Data Augmentation for Sign Language Gloss Translation

Amit Moryossef amitmoryossef@gmail.com
Bar-Ilan University

Kayo Yin kayoy@cs.cmu.edu
Language Technologies Institute, Carnegie Mellon University

Graham Neubig gneubig@cs.cmu.edu
Language Technologies Institute, Carnegie Mellon University

Yoav Goldberg yogo@cs.biu.ac.il
Bar-Ilan University, Allen Institute for AI

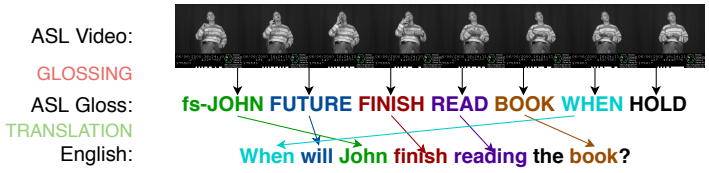
Abstract

Sign language translation (SLT) is often decomposed into *video-to-gloss* recognition and *gloss-to-text* translation, where a gloss is a sequence of transcribed spoken-language words in the order in which they are signed. We focus here on gloss-to-text translation, which we treat as a low-resource neural machine translation (NMT) problem. However, unlike traditional low-resource NMT, gloss-to-text translation differs because gloss-text pairs often have a higher lexical overlap and lower syntactic overlap than pairs of spoken languages. We exploit this lexical overlap and handle syntactic divergence by proposing two rule-based heuristics that generate pseudo-parallel gloss-text pairs from monolingual spoken language text. By pre-training on this synthetic data, we improve translation from American Sign Language (ASL) to English and German Sign Language (DGS) to German by up to 3.14 and 2.20 BLEU, respectively.

1 Introduction

Sign language is the most natural mode of communication for the Deaf. However, in a predominantly hearing society, they often resort to lip-reading, text-based communication, or closed-captioning to interact with others. Sign language translation (SLT) is an important research area that aims to improve communication between signers and non-signers while allowing each party to use their preferred language. SLT consists of translating a sign language (SL) video into a spoken language (SpL) text, and current approaches often decompose this task into two steps: (1) *video-to-gloss*, or continuous sign language recognition (CSLR) (Cui et al., 2017; Camgoz et al., 2018); (2) *gloss-to-text*, which is a text-to-text machine translation (MT) task (Camgoz et al., 2018; Yin and Read, 2020b).

In this paper, we focus on gloss-to-text translation. SL data and resources are often scarce, or nonexistent (§2; Bragg et al. (2019)). Gloss-to-text translation is, therefore, an example of an extremely low-resource MT task. However, while there is extensive literature on low-resource MT between spoken languages (Sennrich et al., 2016a; Zoph et al., 2016; Xia et al., 2019; Zhou et al., 2019), the dissimilarity between sign and spoken languages calls for novel methods. Specifically, as SL glosses borrow the lexical elements from their ambient spoken language, handling syntax and morphology poses greater challenges than lexeme translation (§3).



(a) ASL video with gloss annotation and English translation

(b) Data augmentation and training

Figure 1: Real and synthetic gloss-spoken pairs.

In this work, we (1) discuss the scarcity of SL data and quantify how the relationship between a sign and spoken language pair is different from a pair of two spoken languages; (2) show that the *de facto* method for data augmentation using back-translation is not viable in extremely low-resource SLT; (3) propose two rule-based heuristics that exploit the lexical overlap and handles the syntactic divergence between sign and spoken language, to synthesize pseudo-parallel gloss/text examples (Figure 1b); (4) demonstrate the effectiveness of our methods on two sign-to-spoken language pairs.

2 Background

Sign Language Glossing SLs are often transcribed word-for-word using a spoken language through *glossing* to aid in language learning, or automatic sign language processing (Ormel et al., 2010). While many SL glosses are words from the ambient spoken language, glossing preserves SL's original syntactic structure and therefore differs from translation (Figure 1a).

