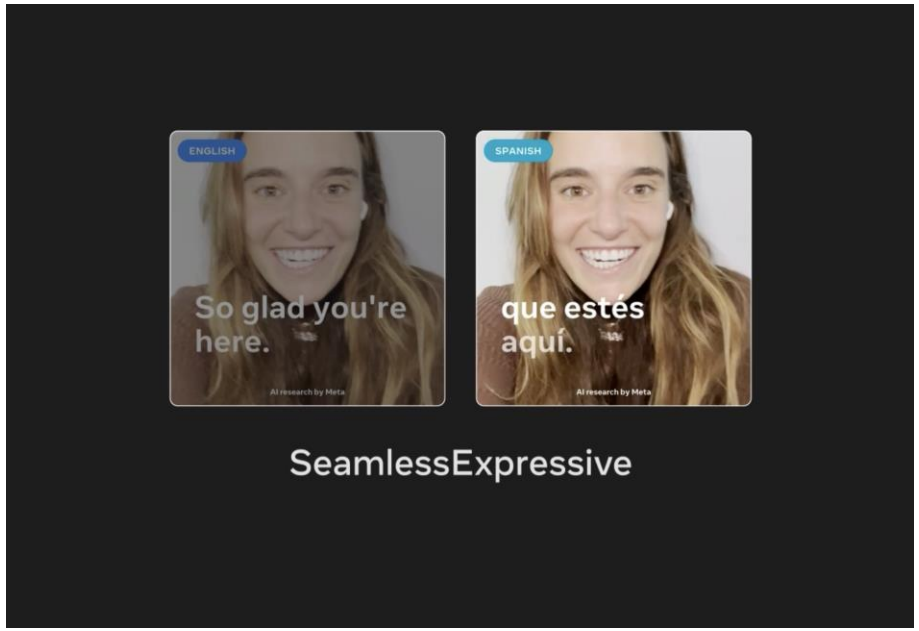
Spoken language translation

Tanel Alumäe

Use cases

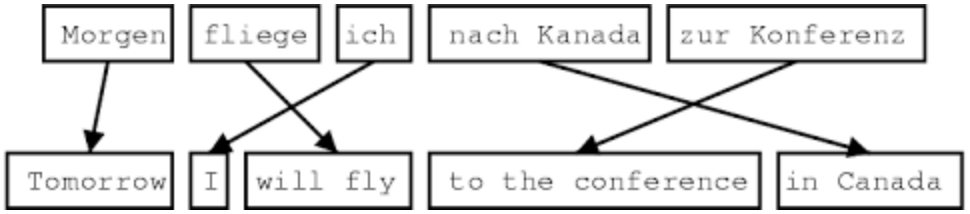
- Speech-to-text translation (e.g. YouTube does this)
- Subtitling
  - Speech translation + compression, optimizing line breaks, applying length constraints
- Speech-to-speech translation
  - Easiest: use a fixed voice for output
  - More difficult: use a voice that is similar to the source voice (even in multi-speaker recordings)
  - Most difficult: retain the expressivity (emotions, stress, prosody) of the source voice
- Even more difficult: do all this in realtime



Text-to-text translation: history

- Statistical machine translation
  - Find **parallel text corpus**
  - Align words (or phrases) with each other
    - Can be done automatically, if corpus is big enough
  - Extract translation units, with scores
    - "fliege -> "will fly" 0.4
    - "fliege" -> "am flying" 0.3
  - Translation:
    - Decompose source sentence into meaningful units (words or phrases)
    - Map units to all possible output units
    - Find the best output sentence, by combining translation scores with language model scores

English	Japanese
look, i don't do that shit anymore.	私は卒業した
thank you! you're so sweet	ありがとう
look, his name is cyrus gold.	いいか 彼の名前はサイラス・ゴールド
is that so? i hate to disappoint you.	そうか それは残念だったな。

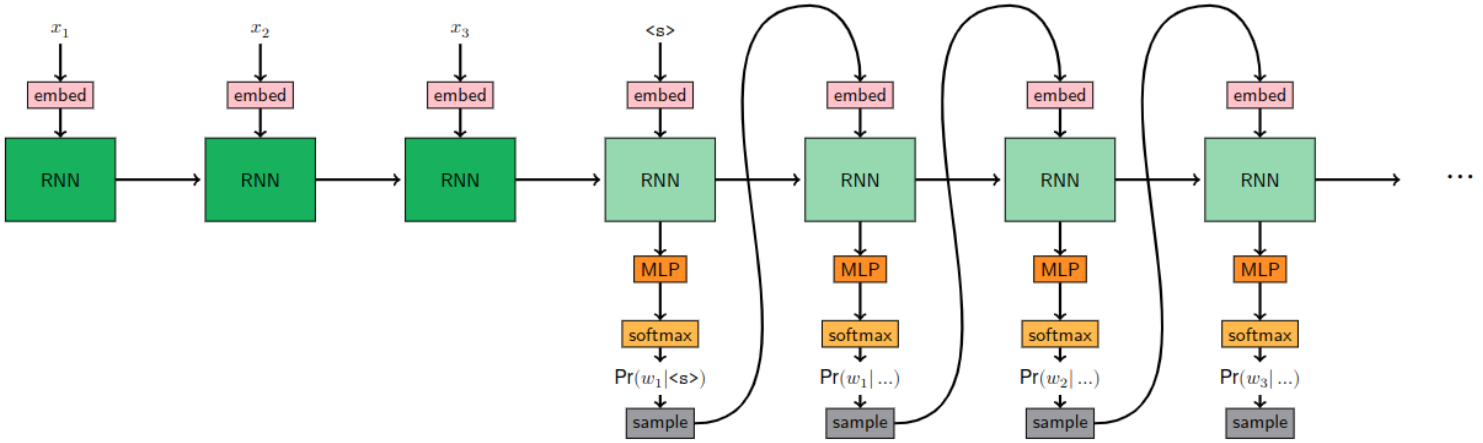


Text-to-text translation: statistical model

- Translating "v*äljas sajab*" -> "*it is raining outside*" (simplified)
  - Deconstruct: "väljas", "sajab"
    - "väljas" -> "outside" 0.7
    - "väljas" -> "in the field" 0.3
    - "sajab" -> "raining" 0.5
    - "sajab" -> "is raining" 0.3
    - "sajab" -> "it is raining" 0.2
  - Create all different **combinations** and **permutations**, compute translation and language model scores
    - "Väljas" -> "outside", "sajab" -> "raining"
      - outside raining $P_{translation}$ = 0.7*0.5 P_{LM}=0.003
      - raining outside $P_{translation}$ = 0.7*0.5 P_{LM}=0.005
    - "Väljas" -> "outside", "sajab" -> "is raining"
      - outside is raining
      - is raining outside
    - …
  - Select the output whose combined translation and language model score is the best

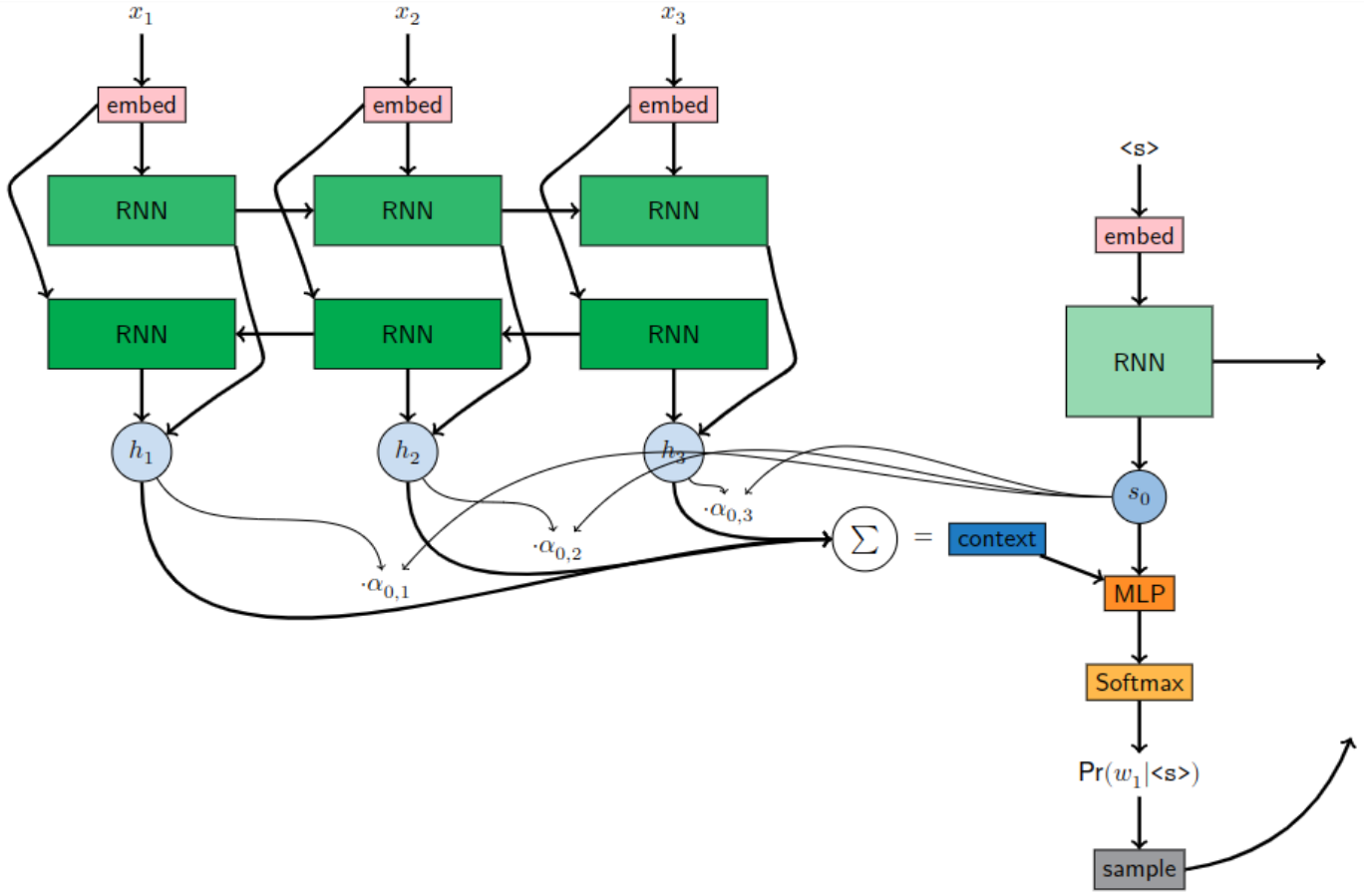
Neural machine translation (NMT)

