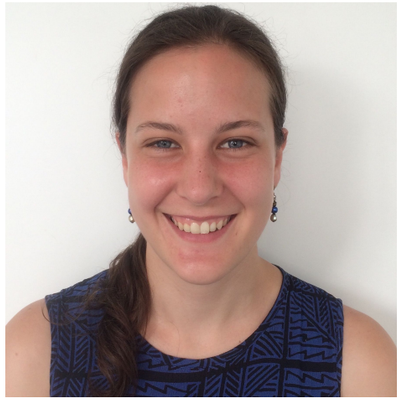
End-to-End Speech Translation

Jan Niehues, Elizabeth Salesky, Marco Turchi and Matteo Negri

EACL 2021

Speakers

Jan Niehues,
Maastricht University
jan.niehues@maastrichtuniversity.nl



Elizabeth Salesky,
Johns Hopkins University
esalesky@jhu.edu

Marco Turchi,
Fondazione Bruno Kessler
turchi@fbk.eu

Matteo Negri,
Fondazione Bruno Kessler
negri@fbk.eu

Outline

Sec 1: ***Introduction***	1.1: Task definition
	1.2: Challenges in translation of speech
	1.3: Traditional cascade approach
Sec 2: ***End-to-End***	2.1: State-of-the-art
	2.2: Input representations
	2.3: Architecture & modifications
	2.4: Output representations
Sec 3: ***Leveraging Data Sources***	3.1: Available data
	3.2: Techniques: Multi-task learning
	Transfer-learning & pretraining
	Knowledge distillation
	3.3: Alternate data representations
Sec 4: ***Evaluation***	4.1: Automatic Metrics
	4.2: Utterance segmentation
	4.3: Mitigating error – gender bias
Sec 5: ***Advanced Topics***	5.1: Utterance segmentation
	5.2: Multilingual ST
	5.3: Under-resourced languages
Sec 6: ***Real-world***	6.1: Automatic generation of subtitles
	6.2: Simultaneous translation
Sec 7: ***Conclusion***	🏆

Sec 1:

