Translating Sign Language Videos to Talking Faces

Seshadri Mazumder
International Institute of Information Technology
Gachibowli, Hyderabad
Telengana, India
seshadri.mazumder@research.iiit.ac.in

Rudrabha Mukhopadhyay
International Institute of Information Technology
Gachibowli, Hyderabad
Telengana, India
radrabha.m@research.iiit.ac.in

Vinay P. Namboodiri
University of Bath
Claverton Down, Bath
United Kingdom
vpn22@bath.ac.uk

C. V. Jawahar
International Institute of Information Technology
Gachibowli, Hyderabad
Telengana, India
jawahar@iiit.ac.in

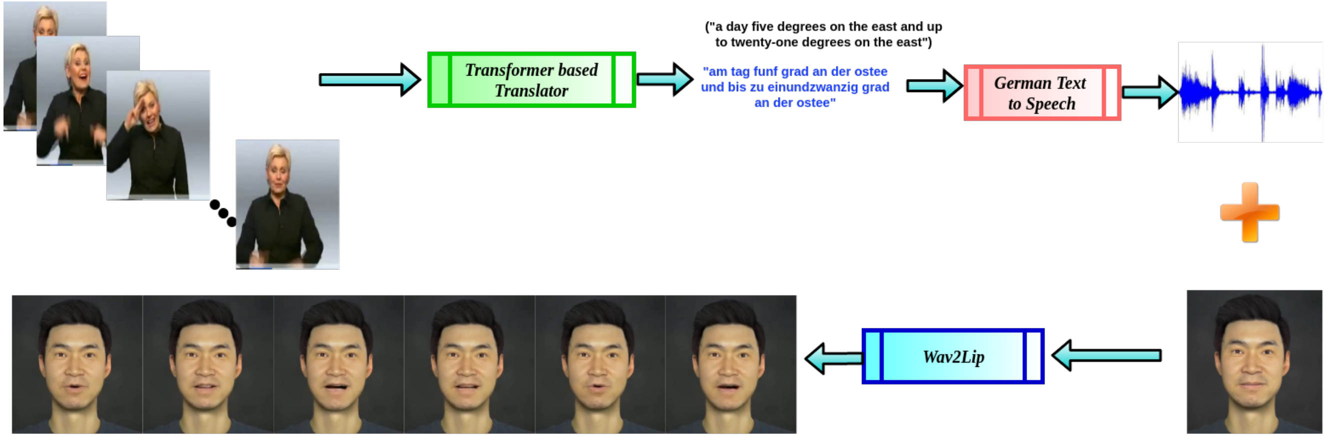
Transformer based Translator
("a day five degrees on the east and up to twenty-one degrees on the east")
"am tag funf grad an der ostee und bis zu einundzwanzig grad an der ostee"
German Text to Speech
Wav2Lip

Figure 1: The pipeline generates a talking face video from a sign language video. Given a sign language video first the transformer based module generates the corresponding translations. Then the translated sentence is fed to a TTS module to generate the corresponding speech. Then given an avatar and the speech to the Wav2Lip module, it generates a Sign Language tallking face video where the lips are in sync with the audio.

ABSTRACT

Communication with the deaf community relies profoundly on the interpretation of sign languages performed by the signers. In light of the recent breakthroughs in sign language translations, we propose a pipeline that we term "Translating Sign Language Videos to Talking Faces". In this context, we improve the existing sign language translation systems by using POS tags to improve language modeling. We further extend the challenge to develop a system that can interpret a video from a signer to an avatar speaking in spoken languages. We focus on the translation systems that attempt to translate sign languages to text without glosses, an expensive annotation form. We critically analyze two state-of-the-art architectures, and based on their limitations, we improvise the systems. We propose a two-stage approach to translate sign language into intermediate text followed by a language model to get the final predictions. Quantitative evaluations on the challenging benchmarks on RWTH-PHOENIX-Weather 2014 T show that the translation accuracy of the texts generated by our translation model improves the state-of-the-art models by approximately 3 points. We then build a working text to talking face generation pipeline by bringing together multiple existing modules. The overall pipeline is capable of generating talking face videos with speech from sign language poses. Additional materials about this project including the codes and a demo video can be found in https://seshadri-c.github.io/SLV2TF/

CCS CONCEPTS

• **Computing methodologies → Computer vision problems.**

ICVGIP '21, December 19–22, 2021, Jodhpur, India

ACM ISBN 978-1-4503-7596-2/21/12.
https://doi.org/10.1145/3490035.3490286

KEYWORDS

Sign Language Translation, Sign Language Recognition, POS Tagging, Sign Language to Text, Sign Language

ACM Reference Format:
Seshadri Mazumder, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, and C. V. Jawahar. 2021. Translating Sign Language Videos to Talking Faces. In *Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP '21), December 19–22, 2021, Jodhpur, India,* Chetan Arora, Parag Chaudhuri, and Subhransu Maji (Eds.). ACM, New York, NY, USA, Article 00, 10 pages. https://doi.org/10.1145/3490035.3490286

1 INTRODUCTION

Sign language is a system of communication using visual gestures and, signs, as used by the deaf community and is the main mode of their communication. They being visual in a sense have multiple channels of input as information. To say they include hand movements, facial expressions, lip movements, head movements and body gestures.

Linguistic system of sign language(s) differs from natural languages and have its own specific linguistic rules [32]. They do not have a syntactical alignment with their natural language counterparts. They differ in word orderings and do not possess a monotonic direction. As a result of such disparities sign language translation(SLT) methods conjointly learn the embedding space of sign sentence videos and natural languages as mappings between them, leading to a complex sequence learning problem. SLT problems can be of two types of which we experiment on the former one:

(i) Natural language text generation from sign language videos, and,
(ii) Sign language video generation from spoken language text.

Generating natural text from sign language has recently gained traction in the computer vision community. Existing Sign Language Translation (SLT) approaches can be categorised into two-staged and one-staged approaches based on whether they need further annotations for video and text alignments.

Sign-to-Gloss & Gloss-to-Text Approaches: Two staged approaches use gloss annotations, which are sign language gestures in a sequence relative to the video. Glosses are used in an intermediate stage and serve as a label. These models first learn to acknowledge gestures exploiting gloss annotations, then rearrange the recognition results into spoken language sentences. In these approaches, gloss annotations significantly facilitate syntactic alignment. However, gloss annotations are not straightforward to amass since annotators need extensive experience in sign languages [4].