Data Scarcity While standard machine translation architectures such as the Transformer (Vaswani et al., 2017) achieve reasonable performance on gloss-to-text datasets (Yin and Read, 2020a; Camgoz et al., 2020), parallel SL and spoken language corpora, especially those with gloss annotations, are usually far more scarce when compared with parallel corpora that exist between many spoken languages (Table 1).

	Language Pair	# Parallel Gloss-Text Pairs	Vocabulary Size (Gloss / Spoken)
Signum (von Agris and Kraiss, 2007)	DGS-German	780	565 / 1,051
NCSLGR (SignStream, 2007)	ASL-English	1,875	2,484 / 3,104
RWTH-PHOENIX-Weather 2014T (Camgoz et al., 2018)	DGS-German	7,096 + 519 + 642	1,066 / 2,887 + 393 / 951 + 411 / 1,001
Dicta-Sign-LSF-v2 (Limsi, 2019)	French SL-French	2,904	2,266 / 5,028
The Public DGS Corpus (Hanke et al., 2020)	DGS-German	63,912	4,694 / 23,404

Table 1: Some publicly available SL corpora with gloss annotations and spoken language translations.

3 Sign vs. Spoken Language

Due to the paucity of parallel data for gloss-to-text translation, we can treat it as a low-resource translation problem and apply existing techniques for improving accuracy in such settings.

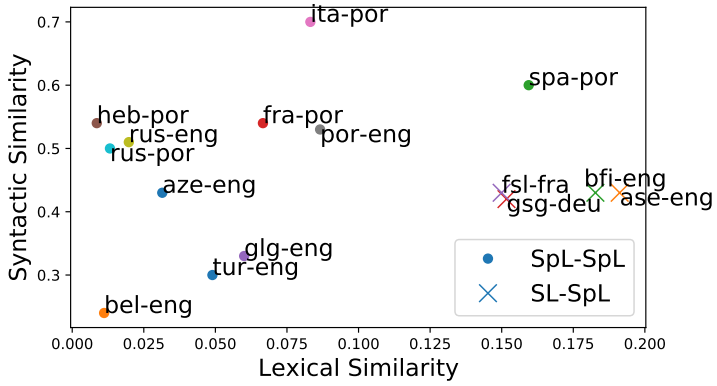


Figure 2: Lexical and syntactic similarity between different language pairs denoted by their ISO639-2 codes.

However, we argue that the relationship between glossed SLs and their spoken counterparts is different from the usual relationship between two spoken languages. Specifically, glossed SLs are *lexically similar but syntactically different* from their spoken counterparts. This contrasts heavily with the relationship among spoken language pairs where lexically similar languages tend also to be syntactically similar the great majority of the time.

To demonstrate this empirically, we adopt measures from (Lin et al., 2019) to measure the lexical and syntactic similarity between languages, two features also shown to be positively correlated with the effectiveness of performing cross-lingual transfer in MT.

Lexical similarity between two languages is measured using word overlap:

$$o_w = \frac{|T_1 \cap T_2|}{|T_1| + |T_2|}$$

where T_1 and T_2 are the sets of types in a corpus for each language. The word overlap between spoken language pairs is calculated using the TED talks dataset (Qi et al., 2018). The overlap between sign-spoken language pairs is calculated from the corresponding corpora in Table 1.

Syntactic similarity between two languages is measured by $1-d_{syn}$ where d_{syn} is the syntactic distance from (Littell et al., 2017) calculated by taking the cosine distance between syntactic features adapted from the World Atlas of Language Structures (Dryer and Haspelmath, 2013).

Figure 2 shows that sign-spoken language pairs are indeed outliers with lower syntactic similarity and higher lexical similarity. We seek to leverage this fact and the high availability of monolingual spoken language data to compensate for the scarcity of SL resources. In the following section, we propose data augmentation techniques using word order modifications to create synthetic sign gloss data from spoken language corpora.

4 Data Augmentation

This section discusses methods to improve gloss-to-text translation through data augmentation, specifically those that take monolingual corpora of standard spoken languages and generate pseudo-parallel "gloss" text. We first discuss a standard way of doing so, back-translation, point out its potential failings in the SL setting, then propose a novel rule-based data augmentation algorithm.

