NLP for Sign Language Translation

Korea University COSE461 Final Project

Robin Dieudonné
Department of Mathematics & Computer Science
Team 5
2024951308

Lee Yunho
Department of Data Science
Team 5
2022320304

Pan Yangcan
Department of Computer Science
Team 5
2021320166

Abstract

Sign Language Machine Translation (SLT) is challenged by the scarcity of annotated data. In this project, we enhance data augmentation techniques and improve *gloss-to-text* translation, where glosses are sequences of transcribed spoken-language words ordered as they are signed. We treat gloss-to-text translation as a low-resource neural machine translation (NMT) problem. Our data augmentation approach focuses on generating glosses from text to address the shortage of professionally annotated SLT data.

Utilizing the PHOENIX-Weather 2014T corpus, we achieve a new state-of-the-art ROUGE-L score of 57.11. Our results demonstrate the potential for further advancements in SLT by jointly training encoder-decoder architectures on both text-to-gloss and gloss-to-text tasks. This combined approach not only augments the data but also improves SLT performance using the newly generated data.

1 Introduction

Sign language translation is a crucial tool for bridging the communication gap between the deaf community and the hearing population. Traditional translation models often rely on a series of processes that first convert sign language videos into gloss annotations and then translate these glosses into natural English sentences. However, this approach faces significant challenges due to the limited availability of gloss-annotated data and the complexity of accurately capturing the nuances of sign language.

To address these challenges, we propose two methods to improve the accuracy of sign language translation. The first method bypasses the gloss representation entirely, converting sign language videos directly into English sentences. This approach aims to streamline the translation process and reduce the dependency on intermediate gloss annotations.

The second method tackles the chronic issue of data scarcity in the artificial intelligence learning process for sign language translation. By augmenting the dataset with additional pairs of gloss and English sentences, we enhance the training data, enabling the model to learn more effectively from a richer set of examples.

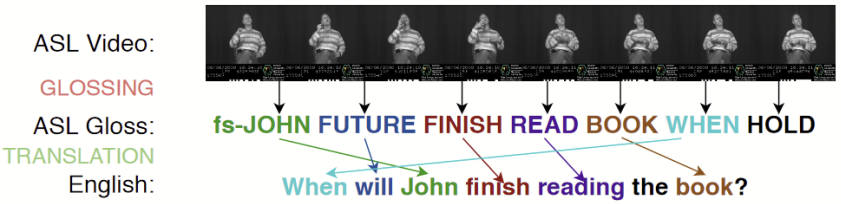

Figure 1: Illustration of the two-stage method for SLT: first extract glosses from the video and then translate glosses to a spoken language.

We conclude by evaluating the effectiveness of these methods in improving translation accuracy. Our findings demonstrate the potential of these approaches to significantly enhance the performance of sign language translation models, making them more reliable and effective tools for communication.

2 Related Work

Our work builds on prior research in sign language translation, leveraging benchmark datasets like PHOENIX-Weather 2014T and advancements in deep learning and NLP.

Traditional data augmentation techniques have primarily focused on images and natural language sentences. However, in this study, we aim to apply these techniques to pairs of glosses and natural English sentences. By incorporating noise addition and cropping techniques, we seek to diversify the data and enhance the performance of the model considering the patterns and Characteristics of glosses.

3 Data Augmentation and SLT

This section discusses methods to improve gloss-to-text translation through data augmentation. The goal is to generate new pairs of sentences and their corresponding glosses.

3.1 A First Algorithm for Data Augmentation

- Word Similarity Analysis: First, we analyze the similarity between words in the sign language sentences and natural language sentences. This includes a process of normalizing the words to lowercase and using the SequenceMatcher to calculate the similarity between words. For each word in the sign language sentence, we find the most similar word in the corresponding natural language sentence. We only match word pairs with a similarity score of 0.8 or higher.
- Noise Injection: To further improve the diversity of the dataset and the robustness of the model, we consider a strategy of intentionally injecting noise into the sentences. This can include shuffling the word order, randomly adding or deleting words. By injecting noise in this way, we can increase the model's resilience to handle various sentence structures and potential errors that may occur in real data.
- Back-Translation: We use Google Translate to translate the sentence into another language and then back into the original language. This method can help introduce natural language variations and paraphrasing.
- Random Deletion: We randomly delete words from the sentence with a certain probability. This technique introduces variability and helps the model handle incomplete sentences.
- Lemmatization and POS Filtering: We apply lemmatization to normalize words to their base forms. This process involves converting words to lowercase, tokenizing, and then using a lemmatizer to reduce words to their root forms. Additionally, we filter words based on their parts of speech (POS), retaining only nouns, verbs, adjectives, and adverbs. This helps in focusing on the most semantically significant words while discarding less important ones like conjunctions and prepositions.

