ZHAW

Project 2

Master of Science in Engineering

Speech Translation for Swiss German
Model output post-processing

Author:
Silas Rudolf

Supervisor
Prof. Dr. Mark Cieliebak

January 16, 2022

Abstract

This project is part of the contribution paper of ZHAW-CAI to the Shared Task "Swiss German Speech to Standard German Text" at the SwissText 2021 conference.

For the task, models based on the Fairseq, Jasper, and Wav2vec architectures were trained on multilingual, German, and Swiss-German data to generate translations for each spoken utterance. As one particular part of the pipeline, this report focuses on post-processing the final model output by applying different spelling-correction models based on the transformer architecture.

The Analysis shows how crucial the source quality of the sentences is when applying the spell-correction models. It displays the difference between low-quality input, where post-processing decreases the output score by up to 8%, to a higher source quality input, where post-processing is achieving an increase in up to 3.7% BLEU score with the BERT architecture.

Table of Contents

1 Introduction **2**

2 Transformers and related work **3**
2.1 Transformers in speech translation 4

3 Datasets **5**
3.1 Datasets for the ST models 5
3.2 Datasets for the Post-processing models 6
3.3 Tokenization 7

4 Experiments **8**
4.1 BERT 10
4.2 DistilBERT 13
4.3 ELECTRA 16
4.4 RoBERTa 18
4.5 SymSpell 21

5 Conclusion **23**

6 Appendix **26**

1 Introduction

In speech-to-text (STT) translation, the goal is to transform the source language's audio into the target language's text. Conventional methods achieve this by performing automatic speech recognition to convert the audio signal into the written source text, then transforming the source text into the target language using a Machine Translation (MT) model.

To successfully perform this approach, a standardized writing system of the source language and source language transcription is always needed. However, most do not have an acknowledged written form among the thousands of languages worldwide, such as Swiss German. Moreover, Swiss German does not have a standardized orthography as a collection of different regional dialects.

As a result, it functions as the default spoke a language in formal and informal situations. In contrast, for writing, the Standard German language is used. This can be a substantial challenge. There is a considerable linguistic distance between Swiss German dialects and Standard German. Developing a model for transcribing Swiss German speech into Standard German text involves end-to-end Speech Translation, which combines STT with Machine Translation (MT).

In order to eliminate the need for source language transcriptions, in addition to other technical considerations such as achieving globally optimized solutions, directly training on source language audio paired with target-language text translations is considered end-to-end Speech Translation.

In the context of the Shared Task "Swiss German Speech to Standard German Text" organized at Swisstext 2021, a solution was provided consisting of three models based on different architectures: Fairseq, Jasper, and Wav2vec XLSR-5, which were trained with various data sets, both in Standard German and Swiss German. Their predictions were subsequently fed into a majority voting algorithm to select the most reliable translation.

Next to the Language Models for Speech Recognition, transcript post-processing, an approach which is the topic of this research, is evaluated using text-only data by training a supervised "spelling correction"(SC) model to correct the errors made by the ST model explicitly. Instead of predicting the likelihood of emitting a word based on the surrounding context, the SC model only needs to identify likely errors in the ST model output and propose alternatives. Intuitively, this task highly depends on the baseline model's quality. If the model transcribes very well, this task can be reduced to simply copying the input transcript directly to the output.

2 Transformers and related work

The Transformer in NLP is a novel architecture that aims to solve sequence-to-sequence tasks while efficiently handling long-range dependencies. It relies entirely on self-attention to compute its input and output representations without using sequence-aligned RNNs or convolution. By design, RNNs require sequential calculations - the calculations of the next timestep cannot start until the result of the previous time step is available. This prohibits parallelization, which is addressed by Transformer architecture.

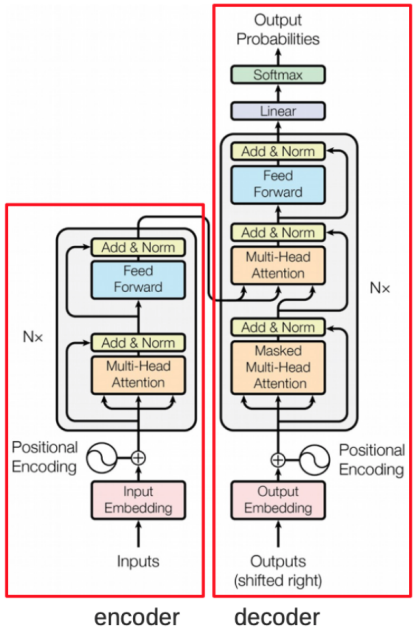

Figure 1: Encoder/Decoder architecture (Devlin et al. [2018])

Training of most transformers is done in two steps:

- Perform an unsupervised pre-training on a large amount of unlabeled data (the idea is to train a general-purpose language understanding model by learning a latent representation of the input text - only done once)
- Reuse the pre-trained models and perform an individual supervised fine-tuning on a small amount of labeled data for the downstream tasks

As such, transformers provide a Transfer Learning platform for NLP tasks (similar to computer vision, where models are pre-trained to learn general-purpose features of

images). More details about the architecture are described within the specific model sections in 4.

2.1 Transformers in speech translation

The first encoder-decoder architecture based on LSTM was introduced for ST by Berard et al. [2016] showing the feasibility of directly translating from the audio signal. Weiss et al. [2017] enhanced this approach by exploring settings with different numbers of layers in encoder and decoder and testing various multitask learning strategies. Bérard et al. [2018] trained a single model to translate English audiobooks into French and shown that pre-training the encoder on ASR data improves the final result. All these works showed that the input sequence length must be reduced to work with recurrent models.

Different directions have been evaluated to cope with the lack of end-to-end data. For instance, Anastasopoulos and Chiang [2018] performed analyses of different multitask settings to leverage more data. Bansal et al. [2019] shown that the pre-training of the encoder is also helpful when performed on a different language, in particular when the source language is low–resourced. Jia et al. [2019] increased the training data by using a large quantity of synthetic data that results in an end-to-end system able to outperform the cascade model. [1]

Also, for transcript post-processing, recent studies show the success in the use of transformer-based methods: Liao et al. [2021] use a modified RoBERTa structure and show an increase of 17.53 BLEU points on the self-augmented English Conversational Telephone Speech data set. On the LibriSpeech dataset, Hrinchuk et al. [2019] show promising results using a pre-trained BERT as initialization for their spell correction model, while Guo et al. [2019] takes a different approach with a bidirectional LSTM.

In addition, as misspellings are one of the challenges for this task, Sun et al. [2020] showed that BERT-style models could be erroneous in some cases so that they do not correctly process word sequences with misspellings.

