Using Neural Machine Translation Methods for Sign Language Translation

Galina Angelova[1] and **Eleftherios Avramidis**[1] and **Sebastian Möller**[1,2]
[1] Technical University of Berlin, [2] German Research Center for Artificial Intelligence (DFKI)
g.angelova@campus.tu-berlin.de
{eleftherios.avramidis,sebastian.moeller}@dfki.de

Abstract

We examine methods and techniques, proven to be helpful for the text-to-text translation of spoken languages in the context of gloss-to-text translation systems, where the glosses are the written representation of the signs. We present one of the first works that include experiments on both parallel corpora of the German Sign Language (PHOENIX14T and the Public DGS Corpus). We experiment with two NMT architectures with optimization of their hyperparameters, several tokenization methods and two data augmentation techniques (back-translation and paraphrasing). Through our investigation we achieve a substantial improvement of 5.0 and 2.2 BLEU scores for the models trained on the two corpora respectively. Our RNN models outperform our Transformer models, and the segmentation method we achieve best results with is BPE, whereas back-translation and paraphrasing lead to minor but not significant improvements.

1 Introduction

Sign languages (SL), the main medium of exchanging information for the deaf and the hard of hearing, are visual-spatial natural languages with their own linguistic rules. In contrast to the spoken ones, they lack a written form, on one hand, and use face, hands and body to convey meaning, on the other. However, in our society, spoken languages are used by and large, leading to social exclusion in the everyday life of the deaf and hard of hearing. Therefore, recent research is making the most out of the technical advances in the fields of Natural Language Processing (NLP), Deep Neural Networks (DNN), and Machine Translation (MT), with the aim to develop systems that are able to translate between signed and spoken languages in order to fill the gap of communication between the SL speaking communities and the people using vocal language. Most latest approaches tackle the problem, dividing it into two sub-tasks: Sign Language Recognition (SLR), also called *video-to-gloss*, and Sign Language Translation (SLT), also known as *gloss-to-text* translation. The latter uses as an intermediate representation the glosses, described in Section 3.1 and Section 4.2.1. Isolating gloss-to-text translation serves as a building block of a bigger project, which considers SL as a whole and is done in direct co-operation with members of the SL community.

For the rest of this work, we focus on the gloss-to-text sub-task and treat it as a low-resource text-to-text machine translation problem. We explore different known techniques for MT of written languages on the glosses, and report our findings during our experiments with:

- two neural architectures (RNN and Transformer)
- several tokenization and sub-word segmentation methods (BPE, unigram and custom tokenization of the gloss annotations)
- two data augmentation techniques (back-translation and paraphrasing)

Preprocessing scripts and data are publicly available.[1]

The rest of our work is organized as follows: In Section 2, there is a review of previous related work in the field. In Section 3, we describe the essence of the gloss-to-text translation task, and briefly present the neural machine translation methods we have used throughout our experiments on the two corpora, both of them introduced in Section 4. Further, we present the experiments in Section 5 as well as all our results and findings, described in Section 6. In the last Section 7 we conclude our work and discuss possibilities for future research.

2 Related work

Sign language translation is a relatively new research field with recent findings made possible

[1] https://github.com/DFKI-SignLanguage/gloss-to-text-sign-language-translation

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Student Research Workshop, pages 273 - 284

thanks to the continuous advances in neural machine translation (NMT). Several experiments with SL gloss-to-text translation have taken place in the previous decade using statistical phrase-based machine translation (Stein et al., 2012; Morrissey et al., 2013). Camgoz et al. (2018) and Camgoz et al. (2020) use the Transformer architecture for SL translation, and are the first to realize an end-to-end system, combining SLR and SLT, jointly training based on both video embeddings, glosses and text, being currently the state-of-the-art work in the field. Yin and Read (2020) employ the Spatial-Temporal Multi-Cue (STMC) Network (Zhou et al., 2020) for the task. There have also been several experiments on the opposite direction: text-to-gloss (Othman and Jemni, 2011; Egea Gómez et al., 2021).

To the best of our knowledge, currently Moryossef et al. (2021) is the only published work experimenting with back-translation in the context of gloss-to-text translation. Their research has been conducted parallel and independent from our studies, and has concluded similar results concerning the use of back-translation in a low-resource SL setting. The main difference is that we further focus on other machine translation techniques, e.g. different models and tokenization schemes, whereas they explore in more detail the gloss-text pairs and their linguistic properties, proposing their own rule-based heuristics with the purpose to generate SL glosses, bearing in mind the specifics of the signed languages. The recent work of Yin et al. (2021), focusing on the problems related to the machine translation between signed and spoken language pairs, reports the first BLEU score on the Public DGS corpus, but contrary to our work, no details are given on how the models were trained and evaluated and therefore there can be no direct comparison of the results.

