Simultaneous Multi-Source Speech Translation

Dominik Macháček, Dec 14 2022, NAIST

Outline

1. はじめまして
2. Book teaser: The reality of Multi-ling. MT
3. **Simultaneous Multi-Source Speech Translation**
4. Human Evaluation of Sim. ST (Continuous Rating)
5. MT Metrics Correlate with CR in Sim. Mode
6. Summary

01
はじめまして!

Introduction

はじめまして

- Dominik Macháček (ドミニク マハーチェク)
  Email: `machacek@ufal.mff.cuni.cz`
  Web: `ufal.cz/dominik-machacek`

はじめまして

- Dominik Macháček (ドミニク マハーチェク)
  Email: `machacek@ufal.mff.cuni.cz`
  Web: `ufal.cz/dominik-machacek`
- 4th year PhD student at ÚFAL departement, Charles University (CUNI), Prague, Czech Republic
- my advisor: Ondřej Bojar

はじめまして

- Dominik Macháček (ドミニク マハーチェク)
  Email: `machacek@ufal.mff.cuni.cz`
  Web: `ufal.cz/dominik-machacek`
- 4th year PhD student at ÚFAL departement, Charles University (CUNI), Prague, Czech Republic
- my advisor: Ondřej Bojar
- my background: Computer Science and Computational Linguistics
- my topics: Machine Translation, Simultaneous Speech Translation

はじめまして

- Dominik Macháček (ドミニク マハーチェク)
  Email: `machacek@ufal.mff.cuni.cz`
  Web: `ufal.cz/dominik-machacek`
- 4th year PhD student at ÚFAL departement, Charles University (CUNI), Prague, Czech Republic
- my advisor: Ondřej Bojar
- my background: Computer Science and Computational Linguistics
- my topics: Machine Translation, Simultaneous Speech Translation
- On-site internship at NICT

はじめまして

- Dominik Macháček (ドミニク マハーチェク)
  Email: `machacek@ufal.mff.cuni.cz`
  Web: `ufal.cz/dominik-machacek`
- 4th year PhD student at ÚFAL departement, Charles University (CUNI), Prague, Czech Republic
- my advisor: Ondřej Bojar
- my background: Computer Science and Computational Linguistics
- my topics: Machine Translation, Simultaneous Speech Translation
- On-site internship at NICT
- Why here? Briefly: our work is related

My Collaborators

- Ondřej Bojar, ass. prof. at ÚFAL – advisor
- Peter Polák, PhD student at ÚFAL – sim. end-to-end ASR and ST
- Dávid Javorský, PhD student at ÚFAL – sim. ST evaluation, IWSLT22
- Raj Dabre – consultant and mentor at NICT

02

The Reality of Multi-Lingual MT

Book teaser

The Reality of Multi-Lingual MT

Kocmi, Macháček, Bojar (2021)

`ufal.cz/books/2021-kocmi`

- Benefits and perils of more than 2 langs. in MT
- Warnings against too optimistic and unjustified explanations!
- Transfer Learning
- Multi-ling. techniques survey
- Practical aspects of deploying
- Good computer cluster
- Inclusivity of research
- Ecological trace, ...

03

Simultaneous Multi-Source ST

Speech Translation from Source AND Interpreter

Image source: https://ec.europa.eu/education/knowledge-centre-interpretation/conference-interpreting/conference-interpreting-explained_en

Speech Translation from Source AND Interpreter

Image source: https://ec.europa.eu/education/knowledge-centre-interpretation/conference-interpreting/conference-interpreting-explained_en

Speech Translation from Source AND Interpreter

Image source: `https://ec.europa.eu/education/knowledge-centre-interpretation/conference-interpreting/conference-interpreting-explained_en`

Speech Translation from Source AND Interpreter

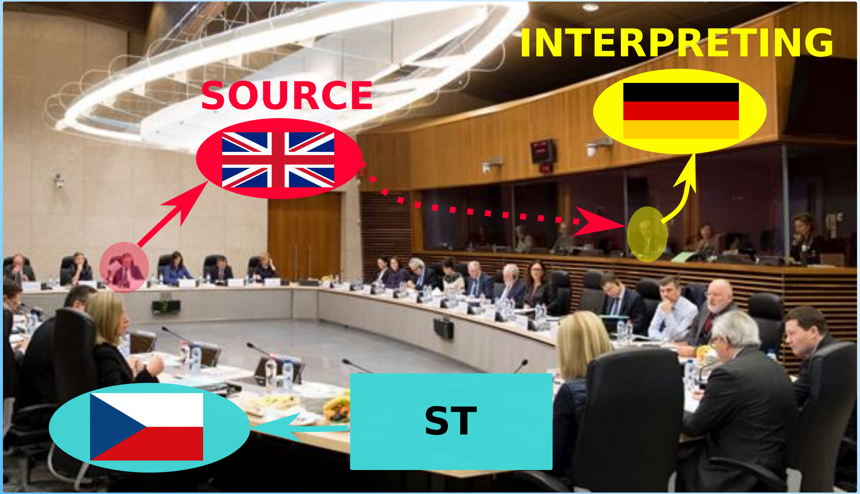


Image source: `https://ec.europa.eu/education/knowledge-centre-interpretation/conference-interpreting/conference-interpreting-explained_en`

Speech Translation from Source AND Interpreter

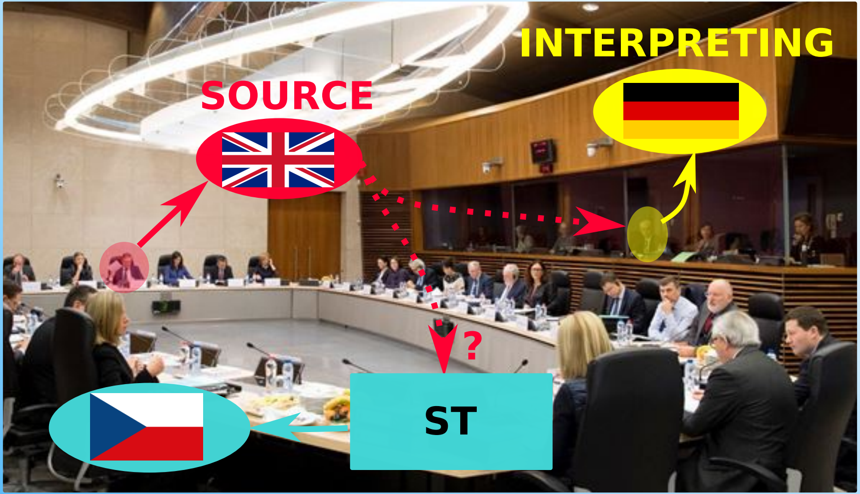


Image source: `https://ec.europa.eu/education/knowledge-centre-interpretation/conference-interpreting/conference-interpreting-explained_en`

Speech Translation from Source AND Interpreter

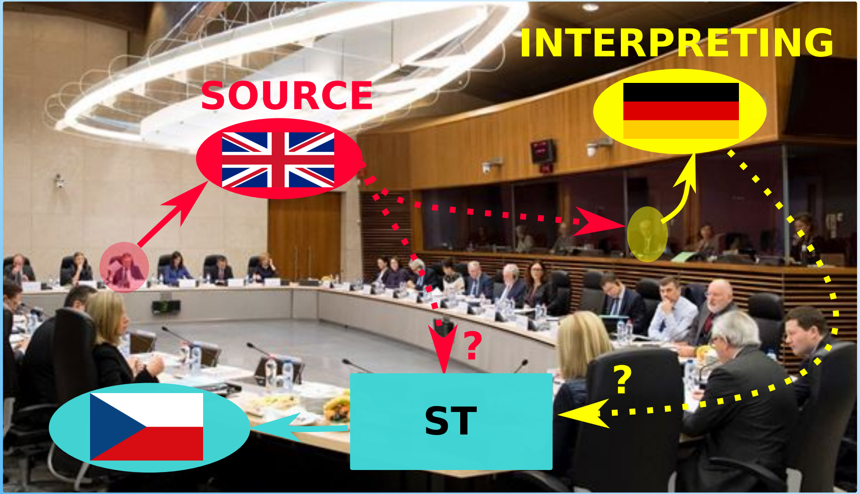


Image source: `https://ec.europa.eu/education/knowledge-centre-interpretation/conference-interpreting/conference-interpreting-explained_en`

Outline

3.1
Motivation

for Sim. Multi-Source ST

Benefits and Risks of Source+Interpreter ST

- Quality
  - Desambiguation: Schloss + lock vs castle

Benefits and Risks of Source+Interpreter ST

- Quality
  - Desambiguation: Schloss + lock vs castle
  - ASR errors complement each other across languages.