Sign-to-Text Approaches: On the other hand, the one-staged approaches learn to translate sign videos to natural languages without the intermediary glosses. To generate spoken language phrases once the information inserted into the sign phrases has been understood by the system. Sign languages have their own unique grammatical and linguistic structures, which often don't have one-to-one characteristics of their counterparts as discussed earlier. This problem really sends a translation task to a machine. Initial studies by computer linguists have been used to map sign glosses and their spoken word translations through text-to-text statistical machine translation models. Glosses, however, are reduced representations of the sign guides and language writers are nonetheless unanimous in their annotation of sign languages.

Few contributions to continuous video SLT were made, especially due to the lack of proper data sets for the formation of such models. Recently [4] released the first sign language video available to the public on the spoken word translation data, namely PHOENIX14T.

The aim of this paper is to increase the performance of the one-staged strategy by splitting the sequencing problem into two problems and using language modelling to improve the performance. The BLEU score improved by almost 2 points, which is an improvement over the available literature based on the current arrangement.We employed an external end-to-end trainable language model to target the tokens that were most frequently misconstrued, and we provided an novel POS tagged technique that substitutes tokens with their associated POS tags first, then recovers them to the same vocab space.

We also generate a talking faces avatar accompanying the signer. This is an application of sign language translation merged with talking face generation models. Generating talking faces avatars provide people with hearing or speech loss a sense of complementary fulfillment and helps them overcome the tag "impaired". It is similar to attaching an artificial leg to a physically immobile person. It has many novel use cases:

(i) People with hearing and speech disorders can teach courses where they will be signing while their avatars perform the complementary actions that a commoner will understand.
(ii) They can speak at various conferences, seminars or workshops but limited to unidirectional interaction.
(iii) Serves as one way interaction for activities like news reading environment.
(iv) Enables the community to create content for different digital mediums.

In this paper, we study two different architectures, Sign Language Translation using Transformers [4] and Sign Language Translation via Temporal Semantic Pyramid Segmentations [27].

The contributions of this paper can be summarized as :

(i) A novel POS-tagged methodology that improvise the scores of the above one-staged SLT models.
(ii) The first successful implementation of an end-to-end pipeline that can translate sign language videos from signers to Speaking Avatar.
(iii) A range of use cases of our implementations discussed earlier that can prove to be a great boon for the hearing or speech impaired community.

The rest of this paper is organized as follows: In Section 2 we survey the works that are related to sign language translation(SLT) and lip synchronization. In Section 4 we discuss the generic SLT problem and discussed the architectures SLT using transformers [4] and SLT using TSPNet [27] and the limitations of these architectures in Section 3.4. In Section 4.1 upon analysing the shortcomings of these models and the sign videos with the vocabulary we propose a novel methodology to improvise their scores. We share our training pipeline in Section 4.4.1. Then we discuss the pipeline to generate talking face avatar in Section 4.5. In Section 5 we discuss the training dataset and the evaluating metrics. We then report the quantitative results in Section 5.2 where we report the previous two baseline results and our improvised scores. In Section 6, we share translation examples generated by our network to give the reader further qualitative insight of how our approach performs. We conclude the paper in Section 6 by discussing our findings and possible future work. Please check our demo video and additional materials for this project in this website[1]

2 RELATED WORKS

We survey and summarize the current works tackling sign language to text conversion that are available in the literature.

2.1 Sign Gesture Segmentations

Firstly, sign sentences, often created using theme-comment structures [38], must be detected by the system from continuous sign language videos. For texts based translation tasks, this is trivial in terms of the models' ability to use punctuation marks in separate phrases. On the other hand, speaking-based recognition and translation systems seek pauses between phonemes to spoken language segment utterances, e.g. silent regions. Studies on automatic sign segmentation [20, 34, 36] were previously carried out in the literature. However, no study is available to the best of the authors to ensure continuous sign video segmentation that enhances sign language translations.

2.2 Sign Language Recognition

SLR methods are roughly divided into isolated SLRs [12, 17, 18] and continuous [4, 19]. SIRs can be divided into an SLR method. Unlike isolated SLRs, most CSLR approach the sequence recognition of the model weakly as a perception.

Some early CSLR methods [9, 23] adopt a paradigm for dividing and winning signs into subunits with HMM-based recognition systems to work on limited data.

The latest computer vision successes of the CNN [10, 25, 30] offer a powerful tool for reproducing visual features. CNNs often need frame-wise annotations, however, which conflict with the CSLR's weakly monitored nature. Koller et al. [13] propose an iterative expectation/maximation approach, which includes an intermediate task of providing frame-level supervision for a classification of hand-shaped shapes in the GMM-HMM model. Some of the work is being extended by proposing frameworks of CNN+LSTM+HMM [14], including further indication [19] and improving the iterative approach to alignment [15]. The iterative CNN-LSTM-HMM configuration offers robust visual features, which many recent works take into account [2, 3].

2.3 Sign Language Translation

The PHOENIX Weather 2014T dataset was formalised at [4], and a 2D CNN model was jointly used for extracting gloss-level features from video frames and a seq model for German sign language translation. Subsequent work on the CSLR component is being carried out in the SLT [35].

A contemporary paper [5] also achieves encouraging results for both tokenization and translation by using a multi-task Transformer, although its CSLR performance is not optimal with a Word Error Rate that is higher than basic models. Another paper [27] introduced an interesting sign gesture segmentation mechanism that improvised the scores of [4] by a great extent.

We also survey related works in neural machine translation and generation of talking faces as they are relevant in the context of our work.

2.4 Neural Machine Translation

The neural machine translation(NMT) task is to translate one language into one another. The majority of NMT models follow a paradigm encoder-decoder. Previous works use RNN for temporary semantic modelling [7, 22, 37]. Mechanisms for the treatment of long-term dependence were subsequently adopted in [1, 29] using attentions. Instead of RNNs, new transformer models [11, 40] are based entirely on attention to sequence modelling and feeding layers. Both translation quality and efficiency are significantly improved.

