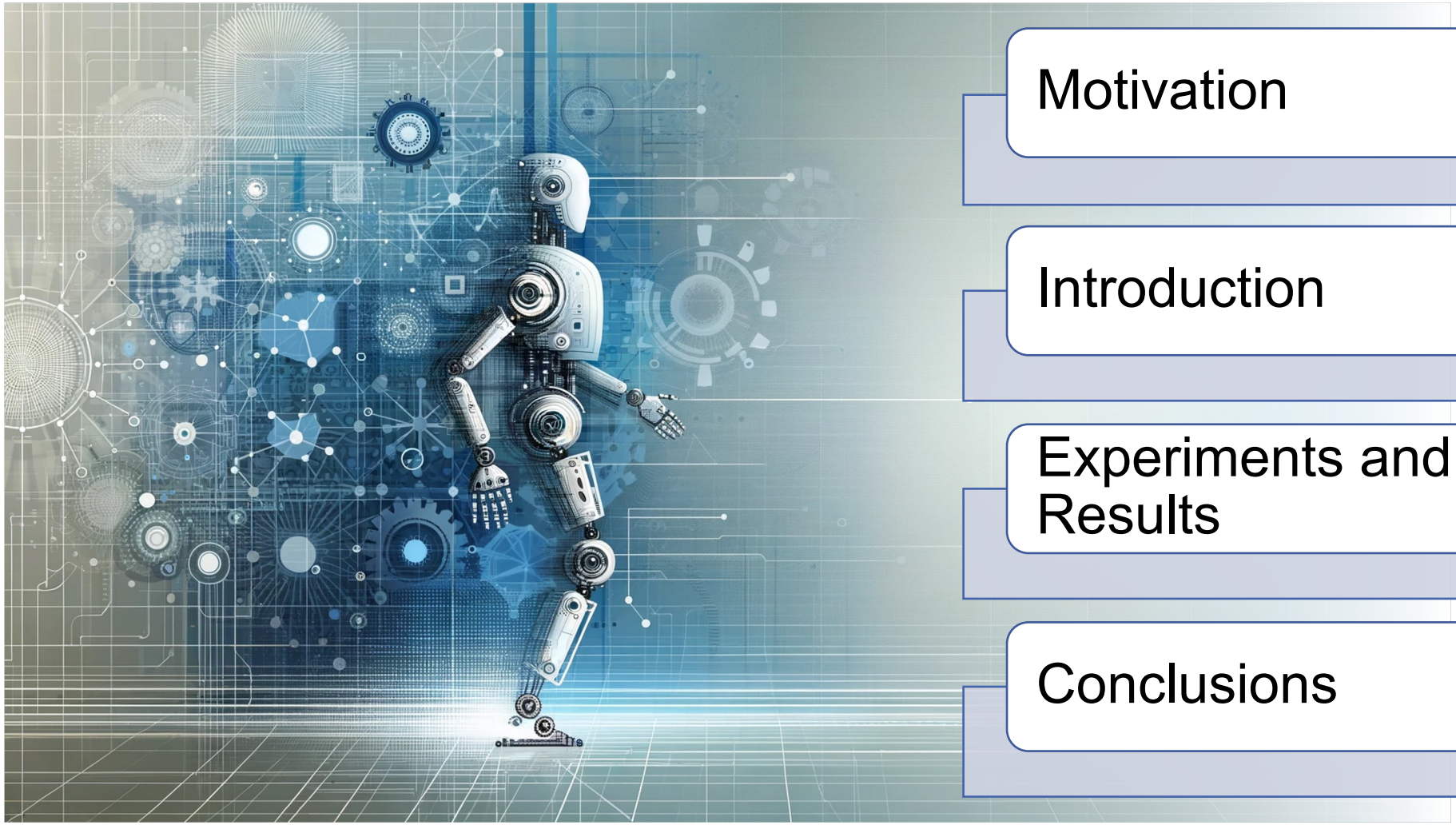
Karlsruher Institut für Technologie

Advanced Speech Translation

Praktikum Speech Translation

Bertil Braun, Arvand Kaveh, Marius Bohnert | February 14, 2024

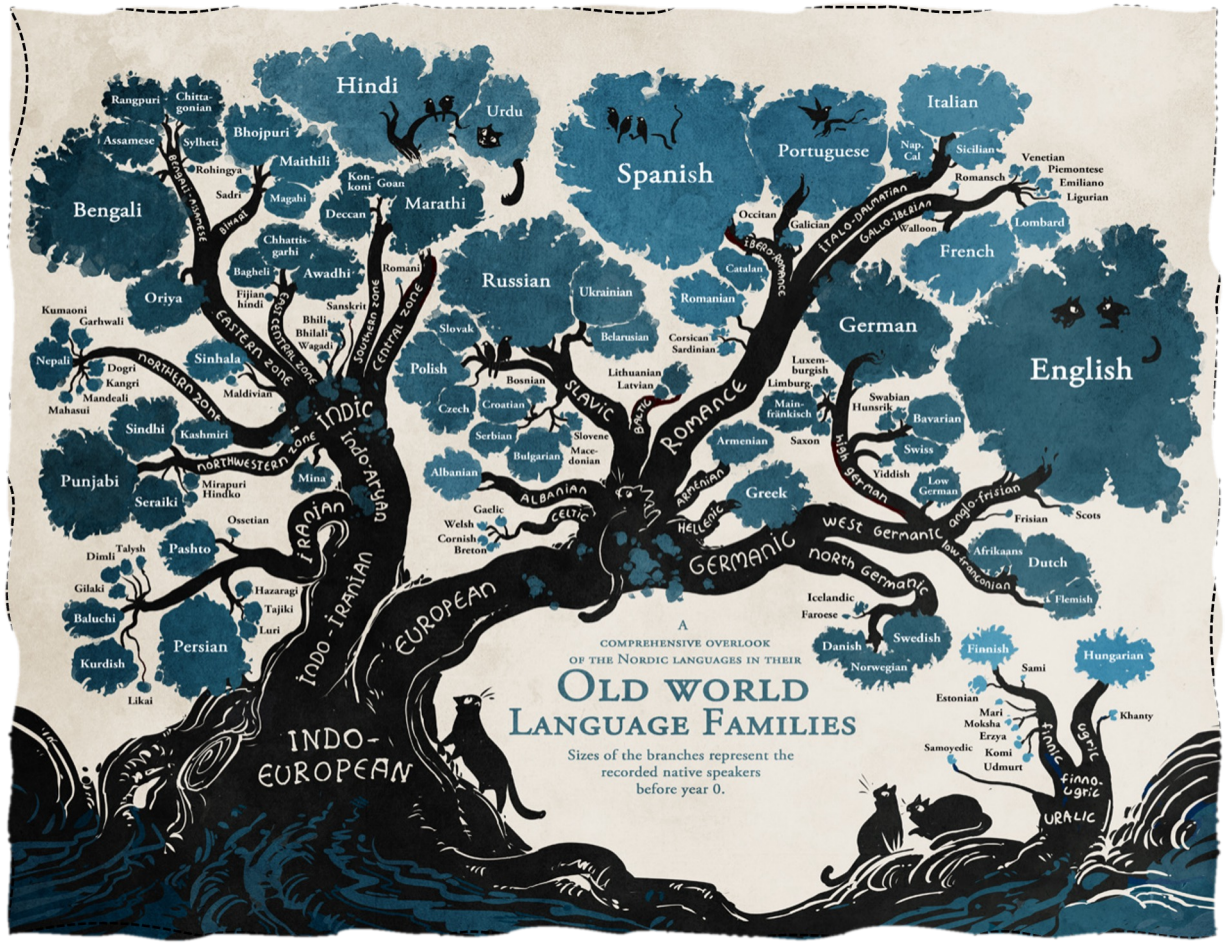
Overview



- Motivation
- Introduction