4.1 Back-translation

Back-translation (Irvine and Callison-Burch, 2013; Sennrich et al., 2016a) automatically creates pseudo-parallel sentence pairs from monolingual text to improve MT in low-resource settings. However, back-translation is only effective with sufficient parallel data to train a functional MT model, which is not always the case in extremely low-resource settings (Currey et al., 2017), and particularly when the domain of the parallel training data and monolingual data to be translated are mismatched (Dou et al., 2020).

4.2 Proposed Rule-based Augmentation Strategies

Given the limitations of standard back-translation techniques, we next move to the proposed method of using rule-based heuristics to generate SL glosses from spoken language text.

General rules The differences in SL glosses from spoken language can be summarized by (1) A lack of word inflection, (2) An omission of punctuation and individual words, and (3) Syntactic diversity.

We, therefore, propose the corresponding three heuristics to generate pseudo-glosses from spoken language: (1) Lemmatization of spoken words; (2) POS-dependent and random word deletion; (3) Random word permutation.

We use spaCy (Honnibal and Montani, 2017) for (1) lemmatization and (2) POS tagging to only keep nouns, verbs, adjectives, adverbs, and numerals. We also drop the remaining tokens with probability $p = 0.2$, and (3) randomly reorder tokens with maximum distance $d = 4$.

Language-specific rules While random permutation allows some degree of robustness to word order, it cannot capture all aspects of syntactic divergence between signed and spoken language. Therefore, inspired by previous work on rule-based syntactic transformations for reordering in MT (Collins et al., 2005; Isozaki et al., 2010; Zhou et al., 2019), we manually devise a shortlist of syntax transformation rules based on the grammar of DGS and German.

We perform lemmatization and POS filtering as before. In addition, we apply compound splitting (Tuggener, 2016) on nouns and only keep the first noun, reorder German SVO sentences to SOV, move adverbs and location words to the start of the sentence, and move negation words to the end. We provide a detailed list of rules in Appendix A.

5 Experimental Setting

5.1 Datasets

DGS & German RWTH-PHOENIX-Weather 2014T (Camgoz et al., 2018) is a parallel corpus of 8,257 DGS interpreted videos from the Phoenix[1] weather news channel, with corresponding SL glosses and German translations.

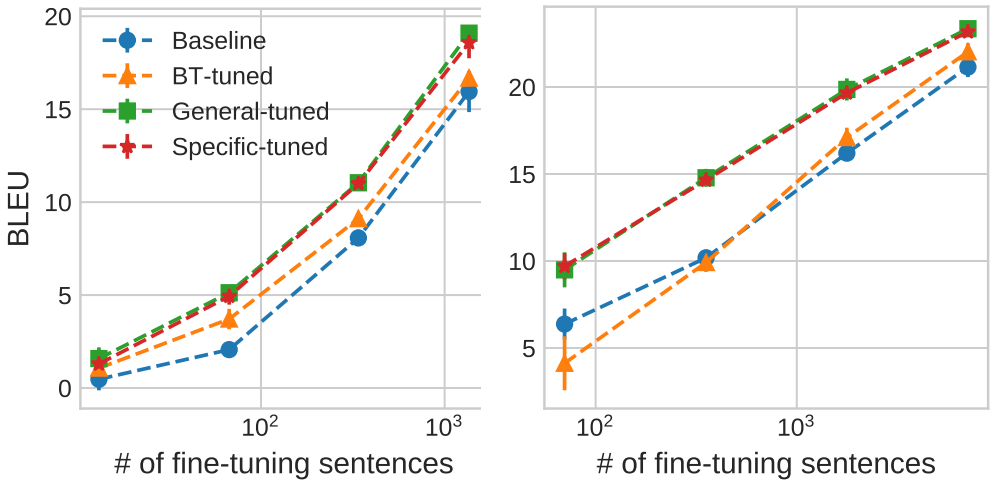
To obtain monolingual German data, we crawled tagesschau[2] and extracted news caption files containing the word "wetter" (German for "weather"). We split the 1,506 caption files into 341,023 German sentences using the spaCy sentence splitter and generate synthetic glosses using our methods described in §4.