[1] The first module, an Automatic Speech Recognition model (ASR) writes a transcription of the spoken sentence, and the second one, a Machine Translation (MT) model, translates the transcription to another language.

3 Datasets

3.1 Datasets for the ST models

The organizers provided a labeled data set containing 293 hours of audio recordings, mainly in the Bernese dialect, transcribed to Standard German. Since the alignment between the recordings and the transcripts was done automatically, each utterance has an Intersection over Union (IoU) score reflecting its alignment quality. Additionally, an unlabelled data set consisted of 1208 hours of recordings, mainly in the Zurich dialect. The solutions were evaluated based on a 13 hours test set, which contains recordings of speakers from all German-speaking parts of Switzerland. The dialect distribution of the test set is close to the actual Swiss German dialect distribution in Switzerland.

The audios used to train the ST models were extracted to 80-dimensional log Mel-scale filterbank features (windows with 25 ms size and 10 ms shift) and saved in NumPy format for the training. In addition, to alleviate overfitting, speech data transforms SpecAugment Park et al. [2019], adopted by Fairseq S2T, were applied. The additional datasets that were used are:

- SwissDial Pelin Dogan-Schönberger [2021]: 26 hours of Swiss German
- ArchiMob Tanja Samardzic [2016]: 80 hours of Swiss German
- Common Voice German v4: 483 hours of German[2]

The SwissDial dataset consists of 26 hours of audios in 8 different Swiss dialects with corresponding transcriptions in Swiss dialect and Standard German translations. The Swiss-German transcription rules differ between dialects.

ArchiMob contains 70 hours of audios in 14 different Swiss dialects with transcription in Swiss German, where each word is additionally provided with a Standard German normalization. The transcription rules are normalized and are equal for all dialects (Dieth transcription, Dieth and Schmid-Cadalbert [1986]). Familiar Voice German v4 consists of 483 hours of audios in Standard German with corresponding transcriptions.

[2] `https://commonvoice.mozilla.org/en/datasets/`

3.2 Datasets for the Post-processing models

For training and evaluation of the post-processing models, the outputs of the STT were used as source sentences, in combination with the German transcripts of the Shared Task training set used as target sentences. In the table below are some of the examples used for training. As can be seen, the mistakes made in the ASR output range from punctuation to grammatical mistakes and can lead to implausible sentences (which are intuitively very hard to correct).

Source (ST output)	Reference
der unsere sympathien auch für die schulden beratungs berner oberland andere	bei uns sind grosse sympathien für die schuldenberatung berner oberland vorhanden.
das unter sind eben andere institutionen abhängig	darunter sind wiederum institutionen von uns abhängig.
das inzwischen noch selber über den rachen	sie können inzwischen über sich selbst lachen.

Table 1: Example ST sentence output

In addition to the ST output, the extension of the data set with synthetic data was evaluated. As an augmentation technique, any number of letters are extended or removed from the target sentence and then used as a source sentence.

```
X ← randomChoices {KEEP, INSERT, DELETE}
if X = KEEP then return sentence
end if
words ← randomChoices of len(sentence)
if X = INSERT then
    for word in words do
        letters ← randomChoices {alphabet} of len(word)
        word ← words + randomPosition{letters}
        if X = DELETE then
            for word in words do
                letters ← randomChoices {word} of len(word)
                word ← words − letters
```

3.3 Tokenization

For encoding of the source sentences, WordPiece tokenization (Huggingface [2020]) is used with a default vocabulary of 30000. WordPiece tokenization is something in-between word-level and character-level tokenization and breaks words like ”playing” into the tokens ”play” and ”##ing.” This allows the model to make inferences based on word structure (two verbs ending in -ing have similar grammatical meaning, and two verbs starting with play - have similar semantic meaning). Beginnings and separation of sentences are tokenized as ”[CLS]” and ”[SEP]” respectively.

4 Experiments

In this section, different Transformer architectures are compared with their corresponding open-sourced pre-trained models and other post-processing methods.

For training the models, the following pipeline is used:

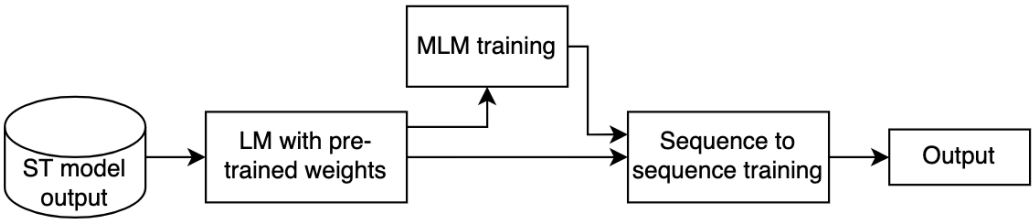

Figure 2: Train pipeline

In the first stage of the pipeline, the Speech Translation model output is fed into a pre-trained Language Model [3]. Next, the LM is used in an encoder-decoder structure for sequence-to-sequence training. Here two approaches are evaluated:

1. Training of the encoder separately by Masked Language Modeling (MLM), this for the encoder to learn the source language specifics (output of the Swiss-German ST models).

2. Using for both encoder and decoder the same pre-trained weights

For MLM, the goal is to solve the problem of working with bi-directional multi-layer architectures where, during training, the model could look up solutions and copy them. It works by randomly masking $k\%$ of the input words and predicting the original words based on the context. For the experiments, k is set to 15%.

In the last step, the output of the post-processing model is measured utilizing the BLEU score (Papineni et al. [2002]), which is a score for comparing a candidate translation of the text to one or more reference translations.

[3]The weights from https://huggingface.co were used as pre-trained embeddings

The transformer architectures and post-processing methods used for comparison are the following:

- BERT (Devlin et al. [2018]), having both encoder and decoder initialised with pre-trained weights.
- DistilBERT (Sanh et al. [2020]), the lightweight alternative to BERT, reducing the training time up to 60%.
- ELECTRA (Clark et al. [2020]), which uses a more sample-efficient pre-training approach for the encoder, called replaced token detection.
- RoBERTa (Liu et al. [2019]) which is similar to BERT with a more robust training approach.
- SymSpell (Garbe [2020]),which is a spelling correction algorithm for correcting spelling errors based on Damerau-Levenshtein distances, stored in a pre-trained dictionary.

The objective for all transformer models is set to next-sentence prediction (sequence to sequence generation) with a vocabulary size of 30'000, batch size of 16, and beam size for beam search set to 5.

Two sources are being considered for reporting the Shared task data set results. The first is the official model used for the competition (Ulasik et al. [2020]) that achieved an overall BLEU score of 38.7 on the test set. The second one was a later improved version of the model (Derju [2021]), with higher performance.