- Experiments and Results
- Conclusions

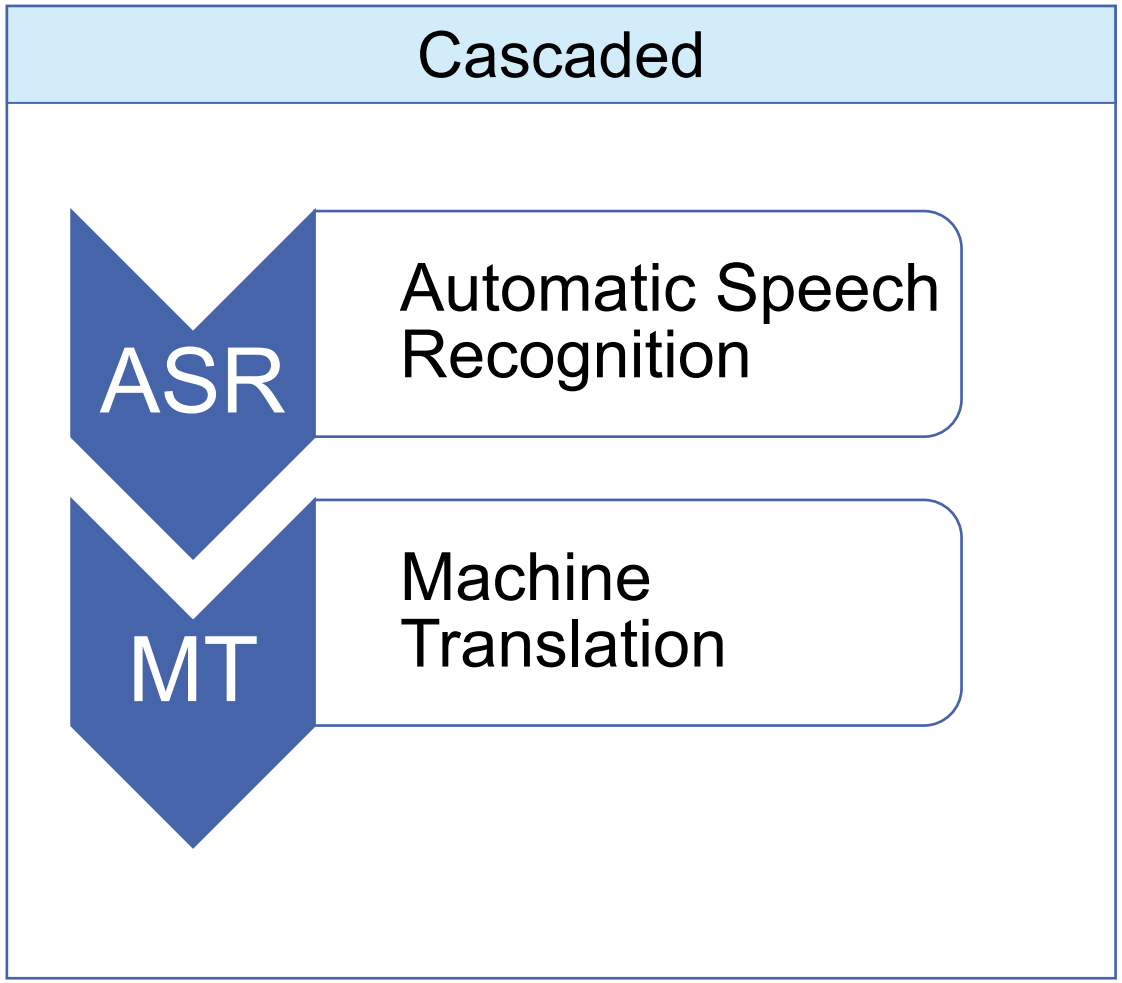
Motivation

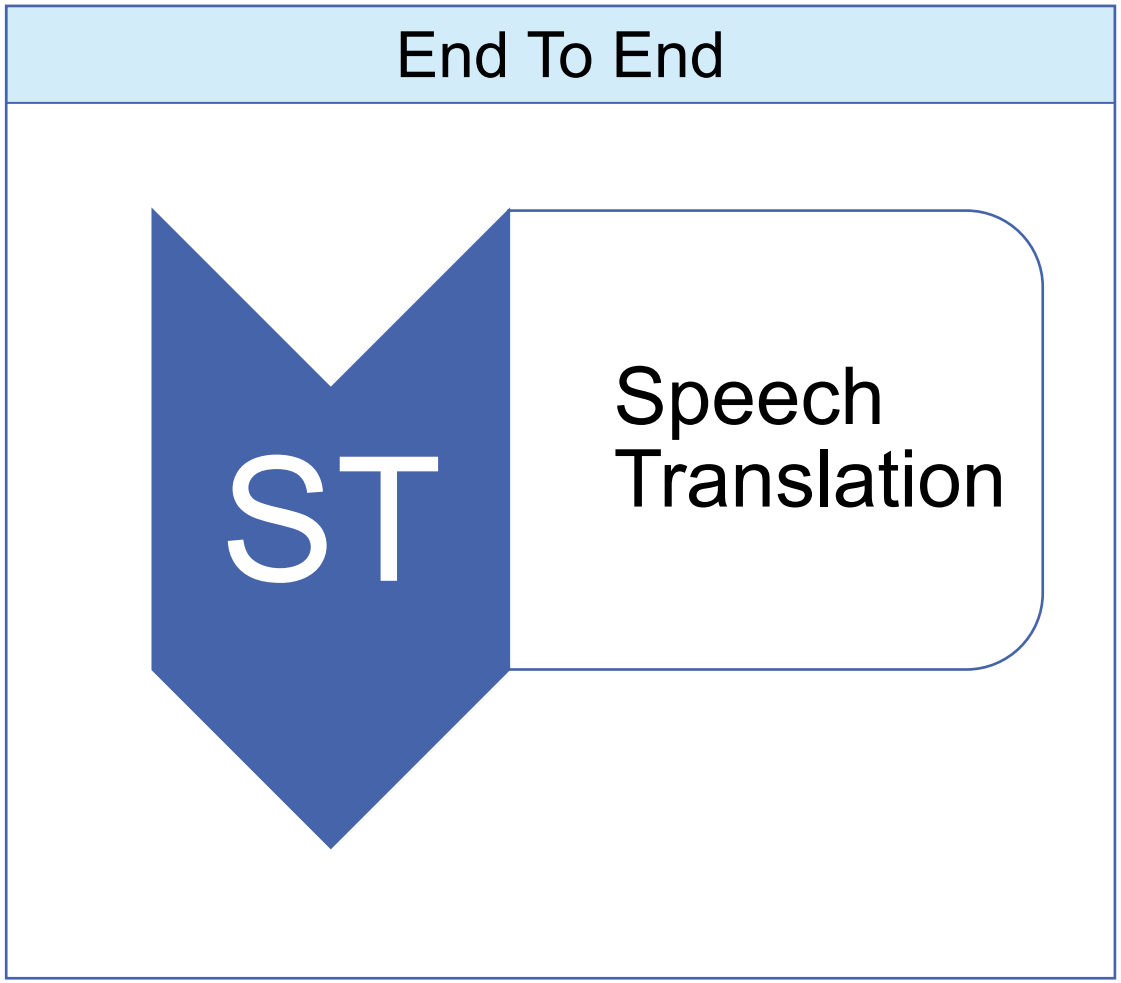
- Ca. 7,000 spoken languages
- Each citizen of EU should know:
  - 1 + 2 Languages
- EU average:
  - 2.32 Languages
- Only solution:
  - Automatic Speech Translation



