KnowComp Submission for WMT23 Sign Language Translation Task

Baixuan Xu[1], Haochen Shi[1], Tianshi Zheng[1], Qing Zong[2],
Weiqi Wang[1], Zhaowei Wang[1], Yangqiu Song[1]
[1]Department of Computer Science and Engineering, HKUST, Hong Kong SAR, China
[2]Harbin Institute of Technology (Shenzhen), Guangzhou, China
{bxuan, hshiah, tzhengad}@connect.ust.hk, zongqing0068@gmail.com
{wwangbw, zwanggy, yqsong}@cse.ust.hk

Abstract

Sign Language Translation (SLT) is a complex task that involves accurately interpreting sign language gestures and translating them into spoken or written language and vice versa. Its primary objective is to facilitate communication between individuals with hearing difficulties using deep learning systems. Existing approaches leverage gloss annotations of sign language gestures to assist the model in capturing the movement and differentiating various gestures. However, constructing a large-scale gloss-annotated dataset is expensive and impractical to cover multiple languages, and pre-trained generative models cannot be efficiently used due to the lack of textual source context in SLT. To address these challenges, we propose a gloss-free framework for the WMT23 SLT task. Our system primarily consists of a visual extractor for extracting video embeddings and a generator responsible for producing the translated text. We also employ an embedding alignment block that is trained to align the embedding space of the visual extractor with that of the generator. Despite undergoing extensive training and validation, our system consistently falls short of meeting the baseline performance. Further analysis shows that our model's poor projection rate prevents it from learning diverse visual embeddings. Our codes and model checkpoints are available at https://github.com/HKUST-KnowComp/SLT.

1 Introduction

Machine translation has significantly improved thanks to the development of pre-trained language models (Mohammadshahi et al., 2022; Huang et al., 2023). While translation within a single modality has been extensively studied, translation involving multiple modalities remains challenging and less explored (Lin et al., 2023). Sign Language Translation (SLT), which translates sign gestures into spoken language, remains an exceedingly complex task due to two fundamental challenges. Firstly, sign languages are visual-gestural languages that rely on manual signs, facial expressions, and body movements to convey information. This fundamental distinction sets them apart from written languages, which consist of word characters and symbols. Consequently, translation models must be able to accurately interpret visual signals and gestures and develop a deep understanding of the semantics involved in producing prompt translations. However, the multimodal nature of sign languages poses a significant challenge for models, requiring them to learn and generalize these complex interactions effectively. Moreover, sign languages are typically represented as exceedingly lengthy sequences of frames, surpassing the number of tokens in a standard sentence (Guo et al., 2018). This requires translation models to grasp the prolonged dependencies within the video to accurately capture the information conveyed through these visual signals.

To tackle these challenges, methods have been proposed that utilize pre-training a visual backbone based on gloss annotations (Camgöz et al., 2020). These approaches have demonstrated exceptional performance in various multimodal translation tasks. Nevertheless, the acquisition of extensive gloss annotations comes with significant cost and practical constraints, making it impractical to cover a wide range of multilingual translation directions (Müller et al., 2023).

In this paper, we propose a gloss-free framework for the SLT task. Our approach combines a pre-trained visual backbone model (Varol et al., 2021), which has been trained to recognize sign gestures, with a GPT2-based language model (Radford et al., 2019) to generate the translated sentence. To align the embedding space between both models, we utilize an embedding alignment block inspired by ClipCap (Mokady et al., 2021). The final translation is produced using converted visual embeddings and text embeddings (Section 3). Despite

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 351–358
December 6–7, 2023.