2.5 Lip Synchronization

The lip synchronization task is to generate talking faces from speech and a reference image. The majority of lip synchronization models follow a paradigm encoder-decoder. Previous works like You Said That [8] simply encodes an audio segments and a face with lower half masked and which targets a face with shape of the lips changed mapping to phoneme levels. Though much recent works ([33], [21]) uses GAN discriminators to improve the quality of video generation.

3 DESIGN ARCHITECTURES

Here we first formulate the problem of sign language translation and then analyse the constraints of two alternative design architectures and proposed a realistic method to enhance the translation scores.

3.1 Problem Formulation

Translating sign videos to spoken language is a sequence-to-sequence learning problem. Our objective is to learn the conditional probability $p(y|x)$, where y is a sentence i.e $y = (y_1, y_2, ..., y_M)$ consisting of M number of words and x is the corresponding video i.e. $x = (x_1, x_2, x_3, ..., x_N)$ with N number of frames. This is not an easy task as the number of frames in sign video is far above the number of words in its speech translation, i.e $N >> M$. In addition, the correspondence between spoken language stream and sign language sequences is usually unknown, non-monotonic and inconsistent. In addition to typical translation activities that operate on text, our source sequences are videos. This makes it difficult utilizing the classical sequence models comparable to the RNNs.

3.2 Sign Language Translation using Transformers [4]

The first architecture that we studied, analysed the shortcomings of the model and later improved the architecture for better scores. In the upcoming section we have discussed the data processing i.e. how the embeddings are generated from raw data. In section 3.2.1, we discuss the spatial and textual embeddings, in section 3.2.2 we brief the architecture of [4] and in section 3.4.1 we discuss its limitations.

[1] https://seshadri-c.github.io/SLV2TF/

3.2.1 ***Spatial and Word Embeddings :*** NMT approaches begin by employing word embeddings for the tokenized source and the destination sequences. These embeddings are either learned from scratch or pretrained on larger datasets and fine-tuned during training. Contrary to text translations, sign language translations, have visual scenes in it. Training the model, needs to learn the spatial embeddings as a form of representation of sign videos.

This is accomplished via 2D CNNs. Given a sign video v, and the CNNs extract spatial domain representations as :

$$f_t = SpatialEmbeddings(v_t) \tag{1}$$

where, f_t represents the spatial embeddings of a frame v_t at the t^{th} time step, when feed forwarded through the network of 2D CNNs.

For the spatial embeddings, ResNet [16] feature extractors have been used to extract the frame level features, which was pretrained on the ImageNet[10] data.

Similarly, for word embeddings, fully connected layers are used to learn the embedding space which is a linear projection from one hot encoding of the words which is represented as :

$$e_t = WordEmbeddings(s_t) \tag{2}$$

where, e_t represents the embeddings of the spoken word s_t which is the t^{th} word of a sentence s.

3.2.2 ***Model Architecture :*** The architecture used here is a sequence-to-sequence modelling architecture i.e., the transformers [40]. The transformer follows an encoder-decoder style. Here the encoders and the decoders have a sequence of layers which cardinally need to be the same for both the blocks. As per design choices $N = 6$, i.e., a $6 - layered$ stack of the network is chosen.

The network also uses attention mechanisms to map the features between the intermediate embedding space of the encoders and decoders. As per design choices, experiments are performed for both Bahadanu's [1] attention and Luong's attention where Bahadanu's attention showed superiority than the other in terms of BLEU scores.

3.3 Sign Language Translation via Temporal Semantic Pyramid Segmentations [27]

In this work a Temporal Semantic Pyramid Network is proposed. Here we analysed the shortcomings of the model and later improved the translation scores by altering the input space in the textual domain. Here we have discussed the data processing i.e., how the embeddings are generated from raw data, and then, discussed the architecture and in latter some light is thrown upon its limitations.

3.3.1 ***Spatial and Word Embeddings :*** Contrary to the previous approach, the model learns sign video representations that encode both spatial appearance and temporal dynamics. However, it is difficult to acquire exact segments of gestures from a continuous sign video, while noisy segments introduce considerable uncertainty to feature learning. So here multi-scale segment representation is proposed.

Earlier SLT methods learn frame-wise video characteristics, the temporal semantics of gestures are neglected for such characteristics from static images.

Given a video of t frames $V = v_0, v_1, ..., v_{t-1}$ with v_i a video frame, a video segment $s_{m,n}$ is a sub-sequence of S, denoted as $s_{m_n} = s_m, s_{m+1}, ..., s_{m+n-1}$. Here each segment is obtained by choosing a fixed window width w and a stride. Since a gesture last for around half a second (i.e. 12 frames) the window length is considered to be $w = 12$ for the PHOENIX dataset. Now for each of these video segments, the I3D [6] feature extractor is used to generate a representation in the embedding space which is a new Two-Stream Inflatable 3D ConvNet (I3D).

Similar to previous approach here the word embeddings are fed to the model which are pretrained and non trainable. Here, the byte-pair encoding (BPE) is used which is trained on Wikipedia and its intended use is as input for neural models in natural language processing. It gives a sub word segmentation that is often good enough, without requiring tokenization or morphological analysis. Pre-trained byte-pair embeddings work surprisingly well as it requires no tokenization and is much smaller than the other alternatives.

3.3.2 ***Model Architecture :*** The architecture is almost the same as that of the SLT [4], except for mapping the attentions Here Inter-Scale and Intra-Scale attentions are used. Inter Scale attentions looks upon to enforce local semantic consistency to compensate for the effect of inaccurate video segmentation, while Intra Scale attentions enhance features across all the local regions.

3.4 Limitations of the Previous Works

3.4.1 ***Limitations of SLT using Transformers [4] :*** Limitations of the model can be critically analysed from the qualitative evaluation of the translations. Evidently from [4] that the model was making mistakes in the translation of dates, places and numbers.

Another concern is about the temporal dynamics in the input feature space. Since frame wise features were extracted instead of sign language gestures using ResNet, so the temporal dynamics have not been taken care off. These actions or gestures in case of the sign language when feed forwarded to the network will improve the translation accuracy.

3.4.2 ***Limitations of SLT using TSPNet [27] :*** The qualitative evaluation of translations is carefully scrutinised to understand the limitations of this model. It was evident that the model mostly made mistakes with the low-frequency words that are very challenging to translate such as city names. Additionally, facial expressions often convey indications that are not clearly represented by the model.

