Tackling Low-Resourced Sign Language Translation: UPC at WMT-SLT 22

Laia Tarrés[1,2*] and **Gerard I. Gállego**[1,*] and **Xavier Giró-i-Nieto**[1] and **Jordi Torres**[2]
[1]*Universitat Politècnica de Catalunya, Spain*   [2]*Barcelona Supercomputing Center, Spain*

Abstract

This paper describes the system developed at the Universitat Politècnica de Catalunya for the Workshop on Machine Translation 2022 Sign Language Translation Task, in particular, for the sign-to-text direction. We use a Transformer model implemented with the Fairseq modeling toolkit. We have experimented with the vocabulary size, data augmentation techniques and pretraining the model with the PHOENIX-14T dataset. Our system obtains 0.50 BLEU score for the test set, improving the organizers' baseline by 0.38 BLEU. We remark the poor results for both the baseline and our system, and thus, the unreliability of our findings.

1 Introduction

The submission of the Universitat Politècnica de Catalunya (UPC) to the WMT22 Sign Language Translation (SLT) Task experimented with vocabulary size, data augmentation techniques and a pretrained system with the PHOENIX-14T dataset (Camgoz et al., 2018). Up to the author's knowledge, our implementation is the first to build on Fairseq, a popular modeling toolkit by Meta AI (Ott et al., 2019).

SLT is a highly complex task because sign language understanding requires a very precise estimation of the signer pose, especially, of its hands. In addition, sign languages have grammatical structures different from spoken languages, which prevents an easy knowledge transfer. Sign languages are represented in continuous and high-dimensional spaces, while the transcribed version of spoken languages are represented by discrete tokens of well-defined vocabularies. Moreover, the few available sign language datasets can be considered very low-resourced, as they typically contain less than one hundred thousand sentences (Goyal et al., 2022).

In particular, the total number of sentences from the two datasets provided in WMT-SLT22 is 17k.

We focus on the sign language translation task of WMT-SLT22 which requires participants to predict the translation in spoken language (written) from a sign language video. Specifically, it consists of translation from Swiss German Sign Language (DSGS) videos to German (DE) text.

The organizers of WMT 2022 SLT track propose a Transformer baseline (Müller et al., 2022) that achieves a very low BLEU (Papineni et al., 2002) score,[1] which indicates a very poor translation quality. The training data provided consists of two datasets: FocusNews (Müller et al., 2022b) and SRF (Müller et al., 2022a). The first one contains 197 episodes in DSGS that have an average length of 5 minutes, which amount for a total duration of 19 hours. The second dataset contains 29 videos from live sign language interpretation, that have an average of 30 minutes length, and totals a duration of 16 hours. The paired subtitles are given in Standard German from Switzerland, which is a dialect of German.

The organizers provide keypoints extracted with two off-the-shelf human pose estimators: OpenPose (Cao et al., 2019) and MediaPipe (Lugaresi et al., 2019).

2 Baseline system

For the sake of self-containment, we briefly define and discuss the baseline model proposed by the organizers for this task (Müller et al., 2022). This solution is built on sockeye (Hieber et al., 2017), with results for the official development (dev) partition.

Regarding the model implementation, they use a Transformer encoder-decoder, with a symmetrical number of layers for the encoder and decoder. The architecture has 6 layers, 8 heads, 2048 neurons in

---

*Equal contribution

[1]For simplicity, we refer to BLEU-4 score as BLEU score.

the feed forward network layers, and an embedding dimension of 512.

The input data are 2D OpenPose keypoints, concatenating the hands and body landmarks. This results in an array $X = x_0, \ldots, x_T$ where $x_t \in \mathbb{R}^{2K}$, K is the number of keypoints selected, and T the number of video frames.