Benefits and Risks of Source+Interpreter ST

- Quality
  - Desambiguation: Schloss + lock vs castle
  - ASR errors complement each other across languages.
- No human interaction for detecting and switching the optimal source.

Benefits and Risks of Source+Interpreter ST

- Quality
  - Desambiguation: Schloss + lock vs castle
  - ASR errors complement each other across languages.
- No human interaction for detecting and switching the optimal source.
- Possibly best from both options:
  - source – word-for-word, faithful = too complex to perceive?, fast, not much controllable
  - interpreter – brief, simpler, inter-culture transfer, but how reliable?, slower, controllable

Benefits and Risks of Source+Interpreter ST

- Quality
  - Desambiguation: Schloss + lock vs castle
  - ASR errors complement each other across languages.
- No human interaction for detecting and switching the optimal source.
- Possibly best from both options:
  - source – word-for-word, faithful = too complex to perceive?, fast, not much controllable
  - interpreter – brief, simpler, inter-culture transfer, but how reliable?, slower, controllable

→ We know little about what do the target users actually need.

Benefits and Risks of Source+Interpreter ST

- Quality
  - Desambiguation: Schloss + lock vs castle
  - ASR errors complement each other across languages.
- No human interaction for detecting and switching the optimal source.
- Possibly best from both options:
  - source – word-for-word, faithful = too complex to perceive?, fast, not much controllable
  - interpreter – brief, simpler, inter-culture transfer, but how reliable?, slower, controllable

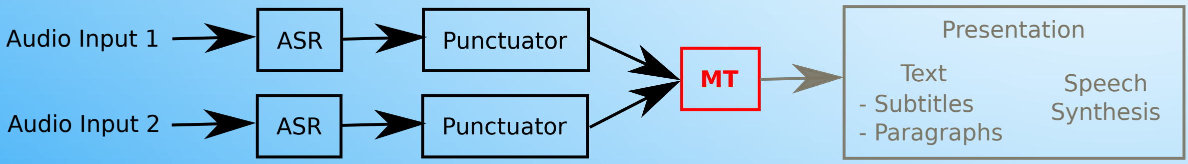
  → We know little about what do the target users actually need.
- Risk of no room for improvement in practice:
  - One source always good enough / more sources never good enough.

3.2
Specification

of Sim. Multi-Source ST

Cascaded Speech Translation (ST)

- I focus on **MT part in cascaded ST** with unspecified output modality

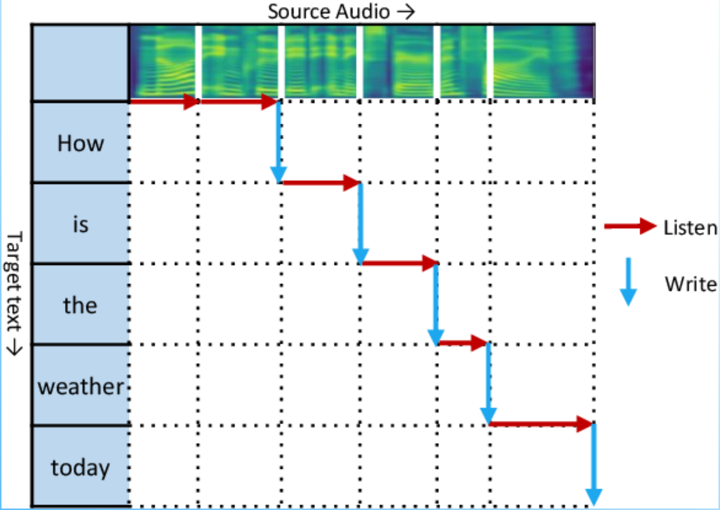


Long-Form Monologue

- **Authentic use-case**
- Often need for simultaneity
- Challenges:
  - Read or spontaneous
  - Disfluencies
  - Native/Non-native
  - Interruptions
  - ...etc.
- No clear sentence boundaries

Image source: `https://www.europarl.europa.eu/`

Simultaneous



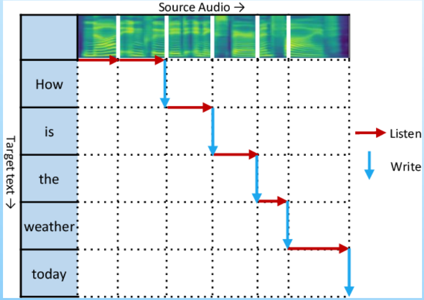
Reprinted from Ren et al., 2020; SimulSpeech: End-to-End Simultaneous Speech to Text Translation

Re-translation vs. Streaming

- Re-translate from beginning of sentence each time: rewrite + append
- Latency vs stability. Top quality.

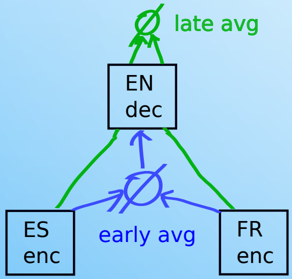
Source	Output								Erasure
1: Neue	New								-
2: Arzneimittel	New	Medicines							0
3: könnten	New	Medicines							0
4: Lungen-	New	drugs	may	be	lung				1
5: und	New	drugs	could	be	lung	and			3
6: Eierstockkrebs	New	drugs	may	be	lung	and	ovarian	cancer	4
7: verlangsamen	New	drugs	may	slow	lung	and	ovarian	cancer	5
Content Delay	1	4	6	7	7	7	7	7	

- MT alternates between reading from ASR and translating: no rewrites, only append
- Latency vs quality. Top stability.



Multi-Source NMT Models

- one encoder, concat sources to one sequence (Dabre et al., 2017), e.g. Hello Bonjour Namaskar Kamusta Hallo → konnichiwa
- multi-encoder NMT (Firat et al., 2016)



3.3

Interpreting

in Sim. Multi-Source ST

Interpreting Analysis

- Shortening: sim. interpreting is by 13% shorter than offline manual translation

Interpreting Analysis

- Shortening: sim. interpreting is by 13% shorter than offline manual translation
  - En-Cs, average document length in number of syllables, ESIC test

Interpreting Analysis

- Shortening: sim. interpreting is by 13% shorter than offline manual translation
  - En-Cs, average document length in number of syllables, ESIC test
- Simplification: words with significantly lower rank in corpus

Interpreting Analysis

- Shortening: sim. interpreting is by 13% shorter than offline manual translation
  - En-Cs, average document length in number of syllables, ESIC test
- Simplification: words with significantly lower rank in corpus
- Latency: inpt. 4 sec. behind src, intp+MT appx. 9.8 sec.

Interpreting Analysis

- Shortening: sim. interpreting is by 13% shorter than offline manual translation
  - En-Cs, average document length in number of syllables, ESIC test
- Simplification: words with significantly lower rank in corpus
- Latency: inpt. 4 sec. behind src, intp+MT appx. 9.8 sec.
  $\rightarrow$ similar to relay interpreting, acceptable

These results are from Macháček et al., INTERSPEECH 2021: Lost in Interpreting: Speech Translation from Source or Interpreter?

