FONDAZIONE
BRUNO KESSLER

HLT
MACHINE
TRANSLATION

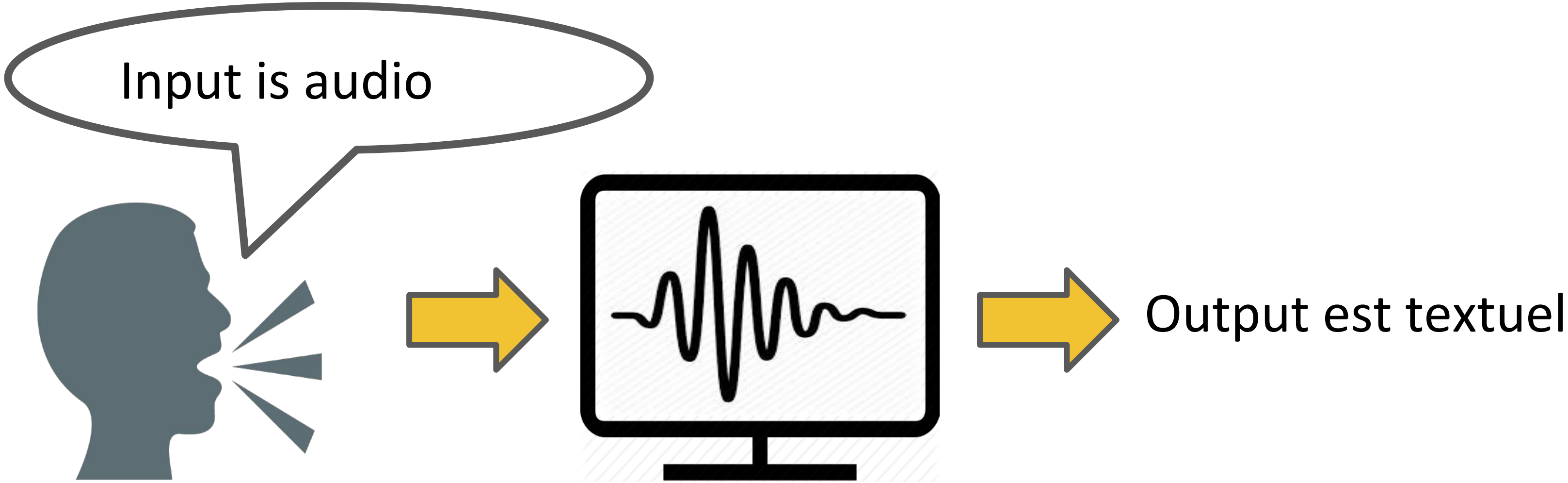
UNIVERSITY
OF TRENTO - Italy

Instance-Based Model Adaptation For Direct Speech Translation

Mattia Antonino Di Gangi, Viet-Nhat Nguyen, Matteo Negri, Marco Turchi

ICASSP 2020
Virtual, 7 May 2020

Speech Translation

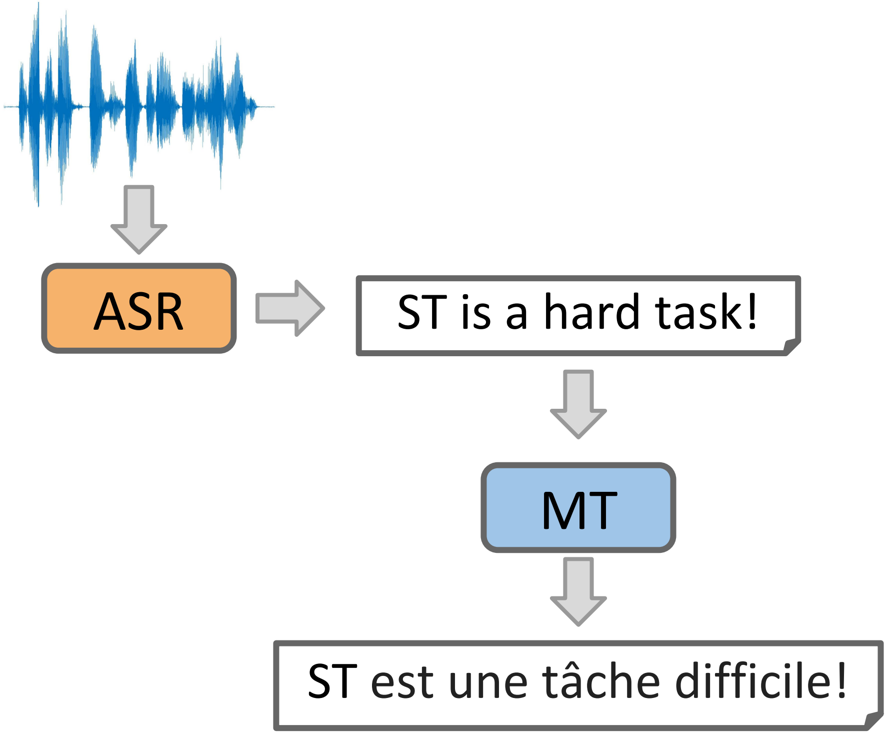


Input is audio

Output est textuel

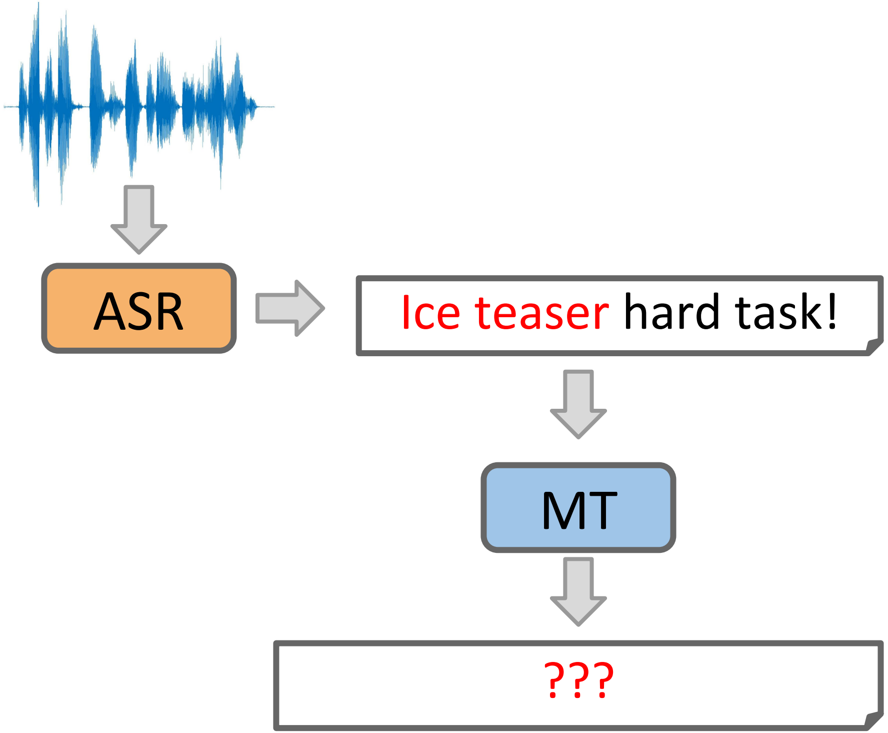
Classic approach: cascade

A pipeline of components



Cascade: limitations

A pipeline of components



ASR

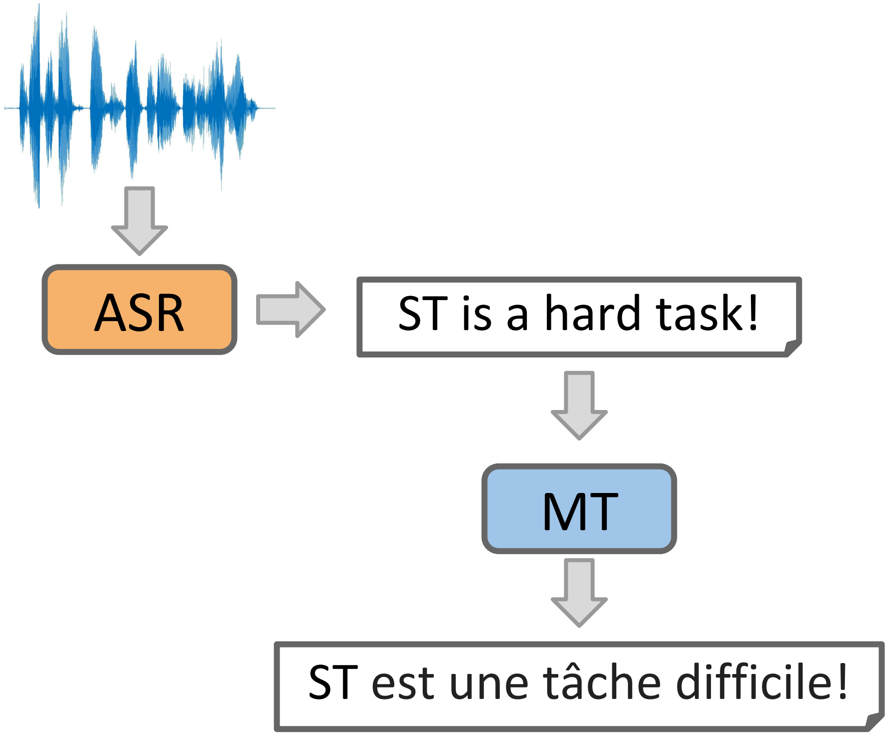
Ice teaser hard task!