Three baseline results were provided by the WMT-SLT22 organizers, shown in Table 1. The first two scores correspond to the models trained individually with each of the two benchmark datasets. The last result is from a model trained on both datasets, which obtains the lowest performance. We hypothesize this might be due to a domain shift between the FocusNews and SRF datasets. These BLEU scores are extremely low, as already noted by Müller et al. (2022). As a comparison, the SLT state-of-the-art BLEU score for PHOENIX-14T dataset (Camgoz et al., 2018) is 25.59 (Voskou et al., 2021) and for the How2Sign dataset (Duarte et al., 2021) is 1.25 (Duarte et al., 2022).

Train dataset	BLEU
FocusNews	0.216
SRF	**0.589**
FocusNews + SRF	0.157

Table 1: Results provided by the organizers for the official dev partition, which contains FocusNews + SRF samples. Scores on test partition, which also contains FocusNews + SRF samples, were not released by the organizers.

3 Method

Our submission also adopts a Transformer architecture, which we implement with the Fairseq sequence modeling toolkit (Ott et al., 2019). Up to the author's knowledge, this is the first time that Fairseq is used for sign language video understanding. We publish our source code [2], offering the SLT community a novel tool widely used in machine translation for spoken languages. On top of this, we experiment with the vocabulary size, data augmentation techniques and pretraining the system with PHOENIX-14T dataset. Details are described in this section.

3.1 Preprocessing steps

The WMT-SLT22 organizers provide both Open-Pose and MediaPipe keypoints from the body pose

[2]https://github.com/mt-upc/fairseq/tree/wmt-slt22

estimators. We choose MediaPipe poses because they provide 3D coordinates (x, y, z) normalized between $[0, 1]$. Moreover, based on our experience, OpenPose is more prone to errors, like detecting several people in videos when there is only one signer in the recording.

Although MediaPipe poses are available from WMT-SLT22, we re-extract them with *pose-format* (Moryossef, 2022). This library defines a standardized way of storing poses, and provides different functionalities to work with them. After the extraction, we obtain an array with the same shape as described in Section 2.

While the video recordings in the SRF dataset have a rate of 25 frames per second (fps), the FocusNews dataset present a frame rates of either 25, 30 or 50 fps. We perform cubic interpolation for the extracted poses, to achieve a unified frame rate of 25fps, using the *interpolate_fps* function from *pose-format*.

We build an independent vocabulary with the training split of each dataset. For the SRF dataset, organizers provide parallel and monolingual data. The latter contains all German subtitles, including much more sentences than the former. Therefore, we choose to build our SRF vocabulary from the monolingual data. Note that this data does not have paired (or parallel) videos, so it can not be used for training the model.

3.2 Architecture

We build a smaller Transformer architecture than the WMT-SLT22 baseline, since we observed signs of overfitting when checking the training losses. With the baseline architecture the system was simply generating the most frequent words from the training set. Therefore, we concluded that training a smaller model would hinder the overfitting and might improve the results.

In particular, our Transformer model has a symmetrical structure for the encoder and decoder, with 3 layers, 4 heads, 1024 neurons in the feed forward network layers, and an embedding dimension of 256.

3.3 Data augmentation

The performance of deep learning models depends on the quality, quantity, and domain of training data, however datasets that provide all qualities needed for models are often not available. To diminish the consequences of the data scarcity, a common practice is to apply data augmentation techniques.

This approach is really useful to improve the performance of models, and makes them more robust to slight changes in the input data.

We made use of the *augment2d* function from the *pose-format* library, which allows applying various transformations directly to the keypoints, such as random rotation, shear effect, and scaling. Specifically, the rotation angle in radians, the shear factor and the scaling factor we apply are obtained by sampling from a normal distribution with zero mean and standard deviation of 0.2. Some examples of augmented poses are shown in Figure 1.
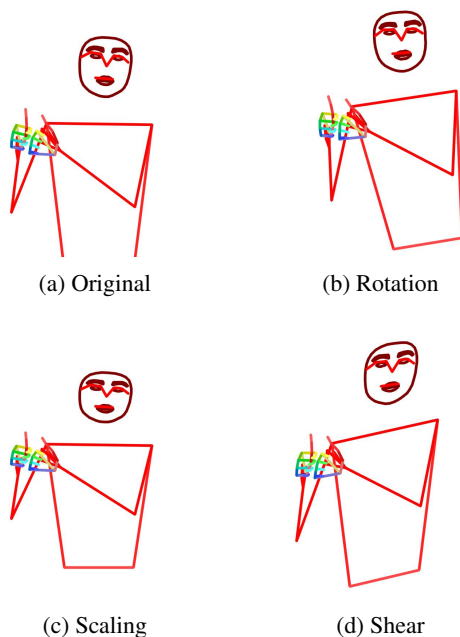