3 Methods

3.1 The gloss-to-text task

Glosses are the most commonly used written form for annotating SL, where each sign has a written gloss transcription. However, a limitation of using them is the fact that they do not sufficiently capture all the information, expressed through body posture, movement of the head and mimics, which also occur in parallel. As a result, there is a loss of information on a semantic level (Camgoz et al., 2020; Yin et al., 2021). Moreover, each SL corpus, offering gloss annotations, uses its own way of glossing, therefore the annotation is not standardized, and as a consequence different SL corpora cannot be concatenated.

Source: HUND3* AUCH1A SPRINGEN1
Target: Der Hund springt hinterher.

Table 1: Example of a parallel gloss sentence - German sentence pair.

In contrast to the classical text-to-text translation task, where the pairs consist of pre-aligned sentences - one in the source language and one in the target language, for our gloss-to-text translation models we work with matching pairs of gloss sentences on the source side, and German sentences on the target side (see Table 1). Hence the name *gloss-to-text*.

3.2 Architectures for neural machine translation

In our work we investigate two model architectures implementing different types of attention mechanisms - RNN and Transformer.

RNN is an encoder-decoder architecture with attention suggested by Sennrich et al. (2017b) (implemented in Nematus), similar to the one proposed in Bahdanau et al. (2014). A key difference is the initialization of the decoder hidden state with the averaged sum of the encoder concatenated hidden states, instead of with the last backward encoder state.

The Transformer is another encoder-decoder architecture (Vaswani et al., 2017), implementing the self-attention function. Without using RNNs the neural system computes representations of the input and output sequences. The encoder and decoder of the Transformer both consist of 6 identical layers, and each of these layers has two sub-layers. The decoder adds one additional sub-layer, which is using *multi-head decoder-encoder attention* on the encoder output helping the decoder to focus on the relevant parts of the input sequence.

3.3 Tokenization

Tokenizing text can be done at word, subword or character level. Investigation of possible tokenization variations for the glosses is particularly relevant in our work, because of the different gloss annotations in the two used corpora (Sections 4.2.1 and 5.2).

	Train	Dev	Test
PHOENIX14T	7,096	519	642
DGS	54,325	4,470	5,113

Table 2: Statistics of the two corpora.

Gloss	Meaning
ZU^3	to squeeze, squeezed
ZU7	closed
ZU9	towards

Table 3: Meaning of different variants of the German word "zu".

Byte Pair Encoding (BPE) is a simple data compression technique that has been succesfully applied to NMT (Sennrich et al., 2016b). The idea behind this algorithm is to replace the most common pairs of consecutive bytes with one single new byte. In order to rebuild the original data, a table of all the replacements is needed (Gage, 1994).

Unigram sub-word segmentation (Kudo, 2018) considers multiple segmentation variations of a word with their respective probabilities calculated based on a unigram language model.

3.4 Back-translation

Back-translation is a semi-supervised method for improving the quality of translation relying on monolingual data (Edunov et al., 2018). It allows using a big amount of monolingual target data, when available, in order to produce synthetic data for the source side. This technique may be beneficial in cases where the bilingual data is scarce, as is the case of the gloss-to-text task.

3.5 Paraphrasing

Paraphrasing is the task of using an alternative formulation to express the same semantic content (Madnani and Dorr, 2010). By using paraphrased sentences in the training set, we hope that the model may be lexically enriched by the provided variations. Here, we follow the paraphrasing method known as *bilingual pivoting* (Mallinson et al., 2017; Turkerud and Mengshoel, 2021).

4 Datasets

For our experiments we utilize the following corpora of the German SL, which due to the different gloss annotations are used only separately for our experiments. Statistics of the two corpora can be seen in Table 2.

4.1 RWTH-PHOENIX-Weather 2014T

Introduced by Camgoz et al. (2018), the corpus contains sign language videos with gloss annotations as well as their corresponding German sentences, and is a popular benchmark in SL translation. The project consists of a training set of 7,097 parallel sentences. For our experiments we used the already publicly available annotated data.[2] Contrary to the DGS corpus, this corpus doesn't contain any gloss suffix annotations.

4.2 The Public DGS Corpus

DGS is the result of a long-term project, conducted at the Institute for German Sign Language and Communication of the Deaf at the Hamburg University. The corpus, introduced by Hanke et al. (2020), is a subset of the full project. All resources are publicly accessible[3] via two formats. Our work will focus on the second one, MY DGS-annotated[4]. The data was extracted via the ELAN[5] format of the files (see Appendix, Figure 3). In the following sub-sections we describe the nature and the format of the DGS corpus as well as the required pre-processing steps. The final version of the corpus consists of 63,908 parallel sentence pairs.

4.2.1 DGS gloss annotation conventions

The gloss annotations of the DGS corpus are far more complex and comprehensive than the ones of the PHOENIX14T corpus. Konrad et al. (2020) give a detailed explanation of the glossing conventions. We use this information to construct the gloss sentences and to build our parallel data set. The glosses are written in capitalized letters - a common convention used for annotating SL. An essential part of the annotations are the gloss suffixes. For instance, they are used to represent lexical variants or to indicate different meanings of a word, as can be seen in the example with the German word "zu" (Konrad et al., 2020). It can be used as a preposition - locative, temporal or causal, as an adverb or as a conjunction. In order to differentiate between

[2] https://www-i6.informatik.rwth-aachen.de/~koller/RWTH-PHOENIX/

[3] https://www.sign-lang.uni-hamburg.de/dgs-korpus/index.php/welcome.html

[4] http://ling.meine-dgs.de

[5] https://archive.mpi.nl/tla/elan

these meanings, a combination of the word itself with a number and, in some cases, a sign, is used (see Table 3).