Interpreting Strategies

(Resource: Interpreting training and theory, e.g. Čeňková, Ešnerová, Olsen)

Interpreting Strategies

(Resource: Interpreting training and theory, e.g. Čeňková, Ešnerová, Olsen)

- **Segmentation** to sentences: prefer simple sentences, avoid long distance dependencies
  $\rightarrow$ not 1:1 sentence alignment as in text-to-text translation

Interpreting Strategies

(Resource: Interpreting training and theory, e.g. Čeňková, Ešnerová, Olsen)

- **Segmentation** to sentences: prefer simple sentences, avoid long distance dependencies
  $\rightarrow$ not 1:1 sentence alignment as in text-to-text translation
- **Language economy**: redundancy reduction (ehm), short variants

Interpreting Strategies

(Resource: Interpreting training and theory, e.g. Čeňková, Ešnerová, Olsen)

- **Segmentation** to sentences: prefer simple sentences, avoid long distance dependencies
  $\rightarrow$ not 1:1 sentence alignment as in text-to-text translation
- **Language economy**: redundancy reduction (ehm), short variants
- **Generalization**: cats and dogs $\rightarrow$ pets ... short
  a carp $\rightarrow$ a freshwater fish ... when forgot translation
  Hallwang $\rightarrow$ some village ... foreign audience doesn't know it anyway

Interpreting Strategies

(Resource: Interpreting training and theory, e.g. Čeňková, Ešnerová, Olsen)

- **Segmentation** to sentences: prefer simple sentences, avoid long distance dependencies
  → not 1:1 sentence alignment as in text-to-text translation
- **Language economy**: redundancy reduction (ehm), short variants
- **Generalization**: cats and dogs → pets ... short
  a carp → a freshwater fish ... when forgot translation
  Hallwang → some village ... foreign audience doesn't know it anyway
- **Grammar constructions**: e.g. passivisation in En-Jap. to overcome word-order diff. (He et al., 2016)

Interpreting Strategies

(Resource: Interpreting training and theory, e.g. Čeňková, Ešnerová, Olsen)

- **Segmentation** to sentences: prefer simple sentences, avoid long distance dependencies
  $\rightarrow$ not 1:1 sentence alignment as in text-to-text translation
- **Language economy**: redundancy reduction (ehm), short variants
- **Generalization**: cats and dogs $\rightarrow$ pets ... short
  a carp $\rightarrow$ a freshwater fish ... when forgot translation
  Hallwang $\rightarrow$ some village ... foreign audience doesn't know it anyway
- **Grammar constructions**: e.g. passivisation in En-Jap. to overcome word-order diff. (He et al., 2016)

Interpreting Strategies

(Resource: Interpreting training and theory, e.g. Čeňková, Ešnerová, Olsen)

- **Segmentation** to sentences: prefer simple sentences, avoid long distance dependencies
  $\rightarrow$ not 1:1 sentence alignment as in text-to-text translation
- **Language economy**: redundancy reduction (ehm), short variants
- **Generalization**: cats and dogs $\rightarrow$ pets ... short
  a carp $\rightarrow$ a freshwater fish ... when forgot translation
  Hallwang $\rightarrow$ some village ... foreign audience doesn't know it anyway
- **Grammar constructions**: e.g. passivisation in En-Jap. to overcome word-order diff. (He et al., 2016)

$\Rightarrow$ Let's use: supervised learning, multi-sequence to sequence processing, NMT across sentence boundaries.

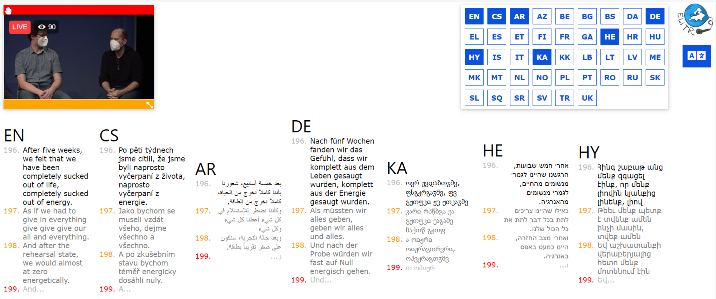
Example

1. Segmentation into sentences
2. Shorter, simpler, removed disfluencies
3. “Cultural independence”

Source (En)	Interpreting (En→Cs)	Gloss to Interpreting
And we try to compare the municipalities with the class of municipalities with the same size,	Zde máme srovnání obcí které mají srovnatelnou velikost.	Here we-have a-comparison of-municipalities, which have a-comparable size.
so we are not comparing Vienna to **Hallwang**, so we are trying to find similar municipalities so em so it will be a fair compare, comparison.	Nesrovnáváme tedy **nějakou vesnici** s Vídní kupříkladu, aby to bylo spravedlivé.	We-are-not-comparing thus **some village** with Vienna for-instance, so-that it was fair.

From Ondřej Bojar, 8. 12. 2022, WMT.

Controllable Speech and Sound



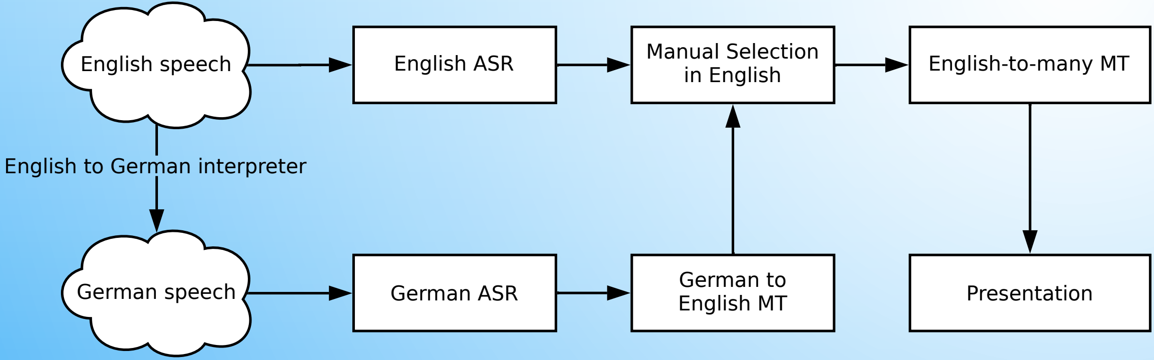
- ELITR demo: A debate after the premiere of GPT-2-written play (`theAItre.com`).
- Orig.: face masks + far microphones + spontaneous speech
- Interpreter instructed to make proper sent. boundaries + good sound $\rightarrow$ saved the sim. ST performance

3.4

“Follow All, Switch”