ASL & English The NCSLGR dataset (SignStream, 2007) is a small, general domain dataset containing 889 ASL videos with 1,875 SL glosses and English translations.

We use ASLG-PC12 (Othman and Jemni, 2012), a large synthetic ASL gloss dataset created from English text using rule-based methods with 87,710 publicly available examples, for our experiments on ASL-English with language-specific rules. We also create another synthetic variation of this dataset using our proposed general rule-based augmentation.

[1] www.phoenix.de

[2] www.tagesschau.de



(a) NCSLGR (ASL)

(b) PHOENIX (DGS)

Figure 3: Translation results using various amounts of annotated parallel data.

5.2 Baseline Setup

We first train a **Baseline** system on the small manually annotated SL dataset we have available in each language pair. The model architecture and training method are based on Yin and Read (2020b)'s Transformer gloss-to-text translation model. While previous work (**Yin and Read Reimpl.**) used word-level tokenization, for Baseline and all other models described below, we instead use BPE tokenization (Sennrich et al. (2016b); with 2,000 BPE codes) for efficiency and simple handling of unknown words. For all tested methods, we repeat every experiment 3 times to account for variance in training.

5.3 Pre-training on Augmented Data

For **General-*pre*** and **Specific-*pre***, we pre-train a tokenizer and translation model on pseudo-parallel data obtained using general and language-specific rules respectively, until the accuracy on the synthetic validation set drops. We test both models on the parallel SL dataset in a zero-shot setting.

For **BT-*tuned***, **General-*tuned*** and **Specific-*tuned***, we take models pre-trained on pseudo-parallel data obtained with either back-translation, general rules, or language-specific rules, and continue training with half of the training data taken from the synthetic pseudo-parallel data and the other half taken from the real SL data. Then, we fine-tune these models on the real SL data and evaluate them on the test set.

6 Results

We evaluate our models across all datasets and sizes using SacreBLEU (v1.4.14) (Post, 2018) and COMET (*wmt-large-da-estimator-1719*) (Rei et al., 2020). We also compare our results to previous work on PHOENIX in Table 2. Detailed scores for each experiment are provided in Appendix C.

First, we note results on General-*pre* and Specific-*pre*. Interestingly, the scores are non-

	PHOENIX		NCSLGR	
	BLEU↑	COMET↑	BLEU↑	COMET↑
Yin and Read Reimpl.[4]	22.17	-2.93	-	-
Baseline	21.15	-5.74	15.95	-61.00
General-*pre* (0-shot)	3.95	-69.09	0.97	-135.99
Specific-*pre* (0-shot)	7.26	-53.14	0.95	-134.13
BT-*tuned*	**22.02**	**6.84**	16.67	**-51.86**
General-*tuned*	**23.35**	**13.65**	**19.09**	**-34.50**
Specific-*tuned*	**23.17**	**11.70**	**18.5**8	**-39.96**

Table 2: Results of our different models on PHOENIX and NCSLGR. We **bold** scores statistically significantly higher than baseline at the 95% confidence level.

negligible, demonstrating that the model can learn with *only* augmented data.[3] Moreover, on PHOENIX Specific-*pre* achieves significantly better performance than General-*pre*, which suggests our hand-crafted syntax transformations effectively expose the model to the divergence between DGS and German during pre-training.

Next, turning to the *tuned* models, we see that Specific and General outperform both the baseline and BT by large margins, demonstrating the effectiveness of our proposed methods. Interestingly, General-*tuned* performs slightly better, in contrast to the previous result. We posit that, similarly to previously reported results on sampling-based back translation (Edunov et al., 2018), General is benefiting from the diversity provided by sampling multiple reordering candidates, even if each candidate is of lower quality.

Looking at Figure 3, we see that the superior performance of our methods holds for all data sizes, but it is particularly pronounced when the parallel-data-only baseline achieves moderate BLEU scores in the range of 5-20. This confirms that BT is not a viable data augmentation method when parallel data is not plentiful enough to train a robust back-translation system.