- Sequence-to-sequence
- Has to compress the full source language meaning to a fixed-size vector that is fed from encoder to decoder



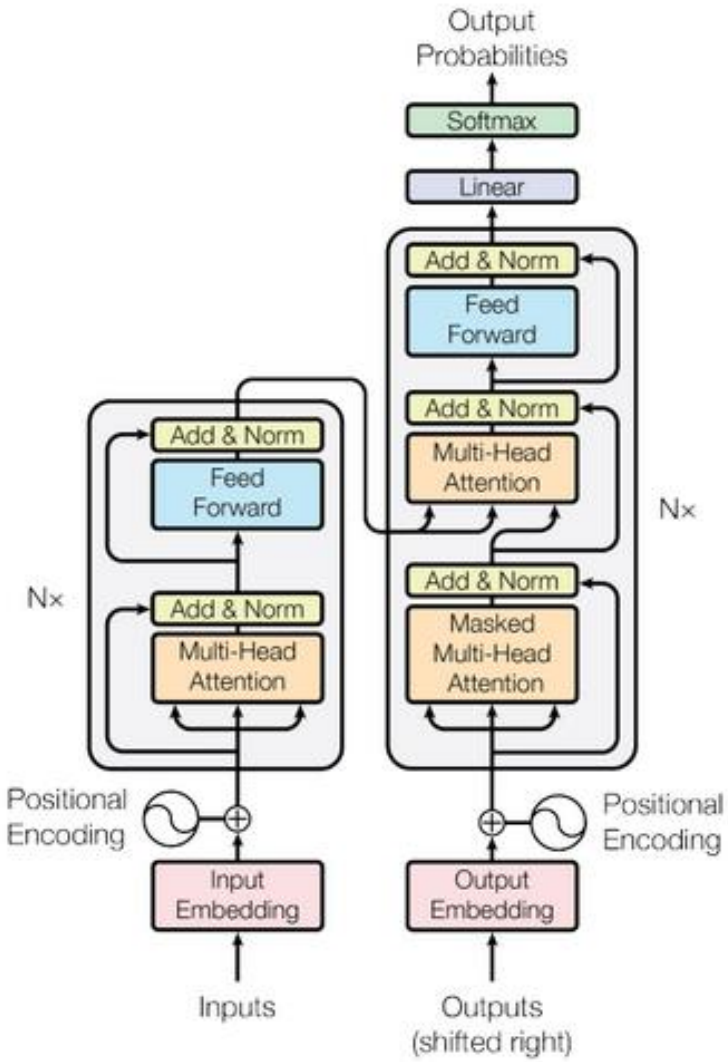
Seq-to-seq with attention

- Encoder: bidirectional RNN
- Decoder state s_0 is used to compute the query vector
- Query vector is used to compute softmax over encoder states
- Weighted average of encoder states: context vector
- Decoder state and context vector are concatenated
- The results is fed through a fully connected layer + softmax, to get the next output word



Transformer

- As in most NLP tasks, the ruling method in MT is currently Transformer (enoder-decoder)
- Decoder-only architectures (i.e., LLMs) are catching up
  - "*Translate to English: Väljas sajab vihma*" -> "*It is raining outside*"



Some tricks: pretraining

- Parallel translation data is usually scarce
- Pretraining helps, e.g. mBART
  - mBART is first pretrained on a denoising task, using multilingual data
    - Each denoising sample is monolingual
  - Pretrained model is then finetuned on translation data
- Gives a nice improvement on low and medium resource language pairs

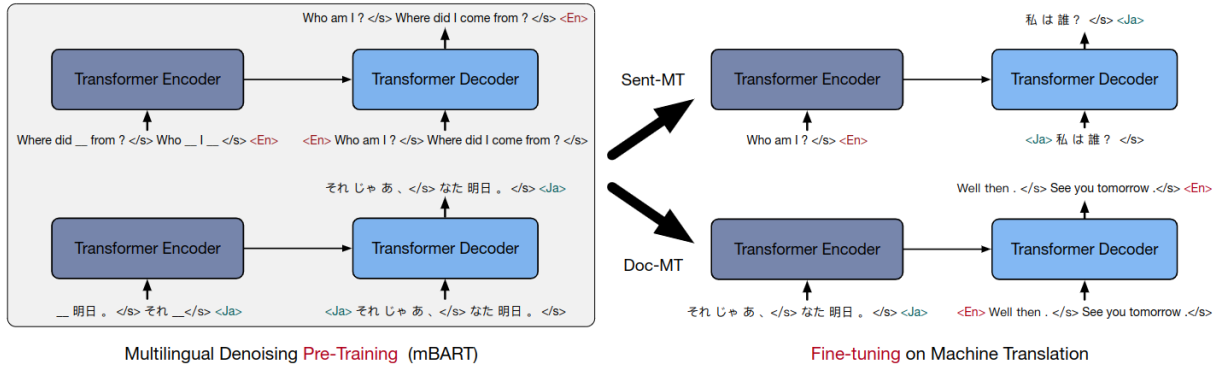


Figure 1: Framework for our Multilingual Denoising Pre-training (left) and fine-tuning on downstream MT tasks (right), where we use (1) sentence permutation (2) word-span masking as the injected noise. A special language id token is added at both the encoder and decoder. One multilingual pre-trained model is used for all tasks.

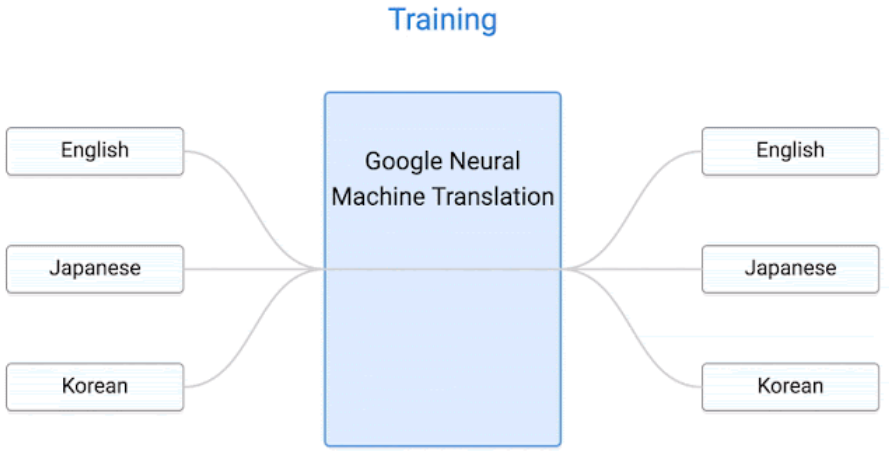
Languages	En-Si		En-Hi		En-Et		En-Lt		En-Fi		En-Lv	
Data Source	FLoRes		ITTB		WMT18		WMT19		WMT17		WMT17	
Size	647K		1.56M		1.94M		2.11M		2.66M		4.50M	
Direction	←	→	←	→	←	→	←	→	←	→	←	→
Random	7.2	1.2	10.9	14.2	22.6	17.9	18.1	12.1	21.8	20.2	15.6	12.9
mBART25	**13.7**	**3.3**	**23.5**	**20.8**	**27.8**	**21.4**	**22.4**	**15.3**	**28.5**	**22.4**	**19.3**	**15.9**

Some tricks: backtranslation

- Parallel texts are hard to find
- Back-translation (BT) is a method to exploit monolingual data for high quality MT
- BT assumes that there is abundant parallel data in the target language
- Steps to improve Estonian->English MT system
  - Train initial MT model on existing parallel data, but in the opposite direction (i.e., English->Estonian)
  - Use it to generate more parallel data
  - Now train Estonian->English on original + synthetic data
  - This can be repeated several times
- Uses the fact that we target side data is more important for translation quality, and therefore uses authentic data for target size
- The synthetic source side (created by MT) might be flawed, ungrammatical to some extent, but hopefully close to human language

Multilingual MT

- With encoder-decoder models, it is easy to build multilingual MT systems
  - E.g., just use a target language tag is the 1st token in the decoder
- Such models can perform zero-shot translation on language pairs whose parallel data is not present in the training corpus
  - e.g. Korean-to-Japanese was not present in training data, but models learned it from other language pairs
- This enables massively multilingual models (e.g. with 200 languages)
- However, zero-shot quality is usually not very good (but better than nothing)



Evaluation

- Evaluation is very important for MT, but also very complicated
  - Every source sentence can have several "perfect" translations
  - What to evaluate?
    - Readability? Work needed to post-edit, in order to create publishable translation?
- Manual evaluation
  - What to show annotators when assessing a translation candidate?
    - Ref-based: only the (human) reference
    - Src-based: only the source sentence
    - Src&ref-based: both
  - Context to consider:
    - Sentence-level: sentences of a document in random order
    - Document-level: obtain a single score per document
    - Document-aware: show whole document, asses per sentence (assumes sentence-per-sentence translation)
  - What to ask:
    - Some relative score over several candidates
    - Some absolute score for a single candidate
    - A more complicated question?

Some problem with manual annotation

- Always time-consuming and expensive
- Relative ranking of candidates: Longer sentences hard to rank. Candidates incomparably poor
- Document-level manual assessment very hard, mental overload

Below are the sentences you have just rated as a single **document**. Please state how much you agree that:

The black text adequately expresses the meaning of the gray text in German (deutsch).