- Synonym Replacement: We replace certain words in the sentence with their synonyms using the WordNet lexical database. This method introduces lexical variety and helps the model generalize better to different vocabulary.

3.2 Data Augmentation and SLT with Transformers

In this section, we introduce a more efficient and scalable approach to data augmentation. We develop two models, D2G-T5 (Deutsch to Gloss) and G2D-T5, based on the pre-trained Flan-Base-T5 architecture. D2G-T5 is trained to translate German language sentences into German Sign Language Glosses, while G2D-T5 performs the reverse translation.

To train D2G-T5, we fine-tune it on the German Sign Language dataset [For+12]. By doing so, it learns to generate glosses for German sentences. This process enables us to augment the data by generating new pairs of sentences and glosses solely from German sentences using D2G-T5. Specifically, we leverage D2G-T5 to expand a set of 519 unannotated German sentences, thus creating additional pairs of glosses and sentences for further training G2D-T5.

Subsequently, we fine-tune G2D-T5 using two approaches: one utilizing the original PHOENIX dataset and the other incorporating the augmented dataset. The augmented dataset is formed by combining and shuffling the base PHOENIX dataset with the newly generated pairs from D2G-T5.

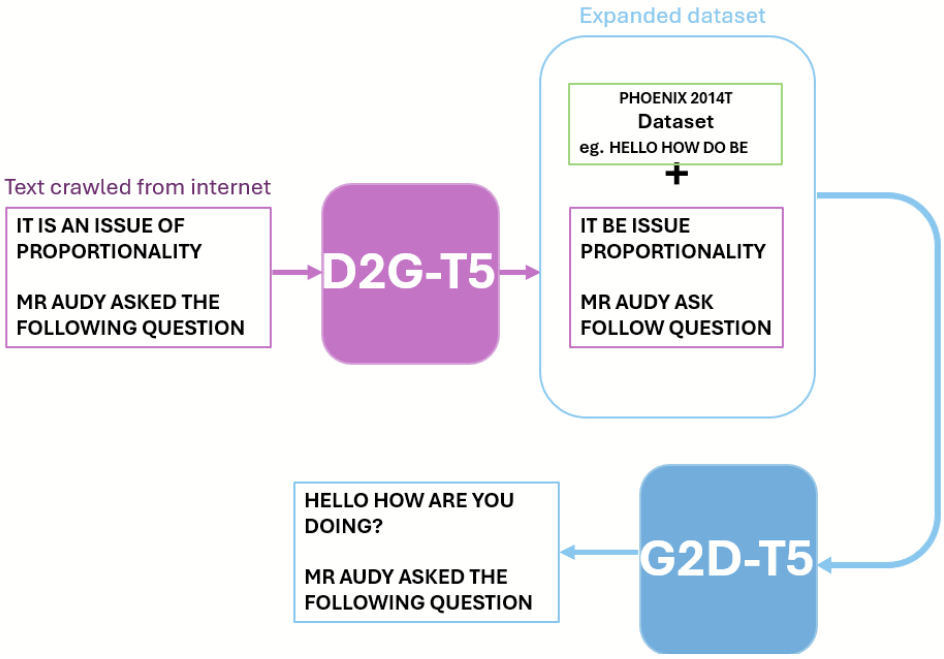

Figure 2: We illustrate our data augmentation and gloss translation combined pipeline. We first augment data using sentences found on internet and inputted to the finetuned D2G-T5 for text-to-gloss. We then combine the new pairs with the PHOENIX dataset and finetune G2D-T5 on gloss-to-text using that new, larger dataset. Note that German text was translated to English here for clarity of understanding.

4 Experiments

To fairly compare our results with other approaches, we use the same train/test splits that they used in our datasets.

4.1 Data

The datasets used are **PHOENIX-Weather 2014T** (PHOENIX), containing German Sign Language (GSL) weather forecast videos with gloss annotations and written German transcriptions. By selecting

only gloss annotations and their corresponding written German transcriptions, we split the dataset into 7096 training pairs and 642 testing pairs. We proceed similarly with **ASLG-PC12** (ASLG), containing American Sign Language (ASL) videos with gloss annotations. We split it into 82,710 training pairs and 1,000 testing pairs.

These datasets are used to train and evaluate models for translating GSL to German and ASL to English. Our models are inputted tokenized versions of glosses (or sentences) and trained to predict their corresponding text (or glosses).

Finally, we set up the AUG dataset, a mixture of PHOENIX and 519 gloss-sentence pairs obtained from our data augmentation method described in 3.2.

4.2 Evaluation method

We evaluate our models across all datasets using the ROUGE-L metric [Lin04] that compares our automatically produced translation against a reference (human-produced) translation.