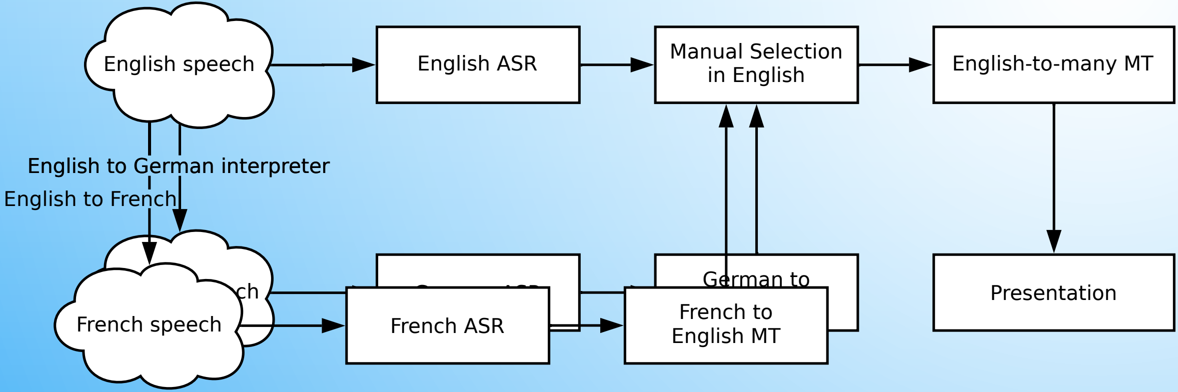
as SOTA Sim. Multi-Source ST

SOTA Multi-Source ST: Follow All, Switch



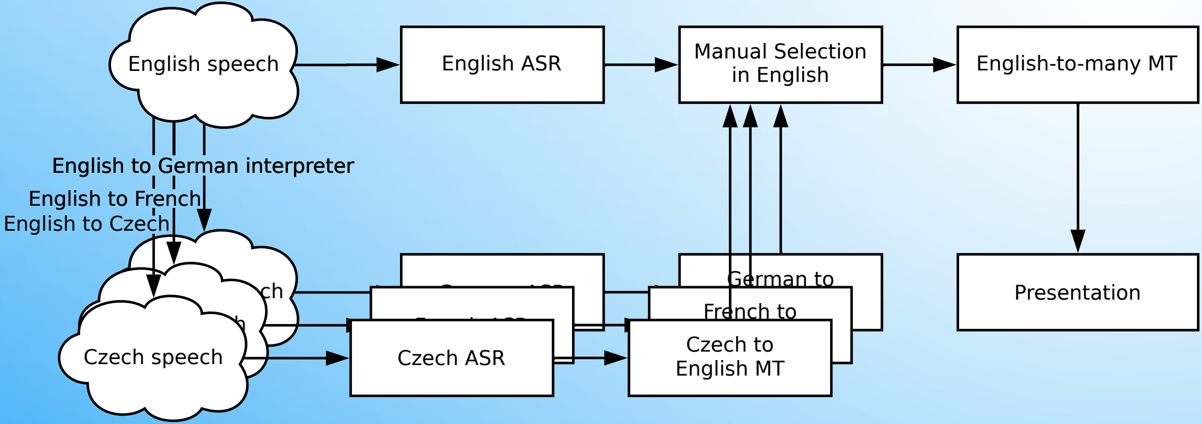
More details in Bojar et al. (2021), **Operating a Complex SLT System with Speakers and Human Interpreters**; https://aclanthology.org/2021.mtsummit-asltrw.3/

SOTA Multi-Source ST: Follow All, Switch



More details in Bojar et al. (2021), **Operating a Complex SLT System with Speakers and Human Interpreters**; `https://aclanthology.org/2021.mtsummit-asltrw.3/`

SOTA Multi-Source ST: Follow All, Switch



More details in Bojar et al. (2021), **Operating a Complex SLT System with Speakers and Human Interpreters**; `https://aclanthology.org/2021.mtsummit-asltrw.3/`

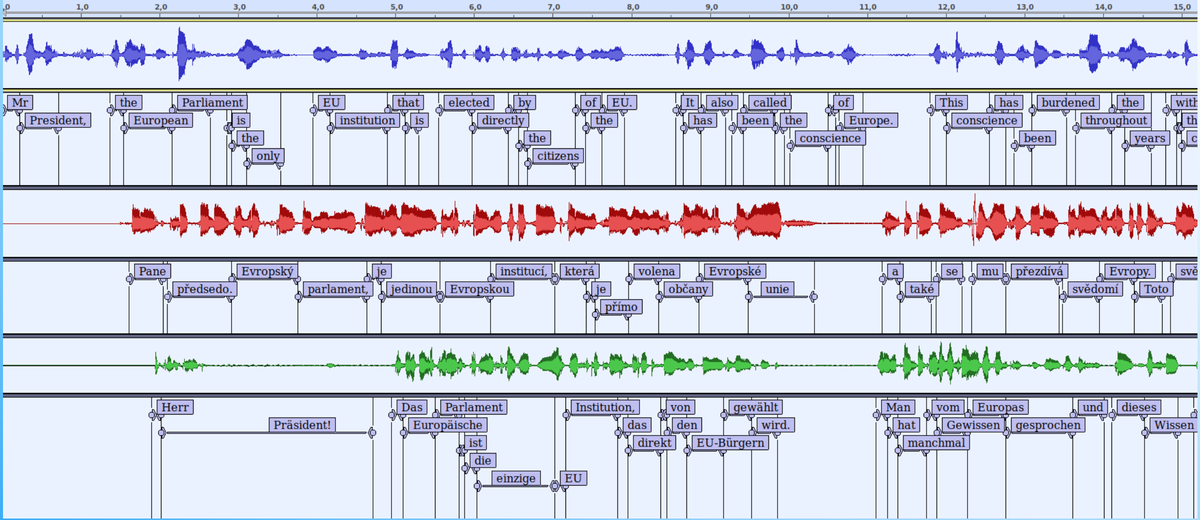
3.5
ESIC Evaluation Corpus

for Sim. Multi-Source ST

ESIC: Europarl Simultaneous Interpreting Corpus

- 10 hours, 370 authentic recordings from European Parliament
  - 2008-2011: EP publishes both **translations** and **interpreting** of plenary sessions into all 23 EU langs.
- original English + simultaneous interpreting into Czech + German
- manual transcriptions, word-level timestamps

ESIC: Europarl Simultaneous Interpreting Corpus



ESIC: Europarl Simultaneous Interpreting Corpus

- video, audio, metadata, parallel translations
- usable by many ways (e.g. interpreting analysis, speech reconstruction, analyzing non-native, fluent vs read speech, ASR, MT, SLT, simultaneous MT evaluation, SOV vs SVO MT...)
- download link: `http://hdl.handle.net/11234/1-3719`
- please cite as
  Macháček et al., 2021, Lost in Interpreting: Speech Translation from Source or Interpreter?, INTERSPEECH 2021

3.6
Mock ASR Results

Sim. Multi-Source ST

Mock ASR for Evaluation

- one source may be always good enough / more sources never enough. Is there any space between, where multi-sourcing is beneficial?

Mock ASR for Evaluation

- one source may be always good enough / more sources never enough. Is there any space between, where multi-sourcing is beneficial?
- possibly, multi-sourcing might work only with ASRs of e.g. 10-15 % WER, but not $< 10\%$ and $> 15\%$

Mock ASR for Evaluation

- one source may be always good enough / more sources never enough. Is there any space between, where multi-sourcing is beneficial?
- possibly, multi-sourcing might work only with ASRs of e.g. 10-15 % WER, but not $< 10\%$ and $> 15\%$
- we may not have such ASRs

Mock ASR for Evaluation

- one source may be always good enough / more sources never enough. Is there any space between, where multi-sourcing is beneficial?
- possibly, multi-sourcing might work only with ASRs of e.g. 10-15 % WER, but not $< 10\%$ and $> 15\%$
- we may not have such ASRs $\rightarrow$ "corrupt" test sources to X% WER

Mock ASR for Evaluation

ÚFAL

- one source may be always good enough / more sources never enough. Is there any space between, where multi-sourcing is beneficial?
- possibly, multi-sourcing might work only with ASRs of e.g. 10-15 % WER, but not $< 10\%$ and $> 15\%$
- we may not have such ASRs $\rightarrow$ "corrupt" test sources to X% WER
  Lexical Modeling of ASR Errors for Robust Speech Translation, Martucci et al., INTERSPEECH 2021

Mock ASR for Evaluation

- one source may be always good enough / more sources never enough. Is there any space between, where multi-sourcing is beneficial?
- possibly, multi-sourcing might work only with ASRs of e.g. 10-15 % WER, but not $< 10\%$ and $> 15\%$
- we may not have such ASRs $\rightarrow$ "corrupt" test sources to X% WER
  Lexical Modeling of ASR Errors for Robust Speech Translation, Martucci et al., INTERSPEECH 2021
  - learn Cp/Sub/Del/Ins on (gold; ASR transcript) pairs
  - rewrite unigrams

Multi-Sourcing

Offline Mode

- En+De→Cs multi-way Transf. NMT, bilingual training
- Marian training, PyTorch decoding
- late averaging – as ensembling, sources need to be parallel sent.!
- two checkpoints from the same training

Simultaneous Mode

- finetuned for stability and quality
- Streaming with LocalAgreement-n
  - ref. CUNI-KIT IWSLT22
  - internally decode every token, commit tgt. prefix of last n
- proportional alignment of src.
- Average Lagging: count only En

Both evaluated on paralell ESIC translations, not on orig+intp. audio!