Focusing in depth on all of the linguistic rules used to create the different gloss annotations is out of scope for this work. Therefore, here we mention briefly some of the main sign categories. The *lexical signs* are approximately equivalent to the commonsense notion of the words, and also form the corpus dictionary. The *productive signs* in combination with other signs illustrate intended meaning, but they do not convey meaning of their own. The *pointing signs* indicate orientation or movement. There are also *fingerspelling signs* for annotating when the signers sketch the form of letters in the air. The *number type* forms a special system for easily representing different kind of numbers.

4.2.2 Creating the parallel corpus

The annotation of the sign language videos is structured in parallel channels, the *tiers*, supporting multi-level and multi-participant annotations (Appendix, Figure 3). The tiers we used to form the parallel sentence pairs are the ones containing the German sentences for each signer and those containing the glosses for the right and for the left hand of each signer. The first step was to access the textual data from all videos, using Beautiful Soup.[6] For this purpose we created a python script, which extracted the links to the files, read the content, and created an XML parse tree of each recording.

The ordering of glosses to a gloss sentence was achieved by considering the starting and the ending time of the corresponding German sentence and of the individual glosses. One particular obstacle we encountered during the formation of the parallel data set were the overlapping timestamps of some glosses done with both hands. Such is the case of the fingerspelling signs. Because signers have a "dominant" and a "non-dominant" hand, the dominant one is usually used for one-handed signs and for fingerspellings (Crasborn, 2011). For the purpose of constructing our gloss sentences we chose a uniform way to order the overlapping signs. We counted all the "left-handed" glosses and all the "right-handed" glosses for each file, and considered files with more "left-handed" ones to have signers with a dominant left hand, whereas files with more "right-handed" glosses to have signers with a dominant right hand. We refer to the glosses as "left" and "right" because of the annotations used in the corpus, although the distinction between "left" and "right" does not seem to have any linguistic role in any SL (Crasborn, 2011). Moreover, the native signers usually don't remember if a new signer is left-handed or right-handed. Thus, we decided to choose a convention for our work so that the gloss sentences formation is consistent, and therefore we always placed the glosses of the dominant hand in front of those of the non-dominant one.

[6] https://www.crummy.com/software/BeautifulSoup/bs4/doc/

5 Experiments

We separate our experiments in three main groups. In the first one, described in Section 5.1 we initially train two baseline models for both corpora and consecutively make changes to them with the goal to investigate how different model architectures and known configurations of neural MT systems influence them. Therefore, we use the best performing models from the first group to further continue our experiments in the second one, described in Section 5.2, where we apply three different tokenization schemes - BPE, unigram and custom tokenization, on the gloss and on the German sides of the corpora. Ultimately, we utilize the models, which produce the best translations up to this point, in the third group of experiments in Section 5.3, where we separately look into two data augmentation techniques - back-translation and paraphrasing. All models are trained using MarianNMT (Junczys-Dowmunt et al., 2018) and all configuration parameters are detailed in our repository.

5.1 Neural MT architecture

Our initial motivation to approach the gloss-to-text translation task as a classical low-resource MT problem were the findings by Koehn and Knowles (2017) and Sennrich and Zhang (2019). Therefore, we compare Transformer and RNN (Sennrich et al., 2017b) on which is the optimal model architecture for gloss-to-text translation. As baselines we train two off-the-shelf models on the PHOENIX14T and the Public DGS corpora separately, using the default parameters of MarianNMT.

We continue the first set of experiments using techniques for improving the MT quality in a low-resource setting (Sennrich and Zhang, 2019). We perform an extensive hyperparameter search, initiating from the configurations suggested by the above authors in order to reduce the chances that

other hyperparameters can lead to different conclusions. We achieve the best configuration when we reduce the size of the encoder to 1 layer, and the size of the decoder to 2 for both types of models. Furthermore, we implement an aggressive dropout of 0.5 and a word dropout of 0.4 on the source and the target sides. We reduce the beam size to 2, as suggested by Camgoz et al. (2018), and keep the learning rate 0.0005, the batch size 32, and the vocabulary size 1,010 and 2,600 for the PHOENIX14T and the DGS corpora, respectively. We use simple word tokenization. For the next group of experiments we continue only with the improved RNN configuration, following our conclusions regarding the best architecture (Section 6.2).

5.2 Tokenization experiments

During the tokenization experiments, using the best performing models up to this point, we investigate if and to what extend existing tokenization methods - BPE, unigram and custom tokenization - proven to be effective for NMT of written natural languages, could be beneficial in the gloss-to-text setting. The tokenization of BPE and unigram was done using SentencePiece (Kudo and Richardson, 2018) with the parameters that have been established as default, due to their good performance in WMT shared tasks (Sennrich et al., 2017a, e.g. 2 BPE iterations).