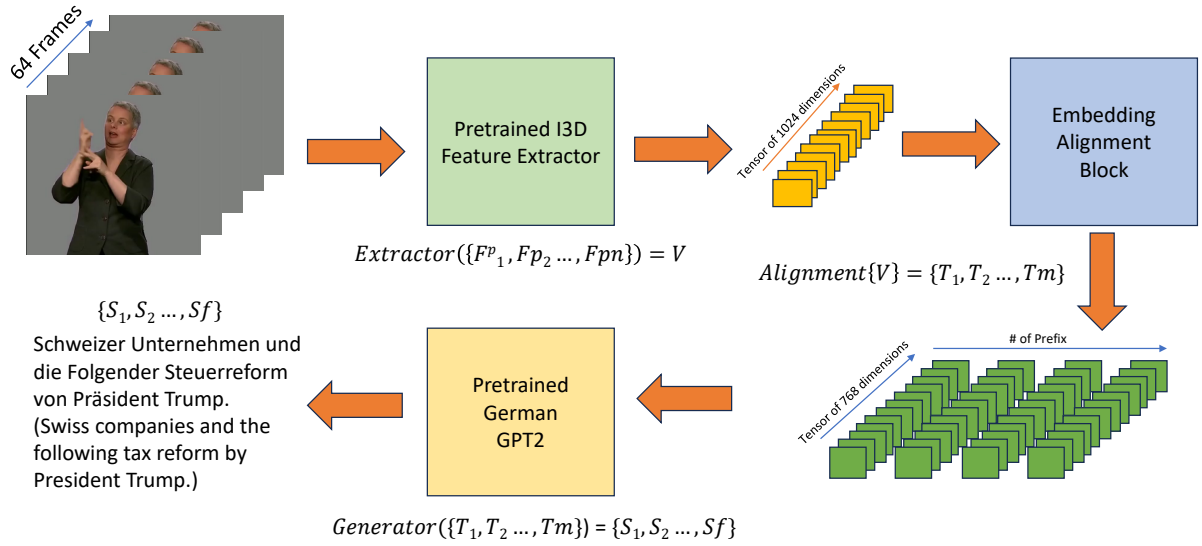

Figure 1: An overview of our framework. We first downsample video data and feed them into the visual feature extractor to obtain the visual embeddings. The embeddings are then passed into the alignment block to project them into embedding inputs of the `German-GPT2`. They are used as the prefix of the GPT2 model to generate the final translation results.

conducting extensive experiments with our system, we consistently achieved a BELU score of 0.1 and a chrF score of 7.6 on the testing set of the SRF dataset, which is below the baseline performance. Further analysis reveals that the embedding alignment block fails to differentiate between different embedding inputs from the visual encoder. As a result, our generation often produces repeated and nonsensical outputs. We will make all codes and results publicly available upon acceptance of this paper.

2 Preliminary

2.1 Task Definition

The objective of the Sign Language Translation (SLT) task (Fang et al., 2017; Kan et al., 2021) is to utilize the model's video understanding ability and language modeling ability to translate meaningful gesture sequence into spoken language (Varol et al., 2021; Hu et al., 2023). Formally, our objective is to learn a conditional probablity $P(S|F^r)$ of generating a natural spoken language, denoted as $\mathcal{S} = \{S_1, S_2..., S_m\}$ with m tokens given the raw sign language video $\mathcal{F}^r = \{F_1^r, F_2^r..., F_n^r\}$ with n frames.

To better elaborate our proposed model, we hereby set some notions for convenience. The aforementioned $\mathcal{S}$ and $\mathcal{F}^r$ refer to the translated spoken language and the sign language video before preprocessing. We use $\mathcal{F}^p = \{F_1^p, F_2^p..., F_n^p\}$ to denote the preprocessed video frames. In our proposed model, we endeavor to optimize the alignment block to yield better translation results while parameters in other modules are frozen for training efficiency.

Dataset	#raw data	#processed data
SRF	771	354901

Table 1: Statistics of the SRF dataset. # raw data refers to the number of videos, and # processed data is the amount of data after video slicing.

2.2 Dataset

We use the datasets provided by Müller et al. (2023) as our primary training and evaluation benchmarks. Our model is exclusively trained on the SRF dataset (Jiang et al., 2023b), while the SignSuisse dataset (Jiang et al., 2023a) is solely utilized for zero-shot evaluation purposes. Both datasets consist of sign language videos accompanied by their corresponding translation text in German. The statistical information for the SRF dataset can be

found in Table 1.

The SRF (Jiang et al., 2023b) dataset comprises videos from Standard German daily news (Tagesschau) and Swiss German weather forecast (Meteo) episodes broadcast. They are further interpreted into Swiss German Sign Language by hearing interpreters via Swiss National TV. In the SRF dataset released by Müller et al. (2022), there are a total of 354901 video slices covering episodes from 2014 to 2022.

The SignSuisse (Jiang et al., 2023a) dataset contains 18221 lexical items in Swiss German Sign Language, French Sign Language of Switzerland, and Italian Sign Language of Switzerland, represented as videos with corresponding spoken language translations.