Results: Offline Mode

		En WER								
	single-src.	0 %	5 %	10 %	15 %	20 %	25 %	30 %	35 %	40 %
s-src.		33.00±0.00	29.40±0.26	26.30±0.35	22.97±0.45	20.40±0.44	18.00±0.46	15.90±0.10	13.93±0.23	12.13±0.06
De WER 0 %	26.00±0.00	31.50±0.00	29.95±0.08	**28.43±0.21**	26.58±0.47	25.20±0.41	23.67±0.23	22.13±0.14	20.57±0.19	19.63±0.19
5 %	23.53±0.12	30.70±0.10	28.85±0.17	27.47±0.29	**25.75±0.47**	24.20±0.36	22.80±0.36	21.07±0.13	19.70±0.14	18.68±0.13
10 %	21.50±0.10	29.80±0.10	28.05±0.17	26.58±0.13	24.75±0.47	23.30±0.36	21.73±0.15	20.27±0.05	18.65±0.25	17.67±0.26
15 %	18.93±0.42	28.63±0.21	27.02±0.15	25.65±0.17	23.85±0.44	**22.52±0.48**	**21.05±0.40**	19.48±0.26	17.90±0.14	16.80±0.24
20 %	17.23±0.25	27.97±0.21	26.35±0.17	24.70±0.16	22.97±0.38	21.82±0.45	**20.17±0.26**	18.30±0.24	17.05±0.10	15.57±0.21
25 %	15.50±0.26	26.87±0.15	25.32±0.10	23.85±0.24	22.15±0.29	20.78±0.51	19.30±0.38	17.65±0.10	16.38±0.32	14.75±0.10
30 %	13.93±0.21	26.23±0.45	24.70±0.23	23.35±0.33	21.20±0.24	19.73±0.34	18.57±0.29	16.52±0.13	15.10±0.20	14.03±0.05
35 %	12.53±0.32	24.60±0.36	22.65±0.17	21.10±0.12	19.30±0.35	18.20±0.24	16.83±0.38	15.32±0.21	14.10±0.16	12.90±0.12
40 %	10.80±0.26	23.33±0.15	21.57±0.15	20.00±0.14	18.30±0.20	16.90±0.38	15.83±0.38	14.60±0.12	13.00±0.24	11.98±0.24

Table 5: BLEU (avg±stddev) with transcription noise on ESIC dev set. Green-backgrounded area is where the English single-source outperforms German single-source. Black underlined numbers indicate the area where multi-sourcing achieves higher score than both single-sourcing options. In **bold** is near maximum gap from single-source, more than 2.1 BLEU. Red-colored numbers are where at least one single-source scores higher.

Results: Simultaneous Mode

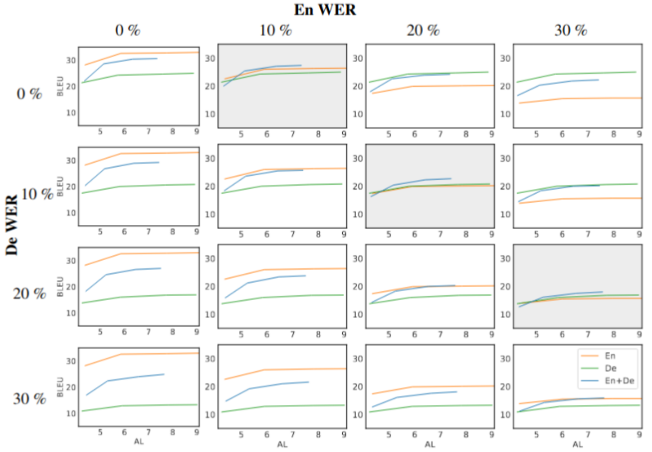


Figure 1: Single-sourcing vs multi-sourcing with different level of artificial ASR noise of the sources (% WER) in simultaneous mode on ESIC dev set. The results are depicted as quality (BLEU) and latency (AL) trade-off of the candidate systems. The plots highlighted by grey background show noise levels where multi-sourcing (En+De, blue line) outperforms both single sources in BLEU at least for AL>5.5.

Limitations

- it depends, from which language the reference was translated
- not realistic use-case, but
  "Robustness of Multi-Source MT to Transcription Errors"
- paper under review
- ...work in progress

3.7

Nearest Plans

with Sim. Multi-Source ST

Nearest plan: Realistic use-case

- Briefly: focus to multi-source for sim. speech + interpreting
  - original + interpreting (not parallel sent.-aligned translations)
  - time offsets

1. evaluation method
2. baseline – late averaging of parallel sources
3. improve baseline:
   - multi-parallel training
   - training with synthetic interpreting?
   - training with ASR noise?
   - ~~quality estimation~~

04

Human Evaluation of Simultaneous ST

aka Continuous Rating

Challenges in MT Evaluation

- Offline text-to-text MT:

Challenges in MT Evaluation

- Offline text-to-text MT:
  - MT quality → direct assessment (DA)

Challenges in MT Evaluation

- Offline text-to-text MT:
  - MT quality $\rightarrow$ direct assessment (DA)
  - competent evaluators $\rightarrow$ bilinguals

Challenges in MT Evaluation

- Offline text-to-text MT:
  - MT quality $\rightarrow$ direct assessment (DA)
  - competent evaluators $\rightarrow$ bilinguals
  - different opinions $\rightarrow$ repetitions, statistics

Challenges in MT Evaluation

- Offline text-to-text MT:
  - MT quality → direct assessment (DA)
  - competent evaluators → bilinguals
  - different opinions → repetitions, statistics
- Offline ST: Speech as input modality

Challenges in MT Evaluation

- Offline text-to-text MT:
  - MT quality $\rightarrow$ direct assessment (DA)
  - competent evaluators $\rightarrow$ bilinguals
  - different opinions $\rightarrow$ repetitions, statistics
- Offline ST: Speech as input modality
- Simultaneous ST:

Challenges in MT Evaluation

- Offline text-to-text MT:
  - MT quality $\rightarrow$ direct assessment (DA)
  - competent evaluators $\rightarrow$ bilinguals
  - different opinions $\rightarrow$ repetitions, statistics
- Offline ST: Speech as input modality
- Simultaneous ST:
  - Simultaneity

Challenges in MT Evaluation

- Offline text-to-text MT:
  - MT quality $\rightarrow$ direct assessment (DA)
  - competent evaluators $\rightarrow$ bilinguals
  - different opinions $\rightarrow$ repetitions, statistics
- Offline ST: Speech as input modality
- Simultaneous ST:
  - Simultaneity
    - only one access to document $\rightarrow$ human memory

Challenges in MT Evaluation

- Offline text-to-text MT:
  - MT quality $\rightarrow$ direct assessment (DA)
  - competent evaluators $\rightarrow$ bilinguals
  - different opinions $\rightarrow$ repetitions, statistics
- Offline ST: Speech as input modality
- Simultaneous ST:
  - Simultaneity
    - only one access to document $\rightarrow$ human memory
    - high demands on concentration

Challenges in MT Evaluation

- Offline text-to-text MT:
  - MT quality → direct assessment (DA)
  - competent evaluators → bilinguals
  - different opinions → repetitions, statistics
- Offline ST: Speech as input modality
- Simultaneous ST:
  - Simultaneity
    - only one access to document → human memory
    - high demands on concentration
    - left-only context, document context