5.2.1 Tokenizing the PHOENIX14T corpus

On the PHOENIX14T corpus we train RNN systems using the same parameters as the ones from the previous group of experiments. The only difference is the way the input and output sentences are tokenized. We conduct additional experiments where we reduce the vocabulary size of the BPE models and compare the translation scores.

5.2.2 Tokenizing the DGS corpus

The DGS corpus has groups of glosses that are more complicated and rich in annotations, which we describe in Section 4. A comparison can be seen in Figure 1. Thus we make the assumption that there should be a difference in the translation quality of the models in favor of the subword tokenization. For our first experiment we use word tokenization and compare the results with the ones of the following models which use either BPE, unigram or custom tokenizations. The vocabulary size is 2,600.

Stripping the gloss parameters In a different, more naive, experiment on the DGS corpus we decide to strip the gloss parameters - such as signs or numbers, as shown in Figure 2, to see if they are making our model too complex, aware of the fact that they convey meaning to each annotation.

Custom tokenization for the glosses For our custom tokenization experimenent on the DGS corpus, we choose to add the token **"@@"** to separate prefix, suffix and compound glosses without losing this information in difference to the above case of leaving only the stem. The chosen custom token is not a part of the gloss parameters.

5.3 Data augmentation

For the last group of experiments we make the assumption that, according to Edunov et al. (2018), on one hand, back-translation has proven to be effective when using strong baselines with a big amount of data, but, on the other hand, it could also have a positive effect in low-resource NMT settings. Thus we decide to try this method for our corpora, together with one additional data augmentation technique - paraphrasing.

5.3.1 Back-translation on the PHOENIX14T corpus

We start with the PHOENIX14T corpus. As a first step, we train a model in the opposite direction, German sentences on the source side and gloss sentences on the target side. Based on the suggestions on back-translation in previous work (Sennrich et al., 2016a; Dou et al., 2019), we focus on in-domain data and we consider filtering sentences from an out-of-domain (ood) corpus separately, as too many out-of-domain sentences would result in adding a lot of noise, which may not be helpful for the translation quality. To confirm our hypothesis for the back-translation experiments, we mainly investigate the quality of the translation when adding in-domain data, different amounts of out-of-domain data or a mixture of in-domain and out-of-domain data.

In-domain back-translation A major challenge for the purpose of using back-translation is to find a big monolingual corpus of the target languages, given the very specific domain of the PHOENIX14T corpus, because it contains strictly weather-related sentences. Our first idea is to try and find weather-related corpus, but unfortunately,

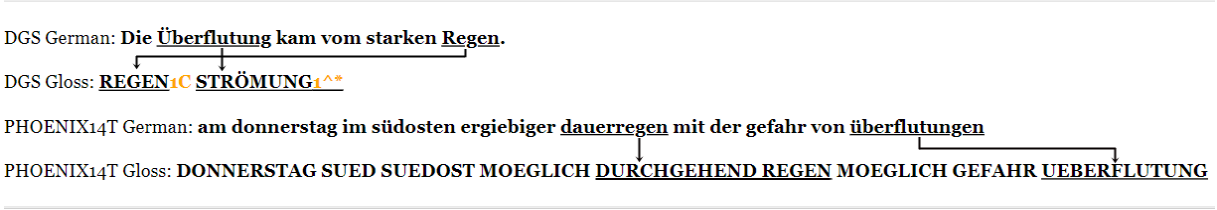

Figure 1: Comparison between gloss annotations for the two different corpora. The specific DGS gloss parameters are shown in orange.

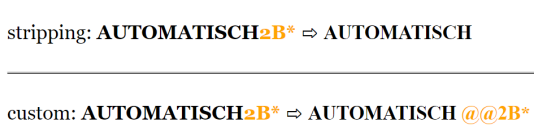

Figure 2: Example of the two manual tokenizations of a gloss in the DGS corpus

popular crawled monolingual corpora do not contain such specific sentences. We collect data manually by selecting sentences from online German weather-related articles or German weather websites. We pay attention to not only choose recent articles, but also to search sentences from some available archive sources. Additionally, we manually process the sentences which includes splitting them in shorter ones, removing some words we know are out-of-vocabulary for our models, rewriting complex verb forms. Needless to say, this process is slow and not scalable. Hence, we stop at 1,202 sentences and add their back-translated variants to our training data.

In the first of the two following experiments we observe the effect of adding filtered out-of-domain back-translated sentences to our training data, and in the second one we combine in-domain and out-of-domain sets.

Filtered sentences from out-of-domain corpus We use crawled data from the German part of the News Crawl corpus (Barrault et al., 2019). We extract 5,000 sentences from the whole dataset using a custom python script. Further, we filter sentences, containing the most frequently used weather-related words in the PHOENIX data set. For example, words or phrases such as: "wetter", "wettervorhersage", "temperatur", "es regnet", "es scheint", "wolken", "böen", "gewitter". It is important to mention here, that even though our filtered sentences contain one of the following words or phrases, these sentences cannot be fully considered in-domain. One reason for this is the fact that the crawled sentences are still different in structure and style than our original training data. Another reason is the fact that many of the words we use for filtering could also have a different not weather-related meaning, depending on the context.