In Addition, performance concerning different source quality inputs is reported to verify the impact of a higher source model performance on the post-processing step. For this, samples of 500 sentences (outputs of the ST model) are taken with different source quality, ranging from low (with a BLEU score of 25-35) to high (with a BLEU score of 45-55).

4.1 BERT

Bidirectional Encoder Representations from Transformers (BERT) is a Transformer-based model, pre-trained on large corpora such as OpenLegalData (Ostendorff et al. [2020]), Wikipedia and News using two training objectives: Masked Language Modeling (MLM), for learning the context of a sentence, and Next Sentence Prediction (NSP) for learning the relationship between two sentences.

The input sentence(s) is decomposed into WordPiece tokens (3.3), which helps with the representation of the input vocabulary, reducing its size by segmenting complex words into sub-words.

By forming new words out of these sub-words that are not seen in the training samples, the model is more robust to out-of-vocabulary words. The WordPiece tokens are represented with three vectors: its token-, segment-, and position-embeddings.

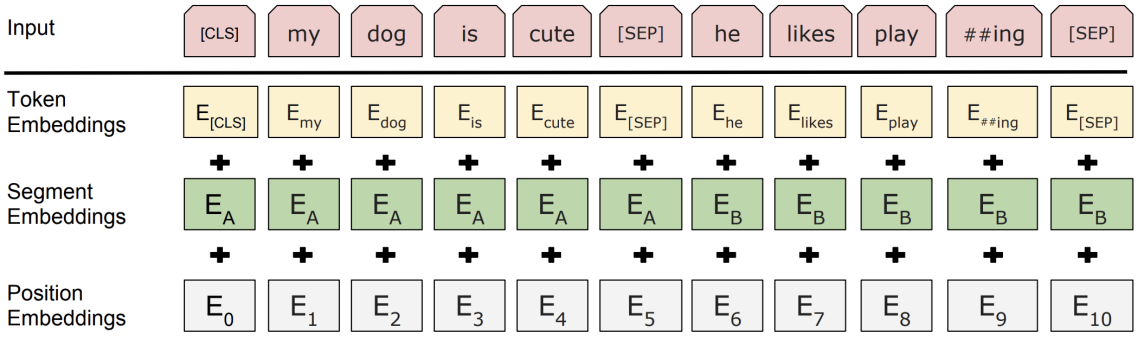

Figure 3: BERT embeddings (Devlin et al. [2018])

These embeddings are summed together and then passed through the transformer backbone (see 2), which produces the output representations fed into the final, application-dependent layer (e.g., a classifier for sentiment analysis).

In the first training round, a language modeling head is used as the final layer for specific fine-tuning on the Shared Task training set with Masked token prediction.

After training the MLM for 50 epochs, the cross-entropy loss between the logits and labels is stable at around 0.85. For evaluating the impact of MLM training, three models are compared on a test subset of 10'000 sentences, predicting one random masked word in each. The first model has no training, the second being trained for 20 epochs and the third being trained for 50 epochs. The masked word in each sentence is left out as a baseline comparison.

	Baseline	BERT	BERT(20)	BERT(50)
Training time	-	-	10.61h	26.37h
WER	0.067	0.051	0.028	**0.025**
BLEU score	83.33	85.25	92.97	**93.73**

Table 2: BERT MLM score and training time with different epochs

As expected, the model trained for 50 epochs performs best, reducing the error rate by 4% and increasing the BLEU score by 10%. However, similar results are already achieved after 20 epochs of training with a substantially lower training time. The loss of the model with its comparison to the other architectures can be found in section 6.

The next training step is combining the trained MLM into an encoder-decoder architecture for sequence-to-sequence training. The inputs for the encoder are the output of the Swiss-German to German ST model. Finally, the loss of the decoder is computed with its reference translation.

The following table shows the result of the original model (Ulasik et al. [2020]) on the Shared Task test set and the improved version (Derju [2021]) on sentence samples of 500, each with different source quality.

	STT output (Baseline)	Post-processed	Delta
BLEU score (ulasik)	38.70	23.26	-15.44
BLEU score (derju)	36.37	37.95	1.58
BLEU score (derju)	40.34	43.8	3.46
BLEU score (derju)	51.49	55.07	3.58

Table 3: BERT post-processing scores for different models and input quality

As the evaluations show, the original post-processing model decreases the overall BLEU score on the Shared Task test set. The main explanation for this is the poor quality of the ST output (that could also not be corrected by the human eye out of context).

For the improved model (derju), the score is increased on all samples, and intuitively, the higher the original ST output score is, the better the post-processing

performs. The following table shows some processed sentences with different source quality scores to visualize this.

Source (STT output)	das ist ein aktiven die umsetzung der stafvorlage mit etwas garnitur
Post-processed sentence	das ist ein **aktives eingreifen** der stafvorlage mit ein paar garnitur
Reference	es ist effektiv eben die umsetzung der stafvorlage mit ein wenig garnitur
BLEU score (source | processed)	36.03 | 23.70
Source (STT output)	das nicht gekümmert als das letzte sparprogramm gemacht haben
Post-processed sentence	das **war** nicht gekümmert als **wir** das letzte sparprogramm gemacht haben
Reference	das hat uns nicht gekümmert als wir das letzte sparprogramm gemacht haben
BLEU score (source | processed)	44.05 | 73.69
Source (STT output)	zum zweiten grossrat das vizepräsidium kommt es wie vor einem jahr zu einer echten auswahl
Post-processed sentence	beim zweiten **grossratsvizepräsidium** kommt es wie **schon** vor einem jahr zu einer echten auswahl
Reference	beim zweiten grossratsvizepräsidium kommt es wie schon vor einem jahr zu einer echten auswahl
BLEU score (source | processed)	50.39 | 100.00

Table 4: BERT post-processed sentences

As the difference between the first sentence with low source quality and the last sentence with high-quality shows, minor spelling and grammatical correctness errors can be improved by the post-processing model, while out of context meanings can not.

4.2 DistilBERT

In general, DistilBERT has the same general architecture as BERT. The main significant difference is that the number of layers is reduced by 2.

As stated by the authors, the reasoning behind this is that "...Most of the operations used in the Transformer architecture (linear layer and layer normalization) are highly optimized in modern linear algebra frameworks, and our investigations showed that variations on the last dimension of the tensor (hidden size dimension) have a smaller impact on computation efficiency (for a fixed parameters budget) than variations on other factors like the number of layers. Thus we focus on reducing the number of layers." (Sanh et al. [2020])

As with the previous model, the first training phase is Masked token prediction. The loss of the model with its comparison to the other architectures can be found in section 6. Again, the epochs are set to 20 since the previous evaluation showed that increasing the number of epochs only increases the precision by a small amount. As a baseline comparison, the masked word in each sentence is left out entirely.