Source: [1] Anderson 2024, [2] 2020, and [3] Young 2015

Introduction
Approaches

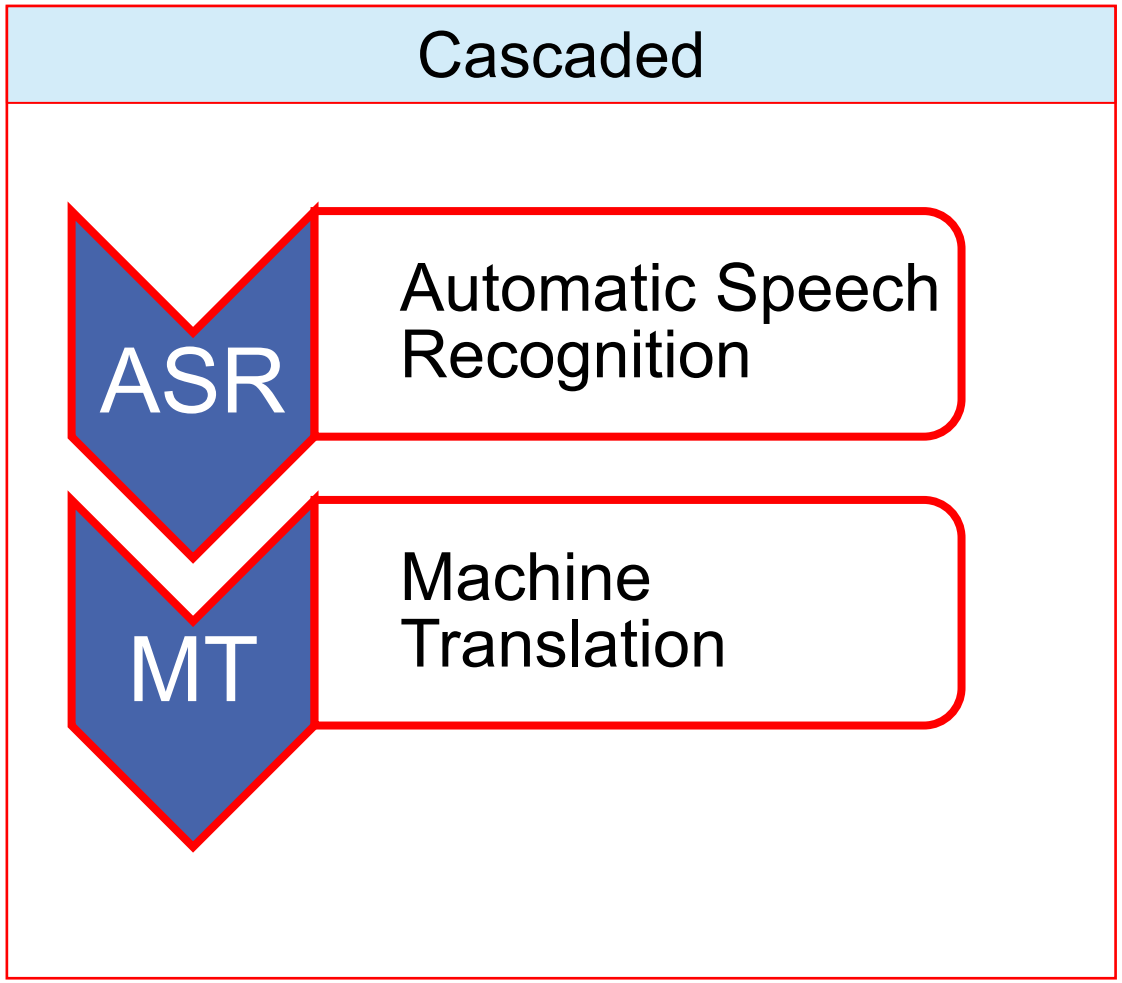


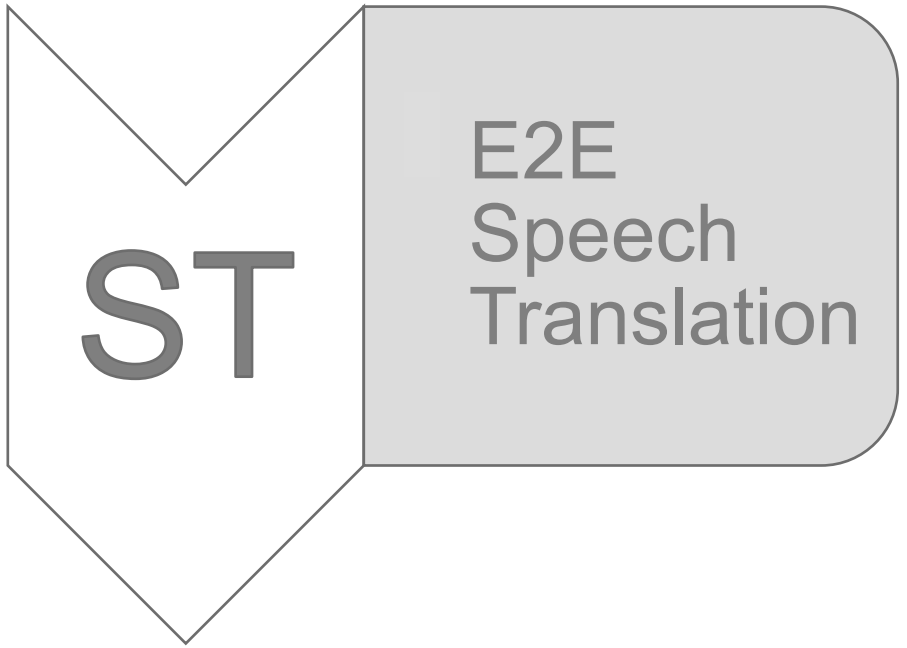


Introduction
Approaches: Pros and Cons

Cascaded		End To End	
+	-	+	-
Modularity	Integration complexity	Simplicity	High quality data
Data utilization	More engineering	Less engineering	Computational resource
Error analysis	Error propagation	Faster inference	Less interpretable
Flexible	Latency		

Introduction
Chosen Approach





Introduction
Cascaded Approach: Process

Introduction
Evaluation's Metrics: WER

- **W**ord **E**rror **R**ate (WER)
  - Formula:
  - The lower the better (↓)

$$Word\ Error\ Rate = \frac{(\boldsymbol{S}ubstitution + \boldsymbol{D}eletion + \boldsymbol{I}nsertion)}{Number\ of\ words\ in\ Reference}$$

Source: [4] Ali 2018

Introduction
Evaluation's Metrics: WER

- **W**ord **E**rror **R**ate (WER)
  - Formula:
  - The lower the better (↓)

$$Word\ Error\ Rate = \frac{(\boldsymbol{S}ubstitution + \boldsymbol{D}eletion + \boldsymbol{I}nsertion)}{Number\ of\ words\ in\ Reference}$$

- Limitations?

Reference
[I love KIT]

I like KIT = I hate KIT

Source: [4] Ali 2018

Introduction
Evaluation's Metrics: BLEU

- **BiL**ingual **E**valuation **U**ndersty Score (BLEU)
  - Multiple references
  - Individual N-Grams scores (BLEU-N)
  - Brevity penalty
  - The higher the better (↑)

Source: [5] Papineni 2002

Introduction
Evaluation's Metrics: BLEU

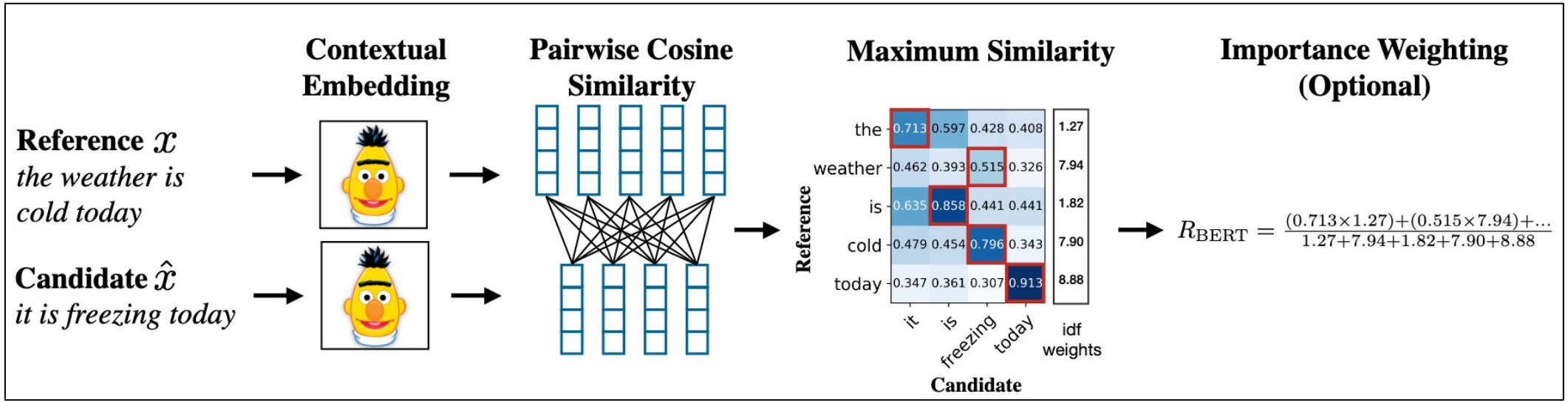
- **Bi**lingual **E**valuation **U**ndersty Score (BLEU)
  - Multiple references
  - Individual N-Grams scores (BLEU-N)
  - Brevity penalty
  - The higher the better (↑)