The BSL-1k (Albanie et al., 2020) is a large-scale sign language recognition dataset constructed based on British Sign Language(BSL) signs. The authors leverage the observation that signers often mouth the word they are signing simultaneously, providing additional visual cues. They use visual keyword spotting to detect the mouthings and align them with the subtitles to determine whether and when a keyword of interest is uttered by a talking face using visual information. The dataset is then used to train a strong sign recognition model for co-articulated signs in BSL and serves as excellent pretraining for other sign languages and benchmarks. Thus, in our paper, it is reasonable for us to use a model pretrained on BSL-1k as our visual feature extractor and expect it to yield meaningful and informative video representations for the model to utilize.

3 Method

This section introduces our proposed framework, which is depicted in Figure 1. While previous systems (Dey et al., 2022; Shi et al., 2022; Tarres et al., 2022) primarily employ an encoder-decoder paradigm and train their models from scratch to address this task, we distinguish ourselves by being the first to utilize a pre-trained language model for this task, as these language models possess strong natural language understanding and generation ability (Wang et al., 2023c, 2022; Fang et al., 2021b,a, 2023; He et al., 2022; Bai et al., 2023a,b). Specifically, we leverage the pre-trained I3D model provided by Varol et al. (2021) as our visual extractor backbone and employ a German-GPT2 model (Schweter, 2020) as the generator's backbone.

3.1 Video Extractor

We use the Two-Stream Inflated 3D ConvNets (I3D; Carreira and Zisserman, 2017) that is pre-trained on the BSL-1k (Albanie et al., 2020) dataset as our visual extractor backbone. I3D was first proposed by Carreira and Zisserman (2017) aiming to mitigate the 2D convolution network failure to capture the temporal information behind the video data. To overcome this, I3D directly expands the original 2D convolution network, which yields significant success in 3-dimensional space by expanding extra dimension to the kernel and pooling layer. When the kernel and pooling layers are extended to 3D in I3D, these layers are initialized using the pre-trained weights from the corresponding 2D image classification networks. Overall, the I3D model offers a powerful framework for action recognition by leveraging the strengths of both image classification architectures and spatio-temporal feature extraction in videos. For the SLT task, we ask the model to transform a 64 frames ($\mathcal{F}_p$) video into a 1024-dimensional tensor ($\mathcal{V}$), denoted as:

$$\text{Extractor}(\{F_1^p, F_2^p ..., F_n^p\}) = \mathcal{V}$$

3.2 Embedding Alignment Block

Inspired by the success of ClipCap (Mokady et al., 2021), we then train an embedding alignment block to project the obtained visual embeddings $\mathcal{V}$ into textual embeddings T for further processing by German-GPT2. ClipCap was originally designed by Mokady et al. (2021) to tackle the task of image captioning (Ou et al., 2023). In the paper, the authors utilized the expressive power of an image feature extractor and a generative language model. By adding an alignment layer in between, the representation of the visual modality can be projected to the text modality for the language model to generate meaningful captions. The extraordinary ability shown by this innovative architecture makes it reasonable for us to adopt it in our framework. We implement the alignment block by stacking six transformer encoder layers together. Two fully connected neural networks are also placed before and after the alignment block to extend the visual embeddings into a sequential format and densify the aligned embeddings into prefix embeddings of German-GPT2, respectively. Formally, this process can be denoted as:

$$\text{Alignment}(\mathcal{V}) = \{T_1, T_2 ..., T_m\}$$

	BLEU			chrF			BLEURT		
Submission	all	SS	SRF	all	SS	SRF	all	SS	SRF
Baseline	0.09±0.03	0.15±0.06	0.10±0.05	12.4±0.4	12.2±0.5	12.5±0.5	0.072±0.003	0.083±0.005	0.060±0.005
KnowComp	0.07±0.05	0.06±0.02	0.11±0.09	7.6±0.3	8.2±0.4	7.2±0.4	0.083±0.005	0.084±0.007	0.081±0.007

Table 2: The experiment result of our proposed model comparing to the baseline released by the shared task organizer. Although our model was trained only on SRF, we still shown stronger performance on BLEURT than the baseline model in domain of SS and all. SS dataset is OOD and all is partially OOD for our model.

