From Sign Spottings to Spoken Language: Fine-Tuning Large Language Models for Improved Translation

Markus Hoehn* **Lane Burgett** **Semyon Zharkov** **Andrei Mazin**

Manhattan High School
Manhattan, KS 66502
{markush, lanbur, semyonzharkov, amazin}@ksu.edu

Abstract

The task of sign language translation is crucial for enhancing communication for the deaf and hard-of-hearing community. In this paper, we build upon previous work by enhancing the large language model (LLM) component of a hybrid approach for sign language translation from continuous video streams. The prior hybrid approach employed a GPT-3.5 Turbo prompt to generate translations from identified sign spottings. We improve this by incorporating fine-tuning techniques and revising the prompt. We employ a sign spotter from existing literature to identify individual gestures within the video stream. These identified gestures are then processed by an LLM, which constructs grammatically correct and coherent sentences. Our evaluation of different models demonstrate significant improvements, with the fine-tuned Gemini-1.0 Pro model being the most successful model in translation accuracy and coherence.

1 Introduction

The main form of communication for millions of deaf and hard of hearing people around the world is sign language. Sign languages are complex communication systems that utilize hand gestures, facial expressions, and body language to convey meaning [11]. Converting sign language to spoken language is a task known as Sign Language Translation (SLT) and can help people communicate to deaf and hard of hearing people.

SLT, when treated as a Neural Machine Translation (NMT) task, struggles due to the challenges of aligning and tokenizing continuous sign language videos and the grammatical differences and word order between spoken and sign languages [3]. To address this issue some researchers approach sign language as a two-step process: first, Sign Language Recognition (SLR) identifies individual signs (i.e., sign spotting) within the video, and then, these signs are translated into meaningful sentences to achieve SLT [4, 5, 16].

Large Language Models (LLMs), trained on extensive web-scale text corpora, have demonstrated significant capabilities in natural language processing tasks, such as multilingual translation. Their ability to understand and generate text across multiple languages with diverse syntactic and lexical properties highlights their versatility and effectiveness [10]. Given the distinct grammatical structures of spoken and sign languages and the different sequences of sign glosses (written representations of signs) and spoken word order, LLMs hold promise for converting sign spottings (identifications of individual signs) into spoken sentences due to their rich semantic understanding.

*Corresponding Author

Preprint. Under review.

A novel approach in the field, introduced recently in [13], involved using a sign spotter to identify hand gestures (glosses) for each video frame, subsequently processed by a large language model (LLM) to convert to spoken language. However, the LLM component was limited to prompting the 'GPT-3.5 Turbo' version of ChatGPT with a brief task description, lacking the integration of fine-tuning [15].

Our contributions to the approach introduced in [13] include: (1) Modifying the system input prompts for better sentence generation. (2) Fine-tuning LLMs specifically for converting prompted sign spottings into coherent sentences [15]. (3) Evaluating the performance of these fine-tuned models to identify the top performers.

2 Related Work

Early tasks in sign language automation primarily focused on SLR. Initially, due to technical limitations, research centered on hand-crafted features that analyzed hand shape and motion. Subsequently, pose, face, and mouth features were extensively incorporated into recognition pipelines [6, 2, 1, 7].

Modern approaches to SLT incorporate deep learning architectures, particularly Convolutional Neural Networks (CNNs), to learn robust features directly from video data. For instance, some papers proposed a CNN-based architecture for sign language translation, achieving significant improvements over traditional methods [3]. These deep learning approaches can capture complex spatial and temporal relationships within sign language gestures, leading to more accurate recognition [5, 16].

More recently, researchers have proposed leveraging LLMs and their large linguistic capability to advance SLT [14]. Approaches utilizing gloss-free translations performed considerably worse in terms of BLEU score [9] compared to their gloss-based counterparts [14]. Most notably, Sincan et al. [13] have proposed a method of translating sign spottings to spoken language through an LLM. In this paper, the authors sent a basic prompt to a pre-trained 'GPT-3.5 Turbo' model. The model did not include fine-tuning. We leverage the approach by Sincan et al. [13] and specifically explore the benefits of fine-tuning on two models.

Fine-tuning refers to a transfer learning approach where a pre-trained model's parameters are adjusted using new data. This adjustment can involve training the entire neural network or only specific layers [15]. For example, an LLM can be provided with a prompted input of sign spottings and fine-tuned to generate coherent spoken language.

3 Method

We use the Spotter from [13] and then feed Spottings over 0.7 confidence into our LLM component.