(a) Original

(b) Rotation

(c) Scaling

(d) Shear

Figure 1: Data augmentation transformations.

3.4 Vocabulary size

State-of-the-art Machine Translation systems use subword dictionaries instead of word-level vocabularies (Tran et al., 2021; Yang et al., 2021). These dictionaries are built by decomposing words into smaller pieces based on their frequency (Sennrich et al., 2016). Analogously to the baseline approach, we use SentencePiece tokenizer to obtain the subword vocabulary (Kudo and Richardson, 2018).

The vocabulary size is a hyperparameter that, in practice, is either chosen arbitrarily or via trial-and-error (Salesky et al., 2020). However, it has been studied that using a greater vocabulary size might help in reducing the class imbalance present in the training dataset (Gowda and May, 2020). The baseline used a vocabulary size of 1000 subwords, and we experiment with 2000 and 4000. Our goal was to detect whether downsampling a vocabulary of 20k unique words, for FocusNews dataset, to

1k subwords may be oversimplifying the problem.

3.5 Pretraining with PHOENIX14-T

We also explore transfer learning to overcome the data scarcity problem. Given that the scope of WMT22 is on Swiss German Sign Language, we chose the PHOENIX14-T dataset (Camgoz et al., 2018). For the three datasets, the target language is either spoken (written) German or Standard German from Switzerland. However, PHOENIX14-T presents an important domain shift with respect to the WMT-SLT22 datasets, since it is limited to live interpretation of weather forecast on broadcast TV.

We pretrained our model with the PHOENIX14-T dataset. In order to implement the transfer learning pipeline, we built a vocabulary by merging the training data from the three available datasets: FocusNews, SRF and PHOENIX14-T.

3.6 Checkpoint Averaging

We choose the best-performing models as those with the best BLEU dev scores. However, in the best-performing cases, and as thee final step in our trainings, we average the weights of the 3 best model checkpoints for each run. This methodology, which was firstly introduced by (Vaswani et al., 2017), proved to be a useful and easy to implement technique to generate more robust predictions on Transformers (Popel and Bojar, 2018), and has been widely used in the Machine Translation field.

4 Results

We train our systems with FocusNews and SRF. For both datasets, we provide results for the dev set, which contains recordings from FocusNews and SRF. We choose the 6 best-performing models for dev to submit to the official challenge submission.

Table 2 shows the results of models trained with FocusNews. We notice an improvement in the performance of our baseline implemented iin Fairseq with respect to the one from the organizers (1-2). We then analyze the effect of the vocabulary size based on the BLEU obtained for the dev set. We notice that using 4000 subwords provides the best results in terms of vocabulary size (2-4). Therefore, we choose this configuration to experiment with pretraining and data augmentation. For this set of experiments, we see a slight improvement when fine-tuning a network that has been pretrained with the PHOENIX14-T dataset (6). However, adding data augmentation (5) or a combination of both

ID	System	Vocab. size	Data Augm.	Pretrain	BLEU (dev)	BLEU (test)
1	Baseline (Müller et al., 2022)	1k			0.22	-
2	Our Baseline (§3.2)	1k			0.47	**0.50**
3	2 + 2k subwords	2k			0.47	-
4	2 + 4k subwords	4k			0.62	-
5	4 + data augmentation	4k	✓		0.51	-
6	4 + pretrain w/ Phoenix	4k		✓	**0.64**	0.41
7	6 + data augmentation	4k	✓	✓	0.48	-
8	6 + checkpoint average	4k		✓	0.57	0.35

Table 2: Results of models trained with the FocusNews dataset. *BLEU (dev)* corresponds to the results obtained in the challenge dev set, and *BLEU (test)* to the results extracted by the organizers using the official test partition (Müller et al., 2022). In bold are the best results for each partition.