3.3 Text Generator

Finally, we leverage a pre-trained `German-GPT2` model as the text generator to generate the final translations by feeding the previously acquired textual prefix embeddings as the input. The German-GPT2 is trained on a large German corpus GC4 and can generate fluent german sentences. This step can be finally denoted as:

$$\text{Generator}(\{T_1, T_2..., T_m\} = \{S_1, S_2..., S_f\}$$

4 Experiments

4.1 Experiment Setup

We first describe our data preprocessing procedure and experiment settings.

4.1.1 Data Preprocessing

We first preprocess the raw data by dividing the video into smaller segments, or video slices, and match them with their corresponding ground truth German translations. To address a potential issue with the video extractor's encoding capacity, we adopt a downsampling strategy. Specifically, we select the first frame from every three frames in each video slice. Doing so reduces the number of frames and alleviates encoding challenges. Additionally, we encounter cases where certain video slices have fewer than 64 frames. To maintain consistency in video length, we append pure black frames to the end of these slices. To ensure compatibility with the video feature extractor's training environment, we resize each video frame to 224×224 dimension. This step guarantees that the model functions effectively within its designated parameters.

4.1.2 Experiment Setting

To enhance training efficiency, the parameters of the two backbone models are frozen, while the parameters of GPT2 are unfrozen after a certain iteration. This ensures that the randomly initialized transformer encoder does not compromise the language modeling ability of the GPT2 model. In our experiment, we set the batch size to 4, the learning rate to 5e-6, and changed the training parameters at iteration 66000. We employ an Adam (Kingma and Ba, 2015) as our optimizer and save the model checkpoint every 1000 iterations. The input and output lengths of GPT2 were fixed at 20, as we observed that most of the ground truth lengths were 20 or less, making this maximum length setting cover a significant portion of the training data. We set the number of heads in the multi-head attention to 8 and the prefix length for GPT2 to 4. Before feeding the embedding to the alignment block, the sequence length for translating the visual embedding was adjusted to 2×4, where 4 represents the GPT2 model's prefix number. Our model consists of 6 stacked encoder layers forming the alignment block. All experiments were conducted on NVIDIA GeForce GTX 1080 Ti with 11G memory.

4.2 Results

After extensive training and evaluation, our system achieves a BLEU (Papineni et al., 2002) score of 0.1 and a Chrf (Popovic, 2015) score of 7.6 in this shared task. These results are obtained from the official result submission platform. We present our experimental findings in comparison to the baseline model provided by the organizers, as shown in Table 2. Despite training our model solely on SRF, we outperform the baseline regarding the BLEURT (Sellam et al., 2020) score in SignSuisee and a combination of both datasets, which are considered out-of-domain evaluations for our model. However, it is important to note that our system falls significantly below the baselines and systems from other submissions. One potential explanation for this discrepancy could be that our system has not yet reached its optimal state, as the alignment block is trained from scratch, which could be quite challenging to converge. We conduct a fine-grained analysis in the following section to further investigate this hypothesis.

Original Subtitles	Generated Subtitles
Das Parlament muss nun auch die Städte ins Boot holen.<pad><pad><pad><pad><pad>	Der Schweiz. Sie werden in der Schweiz geboren. Deutschland.....
da fehlte oft das richtige Timing,<pad><pad><pad><pad><pad>	"Der Stadt Zürich. Zürich. Zürich. Zürich.. die Stadt Zürich. Zürich.. Zürich"
Am Samstagabend zunächst noch Föhn, dann wird es feuchter.<pad><pad><pad>	"Der Schweizer Regierungspräsidentin der Schweiz.. 20. 20. 20. 20. 20."
Danke, Andrea.<pad><pad><pad><pad>	"Der Film die Welt in den Abgrund. Rom. Deutschland....."
Es liegt an uns, Lösungen zu finden, um dieses Spiel zu gewinnen.<pad>	""Ich bin auch nicht, weil ich habe das nicht so viel. West.W.."

Table 3: Examples of our generated subtitles with their corresponding ground truth subtitles. We observe that 4 out of 5 of our generated sentences generate the same token for the first one and keep generating the same token at the end of its sentence. We try to analyze the reason for this in the following section.

4.3 Analysis

To analyze the reasons behind the failure of our system and its tendency to generate repetitive words in translations, we conduct a tSNE plot analysis of the visual embeddings before and after passing through the embedding alignment block. The results are presented in Figure 2. Upon examining the plot, we observe that the orange markers, representing the embeddings before alignment, were scattered, occupying a large area in the plot. In contrast, the blue crosses, corresponding to the embeddings after alignment, are densely concentrated in the middle of the plot. This stark contrast proves that the model loses its ability to differentiate between different visual features after projecting the embeddings from the I3D embedding space to the German-GPT2 embedding space. One potential explanation for this is that the embedding alignment block has not been effectively trained under the current training protocol. Further investigation is required to understand the underlying causes and devise appropriate solutions.