Mixing in and out-of-domain sentences Here we mix our 1,202 in-domain and a part of the out-of-domain back-translated sentences (3,418) from the previous two experiments.

Usage of back-translation tag Since Sennrich et al. (2016a) mix their synthetic data with their original data without distinguishing between them, we conduct a further experiment to investigate if the **bt** tag, indicating synthetic data, is actually helping the neural system or worsening the performance.

5.3.2 Back-translation experiments on the DGS corpus

Considering our low scores on the DGS corpus and the conclusions of Moryossef et al. (2021) regarding the limitations of the back-translation in low resource SL settings, we conduct only one experiment as a proof of our premise that back-translation is not beneficial in a very low-resource setting in combination with a poor model to back-translate. For this purpose we filter the first 10,000 sentences from the news-crawl without taking into account their domains, because the DGS Corpus also does not have a specific domain.

5.3.3 Data augmentation using paraphrasing

For the last experiment we add 3,612 translated sentences from our original training set, using DeepL Translate[7], from German to English and then back from English to German. The paraphrased sentences are firstly reviewed to guarantee their grammatical correctness. Here, our goal is to create more variety in the words (synonyms) on the target side or in their order.

[7] https://www.deepl.com/en/translator

6 Results

In this section we report the results from the three groups of experiments we have conducted.

6.1 Evaluation

We evaluate all our models using SacreBLEU (Post, 2018). We also use the original dev and test sets of the PHOENIX14T corpus. For the DGS corpus we separate our own dev and test sets using 15% of the collected data - 4,470 sentences for the dev set, and 5,113 sentences for the test set.

6.2 Model architecture results

The results from our first group of experiments, described in Section 5.1, where we compare two types of model architecture, combined with adjustment of hyperparameters for improving the translation quality in a low-resource setting, are shown in Table 4. Whereas the baseline models perform better with a Transformer architecture for both corpora, we observe substantial improvements in the BLEU scores for the RNN models, trained on the PHOENIX14T corpus, after optimizing the hyperparameters. These results confirm our hypothesis that the architecture is also a suitable choice for the task of NMT of sign languages.

6.3 Tokenization results

After conducting the first tokenization experiments, described in Section 5.2, we observe the results, shown in Table 5, and conclude that using BPE and unigram, compared to word tokenization, does not lead to a substantial difference in the translation quality of the PHOENIX14T models. We believe that this is a result of the low word inflection in the corpus, and because of that the low number of unique glosses in the training set. Therefore, we decrease the size of the vocabulary for the model with BPE tokenization from 2,600 to 2,000 and gain an increase of 0.5 on the test set. In contrast, in the DGS corpus we have more complicated and rich in annotations different groups of glosses. Our assumption that there should be a greater difference in the translation quality of the models in favor of the subword tokenization is verified by the score we achieve on the test set with BPE (3.7 BLEU) substantially higher score than the previous one (2.7 BLEU) for the model trained with word tokenization, and the highest score we manage to obtain on that corpus. This confirms our hypothesis that subword tokenization is a more suitable choice for machine translation of signed languages with more complex and diverse annotations.

Stripping The BLUE score we achieve on the DGS corpus after stripping the parameters from the glosses is only 2.8 which, we assume, is due to the fact that each gloss annotation consists of important parameters, both contributing to the meaning, and communicating nuances. Removing this information, makes it impossible for our model to learn meaningful and correct representations as the stems of many glosses may be the same, but with added parameters the annotations may have very different meanings.

Custom tokenization By adding a custom token to split the parameters from the stem of the glosses we achieve 3.3 BLEU score on the test set, which is the second best score we manage to obtain. Unfortunately, the translation performance remains low.

6.4 Data augmentation results

Before conducting the back-translation experiments based on previous work (Sennrich et al., 2016a), we consider that (a) when having a very narrow domain, it is useful that the sentences, used for back-translation, are similar in structure and domain to the original ones, and (b) adding a number of sentences less than half of the training set size could not lead to substantial improvements. We also add a tag to each back-translated sentence - **{bt}**, to indicate for the neural system that this data is synthetic. After we train a model with added in-domain synthetic data, we manage to obtain a BLEU score of 22.3 on the test set, which is very close to our current best model (22.5 BLEU), and 22.2 BLEU score on the dev set, where we have a small improvement of 0.3 BLEU, compared to 21.9 BLEU. Results are shown in Table 6. We believe that this is a sign that the performance of our model does not get worse, confirming (a), although with such a small number of data it cannot get substantially better, confirming (b). On the contrary, it is possible that the model is less prone to overfitting, compared to the one without noise from synthetic data.