7 Implications and Future Work

Consistent improvements over the baseline across two language pairs by our proposed rule-based augmentation strategies demonstrate that data augmentation using monolingual spoken language data is a promising approach for sign language translation.

Given the efficiency of our general rules compared to language-specific rules, future work may also include a more focused approach on specifically pre-training the target-side decoder with spoken language sentences so that by learning the syntax of the target spoken language, it can generate fluent sentences from sign language glosses having little to no parallel examples during training.

[3] In contrast, merely outputting the source sentence results in 1.36 BLEU, -90.28 COMET on PHOENIX and 1.5 BLEU, -119.45 COMET on NCSLGR.

[4] The original work achieves 23.32 BLEU; correspondence with the authors has led us to believe that the discrepancy is due to different versions of the underlying software.

References

Bragg, D., Koller, O., Bellard, M., Berke, L., Boudreault, P., Braffort, A., Caselli, N., Huenerfauth, M., Kacorri, H., Verhoef, T., et al. (2019). Sign language recognition, generation, and translation: An interdisciplinary perspective. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility*, pages 16–31.

Camgoz, N. C., Hadfield, S., Koller, O., Ney, H., and Bowden, R. (2018). Neural sign language translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Camgoz, N. C., Koller, O., Hadfield, S., and Bowden, R. (2020). Sign language transformers: Joint end-to-end sign language recognition and translation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Collins, M., Koehn, P., and Kučerová, I. (2005). Clause restructuring for statistical machine translation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 531–540, Ann Arbor, Michigan. Association for Computational Linguistics.

Cui, R., Liu, H., and Zhang, C. (2017). Recurrent convolutional neural networks for continuous sign language recognition by staged optimization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7361–7369.

Currey, A., Miceli Barone, A. V., and Heafield, K. (2017). Copied monolingual data improves low-resource neural machine translation. In *Proceedings of the Second Conference on Machine Translation*, pages 148–156, Copenhagen, Denmark. Association for Computational Linguistics.

Dou, Z.-Y., Anastasopoulos, A., and Neubig, G. (2020). Dynamic data selection and weighting for iterative back-translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5894–5904, Online. Association for Computational Linguistics.

Dryer, M. S. and Haspelmath, M., editors (2013). *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Edunov, S., Ott, M., Auli, M., and Grangier, D. (2018). Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium. Association for Computational Linguistics.

Hanke, T., Schulder, M., Konrad, R., and Jahn, E. (2020). Extending the Public DGS Corpus in size and depth. In *Proceedings of the LREC2020 9th Workshop on the Representation and Processing of Sign Languages: Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives*, pages 75–82, Marseille, France. European Language Resources Association (ELRA).

Honnibal, M. and Montani, I. (2017). spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Irvine, A. and Callison-Burch, C. (2013). Combining bilingual and comparable corpora for low resource machine translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 262–270, Sofia, Bulgaria. Association for Computational Linguistics.

Isozaki, H., Sudoh, K., Tsukada, H., and Duh, K. (2010). Head finalization: A simple reordering rule for sov languages. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 244–251.

Klein, G., Kim, Y., Deng, Y., Senellart, J., and Rush, A. (2017). OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada. Association for Computational Linguistics.

Limsi (2019). Dicta-sign-lsf-v2. ORTOLANG (Open Resources and TOols for LANGuage) –www.ortolang.fr.

Lin, Y.-H., Chen, C.-Y., Lee, J., Li, Z., Zhang, Y., Xia, M., Rijhwani, S., He, J., Zhang, Z., Ma, X., Anastasopoulos, A., Littell, P., and Neubig, G. (2019). Choosing transfer languages for cross-lingual learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3125–3135, Florence, Italy. Association for Computational Linguistics.

Littell, P., Mortensen, D. R., Lin, K., Kairis, K., Turner, C., and Levin, L. (2017). URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 8–14, Valencia, Spain. Association for Computational Linguistics.