Russian Grand Prix: Lewis Hamilton closes in on world title after team orders hand him win over Sebastian Vettel It became clear from the moment that Valtteri Bottas qualified ahead of Lewis Hamilton on Saturday that Mercedes" team orders would play a large part in the race. From pole, Bottas got a good start and almost hung Hamilton out to dry as he defended his position in the first two turns and invited Vettel to attack his teammate. Vettel went into the pits first and left Hamilton to run into the traffic at the tail of the pack, something which should have been decisive. The Mercedes pitted a lap later and came out behind Vettel, but Hamilton went ahead after some wheel-to-wheel action that saw the Ferrari driver reluctantly leave the inside free at risk of holding out after a double-move to defend on the third corner. Max Verstappen started from the back row of the grid and was in seventh by the end of the first lap on his 21st birthday. He then led for a large part of the race as he held onto his tyres to target a quick finish and overtake Kimi Raikkonen for fourth. He eventually came into the pits on the 44th lap but was unable to increase his pace in the remaining eight laps as Raikkonen took fourth. It's a difficult day because Valtteri did a fantastic job all weekend and was a real gentleman told let me by. The team have done such an exceptional job to have a one two," said Hamilton.

— Source text

Großer Preis von Russland: Lewis Hamilton schließt auf Weltmeistertitel ein, nachdem ihm das Team den Sieg über Sebastian Vettel überlassen hat Es wurde von dem Moment an klar, dass Valtteri Bottas sich vor Lewis Hamilton am Samstag qualifiziert hatte, dass die Teamaufträge von Mercedes eine große Rolle im Rennen spielen würden. Von der Pole aus erwischte Bottas einen guten Start und ließ Hamilton fast trocken, als er seine Position in den ersten beiden Kurven verteidigte und Vettel einlud, seinen Teamkollegen anzugreifen. Vettel ging zuerst in die Gruben und verließ Hamilton, um am Rucksack in den Verkehr zu geraten, was entscheidend gewesen sein sollte. Der Mercedes drehte eine Runde später und kam hinter Vettel, aber Hamilton ging nach einigen Rad-an-Rad-Aktion, die sah, dass der Ferrari-Fahrer widerwillig verlassen die Innenseite frei in Gefahr zu halten, nach einem Doppelschlag auf der dritten Ecke zu verteidigen. Max Verstappen startete aus der hinteren Startreihe und wurde am Ende der ersten Runde an seinem 21. Geburtstag Siebter. Er führte dann für einen großen Teil des Rennens, als er auf seinen Reifen hielt, um ein schnelles Ziel zu erreichen und Kimi Räikkönen zum vierten Mal zu überholen. In der 44. Runde kam er schließlich in die Box, konnte aber sein Tempo in den verbleibenden acht Runden nicht erhöhen, da Räikkönen den vierten Platz belegte. Es ist ein schwieriger Tag, denn Valtteri hat das ganze Wochenende einen fantastischen Job gemacht und war ein echter Gentleman, der mir gesagt hat. Das Team hat so einen außergewöhnlichen Job gemacht, um ein, zwei zu haben", sagte Hamilton.

— Candidate translation

Manual relative ranking: example

Defying the shadows, Anto descends the crater and lights the path with a small torch attached to the helmet he bought with his money.

I přes okolní tmu fárá Anto do kráteru a osvětluje si cestu malou svítilnou, kterou má připevněnou na helmě a sám si ji za své peníze koupil.

Best ← Rank 1 ○ Rank 2 ○ Rank 3 ○ Rank 4 ○ Rank 5 ○ → Worst

Vzdoruje stínům Anto, sestupuje z kráteru a svítí cestu s malou pochodní připojenou k helmě, kterou koupil ze svých peněz.

Best ← Rank 1 ○ Rank 2 ○ Rank 3 ○ Rank 4 ○ Rank 5 ○ → Worst

Vzpírat se stínům, Anto sestupuje kráter a osvítí cestu malou baterkou spojenou s helmou, kterou on koupil s jeho penězi.

Best ← Rank 1 ○ Rank 2 ○ Rank 3 ○ Rank 4 ○ Rank 5 ○ → Worst

Odolává stíny, Anto snáší kráter a osvětlí cestu s malou pochodeň na helmu, koupil za své peníze.

Best ← Rank 1 ○ Rank 2 ○ Rank 3 ○ Rank 4 ○ Rank 5 ○ → Worst

Navzdory stínům anto, sestupuje z kráteru a svítí na cestu s malou pochodeň připevněnou na helmu, kterou si koupil ze svých peněz.

Best ← Rank 1 ○ Rank 2 ○ Rank 3 ○ Rank 4 ○ Rank 5 ○ → Worst

Popírání stínovým zpravodajům, Anto nezavládne se crater a svítidla cestu s malou pochodeň oddání helmu koupil s jeho peníze.

Submit

Quiz-based evaluation

- Given: text paragraph in source language
- Create 3 quiz questions about it in the target language
- Present human annotator:
  - Machine-translated text
  - Questions
- Score how well the quiz is performed

Automatic evaluation: BLEU

- Based on geometric mean of n-gram precision.
  - ≈ ratio of 1- to 4-grams of hypothesis confirmed by a ref. Translation

Src	The legislators hope that it will be approved in the next few days .	Confirmed
Ref	Zákonodárci doufají , že bude schválen v příštích několika dnech .	1 2 3 4
Moses	**Zákonodárci doufají , že bude schválen v** nejbližších dnech .	9 7 5 4
TectoMT	**Zákonodárci doufají , že bude** schváleno další páru volna .	6 4 3 2
Google	Zákonodárci naději **, že bude schválen v** několika příštích dnů .	9 4 3 2
PC Tr.	Zákonodárci doufají že to bude schválený v nejbližších dnech .	7 2 0 0

n-grams confirmed: none, unigram, bigram, *trigram*, **fourgram**

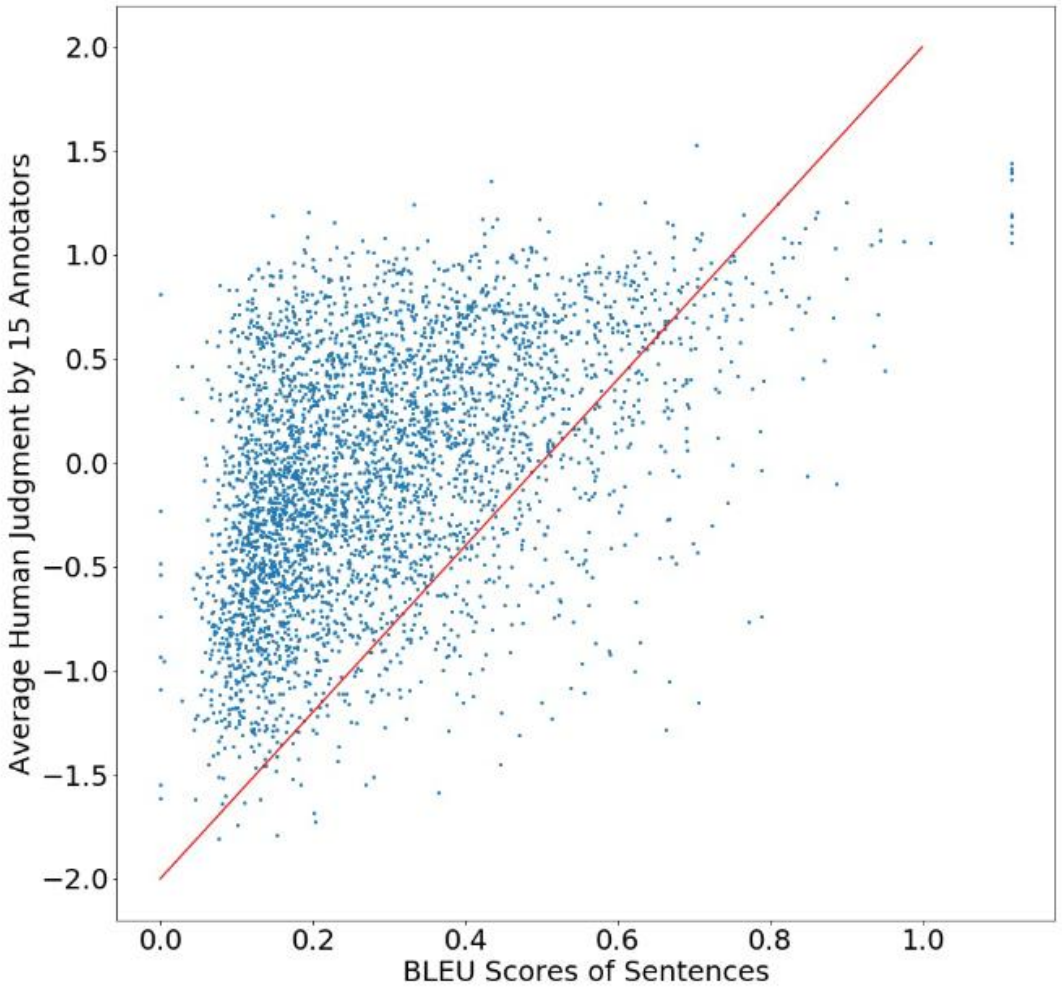
E.g. Moses produced 10 unigrams (9 confirmed), 9 bigrams (7 confirmed), ...

$$\text{BLEU} = \text{BP} \cdot \exp\left(\frac{1}{4}\log\left(\frac{9}{10}\right) + \frac{1}{4}\log\left(\frac{7}{9}\right) + \frac{1}{4}\log\left(\frac{5}{8}\right) + \frac{1}{4}\log\left(\frac{4}{7}\right)\right)$$

BP is "brevity penalty"; $\frac{1}{4}$ are uniform weights, the "denominator" equivalent for $\sqrt[4]{\cdot}$ in geometric mean in the log domain.