So, the previous models cannot capture the low frequency words and which are not explicitly defined by the hand movements and fingers rather by facial expressions and lip movements to be specific. So here upon statistically analysing the vocab we have proposed a POS tagging method which is explained in detailed manner in the following section.

4 PROPOSED METHODOLOGY

In this section we discuss a novel two-staged approach that improves sign language translation with the help of text labelling and language modelling.

4.1 POS Tagging

The practice of labeling a term in a text, based on its definition and context in corpus linguistics, as a part of speech marking is termed as POS or grammatical tagging. An extensive study of the two state-of-the-art sign language translation models and critically analysing their limitations we proposed a novel approach that improves the the translation scores. Both the previous approaches made a substantial amount of error while predicting the low-frequency words which were mostly labelled as proper nouns and numbers. These findings are discussed in the section 4.2

4.2 Frequency Analysis of major POS tags :

The frequency of occurrence of the different parts of speech, namely Pro-noun, Adverb, Noun, Adjective, Verb, Proper Noun, and Number, is shown in table 1. It shows the total number of occurrences of the words w.r.t the different parts of speech and the unique occurrence of the words in a specific POS set.

Parts Of Speech Tags	Tot. Num. of Occur.	Unique Occur.
Pronoun	5680	50
Adverb	15375	259
Noun	23248	1125
Adjective	10360	937
Verb	6950	657
Proper Noun	2963	374
Number	240	48

Table 1: Major Parts of Speech Tags with their total number of occurrences in the train set and their unique occurrences.

Next, we plot the ratio

$$\frac{Total\ Number\ of\ Occurrences}{Total\ Number\ of\ Unique\ words\ in\ speific\ POS\ set},$$

to observe the low-frequency words that have the most number of new occurrences in the specific POS set.

From figure 2, we derive that the POS set *proper noun* and the *number* are shown in green, having the lowest frequencies compared to others.

4.3 Improvement :

We present an algorithm which in simple sense, transforms sentences based on POS tagging with the findings from the previous section 4.2

The vocab space before and after transforming all the sentences in the Train, Dev, and Test set of RWTH-PHOENIX-Weather 2014T dataset is mapped in Table 2. The transformation leads to a substantial rise in BLEU scores. For SLT [4], scores have increased from 9.58 to 12.01, and for TSPNet [27], it increased up to 15.82 from 13.41. The results can be referred from table 3 which is further described quantitatively in the section 5.2 and qualitatively in section 6.

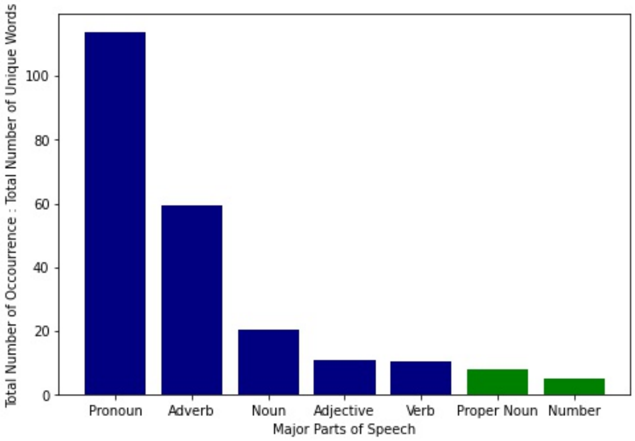

Figure 2: Histogram plot that shows the frequencies of each POS tags based on their unique occurrence. Here the X-axis correspond to the major parts of speech tags and the Y-axis correspond to the ratio of their total number of occurrences : unique occurrences in the training set of sentences. The low frequency POS tags are shown green in colour.

Algorithm 1: Transform Sentences after POS tagging

```
Result: Tokenized Sentence by replacing specific POS
input_sentence;
tokenized_list = [];
tokenized_sentence = list(tokenizer(input_sentence));
for token in tokenized_sentence do
    if POS(token) is 'propn' then
        tokenized_list += 'propn'
    else if POS(token) is 'num' then
        tokenized_list += 'num'
    else
        tokenized_list += token
    end
end
return tokenized_list
```

Category	Train-Set	Dev-Set	Test-Set
Uni. Words in Orig. Sent.	2891	955	1005
Uni. Words in Trans. Sent.	2698	864	916

Table 2: Number of words in the vocab space of train, dev and test set after applying the sentence transformation algorithm using POS tags. First row correspond to the given set of sentences and second row corresponds to the transformed sentences.

4.4 Training Pipeline

There are two steps to our training process. In Stage 1, we predict POS tagged phrases from sign language videos, and in Stage 2, we use a transformer-based language model to substitute the POS tags.

The processing of our source and target data, as well as the training pipeline for both stages, are covered in detail in the subsequent sections.

4.4.1 *Stage 1 : Sign language video to POS tagged sentence*.

This section discusses the aggregated training pipeline of the two state-of-the-art architectures SLT [4] and TSPNet [27], for our processed input sequences. First, we discuss the data processing for the encoder and decoder for the two different aforementioned architectures, and then their respective training paradigms.

Data Processing(Video) : In SLT [4], for each video, framewise features are extracted using Resnet [16] feature extractor. Only the spatial dynamics have been taken care of, and the temporal dynamics are required must be learnt from the transformers. More details about the processing can be found in section 3.2.1.

Contrary to this approach TSPNet [27] processes video much like gesture segments.. Estimating that a sign gesture can last for around 0.5 seconds for the Phoenix dataset with a frame rate of around 25fps, 12 frames are considered. The feature embeddings are extracted, stacked up, and fed to the transformer for training using a pre-trained I3D feature extractor. The temporal dynamics are also taken care of as I3D utilises 3D convolutions with spatial information. Further details related to processing can be found in section 3.3.1.

Data Processing(Text) : Textual embeddings can be trainable or non-trainable. In SLT, spoken language words are first projected as one-hot encoding vectors into the vocab space, and then using two fully connected linear layers, the embeddings are learnt. While in case of TSPNet embeddings are generated using German BPE.

Training : The full training pipeline is shown in figure 3. It is a typical transformer architecture with encoder-decoder blocks. To start with, the video segments pass through the temporal and spatial feature extractors, different for the two architectures, as discussed in previous section. Then the different video feature embeddings are stacked up and feed forwarded to the encoder block for training. Similarly, for the decoder, we transform the spoken language sentences to POS tagged sentences where the corresponding POS tags replace the low-frequency words in the context of the given sentence. Then this modified sentence is subsequently fed to the block of textual embeddings.