- Problem?
  - Lack of semantic evaluation
  - All errors equal
  - Tokenization sensitive

Source: [5] Papineni 2002

Introduction
Evaluation's Metrics: BertScore

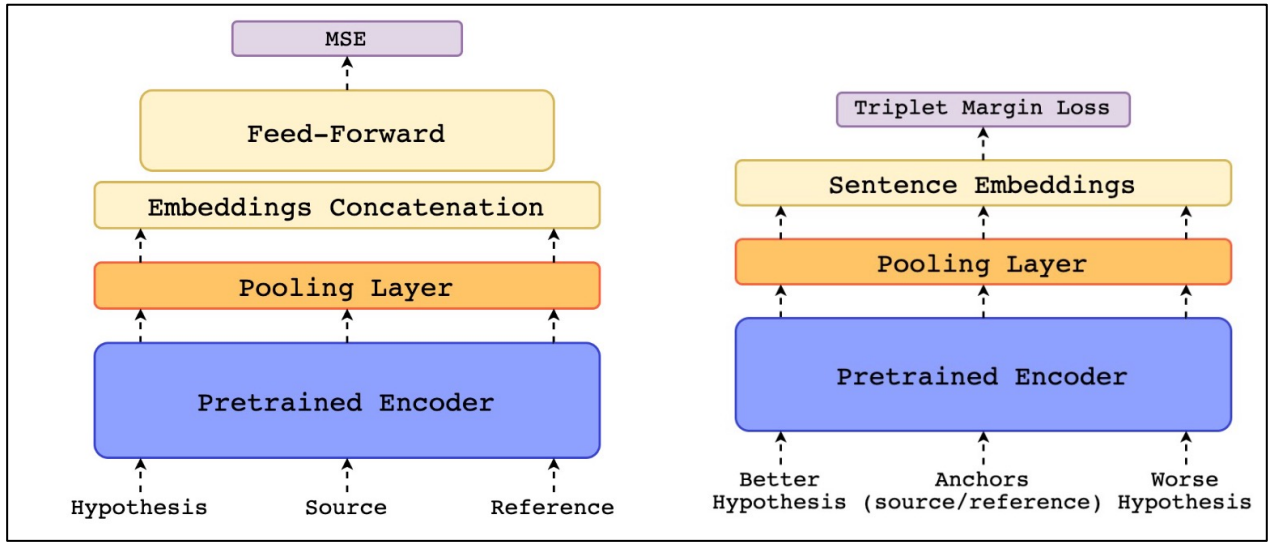
- BertScore
  - Contextualised
  - Importance weight
  - **P**recision, **R**ecall, **F1**
  - The higher the better (↑)



Source: [7] Zhang 2019

Introduction
Evaluation's Metrics: COMET

- **C**rosslingual **O**ptimized **M**etric for **E**valuation of **T**ranslation (COMET)
  - Contextualised
  - Specialised on MT tasks
  - Trained on real human judgements
  - The higher the better (↑)



Source: [7] Rei,2002

Experiments & Results
The Dataset

- **CoVoST 2:** A Massively Multilingual Speech-to-Text Translation Corpus
  - ASR/ST corpus
  - En→De
  - Used as the only training and test dataset

	Hours (CoVoST ext.)			Speakers (CoVoST ext.)			Src./Tgt. Tokens		
	Train	Dev	Test	Train	Dev	Test	Train	Dev	Test
	En→X								
De	364(430)	26(27)	25(472)	10K(10K)	4K(4K)	9K(29K)	3M/3M	156K/155K	4M/4M

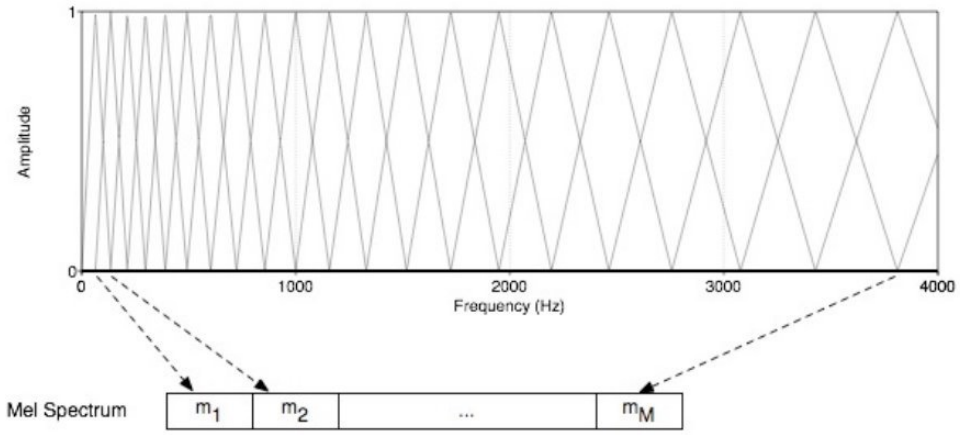
Source: [8] Wang 2020

Experiments & Results
Process: Audio Input

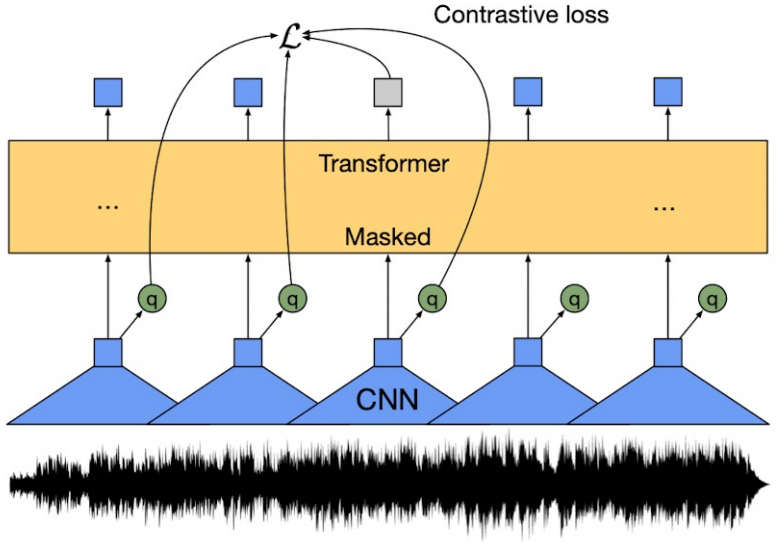
Experiments & Results
Audio Input: Choices

Karlsruher Institut für Technologie

Mel Spectograms



Wav2Vec



Source: [9] 2023

Experiments & Results
Audio Input: Choices

Mel Spectograms		Wav2Vec	
Considerations	Mel WER 21.1 @ 6h ASR training	Considerations	WER 149 @ 6h ASR training
Pre-processing done within hours		WER 7 @120h ASR training	Pre-processing requires > 1 day

Experiments & Results
Audio Input: Chosen One

Mel Spectograms

Wav2Vec

It's faster!

Mel WER 21.1 @6h ASR training

Pre-processing done within hours

Performance good enough!