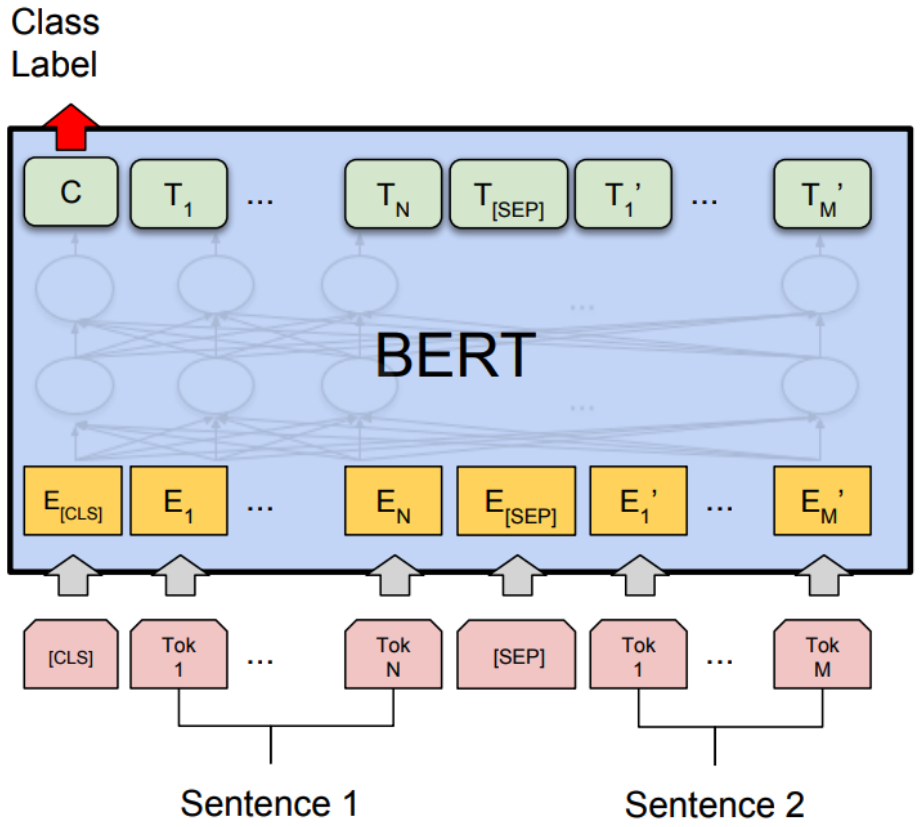
BLEU properties

- BLEU is in the range of 0..1 (often written using %, 0..100%)
- Human translation against other humans ~60%
- Google Chinese→English: ~30%, Arabic→English: ~50%
- BLEU for individual sentences not reliabl
- Not always correlated with human judgements
- BLEU scores are not comparable:
  - across languages
  - on different test sets
  - with different number of reference translations



Automatic evaluation: BLEURT

- BLEU doesn't cosider such things as synonyms at all
- Yet we know that one sentence can have many "perfect" translations
- BLEURT is a trainable multilingual metric
- Multilingual BERT, finetuned for quality estimation
- Input: reference and MT candidate, separated by the [SEP] token
- Output: [CLS] -> 0..1
- BLEURT goes through a few additional phases of training.
  - The “pre-training” phase incorporates synthetic text from Wikipedia, which is randomly perturbed to mimic variations in output
  - Finally trained on translation references + candidates and their human judgements (in several languages)



BLUERT: performance

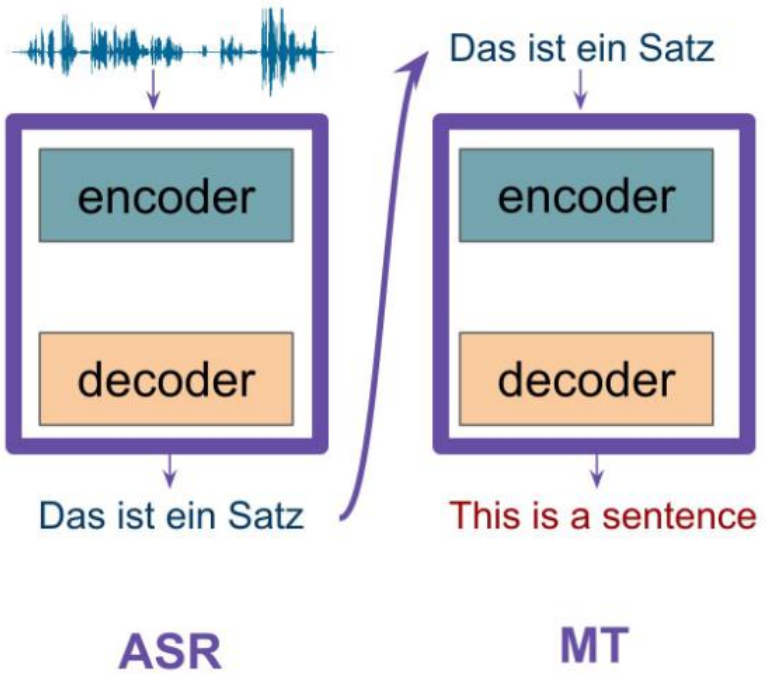
- BLUERT is better correlated with human judegments than BLEU and other metrics

model	cs-en τ / DA	de-en τ / DA	et-en τ / DA	fi-en τ / DA	ru-en τ / DA	tr-en τ / DA	zh-en τ / DA	**avg** τ / DA
sentBLEU	20.0 / 22.5	31.6 / 41.5	26.0 / 28.2	17.1 / 15.6	20.5 / 22.4	22.9 / 13.6	21.6 / 17.6	22.8 / 23.2
BERTscore w/ BERT	29.5 / 40.0	39.9 / 53.8	34.7 / 39.0	26.0 / 29.7	27.8 / 34.7	31.7 / 27.5	27.5 / 25.2	31.0 / 35.7
BERTscore w/ roBERTa	31.2 / 41.1	42.2 / 55.5	37.0 / 40.3	27.8 / 30.8	30.2 / 35.4	32.8 / 30.2	29.2 / 26.3	32.9 / 37.1
Meteor++	22.4 / 26.8	34.7 / 45.7	29.7 / 32.9	21.6 / 20.6	22.8 / 25.3	27.3 / 20.4	23.6 / 17.5*	26.0 / 27.0
RUSE	27.0 / 34.5	36.1 / 49.8	32.9 / 36.8	25.5 / 27.5	25.0 / 31.1	29.1 / 25.9	24.6 / 21.5*	28.6 / 32.4
YiSi1	23.5 / 31.7	35.5 / 48.8	30.2 / 35.1	21.5 / 23.1	23.3 / 30.0	26.8 / 23.4	23.1 / 20.9	26.3 / 30.4
YiSi1 SRL 18	23.3 / 31.5	34.3 / 48.3	29.8 / 34.5	21.2 / 23.7	22.6 / 30.6	26.1 / 23.3	22.9 / 20.7	25.7 / 30.4
BLEURTbase -pre	33.0 / 39.0	41.5 / 54.6	38.2 / 39.6	30.7 / 31.1	30.7 / 34.9	32.9 / 29.8	28.3 / 25.6	33.6 / 36.4
BLEURTbase	34.5 / **42.9**	43.5 / 55.6	39.2 / 40.5	31.5 / 30.9	31.0 / 35.7	35.0 / 29.4	29.6 / **26.9**	34.9 / 37.4
BLEURT -pre	34.5 / 42.1	42.7 / 55.4	39.2 / 40.6	31.4 / 31.6	31.4 / 34.2	33.4 / 29.3	28.9 / 25.6	34.5 / 37.0
BLEURT	**35.6** / 42.3	**44.2 / 56.7**	**40.0 / 41.4**	**32.1 / 32.5**	**31.9 / 36.0**	**35.5 / 31.5**	**29.7** / 26.0	**35.6 / 38.1**

Table 3: Agreement with human ratings on the WMT18 Metrics Shared Task. The metrics are Kendall Tau (τ) and WMT's Direct Assessment metrics divided by 100. The star * indicates results that are more than 0.2 percentage points away from the official WMT results (up to 0.4 percentage points away).

Speech translation: cascade

- Simple: first do speech recognition (ASR), then MT, using decoupled models
  - Usually also applies segmentation/diarization before ASR
- Good:
  - Training data availability
  - can swap out ASR or MT systems and get improvement
- Problem: error propagation
  - Speech recognition errors could be amplified by MT
  - (Partial) solutions:
    - Present MT system n-best hyps from ASR
    - Train MT on noisy source texts, to be robust against ASR errors
- Also: cannot really use use speech->text training data (such as speech + translated subtitles) for training



End-to-end Speech translation: challenges

- "Modeling burden"
  - A single model has to learn both meaningful speech representations (such as phonemes and words) and translation
  - Harder task than ASR, since in ASR there is usually a monothonic alignment between speech and text
- Data scarcity
  - Very few real speech translation corpora availble, while large ASR and MT datasets are more abundant
- Result: almost impossible to train "pure" E2E speech translation models from scratch only on speech translation data

Using synthetic data

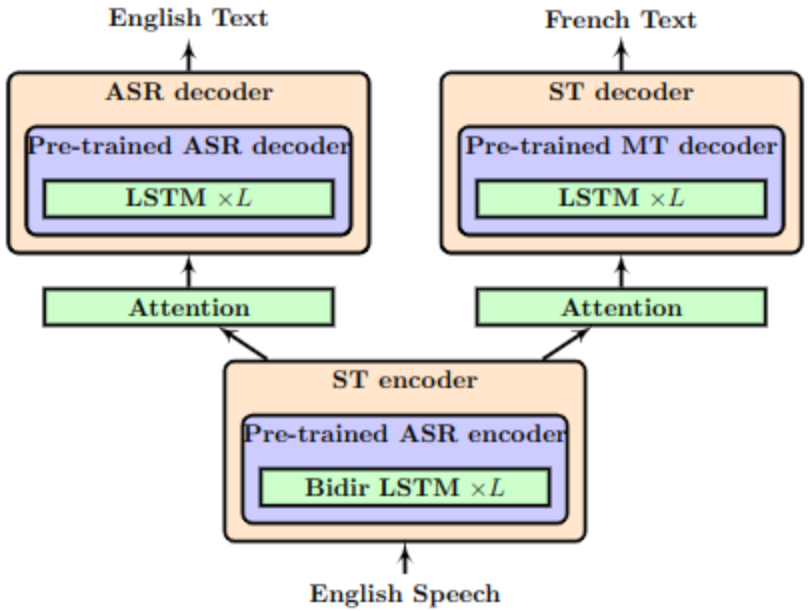
- ASR and MT data is often available
- Can create two types of synthetic data:
  - Use MT to translate ASR training data to target language
  - Use text-to-speech to synthesize audio for source language side ot MT training data
- Problems:
  - Using MT to translate ASR training data puts a "cap" on achievable speech translation quality. Also, cannot take into account prosody, stress, etc.
  - Synthesized speech may lack naturalness, spontaneosness and speaker variability

Pretraining and sharing parts of E2E model

- Most state-of-the art speech translation models make smart use of additonal data or models
- Pretraining: train parts of the model beforehand
- Multi-task learning: train a more complex model to perform more tasks (e.g., speech recognition or machine translation)