SLT uses self-attention mechanism within the encoders while TSPNet uses inter-scale and intra-scale attention mechanisms between the different sign gesture segments. Those attentions are mapped again to the decoder textual embeddings. In the last few decoder layers, the intermediate embeddings are passed through a linear layer and followed to a softmax layer that maps probability values in the vocab space. At each iteration, the decoder works iteratively, giving out the probabilities for the t^{th} word in vocab space at the t^{th} time step. After integrating all the words, the translated natural sentence is generated for its corresponding sign language video given as an input.

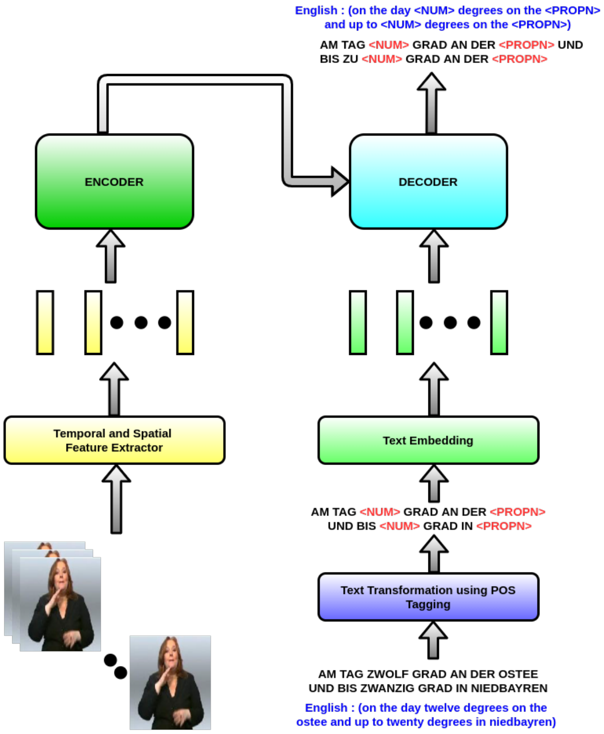
English : (on the day <NUM> degrees on the <PROPN> and up to <NUM> degrees on the <PROPN>)
AM TAG <NUM> GRAD AN DER <PROPN> UND BIS ZU <NUM> GRAD AN DER <PROPN>
ENCODER
DECODER
Temporal and Spatial Feature Extractor
Text Embedding
AM TAG <NUM> GRAD AN DER <PROPN> UND BIS <NUM> GRAD IN <PROPN>
Text Transformation using POS Tagging
AM TAG ZWOLF GRAD AN DER OSTEE UND BIS ZWANZIG GRAD IN NIEDBAYREN
English : (on the day twelve degrees on the ostee and up to twenty degrees in niedbayren)

Figure 3: Training Pipeline, for detailed working read section 4.4.1.

4.4.2 *Stage 2 : POS tagged sentence to Ground Truth sentence*.

This section covers the second stage of our training pipeline, which entails mapping POS-tagged sentences to ground truth statements. The processing of our source and target texts, as well as the training approach, are discussed in the following paragraphs.

Data Processing(Source Text) : We read all of the ground truth sentences from the RWTH-PHOENIX-Weather 2014T(RPWT) dataset and tokenized them with the spacy german tokenizer for preparation. After that, we use the POS tags to replace proper nouns with PROPN and numbers with NUM. This is then fed into the transformer in the order that the time steps dictate.

Data Processing(Target Text) : The target text is taken from the aforementioned dataset's actual ground truth statements. The sentences are tokenized with the same spacy German tokenizer and sent in time steps to the transformer decoder.

Training : The training pipeline of our second stage is shown in figure 4. This architecture is similar to the previous transformer architecture with few differences in the source embedding side. To start with we first feed the processed source text obtained using algorithm 1. Similarly, for the decoder, we tokenize the actual ground truth sentences and are fed to the transformers.

The encoders in the transformer use a self-attention method. Those attentions are remapped to the textual embeddings of the decoder. The intermediate embeddings are sent through a linear layer and then to a softmax layer, which maps probability values in the vocab space, in the last few decoder levels. The decoder works iteratively with each iteration, returning the probabilities for the

t^{th} word in vocab space at the t^{th} time step. The translated natural sentence for the related POS tagged sentence given as an input is generated after all the words have been integrated.

In contrast to stage one, this time we're mapping the decoder's output to the actual vocab space, thus we get sentences back with the POS tags substituted by terms from the train vocab space.

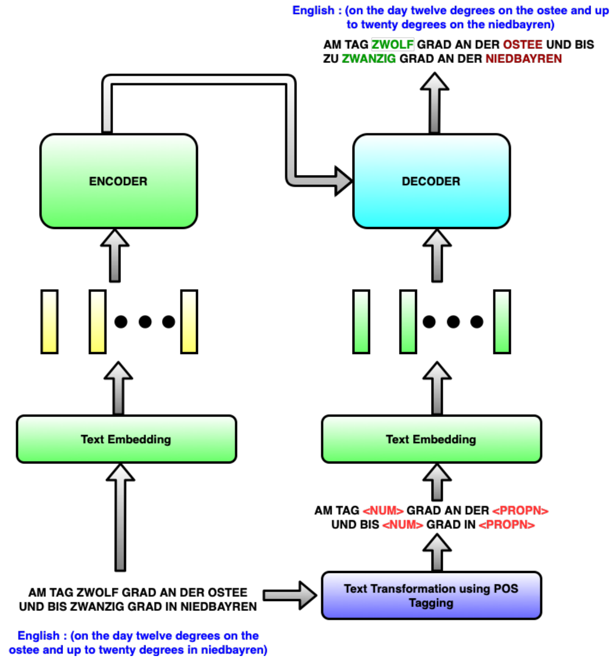
English : (on the day twelve degrees on the ostee and up to twenty degrees on the niedbayren)
AM TAG ZWOLF GRAD AN DER OSTEE UND BIS ZU ZWANZIG GRAD AN DER NIEDBAYREN
ENCODER
DECODER
Text Embedding
Text Embedding
AM TAG <NUM> GRAD AN DER <PROPN> UND BIS <NUM> GRAD IN <PROPN>
AM TAG ZWOLF GRAD AN DER OSTEE UND BIS ZWANZIG GRAD IN NIEDBAYREN
Text Transformation using POS Tagging
English : (on the day twelve degrees on the ostee and up to twenty degrees in niedbayren)

Figure 4: Training Pipeline of our second stage to replace the POS tags, for detailed working read section 4.4.2.