Introduction

Task definition

Challenges in translation of speech

Traditional cascade approaches

Sec 1.1

Task Definition

Speech Translation - Task

Speech input

= *Welcome to this tutorial*

ST system

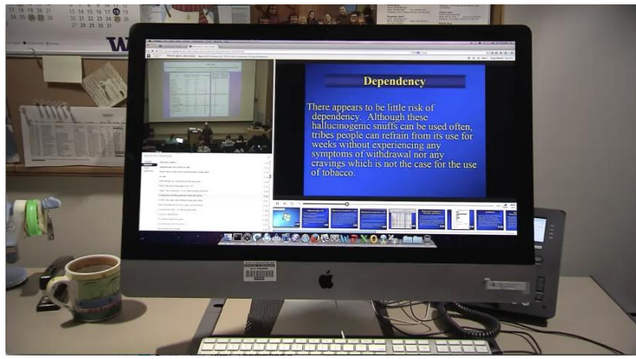
Willkommen zu diesem Tutorial

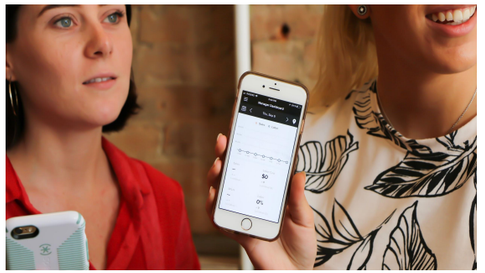
Textual translation

Spoken translation

Speech Translation - Motivation

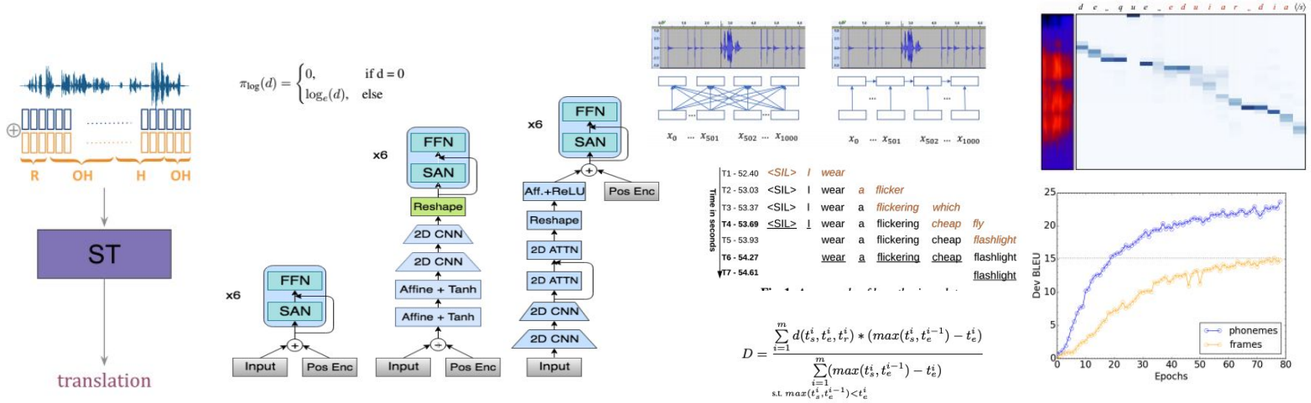
- Break language barriers to communicate, spread information and culture
  - Work
    - Meetings
  - Education and training
    - Lectures, conferences
  - Entertainment
    - Youtube, social media, cinema, tv
  - Everyday communication
    - Tourism, medical care, telephone conversations



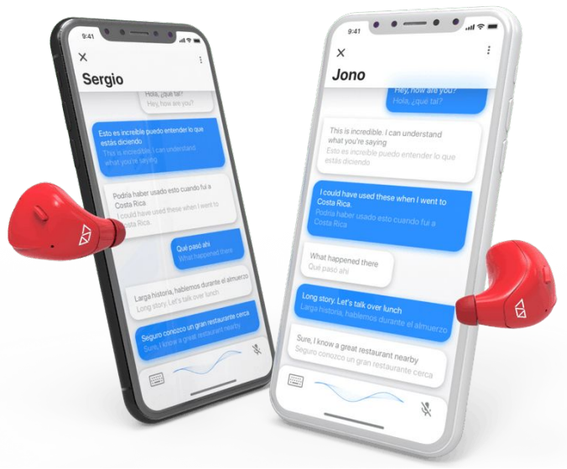


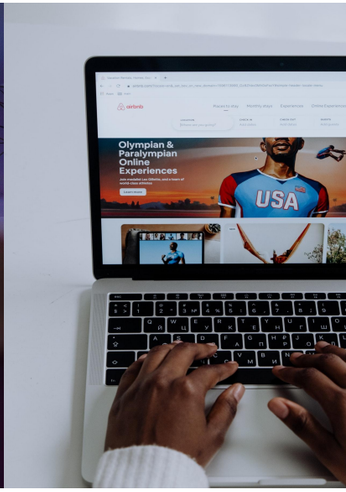
Speech Translation - Motivation

- Room for advanced research...
  - 99% of this tutorial



- ...and for applications
  - Wearable devices
  - Video subtitling
  - Live captioning
  - Human-machine communication





Speech Translation - History (before e2e)

Late ‘80s: first proofs of concept

Constraints to control language ambiguity (phonetics, syntax, semantics)

- Restricted vocabulary
- Controlled speaking style
- Narrow domain
- Offline processing

‘90s: Less constraints (vocabulary, speaking style)

First spontaneous ST systems (C-STAR, Verbmobil, Nespole,...)

2003-2006: Less constraints (domain)

First open-domain ST systems (STR-DUST, TC-STAR, GALE)

- different scenarios (broadcast news, parliamentary speeches, academic lectures)
- different languages (Zh, Ar, Es)

2006: Less constraints (operating conditions)

First simultaneous translator
(real-time translation of spontaneous lectures and presentations)

Speech Translation - History (the e2e era)

2005: first ST corpora

Small size/language coverage

2016-2017: first e2e ST models

(Duong et al., 2016, Berard et al., 2016, Weiss et al., 2017, ...)
encoder-decoder architectures based on RNNs