The model trained with only out-of-domain back-translated data reaches 22.2 BLEU on the test set, and does not improve the BLEU score on the dev set. With these results and the small amount of sentences we have in our original training set, combined with the rather poor quality of translation of

Model	BLEU dev	BLEU test
phoenix-baseline-rnn	18.3	17.7
phoenix-baseline-transformer	18.6	18.2
phoenix-rnn-improved	21.6	22.2
phoenix-transformer-improved	18.8	18.5
dgs-baseline-rnn	1.8	1.6
dgs-baseline-transformer	2.5	2.0
dgs-rnn-improved	2.9	2.7
dgs-transformer-improved	1.9	1.9

Table 4: Model architecture comparison for the baseline and improved systems.

Model	Tokenization	BLEU dev	BLEU test	Vocab size
phoenix-word-tok	word	21.6	22.2	1,010
phoenix-unigram-tok	unigram	22.4	21.5	1,010
phoenix-bpe-tok	bpe	22.5	22	2,600
phoenix-bpe-tok*	bpe	21.9	**22.5**	2,000
dgs-word-tok	word	2.9	2.7	2,600
dgs-bpe-tok	bpe	**4.2**	**3.7**	2,600
dgs-unigram-tok	unigram	3.5	3.2	2,600
dgs-bpe-tok-stemmed	bpe	3.1	2.8	2,600
dgs-custom-tok	word	3.5	**3.3**	2,600

Table 5: Tokenization experiments on PHOENIX14T and The Public DGS Corpus. The last bpe model, marked with *, is indicating the one with reduced vocabulary size.

our back-translating model, our intuition is that adding more sentences, which are poorly back-translated, will not lead to any improvements. It will rather add more noise to the model, which is not beneficial anymore for the diversity of the data.

Our last model that combines in-domain and out-of-domain data, achieves 22.7 BLEU on the test set, which is our best score. It is +0.2 over phoenix-bpe-tok - the best performing model with no synthetic data, but unfortunately, it is not significantly better, based on a bootstrap resampling significance test. It improves the score on the dev set - from 21.9 BLEU to 23.4 BLEU confirming our assumption that this noise is creating some diversity in the data without worsening the performance.

Results from the comparison of models with synthetic sentences, using a tag and not, can also be seen in Table 6. Since they show no substantial difference, we decide that at least in our case the tag does not play an important role for the quality of the translation.

Using back-translation on the DGS corpus we achieve only a small improvement of +0.1 on the test set (results are also shown in Table 6), confirming our hypothesis and the findings of Moryossef et al. (2021) and Edunov et al. (2018) that in a very low-resource setting back-translation cannot be clearly beneficial for the translation quality of the neural systems.

Finally, our model with added grammatically correct paraphrased sentences reaches 22.5 BLEU score on the test set - the same as the PHOENIX14T model without added synthetic data. We believe that the technique does not lead to worse performance. On the contrary, we suppose that it makes a small improvement, which can be again noticed on the dev set in Table 6.

7 Conclusion and Future work

In this work we investigated the effect of several methods used in NMT on the gloss-to-text translation task for a sign language. We present one of the first works that does extensive experiments on both

Model	#added sentences	dev	test
phoenix-bpe-tok	0	21.9	22.5
phoenix-indomain	1,202	23.3	22.3
phoenix-ood	5,000	22.1	22.2
phoenix-mixed	1,202 + 3,418	**23.4**	**22.7**
phoenix-paraphrasing	3,612	23.1	22.5
phoenix-mixed-no-tag	1,202 + 3,418	23	22.3
dgs-bpe-tok	0	4.2	3.7
dgs-bt-10000	10,000	4.2	**3.8**

Table 6: Data augmentation experiments on the PHOENIX14T corpus and the DGS corpus. The "ood" in the names stands for "out-of-domain", the "bt" - for "back-translation".

existing corpora for the German Sign Language - PHOENIX14T and the DGS Corpus. Further, we ran three successive groups of experiments:

Neural MT architectures, contrasting RNN and Transformer, with extensive search of hyperparameters and techniques, proven to be effective in a low-resource setup. In contrary to previous research, we found that RNN performs better than the Transformer.

Tokenization schemes, where our findings were in favor of the BPE tokenization for both corpora. This improved our PHOENIX14T model by 0.3 BLEU on the test set (reaching 22.5 BLEU), and our DGS model by 1 BLEU on the test set (reaching 3.7 BLEU).

Data augmentation techniques, i.e. back-translation and paraphrasing via bilingual pivoting, with the intention to create variance in the data. Back-translation gave small improvements: +0.2 on the PHOENIX14T corpus and +0.1 on the DGS corpus. Further investigation on the reasons for the limited contribution of the above augmentation techniques may be directed to the extremely low-resource scenario, the amount and domain of the data, or the particular nature of the sign language glosses.

All above methods allowed an improvement of 5 BLEU points on the test set (22.7 BLEU) for the PHOENIX14T model, and 2.2 BLEU points on the test set (3.8 BLEU) for the DGS one.

In conclusion, in line with previous research (Yin et al., 2021; Moryossef et al., 2021), we believe that in order to achieve better translation performance, research and experiments should concentrate on two major problems - collecting and annotating more resources, and better understanding the nature of the sign languages with the intention to develop new SL-specific tools.

