Cross-Lingual Data Augmentation Techniques: Insights from Multilingual Back Translation

Harini Champooranan[1] and Solomon Ubani[#]

[1]Coppell High School, USA
[#]Advisor

ABSTRACT

This paper investigates the effectiveness of utilizing multiple chains of back translation compared to the traditional method of single-chain back translation for enhancing data diversity in natural language processing (NLP). We explore how multiple rounds of translation and back translation across different languages contribute to enriching the training dataset with diverse linguistic variations. We evaluate the effectiveness of multilingual back translation in achieving better data diversity by reporting the BLEU scores of different back translation techniques. Additionally, we investigate the impact of using languages from different language families and the resulting effect on the diversity of data. Our findings highlight the importance of leveraging multiple chains and multiple language families of back translation for augmenting datasets and provide insights for future research and advancement in data augmentation techniques for NLP.

Introduction

Data augmentation has emerged as a crucial technique in machine learning to artificially increase the size and diversity of a dataset without actually collecting new data. This is achieved by applying various transformations to existing data. Given rapid advancements in machine learning and artificial intelligence, this process is crucial for diversifying training datasets, improving model performance, and preventing overfitting (models memorizing training data without learning underlying patterns), particularly in contexts where data collection is limited or costly.

In the context of natural language processing, back translation [1] is a technique that can be utilized to achieve data augmentation. This iterative process involves translating previously translated text, known as the target language, back to its original language, the source language. This iterative process can introduce subtle variations in phrasing and word choice, allowing for greater diversity in text and hence better training data for downstream natural language processing tasks. This diversity is crucial for developing generalizable NLP models capable of handling a wide range of expressions.

To illustrate the back translation process, consider the following sentence in English (the source language): “The quick brown fox jumps over the lazy dog.” When this is translated to, for example, French (the target language), it produces: “Le renard brun rapide saute par-dessus le chien paresseux.” Translating the French sentence back to English produces: “The fast brown fox jumps over the lazy dog.” In this example, the final sentence is the back-translated sentence. It conveys the same message but has variations such as “fast” in the resulting translated sentence, as opposed to “quick” used in the original sentence. Variations such as this on a larger scale can be valuable for training language processing models as no new data had to be collected.

This research explores approaching data augmentation for NLP processing through multilingual back translation. Unlike traditional angle-chain back translation methods that utilize only one intermediary language, this approach uses multiple languages from diverse language families during the back translation process. This is done by the introduction of multiple “hops” across languages from families such as Indo-European, Sino-

Tibetan, and others. This way, we are able to capture a broader spectrum of linguistic variations and phrasings that may be altered differently when compared to single-language translations. This research also compares scenarios where all intermediary languages belong in the same language family versus cases where languages from varied language families are incorporated. This approach aims to understand the impact of language family choice on data diversity. Moreover, this research analyzes the effect of increasing the number of translation "hops" beyond the traditional two-hop approach (source -> intermediary -> target). By introducing three, four, or more intermediary languages in the back translation chain, we explore if additional hops contribute to increased data diversity. The process is done by selecting a random sample of 2000 sentences from the SST-2 dataset [2]. For each sentence in the sample, the sentence is translated from English to different intermediary languages, and finally back to English. This final translation back to English is the back-translated sentence. After the back translation process is complete, the BLEU score [3] is calculated between the original sentences and the resultant sentences to evaluate how similar they are. The BLEU score is calculated for each pair of original and resultant sentences, and then the average is found among each of the individual BLEU scores. The final BLEU score is then displayed, and lower score suggesting greater diversity in the data.

Methodology

Research Question 1

Do we achieve more diversity in training data utilizing multiple chains of back translation compared to the traditional method of using one-chain back translation?

Research Question 2

Do we achieve more diversity in training data utilizing languages from different language families for back translation compared to using languages from the same language family?

Data Acquisition

In this study, we assess the data augmentation approach using the widely recognized Stanford Sentiment Treebank (SST-2) dataset, a benchmark in natural language understanding commonly employed by researchers to evaluate various data augmentation techniques. The SST-2 dataset comprises movie reviews extracted from the Rotten Tomatoes website, with each review instance annotated with its sentiment polarity, categorized as either positive or negative.

Back Translation Pipeline

To facilitate the back translation process, we implemented a pipeline utilizing Google Translate, a state-of-the-art translation model. The source and target language were always the English language, while the intermediary language varied depending on the specific experiment being conducted. This was done to explore the impact of different language combinations on the resulting data diversity.

Experimental Setup

We conducted experiments to compare the effectiveness of multiple chains of back translation against the traditional single-chain approach. These experiments were designed to systematically investigate the impact of varying the number of translation "hops" and the language family combinations used in the back translation process.