	Baseline	DistilBERT	DistilBERT(20)
Training time	-	-	6.40h
WER	0.067	0.061	**0.030**
BLEU score	83.33	86.21	**92.48**

Table 5: DistilBERT MLM score and training time with different epochs

Having a similar performance as the typical BERT architecture, the significant difference is in the reduced training time, at about 60% of the original BERT model.

Also, the following training step for this architecture is to combine the trained MLM into an encoder-decoder system for sequence-to-sequence training. The source STT models and baselines are as described in 4.

	STT output (Baseline)	Post-processed	Delta
BLEU score (ulasik)	38.70	26.66	-12.04
BLEU score (derju)	36.37	37.21	1.58
BLEU score (derju)	40.34	42.51	2.17
BLEU score (derju)	51.49	47.96	-3.53

Table 6: DistilBERT post-processing scores for different models and input quality

For the original DistilBERT post-processing model, the overall BLEU score is decreased on the Shared Task test set, however being slightly better than the previous BERT model.

For the improved model (derju), the score is increased up to the last baseline, having a score of 51.49, where post-processing decreases the score. A closer look at some of the processed sentences might explain this behavior.

Source (STT output)	das ist ein aktiven die umsetzung der stafvorlage mit etwas garnitur
Post-processed sentence	das ist ein **wenig der** umsetzung der stafvorlage mit etwas garnitur
Reference	es ist effektiv eben die umsetzung der stafvorlage mit ein wenig garnitur
BLEU score (source | processed)	36.03 | 27.37
Source (STT output)	das nicht gekümmert als das letzte sparprogramm gemacht haben
Post-processed sentence	das **hat uns** nicht gekümmert als das letzte sparprogramm gemacht wurde
Reference	das hat uns nicht gekümmert als wir das letzte sparprogramm gemacht haben
BLEU score (source | processed)	44.05 | 71.03
Source (STT output)	zum zweiten grossrat das vizepräsidium kommt es wie vor einem jahr zu einer echten auswahl
Post-processed	zum zweiten **grossrat das heisst** kommt es wie vor einem jahr zu einer echten **wahl**
Reference	beim zweiten grossratsvizepräsidium kommt es wie schon vor einem jahr zu einer echten auswahl
BLEU score (source | processed)	50.39 | 42.3

Table 7: DistilBERT post-processed sentences

With the second and third sentences, some problems are visible such as changing the grammatical tense and not having learned specific words of the dataset (such as "grossratsvizepräsidium") and splitting the words into more often occurring ones.

4.3 ELECTRA

Reported in 2020, ELECTRA uses a different pre-training method as the other Transformer architectures called *replaced token detection*. Here, instead of masking $k\%$ of the input, some of the tokens are replaced with samples from a proposal distribution (typically the output of a small masked language model.

The corruption process solves a mismatch in BERT where the network sees artificial [MASK] tokens during pre-training but not when being fine-tuned on downstream tasks. The network is then pre-trained as a discriminator that predicts whether every token is an original or a replacement. In In contrast, MLM trains the network as a generator that predicts the original identities of the corrupted tokens. The intended key advantage is that the model learns from all input tokens instead of just the tiny masked-out subset, making it more computationally efficient.

Compared to the other models, the number of epochs for the replaced token detection pre-training task is kept at 20. The loss can be found in section 6. The masked word in each sentence is left out as a baseline comparison.

	Baseline	ELECTRA	ELECTRA(20)
Training time	-	-	4.37h
WER	0.067	0.042	**0.031**
BLEU score	83.33	90.12	**92.21**

Table 8: ELECTRA MLM score and training time with different epochs

Here we again get similar performance with the previous architectures. A slight difference is that the performance out-of-the-box is already relatively high and is only slightly increased with training. Following the sequence-to-sequence evaluation is displayed. The source STT models and baselines are as described in 4.

	STT output (Baseline)	Post-processed	Delta
BLEU score (ulasik)	38.70	14.77	-23.93
BLEU score (derju)	36.37	35.52	-0.85
BLEU score (derju)	40.34	42.55	2.21
BLEU score (derju)	51.49	49.19	-2.3

Table 9: ELECTRA post-processing scores for different models and input quality

The evaluations show that this architecture performs poorly on the sequence-to-sequence post-processing task. The processing step decreases the overall score for the original model and 2 of 3 improved ST model outputs. The improvement on one step might be out of pure luck due to the limited sample size.

The processed sentences for comparison are the following:

Source (STT output)	das ist ein aktiven die umsetzung der stafvorlage mit etwas garnitur
Post-processed sentence	das ist ein aktiven die umsetzung der stafvorlage **abend** etwas garnitur
Reference	es ist effektiv eben die umsetzung der stafvorlage mit ein wenig garnitur
BLEU score (source | processed)	36.03 | 24.64
Source (STT output)	das nicht gekümmert als das letzte sparprogramm gemacht haben
Post-processed sentence	das nicht gekümmert als das letzte sparprogramm gemacht haben
Reference	das hat uns nicht gekümmert als wir das letzte sparprogramm gemacht haben
BLEU score (source | processed)	44.05 | 44.05
Source (STT output)	zum zweiten grossrat das vizepräsidium kommt es wie vor einem jahr zu einer echten auswahl
Post-processed	zum zweiten grossrat das vizepräsidium kommt es wie vor einem jahr zu einer echten auswahl
Reference	beim zweiten grossratsvizepräsidium kommt es wie schon vor einem jahr zu einer echten auswahl
BLEU score (source | processed)	50.39 | 50.39

Table 10: ELECTRA post-processed sentences

Analyzing the sentences shows particular behavior. For the poor source sentences, the processing decreases the score even more, and for the better translations, nothing is changed.

This could mean that the short training time might not be enough in this task to adjust the language model to the context of the dataset.

4.4 RoBERTa

Although models such as XLNet and BERT brought significant performance gains, it is difficult to determine which aspects contributed the most, especially as training is computationally expensive and only limited tuning can be done. RoBERTa represents a replication study of BERT, emphasizing the impact of many vital hyperparameters and training data size. Key takeaways found BERT to be undertrained. Thus RoBERTa is an improved method for pre-training BERT to increase performance. Modifications include

1. More extended model training on larger data sets
2. Changing the masking pattern applied to training data
3. Training on longer sequences
4. Removing next-sentence prediction

The embeddings used for this experiment were pre-trained on 2.5TB of filtered CommonCrawl data containing 100 languages. This means it was pre-trained on the raw texts only, with no human labeling (which is why it can use lots of publicly available data), in an automatic process to generate inputs and labels from those texts.