Acknowledgements

The work was accomplished as a Bachelor thesis, as part of the program Digital Media and Technology at the Technical University of Berlin. Supervision, computational resources and conference participation were supported by the project SocialWear (German Ministry of Research and Education; BMBF). We would also like to thank the anonymous reviewers for critically reading this work and suggesting improvements.

Ethical considerations

Our work concerns the German Sign Language (DGS) and is only a part of a bigger research line that intends to provide communication improvements for the the deaf and hard of hearing communities. In order to respect these communities and ensure proper representation of their interests, members of them have been included in our project as part of the research team, consultants or participants in user studies and ELSI workshops (e.g. Nguyen et al., 2021), as per the recommendations of the SLLS ethics statement. The isolation of glosses, known to be inferior to the full linguistic capacity of the sign language does not intend to simplify the language but is rather used as a tool for aiding further research.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

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.

Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden. 2020. 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.

Onno Crasborn. 2011. The other hand in sign language phonology. In Keren van Oostendorp, Marc ; Ewen, Colin J.; Hume, Elizabeth V.; Rice, editor, *The Blackwell companion to phonology*, chapter 5, pages 223–240. Wiley-Blackwell.

Zi-Yi Dou, Junjie Hu, Antonios Anastasopoulos, and Graham Neubig. 2019. Unsupervised domain adaptation for neural machine translation with domain-aware feature embeddings. 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 1417–1422, Hong Kong, China. Association for Computational Linguistics.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 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.

Santiago Egea Gómez, Euan McGill, and Horacio Saggion. 2021. Syntax-aware Transformers for Neural Machine Translation: The Case of Text to Sign Gloss Translation. In *Proceedings of the 14th Workshop on Building and Using Comparable Corpora (BUCC 2021)*, pages 18–27, Online (Virtual Mode). INCOMA Ltd.

Philip Gage. 1994. A new algorithm for data compression. *C Users Journal*, 12(2):23–38.

Thomas Hanke, Marc Schulder, Reiner Konrad, and Elena Jahn. 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).

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Aji, Nikolay Bogoychev, André Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in c++. pages 116–121.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. Association for Computational Linguistics.

Reiner Konrad, Thomas Hanke, Gabriele Langer, Susanne König, Lutz König, Rie Nishio, and Anja Regen. 2020. Public DGS Corpus: Annotation Conventions / Öffentliches DGS-Korpus: Annotationskonventionen.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia. Association for Computational Linguistics.

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.

Nitin Madnani and Bonnie J. Dorr. 2010. Generating phrasal and sentential paraphrases: A survey of data-driven methods. *Computational Linguistics*, 36(3):341–387.

Jonathan Mallinson, Rico Sennrich, and Mirella Lapata. 2017. Paraphrasing revisited with neural machine translation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 881–893, Valencia, Spain. Association for Computational Linguistics.

Sara Morrissey, Andy Way, S Morrissey, and A Way. 2013. Manual labour: tackling machine translation for sign languages. *Machine Translation 2013 27:1*, 27(1):25–64.

Amit Moryossef, Kayo Yin, Graham Neubig, and Yoav Goldberg. 2021. Data augmentation for sign language gloss translation. In *Proceedings of the 1st*

International Workshop on Automatic Translation for Signed and Spoken Languages (AT4SSL), pages 1–11, Virtual. Association for Machine Translation in the Americas.

Lan Thao Nguyen, Florian Schicktanz, Aeneas Stankowski, and Eleftherios Avramidis. 2021. Evaluating the translation of speech to virtually-performed sign language on ar glasses. In *2021 13th International Conference on Quality of Multimedia Experience (QoMEX)*, pages 141–144.

Achraf Othman and Mohamed Jemni. 2011. Statistical Sign Language Machine Translation: from English written text to American Sign Language Gloss. *International Journal of Computer Science Issues*, 8(5):65–73.

Matt Post. 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.

Rico Sennrich, Alexandra Birch, Anna Currey, Ulrich Germann, Barry Haddow, Kenneth Heafield, Antonio Valerio Miceli Barone, and Philip Williams. 2017a. The University of Edinburgh's neural MT systems for WMT17. In *Proceedings of the Second Conference on Machine Translation*, pages 389–399, Copenhagen, Denmark. Association for Computational Linguistics.

Rico Sennrich, Orhan Firat, Kyunghyun Cho, Alexandra Birch, Barry Haddow, Julian Hitschler, Marcin Junczys-Dowmunt, Samuel Läubli, Antonio Valerio Miceli Barone, Jozef Mokry, and Maria Nădejde. 2017b. Nematus: a toolkit for neural machine translation. In *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 65–68, Valencia, Spain. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 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.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 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.

Rico Sennrich and Biao Zhang. 2019. Revisiting low-resource neural machine translation: A case study. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 211–221, Florence, Italy. Association for Computational Linguistics.