Ormel, E., Crasborn, O., van der Kooij, E., van Dijken, L., Nauta, E., Forster, J., and Stein, D. (2010). Glossing a multi-purpose sign language corpus. In *Proceedings of the 4th Workshop on the Representation and Processing of Sign Languages: Corpora and sign language technologies*, pages 186–191.

Othman, A. and Jemni, M. (2012). English-asl gloss parallel corpus 2012: Aslg-pc12. In *5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon LREC*.

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

Qi, Y., Sachan, D., Felix, M., Padmanabhan, S., and Neubig, G. (2018). When and why are pre-trained word embeddings useful for neural machine translation? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 529–535, New Orleans, Louisiana. Association for Computational Linguistics.

Rei, R., Stewart, C., Farinha, A. C., and Lavie, A. (2020). COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Sennrich, R., Haddow, B., and Birch, A. (2016a). Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Sennrich, R., Haddow, B., and Birch, A. (2016b). 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.

SignStream, N. (2007). Volumes 2–7.

Tuggener, D. (2016). *Incremental coreference resolution for German*. PhD thesis, University of Zurich.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008. Curran Associates, Inc.

von Agris, U. and Kraiss, K. (2007). Towards a video corpus for signer-independent continuous sign language recognition. In *Gesture in Human-Computer Interaction and Simulation*.

Xia, M., Kong, X., Anastasopoulos, A., and Neubig, G. (2019). Generalized data augmentation for low-resource translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5786–5796, Florence, Italy. Association for Computational Linguistics.

Yin, K. and Read, J. (2020a). Attention is all you sign: Sign language translation with transformers. In *Sign Language Recognition, Translation and Production (SLRTP) Workshop - Extended Abstracts*.

Yin, K. and Read, J. (2020b). Better sign language translation with stmc-transformer. In *Proceedings of the 28th International Conference on Computational Linguistics*.

Zhou, C., Ma, X., Hu, J., and Neubig, G. (2019). Handling syntactic divergence in low-resource machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1388–1394, Hong Kong, China. Association for Computational Linguistics.

Zoph, B., Yuret, D., May, J., and Knight, K. (2016). Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575, Austin, Texas. Association for Computational Linguistics.

A Data Augmentation Rules

A.1 General Rules

For a given sentence $\mathcal{S}$:

1. Discard all tokens $t \in \mathcal{S}$ if $\textbf{POS}(t) \notin \{\text{noun, verb, adjective, adverb, numeral}\}$
2. Discard remaining tokens $t \in \mathcal{S}$ with probability $p = 0.2$
3. Lemmatize all tokens $t \in \mathcal{S}$
4. Apply a random permutation σ to $\mathcal{S}$ verifying $\forall i \in \{1, n\}, |\sigma(i) - i| \leq 4$

where n is the number of tokens in $\mathcal{S}$ at step 4 and **POS** is a part-of-speech tagger.

A.2 German-DGS Rules

For a given sentence $\mathcal{S}$:

1. For each subject-verb-object triplet $(s, v, o) \in \mathcal{S}$, swap the positions of v and o in $\mathcal{S}$
2. Discard all tokens $t \in \mathcal{S}$ if $\textbf{POS}(t) \notin \{\text{noun, verb, adjective, adverb, numeral}\}$
3. For $t \in \mathcal{S}$, if $\textbf{POS}(t) = \text{adverb}$, then move t to the start of s
4. For $t \in \mathcal{S}$, if $\textbf{NER}(t) = \text{location}$, then move t to the start of s
5. For $t \in \mathcal{S}$, if $\textbf{DEP}(t) = \text{negation}$, then move t to the end of s
6. For $t \in \mathcal{S}$, if t is a compound noun $c_1 c_2 ... c_n$, replace t by c_1
7. Lemmatize all tokens $t \in \mathcal{S}$

where **POS** is a part-of-speech tagger, **NER** is a named entity recognizer and **DEP** is a dependency parser.

B Model Reproduction

For reproduction purposes, here we lay the exact commands for training a single model using OpenNMT 1.2.0 (Klein et al., 2017). These commands are taken from (Yin and Read, 2020b).