ID	System	Vocab. size	Data Augm.	Pretrain	BLEU (dev)	BLEU (test)
1	Baseline (Müller et al., 2022)	1k			0.59	0.12
2	Baseline (§3.2)	1k			0.64	-
3	2 + 2k subwords	2k			**0.69**	**0.28**
4	2 + 4k subwords	4k			0.63	**0.28**
5	3 + checkpoint average	2k			0.60	0.24

Table 3: Results of models trained with the SRF dataset. *BLEU (dev)* corresponds to the results obtained in the challenge dev set, and *BLEU (test)* to the results extracted by the organizers using the official test partition (Müller et al., 2022). In bold are the best results for each partition.

data augmentation and pretraining (7) does not improve the results. Checkpoint averaging does not bring an improvement either (8). Surprisingly, we cannot extract the same conclusions from the test results. After we received the preliminary findings from the organizers (Müller et al., 2022), we found that the best-performing model for dev (6) was not the best for test, but the simplest one (2).

Results of models trained with SRF are presented in Table 3. Similarly to FocusNews, our proposed baseline improves the organizers' in this dataset (1-2). However, in this case, the optimal number of subwords is 2000, with a slight improvement over other vocabulary sizes (2-4). Similar to the FocusNews case, we observe that checkpoint averaging does not improve results. Due to technical issues and time limitations, experiments with SRF are limited to analyzing the vocabulary size, hence the best experiment with FocusNews could not be replicated using SRF. Although results for dev are better for models trained with SRF, results

for test show a poorer performance than Focus-News models.

We optimized our systems to obtain the best BLEU metric, without taking other metrics into consideration. However, organizers also compute chrF++ (Popović, 2017) and BLEURT (Sellam et al., 2020) metrics (Müller et al., 2022). We find that the BLEURT score shows a similar performance than BLEU. However, for the chrF++ metric, which correlates better with respect to human relative rankings, our models score lower compared to other submissions.

We provide some examples of the sentences generated by our best-performing model in Table 4. We see that the translations are poor and lack correlation with the video, which relates to the poor overall performance in the BLEU metric.

5 Discussion and Conclusion

We proposed a pipeline to tackle the Sign Language Translation Task for the newly released datasets:

Ref.:	*Letztes Jahr haben viele Gehörlosen-Medien über die erste Gehörlosen Universität in Europa in Bad Kreuznach in Deutschland berichtet.*
Pred.:	*Am letzten Samstag, 22. Mai, in der Schweiz, in der Schweiz, in der Schweiz, GSC Aarau, GSC Aarau.*
Ref.:	*Dazu sind 4 Politiker eingeladen, die über für Behinderte wichtige Themen diskutieren werden, wie zum Beispiel TV-Untertitel, UNO Konvention für Behinderte usw.*
Pred.:	*Das Ziel ist es, dass es, dass die nempflichkeit für die nempflichkeitssetzen kann.*
Ref.:	*Der Deutsche Fernsehsender ZDF bietet Filme im Internet mit Untertitel an, sofern der Film vorher im Fernseher mit Untertitel ausgestrahlt wurde.*
Pred.:	*Der Schweizerische Gehörlosenbund SGB-FSS organisiert mit dem Schweizerischen Gehörlosenbund SGB-FSS, der Gehörlosen Sportverband der Gehörlosen Sportverband der Gehörl osen Sportverband der Gehörlosen Sportverband der Gehörlosen Sportverband der Gehörlosen Sportverband.*

Table 4: Example reference and predictions from our best-performing model for the official dev partition.

Focusnews and SRF. Our fresh implementation with Fairseq slightly improved the baseline provided by the organizers.