Multi-task learning: one-to-many

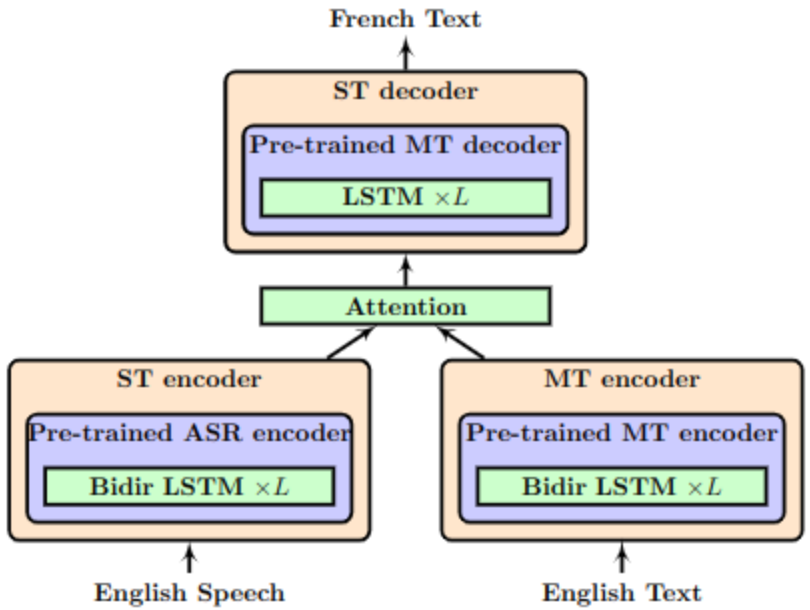
- Weis et al 2017
  - One encoder
    - Source language audio
  - Two decoders
    - Source language text
    - Target language text
  - Train on both AST and ST data
- If source language text is available for speech translation data, then we can use CTC to co-train encoder for doing ASR



(b) multitask-one2many

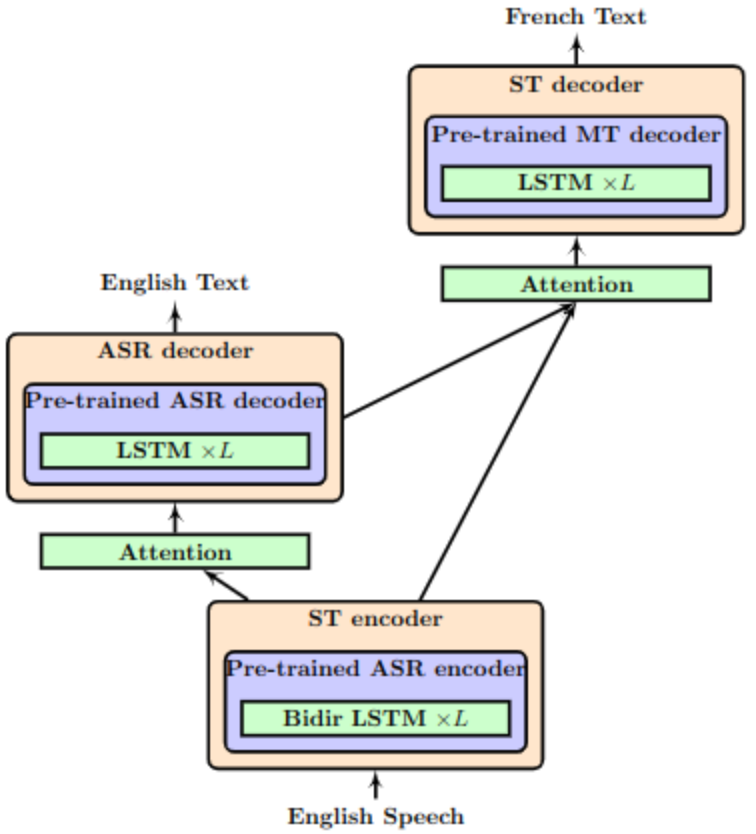
Multi-task learning: many-to-one

- Train on 2 tasks:
  - Text-to-text translation
  - Speech-to-text translation
- 2 encoders, 1 decoder
- Speech encoder applies so-called adaptors (typically just strided convolution) in the last layers to make the frame rate of speech comparable to token rate of text



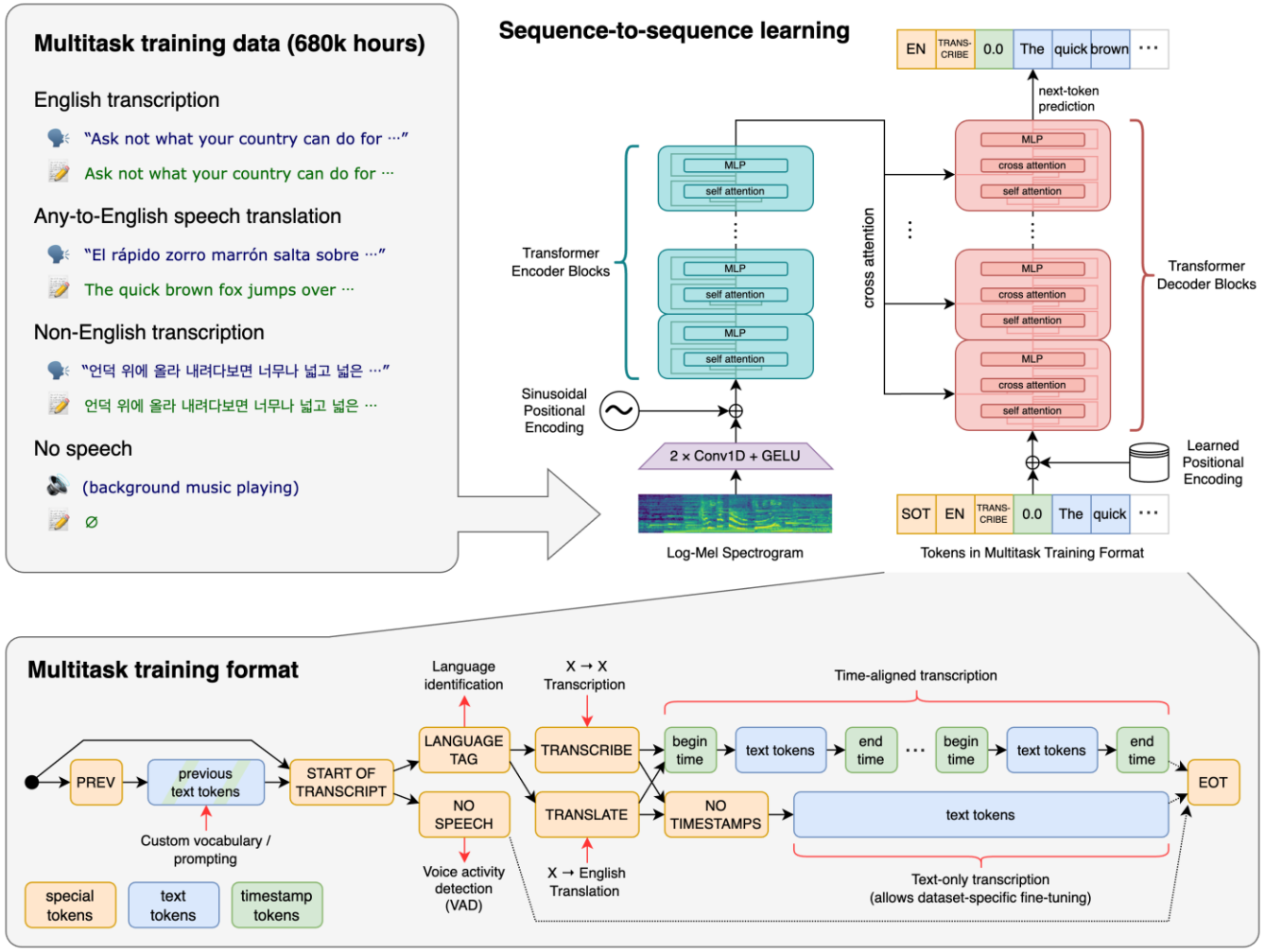
Multitask learning: triangle

- Speech recognition and speech translation are trained simultaneously
- Target language decoder attends over encoder outputs as well as decoder outputs
- Requires ST data with source language transcription



Multi-task learning: single encoder&decoder

- Like Whisper model
- Use a single encoder and decoder, but use a special token prefix to denote the task (transcribe or translate)
- Benefit over 2-decoder approach: decoder can benefit from ASR data in the target language
- E.g., to train Estonian->Latvian system, use the following data:
  - Estonian and Latvian ASR data
  - Estonian-to-Latvian speech translation data



Pretraining: using unlabelled data

- We usually have abundant unlabelled training data:
  - Source language speech
  - Target language text
- Let's pretrain two models and bridge them!
  - Wav2vec2.0 on source language speech
    - Or, just use a multilingual wav2vec2.0 -- might work as well and is readily available!
  - BART -- text-based encoder-decoder model trained to denoise artificially scrambled text
    - Or, just use the multingual mBART, unless the target language is very exotic
- If source-to-target language MT data is available, pre-finetune BART for text-based MT
- Length adaptor is used to reduce the output frame rate of speech encoder
- This approach is perfected by Meta's Seamless4MT model

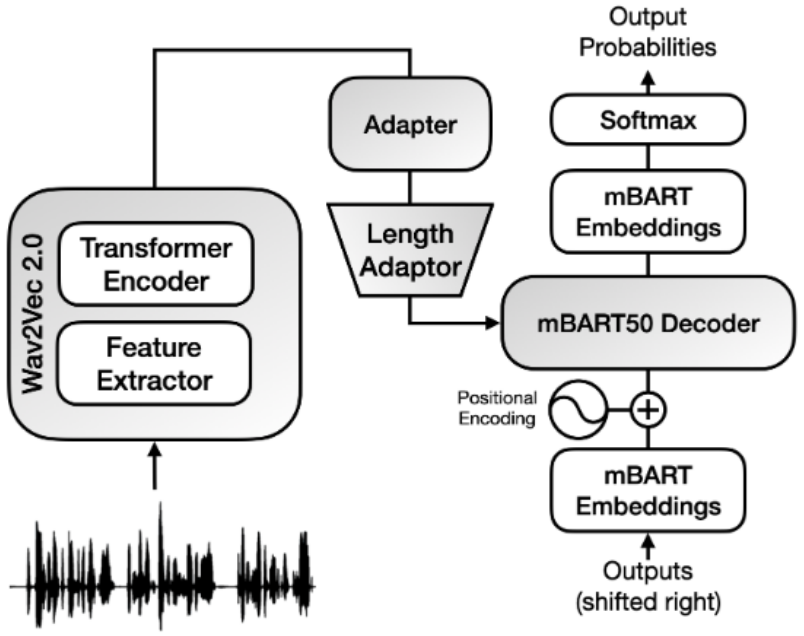


Figure 1: System overview. The original architecture proposed by Li et al. (2021) includes a pre-trained Wav2Vec 2.0 as the encoder, a pre-trained mBART decoder and a Length Adaptor. In this work, we add an Adapter module after the encoder.