2018: first e2e models at IWSLT

8.7 BLEU points below cascade ST solutions on En-De

2019: significant gap reduction at IWSLT

1.6 BLEU points below cascade ST solutions on En-De

2019: ST adaptation of Transformer

(Di Gangi et al., 2019)

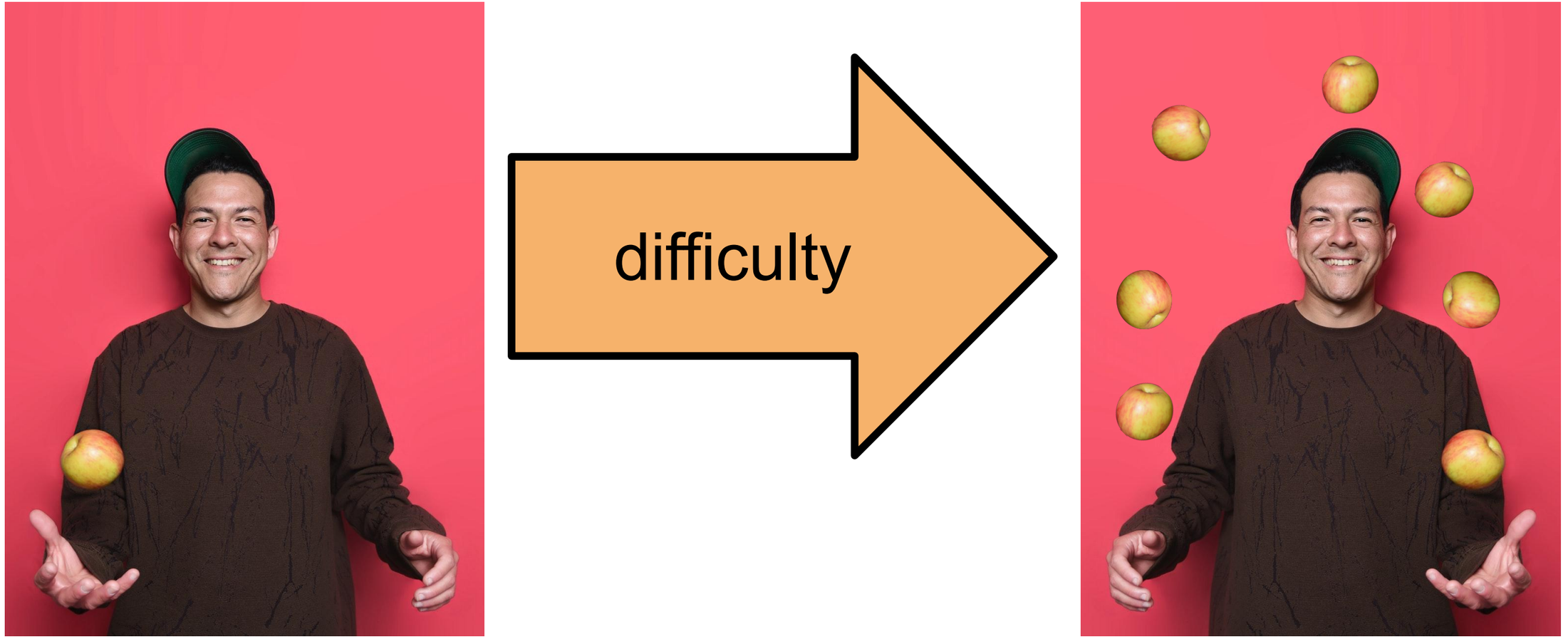
2019-2020: new ST corpora

Larger size/language coverage

2020: the gap almost closed?