As with the other experiments, the first training phase is Masked token prediction. The loss of the model with its comparison to the other architectures is found in section 6.

	Baseline	RoBERTa	RoBERTa(20)
Training time	-	-	33.06h
WER	0.067	0.047	**0.031**
BLEU score	89.13	90.12	**92.43**

Table 11: RoBERTa MLM score and training time with different epochs

RoBERTa has a similar performance as its predecessor, BERT, with the longest training time due to its size. Next, the model is combined to an encoder-decoder system for sequence-to-sequence training. The source STT models and baselines are as described in 4.

	STT output (Baseline)	Post-processed	Delta
BLEU score (ulasik)	38.70	25.3	-13.4
BLEU score (derju)	36.37	31.71	-4.66
BLEU score (derju)	40.34	40.55	0.21
BLEU score (derju)	51.49	53.02	1.53

Table 12: RoBERTa post-processing scores for different models and input quality

As with all post-processing methods, the overall BLEU score for the RoBERTA architecture is decreased on the Shared Task test set. Also, the improved model only shows a slight increase in performance, getting better with higher quality source sentences.

The processed sentences are in the table below:

Source (STT output)	das ist ein aktiven die umsetzung der stafvorlage mit etwas garnitur
Post-processed sentence	das ist ein aktives umsetzung der stafvorlage mit etwas garnitur
Reference	es ist effektiv eben die umsetzung der stafvorlage mit ein wenig garnitur
BLEU score (source | processed)	36.03 | 23.12
Source (STT output)	das nicht gekümmert als das letzte sparprogramm gemacht haben
Post-processed sentence	das nicht gekümmert als wir letzte sparprogramm gemacht wurde
Reference	das hat uns nicht gekümmert als wir das letzte sparprogramm gemacht haben
BLEU score (source | processed)	44.05 | 31.98
Source (STT output)	zum zweiten grossrat das vizepräsidium kommt es wie vor einem jahr zu einer echten auswahl
Post-processed	beim zweiten grossrat für das vizepräsidium kommt es wie vor einem jahr zu einer echten auswahl
Reference	beim zweiten grossratsvizepräsidium kommt es wie schon vor einem jahr zu einer echten auswahl
BLEU score (source | processed)	50.39 | 67.83

Table 13: RoBERTa post-processed sentences

As expected, the post-processing only improves the higher quality source sentences.

4.5 SymSpell

SymSpell is a Spelling correction algorithm that has, as of now, no published paper. However, about its functionality, the author writes: "..Utilizing Symmetric Deletion SymSpell reduces the complexity of edit candidate generation and dictionary lookup for a given Damerau-Levenshtein distance. In addition, it is six orders of magnitude faster (than the standard approach with deletes, transposes, replaces, inserts) while being language independent.

Opposite to other algorithms, only deletes are required. No transposes replace or inserts. Instead, transposes replace and inserts of the input term are transformed into deletes of the dictionary term. Replaces and inserts are expensive and language-dependent: e.g., Chinese with 70,000 Unicode Han characters.

The speed comes from the inexpensive delete-only edit candidate generation and the pre-calculation. An average five-letter word has about 3 million possible spelling errors within a maximum edit distance of 3. However, SymSpell needs to generate only 25 deletes to cover them all, both at pre-calculation and lookup time..."

Compared to the transformer models, the SymSpell algorithm is also applied to the original Shared Task test set and the samples with different baseline scores.

	STT output (Baseline)	Post-processed	Delta
BLEU score (ulasik)	38.70	30.65	-8.05
BLEU score (derju)	36.37	37.06	0.69
BLEU score (derju)	40.34	41.26	0.92
BLEU score (derju)	51.49	53.27	1.78

Table 14: SymSpell post-processing scores for different models and input quality

While performing best on the original dataset, the algorithm still decreases the overall performance by 8%. However, on the other samples, a slight increase can be found. The following table shows the same sentences compared to different source scores as with the previous models.

Source (STT output)	das ist ein aktiven die umsetzung der stafvorlage mit etwas garnitur
Post-processed sentence	das ist ein aktiven die umsetzung der stafvorlage mit etwas garnitur
Reference	es ist effektiv eben die umsetzung der stafvorlage mit ein wenig garnitur
BLEU score (source | processed)	36.03 | 36.03
Source (STT output)	das nicht gekümmert als das letzte sparprogramm gemacht haben
Post-processed sentence	das nicht gekümmert als das letzte sparprogramm gemacht haben
Reference	das hat uns nicht gekümmert als wir das letzte sparprogramm gemacht haben
BLEU score (source | processed)	44.05 | 44.05
Source (STT output)	zum zweiten grossrat das vizepräsidium kommt es wie vor einem jahr zu einer echten auswahl
Post-processed	zum zweiten grossrat das vizepräsidium kommt es wie vor einem jahr zu einer echten auswahl
Reference	beim zweiten grossratsvizepräsidium kommt es wie schon vor einem jahr zu einer echten auswahl
BLEU score (source | processed)	50.39 | 50.39

Table 15: SymSpell post-processed sentences

None of the displayed post-processed sentences differ from the source output. However, a closer analysis of the samples shows slight differences within longer sentences, where the SymSpell algorithm can correct some misspellings.

5 Conclusion

As the evaluations show, most post-processing attempts decrease the overall BLEU score on the original test set, with SymSpell as the most straightforward approach performing best. Compared with previous work in this area, this could be explained by the limited amount of data available for training the transformer models and low-quality STT outputs.

Baseline	BERT	DistilBERT	ELECTRA	RoBERTA	SymSpell
38.70	23.26	26.66	14.77	25.3	**30.65**

Table 16: Post-processing BLEU scores with the Ulasik et al. [2020] source model and the public test set reference

Therefore, due to lack of performance, the post-processing step is excluded in the final submission for the SharedTask.

For the improved model, some of the post-processing approaches have a good outline, with BERT (3) showing the most promising results. The summary performance of the different models is shown below. The complete list of comparison sentences can be found in 19.

Baseline	BERT	DistilBERT	ELECTRA	RoBERTA	SymSpell
36.37	**37.95**	37.21	35.52	31.71	37.06
40.34	**43.8**	42.51	42.55	40.55	41.26
51.49	**55.07**	47.96	49.19	53.02	53.27

Table 17: Post-processing BLEU scores with the Derju [2021] source model and sample sentences with different input quality

As for usage in a practical environment, it is essential to check individual scores on a sentence level. For this, an additional analysis is made with BERT as the best performing model on a subset of 500 sentences. Here the score is computed sentence by sentence to see how much improvement is achieved on average and if negative effects appear (good sentences being worsened to a non-comprehensible state).