Baseline (Two-Hops)
English -> Indo-European -> English
English -> Sino-Tibetan -> English
English -> Niger-Congo -> English
English -> Afro-Asiatic -> English
English -> Austronesian -> English

Three-Hops Back Translation (Same Language Family)
English -> Indo-European-> Indo-European -> English
English -> Sino-Tibetan -> Sino-Tibetan -> English
English -> Niger-Congo -> Niger-Congo -> English
English -> Afro-Asiatic -> Afro-Asiatic -> English
English -> Austronesian -> Austronesian -> English

Three-Hops Back Translation (Different Language Family)
English -> Indo-European-> Sino-Tibetan -> English
English -> Sino-Tibetan-> Niger-Congo -> English
English -> Niger-Congo -> Afro-Asiatic -> English
English -> Afro-Asiatic-> Austronesian -> English
English -> Austronesian -> Niger-Congo-> English

Four-Hops Back Translation (Same Language Family)
English -> Indo-European-> Indo-European-> Indo-European -> English
English -> Sino-Tibetan -> Sino-Tibetan -> Sino-Tibetan -> English
English -> Niger-Congo -> Niger-Congo -> Niger-Congo -> English
English -> Afro-Asiatic -> Afro-Asiatic -> Afro-Asiatic -> English
English -> Austronesian -> Austronesian -> Austronesian -> English

Four-Hops Lingual Back Translation (Different Language Family)
English -> Indo-European-> Sino-Tibetan-> Niger-Congo -> English
English -> Sino-Tibetan-> Niger-Congo -> Afro-Asiatic -> English
English -> Niger-Congo -> Afro-Asiatic-> Austronesian -> English
English -> Afro-Asiatic-> Austronesian -> Sino-Tibetan--> English
English -> Austronesian -> Niger-Congo -> Sino-Tibetan--> English

In our experiments, languages from each language family were selected at random.

Evaluation Metric

To evaluate the effectiveness of the different data augmentation techniques, we employed the BLEU (Bilingual Evaluation Understudy) score, a widely adopted metric for assessing the quality of the machine-translated text. The BLEU score measures the similarity between the original training data and the augmented data by comparing the n-gram overlap between them and one or more human reference translations. The BLEU score ranges from zero to one, where a higher score indicates a better match between the training data and the augmented data. However, in the context of data augmentation, the goal is to increase the diversity of the data rather than to match the original data. Therefore, a lower BLEU score is desired as it signifies greater diversity in the new data opposed to the original data set, indicating a better data augmentation method.

Results

As seen in Table 1 and Fig 1, three hops of the same language family achieves the lowest BLEU score which is a 71% relative reduction in BLEU score when compared to the baseline (2 hops, same language family).

Table 1. Results showing the BLEU scores for each back translation method.

Experiment	Same Language Family	Different Language Family
Baseline (Two hops)	0.2917	0.2546
Three Hops	0.1190	0.0942
Four Hops	0.0848	0.0939

Absolute Decrease in BLEU Score based on Data Augmentation Method

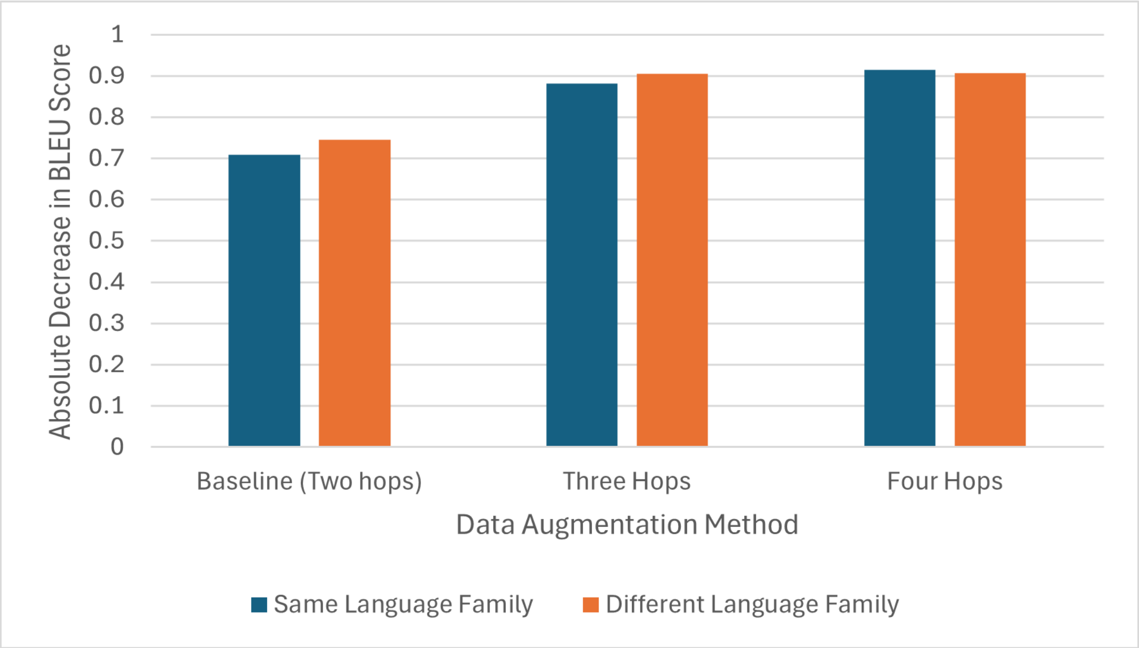

Figure 1. Graph that shows the absolute decrease in BLEU score for each back translation method

Discussion

To Answer Research Question 1: Do we achieve more diversity in training data utilizing multiple chains of back translation compared to the traditional method of using one-chain back translation?