Experiments & Results

Process: Automatic Speech Recognition (ASR)

Experiments & Results
ASR: Architecture

- Initial
  - S2T transformer small
    - 31 M parameters
- Improvement
  - Convolution-augmented transformer (S2T conformer)
    - Combine strengths of CNNs and transformer models
    - 16x encoder layers, 6x decoder layers
    - Convolutional downsampler
    - Relative positional encoding
      - 42.9 M parameters

Source: [10] Popuri 2022 and [11] Zhou 2019

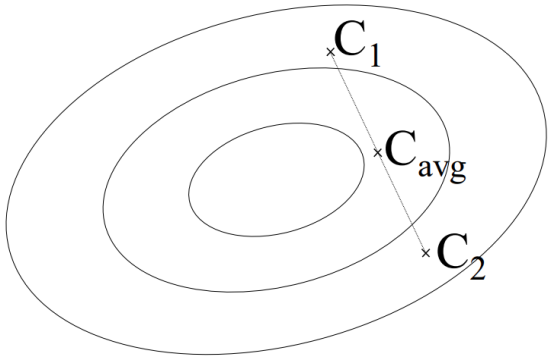
Experiments & Results
ASR: Hyperparamters

- Criterion: Cross entropy with label smoothing
- Optimizer: Adam (b1 = 0.9, b2 = 0.999)
- Learning rate: 0.002
- Learning rate scheduler: Cosine

Experiments & Results
Inference: Checkpoint Averaging

- What is it?
  - Combines multiple model states (checkpoints)
  - Boost performance of ASR models
  - Low calculation effort

- Why it helps?
  - Improves generalisation
  - Offset unwanted fluctuations in weights



Source: [12] Gao 2022

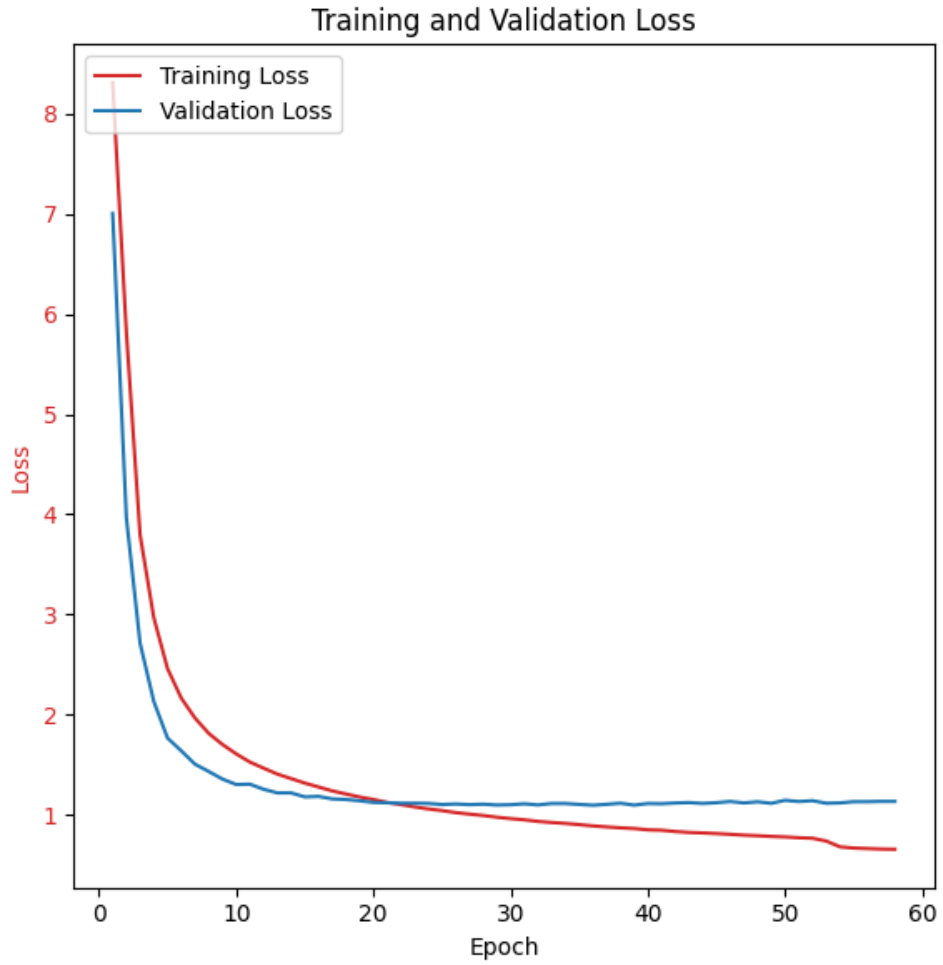
Experiments & Results

ASR: Results

Model	BLEU (↑)	WER (↓)
6h Train time	48.7	36.4
32h Train time	**55.7**	**30.3**
Baseline (CoVoST 2 Paper*)	-	**25.6**

*Baseline Architecture ASR:

- Vanilla Transformer encoder-decoder architecture.
- 12 encoder layers and 6 decoder layers.
- Convolutional downsampler



Experiments & Results
Process: Machine Translation (MT)

Experiments & Results

Machine Translation (MT) – Architecture & Hyperparmeters

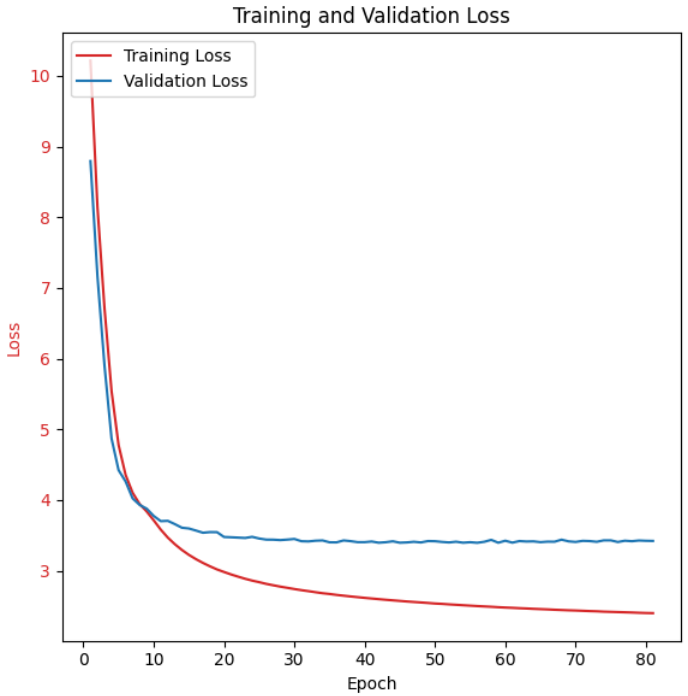
- Simple Transformer
  - Layers
    - 6x encoder
    - 6x decoder
  - Shared embedding layer between the decoder input and the output softmax layer
  - Regularization with 30 % dropout
  - 51.3 M parameters

- Hyperparameters:
  - Criterion: Cross entropy with label smoothing
  - Optimizer: Adam (b1 = 0.9, b2 = 0.98)
  - Learning rate: 0.001

Experiments & Results

Machine Translation (MT) – Problems

- Model initially trains very quickly but then does not continue to learn
- Assumption
  - This is probably because the dataset simply does not provide more information
- Approaches to fix this
  - Larger dataset (relatively uninteresting for this practical)
  - Some sort of data augmentation
    - Backtranslation
    - Paraphrasing



Experiments & Results

Machine Translation (MT) – Paraphrase Generation

Reason:

- CoVoST dataset only contained about 290k Samples which are all relatively short (~10 tokens)

Idea:

- Generate 5 extra paraphrases for each sentence in src and tgt
- Write out each combination as new training data
- Effectively increase the size about: $(5 + 1) * (5 + 1) = 36\,x$

Experiments & Results

Machine Translation (MT) – Paraphrase Generation

Prompt:

- Zero shot prompt lead to diffuse output with little coherency and difficulty to parse paraphrases from LLM output
- One shot prompt for consistent output to be able to extract
- Two shot prompt, even though it improved the output marginally, was even slower and therefore not applicable

Experiments & Results

Machine Translation (MT) – Paraphrase Generation – Prompt

- Shown in English, is analogous in German

Prompt

```
[INST] <<SYS>>
You are a professional writer, your expertise is in crafting accurate and engaging paraphrases.
<</SYS>>
Generate five distinct English paraphrases of the sentence provided below. Each paraphrase should fully
convey the original meaning without adding extraneous information. Aim for a balance between retaining the
essence of the sentence and presenting it in a fresh, clear manner.

Original Sentence: 'We know that, right? We've experienced that.'
[/INST]Paraphrases:
1. "That's something we understand, isn't it? It's been part of our experiences."
2. "We're aware of that, correct? We've gone through it ourselves."
3. "That's known to us, right? We have lived through that."
4. "We are cognizant of that, aren't we? We've felt that in our own lives."
5. "Isn't that familiar to us? We've personally encountered it."
</s><s>[INST]

Original Sentence: '{ORIGINAL_SENTENCE}'
[/INST]Paraphrases:
1.
```

Experiments & Results

Machine Translation (MT) – Paraphrase Generation

Extraction:

- Manual output cleanup based on desired output from prompt
- Neural comparison between paraphrase and reference and only accept paraphrase if not too different but also not the same

Experiments & Results

Machine Translation (MT) – Paraphrase Generation

Karlsruher Institut für Technologie

Resulted Dataset:

- 65k paraphrases generated in 24h
- Total paraphrases written: 225.000
  - Should have been 2.340.000
- New Dataset size 510.000 Sentences

Paraphrasing Samples

'I come here, and your wife is not here.'
Paraphrases:
1) I visit, but your wife is nowhere to be found.
2) When I arrive, your wife is not present.
3) I find that your wife is not with you when I come.
4) Upon my arrival, I discover that your wife is not available.
5) Your wife is not here at the moment.