Challenges in MT Evaluation

- Offline text-to-text MT:
  - MT quality → direct assessment (DA)
  - competent evaluators → bilinguals
  - different opinions → repetitions, statistics
- Offline ST: Speech as input modality
- Simultaneous ST:
  - Simultaneity
    - only one access to document → human memory
    - high demands on concentration
    - left-only context, document context
  - Presentation options influence latency and readability (in re-translating)

Challenges in MT Evaluation

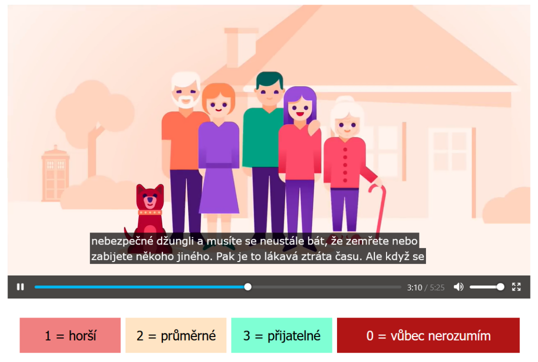
- Offline text-to-text MT:
  - MT quality $\rightarrow$ direct assessment (DA)
  - competent evaluators $\rightarrow$ bilinguals
  - different opinions $\rightarrow$ repetitions, statistics
- Offline ST: Speech as input modality
- Simultaneous ST:
  - Simultaneity
    - only one access to document $\rightarrow$ human memory
    - high demands on concentration
    - left-only context, document context
  - Presentation options influence latency and readability (in re-translating)

Challenges in MT Evaluation

- Offline text-to-text MT:
  - MT quality $\rightarrow$ direct assessment (DA)
  - competent evaluators $\rightarrow$ bilinguals
  - different opinions $\rightarrow$ repetitions, statistics
- Offline ST: Speech as input modality
- Simultaneous ST:
  - Simultaneity
    - only one access to document $\rightarrow$ human memory
    - high demands on concentration
    - left-only context, document context
  - Presentation options influence latency and readability (in re-translating)

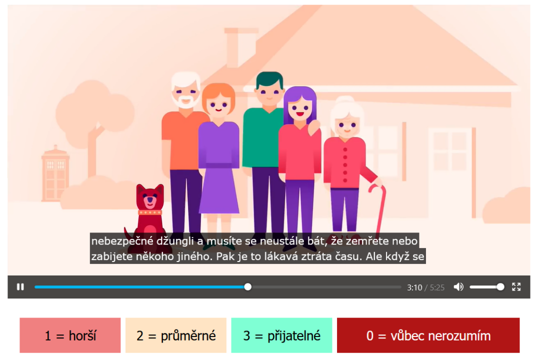
Let's simulate and collect ratings = **Continuous Rating**.

Continuous Rating



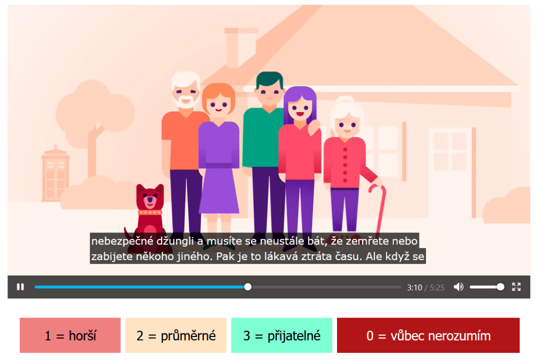
- Continuous Rating captures **current satisfaction** of users.

Continuous Rating



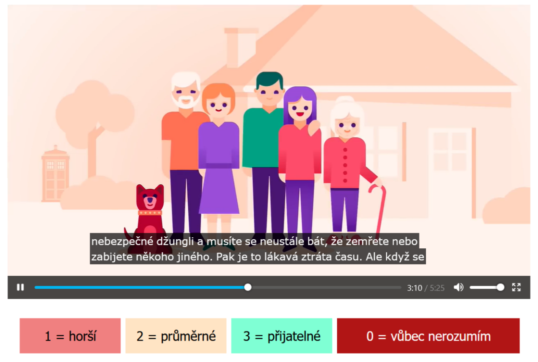
- Continuous Rating captures **current satisfaction** of users.
- 4 buttons below the audio/video document.

Continuous Rating



- Continuous Rating captures **current satisfaction** of users.
- 4 buttons below the audio/video document.
- The scores of the rating ranged between 0 (the worst) and 3 (the best).

Continuous Rating



- Continuous Rating captures **current satisfaction** of users.
- 4 buttons below the audio/video document.
- The scores of the rating ranged between 0 (the worst) and 3 (the best).
- First published by Macháček and Bojar, 2020. Presenting simultaneous translation in limited space. ITAT

Reliability of Continuous Rating (CR)

Javorský, Macháček, Bojar, Continuous Rating as Reliable Human Evaluation of Simultaneous Speech Translation, WMT 2022

- Let's evaluate CR on a downstream task: Comprehension. Factual questions:
  - **open style**, instead of yes/no or multiple choice
  - prepared from every **30 seconds** of the source document
  - evaluated manually against reference key

Reliability of Continuous Rating (CR)

Javorský, Macháček, Bojar, Continuous Rating as Reliable Human Evaluation of Simultaneous Speech Translation, WMT 2022

- Let's evaluate CR on a downstream task: Comprehension.

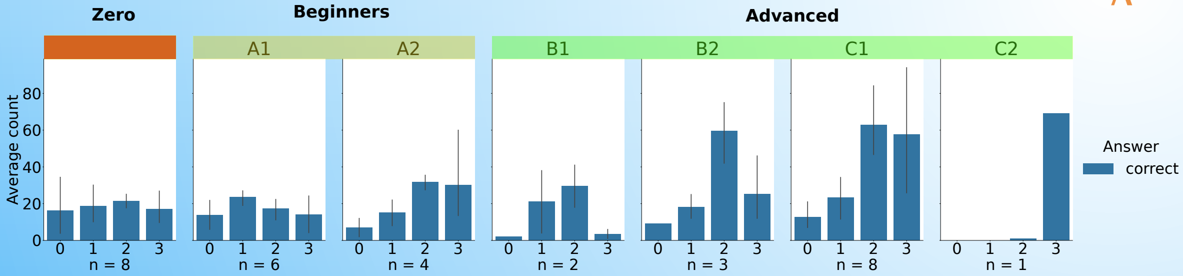
  Factual questions:
  - **open style**, instead of yes/no or multiple choice
  - prepared from every **30 seconds** of the source document
  - evaluated manually against reference key

  Collected feedback:
  - correct/partially correct/incorrect answer
  - or "unknown" answer (no guessing), or "forgot"
  - ...etc.

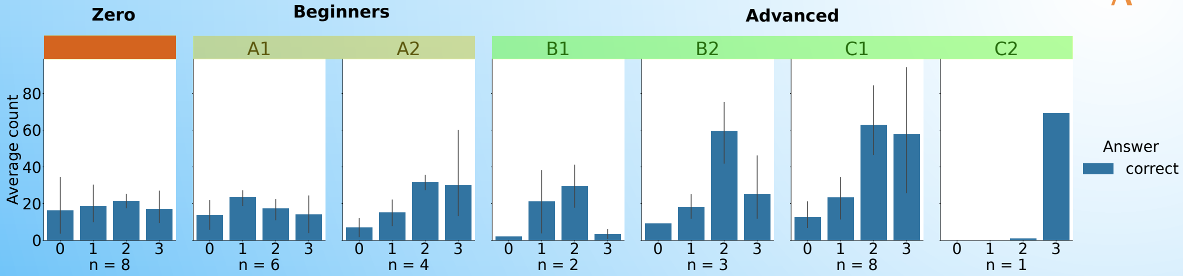
Documents, ST and judges

- German→Czech, one re-translating Sim. ST system
- 15 German docs., 5-10 min. each, 2h total, informative content, not too technical, audios or videos
- 32 judges, native Czech speakers, different levels of German proficiency

CR vs Answer Correctness



CR vs Answer Correctness



	χ^2-test p-values		
	Zero level	**Beginners**	**Advanced**
OK/OK-	0.24	$\mathbf{1.8 \cdot 10^{-5}}$	$\mathbf{5.6 \cdot 10^{-5}}$
unknown	0.033	$\mathbf{1.7 \cdot 10^{-4}}$	$\mathbf{9.1 \cdot 10^{-4}}$
wrong	0.59	0.45	$\mathbf{2.9 \cdot 10^{-3}}$
forgot	0.9	0.48	0.019

Table: **Bold**: CR and answers are **significantly dependent** ($p < 0.01$).

