# The Interpreter Understands Your Meaning: End-to-end Spoken Language Understanding Aided by Speech Translation

Mutian He[1,2], Philip N. Garner[1]
[1] Idiap Research Institute [2] EPFL On Findings of EMNLP'23

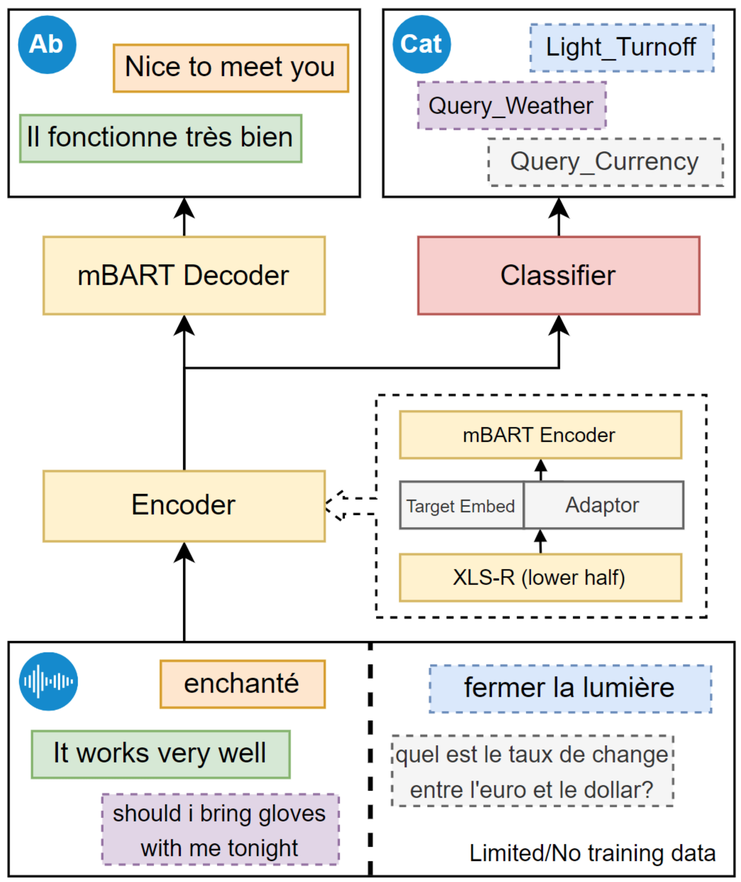
Code, Data, Model

## Background

- Trend towards difficult & informative LM pretraining task
  - From MLM to whole-word-masking, span MLM, token permutation, text infilling, …
- Goal: capture more sophisticated/high-level semantics
  - Speech models like Wav2Vec2 are pretrained by low-level acoustic modelling
  - Often further fine-tuned on ASR before used on SLU
  - … in order to connect with the more information-dense textual space
- PTLMs like CoVe and T5 use supervised pretraining

## Motivation & Goal

- A desirable pretraining task for speech models can help …
  - Understand high-level meanings
  - Capture not only local information
  - Enable multilinguality by language-agnostic representations
- Speech translation!




## Methods

- Pretrain on Speech Translation (ST)
  - Common arch: XLS-R (mWav2Vec2) + mBART
  - On, e.g., En↔Fr ST or/and En+Fr ASR
- Transfer to downstream tasks
  - Reusing encoder
- Cross-lingual transfer to, e.g. French
  - From English downstream tasks, with zero/few-shots

## Knowledge Preservation

- ST knowledge is valuable!
  - Preserve it by joint/multitask training
    - Computational intensive, require source data
  - Bayesian regularizers: L2-SP & EWC
    - Adaptively limit parameter shift from pretraining during fine-tuning, as used for continual learning

## Downstream Tasks

- Intent classification
  - English IC: SLURP benchmark
  - Multilingual IC: MINDS-14 benchmark
    - Low-resource (~600 utterances) per language; use English variants & French
  - Cross-lingual IC: SLURP-Fr & -Es
    - New dataset: Synthetic + Real samples in French & Spanish, based on MASSIVE
- Spoken QA: NMSQA
- Speech summarization: Spoken Gigaword
  - New synthetic dataset