'Your words are poetic and memorable.'
Paraphrases:
1) The way you express yourself is both poignant and lasting.
2) Your poetry is not only beautiful but also long-lasting.
3) Your words are melodic and leave a lasting impression.
4) Your language is artistic and enduring.
5) Your words are rhythmic and unforgettable.

Experiments & Results
Machine Translation (MT) – Paraphrase Generation

Problems:

- Paraphrase generation is slow: ~1s per paraphrase
- Probability tensor contains:

  $\inf,\ nan$ or $element < 0$
- Solution?
  - Increase temperature
- Unwanted side effects?
  - some paraphrases might change the sentences meaning

Paraphrasing Samples

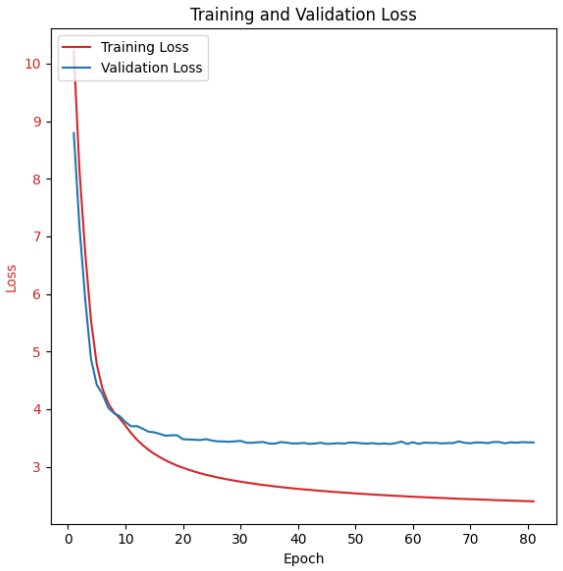
'Reductants other than hydrazine can be employed.'
Paraphrases:
1) Hydrazine is not the only option for reductants.
2) Depending on the situation, other reductants like... (list specific alternatives).
3) Alternative reducing agents besides hydrazine can be used.
4) In some cases, non-hydrazine reductants may be preferred.
5) Other reductants, such as... (list specific examples).

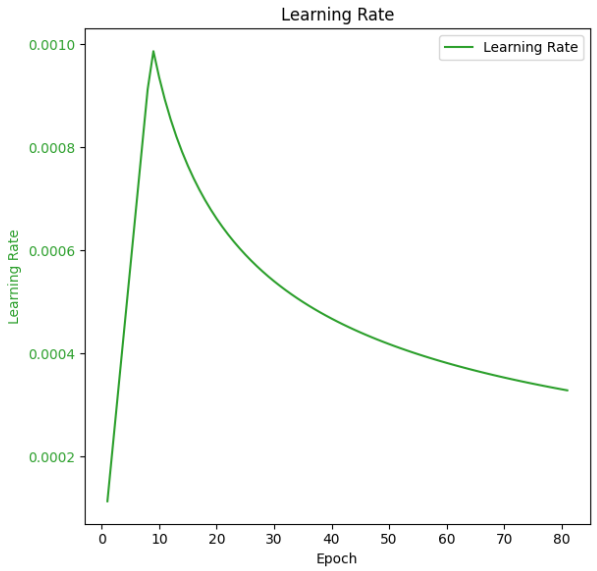
'That seemed to Celia almost the cruelest part of the whole tragedy.'
Paraphrases:
1) For Celia, the most painful element of the tragedy was...
2) The thought of it filled Celia with a sense of cruelty, as if...
3) Celia found it particularly cruel that...
4) In Celia's eyes, the most heinous part of the tragedy was...

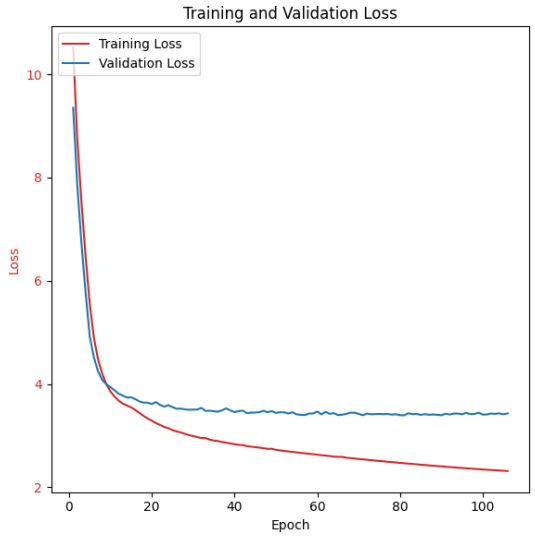
Experiments & Results

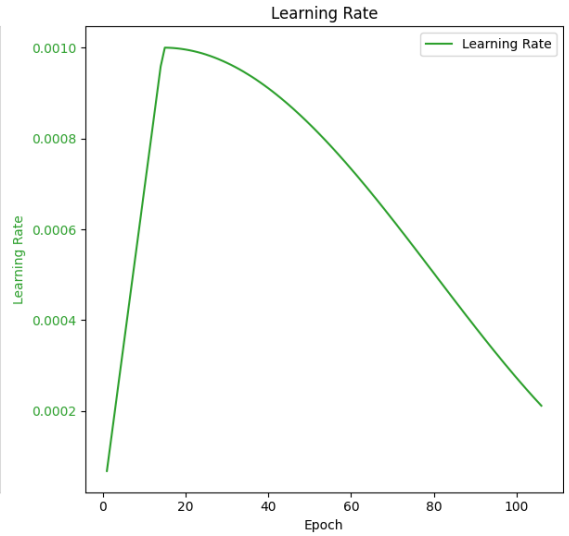
Machine Translation (MT) – Cosine Learning Rate (LR) Scheduler

- Possible reason: the learning rate decreased too quickly
- Longer training with higher LR might help → Cosine LR Scheduler









Experiments & Results

Machine Translation (MT) – Results

Variant	BLEU (↑)	Beam
Starting Point	28.8	64
Cosine LR Scheduler	**29.4**	64
Paraphrases	27.7	64
Paraphrases + Cosine LR Scheduler	27.4	64
Baseline (CoVoST 2 Paper)*	29.0	5

***Baseline Architecture MT:**

- Transformer base architecture (Vanilla).
- Dropout rate of 30%
- Use shared embeddings for both encoder/decoder inputs and decoder outputs to enhance performance.