Conclusion

- CR means satisfaction with subtitling

Conclusion

- CR means satisfaction with subtitling
- advanced bilinguals:

Conclusion

- CR means satisfaction with subtitling
- advanced bilinguals:
  - CR is dependent to comprehension

Conclusion

- CR means satisfaction with subtitling
- advanced bilinguals:
  - CR is dependent to comprehension
  - they can listen to speech and rate adequacy by CR

Conclusion

- CR means satisfaction with subtitling
- advanced bilinguals:
  - CR is dependent to comprehension
  - they can listen to speech and rate adequacy by CR
- $\Rightarrow$ CR can be used to reliably assess satisfaction with subtitling (no questionnaires needed)

Other Results in Jávorský et al., WMT22

- Comprehension levels: Users understand a foreign lang. document...

Other Results in Jávorský et al., WMT22

- Comprehension levels: Users understand a foreign lang. document...
  - 33% with re-translating sim. ST

Other Results in Jávorský et al., WMT22

- Comprehension levels: Users understand a foreign lang. document...
  - 33% with re-translating sim. ST
  - 36% in sim. mode, without re-translations (oracle)

Other Results in Jávorský et al., WMT22

- Comprehension levels: Users understand a foreign lang. document...
  - 33% with re-translating sim. ST
  - 36% in sim. mode, without re-translations (oracle)
  - 59% (signif.) with MT in offline mode

Other Results in Jávorský et al., WMT22

- Comprehension levels: Users understand a foreign lang. document…
  - 33% with re-translating sim. ST
  - 36% in sim. mode, without re-translations (oracle)
  - 59% (signif.) with MT in offline mode
  - 81% (signif.) with MT in offline mode in a team
    (at least one of two persons is correct)

Other Results in Jávorský et al., WMT22

- Comprehension levels: Users understand a foreign lang. document...
  - 33% with re-translating sim. ST
  - 36% in sim. mode, without re-translations (oracle)
  - 59% (signif.) with MT in offline mode
  - 81% (signif.) with MT in offline mode in a team (at least one of two persons is correct)
  - 100% (assumingly) if the language is not foreign = no MT, offline access

Other Results in Jávorský et al., WMT22

- Comprehension levels: Users understand a foreign lang. document...
  - 33% with re-translating sim. ST
  - 36% in sim. mode, without re-translations (oracle)
  - 59% (signif.) with MT in offline mode
  - 81% (signif.) with MT in offline mode in a team (at least one of two persons is correct)
  - 100% (assumingly) if the language is not foreign = no MT, offline access
- Subtitling layout gives negligible difference on understanding

Other Results in Jávorský et al., WMT22

- Comprehension levels: Users understand a foreign lang. document...
  - 33% with re-translating sim. ST
  - 36% in sim. mode, without re-translations (oracle)
  - 59% (signif.) with MT in offline mode
  - 81% (signif.) with MT in offline mode in a team (at least one of two persons is correct)
  - 100% (assumingly) if the language is not foreign = no MT, offline access
- Subtitling layout gives negligible difference on understanding
- Advanced bilinguals understand more with low latency despite high flicker (sign.)

Other Results in Jávorský et al., WMT22

- Comprehension levels: Users understand a foreign lang. document...
  - 33% with re-translating sim. ST
  - 36% in sim. mode, without re-translations (oracle)
  - 59% (signif.) with MT in offline mode
  - 81% (signif.) with MT in offline mode in a team (at least one of two persons is correct)
  - 100% (assumingly) if the language is not foreign = no MT, offline access
- Subtitling layout gives negligible difference on understanding
- Advanced bilinguals understand more with low latency despite high flicker (sign.)
- Published: data, subtitler implem., evaluation campaign web app

05

MT Metrics Correlate with CR in Sim. Mode

Why CR vs MT Metrics?

Why CR vs MT Metrics?

- CR is expensive

Why CR vs MT Metrics?

- CR is expensive
- MT Metrics are designed for offline text-to-text MT, not Sim. ST. Differences:

Why CR vs MT Metrics?

- CR is expensive
- MT Metrics are designed for offline text-to-text MT, not Sim. ST. Differences:
  - no document context, no left only context

Why CR vs MT Metrics?

- CR is expensive
- MT Metrics are designed for offline text-to-text MT, not Sim. ST. Differences:
  - no document context, no left only context
  - no speech modality

Why CR vs MT Metrics?

- CR is expensive
- MT Metrics are designed for offline text-to-text MT, not Sim. ST. Differences:
  - no document context, no left only context
  - no speech modality
  - no simultaneity = one access, limited time

Why CR vs MT Metrics?

- CR is expensive
- MT Metrics are designed for offline text-to-text MT, not Sim. ST. Differences:
  - no document context, no left only context
  - no speech modality
  - no simultaneity = one access, limited time
  - no latency, reading comfort, frequency of updates

Why CR vs MT Metrics?

- CR is expensive
- MT Metrics are designed for offline text-to-text MT, not Sim. ST. Differences:
  - no document context, no left only context
  - no speech modality
  - no simultaneity = one access, limited time
  - no latency, reading comfort, frequency of updates
  - no other than translation aspects (if there are any)

Why CR vs MT Metrics?

- CR is expensive
- MT Metrics are designed for offline text-to-text MT, not Sim. ST. Differences:
  - no document context, no left only context
  - no speech modality
  - no simultaneity = one access, limited time
  - no latency, reading comfort, frequency of updates
  - no other than translation aspects (if there are any)
  - ...but **easy, cheap, automatic**

Why CR vs MT Metrics?

- CR is expensive
- MT Metrics are designed for offline text-to-text MT, not Sim. ST. Differences:
  - no document context, no left only context
  - no speech modality
  - no simultaneity = one access, limited time
  - no latency, reading comfort, frequency of updates
  - no other than translation aspects (if there are any)
  - ...but **easy, cheap, automatic**

MT Metrics are good for translation quality.

Why CR vs MT Metrics?

- CR is expensive
- MT Metrics are designed for offline text-to-text MT, not Sim. ST. Differences:
  - no document context, no left only context
  - no speech modality
  - no simultaneity = one access, limited time
  - no latency, reading comfort, frequency of updates
  - no other than translation aspects (if there are any)
  - ...but **easy, cheap, automatic**

MT Metrics are good for translation quality.

Translation is important in Sim. ST.

Why CR vs MT Metrics?

- CR is expensive
- MT Metrics are designed for offline text-to-text MT, not Sim. ST. Differences:
  - no document context, no left only context
  - no speech modality
  - no simultaneity = one access, limited time
  - no latency, reading comfort, frequency of updates
  - no other than translation aspects (if there are any)
  - ...but **easy, cheap, automatic**

MT Metrics are good for translation quality.

Translation is important in Sim. ST.

Can we replace CR by MT Metrics?

Why CR vs MT Metrics?

- CR is expensive
- MT Metrics are designed for offline text-to-text MT, not Sim. ST. Differences:
  - no document context, no left only context
  - no speech modality
  - no simultaneity = one access, limited time
  - no latency, reading comfort, frequency of updates
  - no other than translation aspects (if there are any)
  - ...but **easy, cheap, automatic**

MT Metrics are good for translation quality.