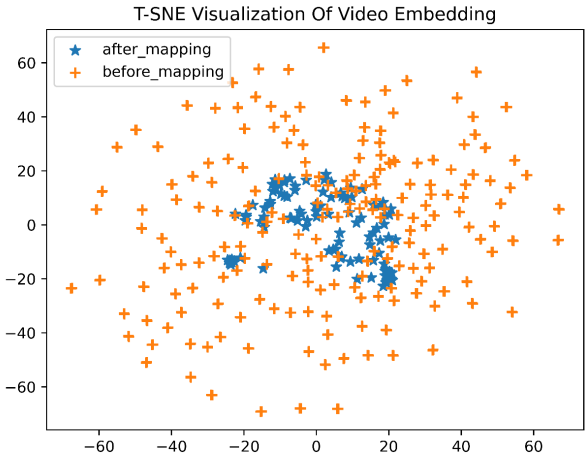

Figure 2: The tSNE comparison plot of the video embeddings before and after the embedding alignment block. We observe that the embeddings of different videos are dispersely distributed. However, they exhibit a denser distribution after alignment, which challenges generating coherent natural language descriptions.

4.4 Case Study

In Table 3, we present several instances of our generation using the data from the SRF dataset. The left column displays the ground truth sentences with a pad token appended at the end. In the right column, we showcase the generated sentences. Notably, 4 out of 5 of these sentences begin with "Der," and some consistently produce the same token, particularly in the final few positions. This further illustrates the subpar performance of our model. One possible explanation for this issue is the concentration of embeddings after the alignment block, which increases the likelihood of generating similar tokens. In the future, large-scale pertaining and appropriately leveraging large language models (OpenAI, 2023; Chan et al., 2023; Yu et al., 2023) and large multimodal foundation models (Zhu et al., 2023) may also be considered to improve the performance of this task further.

5 Conclusions

In conclusion, this paper presents the KnowComp system for the WMT23-SLT Sign Language Translation Shared Task. Our system utilizes two pretrained backbone models for visual feature extraction and translation text generation. However, this architecture fails, resulting in unsatisfactory perfor-

mance across all evaluation datasets. Our system's performance is significantly below the baseline's performance. We have identified a critical weakness in our model through further analysis, including embedding t-SNE plots and case studies. The embedding alignment block unexpectedly densifies all visual embeddings together, leading to the generator generating repeated tokens. To enhance our model's performance in future work, an appropriate data augmentation technique (Wang et al., 2023b,a; Gowda et al., 2022) can be implemented to help the alignment block distinguish different input features more efficiently. Also, future works can focus on whether further increasing the model capacity could help to mitigate the issue shown in the analysis section considering the advancing computation resources.

Acknowledgements

The authors would like to thank the committee of WMT2023, the organizers of the SLT task, and the anonymous reviewers. The authors of this paper were supported by the NSFC Fund (U20B2053) from the NSFC of China, the RIF (R6020-19 and R6021-20), and the GRF (16211520 and 16205322) from RGC of Hong Kong. We also thank the support from the UGC Research Matching Grants (RMGS20EG01-D, RMGS20CR11, RMGS20CR12, RMGS20EG19, RMGS20EG21, RMGS23CR05, RMGS23EG08).

References

Samuel Albanie, Gül Varol, Liliane Momeni, Triantafyllos Afouras, Joon Son Chung, Neil Fox, and Andrew Zisserman. 2020. BSL-1K: scaling up co-articulated sign language recognition using mouthing cues. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XI*, volume 12356 of *Lecture Notes in Computer Science*, pages 35–53. Springer.

Jiaxin Bai, Xin Liu, Weiqi Wang, Chen Luo, and Yangqiu Song. 2023a. Complex query answering on eventuality knowledge graph with implicit logical constraints. *CoRR*, abs/2305.19068.

Jiaxin Bai, Tianshi Zheng, and Yangqiu Song. 2023b. Sequential query encoding for complex query answering on knowledge graphs. *CoRR*, abs/2302.13114.