+0.24 BLEU on unsegmented En-De test data

Speech Translation - a Multi-faceted Problem



Speech Translation - a Multi-faceted Problem

Offline

Single-speaker

Clean audio

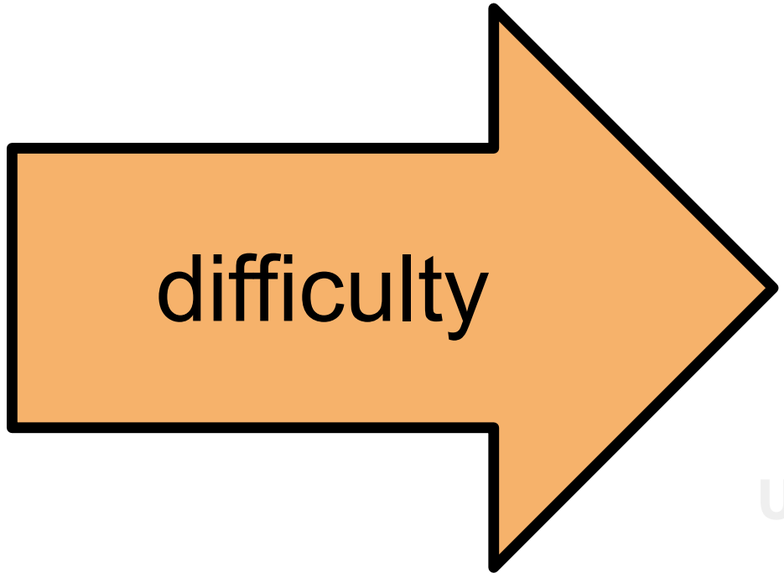
Restricted domain

Resource-rich languages

Low speaker variety (gender, accent, ...)

Unconstrained

...



Simultaneous

Multi-speaker

Noisy conditions

Open domain

Under-resourced languages

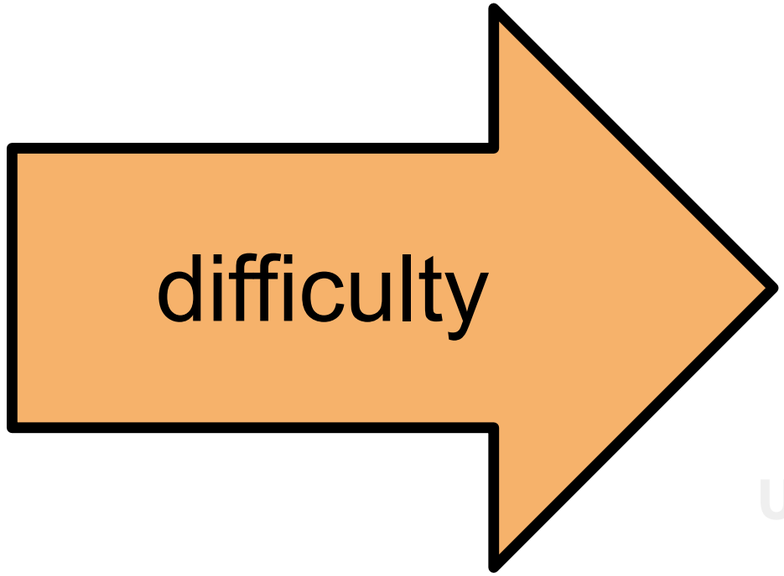
High speaker variety

Constrained (e.g. subtitling)

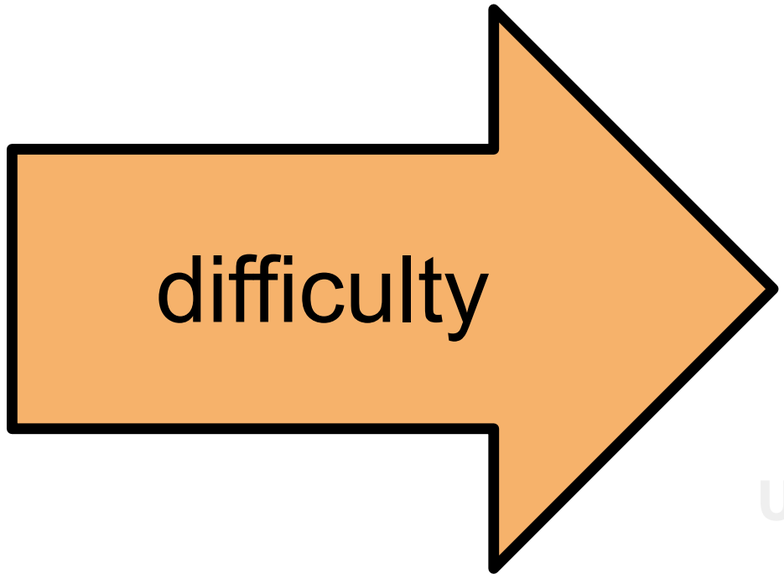
...