Daniel Stein, Christoph Schmidt, and Hermann Ney. 2012. Analysis, preparation, and optimization of statistical sign language machine translation. *Machine Translation 2012 26:4*, 26(4):325–357.

Ingrid Ravn Turkerud and Ole Jakob Mengshoel. 2021. Image captioning using deep learning: Text augmentation by paraphrasing via backtranslation. In *2021 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 01–10.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Kayo Yin, Amit Moryossef, Julie Hochgesang, Yoav Goldberg, and Malihe Alikhani. 2021. Including signed languages in natural language processing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7347–7360, Online. Association for Computational Linguistics.

Kayo Yin and Jesse Read. 2020. Better sign language translation with STMC-transformer. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5975–5989, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Hao Zhou, Wengang Zhou, Yun Zhou, and Houqiang Li. 2020. Spatial-temporal multi-cue network for continuous sign language recognition. In *Proceedings of the Fourth AAAI Conference on Artificial Intelligence*, volume 34, pages 13009–13016.

Appendix

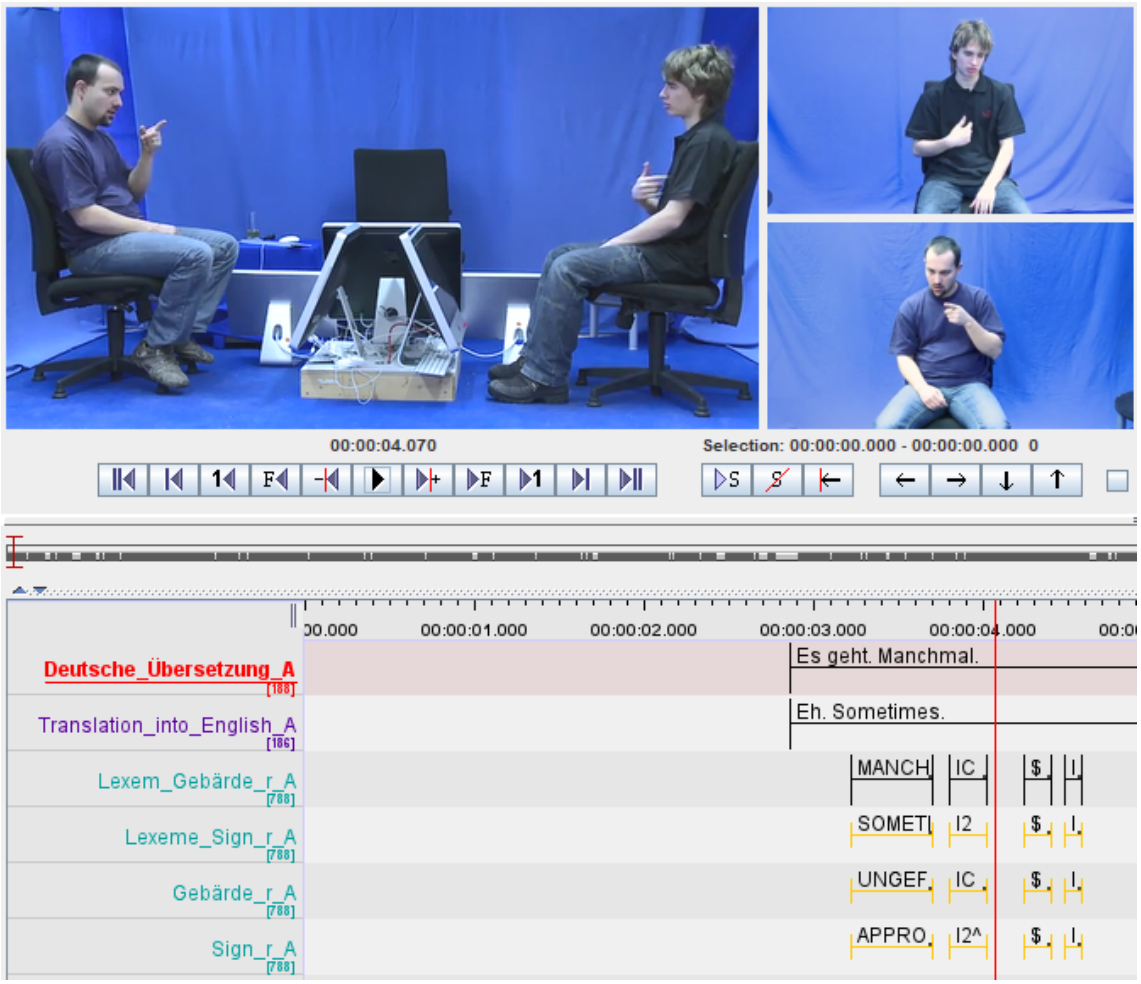

Figure 3: Sample of a short sentence form the DGS corpus in the ELAN software. Video with participants is shown above, and the different tiers can be seen underneath - e.g. “Deutsche_Übersetzung_A” for the German sentence, “Lexem_Gebärde_l_A” and “Lexem_Gebärde_r_A” for the gloss annotations for the left and right hands of signer A. Source: Hanke et al. (2020)