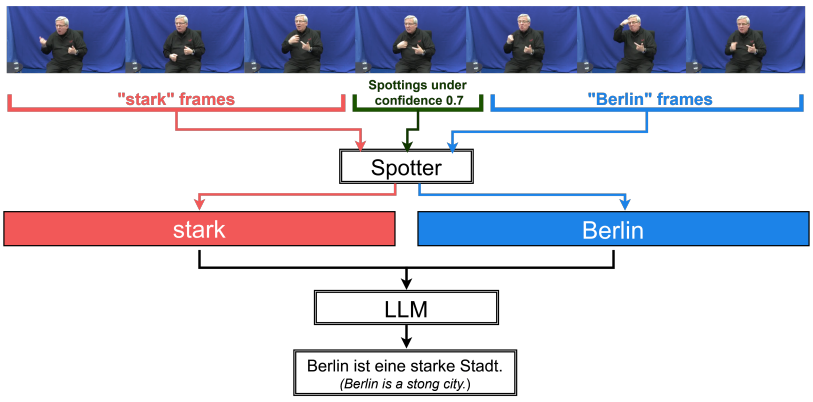

Figure 1: An overview of the proposed sign language translation architecture.

We use a modified system prompt for all our LLM components. For Gemini-1.0 Pro, it is added as another user prompt. Adapted from [13], this prompt is altered to always generate a sentence, eliminating the "No Translation" option. This ensures better comparison with fine-tuned models, which do not produce "No Translation" due to their training data.

Prompt: `"You are a helpful assistant designed to generate a sentence based on the list of words entered by the user. You need to strictly follow these rules:`

1. `The user will only give the list of German words separated by a space. You must generate a meaningful German sentence from them, even if you have to guess.`
2. `Only provide a response containing the generated sentence. If you cannot create a coherent German sentence from the words, still make an attempt to form a German sentence using the given words."`

We fine-tune two models, 'GPT-3.5 Turbo' and 'Gemini-1.0 Pro,' using glosses (except the INDEX gloss which stands for pointing) directly from the dataset and separately the spotter's sign spottings on the dataset's videos.

4 Experiments

We use the MeineDGS dataset which is a comprehensive linguistic resource for German Sign Language (DGS) [8]. The videos feature natural conversations between two deaf participants. We adhere to a sign language translation protocol for our splits [11] to compare to [13], which includes 40,230 training sentences, 4,996 development sentences, and 4,997 test sentences.

We fine-tune models (with hyperparameters: epochs = 1, batch size = 26, LR multiplier = 2) using the protocol's train split for training and for GPT-3.5 Turbo we use the protocol's dev split for validation [11]. For user input, we use the glosses in the split (excluding the INDEX gloss) or the predicted sign spottings. For model output, we use the provided spoken language directly from [8].

We use BLEU [9] and BLEURT [12] metrics to evaluate our approach.[2] We use the sacreBLEU implementation for BLEU scores and BLEURT-20 checkpoints for BLEURT scores, evaluating various LLM implementations [9, 12]. The evaluation are conducted on the protocol's test split [11], comparing to the original spoken language sentences from [8].

Table 1: Benchmarks of different models

Method	BLEU1	BLEU2	BLEU3	BLEU4	BLEURT
On Spottings					
GPT-3.5 Turbo (scores from [13])	14.82	4.19	**1.45**	**0.64**	21.62
GPT-3.5 Turbo (modified prompt)	19.23	1.32	0.21	0.04	27.26
GPT-3.5 Turbo (fine-tuned on glosses)	14.97	0.99	0.11	0.01	27.54
GPT-3.5 Turbo (fine-tuned on spottings)	14.38	0.80	0.12	0.02	25.59
Gemini-1.0 Pro (modified prompt)	19.28	1.40	0.24	0.05	27.72
Gemini-1.0 Pro (fine-tuned on glosses)	24.09	3.13	0.80	0.29	35.03
Gemini-1.0 Pro (fine-tuned on spottings)	**26.73**	**3.60**	0.92	0.37	**35.41**
On Glosses					
GPT-3.5 Turbo (modified prompt)	28.38	4.50	1.30	0.48	40.73
GPT-3.5 Turbo (fine-tuned on glosses)	20.87	2.80	0.69	0.20	39.85
GPT-3.5 Turbo (fine-tuned on spottings)	18.15	1.95	0.46	0.14	33.85
Gemini-1.0 Pro (modified prompt)	28.70	4.74	1.36	0.48	42.54
Gemini-1.0 Pro (fine-tuned on glosses)	36.55	9.86	**4.09**	**1.98**	**51.65**
Gemini-1.0 Pro (fine-tuned on spottings)	**37.47**	**10.05**	4.08	1.95	49.30

[2]The BLEU and BLEURT metrics were computed on a system with a Ryzen 9 3950X processor and 128GB of RAM. Sign spottings were computed on an NVIDIA A40 GPU.