Speech Translation - a Multi-faceted Problem

Offline	Simultaneous
Single-speaker	Multi-speaker
Clean audio	Noisy conditions
Restricted domain	Open domain
Resource-rich languages	Under-resourced languages
Low speaker variety (gender, accent, ...)	High speaker variety
Unconstrained	Constrained (e.g. subtitling)
...	...

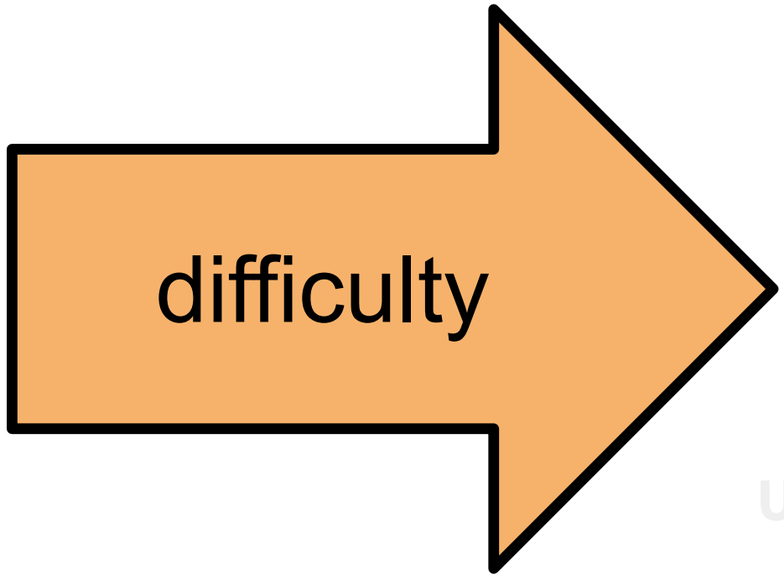


Speech Translation - a Multi-faceted Problem



Offline	Simultaneous
Single-speaker	Multi-speaker
Clean audio	Noisy conditions
Restricted domain	Open domain
Resource-rich languages	Under-resourced languages
Low speaker variety (gender, accent, ...)	High speaker variety
Unconstrained	Constrained (e.g. subtitling)
...	...

Speech Translation - a Multi-faceted Problem



Offline

Single-speaker

Clean audio

Restricted domain

Resource-rich languages

Low speaker variety (gender, accent, ...)

Unconstrained

...

Simultaneous

Multi-speaker

Noisy conditions

Open domain

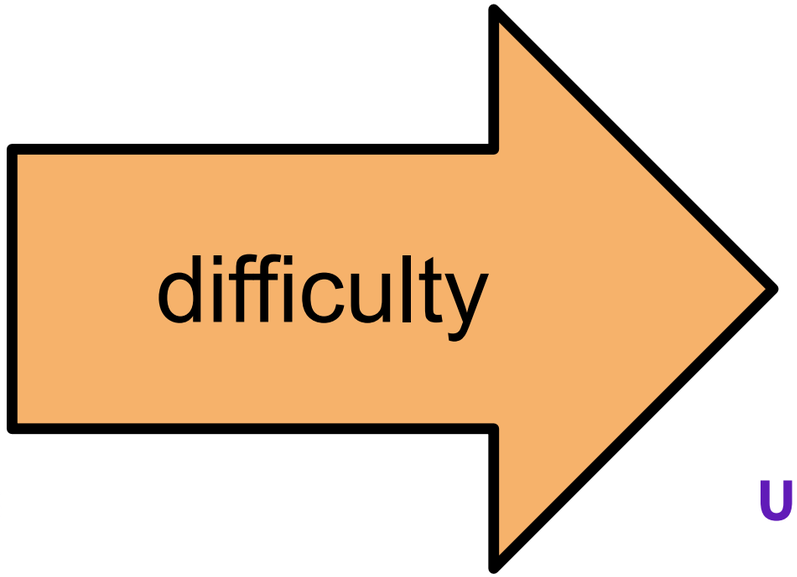
Under-resourced languages

High speaker variety

Constrained (e.g. subtitling)