Seamless4MT: more details

- Starts with 2 models:
  - w2v-BERT 2.0: improved wav2vec20 model, trained on 1 million hours of open speech audio data that covers over 143 languages
  - Seamless4MT-NLLB text-to-text translation model: massive transformer-based MT model trained on 200 languages

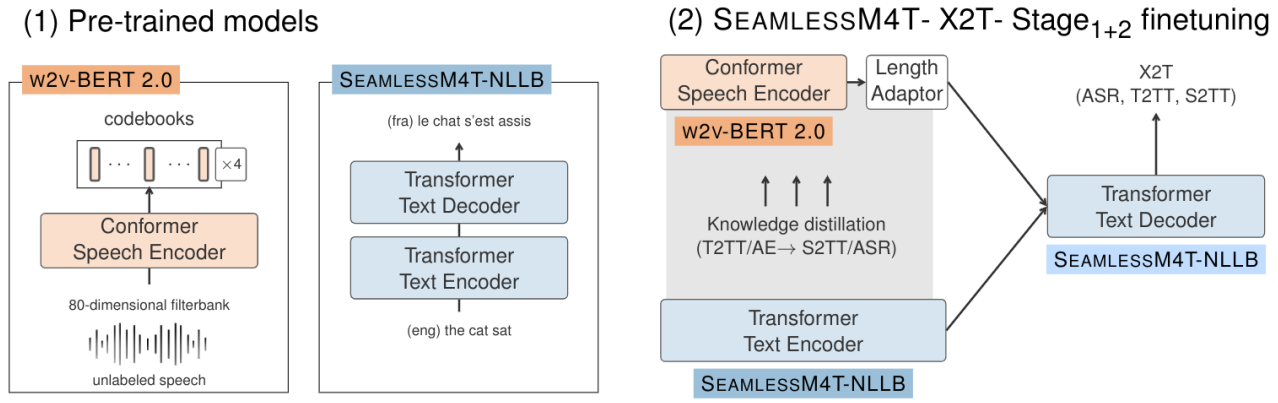
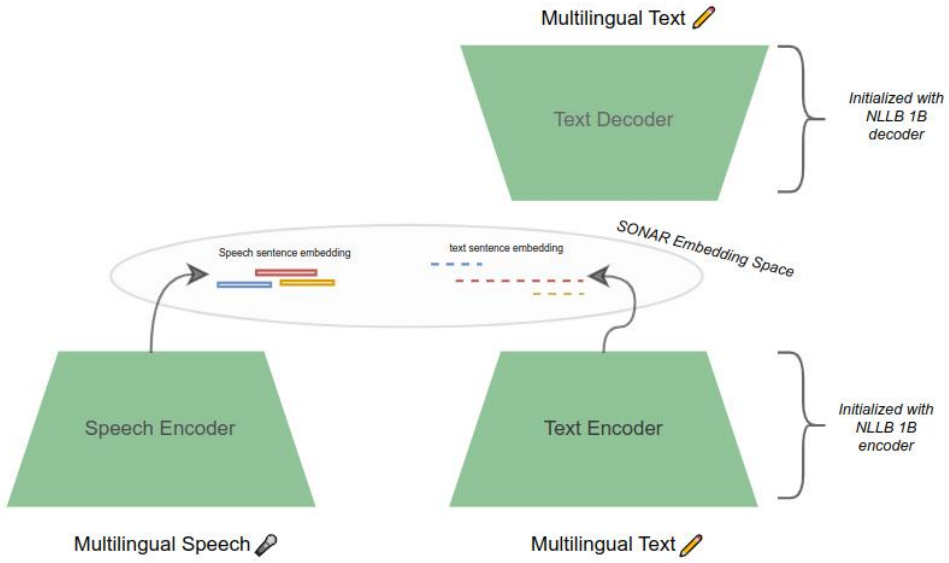


Figure 5: Overview of the SeamlessM4T X2T model. (1) describes the main two building blocks: w2v-BERT 2.0 and SeamlessM4T-NLLB. (2) describes the training of the X2T model. In Stage_1, the model is trained on X–eng directions and in Stage_2, eng–X directions are added.

Seamless4MT: speech and text data mining

- Starts with some initial amount of speech and corresponding text data
- This is used to train so-called SONAR embedding model
  - SONAR can map multilingual speech and text to a joint embedding space
  - If a speech segment (in Estonian) and the corresponding text (in English) map to close embeddings, then the English text is likely a translation
- SONAR is then used to mine for parallel speech-text data from the internet
- The resulting dataset is used to train the final model



ISO	Raw	Train	X–eng (↑BLEU)		Mined audio [h]		
	audio [h]	ASR [h]	Ours	Whisper	Sen2Txx	Sxx2Ten	Sxx2Sen
arb	106755	822	28.7	25.5	1568	8072	776
ben	7012	335	18.9	13.2	606	1345	263
cat	43531	1738	35.1	34.2	1570	4411	354
ces	41318	181	29.2	27.8	1454	6905	602
cmn	79772	9320	16.2	18.4	5440	18760	1570
cym	24161	99	14.5	13.0	–	4411	278
dan	34300	115	31.9	32.7	2499	6041	583
deu	490604	3329	32.7	34.6	91715	17634	1921
est	12691	131	23.8	18.7	1022	3346	607

Some recent results on Estonian speech translation to English & Russian

BLEU scores. Data: TV news, talkshows, press conferences

Model	Estonian-to-English	Estonian-to-Russian
Estonian ref transcripts -> Google Translate	38.9	26.1
Estonian ref transcripts -> UTartu MT	34.8	29.3
Whisper-large-v3	14.9	-
SeamlessM4T v2	13.2	16.2
Cascade: Estonian ASR -> Google Translate	34.7	23.4
SeamlessM4T v2, finetuned on found internet data (Estonian speech with English and Russian subtitles)	19.3	14.4
SeamlessM4T v2, finetuned on synth data (Estonian ASR data, translated using MT)	35.4	26.8
Whisper-large-v3, finetuned on synth data (Estonian ASR data, translated using MT)	33.2	26.1

Demo

Automatic translation (whisper-large-v3 model):

Good afternoon, dear viewers here in Tallinn's press conference, all of you who are watching us.

The webinar will be held on December 16th.

Tallinn City Council has held its inaugural meeting and today the City Council press conference will be attended by the mayor Mihail Kõlvart.

The director of communication in Tallinn from January 1, 2021, Keesti Ruul.

Our good friend from Tervisaameti, the head of the epidemiology department, Riina Tantsenko.

And Kaisa Kamm, the director of emergency care in Estonia.

And we have already heard and read the news today that there were 590 positive cases in the evening.

And that's why the Tervisaameti epidemiology department head, Riina Tantsenko, also starts today.

Demo

Automatic translation. Whisper, finetuned on synth data (Estonian ASR data, translated using MT):

Good day, dear journalists here at the Tallinn presentation and in the press center, all interested are watching us via online transmission, it is the sixteenth of December.

The Tallinn city government has held its regular session, and today the mayor Mihhail Kõlvart is in front of you at the city government's press conference.

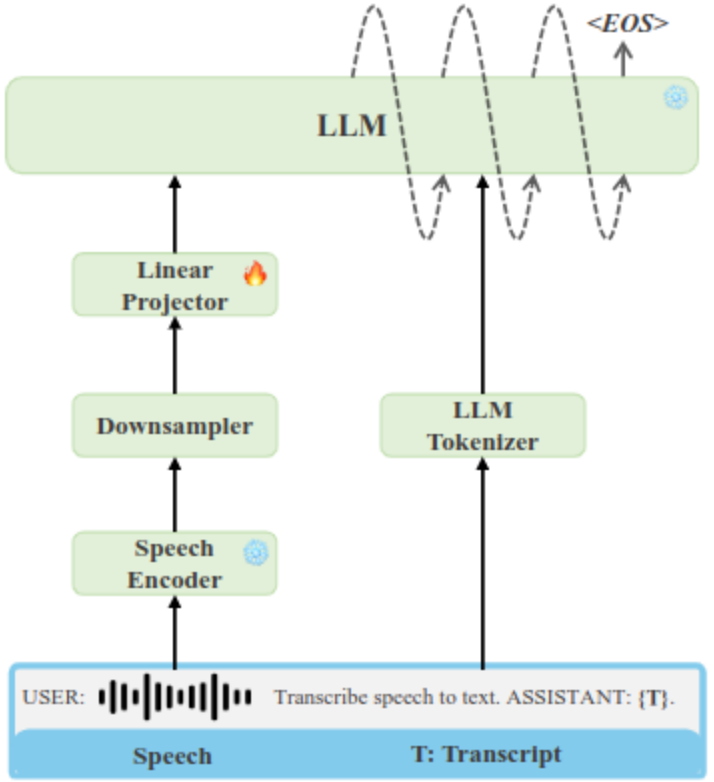
The director of city communication from the first of January, two thousand and twenty one, Kersti Ruu, our good acquaintance from the Board of Health, Irina Tontšenko, adviser to the Epidemic Control Department, and Kaisa Kamm, head of direct assistance of the Estonian Animal Protection Society.

And we have already heard and read the news today that five hundred and ninety corona positives were added in a day.

And that's why Irina Tontšenko, adviser to the Epidemic Control Department of the Board of Health, starts today as well.