MT

???

Emerging approach: end-to-end

A pipeline of components



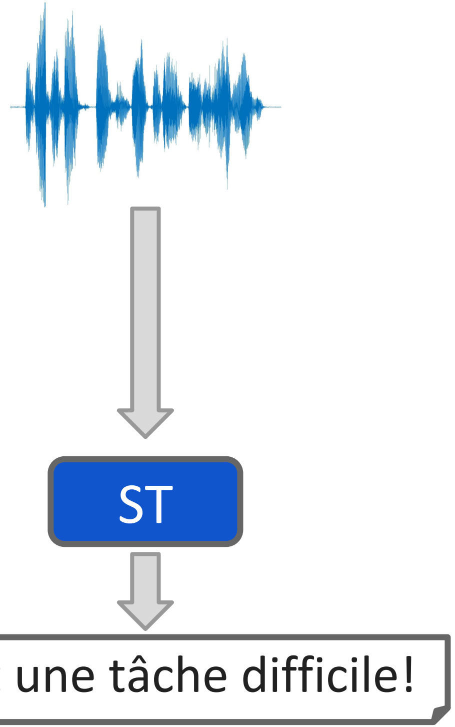
ASR

ST is a hard task!

MT

ST est une tâche difficile!

Direct, sequence to sequence



ST

ST est une tâche difficile!

Problem: small data available

- Strong MT models trained on tens of millions sentence pairs
- Strong ASR models trained on thousands of hours of speech
- Only few hundreds hours of speech for direct ST in the best cases

⇨ Target vocabulary is limited

⇨ Few samples for each variety of dialects, voices, noise conditions

⇨ Reduced generalization capabilities!

Motivation: better use of available data

- Datasets are small but diverse
- Can have few samples similar to the test sample
- Exploit similarity between test and training samples!

Instance-Based Model Adaptation (IBMA)

- Previously used in NMT for on-the-fly domain adaptation
- Idea:
    - Given an input sentence *s* to be translated, fine-tune the model on training pairs with *source text* similar to *s*
- Rationale:
    - Fine-tuning on similar data = “positive” overfitting for on-the-fly model adaptation

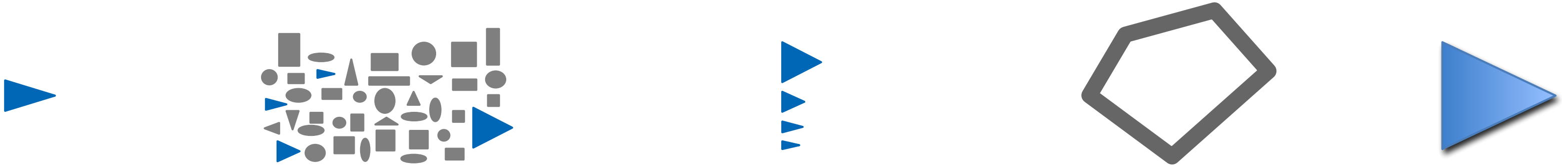
Src Audio

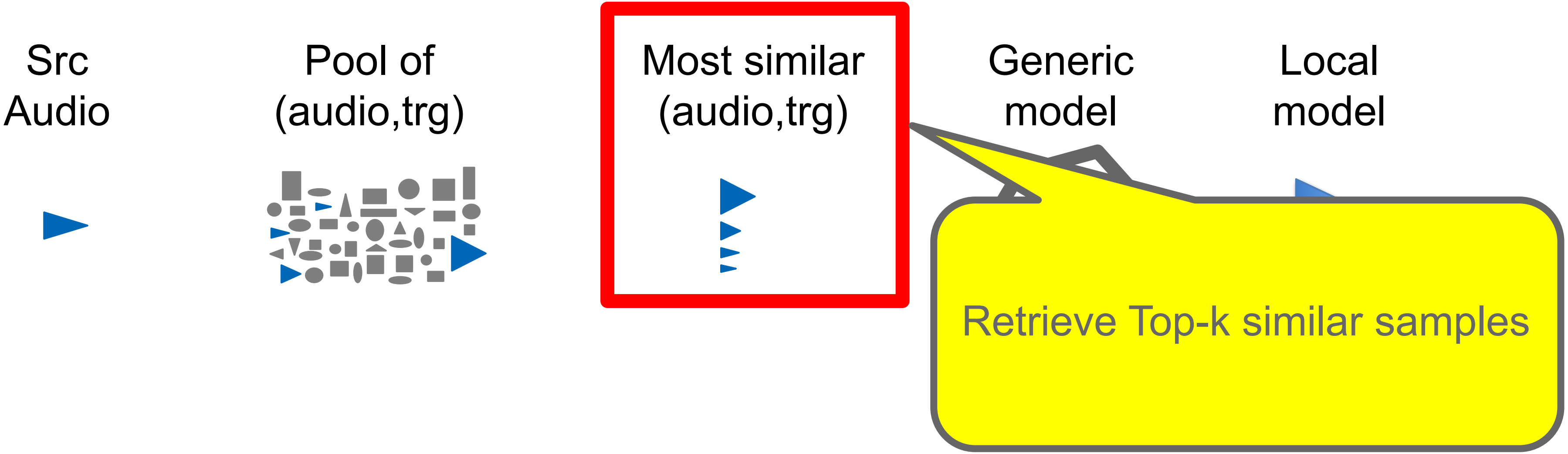
Pool of (audio,trg)

Most similar (audio,trg)

Generic model

Local model





Src Audio

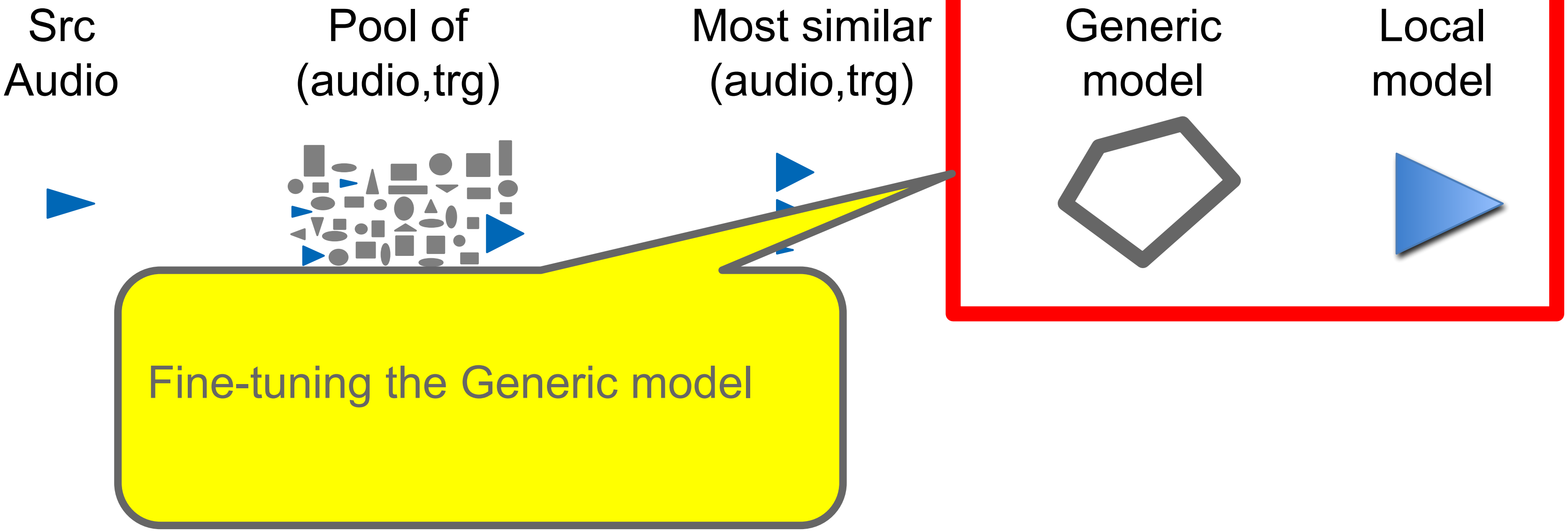
Pool of (audio,trg)