The following histogram shows the delta for each sentence between the source BLEU score of the ST model and the BLEU score of the post-processing model.

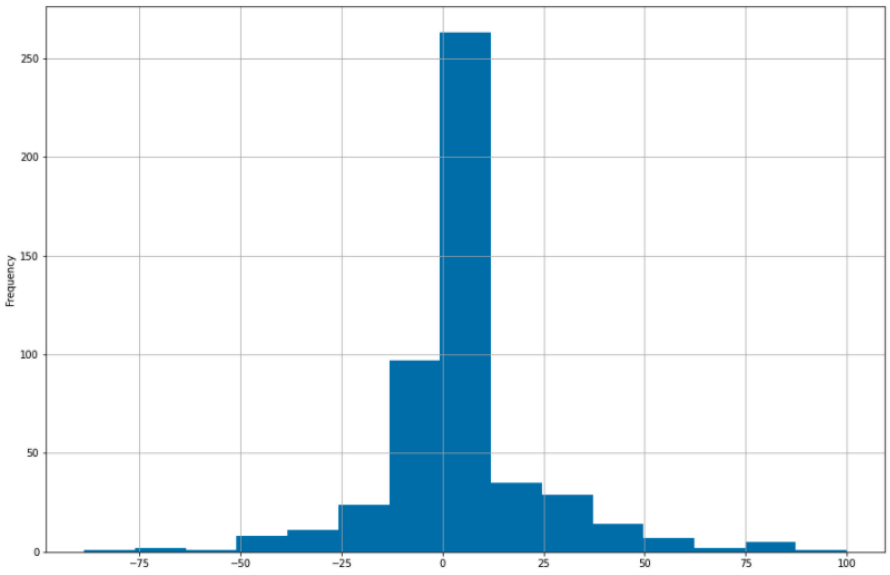

Figure 4: Histogram with sentence-level deltas between the ST source output score and the post-processing score

The slightly right-skewed histogram shows that most post-processing improves/slightly decreases the sentence quality, with some outliers on both tails. This is also shown in the box-plot below:

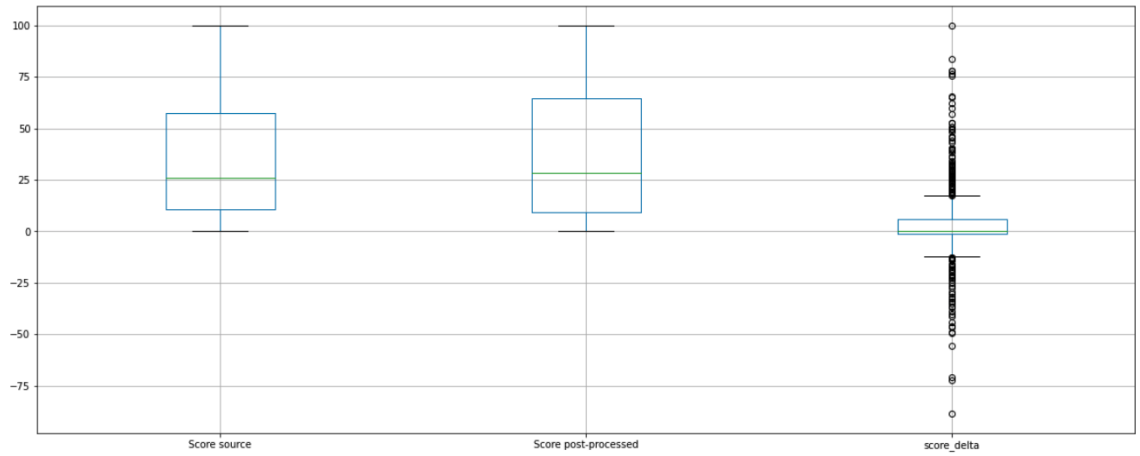

Figure 5: Box-plots with sentence-level deltas between the ST source output score and the post-processing score

For a more detailed view of the tails, the number of sentences is counted that decrease and improve the score by over 20 and 50 BLEU points. Out of the 500 sample sentences, 101 have a change in absolute score of over 20.

Decrease: $<$ **-50**	Decrease: $<$ **-20**	Improvement: $>$ **20**	Improvement: $>$ **50**
4	31	70	15

Table 18: Number of sentences with post processing score change of 20 and 50

As the final evaluations show, post-processing can be helpful with minor improvements when attached to an ST model output. However, there are some drawbacks, the most important being the high dependency on good quality source output and some unrecognized sentences being worsened by the model.

In general, post-processing could see an application in the future as the last step of a Speech-Translation pipeline for established models with high translation scores.