Table 2: Selected Example from MeineDGS test split

Glosses	TAUB-GEHORLOS KENNENLERNEN BEIDE GUT
Spotter	TAUB-GEHORLOS KENNENLERNEN BEIDE GUT
Spoken Language	Er war gehörlos, wir haben uns kennengelernt und gut verstanden. *(He was deaf, we met and got along well.)*
GPT-3.5 Turbo ([13] prompt)	No Translation
GPT-3.5 Turbo (modified prompt)	Beide können sich gut kennenlernen, obwohl einer taub und der andere gehörlos ist. *(They both get to know each other well, even though one is deaf and the other is deaf.)*
GPT-3.5 Turbo (fine-tuned on glosses)	Und ich habe viele taube Kinder kennengelernt, es war alles gut. *(And I met a lot of deaf children, everything was good.)*
GPT-3.5 Turbo (fine-tuned on spottings)	Meine Freundin und ich kennen uns. *(My girlfriend and I know each other.)*
Gemini-1.0 Pro ([13] prompt)	Taub-Gehörlose lernen sich beide gut kennen. *(Both deaf people get to know each other well.)*
Gemini-1.0 Pro (modified prompt)	Taub-Gehörlose kennenlernen beide gut. *(Deaf people get to know both people well.)*
Gemini-1.0 Pro (fine-tuned on glosses)	Ich habe einen Gehörlosen kennengelernt und wir verstehen uns gut. *(I met a deaf person and we get along well.)*
Gemini-1.0 Pro (fine-tuned on spottings)	Ich habe einen Gehörlosen kennengelernt, der war sehr nett. *(I met a deaf person who was very nice.)*

Table 2 includes results from using the prompt in [13] via our calling code.

5 Discussion

Fine-tuning improvements among BLEURT and BLEU scores, particularly with Gemini-1.0 Pro, show potential for application to other datasets with different spotters. This facilitates a complete gloss-based sign language translation system, enhancing the LLM's ability to generate coherent and accurate translations, proving effective in SLT. Fine-tuning also provides qualitative improvements and prevents outputs like "The generated sentence is..." and other unwanted text, such as English text, which can occur even with an explicit system prompt on a pre-trained model. However, fine-tuning has a drawback: without training data designating some spottings as untranslatable, the model won't acknowledge its inability to translate and may generate long sentences, even with a system prompt addressing this limitation. We observe that fine-tuning increases the use of some words and phrases disconnected from the input spottings/glosses. Further exploration of prompt engineering to discourage these behaviors is warranted. These behaviors may also explain the differences between BLEU scores among different n-gram precisions. Our benchmarks on the dataset's glosses show capabilities for a theoretically perfect spotter. We recognize that Gemini-1.0 Pro outperforms GPT-3.5 Turbo for this task when fine-tuned. Finally, SLT has significant potential for improving communication between deaf and hearing people. One promising application is sign language teaching, where models like ours can act as virtual tutors, providing instant feedback and translation during practice sessions.

References

[1] Epameinondas Antonakos, Anastasios (Tassos) Roussos, and Stefanos Zafeiriou. A survey on mouth modeling and analysis for sign language recognition. 05 2015.

[2] Patrick Buehler, Andrew Zisserman, and Mark Everingham. Learning sign language by watching tv (using weakly aligned subtitles). In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2961–2968, 2009.

[3] Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. Neural sign language translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[4] Necati Cihan Camgöz, Oscar Koller, Simon Hadfield, and Richard Bowden. Sign language transformers: Joint end-to-end sign language recognition and translation. *CoRR*, abs/2003.13830, 2020.

[5] Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden. Sign language transformers: Joint end-to-end sign language recognition and translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10023–10033, 2020.

[6] Ali Farhadi, David Forsyth, and Ryan White. Transfer learning in sign language. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2007.

[7] Oscar Koller, Jens Forster, and Hermann Ney. Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers. *Computer Vision and Image Understanding*, 141:108–125, 12 2015.

[8] Reiner Konrad, Thomas Hanke, Gabriele Langer, Dolly Blanck, Julian Bleicken, Ilona Hofmann, Olga Jeziorski, Lutz König, Susanne König, Rie Nishio, Anja Regen, Uta Salden, Sven Wagner, Satu Worseck, Oliver Böse, Elena Jahn, and Marc Schulder. Meine dgs – annotiert. öffentliches korpus der deutschen gebärdensprache, 3. release / my dgs – annotated. public corpus of german sign language, 3rd release, 2020.

[9] Matt Post. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels, October 2018. Association for Computational Linguistics.

[10] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9, 2019.

[11] Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. Signing at Scale: Learning to Co-Articulate Signs for Large-Scale Photo-Realistic Sign Language Production. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[12] Thibault Sellam, Dipanjan Das, and Ankur P. Parikh. BLEURT: learning robust metrics for text generation. *CoRR*, abs/2004.04696, 2020.

[13] Ozge Mercanoglu Sincan, Necati Cihan Camgoz, and Richard Bowden. Using an llm to turn sign spottings into spoken language sentences. *arXiv preprint arXiv:2403.10434*, 2024.

[14] Ryan Wong, Necati Cihan Camgoz, and Richard Bowden. Sign2GPT: Leveraging large language models for gloss-free sign language translation. In *The Twelfth International Conference on Learning Representations*, 2024.

[15] Aston Zhang, Zachary C. Lipton, Mu Li, and Alexander J. Smola. *Dive into Deep Learning*. Cambridge University Press, 2023. `https://D2L.ai`.

[16] Biao Zhang, Mathias Müller, and Rico Sennrich. Sltunet: A simple unified model for sign language translation. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2022.