Most similar (audio,trg)

Generic model

Local model

Retrieve Top-k similar samples



Src Audio

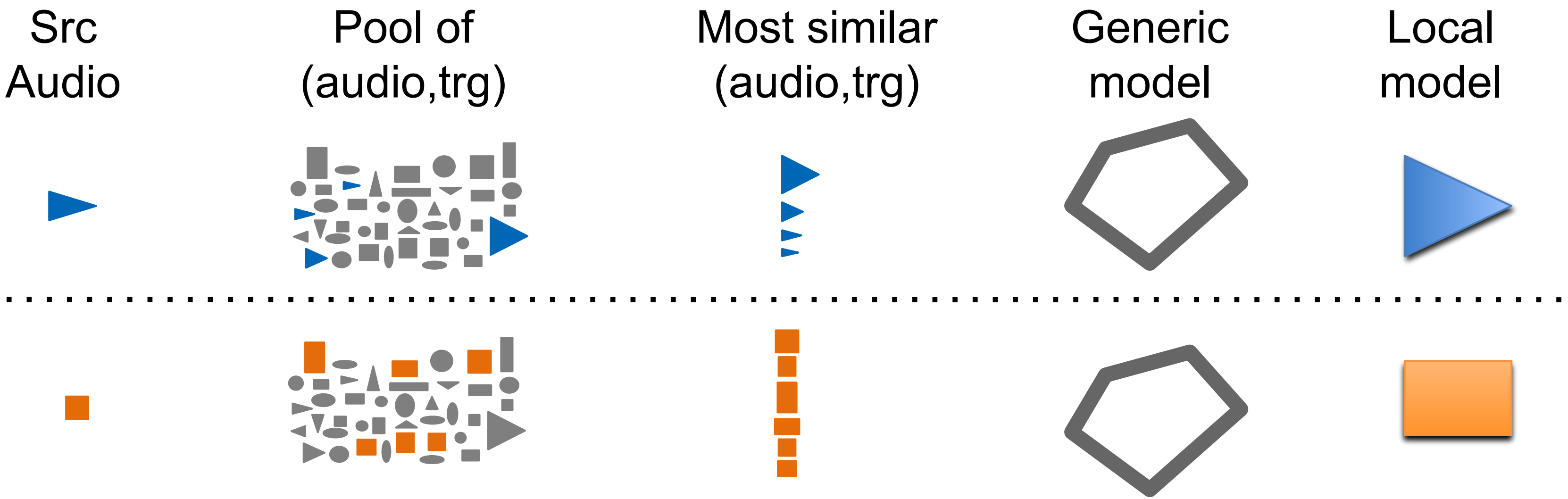
Pool of (audio,trg)

Most similar (audio,trg)

Generic model

Local model

Fine-tuning the Generic model



Src Audio

Pool of (audio,trg)

Most similar (audio,trg)

Generic model

Local model

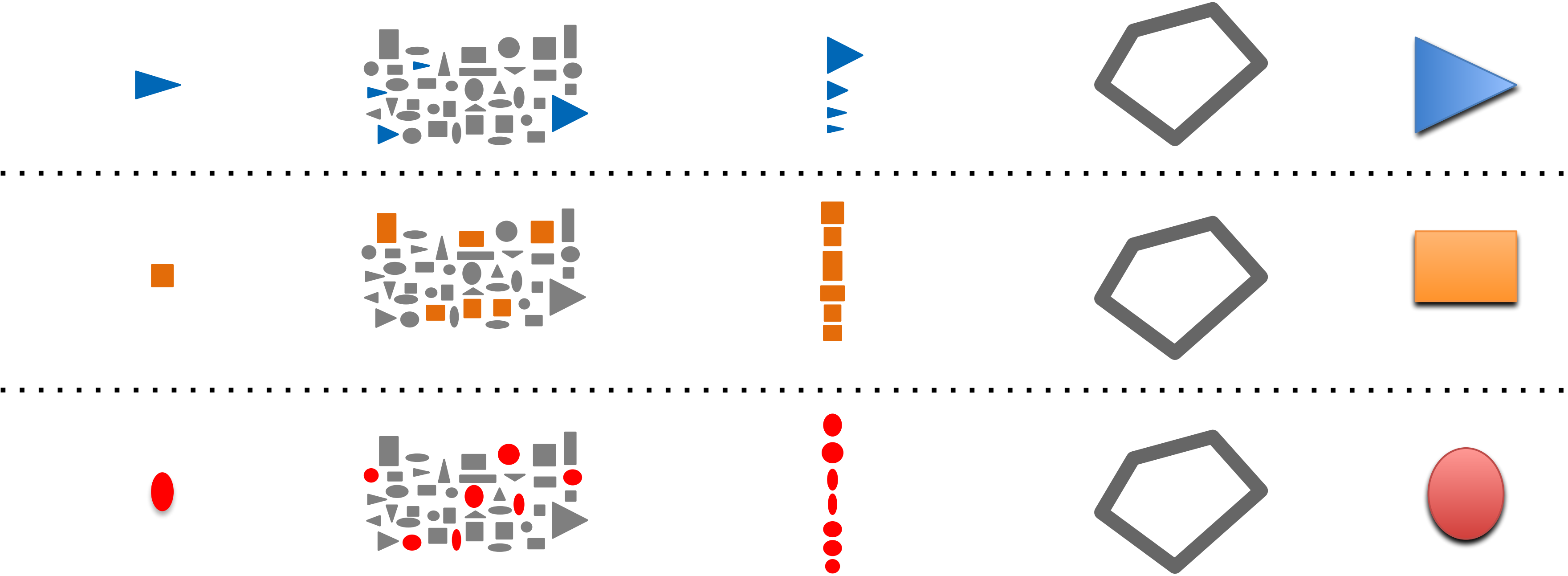
Src Audio

Pool of (audio,trg)

Most similar (audio,trg)

Generic model

Local model



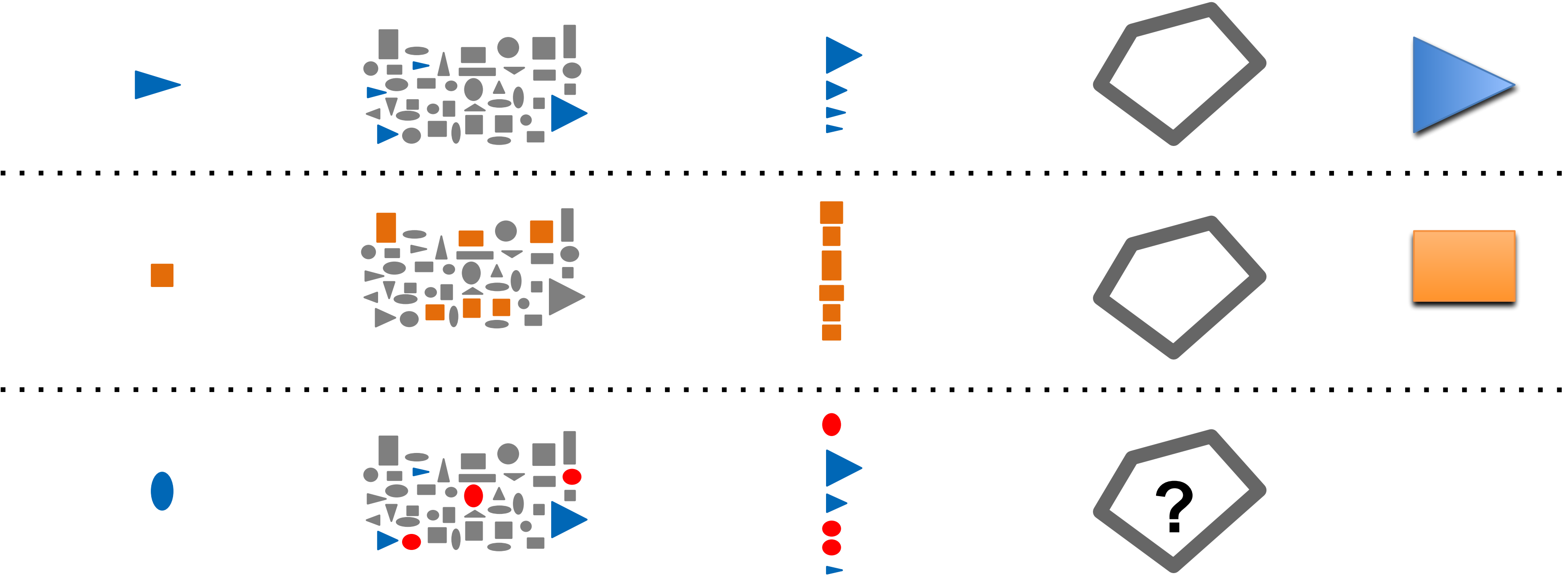
Src Audio

Pool of (audio,trg)