6 Appendix

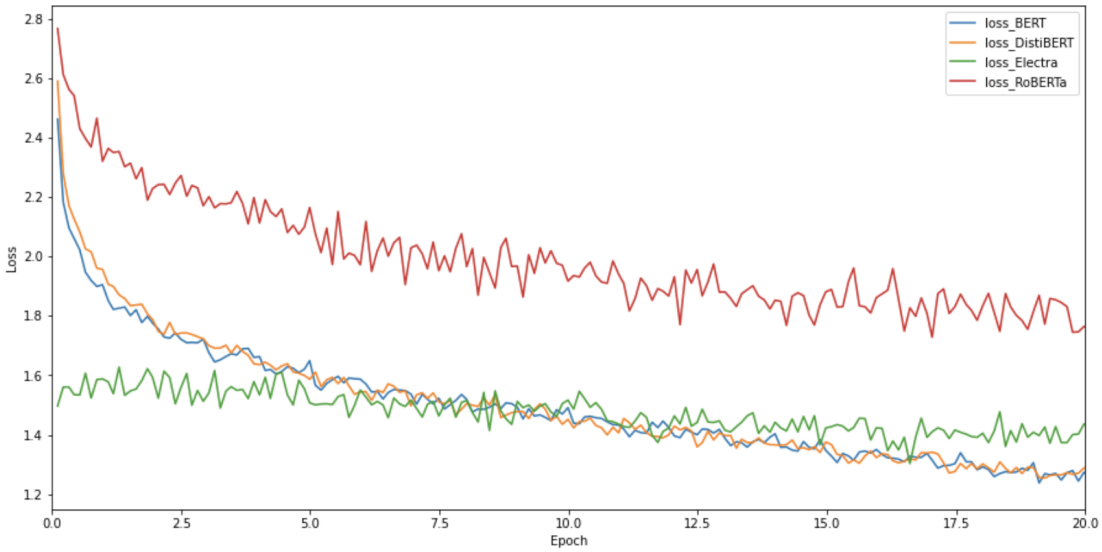

Figure 6: MLM losses of the different architectures

Source (STT output)	das ist ein aktiven die umsetzung der stafvorlage mit etwas garnitur
BERT	das ist ein **aktives eingreifen** der stafvorlage mit **ein paar** garnitur
DistilBERT	das ist ein **wenig der** umsetzung der stafvorlage mit etwas garnitur
ELECTRA	das nicht gekümmert als das letzte sparprogramm gemacht haben
RoBERTa	das ist ein aktives umsetzung der stafvorlage mit etwas garnitur
SymSpell	das ist ein aktiven die umsetzung der stafvorlage mit etwas garnitur
Reference	es ist effektiv eben die umsetzung der stafvorlage mit ein wenig garnitur
Source (STT output)	das nicht gekümmert als das letzte sparprogramm gemacht haben
BERT	das **war** nicht gekümmert als **wir** das letzte sparprogramm gemacht haben
DistilBERT	das **hat uns** nicht gekümmert als das letzte sparprogramm gemacht wurde
ELECTRA	das nicht gekümmert als das letzte sparprogramm gemacht haben
RoBERTa	das nicht gekümmert als wir letzte sparprogramm gemacht wurde
SymSpell	das nicht gekümmert als das letzte sparprogramm gemacht haben
Reference	das hat uns nicht gekümmert als wir das letzte sparprogramm gemacht haben
Source (STT output)	zum zweiten grossrat das vizepräsidium kommt es wie vor einem jahr zu einer echten auswahl
BERT	beim zweiten **grossratsvizepräsidium** kommt es wie **schon** vor einem jahr zu einer echten auswahl
DistilBERT	zum zweiten **grossrat das heisst** kommt es wie vor einem jahr zu einer echten **wahl**
ELECTRA	zum zweiten grossrat das vizepräsidium kommt es wie vor einem jahr zu einer echten auswahl
RoBERTa	beim zweiten grossrat für das vizepräsidium kommt es wie vor einem jahr zu einer echten auswahl
SymSpell	zum zweiten grossrat das vizepräsidium kommt es wie vor einem jahr zu einer echten auswahl
Reference	beim zweiten grossratsvizepräsidium kommt es wie schon vor einem jahr zu einer echten auswahl

Table 19: Post-processed sentences

References

Alexis Conneau, Alexei Baevski, R.C., 2020. Unsupervised Cross-lingual Representation Learning for Speech Recognition. Facebook AI. URL: https://arxiv.org/pdf/2006.13979.pdf.

Anastasopoulos, A., Chiang, D., 2018. Tied multitask learning for neural speech translation. arXiv:1802.06655.

Ando, R.K., Zhang, T., 2005. A Framework for Learning Predictive Structures from Multiple Tasks and Unlabeled Data. Journal of Machine Learning Research 6, 1817–1853.

Baevski, A., Zhou, H., Mohamed, A., Auli, M., 2020. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. Facebook AI. URL: https://arxiv.org/pdf/2006.11477.pdf.

Bansal, S., Kamper, H., Livescu, K., Lopez, A., Goldwater, S., 2019. Pre-training on high-resource speech recognition improves low-resource speech-to-text translation. arXiv:1809.01431.

Berard, A., Pietquin, O., Servan, C., Besacier, L., 2016. Listen and translate: A proof of concept for end-to-end speech-to-text translation. arXiv:1612.01744.

Büchi, M., Ulasik, M.A., Hürlimann, M., Benites, F., von Däniken, P., Cieliebak, M., 2020. ZHAW-InIT at GermEval 2020 Task 4: Low-Resource Speech-to-Text, in: Proceedings of the 5th Swiss Text Analytics Conference (SwissText) & 16th Conference on Natural Language Processing (KONVENS), CEUR-WS.

Bérard, A., Besacier, L., Kocabiyikoglu, A.C., Pietquin, O., 2018. End-to-end automatic speech translation of audiobooks. arXiv:1802.04200.

Changhan Wang, Juan Pino, J.G., 2020. Improving Cross-Lingual Transfer Learning for End-to-End Speech Recognition with Speech Translation. arXiv:2006.05474.

Clark, K., Luong, M.T., Le, Q.V., Manning, C.D., 2020. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. arXiv:2003.10555.

Derju, J.M., 2021. Unpublished work - improved speech to text model for swiss german. URL: derju@zhaw.ch.

Devlin, J., Chang, M.W., Lee, K., Toutanova, K., 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805.

Dieth, E., Schmid-Cadalbert, C., 1986. Schwyzertütschi dialäktschrift. Sauerländer, Aarau 2.

Garbe, W., 2020. SymSpell: Fast spell correction algorithm. URL: https://github.com/wolfgarbe/symspell.

Guo, J., Sainath, T.N., Weiss, R.J., 2019. A Spelling Correction Model for End-to-End Speech Recognition. arXiv:1902.07178.

Hrinchuk, O., Popova, M., Ginsburg, B., 2019. Correction of Automatic Speech Recognition with Transformer Sequence-to-sequence Model. arXiv:1910.10697.

Huggingface, 2020. WordPiece Tokenizer. URL: https://huggingface.co/docs/tokenizers/python/latest/api/reference.html#tokenizers.models.WordPiece.

Iranzo-Sánchez, J., Silvestre-Cerdà, J.A., Jorge, J., Roselló, N., Giménez, A., Sanchis, A., Civera, J., Juan, A., 2020. Europarl-ST: A Multilingual Corpus for Speech Translation of Parliamentary Debates, in: ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 8229–8233. doi:10.1109/ICASSP40776.2020.9054626.

Jia, Y., Johnson, M., Macherey, W., Weiss, R.J., Cao, Y., Chiu, C.C., Ari, N., Laurenzo, S., Wu, Y., 2019. Leveraging weakly supervised data to improve end-to-end speech-to-text translation. arXiv:1811.02050.

Kahn, J., Lee, A., Hannun, A., 2020. Self-training for End-to-End Speech Recognition, in: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE. pp. 7084–7088.

Li, J., Lavrukhin, V., Ginsburg, B., Leary, R., Kuchaiev, O., Cohen, J.M., Nguyen, H., Gadde, R.T., 2019. Jasper: An End-to-End Convolutional Neural Acoustic Model, in: Proceedings of Interspeech 2019, pp. 71–75.