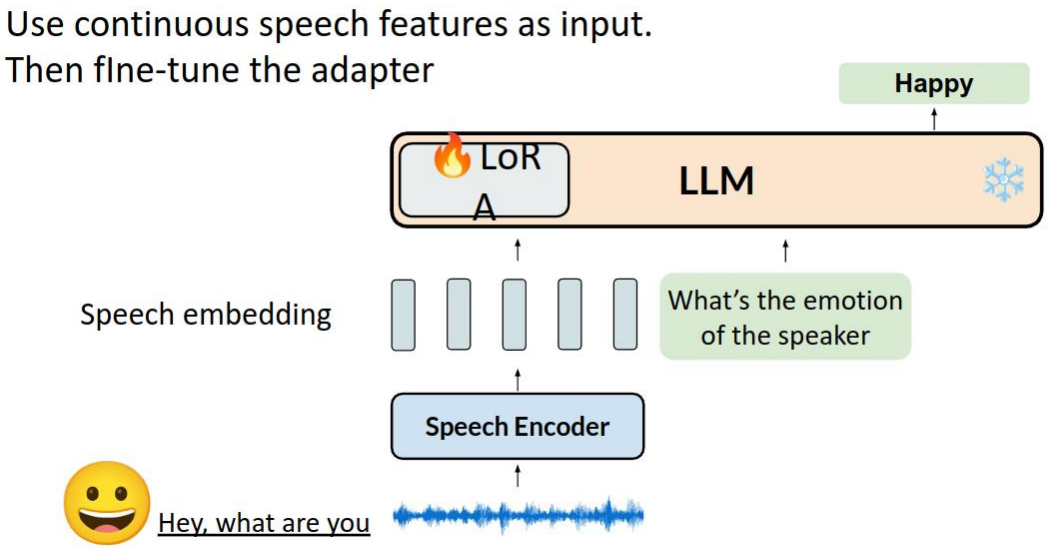
Future: Speech + LLMs

- Very recent works have shown that LLMs can be effectively used for speech processing, including end-to-end ASR, speech translation
- Very strong results
- How?
  - Use pretrained instruction-tuned LLM
  - And a speech encoder, e.g encoder part of Whisper-like model, or a wav2vec2 model
  - Downsample speech features
  - Train a projection layer that maps speech "tokens" to the same semantic space as text tokens



Speech features to LLM

- Speech features typically use a much higher frame rate, like 50-100 per second
- Typical read speech is only ~2 text tokens per second!
- Solution: downsample speech features, and train a mapping to LLM token space
- Optionally, also finetune the LLM



Slam-ASR:

- "*An Embarrassingly Simple Approach for LLM with Strong ASR Capacity*", Feb 2024
- Speech encoder: encoder module from OpenAI Whisper, or self-supervised encoder (like wav2vec), finetuned for speech recognition
- Downsampler: concatenate every 5 features, leading to 10 "superfeatures" per second
- Linear projector: simple feed-forward neural network with one hidden layer (dim=2048)
- LLM and speech encoder are pretrained
- **Linear projector is the only trainable module!**

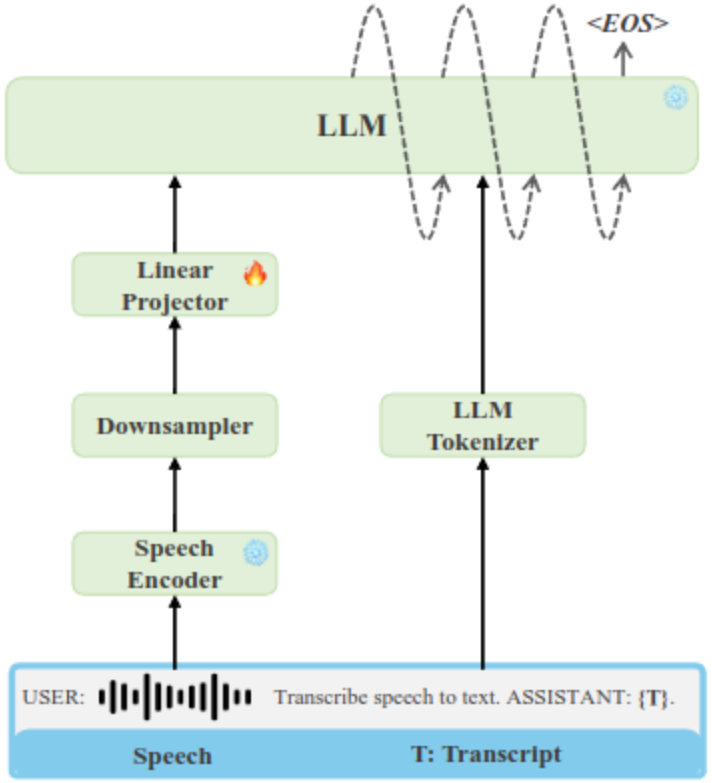


Figure 1: A brief pipeline of SLAM-ASR, at the core of which is a frozen speech encoder and a frozen LLM, with the only trainable linear projector to align between speech and text modalities.

Slam-ASR: results

- Best results when using self-supervised model (HuBERT) as encoder, finetuned for ASR

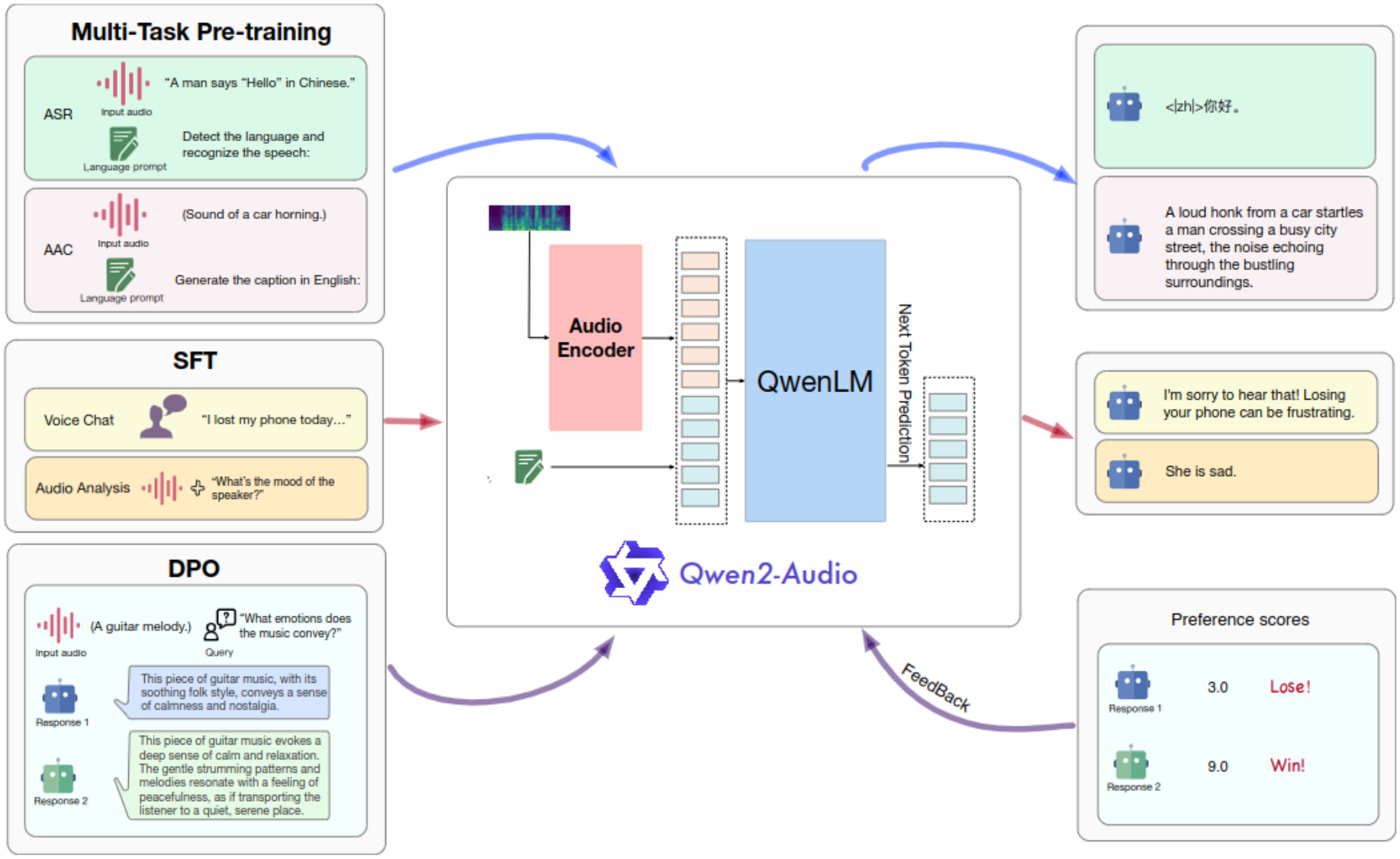
Table 4: Explore the performance with different speech encoders for LLM-based ASR. The projector is fixed with linear layers and LLM is fixed with Vicuna-7B-v1.5. LS-960 means the Librispeech 960 hours dataset.