Most similar (audio,trg)

Generic model

Local model



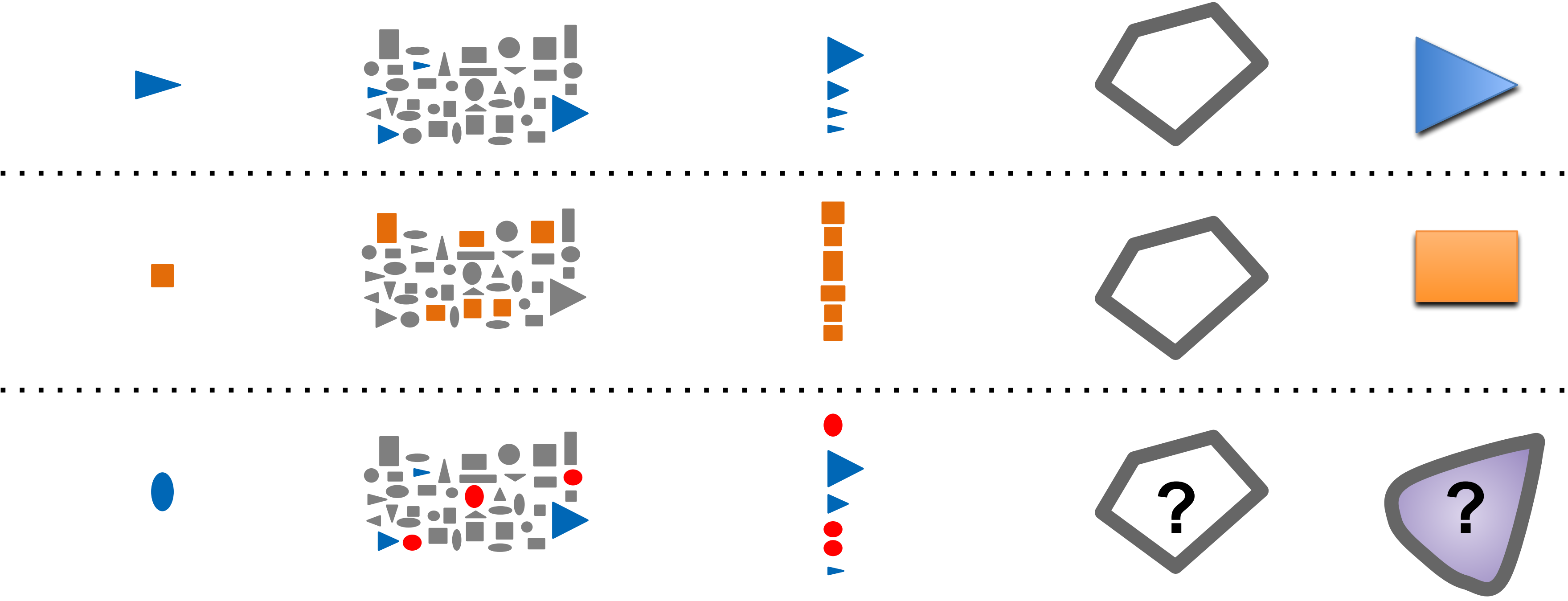
Src Audio

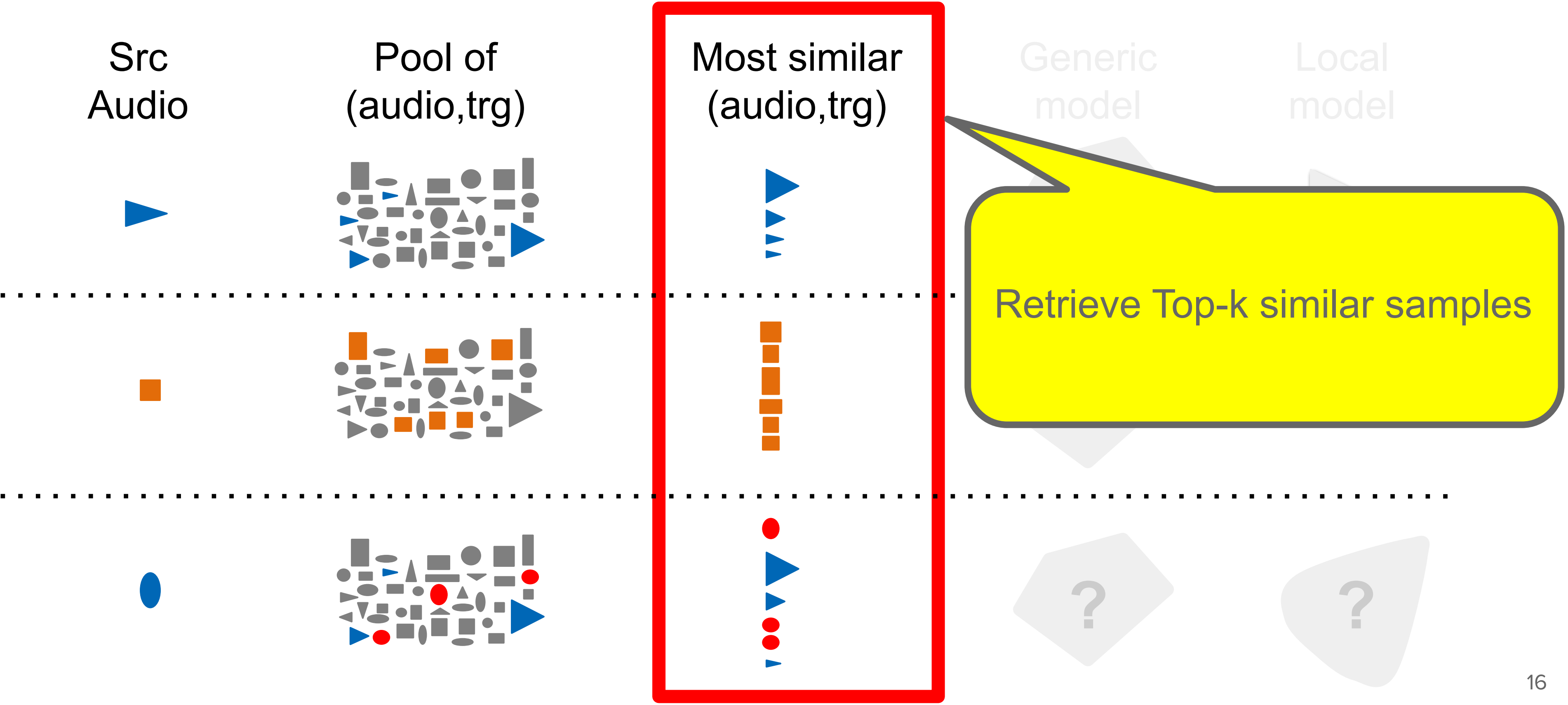
Pool of (audio,trg)

Most similar (audio,trg)