Necati Cihan Camgöz, Oscar Koller, Simon Hadfield, and Richard Bowden. 2020. Sign language transformers: Joint end-to-end sign language recognition and translation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 10020–10030. Computer Vision Foundation / IEEE.

João Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? A new model and the kinetics dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 4724–4733. IEEE Computer Society.

Chunkit Chan, Jiayang Cheng, Weiqi Wang, Yuxin Jiang, Tianqing Fang, Xin Liu, and Yangqiu Song. 2023. Chatgpt evaluation on sentence level relations: A focus on temporal, causal, and discourse relations. *CoRR*, abs/2304.14827.

Subhadeep Dey, Abhilash Pal, Cyrine Chaabani, and Oscar Koller. 2022. Clean text and full-body transformer: Microsoft's submission to the WMT22 shared task on sign language translation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 969–976, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Biyi Fang, Jillian Co, and Mi Zhang. 2017. Deepasl: Enabling ubiquitous and non-intrusive word and sentence-level sign language translation. In *Proceedings of the 15th ACM Conference on Embedded Network Sensor Systems, SenSys 2017, Delft, Netherlands, November 06-08, 2017*, pages 5:1–5:13. ACM.

Tianqing Fang, Quyet V. Do, Sehyun Choi, Weiqi Wang, and Yangqiu Song. 2023. CKBP v2: An expert-annotated evaluation set for commonsense knowledge base population. *CoRR*, abs/2304.10392.

Tianqing Fang, Weiqi Wang, Sehyun Choi, Shibo Hao, Hongming Zhang, Yangqiu Song, and Bin He. 2021a. Benchmarking commonsense knowledge base population with an effective evaluation dataset. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 8949–8964. Association for Computational Linguistics.

Tianqing Fang, Hongming Zhang, Weiqi Wang, Yangqiu Song, and Bin He. 2021b. DISCOS: bridging the gap between discourse knowledge and commonsense knowledge. In *WWW '21: The Web Conference 2021, Virtual Event / Ljubljana, Slovenia, April 19-23, 2021*, pages 2648–2659. ACM / IW3C2.

Shreyank N. Gowda, Marcus Rohrbach, Frank Keller, and Laura Sevilla-Lara. 2022. Learn2augment: Learning to composite videos for data augmentation in action recognition. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXXI*, volume 13691 of *Lecture Notes in Computer Science*, pages 242–259. Springer.

Dan Guo, Wengang Zhou, Houqiang Li, and Meng Wang. 2018. Hierarchical LSTM for sign language

translation. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 6845–6852. AAAI Press.

Mutian He, Tianqing Fang, Weiqi Wang, and Yangqiu Song. 2022. Acquiring and modelling abstract commonsense knowledge via conceptualization. *CoRR*, abs/2206.01532.

Hezhen Hu, Weichao Zhao, Wengang Zhou, and Houqiang Li. 2023. Signbert+: Hand-model-aware self-supervised pre-training for sign language understanding. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(9):11221–11239.

Kaiyu Huang, Peng Li, Jin Ma, Ting Yao, and Yang Liu. 2023. Knowledge transfer in incremental learning for multilingual neural machine translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 15286–15304. Association for Computational Linguistics.

Zifan Jiang, Amit Moryossef, Mathias Müller, and Sarah Ebling. 2023a. Signsuisse dsgs/lsf/lis lexicon.

Zifan Jiang, Mathias Müller, Sarah Ebling, Amit Moryossef, and Robin Ribback. 2023b. Srf dsgs daily news broadcast: video and original subtitle data.

Jichao Kan, Kun Hu, Markus Hagenbuchner, Ah Chung Tsoi, Mohammed Bennamoun, and Zhiyong Wang. 2021. Sign language translation with hierarchical spatio-temporalgraph neural network. *CoRR*, abs/2111.07258.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Kezhou Lin, Xiaohan Wang, Linchao Zhu, Ke Sun, Bang Zhang, and Yi Yang. 2023. Gloss-free end-to-end sign language translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 12904–12916. Association for Computational Linguistics.

Alireza Mohammadshahi, Vassilina Nikoulina, Alexandre Berard, Caroline Brun, James Henderson, and Laurent Besacier. 2022. Small-100: Introducing shallow multilingual machine translation model for low-resource languages. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 8348–8359. Association for Computational Linguistics.

Ron Mokady, Amir Hertz, and Amit H. Bermano. 2021. Clipcap: CLIP prefix for image captioning. *CoRR*, abs/2111.09734.