Given 6 files—*train.gloss* / *train.txt*, *dev.gloss* / *dev.txt*, *test.gloss* / *test.txt*—we start by preprocessing the data using the following command:

```
onmt_preprocess –dynamic_dict –save_data processed_data \
–train_src train.gloss –train_tgt train.txt –valid_src dev.gloss –valid_tgt dev.txt
```

Then, we train a translation system using the train command:

```
onmt_train –data processed_data –save_model model –layers 2 –rnn_size 512 –word_vec_size 512 –heads 8 \
–encoder_type transformer –decoder_type transformer –position_encoding –transformer_ff 2048 –dropout 0.1 \
–early_stopping 3 –early_stopping_criteria accuracy ppl –batch_size 2048 –accum_count 3 –batch_type tokens \
–max_generator_batches 2 –normalization tokens –optim adam –adam_beta2 0.998 –decay_method noam \
–warmup_steps 3000 –learning_rate 0.5 –max_grad_norm 0 –param_init 0 –param_init_glorot –label_smoothing 0.1 \
–valid_steps 100 –save_checkpoint_steps 100 –world_size 1 –gpu_ranks 0
```

At the end of the training procedure, it prints to console "Best model found at step X". Locate it, and use it for translating the data:

```
onmt_translate –model model_step_X.pt –src test.gloss –output hyp.txt –gpu 0 –replace_unk –beam_size 4
```

Finally, evaluate the output using SacreBLEU:

```
cat hyp.txt | sacrebleu test.txt
```

C Full Experimental Results

Table 3 includes the evaluation scores for all of our experiments, ran three times.

% of available annotated data used		1%		5%		25%		100%	
		BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET
PHOENIX	Baseline	6.37 ± 0.89	-89.21 ± 12.82	10.18 ± 0.40	-71.37 ± 2.86	16.20 ± 0.27	-33.88 ± 4.35	21.15 ± 0.58	-5.74 ± 2.35
	BT-*tuned*	4.12 ± 1.55	-91.87 ± 16.35	9.91 ± 0.54	**-53.38 ± 4.04**	**17.10 ± 0.56**	**-16.46 ± 2.52**	**22.02 ± 0.50**	**6.84 ± 0.34**
	General-*tuned*	**9.49 ± 1.01**	**-52.23 ± 6.31**	**14.78 ± 0.51**	**-27.13 ± 2.29**	**19.86 ± 0.64**	**-0.72 ± 2.44**	**23.35 ± 0.22**	**13.65 ± 1.68**
	Specific-*tuned*	**9.70 ± 0.75**	**-55.94 ± 2.08**	**14.65 ± 0.29**	**-30.85 ± 1.45**	**19.66 ± 0.08**	**-5.62 ± 0.51**	**23.17 ± 0.30**	**11.70 ± 1.20**
NCSLGR	Baseline	0.47 ± 0.60	-153.90 ± 11.89	2.07 ± 0.32	-145.14 ± 1.15	8.07 ± 0.43	-101.24 ± 5.14	15.95 ± 1.11	-61.00 ± 6.86
	BT-*tuned*	1.07 ± 0.47	**-139.80 ± 3.78**	**3.71 ± 0.55**	**-117.33 ± 3.03**	**9.11 ± 0.05**	**-82.41 ± 2.29**	16.67 ± 0.32	**-51.86 ± 0.66**
	General-*tuned*	1.58 ± 0.60	**-134.22 ± 1.73**	**5.13 ± 0.15**	**-106.59 ± 1.56**	**11.04 ± 0.04**	**-66.35 ± 2.00**	**19.09 ± 0.20**	**-34.50 ± 1.19**
	Specific-*tuned*	1.30 ± 0.52	**-128.14 ± 1.58**	**4.94 ± 0.45**	**-107.60 ± 4.01**	**10.99 ± 0.12**	**-71.37 ± 1.01**	**18.58 ± 0.84**	**-39.96 ± 1.91**

Table 3: Mean and standard deviation of BLEU and COMET over different experimental settings. We **bold** scores statistically significantly higher than baseline at the 95% confidence level.