Generic model

Local model





Src Audio

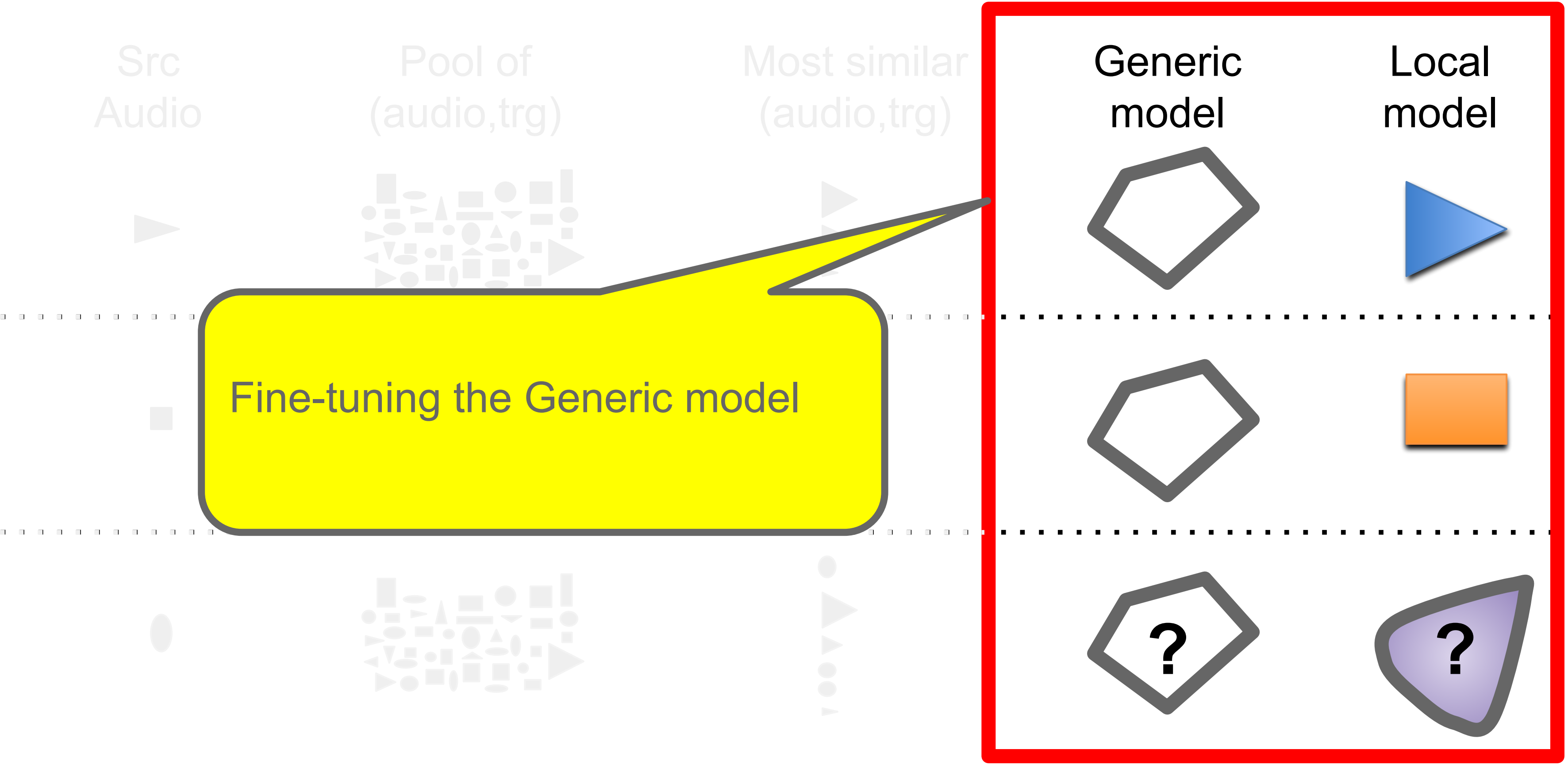
Pool of (audio,trg)

Most similar (audio,trg)

Generic model

Local model

Retrieve Top-k similar samples



Src Audio

Pool of (audio,trg)

Most similar (audio,trg)

Generic model

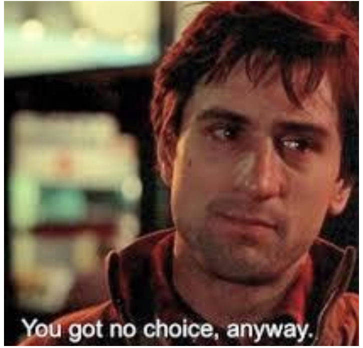
Local model

Fine-tuning the Generic model

IBMA for direct ST

- Source segment is **audio,** not **text.**
- Computing similarity is a research problem
- Similarity involves at least two dimensions:

1. Content similarity (*what* is said)
2. Voice similarity (*how* it is said)



Computing audio similarity

We propose a simple similarity for a proof of concept:

1. Input reduced to a fixed-size vector
   a. Input can be raw spectrogram (raw) or output of ST encoder (encoder features)
2. Similarity is computed with cosine distance
3. Cosine similarity lower than 0.5 is filtered out

Experiments

Data

MuST-C 8 languages (De, Es, Fr, It, Nl, Pt, Ro, Ru):

- 385-500 hours of speech
- TED talks

Di Gangi, Mattia A., et al. "MuST-C: a multilingual speech translation corpus." *NAACL* 2019.

How2 (En-Pt):

- 300 hours of speech
- Video tutorials downloaded from Youtube

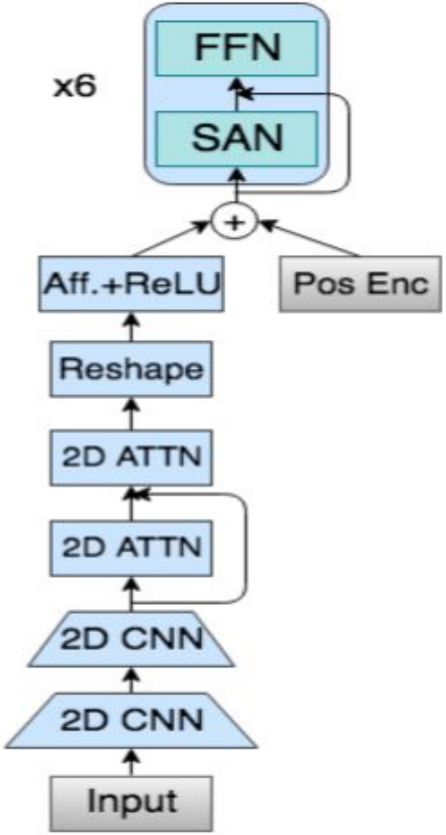
Sanabria, R., et al. "How2: a large-scale dataset for multimodal language understanding." *ICLR 2018.*

Experiments

1. IBMA within one dataset (MuST-C or How2)
2. Similarity check:
   a. Comparison between most and least similar pairs
3. Multi-domain experiments:
   a. MuST-C En-Pt + How2
   b. Different combinations of train and test domains

Model - S-Transformer

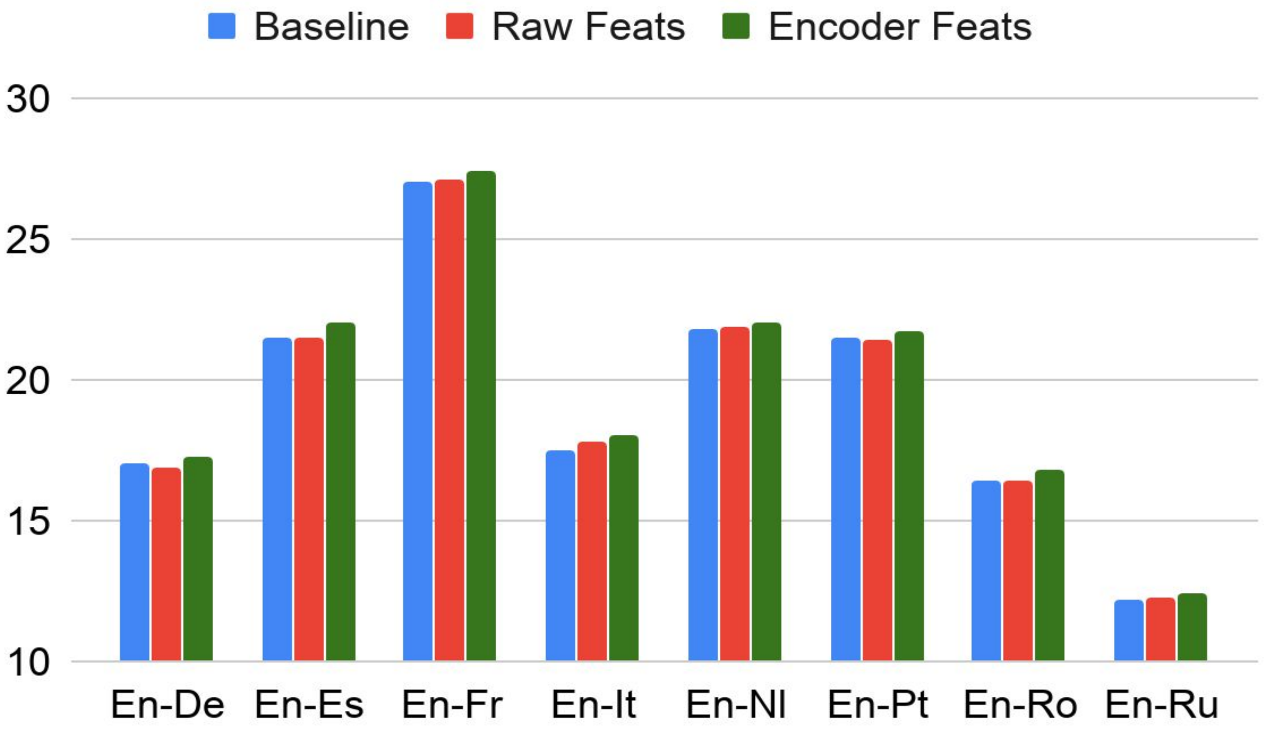
- Adaptation of Transformer to ST task
- Good results on MuST-C and How2
- Fast to train