Experiments & Results
Machine Translation (MT) – Discussion

Karlsruher Institut für Technologie

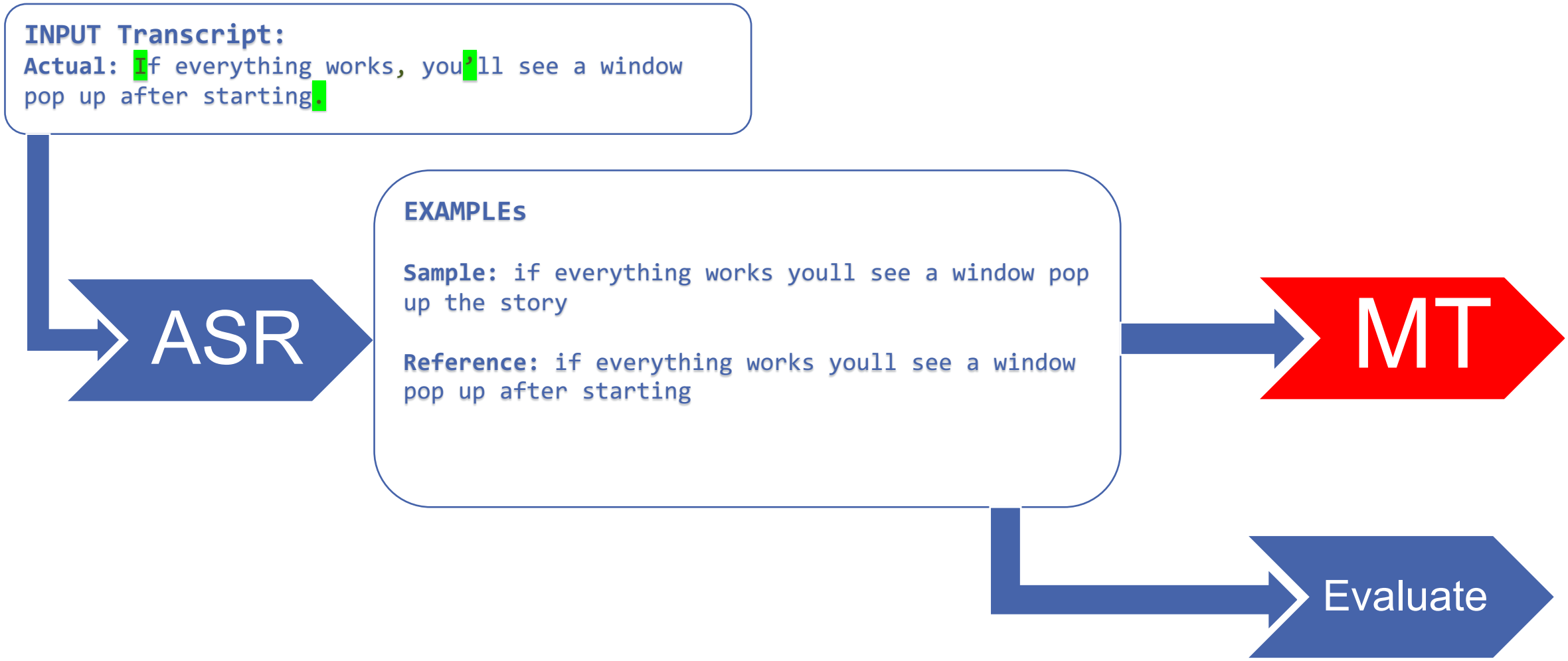
Why did Paraphrasing apparently not help in translation score improvements?

- Quality of paraphrases
- Semantic divergence
  - Subtle semantic shifts that aren't matched with correct translations
- Domain-specific language and context
  - Atypical phrases or usage patterns compared to CoVoST
- Data balance and representation
  - Overrepresenting certain structures or concepts from failed paraphrasing

Experiments & Results

Process: Connection (from ASR to MT)

Experiments & Results
Connection From ASR to MT – The Problem



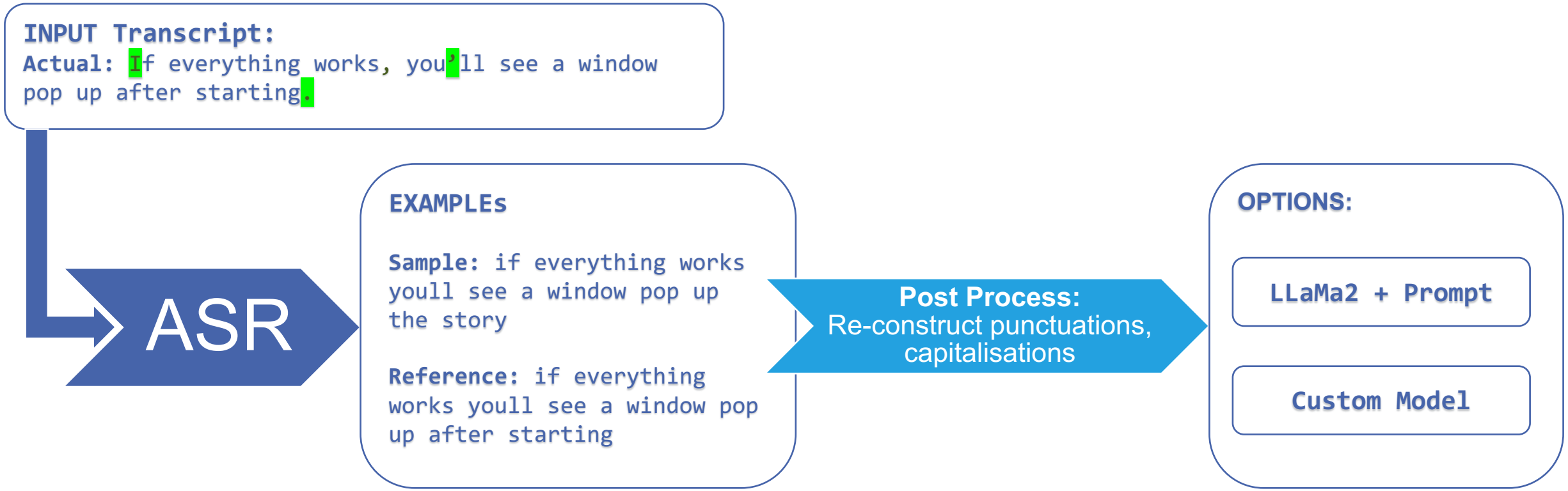
Experiments & Results
Connection From ASR to MT – Adjusting MT

- Train the MT System with the simplified input samples
  - Remove modularization
  - Tightly couple the systems

- Do the simplification step for all MT inputs
  - Removes a lot of information before translating the inputs

Experiments & Results

Connection From ASR to MT – Approaches



Experiments & Results

Connection From ASR to MT – Custom Model

Clean and Add Noise(
If everything works, you'll see a window pop up after starting.
)
= INPUT:
if everything quirks youll see a window pop up after darting

Custom Model: Transformer

OUTPUT:
If everything works, you'll see a window pop up after starting.

Experiments & Results
Connection From ASR to MT – Custom Model

Karlsruher Institut für Technologie

```
Clean and Add Noise(
      If everything works, you'll see a window pop
      up after starting.
)
= INPUT:
      if everything quirks youll see a window pop
up after darting
```

Custom Model: Transformer

```
OUTPUT:
If everything works, you'll see a window pop up after starting.
```

Fast Training

BLEU Score: 85 @ 6h

≈ 190 Sentences/Second

Experiments & Results

Connection From ASR to MT – LLaMa2 + Prompt

```
PROMPT(
        INPUT:
        if all works youll see a window pop up after
        starting
)
```

Slower Inference

≈ 1.7 Sentences / Second

LLaMa2

```
OUTPUT:
If everything works, you'll see a window pop up after starting.
```

Experiments & Results

Connection From ASR to MT – LLaMa2 + Prompt – Prompt

Prompt

<<SYS>>
You are a professional specialized in ASR (Automatic Speech Recognition) transcription enhancement.
<</SYS>>
[INST]Task:
1. Punctuation Restoration: Add necessary punctuation to make the sentences grammatically correct and more readable.
2. Evaluate the hypotheses for accuracy and likelihood of being the correct representation of the spoken text.
3. Select the best hypothesis that accurately reflects the original spoken text.

Input list of the 10 best ASR hypotheses:
1. "mr jones said meet me at 10 am in the conference room"
2. "mister jones said meet me at ten a m in the conference room"
3. "mr jones said meet me at ten am in the conference room"
4. "mister jones said meet me at 10 am in the conference room"
5. "mr jones said meet me at ten in the morning in the conference room"
6. "mr jones said meet me at 10 in the conference room"
7. "mister jones said meet me at ten in the conference room"
8. "mr jones said meet me at 10 in the conference room"
9. "mr jones said meet me at ten a m in the conference room"
10. "mister jones said meet me at 10 a m in the conference room"[/INST]
Best Hypothesis: "Mr. Jones said, 'Meet me at 10 a.m. in the conference room.'"