Liao, J., Shi, Y., Gong, M., Shou, L., Eskimez, S., Lu, L., Qu, H., Zeng, M., 2021. Generating Human Readable Transcript for Automatic Speech Recognition with Pre-trained Language Model. arXiv:2102.11114.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V., 2019. Roberta: A robustly optimized bert pretraining approach. arXiv:1907.11692.

Ney, H., 1999. Speech Translation: coupling of recognition and translation, in: 1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No.99CH36258), pp. 517–520 vol.1. doi:10.1109/ICASSP.1999.758176.

Ostendorff, M., Blume, T., Ostendorff, S., 2020. Towards an open platform for legal information, in: Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020, Association for Computing Machinery, New York, NY, USA. p. 385–388. URL: https://doi.org/10.1145/3383583.3398616, doi:10.1145/3383583.3398616.

Panayotov, V., Chen, G., Povey, D., Khudanpur, S., 2015. Librispeech: An ASR corpus based on public domain audio books, in: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5206–5210. doi:10.1109/ICASSP.2015.7178964.

Papineni, K., Roukos, S., Ward, T., Zhu, W.J., 2002. BLEU: a method for automatic evaluation of machine translation, in: Proceedings of the 40th annual meeting of the Association for Computational Linguistics, pp. 311–318.

Park, D.S., Chan, W., Zhang, Y., Chiu, C.C., Zoph, B., Cubuk, E.D., Le, Q.V., 2019. SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. arXiv:1904.08779.

Pelin Dogan-Schönberger, Julian Mäder, T.H., 2021. SwissDial: Parallel Multidialectal Corpus of Spoken Swiss German. arXiv:2103.11401v1.

Plüss, M., Neukom, L., Vogel, M., 2021. SwissText 2021 Task 3: Swiss German Speech to Standard German Text. In preparation.

Sanh, V., Debut, L., Chaumond, J., Wolf, T., 2020. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv:1910.01108.

Schultz, T., Jou, S., Vogel, S., Saleem, S., 2004. Using Word Lattice Information for a Tighter Coupling in Speech Translation Systems, in: INTERSPEECH.

Siebenhaar, B., 2003. Sprachgeographische Aspekte der Morphologie und Verschriftung in schweizerdeutschen Chats. Linguistik online 15.

Siebenhaar, B., Wyler, A., 1997. Dialekt und Hochsprache in der deutschsprachigen Schweiz. Pro Helvetia.

Sperber, M., Paulik, M., 2020. Speech Translation and the End-to-End Promise: Taking Stock of Where We Are, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 7409–7421.

Sun, L., Hashimoto, K., Yin, W., Asai, A., Li, J., Yu, P., Xiong, C., 2020. Adv-bert: Bert is not robust on misspellings! generating nature adversarial samples on bert. arXiv:2003.04985.

Taku Kudo, J.R., 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. arXiv:1808.06226.

Tanja Samardzic, Yves Scherrer, E.G., 2016. ArchiMob - A Corpus of Spoken Swiss German. URL: https://www.aclweb.org/anthology/L16-1641.

Ulasik, A., Hurlimann, M., Dubel, B., Kaufmann, Y., Rudolf, S., 2020. Zhawcai: Ensemble method for swiss german speech to standard german text. URL: http://ceur-ws.org/Vol-2957/sg_paper3.pdf.

Waibel, A., Jain, A., McNair, A., Saito, H., Hauptmann, A., Tebelskis, J., 1991. JANUS: a speech-to-speech translation system using connectionist and symbolic processing strategies, in: [Proceedings] ICASSP 91: 1991 International Conference on Acoustics, Speech, and Signal Processing, pp. 793–796 vol.2. doi:10.1109/ICASSP.1991.150456.

Wang, C., Tang, Y., Ma, X., Wu, A., Okhonko, D., Pino, J., 2020a. fairseq S2T: Fast Speech-to-Text Modeling with fairseq. arXiv:2010.05171.

Wang, C., Wu, Y., Liu, S., Yang, Z., Zhou, M., 2020b. Bridging the Gap between Pre-Training and Fine-Tuning for End-to-End Speech Translation, in: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 9161–9168. URL: https://ojs.aaai.org/index.php/AAAI/article/view/6452, doi:10.1609/aaai.v34i05.6452.

Weiss, R.J., Chorowski, J., Jaitly, N., Wu, Y., Chen, Z., 2017. Sequence-to-sequence models can directly translate foreign speech. arXiv:1703.08581.

Woszczyna, M., Coccaro, N., Eisele, A., Lavie, A., McNair, A., Polzin, T., Rogina, I., Rose, C.P., Sloboda, T., Tomita, M., Tsutsumi, J., Aoki-Waibel, N., Waibel,

A., Ward, W., 1993. Recent Advances in Janus: A Speech Translation System, in: Proceedings of the Workshop on Human Language Technology, Association for Computational Linguistics, USA. p. 211–216. URL: https://doi.org/10.3115/1075671.1075718, doi:10.3115/1075671.1075718.

Xu, Q., Likhomanenko, T., Kahn, J., Hannun, A., Synnaeve, G., Collobert, R., 2020. Iterative Pseudo-Labeling for Speech Recognition, in: Proceedings of Interspeech 2020, pp. 1006–1010.

List of Figures

1 Encoder/Decoder architecture (Devlin et al. [2018]) 3
2 Train pipeline 8
3 BERT embeddings (Devlin et al. [2018]) 10
4 Histogram with sentence-level deltas between the ST source output score and the post-processing score 24
5 Box-plots with sentence-level deltas between the ST source output score and the post-processing score 24
6 MLM losses of the different architectures 26

List of Tables

1 Example ST sentence output 6
2 BERT MLM score and training time with different epochs 11
3 BERT post-processing scores for different models and input quality . 11
4 BERT post-processed sentences 12
5 DistilBERT MLM score and training time with different epochs . . . 13
6 DistilBERT post-processing scores for different models and input quality 14
7 DistilBERT post-processed sentences 15
8 ELECTRA MLM score and training time with different epochs 16
9 ELECTRA post-processing scores for different models and input quality 16
10 ELECTRA post-processed sentences 17
11 RoBERTa MLM score and training time with different epochs 18
12 RoBERTa post-processing scores for different models and input quality 19
13 RoBERTa post-processed sentences 20
14 SymSpell post-processing scores for different models and input quality 21
15 SymSpell post-processed sentences 22
16 Post-processing BLEU scores with the Ulasik et al. [2020] source model and the public test set reference 23
17 Post-processing BLEU scores with the Derju [2021] source model and sample sentences with different input quality 23
18 Number of sentences with post processing score change of 20 and 50 . 25
19 Post-processed sentences 27