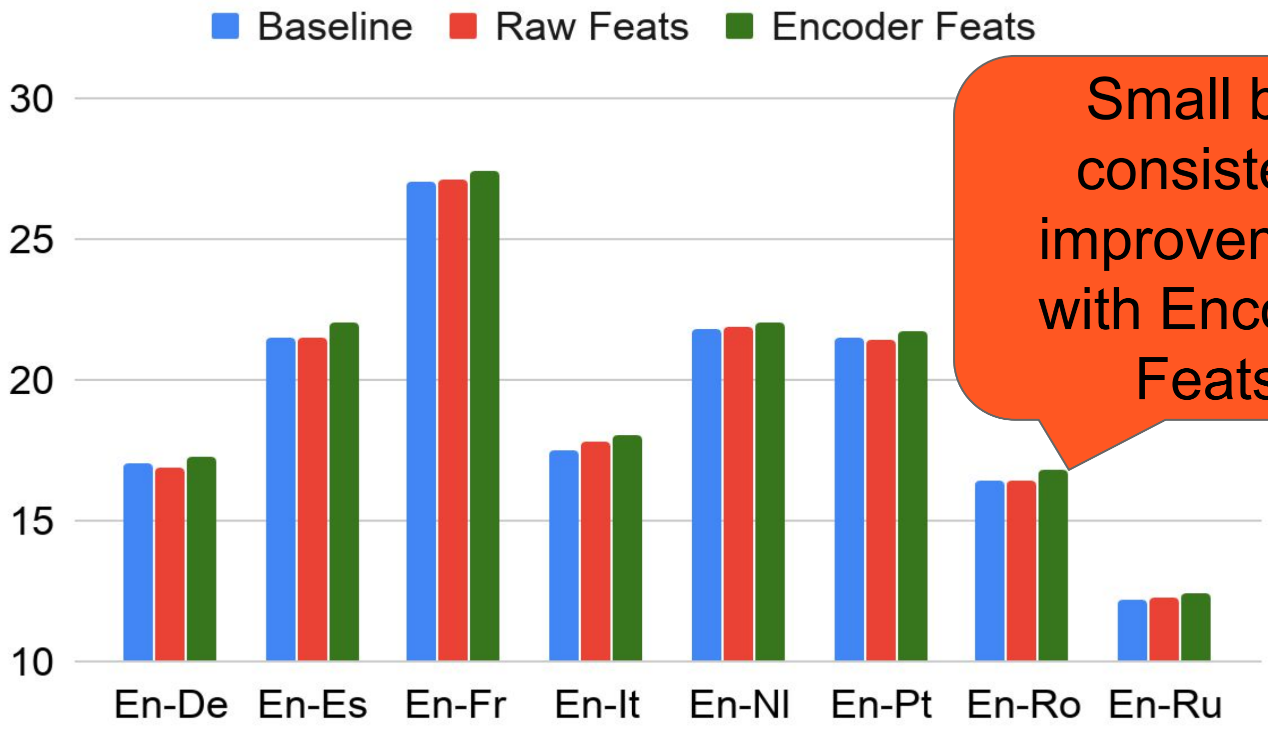
Di Gangi, M. A., et al. "Adapting transformer to end-to-end spoken language translation." ***INTERSPEECH 2019*****.**