Speech Encoder	#Encoder Params	Hidden Size	#Projector Params	WER(%) ↓	
				test-clean	test-other
Acoustic Feature					
FBank	-	80	10.03M	68.95	99.37
Supervised Speech Encoder					
Whisper-tiny	7.63M	394	12.33M	7.07	16.01
Whisper-base	19.82M	512	13.64M	5.07	13.07
Whisper-small	87.00M	768	16.26M	4.19	9.50
Whisper-medium	305.68M	1024	18.88M	2.72	6.79
Whisper-large	634.86M	1280	21.50M	2.58	6.47
+ Qwen-Audio Fine-tuning	634.86M	1280	21.50M	2.52	6.35
Self-supervised Speech Encoder					
HuBERT Base	94.70M	768	16.26M	5.39	11.99
WavLM Base	94.38M	768	16.26M	4.14	9.66
HuBERT Large	316.61M	1024	18.88M	4.53	8.74
+ LS-960 Fine-tuning	316.61M	1024	18.88M	2.30	4.53
WavLM Large	315.45M	1024	18.88M	2.37	4.90
HuBERT X-Large	964.32M	1280	21.50M	4.29	6.66
+ LS-960 Fine-tuning (SLAM-ASR)	964.32M	1280	21.50M	1.94	3.81

Qwen2-Audio: chat using audio

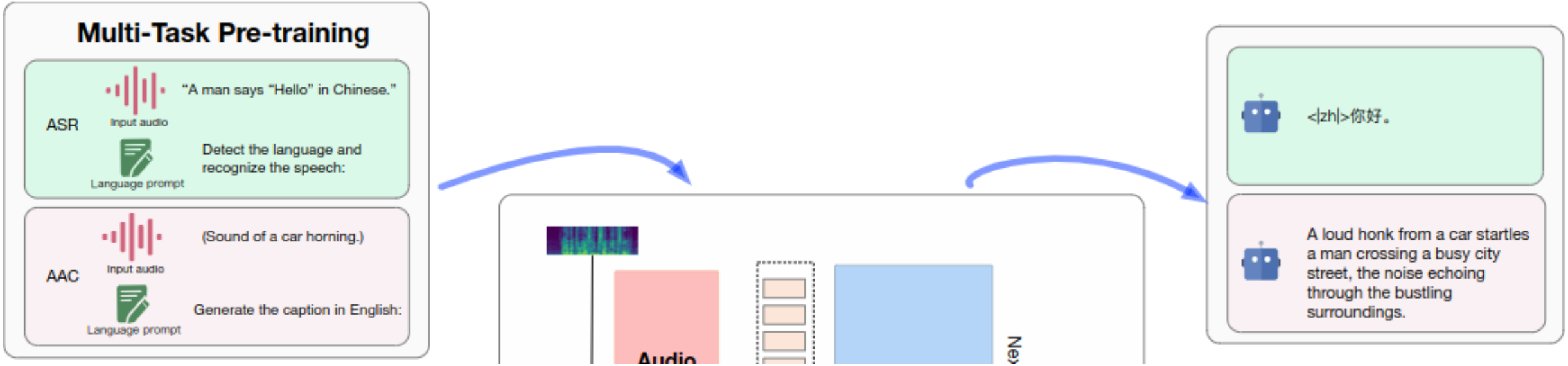
Qwen2 from Alibaba as LLM

Audio encoder from Whisper



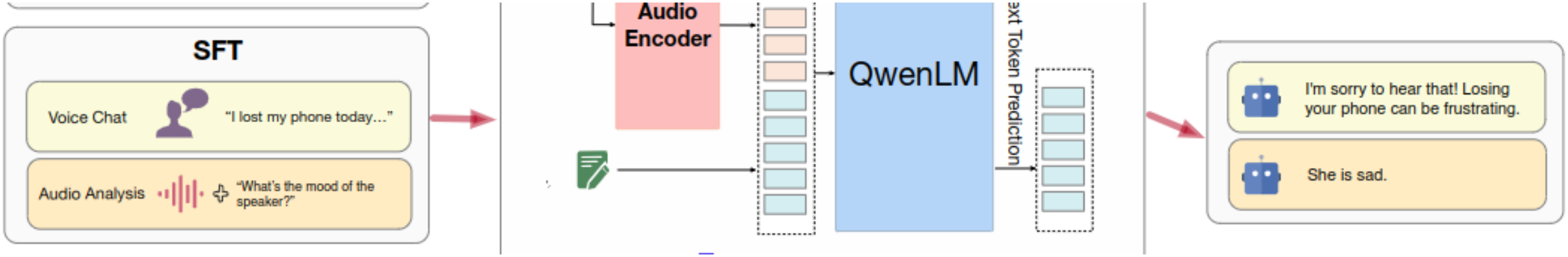
Qwen2-Audio, cont.

- Pretraining phase: use ASR, speech translation, emotion recognition, audio captioning datasets to do large-scale pretraining



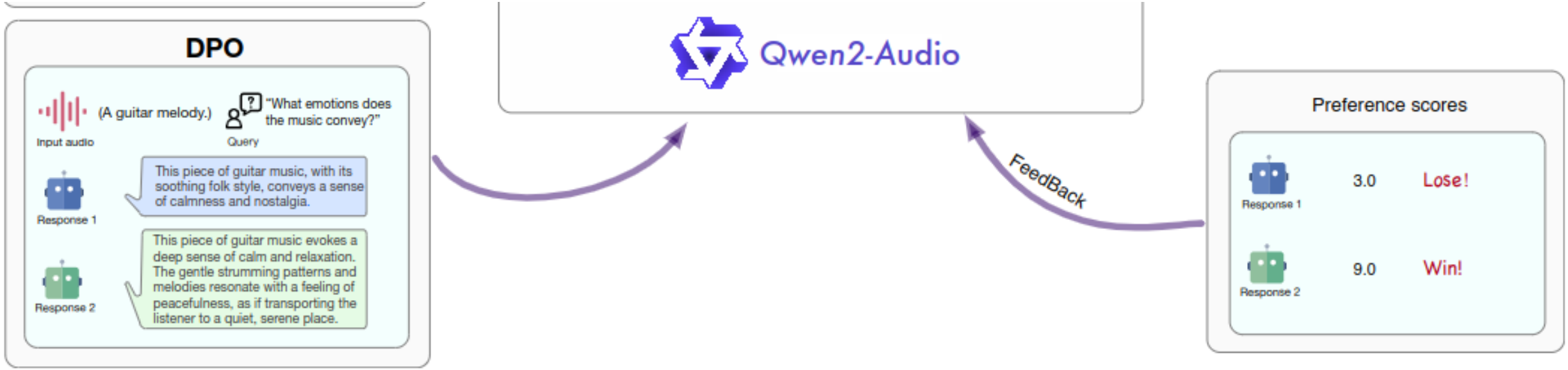
Qwen2-Audio, cont.

- Supervised finetuning phase: use annotators to generate new training data
- Two distinct modes for human interactions:
  - Audio Analysis: In the audio analysis mode, annotators are asked to use Qwen2-Audio analyze a diverse array of audio. User instructions can be given either through audio or text.
    - This mode is often used for offline analysis of audio files.
  - Voice Chat: In the voice chat mode, users are encouraged to engage in voice conversations with Qwen2-Audio, asking a wide range of questions. "Please feel free to consider it your voice chat assistant."

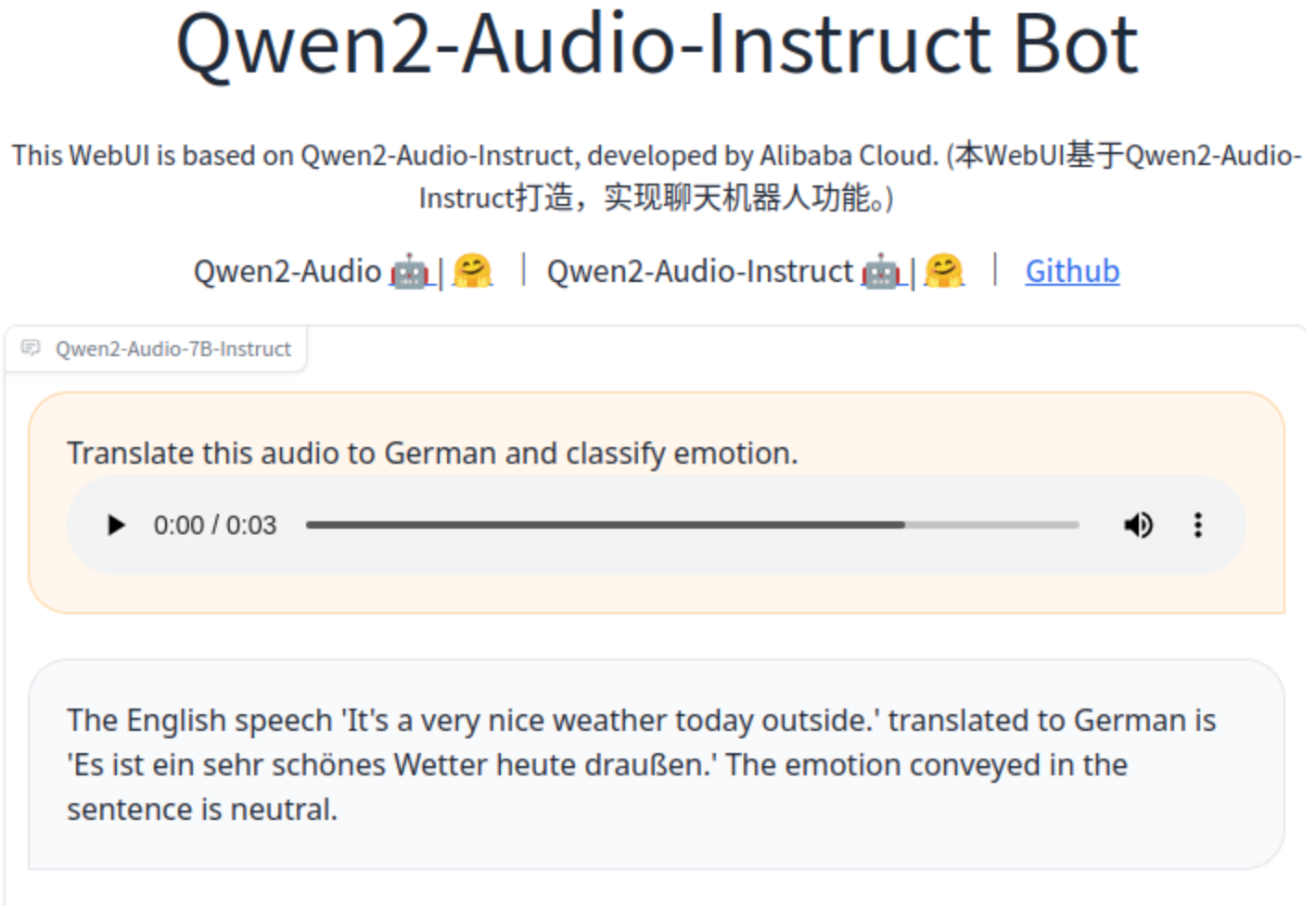


Qwen2-Audio, cont.

- DPO phase: further optimizes models to follow human preferences.



Qwen2-Audio, cont.



https://huggingface.co/spaces/Qwen/Qwen2-Audio-Instruct-Demo

Conclusion

- Speech translation is one of the "holy grails" of NLP
- Future: emotion-preserving, voice-preserving simultenaous speech-to-speech translation (like movie dubbing)