</s><s>[INST]
Input list of the 10 best ASR hypotheses:
{HYPOTHESES}[/INST]
Best Hypothesis:

Source: [14] Boraks 2023

Experiments & Results
Process: Evaluation

Experiments & Results

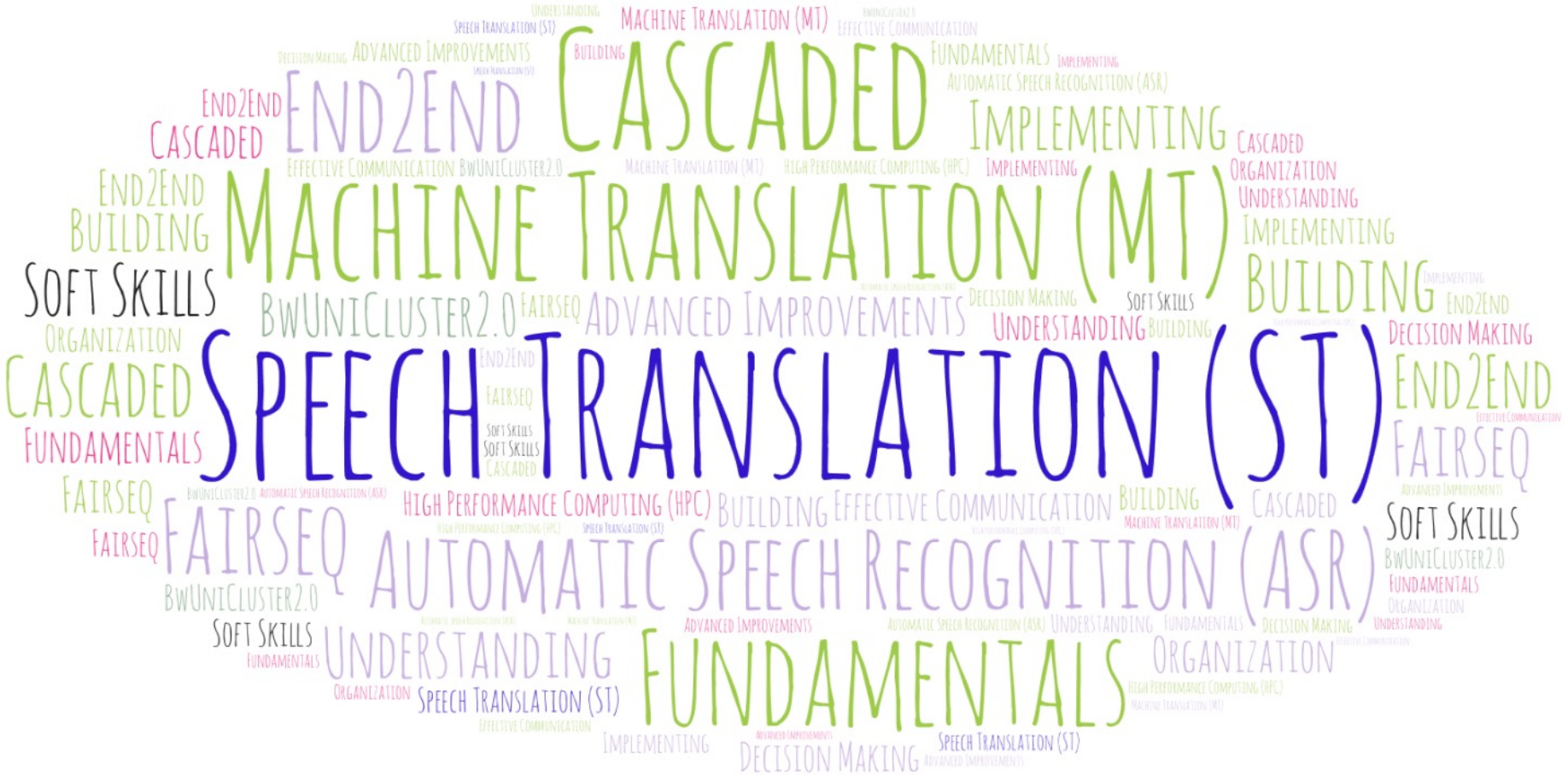
Evaluation – Ablation

Variant	Sub Set	BLEU (↑)	COMET (↑)	BertScore (↑)		
				Precision	Recall	F1
No Postediting	256	10.6	56.4	0.78	0.77	0.77
Llama2+Prompt	256	10.6	**57.8**	0.79	0.78	0.79
Custom Postediting	256	**12.6**	57.5	**0.80**	**0.79**	**0.80**
No Postediting	all	12.5	58.9	0.78	0.78	0.78
Custom Postediting	all	**14.1**	**59.8**	**0.79**	**0.79**	**0.79**
Baseline (CoVoST 2 Paper)*	all	**18.3**	-	-	-	-

* Baseline architecture identical to baseline ASR/MT architectures

Postediting does increase the ST Score as described on earlier slides

Conclusions
What did we learn?



Conclusions
Challenges

- BW Uni Cluster 2.0
  - High queue times up to 3 days

Slurm Job_id=23101713 Name=finetune_asr_covost Began, Queued time 3-00:34:05

Slurm Job_id=23113156 Name=generate_paraphrases Began, Queued time 2-07:33:33

Conclusions
Further Improvements

More data & Bigger model

Better filtering/lower temperature for paraphrases

Larger LLM for paraphrase generation

Backtranslation for data augmentation instead of paraphrasing

ASR data augmentation

E2E training with data augmentation


ممنون
you!
Thank
خیلی
schön!
Danke


Fragen?
سوال؟
Questions?
Any

References

- [1] Stephen R. Anderson. "How many languages are there in the world?", https://www.linguisticsociety.org/content/how-many-languages-are-there-world, last accessed: February 11, 2024, (2024).
- [2] "How many languages do Europeans speak?", https://examenexam.com/in/en/blog/how-many-languages-do-europeans-speak, last accessed: February 11, 2024, (2020).
- [3] Holly Young, "A language family tree – in pictures", https://www.theguardian.com/education/gallery/2015/jan/23/a-language-family-tree-in-pictures, last accessed: February 11, 2024, (2015).
- [4] Ahmed Ali and Steve Renals. 2018. Word error rate estimation for speech recognition: e-wer. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 20–24.
- [5] Kishore Papineni, Salim Roukos, Todd Ward, and Wei Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting of the Association for Computational Linguistics, pages 311–318.
- [6] Snover, Matthew, et al. "A study of translation edit rate with targeted human annotation." *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*. 2006.
- [7] Zhang, Tianyi, et al. "Bertscore: Evaluating text generation with bert." arXiv preprint arXiv:1904.09675 (2019).
- [8] Wang, Changhan, Anne Wu, and Juan Pino. "Covost 2 and massively multilingual speech-to-text translation." *arXiv preprint arXiv:2007.10310* (2020).
- [9] Praktikum Speech Translation (2023)
- [10] Sravya Popuri, "S2T Example: ST on CoVoST", https://github.com/facebookresearch/fairseq/blob/main/examples/speech_to_text/docs/covost_example.md, last accessed: February 11, 2024, (2022).
- [11] Zhou, Pan, et al. "Improving generalization of transformer for speech recognition with parallel schedule sampling and relative positional embedding." arXiv preprint arXiv:1911.00203 (2019).
- [12] Gao, Yingbo, et al. "Revisiting checkpoint averaging for neural machine translation." arXiv preprint arXiv:2210.11803 (2022).
- [13] covost, https://github.com/facebookresearch/covost, last accessed: February 11, 2024, (2020).
- [14] Eliran Boraks, "Llama 2 Prompt Engineering — Extracting Information From Articles Examples, https://medium.com/@eboraks/llama-2-prompt-engineering-extracting-information-from-articles-examples-45158ff9bd23, last accessed: February 12, 2024, (2023).
- [15] Bertil Braun, "llama_outputs.txt", https://raw.githubusercontent.com/BertilBraun/Advanced-Improvement-in-Speech-Translation/master/src/logs/st_eval/llama_outputs.txt, last accessed: February 12, 2024, (2024).
- [16] Rei, Ricardo, et al. "COMET: A neural framework for MT evaluation." *arXiv preprint arXiv:2009.09025* (2020).