Results

1) IBMA Within One Dataset - MuST-C

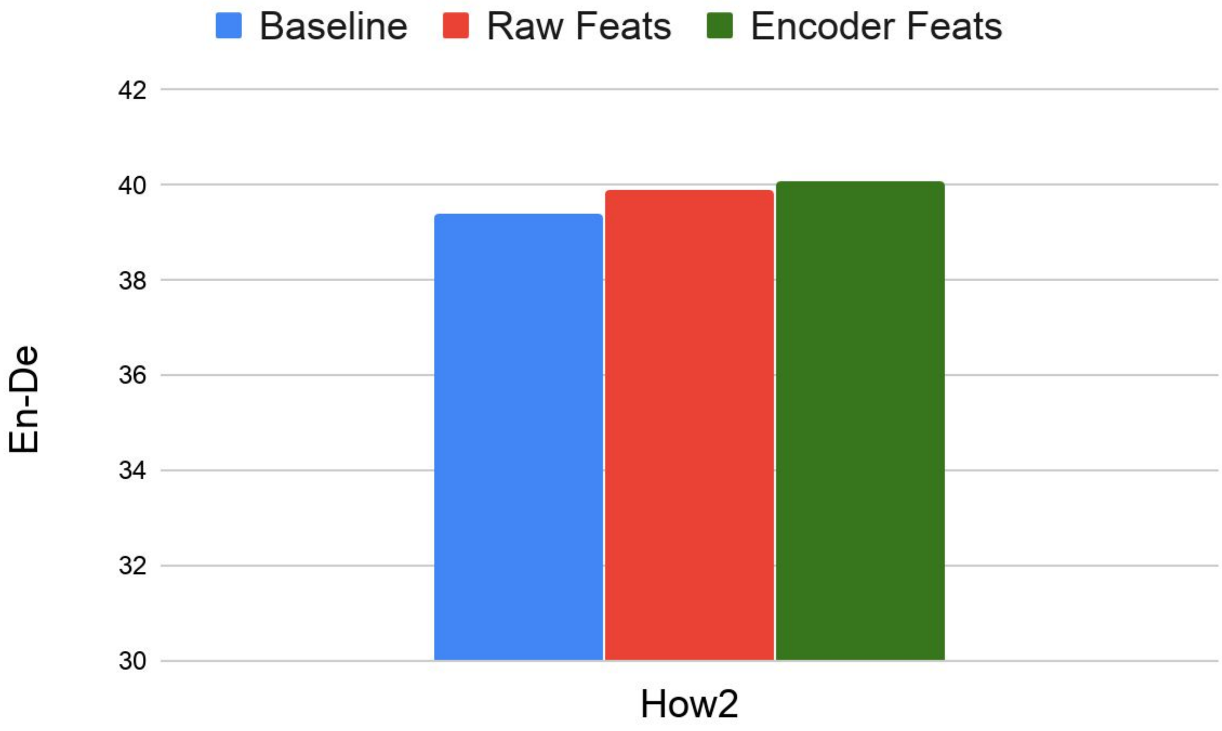


1) IBMA Within One Dataset - MuST-C

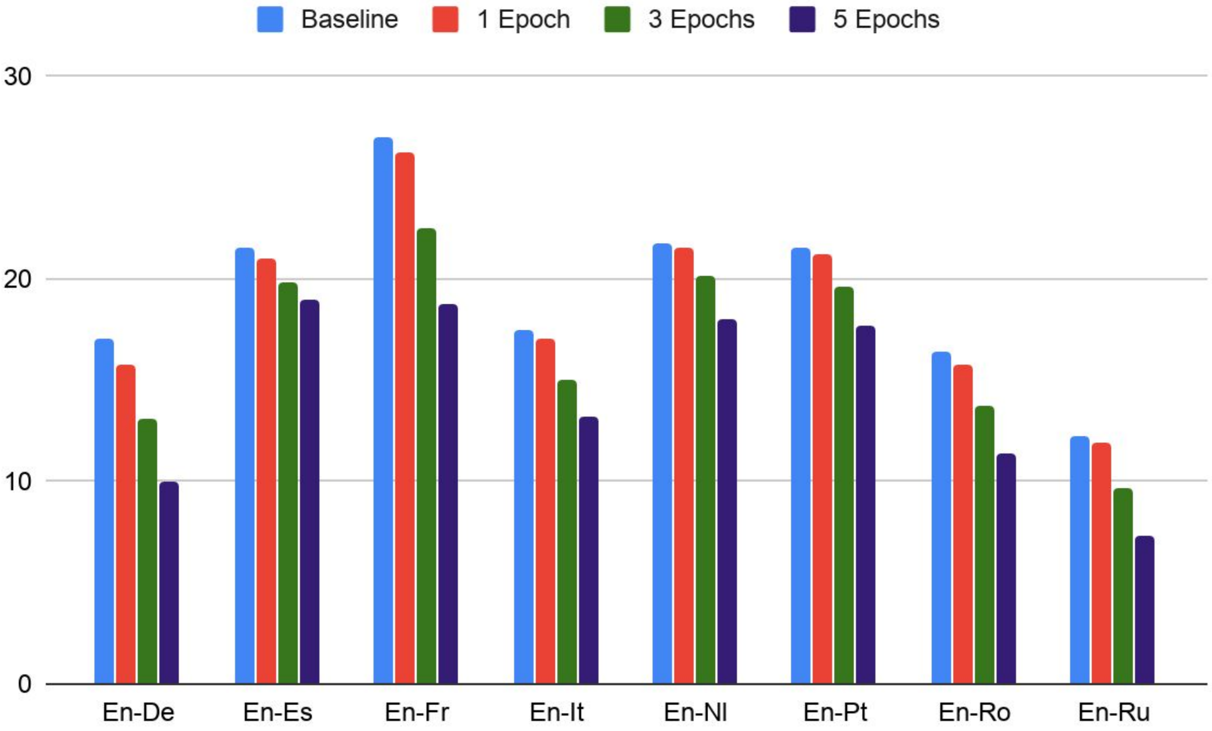


Small but consistent improvement with Encoder Feats

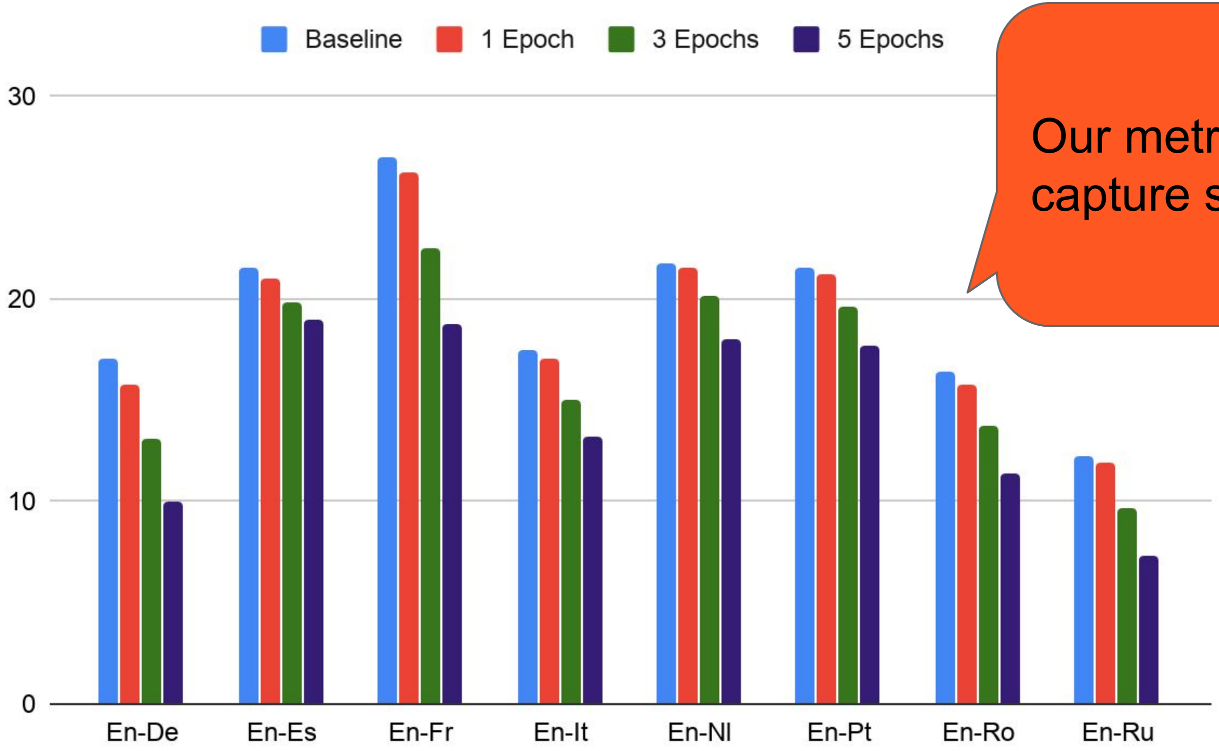
1) IBMA Within One Dataset - How2



2) Similarity Check

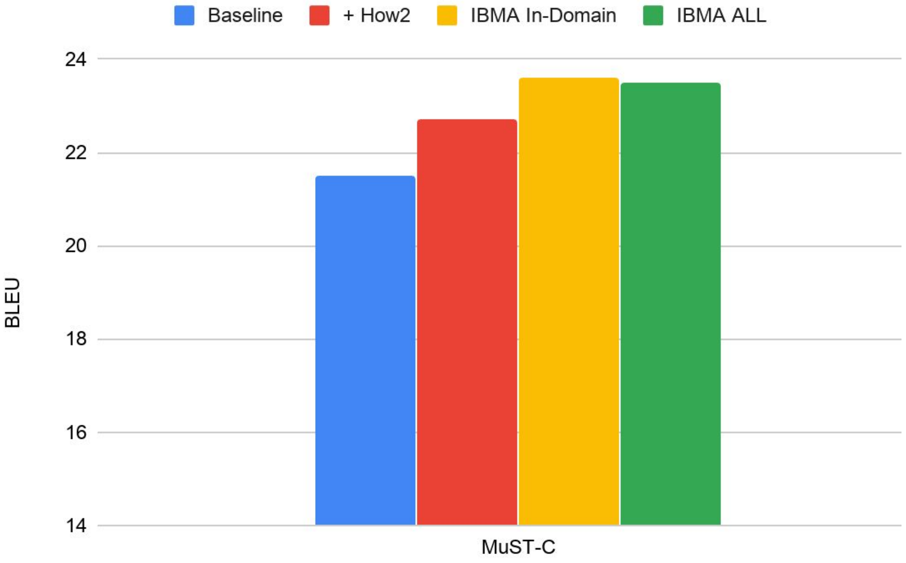


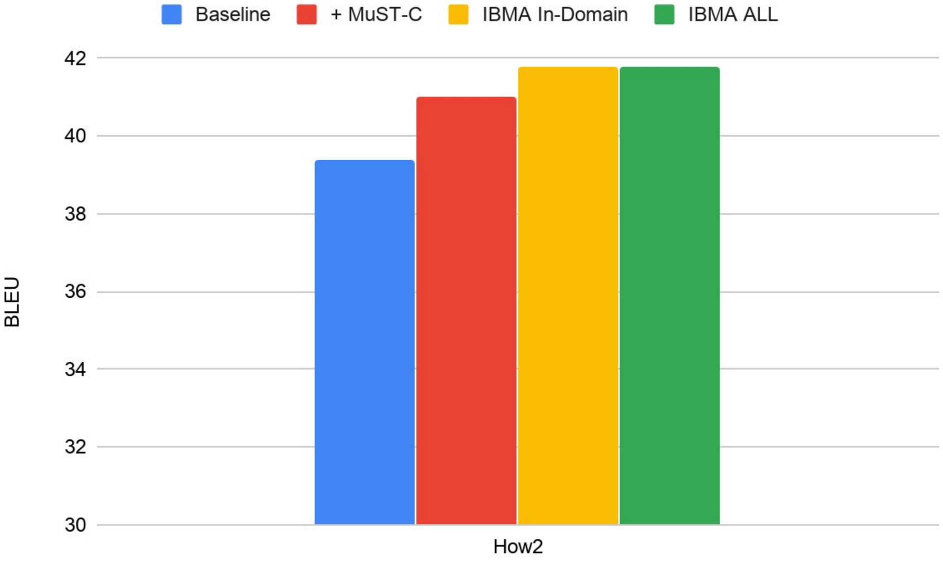
2) Similarity Check



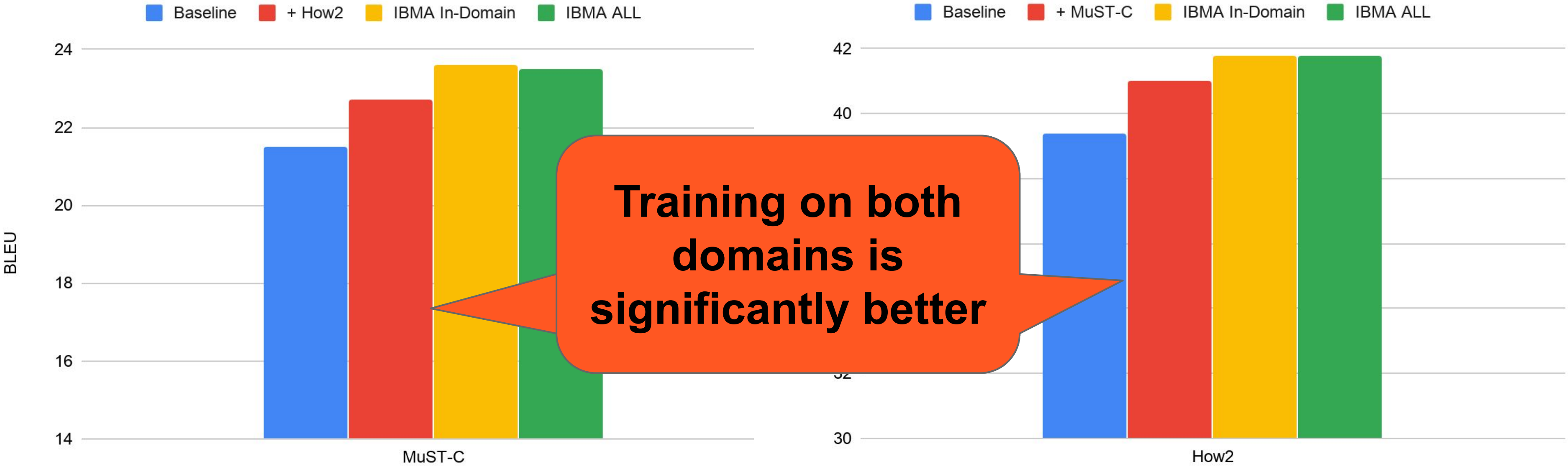
Our metric can capture similarity

3) Multi-Domain Experiments



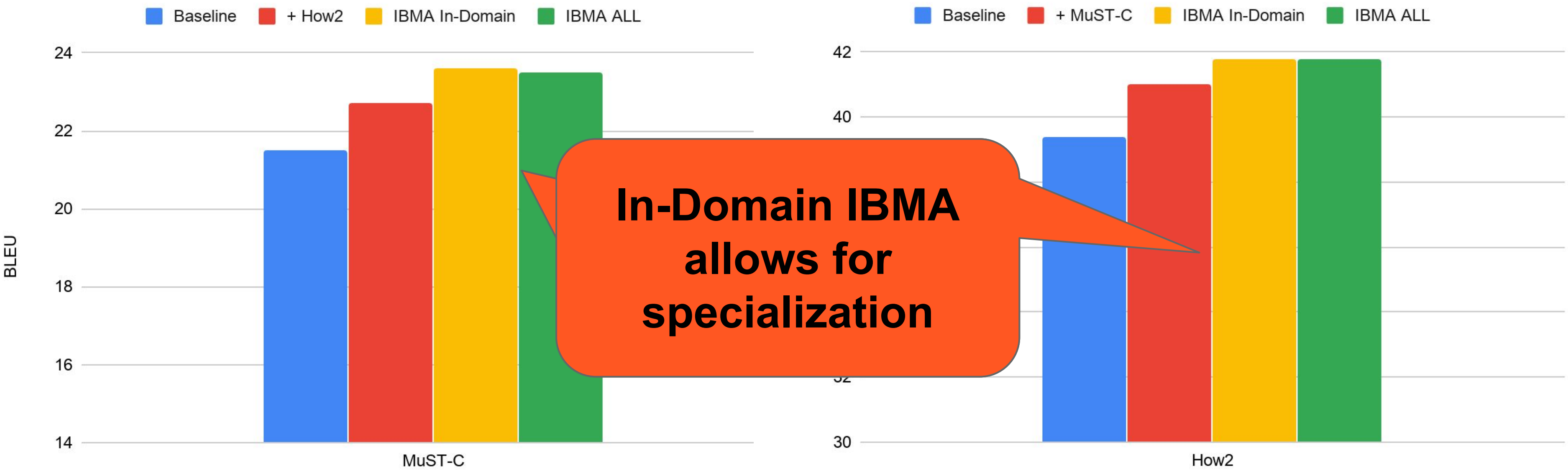


3) Multi-Domain Experiments



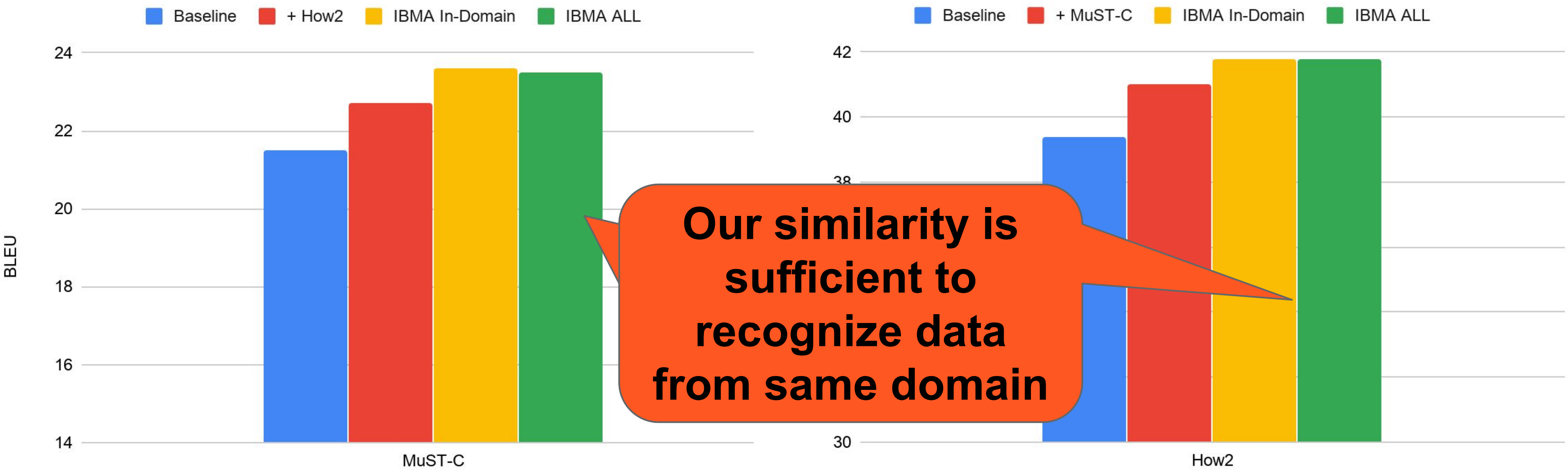
Training on both domains is significantly better

3) Multi-Domain Experiments



In-Domain IBMA allows for specialization

3) Multi-Domain Experiments



Our similarity is sufficient to recognize data from same domain

Findings

- IBMA provides small but consistent improvements
- Our similarity, though simple, can filter out unrelated samples
- In a multi-domain scenario, it can correctly identify samples from the same domain

Open Problems

- Audio similarity (ST) is a different beast compared to text similarity (MT): multi-faceted and more subtle
    - What are we capturing/exploiting when computing similarity?
    - Content (*what is said*) or voice (*how is said)*?
    - Which one is better?
    - Can we mix them for larger improvements?

Conclusions

- Data paucity is the main bottleneck in direct ST
- IBMA for “positive” model overfitting (performed on-the-fly!):
    - Audio-based similarity to retrieve training samples similar to the input sentence
    - Fine-tune on the retrieve samples & reset the model
- Small but consistent improvements on different language pairs
- New exciting open problems

FONDAZIONE BRUNO KESSLER

HLT MACHINE TRANSLATION

UNIVERSITY OF TRENTO - Italy

Instance-Based Model Adaptation For Direct Speech Translation

Mattia Antonino Di Gangi, Viet-Nhat Nguyen, Matteo Negri, Marco Turchi

{digangi,negri,turchi}@fbk.eu

Follow our work!

Medium

@fbk_mt @mdigangiPA

Machine Translation @ FBK @mattiadigangi