Comprehensive Evaluation of Spoken Language Translation

Ansari E, Bojar O, Haddow B, Mahmoudi M
Spoken Language Translation
Budeme si o projektu divadelní hry, o kterou...
We'll be talking about a project of a theatre play which...
We'll be talking about a project of a theatre play which...
We'll be talking about a project of a theatre play which...
We'll be talking about a project of a theatre play which...
We'll be talking about a project of a theatre play which...
Spoken Language Translation with Revisions

**SRC**
Budeme si povídat o projektu divadelní hry kterou...

**HYP**
We will talk about a project of a theatre play which...

**REF**
We'll be talking about a project of a theatre play which...
Spoken Language Translation with Revisions

SRC
Budeme si o projektu divadelní hry kterou...

HYP
We will
We will talk
We will talk about a programme
We will talk about a project
We will talk about a play project

REF
We'll be talking about a project of a theatre play which...
Spoken Language Translation Errors

We'll be talking about a project of a theatre play which...

We will talk about a project
We will talk about a programme
We will talk about a project
We will talk about a play project
We will talk about a programme
We will talk about a project

We'll be talking about a project of a theatre play which...
We'll be talking about a project of a theatre play which...
We'll be talking about a project of a theatre play which...
SLTev – Comprehensive Evaluation of SLT

https://github.com/ELITR/SLTev

Installation:
- pip3 install SLTev

Usage:
- Provide timestamped logs, no server-client API.

SLTev evaluates:
- Quality = BLEU or chrF3 (via sacreBLEU).
- Delay = How long we had to wait beyond necessary word reordering.
- Flicker = How much of our reading was wasted.
SLTev Comes with elitr-testset

- Developed by the EU project ELITR.
  (See our other demo at EACL 2021: ELITR Multilingual Subtitling.)
- elitr-testset focuses on:
  - Multi-lingual ASR, MT and (simultaneous) SLT.
  - Continuous growth and yet reproducibility:
    - SLTev emits fingerprints with git commit IDs.
  - Non-native speakers and not only English source.
  - Realistic (i.e. bad) sound quality.
  - Diverse domains:
    - auditing, computational linguistics, NLP . . .
    - oral history, debates on AI, student’s mock business presentations . . .
Structure of elitr-testset

To facilitate growth and variability, elitr-testset consists of:

- Assorted collection of documents.
- Collection of indices of these documents.
Structure of elitr-testset

- Assorted collection of **documents**.
  - Each document comes in several files, with different modalities, e.g.:
    - Original sound.
    - Original Speech Transcribed (OSt).
    - Original Speech Transcribed with Timestamps (OStt).
    - Text-based Translation into a target language (TT, e.g. en.TTcs).
    - sometimes even Interpreter’s Speech transcribed (IS, IST, ISTt).

- Collection of **indices** of these documents.
Structure of elitr-testset

- Assorted collection of documents.
  - Each document comes in several files, with different modalities, e.g.:
    - Original sound.
    - Original Speech Transcribed (OSt).
    - Original Speech Transcribed with Timestamps (OStt).
    - Text-based Translation into a target language (TT, e.g. en.TTcs).
    - sometimes even Interpreter’s Speech transcribed (IS, IST, ISTt).

- Collection of indices of these documents.
  - An index is a set of documents useful for a particular evaluation purpose.
  - Sample indices:
    - Evaluation of Czech off-line speech recognition.
    - Evaluation of English-to-Czech simultaneous speech-to-text translation.
    - Evaluation of English-to-Serbian text-only translation.
  - An index defines what are source and target/reference modalities and files.
Using elitr-testset

- Browse on github:
  
  https://github.com/ELITR/elitr-testset
Using elitr-testset

- Browse on github:
  
  https://github.com/ELITR/elitr-testset

- Use SLTev to evaluate an index for you:
  1. Obtain all documents for a given index, e.g. “SLTev-sample“ index:
     SLTev -g SLTev-sample --outdir mytestdir
  2. Run your system on files in mytestdir.
  3. Run SLTev to get the scores:
     SLTev -e mytestdir

Useful options:
  - -T your-clone-of-elitr-testset ... so that SLTev does not download it
  - --aggregate ... to aggregate scores over individual files
  - --simple ... to show the most common scores only

Anything unclear or wrong? Please create github issues.
Using elitr-testset

- Browse on github:
  https://github.com/ELITR/elitr-testset

- Use SLTev to evaluate an index for you:
  1. Obtain all documents for a given index, e.g. “SLTev-sample“ index:
     SLTev -g SLTev-sample --outdir mytestdir
  2. Run your system on files in mytestdir.
  3. Run SLTev to get the scores:
     SLTev -e mytestdir

Useful options:
  -T your-clone-of-elitr-testset ... so that SLTev does not download it
  --aggregate ... to aggregate scores over individual files
  --simple ... to show the most common scores only

- Anything unclear or wrong? Please create github issues.
Summary

SLTev would like to be the “sacreBLEU” for (simultaneous) SLT.

- Open-source, pip-installed.
- Estimates quality, delay and flicker.
- Allows output revisions.

SLTev is complemented by elitr-testset:

- Diverse collection of speech, transcripts, translations, interpretations.
- Versioned index files to allow for reproducible benchmarking.

Enjoy and contribute!
- Send pull requests, file issues.
- Report bugs, fix existing data, donate new documents.
- Add indices for your papers, for and easy comparison.
Summary
SLTev would like to be the “sacreBLEU” for (simultaneous) SLT.
▶ Open-source, pip-installed.
▶ Estimates **quality**, **delay** and **flicker**.
▶ Allows **output revisions**.
SLTev is complemented by elitr-testset:
▶ Diverse collection of speech, transcripts, translations, interpretations.
▶ Versioned **index files** to allow for reproducible benchmarking.
Summary

SLTev would like to be the “sacreBLEU” for (simultaneous) SLT.

- Open-source, pip-installed.
- Estimates **quality**, **delay** and **flicker**.
- Allows **output revisions**.

SLTev is complemented by elitr-testset:

- Diverse collection of speech, transcripts, translations, interpretations.
- Versioned **index files** to allow for reproducible benchmarking.

Enjoy and contribute!

- Send pull requests, file issues.
- Report bugs, fix existing data, donate new documents.
- Add indices for your papers, for easy comparison.