European Live Translator

Ondřej Bojar; Sep 16, Pec pod Sněžkou

H2020 RIA ELITR (2019–2021)

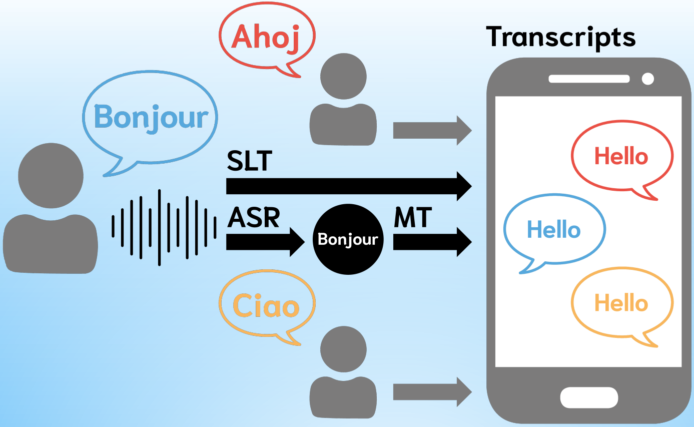
ELITR (European Live Translator) aims:

- Highly multi-lingual machine and speech translation.
- Document-level machine translation.
- Automatic meeting summarization, “Minuting”.

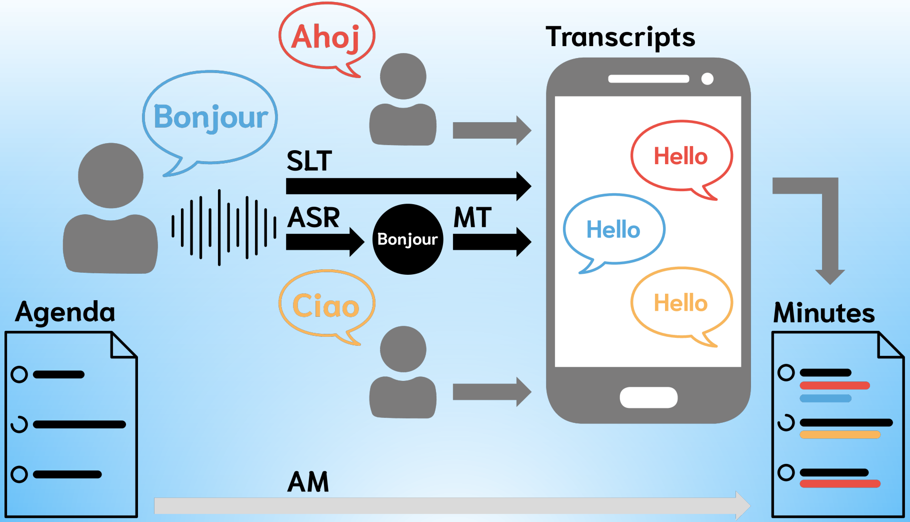
. . . and the Supreme Audit Office of the Czech Republic as affiliated user partner.

Main ELITR event: Interpreting at EUROSAI Congress (Prague, May 2020).

ELITR Technologies: ASR, MT, SLT, Minuting



ELITR Technologies: ASR, MT, SLT, Minuting



ELITR Languages

Languages to be supported by the ELITR project:

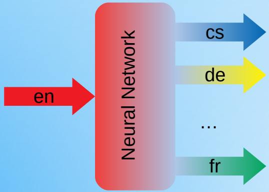
Technology	Primary Focus	Covered	Experimental
ASR	En, De	Fr, Sp, It, Ru	Cs
SLT & MT	{ En, De }	all EU languages	all EUROSAI langs
	→ { En, De, Cs }	→ all EU languages	→ all EUROSAI langs
Summarization	English, Czech	–	–

24 EU languages: Bulgarian, Croatian, Czech, Danish, Dutch, English, Estonian, Finnish, French, German, Greek, Hungarian, Irish, Italian, Latvian, Lithuanian, Maltese, Polish, Portuguese, Romanian, Slovak, Slovene, Spanish, and Swedish.
Further 19 EUROSAI languages: Albanian, Arabic, Armenian, Azerbaijani, Belorussian, Bosnian, Georgian, Hebrew, Icelandic, Kazakh, Luxembourgish, Macedonian, Moldovan, Montenegrin, Norwegian, Russian, Serbian, Turkish, and Ukrainian.

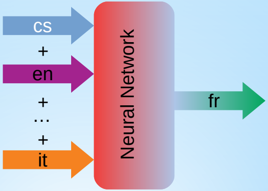
MT Challenges (WP4)

- Document-level translation:
  - Lexical and structural coherence.
  - Context perhaps even more important for speech translation.
  - **Very hard to evaluate; errors too diverse and not frequent enough**
- High multi-linguality:

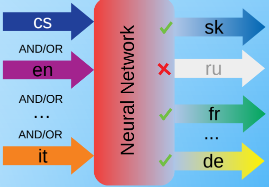
Multi-target
(Efficiency)



Multi-source
(Ambiguity resolution)



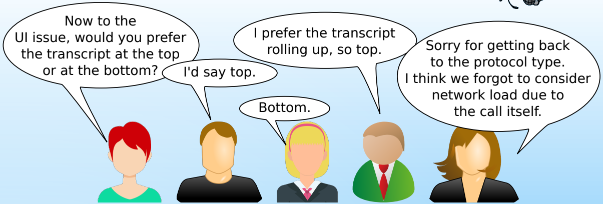
Flexible
Multilingual

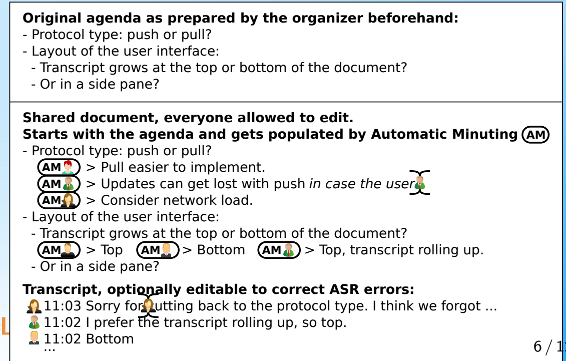


Meeting Summarization (“Minuting”; WP5)

High-risk goal:

- ▶ Given agenda and transcript.
- ▶ Populate the agenda with items from the transcript to obtain **meeting minutes**.
- = Structured summarization of discussions.
- ▶ “De-duplicating” rather than reducing content.
- ▶ Critical: Gather such data.