4.5 Talking Face Generation

Generating spoken languages from sign language videos was our primary interest which we have discussed in our introduction. We discuss the two-staged method to generate talking faced avatars from sentences with the talking face video of its corresponding sign language counterpart. In the upcoming section 4.5.1 we discuss the generation of speech from sign language translation, and in section 4.5.2 we discuss the generation of the talking face avatar.

4.5.1 Text-to-Speech. A German Text-To-Speech (TTS)is used to generate spoken language sentences from the translation sentences. Amazon Polly, a service that turns text into life-like speech, allows us to create applications that talk and build entirely new categories of speech-enabled products

4.5.2 Wav2Lip. Realistic generation of talking face videos was achieved by a few recent works [26, 39] We have used Wav2Lip [33] for generating speaking faces video, which focuses on lip-syncing unconstrained talking face videos to match any target speech, not limited by identities, voices, or vocabulary.

5 EXPERIMENTS

5.1 Dataset and Metrics

We evaluate the RWTH-PHOENIX-Weather 2014T (RPWT) [4]. It is the only standard sign language translation dataset publicly available for training and inference. We use 7096, 519, 642 videos for training, validation and test sets, respectively, to follow the official RPWT data partition protocol. Nine signers in German Sign Language (GSL) perform, these samples and translations in English are also made available. The RPWT dataset contains a varied 3k German word vocabulary. This distinguishes sign language translation from most tasks of vision and language which usually have a small vocabulary and a simple phrase structure.

We use BLEU [31]] scores, which is commonly used for machine translation, to measure our translation performance. As a BLEU score, we report BLEU-1,2,3,4 to provide better insights into the performance of the translation at different sentence levels. We also use ROUGE-L [28] that measures the F1 score based on the longest common sub-sequences between predictions and ground-truth translations.

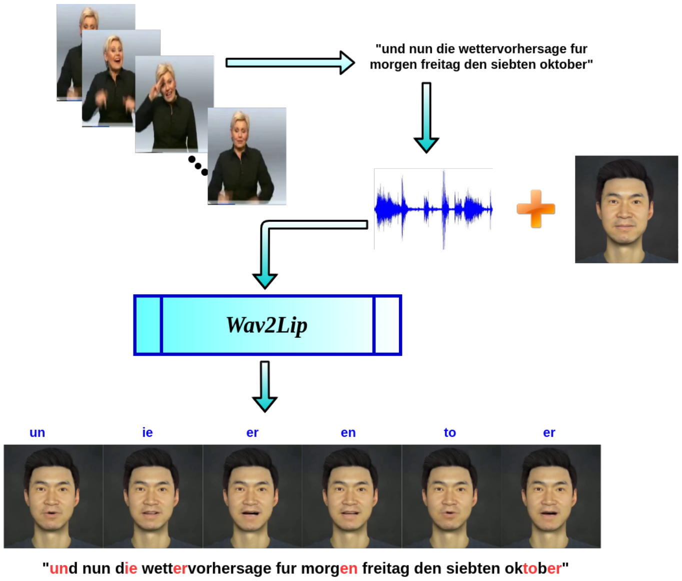
"und nun die wettervorhersage fur morgen freitag den siebten oktober"
Wav2Lip
un
ie
er
en
to
er
"und nun die wettervorhersage fur morgen freitag den siebten oktober"

Figure 5: Qualitative Evaluation of the Lip Synchronization module i.e. the Wav2Lip. Here as the output from Wav2Lip module, we show 6 faces that corresponds to the 6 red lettered phonemes in the words in the German sentence given below the faces. The individual phonemes are also shown in blue for each of the corresponding faces.

5.2 Quantitative Results

We perform experiments to improve the baseline scores for both the architectures SLT [4] and TSPNet [27] using the RWTH-PHOENIX-Weather 2014T [4] data set. All the networks are trained until the perplexity of the training converges. We evaluate our dev/test models for each epoch and report results (in Table : 3) using the model with the best results on the dev set for both the architectures.

Using POS tagging and a language model, we report that our scores are better than the SLT [4] by 3 points BLEU increase and TSPNet [27] by 2 points BLEU increase. There is also substantial increase in the Rouge score for both the architectures, evident from table 3.

5.3 Qualitative Results

5.3.1 Sign Language Translation. Our qualitative results are shared in this section. The resulting translations are one of the obvious ways to qualify for them. We share samples from our SLT [4] and TSPNet [27]networks with POS tagging, in Table 4, together with basic German ground truth and the transformed ground truth. Here in the ground truth sentences, proper-nouns are shown green in colour and numbers are shown brown in colour. After processing the ground truth sentences using algorithm 1 each contextual proper-nouns are replaced by <PROPN> and numbers are replaced by <NUM>. Then using a language model we try to recover the proper nouns and numbers. For a better comprehension, the reader is given English translations for each German phrase. Our translations are pretty close to that of the transformed ground truth, and are grammatically correct with semantic consistency.

5.3.2 Talking Faces Generation. In this section, we share the qualitative results of our talking face generation avatar. In figure 5 we share the key frames uttering the phonemes. It can be seen evidently that the lip shapes are in sync with the phonemes. The generation of theses talking faces will be helpful for the deaf community to lipread and get a better understanding of the translated sentences. From the image it is readily evident that the lip movements for the corresponding phonemes that are highlighted match the actual human lip utterances.

6 CONCLUSION & FUTURE WORK

While it is effective to model sign language videos by using the aforementioned transformer based architectures, we notice adequate limitations in our model. From the qualitative evaluation section and table 4 it's already evident that our predicted translations are grammatically correct and are maintaining semantic consistency. From the qualitative we can see that we tried to get back the low frequency words i.e the proper nouns and numbers by training our language based transformer. The low frequency words are mostly uttered by lips or particularly by fingers. We also note that the work [5] obtains a score of 20.17 BLEU-4 using the glossary backbone networks of [24]. Our proposed procedure makes it much easier to minimise the need of expensive annotations to learn sign language translation models directly from natural language sources, e.g. subtitled TV News or Films.