Mathias Müller, Malihe Alikhani, Eleftherios Avramidis, Richard Bowden, Annelies Braffort, Necati Cihan Camgöz, Sarah Ebling, Cristina España-bonet, Anne Goering, Roman Grundkiewicz, Mert Inan, Zifan Jiang, Oscar Koller, Amit Moryossef, Annette Rios, Dimitar Shterionov, Sandra Sidler-miserez, and Katja Tissi. 2023. Findings of the second WMT shared task on sign language translation (WMT-SLT23). In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Mathias Müller, Sarah Ebling, Eleftherios Avramidis, Alessia Battisti, Michèle Berger, Richard Bowden, Annelies Braffort, Necati Cihan Camgöz, Cristina España-Bonet, Roman Grundkiewicz, Zifan Jiang, Oscar Koller, Amit Moryossef, Regula Perrollaz, Sabine Reinhard, Annette Rios, Dimitar Shterionov, Sandra Sidler-Miserez, and Katja Tissi. 2022. Findings of the first WMT shared task on sign language translation (WMT-SLT22). In *Proceedings of the Seventh Conference on Machine Translation, WMT 2022, Abu Dhabi, United Arab Emirates (Hybrid), December 7-8, 2022*, pages 744–772. Association for Computational Linguistics.

Mathias Müller, Zifan Jiang, Amit Moryossef, Annette Rios, and Sarah Ebling. 2023. Considerations for meaningful sign language machine translation based on glosses. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 682–693. Association for Computational Linguistics.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Jiefu Ou, Benno Krojer, and Daniel Fried. 2023. Pragmatic inference with a CLIP listener for contrastive captioning. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 1904–1917. Association for Computational Linguistics.

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, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL.

Maja Popovic. 2015. chrf: character n-gram f-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation, WMT@EMNLP 2015, 17-18 September 2015, Lisbon, Portugal*, pages 392–395. The Association for Computer Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Stefan Schweter. 2020. German gpt-2 model.

Thibault Sellam, Dipanjan Das, and Ankur P. Parikh. 2020. BLEURT: learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7881–7892. Association for Computational Linguistics.

Bowen Shi, Diane Brentari, Gregory Shakhnarovich, and Karen Livescu. 2022. TTIC's WMT-SLT 22 sign language translation system. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 989–993, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Laia Tarres, Gerard I. Gállego, Xavier Giro-i nieto, and Jordi Torres. 2022. Tackling low-resourced sign language translation: UPC at WMT-SLT 22. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 994–1000, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Gül Varol, Liliane Momeni, Samuel Albanie, Triantafyllos Afouras, and Andrew Zisserman. 2021. Read and attend: Temporal localisation in sign language videos. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 16857–16866. Computer Vision Foundation / IEEE.

Weiqi Wang, Tianqing Fang, Wenxuan Ding, Baixuan Xu, Xin Liu, Yangqiu Song, and Antoine Bosselut. 2023a. CAR: conceptualization-augmented reasoner for zero-shot commonsense question answering. *CoRR*, abs/2305.14869.

Weiqi Wang, Tianqing Fang, Baixuan Xu, Chun Yi Louis Bo, Yangqiu Song, and Lei Chen. 2023b. CAT: A contextualized conceptualization and instantiation framework for commonsense reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 13111–13140. Association for Computational Linguistics.

Zhaowei Wang, Quyet V. Do, Hongming Zhang, Jiayao Zhang, Weiqi Wang, Tianqing Fang, Yangqiu Song, Ginny Y. Wong, and Simon See. 2023c. COLA: contextualized commonsense causal reasoning from the causal inference perspective. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 5253–5271. Association for Computational Linguistics.

Zhaowei Wang, Hongming Zhang, Tianqing Fang, Yangqiu Song, Ginny Y. Wong, and Simon See. 2022. Subeventwriter: Iterative sub-event sequence generation with coherence controller. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 1590–1604. Association for Computational Linguistics.

Changlong Yu, Weiqi Wang, Xin Liu, Jiaxin Bai, Yangqiu Song, Zheng Li, Yifan Gao, Tianyu Cao, and Bing Yin. 2023. Folkscope: Intention knowledge graph construction for e-commerce commonsense discovery. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 1173–1191. Association for Computational Linguistics.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *CoRR*, abs/2304.10592.