Now to the UI issue, would you prefer the transcript at the top or at the bottom?
I'd say top.
Bottom.
I prefer the transcript rolling up, so top.
Sorry for getting back to the protocol type. I think we forgot to consider network load due to the call itself.



Original agenda as prepared by the organizer beforehand:

- Protocol type: push or pull?
- Layout of the user interface:
  - Transcript grows at the top or bottom of the document?
  - Or in a side pane?

Shared document, everyone allowed to edit.
Starts with the agenda and gets populated by Automatic Minuting (AM)

- Protocol type: push or pull?
  - (AM) > Pull easier to implement.
  - (AM) > Updates can get lost with push *in case the user*
  - (AM) > Consider network load.
- Layout of the user interface:
  - Transcript grows at the top or bottom of the document?
    (AM) > Top (AM) > Bottom (AM) > Top, transcript rolling up.
  - Or in a side pane?

Transcript, optionally editable to correct ASR errors:

11:03 Sorry for getting back to the protocol type. I think we forgot ...
11:02 I prefer the transcript rolling up, so top.
11:02 Bottom
...

ELITR Key Event: EUROSAI Congress 2020

- The main congress of supreme audit institutions (EUROSAI).
- May 31 to June 5, 2020.
- 200-250 people from 50 different countries will attend.

This event has two radically different “mode”s:

- Plenary sessions: More languages, one room.
- Workshops: More rooms, one language.

We run our workshop on NLP tools for auditors.

ÚFAL Focus (1/2)

- Ensure that everything works:
- Ondřej, Dominik→Sangeet Sagar (intern)
- MT (multi-lingual, multi-source): Dominik
- MT (multi-target): Bohdan Ihnatchenko (MFF Mgr)
- MT (domain adaptation): Daniel Suchý (FF Mgr)
- MT (doc-level): Tomáš Musil
    - **Still searching for people!!**
- ASR: Jonáš Kratochvíl (MFF Mgr)
- SLT (simultaneous translation): Ebrahim Ansari

ÚFAL Focus (2/2)

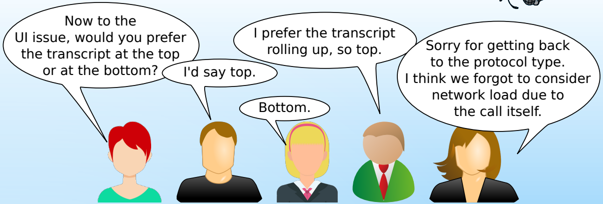
- Minuting (corpus): Anja, Ondřej, 2 annotators
  - **More data and annotators needed.**
- Minuting (models): Erion (over the summer, we had Ahmad)
- Minuting (segmentation): Michal Auersperger
- Ethics (incl. GDPR): Libuška Kaprálová (external)
- Dissemination (not only MT, but all NLP): Tea
  - **We need people to help us popularize our tools in May.**
- Ensure that people deliver what they have to: **Tea**

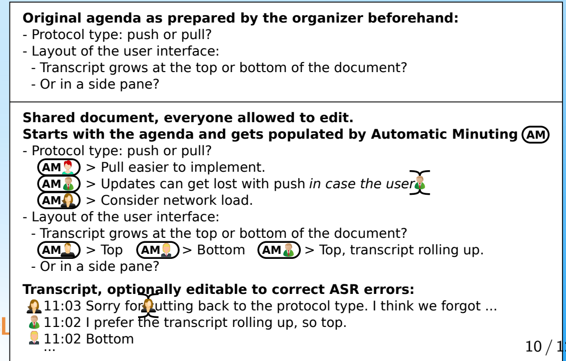
ELITR Needs Meeting Data for Minuting

We need:

- Audio and/or Transcripts.
- Agenda
  - Ideally hierarchical.
- Minutes
  - Ideally the Agenda populated with items.
- Any supplementary files:
  - Slides, related reports, ...

Please contact us at `info@elitr.eu`.


Now to the UI issue, would you prefer the transcript at the top or at the bottom?
I'd say top.
Bottom.
I prefer the transcript rolling up, so top.
Sorry for getting back to the protocol type. I think we forgot to consider network load due to the call itself.



Original agenda as prepared by the organizer beforehand:
- Protocol type: push or pull?
- Layout of the user interface:
  - Transcript grows at the top or bottom of the document?
  - Or in a side pane?

Shared document, everyone allowed to edit.
Starts with the agenda and gets populated by Automatic Minuting AM
- Protocol type: push or pull?
  - AM > Pull easier to implement.
  - AM > Updates can get lost with push *in case the user*
  - AM > Consider network load.
- Layout of the user interface:
  - Transcript grows at the top or bottom of the document?
    AM > Top AM > Bottom AM > Top, transcript rolling up.
  - Or in a side pane?

Transcript, optionally editable to correct ASR errors:
11:03 Sorry for getting back to the protocol type. I think we forgot ...
11:02 I prefer the transcript rolling up, so top.
11:02 Bottom
...

Thank you!