Apart from having a substantial increase in the scores as evident from table 3 we also generated a talking face avatar for the corresponding sign videos. This provide people with hearing or speech loss a sense of complementary fulfillment and helps them overcome the tag "impaired". Using this they can participate in any events that include one way interaction, like TV-news reporting, teaching and attending conferences. Being a teacher they will be signing, while their avatars perform the complementary actions that a commoner will understand. As a speaker they can attend conferences where they will be able to speak, i.e. they can attend any events where one way interaction is the key.

As a part of future works we will include further modalities and will mitigate the gloss-less and gloss-reliant performance gap so that it will be useful in the wild. As a part of our second work generating speech to sign language performing avatar is a future scope for this paper.

We believe that our work sets a new paradigm for the sign language community. Our generated talking face avatars will be a great boon to the speech impaired community which will give them a sense of complementary fulfillment and help them overcome the tag "impaired".

REFERENCES

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Y. Bengio. 2014. Neural Machine Translation by Jointly Learning to Align and Translate. *ArXiv* (09 2014).

[2] Mark Borg and Kenneth Camilleri. [n.d.]. Sign Language Detection "in the Wild" with Recurrent Neural Networks.

[3] Jan Bungeroth and Hermann Ney. [n.d.]. Statistical sign language translation. ([n. d.]).

[4] Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. 2018. Neural Sign Language Translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

[5] 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 (CVPR).*

[6] J. Carreira and Andrew Zisserman. 2017. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. 4724–4733.

[7] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Fethi Bougares, Holger Schwenk, and Y. Bengio. 2014. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. (06 2014).

[8] Joon Son Chung, Amir Jamaludin, and Andrew Zisserman. 2017. You said that? *ArXiv* abs/1705.02966 (2017).

[9] Runpeng Cui, Hu Liu, and Changshui Zhang. [n.d.]. A Deep Neural Framework for Continuous Sign Language Recognition by Iterative Training. ([n. d.]).

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition.*

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

[12] Jens Forster, Christoph Schmidt, Thomas Hoyoux, Oscar Koller, Uwe Zelle, Justus Piater, and Hermann Ney. [n.d.]. RWTH-PHOENIX-Weather: A Large Vocabulary Sign Language Recognition and Translation Corpus.

[13] Jens Forster, Christoph Schmidt, Oscar Koller, Martin Bellgardt, and Hermann Ney. [n.d.]. Extensions of the Sign Language Recognition and Translation Corpus RWTH-PHOENIX-Weather.

[14] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. [n.d.]. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural 'networks.

[15] Thomas Hanke, Lutz König, Sven Wagner, and Silke Matthes. [n.d.]. DGS Corpus Dicta-Sign: The Hamburg Studio Setup.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. 770–778.

[17] Jie Huang, Wengang Zhou, Qilin Zhang, Houqiang Li, and Weiping Li. [n.d.]. Video-based Sign Language Recognition without Temporal Segmentation.

[18] Sergey Ioffe and Christian Szegedy. [n.d.]. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. ([n. d.]).

[19] Sergey Ioffe and Christian Szegedy. 2015. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. (2015).

[20] Hamid Joze and Oscar Koller. 2018. MS-ASL: A Large-Scale Data Set and Benchmark for Understanding American Sign Language.

[21] Prajwal K R, Rudrabha Mukhopadhyay, Jerin Philip, Abhishek Jha, Vinay Namboodiri, and C V Jawahar. 2019. Towards Automatic Face-to-Face Translation *(MM '19).* Association for Computing Machinery, 1428–1436.

[22] Nal Kalchbrenner and P. Blunsom. 2013. Recurrent continuous translation models. *EMNLP 2013 - 2013 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference* 3 (01 2013), 1700–1709.

[23] Diederik Kingma and Jimmy Ba. [n.d.]. Adam: A Method for Stochastic Optimization. ([n. d.]).

Methods	Rouge score	BLEU-1	BLEU-2	BLEU-3	BLEU-4
SLT [4]	25.28	32.24	19.03	12.83	9.58
TSPNet [27]	34.96	36.10	23.12	16.88	13.41
SLT [4] + POS tagging + Lang. model	**31.43**	**40.13**	**16.37**	**15.69**	**12.01**
TSPNet [27] + POS tagging + Lang. model	**40.20**	**41.86**	**41.11**	**19.80**	**15.82**

Table 3: Comparisons of translation results on RWTH-PHOENIX-Weather 2014T dataset from SLT, TSPNet and our improvements using POS tags. Here we give the Rouge-L and the BLEU scores. BLEU-1,2,3,4 means the BLEU score calculated using 1-gram, 2-gram, 3-gram and 4-gram respectively. Our scores are made bold for better understanding.

[24] Oscar Koller, Necati Camgoz, Hermann Ney, and Richard Bowden. 2019. Weakly Supervised Learning with Multi-Stream CNN-LSTM-HMMs to Discover Sequential Parallelism in Sign Language Videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence* PP (04 2019).

[25] Julia Kreutzer, Jasmijn Bastings, and Stefan Riezler. 2019. Joey NMT: A Minimalist NMT Toolkit for Novices.

[26] Rithesh Kumar, Jose M. R. Sotelo, Kundan Kumar, A. D. Brébisson, and Yoshua Bengio. 2018. ObamaNet: Photo-realistic lip-sync from text. *ArXiv* (2018).

[27] DONGXU LI, Chenchen Xu, Xin Yu, Kaihao Zhang, Benjamin Swift, Hanna Suominen, and Hongdong Li. 2020. TSPNet: Hierarchical Feature Learning via Temporal Semantic Pyramid for Sign Language Translation. In *Advances in Neural Information Processing Systems*.

[28] Chin-Yew Lin and Franz Josef Och. [n.d.]. Automatic Evaluation of Machine Translation Quality Using Longest Common Subsequence and Skip-Bigram Statistics.

[29] Minh-Thang Luong, Hieu Pham, and Christopher Manning. 2015. Effective Approaches to Attention-based Neural Machine Translation. (08 2015).

[30] Sara Morrissey. [n.d.]. Data-driven machine translation for sign languages. *Morrissey, Sara (2008) Data-driven machine translation for sign languages. PhD thesis, Dublin City University.* ([n. d.]).