Translation is important in Sim. ST.

Can we replace CR by MT Metrics?

Why CR vs MT Metrics?

- CR is expensive
- MT Metrics are designed for offline text-to-text MT, not Sim. ST. Differences:
  - no document context, no left only context
  - no speech modality
  - no simultaneity = one access, limited time
  - no latency, reading comfort, frequency of updates
  - no other than translation aspects (if there are any)
  - ...but **easy, cheap, automatic**

MT Metrics are good for translation quality.

Translation is important in Sim. ST.

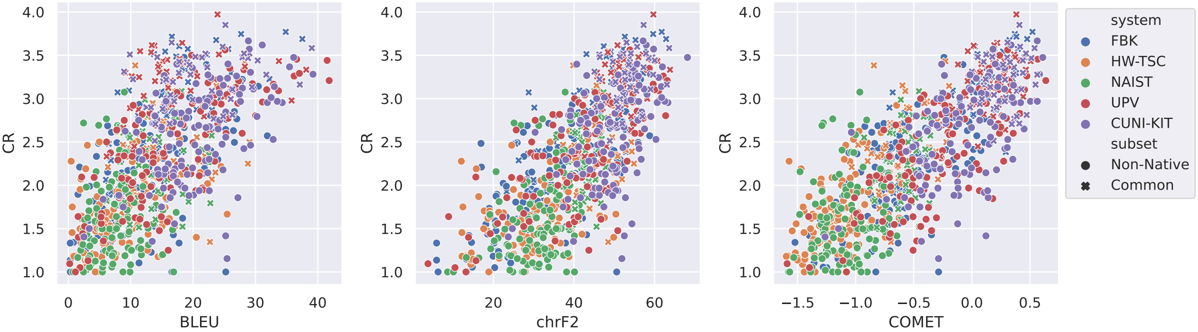
Can we replace CR by MT Metrics? If yes, why?

CR in IWSLT22 En-De Sim. ST

Macháček, Bojar, Dabre (2022): MT Metrics Correlate with Human Ratings of Simultaneous Speech Translation. `arxiv.org/abs/2211.08633`

- FBK, NAIST, UPV, HW-TSC, CUNI-KIT, each in 3 latency regimes
- 2 subsets: Common TED talks, Non-Native, 60 documents
- in total 900 document candidate translations
- 1584 rating sessions – each is one evaluator, one document, one system and latency candidate

Doc-Level Correlation



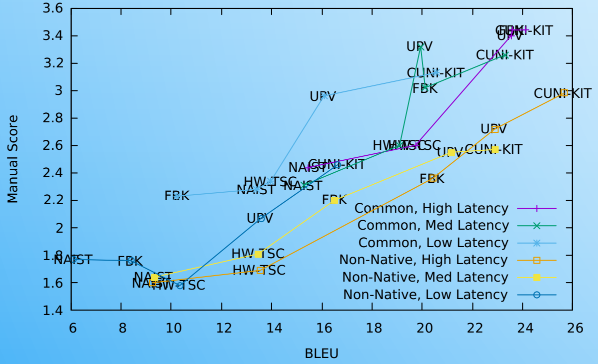
Doc-Level Correlation

Averaged document ratings

subsets	num.	BLEU	chrF2	COMET
both	823	0.65	0.73	0.80
Common	228	0.42	0.63	0.76
Non-native	595	0.70	0.70	0.75

Table: Pearson correlation coefficients for CR vs MT metrics BLEU, chrF2 and COMET for averaged document ratings by all 5 SST systems and 3 latency regimes. When the coefficient is less than 0.6 (in gray), the correlation is not considered as strong. Significance values are $p < 0.01$ in all cases, meaning strong confidence.

Test-Level Correlation



Slide from Ondřej Bojar, WMT22. Available in Findings IWSLT22.

- BLEU correlates very well with continuous rating on each test set part.
- Pearson across systems and latency regimes is:
  - .898 for the Common part.
  - .933 for the Non-native part.
  - .858 when considered together.

Conclusion on CR vs MT Correlation

- BLEU, chrF2 and COMET can be used to assess CR at least on the level of test sets
- COMET also on level of documents

Conclusion on CR vs MT Correlation

- BLEU, chrF2 and COMET can be used to assess CR at least on the level of test sets
- COMET also on level of documents

Limitations:

  - En-De only, 5 systems from IWSLT 2022 only
  - maybe future Sim. ST systems show divergence of CR and offline MT metrics

Remark: Unfair Comparison

- 900 document candidate transl., 823 document CR
  ⇒ some are not rated at all!
- It is **unfair** to put them on one scale:

System	Common Low	Common Medium	Common High	Non-native Low	Non-native Medium	Non-native High
CUNI-KIT	**3.13**	3.26	**3.44**	**2.46**	**2.57**	**2.98**
UPV	2.96	**3.32**	3.40	2.07	2.55	2.72
FBK	2.23	3.02	**3.44**	1.76	2.20	2.36
HW-TSC	2.34	2.60	2.60	1.58	1.81	1.69
NAIST	2.28	2.31	2.44	1.77	1.64	1.60
Avg±Std.d.	2.59±0.38	2.90±0.39	3.06±0.45	1.93±0.31	2.15±0.38	2.27±0.55
Interpreting		2.99			3.22	

Table: Test-level aggregated En-De CR scores from IWSLT22 Findings. It is unfair comparison because it is not ensured that all systems are rated on the same documents.

Next Ideas with CR from IWSLT22

- Reference for Sim. ST: translation, or interpreting?

Next Ideas with CR from IWSLT22

- Reference for Sim. ST: translation, or interpreting?
  What correlates better to CR.

Next Ideas with CR from IWSLT22

- Reference for Sim. ST: translation, or interpreting?
  What correlates better to CR.
  Probably depends on domain.

Next Ideas with CR from IWSLT22

- Reference for Sim. ST: translation, or interpreting?
  What correlates better to CR.
  Probably depends on domain.
- Candidate-reference sent. segm. mismatch – how to use BLEU and chrF2?

Next Ideas with CR from IWSLT22

- Reference for Sim. ST: translation, or interpreting?
  What correlates better to CR.
  Probably depends on domain.
- Candidate-reference sent. segm. mismatch – how to use BLEU and chrF2?
  mwerSegmenter? doc-level sequences?

Next Ideas with CR from IWSLT22

- Reference for Sim. ST: translation, or interpreting?
  What correlates better to CR.
  Probably depends on domain.
- Candidate-reference sent. segm. mismatch – how to use BLEU and chrF2?
  mwerSegmenter? doc-level sequences?
  What correlates better to CR.

06
Summary

Summary

- Book: The reality of Multi-lingual MT
- **Simultaneous Multi-Source Speech Translation**
- Continuous Rating, MT Metrics Correlate in Sim. Mode

Summary

- Book: The reality of Multi-lingual MT
- **Simultaneous Multi-Source Speech Translation**
- Continuous Rating, MT Metrics Correlate in Sim. Mode

User satisfaction

Summary

- Book: The reality of Multi-lingual MT
- **Simultaneous Multi-Source Speech Translation**
- Continuous Rating, MT Metrics Correlate in Sim. Mode

User satisfaction > Comprehension questionnaires

Summary

- Book: The reality of Multi-lingual MT
- **Simultaneous Multi-Source Speech Translation**
- Continuous Rating, MT Metrics Correlate in Sim. Mode

User satisfaction > Comprehension questionnaires >
> Continuous Rating

Summary

- Book: The reality of Multi-lingual MT
- **Simultaneous Multi-Source Speech Translation**
- Continuous Rating, MT Metrics Correlate in Sim. Mode

User satisfaction > Comprehension questionnaires >
> Continuous Rating > MT Metrics

Thank you!