Our findings showed that when training with FocusNews, our baseline system has the best performance for test. The changes in vocabulary size did not affect the test performance. Furthermore, we showed that using checkpoint averaging does not help for this task. In all cases, we still think that the results are extremely low, which indicate a really poor translation, and there is potential unreliability of the findings due to the close to 0 BLEU score.

We consider the results we obtained can be further improved, so we leave some experiments for future work. Firstly, we believe that a joint training from the two provided datasets could boost the per-

formance of the models by bridging the domain gap between these datasets. Secondly, we did not see any improvement by pretraining the models with PHOENIX14-T dataset. However, we think that solving the WMT-SLT22 task must require some sort of transfer learning from a pretrained model.

Limitations

As stated by the organizers, results are still poor. When inspecting the predictions, it seems evident that the model is learning the most frequent words in the vocabulary, thus failing to provide meaningful predictions from the video. We consider that this is due to the high complexity of the task paired with a lack of data.

	Length	**Words**	**Ratio**
FocusNews	19 h	21 k	0.90
SRF	16 h	19 k	0.84
How2Sign	79 h	16 k	4.93
PHOENIX14-T	11 h	3 k	3.67

Table 5: Comparison between SLT datasets based on the duration of the videos (in hours) and number of unique words (in thousands) in the vocabulary. The Ratio column provides an indication of the difficulty of solving the SLT task for each dataset.

As shown in Table 5, the ratio between the training data and vocabulary size is much lower compared to other SLT datasets such as How2Sign and PHOENIX14-T. We take these results as an indication of the complexity of the datasets. We hypothesize that the low BLEU scores reported in the baseline, may be caused by the low ratio between video hours per unique words in the vocabulary, hence the dataset might be too complex. Therefore, we decide to experiment with data augmentation since it artificially improves the amount of training samples.

Our experiments with the SRF dataset, have been computationally expensive. Due to technical details, we have to read full sequences of around 30 minutes every time we load a sample. Processing them, even with the dimensional reduction provided by pose estimators, has been a challenge for the machines of our academic lab. A slightly better set of results might have been produced, but we still think it would not have made a significant difference.

In addition, we present the input poses as a sequence of one-dimensional arrays with the XYZ

coordinates. We think that this may not be the optimal way of processing the graph-like structure from poses. Using graph neural networks to preprocess input poses (Yan et al., 2018, 2019; Bull et al., 2020; Jiang et al., 2021) might be an interesting approach to improve SLT results.

We also lacked the time to experiment with other features, such as processing the RGB videos with a convolutional network (Vaezi Joze and Koller, 2019; Li et al., 2020; Albanie et al., 2020). We tried extracting i3d features fine-tuned on PHOENIX14-T, but the output features contained an excessive number of 0's, and we never run a proper experiment with this setup. This might happen because the visual appearance of the videos is too different between PHOENIX14-T and the WMT-SLT22 datasets, specifically due to the spatial segmentation of the signer in the frames provided in these datasets.

Acknowledgements

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 *European Conference on Computer Vision (ECCV)*. Springer.

Hannah Bull, Michèle Gouiffès, and Annelies Braffort. 2020. Automatic segmentation of sign language into subtitle-units. In *European Conference on Computer Vision*, pages 186–198. Springer.

Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. 2018. Neural sign language translation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7784–7793.

Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. 2019. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Amanda Duarte, Samuel Albanie, Xavier Giró-i Nieto, and Gül Varol. 2022. Sign language video retrieval with free-form textual queries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14094–14104.

Amanda Duarte, Shruti Palaskar, Lucas Ventura, Deepti Ghadiyaram, Kenneth DeHaan, Florian Metze, Jordi Torres, and Xavier Giro-i Nieto. 2021. How2Sign: A Large-scale Multimodal Dataset for Continuous American Sign Language. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Thamme Gowda and Jonathan May. 2020. Finding the optimal vocabulary size for neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3955–3964, Online. Association for Computational Linguistics.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Felix Hieber, Tobias Domhan, Michael Denkowski, David Vilar, Artem Sokolov, Ann Clifton, and Matt Post. 2017. Sockeye: A toolkit for neural machine translation.