As seen in Table 1, when we average the results for the same language family and different language families, four hops achieved 61% reduction in BLEU score compared to three hops and three hops achieved a 16.2% reduction in BLEU score compared to two hops. This indicates that utilizing multiple chains of back translations leads to more diversity in data when compared to the traditional method of using one-chain back translation.

To Answer Research Question 2: Do we achieve more diversity in training data utilizing languages from different language families for back translation compared to using languages from the same language family?

Across the different experiments, there is no conclusive evidence that using language from different language families leads to more diversity in data when compared to using languages from the same language family. While using languages from different language families led to more diversity in two hops and three hops, using language from the same language family led to more diversity in four hops.

Conclusion and Future Work

The findings from this research indicate that the novel multi-chain back translation can be a more effective data augmentation strategy for natural language processing tasks compared to traditional single-chain methods. By integrating multiple language families, the augmented dataset can create a broader spectrum of linguistic diversity in some scenarios, potentially leading to improved model performance when trained with diverse datasets.

The experiments conducted thus far have utilized Google Translate for back translation. To account for the variability in different translation models, future tests could expand to include systems such as Microsoft Translator or ChatGPT Translator. Given that these models are trained on distinct datasets and have unique algorithms, they may produce varied translation outputs. Incorporating a range of translation systems will allow us to assess if they produce varied results.

Additionally, current experiments used languages from each language family that were selected at random. Future tests could take a more systematic approach by carefully selecting languages based on specific criteria, such as language family characteristics or linguistic properties. This could help identify if certain languages or language families inherently introduce more or less diversity during the back translation process. Additionally, analyzing the translation patterns across different language family combinations could reveal if particular combinations lead to unique variations in the augmented data. Such insights could help identify if there is an optimal selection of languages and combinations that increase diversity.

Furthermore, measuring the impact of length for the translation process by adding more languages to the translation chain could provide insight into whether there is a direct relationship between number of translation "hops" and the amount of diversity, or if there is a point where a certain amount of translation "hops" no longer produces adequate diversity.

By implementing these future tests and more, we aim to continue refining our understanding of the back translation process and its impact on data augmentation. Future tests and insights will help contribute to the development of diversified natural language processing models that can be used in a wide range of applications.

Limitations

The augmented data derived from the novel data augmentation techniques investigated in this research study has not been utilized in downstream machine learning tasks to measure how much performance improvement they provide in downstream tasks. Future researchers could explore the potential benefits of integrating this augmented data into their models to assess its impact on relevant performance metrics. Furthermore, the novel data augmentation techniques were evaluated on one dataset (SST-2). Future researchers could explore evaluating these novel data augmentation techniques on diverse datasets spanning various domains to ascertain their generalizability and effectiveness across different tasks and contexts.

Acknowledgments

First and foremost, I would like to thank the University of North Texas and Dr. Rodney Nielsen for this opportunity, and for connecting me to my mentor, Dr. Solomon Ubani, whose unwavering support and guidance have been a driving force behind this research project. Their insight, feedback, dedication, time, and help have been instrumental in shaping the direction and outcome of this project. I am truly fortunate to have been able to work and learn from their mentorship, and I am forever grateful for their belief in my abilities and the lasting impact they left on my path to further research endeavors.

I would also like to express my gratitude to my family, Meera, Salai, and Praneeth Champooranan, for always supporting me and encouraging me to seek new opportunities to grow myself. Their motivation has been a constant source of strength, enabling me to navigate through challenges and embrace new opportunities to learn. Their faith in me has been the driving force behind my research pursuits.

Lastly, I would like to thank my friend, Alisma Adhikari, for introducing me to the realm of research and igniting my passion for this field. My research pursuits would not have been possible without her and her encouragement to take chances and step out of my comfort zone. Her support has been a constant source of inspiration and motivation throughout my research journey.

Without these individuals, this research project would not have been possible. I am grateful for their guidance, support, belief, and everything they have done for me.

Acknowledgments

I would like to thank my advisor for the valuable insight provided to me on this topic.

References

[1] Hayashi, T., Watanabe, S., Zhang, Y., Toda, T., Hori, T., Astudillo, R., & Takeda, K. (2018, December). Back-translation-style data augmentation for end-to-end ASR. In 2018 IEEE Spoken Language Technology Workshop (SLT) (pp. 426-433). IEEE.

[2] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. "Recursive deep models for semantic compositionality over a sentiment treebank". In: Proceedings of the 2013 conference on empirical methods in natural language processing. 2013, pp. 1631–1642

[3] Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002, July). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics* (pp. 311-318).