zhang, Furu Wei, and Zhoujun Li. 2023. Gtrans: Grouping and fusing transformer layers for neural machine translation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:1489–1498.