## Results

- On SLURP and MINDS-14
  - ST leads to better results cf. ASR pretraining
  - Joint training of both pretraining and target tasks help

| Pretraining task | en-AU | en-GB | en-US | fr-FR | Average |
|---|---|---|---|---|---|
| ASR | 95.7% | 97.3% | 96.5% | 95.2% | 96.2% |
| w/ Joint training | 96.3% | 98.3% | 98.2% | 93.7% | 96.6% |
| ST | 96.9% | **99.0%** | 98.2% | 97.8% | 98.0% |
| w/ Joint training | **97.3%** | 98.7% | **99.3%** | 98.2% | **98.3%** |
| ST+ASR | 95.4% | 98.3% | 97.5% | 95.6% | 96.7% |
| w/ Joint training | 96.3% | 98.3% | 98.9% | **98.5%** | 98.0% |
| XLS-R (Lozhkov, 2022) | 92.4% | 93.2% | 93.3% | 94.4% | 93.3% |

Table 3: Test accuracies for models on MINDS-14 multilingual IC, comparing with directly fine-tuning the full XLS-R model. Both ST pretraining and joint training show benefits.

- On cross-lingual IC to SLURP-Fr
  - ST leads to much better results for transfer to French in 100- or zero-shot, possibly with additional language adversarial training

| | Full | | | 100-shot | | | zero-shot | | |
|---|---|---|---|---|---|---|---|---|---|
| Pretraining task | *dev* | *test* | *real* | *dev* | *test* | *real* | *dev* | *test* | *real* |
| ASR | 83.3% | 84.0% | 79.7% | 69.6% | 69.0% | 71.3% | 39.0% | 39.8% | 39.4% |
| ST | 85.2% | 86.1% | 84.9% | 78.1% | 77.0% | 79.0% | 58.9% | 58.9% | 56.6% |
| ST+ASR | 85.8% | 85.7% | 82.4% | 78.0% | 77.0% | 78.8% | 63.9% | 62.6% | 59.1% |
| ST+Adv. | 86.4% | 84.9% | 84.1% | 78.3% | 78.1% | 80.9% | 67.0% | 67.7% | 63.7% |
| None | 75.9% | 74.0% | 65.4% | 57.5% | 52.4% | 53.5% | 15.3% | 16.1% | 13.6% |
| Cascaded | | — | | | — | | 66.2% | 67.0% | 62.2% |

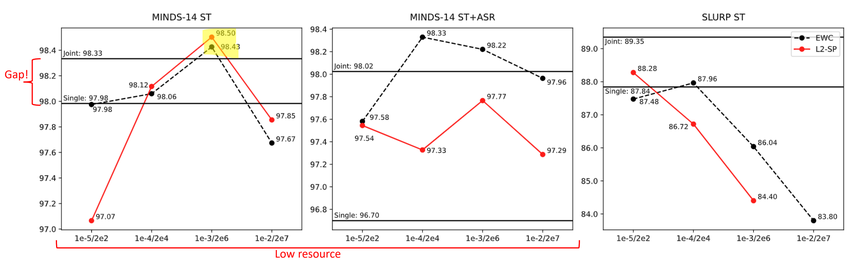
Table 4: Accuracies for models on SLURP-Fr cross-lingual IC transferred from SLURP with different amounts of data, accompanied with a cascaded system. Results highlight ST pretraining and language adversarial training.

- Similar improvements on spoken QA and summarization
  - Consistent improvements brought by ST
  - ST+ASR more helpful on summarization

| Pretraining task | *dev* | *test* |
|---|---|---|
| ASR | 54.6% | 53.0% |
| ST | **58.2%** | **59.4%** |
| ST+ASR | 57.8% | 58.0% |
| DUAL (Lin et al., 2022) | 48.5% | 49.1% |
| Cascaded | 58.3% | 57.4% |

Table 5: AOS (↑) scores for models on NMSQA. The pretraining tasks prove helpful, particularly ST pretraining, which outperforms the cascaded system.

- Model-preserving Bayesian transfer
  - Multi-task joint training already shows better results
  - EWC/L2-SP regularizers can also catch up with the joint training results and well preserve ST pretraining knowledge
    - In low resource cases, with appropriate weights




## Takeaway

- ST is a powerful pretraining means for E2E SLU
  - With strong multilinguality
- Preserving the ST knowledge during fine-tuning helps
  - Bayesian model-preserving methods works
- Two new datasets for E2E SLU