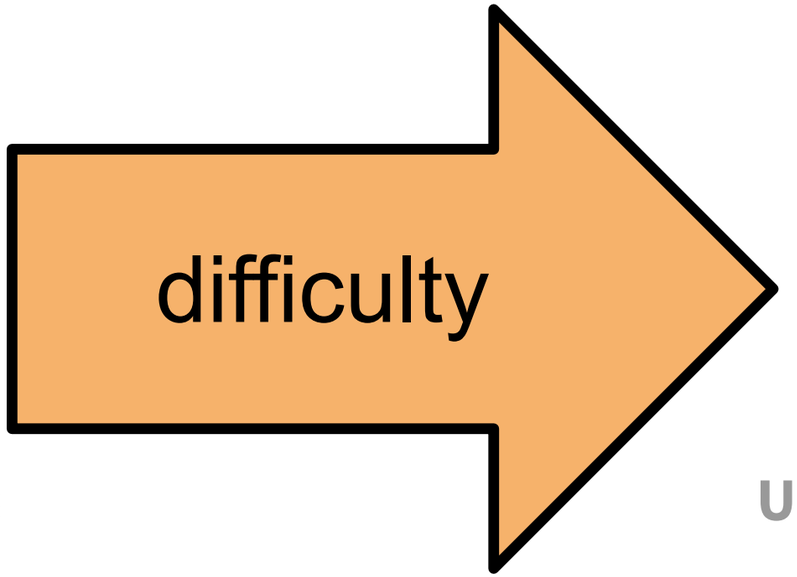
...

Speech Translation - a Multi-faceted Problem



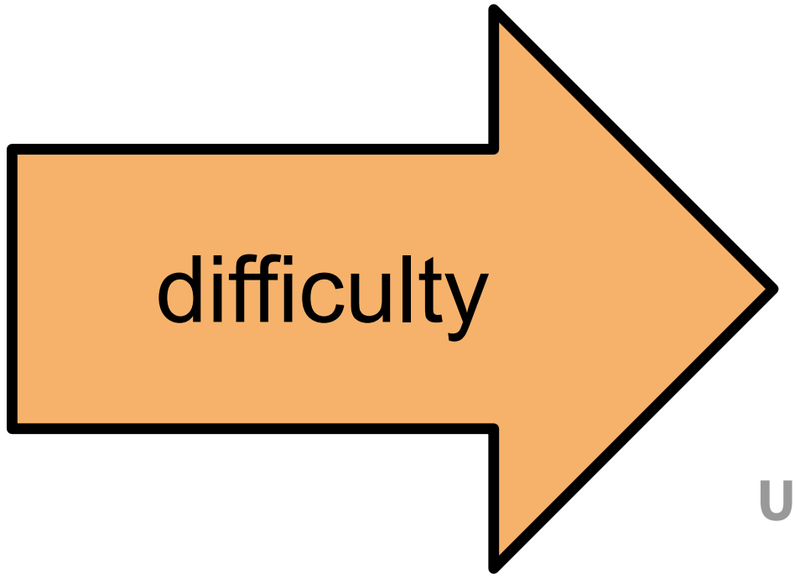
Offline	Simultaneous
Single-speaker	Multi-speaker
Clean audio	Noisy conditions
Restricted domain	Open domain
Resource-rich languages	**Under-resourced languages**
Low speaker variety (gender, accent, ...)	High speaker variety
Unconstrained	Constrained (e.g. subtitling)
...	...

Speech Translation - a Multi-faceted Problem



Offline	Simultaneous
Single-speaker	Multi-speaker
Clean audio	Noisy conditions
Restricted domain	Open domain
Resource-rich languages	Under-resourced languages
Low speaker variety (gender, accent, ...)	**High speaker variety**
Unconstrained	Constrained (e.g. subtitling)
...	...

Speech Translation - a Multi-faceted Problem



Offline	Simultaneous
Single-speaker	Multi-speaker
Clean audio	Noisy conditions
Restricted domain	Open domain
Resource-rich languages	Under-resourced languages
Low speaker variety (gender, accent, ...)	High speaker variety
Unconstrained	**Constrained (e.g. subtitling)**
...	...