Songyao Jiang, Bin Sun, Lichen Wang, Yue Bai, Kunpeng Li, and Yun Fu. 2021. Skeleton aware multimodal sign language recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Dongxu Li, Cristian Rodriguez, Xin Yu, and Hongdong Li. 2020. Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1459–1469.

Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. 2019. Mediapipe: A framework for building perception pipelines. *ArXiv*, abs/1906.08172.

Amit Moryossef. 2022. Complete toolkit for working with poses. https://github.com/AmitMY/pose-format/.

Mathias Müller, Sarah Ebling, Necati Cihan Camgöz, Zifan Jian, Alessia Battisti, Katja Tissi, Sandra Sidler-Miserez, Regula Perrollaz, Michèle Berger, Sabine Reinhard, Amit Bar-Ilan University Moryossef, Annette Rios, Richard Bowden, Ryan Wong, Robin Ribback, and Severine Schori. 2022a. WMT-SLT SRF: Training data for the WMT shared task on sign

language translation (videos, subtitles). We additionally acknowledge funding through the Innosuisse Flagship "Inclusive Information and Communication Technologies" (IICT) (grant agreement no. PFFS-21-47).

Mathias Müller, Sarah Ebling, Necati Cihan Camgöz, Zifan Jiang, Alessia Battisti, Amit Moryossef, Annette Rios, Richard Bowden, and Ryan Wong. 2022b. WMT-SLT FocusNews: Training data for the WMT shared task on sign language translation. We additionally acknowledge funding through the Innosuisse Flagship "Inclusive Information and Communication Technologies" (IICT) (grant agreement no. PFFS-21-47).

Mathias Müller, Sarah Ebling, Avramidis Eleftherios, 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, Katja Tissi, and Davy Van Landuyt. 2022. Findings of the WMT 2022 shared task on sign language translation. In *Proceedings of the Seventh Conference on Machine Translation*, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Mathias Müller, Annette Rios, and Amit Moryossef. 2022. Sockeye baseline models for sign language translation. https://github.com/bricksdont/sign-sockeye-baselines.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

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.

Martin Popel and Ondřej Bojar. 2018. Training tips for the transformer model. *The Prague Bulletin of Mathematical Linguistics*, 110:43–70.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Elizabeth Salesky, Andrew Runge, Alex Coda, Jan Niehues, and Graham Neubig. 2020. Optimizing segmentation granularity for neural machine translation. *Machine Translation*, 34(1):41–59.

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

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Chau Tran, Shruti Bhosale, James Cross, Philipp Koehn, Sergey Edunov, and Angela Fan. 2021. Facebook AI's WMT21 news translation task submission. In *Proceedings of the Sixth Conference on Machine Translation*, pages 205–215, Online. Association for Computational Linguistics.

Hamid Vaezi Joze and Oscar Koller. 2019. Ms-asl: A large-scale data set and benchmark for understanding american sign language. *The British Machine Vision Conference (BMVC)*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Andreas Voskou, Konstantinos P. Panousis, Dimitrios Kosmopoulos, Dimitris N. Metaxas, and Sotirios Chatzis. 2021. Stochastic transformer networks with linear competing units: Application to end-to-end sl translation.

Sijie Yan, Yuanjun Xiong, Wangm Jingbo, and Dahua Lin. 2019. Mmskeleton. https://github.com/open-mmlab/mmskeleton.

Sijie Yan, Yuanjun Xiong, and Dahua Lin. 2018. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *AAAI*.

Jian Yang, Shuming Ma, Haoyang Huang, Dongdong Zhang, Li Dong, Shaohan Huang, Alexandre Muzio, Saksham Singhal, Hany Hassan, Xia Song, and Furu Wei. 2021. Multilingual machine translation systems from Microsoft for WMT21 shared task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 446–455, Online. Association for Computational Linguistics.