[31] 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*.

[32] Roland Pfau, Martin Salzmann, and Markus Steinbach. 2018. The syntax of sign language agreement: Common ingredients, but unusual recipe. *Glossa: a journal of general linguistics* 3 (10 2018).

[33] K R Prajwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, and C.V. Jawahar. 2020. A Lip Sync Expert Is All You Need for Speech to Lip Generation In the Wild. In *Proceedings of the 28th ACM International Conference on Multimedia*. Association for Computing Machinery.

[34] Pınar Santemiz, Oya Aran, Murat Saraclar, and Lale Akarun. 2009. Automatic sign segmentation from continuous signing via multiple sequence alignment.

[35] Ben Saunders, Necati Camgoz, and Richard Bowden. 2021. Continuous 3D Multi-Channel Sign Language Production via Progressive Transformers and Mixture Density Networks. 129 (2021).

[36] Frank Shipman, Satyakiran Duggina, Caio Monteiro, and Ricardo Gutierrez-Osuna. 2017. Speed-Accuracy Tradeoffs for Detecting Sign Language Content in Video Sharing Sites.

[37] Ilya Sutskever, Oriol Vinyals, and Quoc Le. 2014. Sequence to Sequence Learning with Neural Networks. *Advances in Neural Information Processing Systems* 4 (09 2014).

[38] Rachel Sutton-Spence and Bencie Woll. 1999. *The Linguistics of British Sign Language: An Introduction.* Cambridge University Press. https://doi.org/10.1017/CBO9781139167048

[39] Supasorn Suwajanakorn, S. Seitz, and Ira Kemelmacher-Shlizerman. 2017. Synthesizing Obama. *ACM Transactions on Graphics (TOG)* (2017).

[40] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All You Need.

GT	und nun die wettervorhersage fur morgen freitag den **zwanzigsten** november (*and now the weather forecast for tomorrow, friday the twentieth of november*)
GT Tranformed	und nun die wettervorhersage fur morgen freitag den **<NUM>** november (*and now the weather forecast for tomorrow, friday the* ***<NUM>*** *of november*)
SLT + POS tag	und nun die wettervorhersage fur morgen freitag den **<NUM>** oktober (*and now the weather forecast for tomorrow, friday the* ***<NUM>*** *of october*)
SLT + POS tag + Lang. model	und nun die wettervorhersage fur morgen freitag den **zwanzigsten** oktober (*and now the weather forecast for tomorrow, friday the* ***twentieth*** *of october*)
TSPNet + POS tag	und nun die wettervorhersage für morgen freitag den **<NUM>** juli (*and now the weather forecast for tomorrow, friday the* ***<NUM>*** *of july*)
TSPNet + POS tag + Lang. model	und nun die wettervorhersage für morgen freitag den **zwanzig** juli (*and now the weather forecast for tomorrow, friday the* ***twenty*** *of july*)
GT	am tag **zwolf** grad an der **ostee** und bis **zwanzig** grad in **niedbayren** (*on the day twelve degrees on the ostee and up to twenty degrees in niedbayren*)
GT Transformed	am tag **<NUM>** grad an der **<PROPN>** und bis **<NUM>** grad in **<PROPN>** (*on the day* ***<NUM>*** *degrees on the* ***<PROPN>*** *and up to* ***<NUM>*** *degrees in* ***<PROPN>***)
SLT + POS tag	am tag **<NUM>** grad an der **<PROPN>** und bis zu **<NUM>** grad an der **<PROPN>** (*on the day* ***<NUM>*** *degrees on the* ***<PROPN>*** *and up to* ***<NUM>*** *degrees on the* ***<PROPN>***)
SLT + POS tag + Lang. model	am tag **dreissig** grad an der **schwächer** und bis zu **zwolf** grad an der **niedbayren** (*on the day* ***thirty*** *degrees on the* ***schwächer*** *and up to* ***tweleve*** *degrees on the* ***niedbayren***)
TSPNet + POS tag	es gelten entsprechende unwetterwarnungen des **<PROPN>** wetterdienstes im süden bis **<NUM>** grad (*corresponding storm warnings from the* ***<PROPN>*** *weather service apply in the south up to* ***<NUM>*** *degrees*)
TSPNet + POS tag + Lang. model	es gelten entsprechende unwetterwarnungen des **niedbayren** wetterdienstes im süden bis **zwanzig** grad (*corresponding storm warnings from the* ***niedbayren*** *weather service apply in the south up to* ***twenty*** *degrees*)
GT	und nun die wettervorhersage fur morgen mittowoch den **dreissigsten** marz (*and now the weather forecast for tomorrow wednesday the thirtieth of march*)
GT Transformed	und nun die wettervorhersage fur morgen mittowoch den **<NUM>** marz (*and now the weather forecast for tomorrow wednesday the* ***<NUM>*** *of march*)
SLT + POS tag	und nun die wettervorhersage fur morgen mittowoch den **<NUM>** marz (*and now the weather forecast for tomorrow wednesday the* ***<NUM>*** *of march*)
SLT + POS tag + Lang. model	und nun die wettervorhersage fur morgen mittowoch den **dreissig** marz (*and now the weather forecast for tomorrow wednesday the* ***thirty*** *of march*)
TSPNet + POS tag	und nun die wettervorhersage für morgen montag den **<NUM>** januar (*and now the weather forecast for tomorrow monday the* ***<NUM>*** *of january*)
TSPNet + POS tag + Lang. model	und nun die wettervorhersage für morgen montag den **dressigsten** januar (*and now the weather forecast for tomorrow monday the* ***thirtieth*** *of january*)

Table 4: Qualitative Evaluation of Translation Results on RWTH-PHOENIX-Weather 2014T(TEST) dataset from our improvised networks.GT refers to the ground truth translations. GT Transformed refers to the ground truth translations after implementation of algorithm 1. SLT + POS tag refers to the predicted translation by our improvised SLT [4] network. SLT + POS tag + Lang. model refers to the predicted translation by our trained language model on ground truth, which replaces the POS tags. TSPNet + POS tag refers to the predicted translation by our improvised TSPNet [27] network. TSPNet + POS tag + Lang. model refers to the predicted translation by our proposed language model. Each German sentence is accompanied by its equivalent English translation for the ease of understanding.