4.3 Experimental details

The Flan-T5 [Chu+22] encoder-decoder base model contains 250M parameters and is pretrained on dozens of languages including English and German. We finetune it from the Flan-T5-base pretrained checkpoint with a batch size of 8 and a learning rate of 0.0004 with AdamW [LH19]; other parameters are the T5 defaults. For models trained on ASLG, we finetune for 31,017 steps. For models trained on PHOENIX, we finetune for 2856 steps. We train for 6.0 L4-GPU hours on ASLG and 1.0 on PHOENIX.

4.4 Results

Approach	Training Schedule	ROUGE-L	ROUGE-1	ROUGE-2
Yin et al. [YR20]	PHOENIX	48.51	-	-
G2D-T5	PHOENIX	56.71	59.13	28.37
AUG-G2D	PHOENIX+AUG	**57.11**	61.05	30.24
Yin et al. [YR20]	ASLG	96.22	-	-
E2G-T5	ASLG	96.02	95.93	92.84

Metrics for ASL to English and GSL to German translation. Our models finetune a pretrained Flan-T5 checkpoint.

We ablate the effect of our new architecture and of our data augmentation method.

- PHOENIX: We only train on PHOENIX
- PHOENIX+AUG: We train on a mixture of PHOENIX and the AUG=augmented dataset generated by D2G-T5
- ASLG: We only train on ASLG

5 Analysis

Our models were evaluated on the PHOENIX-Weather 2014T and ASLG-PC12 datasets using the ROUGE-L metric. For the PHOENIX dataset, the G2D-T5 model achieved a ROUGE-L score of 56.71, significantly outperforming the baseline. The AUG-G2D model, incorporating data augmentation, further improved to 57.11, with higher ROUGE-1 and ROUGE-2 scores indicating better word and bi-gram accuracy. However, both models occasionally struggled with idiomatic expressions and context-dependent signs.

For the ASLG dataset, the E2G-T5 model achieved a ROUGE-L score of 96.02, close to the state-of-the-art, demonstrating strong performance in translating ASL to English. The high ROUGE-1 and ROUGE-2 scores reflect its proficiency in capturing ASL semantics and syntax, though it also faced challenges with context-dependent signs and idiomatic expressions.

Overall, the Flan-T5 model showed robust translation capabilities, enhanced by data augmentation. Future work should focus on improving the handling of idiomatic and context-rich expressions in sign language.

6 Conclusion

In this project, we presented four models for *gloss-to-text* translation, *text-to-gloss* translation and data augmentation. Our key improvement over prior work is our capacity to generate arbitrarily many new gloss-sentence pairs thanks to our *text-to-gloss* models that help augment the training data size for our *gloss-to-text* models and their scores. We showed the importance of data with our pipeline for SLT by achieving a better score with G2D-T5 finetuned on PHOENIX+AUG, compared to the score after finetuning on PHOENIX only. Our *gloss-to-text* G2D-T5 model achieves SOTA on PHOENIX, proving the capability of encoder-decoder architecture for SLT.

However, our work leverages heavy transformer architectures which are not adapted to efficient translation if not used on powerful hardware. Furthermore, our finetuning required many computational resources. Finally, our base G2E-T5 does not manage to score better ROUGE-L than in [YR20]. Data augmentation could have helped but limited computational resources denied us from doing it.

While our methods improve upon prior work, the translations are still subjectively low-quality and are not ready for real-time end-to-end SLT. We also hope for better quality and curated datasets, as we observed that some of them contained non-sense English sentences and some foreign languages too. Future work may look to add our G2D-T5 or E2G-T5 to complete SLT pipelines, combined with a video-to-glosses module.

References

[Lin04] Chin-Yew Lin. "ROUGE: A Package for Automatic Evaluation of Summaries". In: *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, July 2004, pp. 74–81. URL: https://www.aclweb.org/anthology/W04-1013.

[For+12] Jens Forster et al. "RWTH-PHOENIX-Weather: A Large Vocabulary Sign Language Recognition and Translation Corpus". In: May 2012.

[LH19] Ilya Loshchilov and Frank Hutter. *Decoupled Weight Decay Regularization*. 2019. arXiv: 1711.05101 [cs.LG].

[YR20] Kayo Yin and Jesse Read. "Attention is All You Sign: Sign Language Translation with Transformers". In: 2020. URL: https://api.semanticscholar.org/CorpusID:231609254.

[Chu+22] Hyung Won Chung et al. *Scaling Instruction-Finetuned Language Models*. 2022. DOI: 10.48550/ARXIV.2210.11416. URL: https://arxiv.org/abs/2210.11416.

A Appendix: Team contributions

Robin Dieudonné: Code Implementation, SLT with T5, Data Augmentation using Transformers, Experiment, Evaluation & Analysis, Report Writing

Pan Yangcan: Data Augmentation, Report Writing

Lee Yunho: Data Augmentation, Report Writing